

## RESEARCH ARTICLE OPEN ACCESS

# Detection of Dense Subhypergraphs by Low-Degree Polynomials

Abhishek Dhawan<sup>1</sup>  | Cheng Mao<sup>1</sup> | Alexander S. Wein<sup>2</sup>

<sup>1</sup>School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia, USA | <sup>2</sup>Department of Mathematics, University of California, Davis, California, USA

**Correspondence:** Abhishek Dhawan ([abhishekdhawan@gatech.edu](mailto:abhishekdhawan@gatech.edu))

**Received:** 8 May 2023 | **Revised:** 5 August 2024 | **Accepted:** 9 December 2024

**Funding:** A.D. was supported in part by NSF grant DMS-2053333. C.M. was supported in part by NSF grants DMS-2053333 and DMS-2210734. Part of this work was done while A.S.W. was with the Algorithms and Randomness Center at Georgia Tech, supported by NSF grants CCF-2007443 and CCF-2106444.

## ABSTRACT

Detection of a planted dense subgraph in a random graph is a fundamental statistical and computational problem that has been extensively studied in recent years. We study a hypergraph version of the problem. Let  $G^r(n, p)$  denote the  $r$ -uniform Erdős–Rényi hypergraph model with  $n$  vertices and edge density  $p$ . We consider detecting the presence of a planted  $G^r(n^\gamma, n^{-\alpha})$  subhypergraph in a  $G^r(n, n^{-\beta})$  hypergraph, where  $0 < \alpha < \beta < r - 1$  and  $0 < \gamma < 1$ . Focusing on tests that are degree- $n^{o(1)}$  polynomials of the entries of the adjacency tensor, we determine the threshold between the easy and hard regimes for the detection problem. More precisely, for  $0 < \gamma < 1/2$ , the threshold is given by  $\alpha = \beta\gamma$ , and for  $1/2 \leq \gamma < 1$ , the threshold is given by  $\alpha = \beta/2 + r(\gamma - 1/2)$ . Our results are already new in the graph case  $r = 2$ , as we consider the subtle *log-density regime* where hardness based on average-case reductions is not known. Our proof of low-degree hardness is based on a *conditional* variant of the standard low-degree likelihood calculation.

## 1 | Introduction

Finding a dense subgraph in a given undirected graph is an iconic problem at the intersection of graph theory, computer science, and statistics, and it finds broad applications in social and biological sciences. The last few decades have observed a wide range of research on multiple variants of the problem, including the planted clique problem [1, 2], densest  $k$ -subgraph problem [3, 4], community detection [5, 6], and hypergraph versions [7–9], among others. In this work, we focus on the problem of detecting the presence of a *planted dense subhypergraph* in a given  $r$ -uniform hypergraph  $H$  on  $n$  vertices. More precisely, an  $r$ -uniform hypergraph  $H$  is a tuple  $(V, E)$  of vertices and edges, where  $E$  is a collection of  $r$ -element subsets of  $V$ . In the language of statistical hypothesis testing, we consider the task of testing

between the following two distributions on  $H$  (defined formally in Section 2):

- the *null distribution* is the  $r$ -uniform Erdős–Rényi hypergraph  $G^r(n, q)$  with edge density  $q$ ;
- the *planted distribution* randomly selects  $\approx \rho n$  vertices to lie in the planted subhypergraph, the edges within the subhypergraph occur independently with probability  $p$ , and all other edges occur independently with probability  $q$ .

Given a single hypergraph  $H$  drawn from one of these two distributions, the goal is to distinguish the two cases with high probability (w.h.p.), that is, probability  $1 - o(1)$  as  $n \rightarrow \infty$ , where  $p, q, \rho$  may scale with  $n$ . We assume the parameters  $p, q, \rho$  are known.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2025 The Author(s). *Random Structures & Algorithms* published by Wiley Periodicals LLC.

Our focus is on understanding the power and limitation of *computationally efficient* tests, that is, polynomial-time algorithms. We currently lack complexity-theoretic tools to prove computational hardness of *average-case* problems like this one (where the input is random), so the leading approaches are either based on average-case reductions which formally relate different average-case problems to each other (see e.g., [10] and references therein) or unconditional lower bounds against restricted classes of algorithms.

Our main result establishes sharp necessary and sufficient conditions on  $p, q, \rho$  for success of *low-degree polynomial tests*. This is a powerful class of tests including statistics like small subgraph counts (edges, triangles, etc.). It is by now well-established that these low-degree tests are a useful proxy for computationally efficient tests, in the sense that the best-known polynomial-time algorithms for a wide variety of high-dimensional testing problems are captured by the low-degree class; see for example, [11, 12].

More specifically, we focus on the so-called *log-density* regime [3] where  $p = n^{-\alpha}$ ,  $q = n^{-\beta}$ , and  $\rho = n^{\gamma-1}$  for constants  $0 < \alpha < \beta < r-1$  and  $\gamma \in (0, 1)$ . The assumption  $\beta < r-1$  is a natural non-degeneracy condition: if  $\beta > r-1$  then the expected number of hyperedges incident to a fixed non-planted vertex is  $o(1)$ , so the vast majority of non-planted vertices will be isolated. We show that if  $0 < \gamma < 1/2$  and  $\alpha < \beta\gamma$ , or if  $1/2 \leq \gamma < 1$  and  $\alpha < \beta/2 + r(\gamma - 1/2)$ , there is a constant-degree polynomial test that distinguishes the null and the planted distribution w.h.p. On the other hand, if  $0 < \gamma < 1/2$  and  $\alpha > \beta\gamma$ , or if  $1/2 \leq \gamma < 1$  and  $\alpha > \beta/2 + r(\gamma - 1/2)$ , then no degree- $n^{o(1)}$  polynomial can separate the null and the planted distribution.

## 1.1 | Relation to Prior Work

### 1.1.1 | Planted Dense Subgraph

Our results are already interesting in the graph case  $r = 2$ , in which case we are considering the *planted dense subgraph detection* problem. The statistical threshold is well-studied [5, 13–15], and the computational threshold has also been established in various parameter regimes via reduction from the presumed-hard *planted clique* problem [14, 16, 17]. However, existing reductions assume that  $p$  and  $q$  are of the same order, and so we are currently lacking reduction-based evidence for hardness in the log-density regime where  $p = n^{-\alpha}$ ,  $q = n^{-\beta}$ , and  $\rho = n^{\gamma-1}$  for constants  $0 < \alpha < \beta < 1$  and  $\gamma \in (0, 1)$ . Here the problem undergoes a qualitative change at  $\gamma = 1/2$ . Namely, when  $\gamma \geq 1/2$ , the best-known polynomial-time algorithm is simply to threshold the total number of edges in the graph, and this succeeds when  $\alpha < \beta/2 + 2(\gamma - 1/2)$ . On the other hand, when  $\gamma < 1/2$ , the best-known polynomial-time algorithm is a more subtle subgraph-counting procedure which succeeds when  $\alpha < \beta\gamma$ ; see section 3.2 of [3] (they give only a proof sketch, as their main goal is to give guarantees for approximating the densest  $k$ -vertex subgraph in *worst-case* graphs).

Our main result shows that for every  $\gamma \in (0, 1)$ , low-degree tests cannot surpass the thresholds above, providing evidence for optimality of the existing algorithms. We also give a matching

low-degree testing upper bound, confirming that the above algorithms are captured by the low-degree framework; in the case  $\gamma < 1/2$ , our test is simpler than that of [3], although their algorithm can also *recover* the planted subgraph.

In contrast, the *statistical* (i.e., information-theoretic) threshold for detection (i.e., testing) is distinct from the computational threshold above, at least in some parameter regime s [5, 15]. Specifically, there is a gap between the statistical and computational thresholds whenever  $\gamma \in (0, 1/2]$ . Also, when  $\gamma \in (1/2, 2/3)$  there is a gap for some values of  $\beta$ . We will elaborate further in our discussion of the hypergraph case below.

While it is not our main focus, we note that the statistical and computational aspects of *recovering* a planted dense subgraph (i.e., identifying the planted vertices) are also well-studied [18–21]. Formally, the recovery problem is at least as hard as the detection problem (via a polynomial-time reduction in the style of [22, section 5.1]). Notably, when  $\gamma > 1/2$ , the recovery problem is *strictly* harder, that is, the computational thresholds for detection and recovery are different (at least for low-degree algorithms), with recovery requiring  $\alpha < \beta/2 + \gamma - 1/2$  [23]. When  $\gamma < 1/2$ , our result shows low-degree hardness of detection at the same threshold as the recovery algorithm of [3], so there is no *detection-recovery gap* in this regime, resolving a question left open by [23, section 2.4.1].

### 1.1.2 | Low-Degree Testing

A successful degree- $D$  test is a degree- $D$  multivariate polynomial in the input variables (in our case, the  $\binom{n}{r}$  hyperedge-indicator variables) whose real-valued output *separates* (see Definition 2.1) samples from the planted and null distributions. The idea to study this class of tests emerged from the line of work [11, 24–26]; see also [12] for a survey. Tests of degree  $O(\log n)$  are generally taken as a proxy for polynomial-time tests, as they capture leading algorithmic approaches such as spectral methods. Our upper bound will give a constant-degree test, which yields a polynomial-time algorithm for testing; our lower bound will rule out all tests of degree  $n^{o(1)}$ .

There is by now a standard method for proving low-degree testing bounds based on the *low-degree likelihood ratio* (see section 2.3 of [11]), which boils down to finding an orthonormal basis of polynomials with respect to the null distribution, and computing the expectations of these basis polynomials under the planted distribution. However, our setting is more subtle because (for  $\gamma < 1/2$ ) the second moment of the low-degree likelihood ratio diverges due to rare “bad” events under the planted distribution. We therefore need to carry out a conditional low-degree argument whereby the planted distribution is conditioned on some “good” event.

Conditional low-degree arguments of this kind have appeared before in a few instances [27, 28], but our argument differs on a technical level. Prior work chose to condition on an event that would seem to make direct computations with the orthogonal polynomials very complicated; to overcome this, they bound the conditional low-degree likelihood ratio in an indirect way by first relating it to a certain “low-overlap” second moment (also called

the *Franz-Parisi criterion* in [27]). In contrast our approach is more direct: we are careful to condition on an event for which computations with the orthogonal polynomials remain tractable.

### 1.1.3 | Integrality Gaps for Densest $k$ -Subgraph

Complementary to our results is a line of work on establishing integrality gaps for convex relaxations of densest  $k$ -subgraph [29–31] (and [30] also considers the extension to hypergraphs). Here the setting of interest is to find, in a *worst-case* graph, a subgraph induced on  $k$  vertices (for a given  $k$ ) whose density of edges is guaranteed to be within some factor of the densest such  $k$ -subgraph. The best-known polynomial-time approximation factor is  $\approx n^{1/4}$ , due to [3]. The integrality gap results mentioned above show that powerful families of convex relaxations fail to improve upon this.

In fact, the proofs of these integrality gaps show that the convex relaxations fail on certain random distributions such as the ones we study. The hard instance for approximation factor  $n^{1/4-\epsilon}$  is precisely our testing problem with parameters  $\alpha = 1/4 + \epsilon$ ,  $\beta = 1/2$ ,  $\gamma = 1/2$  (and so our result provides additional evidence that this approximation factor is unimprovable). The work on convex relaxations that is most relevant to our result is the recent result of [31] which was obtained independently from ours. We provide a detailed comparison below.

### 1.1.4 | Comparison to Independent Work on Sum-Of-Squares [31]

In concurrent and independent work, [31] proved that the *sum-of-squares (SoS) hierarchy* (a powerful family of semidefinite programming relaxations) at degree  $n^{\Omega(1)}$  fails to improve the approximation factor  $n^{1/4}$  discussed above. This is the strongest known result on integrality gaps for the densest  $k$ -subgraph problem. While they are conjectured to be closely related (see [11, 25]), we are not aware of any formal implications in either direction between failure of SoS and failure of low-degree tests.

More specifically, the authors of [31] consider the same hypothesis testing problem that we do (in the graph case  $r = 2$ ) and write down a particular SoS program that takes as input a sample from the null distribution  $G(n, q)$  and attempts to *refute* the existence of a  $(1 + o(1))pn$ -vertex subgraph with  $(1 + o(1))p(pn)^2/2$  edges. Note that a subgraph of this density exists in the planted distribution with high probability, and thus, if SoS succeeds at the refutation task (w.h.p.) this also gives an algorithm for the detection (i.e., testing) task with the same parameters  $p, q, \rho$ . The result of [31] shows that SoS *fails* to solve the refutation task when  $\gamma < 1/2$  and  $\alpha > \beta\gamma$ , matching the threshold in our result. This gives strong evidence for computational hardness of the *refutation* problem in this regime. However, this need not imply hardness of the *detection* problem (the focus of our work), since detection is a formally easier problem.

To illustrate the previous point, when  $\gamma > 1/2$  (a regime not covered by [31]) we expect an inherent gap between the detection and refutation tasks, with polynomial-time detection requiring  $\alpha < \beta/2 + 2(\gamma - 1/2)$  and polynomial-time refutation requiring

$\alpha < \beta/2 + \gamma - 1/2$ ; see Appendix A for more details. Therefore, in this regime, we expect that checking feasibility of the SoS program of [31] is a strictly suboptimal algorithm for the *detection* problem. Accordingly, one strength of our result is that it directly addresses the detection problem and captures the best-known computational threshold for all  $\gamma \in (0, 1)$ .

The above discussion warns that one should be careful when taking an SoS lower bound as evidence for hardness of *detection* (rather than refutation). However, in this case the situation is more subtle because the proof of [31] uses the *pseudo-calibration* approach [24], and so a key step in their analysis is closely related to the low-degree testing lower bounds that we prove. Namely they prove that the low-degree likelihood ratio  $L_{\leq D}$  (in their notation,  $\tilde{E}[1]$ ) is  $1 + o(1)$  with high probability over the null distribution (and like our result, this involves some conditioning arguments). Our approach is related but distinct: we show that the second moment of the *conditional* low-degree likelihood ratio (see Section 3) is  $1 + o(1)$ , and this is what implies our desired result (failure of weak separation). The proof of [31] also requires some additional steps, notably a rather involved and impressive analysis to show that their moment matrix is positive semidefinite.

### 1.1.5 | Planted Dense Subhypergraphs

Compared to planted dense subgraphs, the generalized problem for hypergraphs is less well-studied. For the worst-case densest subhypergraph problem [32], shows that it is hard to approximate the densest subhypergraph to within an  $n^\epsilon$  factor for a fixed  $\epsilon > 0$  based on a pseudorandomness assumption. In the line of research on integrality gaps discussed above, the result of [30] holds for  $r$ -uniform hypergraphs, establishing an integrality gap of  $\approx n^{(r-1)/4}$  for the Sherali-Adams hierarchy, which matches the log-density threshold. However, unlike the graph case  $r = 2$ , no polynomial-time approximation algorithm is known to match the above threshold for  $r \geq 4$ , and partial progress has been made for  $r = 3$  by [7].

Considering the same planted dense subhypergraph model as ours [9], studies both the detection and the recovery problem. For the detection problem with  $p = 1$  and  $q = 1/2$  (i.e., planted clique), it is proved that degree- $O(\log n)$  polynomial tests fail if  $\gamma < 1/2$ . This is complementary to our result as we consider  $p = n^{-\alpha}$  and  $q = n^{-\beta}$  with  $0 < \alpha < \beta < r - 1$ . For the recovery problem [9], focuses on the regime  $\gamma \geq 1/2$  and proposes an efficient algorithm succeeding when  $\alpha < \beta/2 + (r - 1)(\gamma - 1/2)$ . In addition, if  $\alpha > \beta/2 + (r - 1)(\gamma - 1/2)$ , it is shown that polynomials of degree  $\text{polylog}(n)$  fail to recover the planted dense subhypergraph. Note that this recovery threshold is different from the detection threshold  $\alpha = \beta/2 + r(\gamma - 1/2)$  proved in this work; therefore, these two results together generalize the detection-recovery gap in the regime  $\gamma \geq 1/2$  shown in [23] for  $r = 2$  to the case of hypergraphs.

Another line of research [33–35] concerns the statistical thresholds for planted dense subhypergraphs. In particular, in the same vein as the results from [5], it is shown in [35] that the optimal statistical threshold for detection is achieved either by a total degree test or a computationally inefficient scan test.

When the former dominates, the detection threshold naturally matches our result, but when the scan test performs better, there is a statistical-computational gap. To be more precise, the total degree test succeeds when  $\alpha < \alpha_{\deg} := \beta/2 + r(\gamma - 1/2)$  and the scan test succeeds when  $\alpha < \alpha_{\text{scan}} := \gamma(r - 1)$ . If  $\gamma \leq 1/2$ , the scan test strictly outperforms our condition  $\alpha < \beta\gamma$ , so there is a statistical-computational gap. Moreover, we also have a gap when  $1/2 < \gamma < r/(r + 1)$  and  $2r(\gamma - 1/2) < \beta < r - 2\gamma$  (where the first condition on  $\gamma$  ensures the second interval for  $\beta$  is nonempty), since in this case  $\alpha_{\text{scan}} > \alpha_{\deg}$  and  $\alpha_{\deg} < \beta$  (note that when  $\alpha_{\deg} \geq \beta$ , the total degree test succeeds for all  $\alpha < \beta$  and there is no gap).

In addition [8, 36], study the recovery of a planted dense subhypergraph via a tensor PCA model with additive Gaussian noise. They probe the computational threshold of the problem using Approximate Message Passing, but the result is not directly comparable to ours. Finally, we refer the reader to a recent survey [37, section 5.7] for more related works on dense subhypergraphs.

## 1.2 | Notation

Let  $\mathbb{N}$  denote the set of positive integers. For any  $n \in \mathbb{N}$ , let  $[n] := \{1, 2, \dots, n\}$ . Throughout this work, we consider  $n \rightarrow \infty$  and use the asymptotic notation  $O(\cdot)$ ,  $o(\cdot)$ , and so forth.

For a fixed integer  $r \geq 2$ , let  $K_n^r$  denote the complete  $r$ -uniform hypergraph on  $n$  vertices. For any hypergraph  $H$ , let  $V(H)$  denote its vertex set and let  $E(H)$  denote its edge set. For a hypergraph  $H$  and  $S \subseteq E(H)$ , let  $H[S]$  denote the subgraph of  $H$  induced by  $S$ . Note that any set of hyperedges  $S \subseteq E(K_n^r)$  can be identified with the subhypergraph  $K_n^r[S]$ , so we sometimes view  $S$  as a subhypergraph without ambiguity. For brevity, we often refer to hypergraphs and hyperedges simply as graphs and edges, respectively.

For a distribution  $\mathcal{P}$ , let  $\mathbb{E}_{\mathcal{P}}$  denote the expectation under  $\mathcal{P}$ , and with slight abuse of notation, let  $\mathcal{P}$  also denote the associated probability. Let  $\text{Var}_{\mathcal{P}}(\cdot)$  denote the variance of a random variable under  $\mathcal{P}$ . Let  $\text{Ber}(p)$  denote the Bernoulli distribution with parameter  $p \in [0, 1]$ , and let  $\text{Bin}(n, p)$  denote the Binomial distribution with parameters  $n \in \mathbb{N}$  and  $p \in [0, 1]$ .

## 2 | Main Results

Let  $n, r \in \mathbb{N}$  with  $n \geq r \geq 2$ , and let  $0 < \alpha < \beta < r - 1$  and  $\gamma \in (0, 1)$ . Define  $M := \binom{[n]}{r}$ . We will be considering random  $n$ -vertex  $r$ -uniform hypergraph models where the output is an undirected hypergraph  $H$  with adjacency tensor  $Y \in \{0, 1\}^M$ . To be more precise, a hyperedge  $e$  is a subset of  $[n]$  of cardinality  $r$ ; we let  $[M] := \binom{[n]}{r}$  denote the set of all hyperedges of  $K_n^r$  and write  $e \in [M]$ . The adjacency tensor  $Y$  is indexed by  $e \in [M]$ , and  $Y_e \in \{0, 1\}$  indicates the presence of hyperedge  $e$ .

We formulate the detection of a dense random subhypergraph as a statistical hypothesis testing problem between distributions  $\mathcal{P}$  and  $\mathcal{Q}$ , defined as follows:

- Under  $\mathcal{P}$ , let  $z_1, \dots, z_n$  be i.i.d. random variables from the Bernoulli distribution  $\text{Ber}(\rho)$ , where  $\rho := n^{\gamma-1}$ , and let

$Z := \{i \in [n] : z_i = 1\}$ . Conditional on  $Z$ , we observe an  $r$ -uniform hypergraph  $H$  with independent hyperedges

$$Y_e \sim \begin{cases} \text{Ber}(q) & \text{if } e \not\subseteq Z \\ \text{Ber}(p) & \text{if } e \subseteq Z \end{cases}$$

where  $p := n^{-\alpha}$  and  $q := n^{-\beta}$ .

- Under  $\mathcal{Q}$ , we observe an  $r$ -uniform Erdős–Rényi hypergraph  $H$  with adjacency tensor  $Y$ , where the hyperedges are i.i.d.  $Y_e \sim \text{Ber}(q)$  with  $q := n^{-\beta}$ .

There are two points in the above definition that can be relaxed. First, in the definition of  $\mathcal{P}$ , instead of including each of the  $n$  vertices with probability  $\rho$  in the set  $Z$ , we could define  $Z$  to be a uniformly random subset of  $[n]$  of a fixed size  $k = n\rho$ . Our main results would still hold under this alternative definition of  $Z$ . We choose the current definition for technical convenience. Second, our results would remain valid if we assumed  $\rho = n^{\gamma-1+o(1)}$ ,  $p = n^{-\alpha+o(1)}$ , and  $q = n^{-\beta+o(1)}$  with extra  $o(1)$  terms that vanish as  $n \rightarrow \infty$  in the exponents. This will be clear from the proofs and we omit the  $o(1)$  terms for brevity.

To probe the computational threshold for testing between  $\mathcal{P}$  and  $\mathcal{Q}$ , we focus on low-degree polynomial algorithms (e.g., [11, 12]). Let  $\mathbb{R}[Y]_{\leq D}$  denote the set of multivariate polynomials in the entries of  $Y$  with degree at most  $D$ . With some abuse of notation, we will often say “a polynomial” to mean a sequence of polynomials  $f = f_n \in \mathbb{R}[Y]_{\leq D}$ , one for each problem size  $n$ ; the degree  $D = D_n$  of such a polynomial may scale with  $n$ . To study the ability of a polynomial in testing  $\mathcal{P}$  against  $\mathcal{Q}$ , we consider the notions of *strong separation* and *weak separation* defined in [27], with the former being stronger than the latter.

**Definition 2.1.** ([27], Definition 1.6). As  $n \rightarrow \infty$ , a polynomial  $f \in \mathbb{R}[Y]_{\leq D}$  is said to

- strongly separate  $\mathcal{P}$  and  $\mathcal{Q}$  if  $\sqrt{\text{Var}_{\mathcal{P}}(f(Y)) \vee \text{Var}_{\mathcal{Q}}(f(Y))} = o(|\mathbb{E}_{\mathcal{P}}[f(Y)] - \mathbb{E}_{\mathcal{Q}}[f(Y)]|)$ ;
- weakly separate  $\mathcal{P}$  and  $\mathcal{Q}$  if  $\sqrt{\text{Var}_{\mathcal{P}}(f(Y)) \vee \text{Var}_{\mathcal{Q}}(f(Y))} = O(|\mathbb{E}_{\mathcal{P}}[f(Y)] - \mathbb{E}_{\mathcal{Q}}[f(Y)]|)$ ,

where we use the notation  $a \vee b := \max\{a, b\}$ .

See [27] for a detailed discussion on why these conditions are natural for hypothesis testing. In particular, by Chebyshev’s inequality, strong separation implies that we can threshold  $f(Y)$  to test  $\mathcal{P}$  against  $\mathcal{Q}$  with vanishing type I and type II errors. Our main results are the following.

**Theorem 2.1.** Suppose that we observe a random  $n$ -vertex  $r$ -uniform hypergraph  $Y \in \{0, 1\}^M$  from either  $\mathcal{P}$  or  $\mathcal{Q}$  with parameters  $p = n^{-\alpha}$ ,  $q = n^{-\beta}$ , and  $\rho = n^{\gamma-1}$  for fixed  $r \geq 2$ ,  $0 < \alpha < \beta < r - 1$ ,  $\gamma \in (0, 1)$ ,  $n \rightarrow \infty$ , and  $M = \binom{[n]}{r}$ .

- (Lower bound) Suppose that either (1)  $\gamma \geq 1/2$  and  $\alpha > \beta/2 + r(\gamma - 1/2)$ , or (2)  $\gamma < 1/2$  and  $\alpha > \beta\gamma$ . If  $D = n^{o(1)}$  then no polynomial in  $\mathbb{R}[Y]_{\leq D}$  weakly separates  $\mathcal{P}$  and  $\mathcal{Q}$ .
- (Upper bound) Suppose that either (1)  $\gamma \geq 1/2$  and  $\alpha < \beta/2 + r(\gamma - 1/2)$ , or (2)  $\gamma < 1/2$  and  $\alpha < \beta\gamma$ . There exists a positive



integer  $D$  depending only on  $(\alpha, \beta, \gamma)$  and a polynomial in  $\mathbb{R}[Y]_{\leq D}$  that strongly separates  $\mathcal{P}$  and  $\mathcal{Q}$ .

We have therefore completely characterized the low-degree detection threshold for both  $\gamma \geq 1/2$  and  $\gamma < 1/2$ . We choose not to consider the boundary cases such as  $\alpha = \beta$  to simplify the presentation, but the method of low-degree polynomials can be used in these cases too. The lower bound in Theorem 2.1 is proved at the beginning of Section 3. The upper bound is proved at the beginning of Section 4. In particular, in each regime stated in the upper bound, the strong separation is achieved by a constant-degree polynomial in the entries of  $Y$ , so the testing algorithm is polynomial-time.

### 3 | Detection Lower Bound

Following the framework of low-degree polynomial algorithms [11, 12, 27], we first introduce the notation used in this section. Let  $\tilde{Y}_e$  be the standardized hyperedge under  $\mathcal{Q}$ , that is,

$$\tilde{Y}_e := \frac{Y_e - q}{\sigma} \quad (1)$$

where  $\sigma := \sqrt{\text{Var}_{\mathcal{Q}}(Y_e)} = \sqrt{q(1-q)}$ . We define

$$\phi_S(Y) := \prod_{e \in S} \tilde{Y}_e, \quad \forall S \subseteq [M] \quad (2)$$

Note that  $Y_e$  and  $Y_f$  are independent under  $\mathcal{Q}$  for  $e \neq f$ . Therefore, as  $\tilde{Y}_e$  is centered, we have

$$\mathbb{E}_{\mathcal{Q}}[\phi_S(Y)\phi_{S'}(Y)] = \begin{cases} 0 & \text{if } S \neq S', \\ 1 & \text{if } S = S', \end{cases}$$

so that  $\{\phi_S : S \subseteq [M]\}$  is an orthonormal basis of the set of real-valued functions of  $Y$ .

In order to study polynomials of degree at most  $D > 0$ , we define the following quantity

$$\|L_{\leq D}\|^2 := \sum_{\substack{S \subseteq [M], \\ |S| \leq D}} (\mathbb{E}_{\mathcal{P}}[\phi_S(Y)])^2 \quad (3)$$

When considering an event  $\mathcal{E}$  and the conditional distribution  $\mathcal{P}'$  on  $\mathcal{E}$ , we define analogously

$$\|L'_{\leq D}\|^2 := \sum_{\substack{S \subseteq [M], \\ |S| \leq D}} (\mathbb{E}_{\mathcal{P}'}[\phi_S(Y)])^2 \quad (4)$$

The notation  $\|L_{\leq D}\|$  stands for the norm of the degree- $D$  likelihood ratio between  $\mathcal{P}$  and  $\mathcal{Q}$ . We remark that  $\|L_{\leq D}\|$  is the same as the quantity  $\text{LD}(D)$  in [27, Definition 1.3], and the equivalence is justified in [11, section 2.3].

**Proof of Theorem 2.1. (lower bound).** Let  $\mathcal{E}$  be an event such that  $\mathcal{P}(\mathcal{E}) = 1 - o(1)$ . Let  $\mathcal{P}'$  be the distribution obtained from  $\mathcal{P}$  by conditioning on the event  $\mathcal{E}$ . By [27, Proposition 6.2], to prove that no polynomial in  $\mathbb{R}[Y]_{\leq D}$  weakly separates  $\mathcal{P}$  and  $\mathcal{Q}$ , it suffices to show that

$$\|L'_{\leq D}\|^2 = 1 + o(1) \quad (5)$$

If  $\gamma \geq 1/2$  and  $\alpha > \beta/2 + r(\gamma - 1/2)$ , we take  $\mathcal{E}$  to be the full sample space so that  $\mathcal{P} = \mathcal{P}'$  and establish (5) in Proposition 3.1. If  $\gamma < 1/2$  and  $\alpha > \beta\gamma$ , we define the high-probability event  $\mathcal{E}$  in (11) and establish (5) in Proposition 3.2.

**Remark 3.1.** We attempt to briefly explain why conditioning is needed. The quantity  $\|L_{\leq D}\|$  is equal to the supremum over degree- $D$  polynomials  $f$  of the ratio  $\mathbb{E}_{\mathcal{P}}[f(Y)]/\sqrt{\mathbb{E}_{\mathcal{Q}}[f(Y)^2]}$ . While boundedness of this ratio is a sufficient condition to rule out degree- $D$  separation, it may not be a necessary condition: there can be polynomials that achieve a large value for this ratio (diverging to  $\infty$ ) yet do not work as distinguishers. For a simple example, consider the planted dense subgraph model ( $r = 2$ ) with  $q = 0$ ,  $p = n^{-\alpha}$ , and  $\rho = n^{\gamma-1}$ , and let  $f$  be the degree- $\binom{m}{2}$  polynomial that counts the number of  $m$ -vertex cliques in the graph, for some constant  $m$ . Now the ratio is infinite since  $\mathbb{E}_{\mathcal{Q}}[f(Y)^2] = 0$  while  $\mathbb{E}_{\mathcal{P}}[f(Y)] > 0$ ; however, for  $\gamma m < \alpha \binom{m}{2}$ ,  $f$  is not a good distinguisher because  $\mathbb{E}_{\mathcal{P}}[f(Y)] \ll 1$  and so  $f = 0$  with high probability under both the planted and null distributions. This type of phenomenon can persist in cases where  $q$  is nonzero, including cases where the testing problem is low-degree hard. One can imagine “fixing” this issue by conditioning  $\mathcal{P}$  on the high-probability event that there are no  $m$ -cliques. This is essentially what we will need to do, but for all “problematic” subgraphs, not just cliques.

#### 3.1 | Lower Bound for Large $\gamma$

We will first consider the case that  $\gamma \geq 1/2$ , that is, the case that the set  $Z$  is expected to be “large”.

**Proposition 3.1.** Suppose  $p = n^{-\alpha}$ ,  $q = n^{-\beta}$ , and  $\rho = n^{\gamma-1}$  with fixed  $r \geq 2$ ,  $0 < \alpha < \beta < r - 1$ , and  $\gamma \in (0, 1)$  such that

$$\gamma \geq \frac{1}{2}, \quad \alpha > \frac{\beta}{2} + r\left(\gamma - \frac{1}{2}\right). \quad (6)$$

If  $D = n^{o(1)}$ , then we have

$$\|L_{\leq D}\|^2 = 1 + o(1).$$

Let us consider an arbitrary subgraph  $S \subseteq [M]$ . We first compute  $\mathbb{E}_{\mathcal{P}}[\phi_S(Y)]$ .

**Lemma 3.1.** For  $S \subseteq [M]$ , let  $\phi_S(Y)$  be as defined in (2), and let  $V(S) \subseteq [n]$  denote the vertex set of the hypergraph induced by  $S$ . Then we have

$$\mathbb{E}_{\mathcal{P}}[\phi_S(Y)] = \rho^{|V(S)|} \left(\frac{p-q}{\sigma}\right)^{|S|}.$$

**Proof.** We note the following:

$$\mathbb{E}_{\mathcal{P}}[\phi_S(Y)] = \mathbb{E}_Z[\mathbb{E}_{\mathcal{P}}[\phi_S(Y)|Z]].$$

We note that  $Y_e$  and  $Y_f$  are conditionally independent for  $e \neq f$ , given  $Z$ . In particular, we have

$$\mathbb{E}_{\mathcal{P}}[\phi_S(Y)|Z] = \prod_{e \in S} \mathbb{E}_{\mathcal{P}}[\tilde{Y}_e|Z].$$

Furthermore, by definition of  $\mathcal{P}$ , we have

$$\mathbb{E}_P[\tilde{Y}_e | Z] = \begin{cases} \frac{p-q}{\sigma} & e \subseteq Z \\ 0 & e \not\subseteq Z \end{cases}$$

With this in hand, we have

$$\mathbb{E}_P[\phi_S(Y) | Z] = \left( \frac{p-q}{\sigma} \right)^{|S|} \mathbb{1}\{V(S) \subseteq Z\},$$

from where we get

$$\mathbb{E}_P[\phi_S(Y)] = \rho^{|V(S)|} \left( \frac{p-q}{\sigma} \right)^{|S|},$$

as desired.  $\square$

Let  $S_{\ell,m}$  be the set of all edge-induced subhypergraphs of  $K_n^r$  having  $\ell$  vertices and  $m$  edges, or, more formally,

$$S_{\ell,m} := \{S \subseteq K_n^r : |V(S)| = \ell, |E(S)| = m\} \quad (7)$$

The following inequalities will be used throughout the proofs and hold for  $n$  sufficiently large:

$$\frac{q}{e} \leq \sigma^2 \leq q, \quad p-q \geq \frac{p}{e} \quad (8)$$

*Proof of Proposition 3.1.* Note that as the graphs in  $S_{\ell,m}$  are edge induced and we only consider  $S \subseteq [M]$  such that  $|S| \leq D$ , we have

$$r \leq \ell \leq rD, \quad \ell/r \leq m \leq D.$$

With the result of Lemma 3.1, we can conclude

$$\|L_{\leq D}\|^2 = 1 + \sum_{\ell=r}^{rD} \sum_{m=\ell/r}^D \sum_{S \in S_{\ell,m}} \rho^{2\ell} \left( \frac{p-q}{\sigma} \right)^{2m},$$

where the summand 1 comes from the term  $S = \emptyset$ . Let us compute an upper bound for  $|S_{\ell,m}|$ . To do so, we first consider the possible choices for the  $\ell$  vertices  $V(S)$  among all  $n$  vertices, and then the  $m$  edges  $E(S)$  among all  $\binom{\ell}{r}$  possible edges. We have

$$|S_{\ell,m}| \leq \binom{n}{\ell} \binom{\ell}{m} \leq n^\ell \binom{\ell}{r}^m \leq n^\ell \left( \frac{e\ell}{r} \right)^{rm} \leq n^\ell (eD)^{rm}, \quad (9)$$

since  $\ell \leq rD$ . Furthermore, as  $p > q$ , we have  $(p-q)^2 \leq p^2$ . We can now get

$$\begin{aligned} \|L_{\leq D}\|^2 &\leq 1 + \sum_{\ell=r}^{rD} \sum_{m=\ell/r}^D n^\ell (eD)^{rm} \rho^{2\ell} \left( \frac{p-q}{\sigma} \right)^{2m} \\ \text{[by (8)]} \quad &\leq 1 + \sum_{\ell=r}^{rD} (n\rho^2)^\ell \sum_{m=\ell/r}^D \left( \frac{e^{r+1} D^r p^2}{q} \right)^m \\ \text{[by } D = n^{o(1)}] \quad &\leq 1 + \sum_{\ell=r}^{rD} n^{(2\gamma-1)\ell} \sum_{m=\ell/r}^D n^{(\beta-2\alpha+o(1))m}. \end{aligned}$$

By (6), we have  $\beta - 2\alpha + o(1) < -r(2\gamma - 1) - \delta$  for a fixed constant  $\delta = \delta(\alpha, \beta, \gamma, r) > 0$ . It follows that

$$\|L_{\leq D}\|^2 \leq 1 + \sum_{\ell=r}^{rD} n^{(2\gamma-1)\ell} \sum_{m=\ell/r}^D n^{(-r(2\gamma-1)-\delta)m}.$$

Since  $\gamma \geq 1/2$ , we have  $n^{-r(2\gamma-1)-\delta} < 1/2$  for  $n$  sufficiently large. Therefore,

$$\begin{aligned} \|L_{\leq D}\|^2 &\leq 1 + 2 \sum_{\ell=r}^{rD} n^{(2\gamma-1)\ell} n^{(-r(2\gamma-1)-\delta)\ell/r} \\ &= 1 + 2 \sum_{\ell=r}^{rD} n^{-\delta\ell/r} \leq 1 + 4n^{-\delta}, \end{aligned}$$

as desired.  $\square$

### 3.2 | Lower Bound for Small $\gamma$

We now consider the regime  $\gamma < 1/2$  and show that the desired threshold is  $\alpha > \beta\gamma$ . In this case, the degree- $D$  likelihood ratio  $L_{\leq D}$  between  $\mathcal{P}$  and  $\mathcal{Q}$  has unbounded norm due to certain rare events under  $\mathcal{P}$ . To mitigate this issue, we modify our approach by conditioning on the complement of these rare events. We then study the degree- $D$  likelihood ratio  $L'_{\leq D}$  between the conditional distribution  $\mathcal{P}'$  and  $\mathcal{Q}$ .

#### 3.2.1 | A High-Probability Event

Recall that  $S_{\ell,m}$  is defined in (7). Fix an arbitrarily small constant  $\delta = \delta(\alpha, \beta, \gamma) > 0$  to be chosen later. For  $\ell \in \mathbb{N}$ , define

$$m_\ell := \left\lceil \ell \left( \frac{\gamma}{\alpha} + \delta \right) \right\rceil \quad (10)$$

Moreover, we define a set of pairs of integers

$$I := \{(\ell, m) \in \mathbb{N}^2 : m_\ell \leq m \leq D, S_{\ell,m} \neq \emptyset\}.$$

Let  $C := H[Z]$  denote the subhypergraph of the observed graph  $H$  induced by  $Z$ . In other words,  $C$  is the planted subhypergraph to be detected. We define the following events:

$$\begin{aligned} \mathcal{E}_{\ell,m} &:= \{\exists S \in S_{\ell,m} \text{ such that } S \subseteq C\} \\ \mathcal{E} &:= \bigcap_{(\ell,m) \in I} \mathcal{E}_{\ell,m}^c \end{aligned} \quad (11)$$

where  $\mathcal{E}_{\ell,m}^c$  is the complement of the event  $\mathcal{E}_{\ell,m}$ . In particular, under  $\mathcal{E}$ , every subset of edges  $S$  that induces a dense hypergraph does *not* appear in  $C$ , where “dense” means that the edge-to-vertex ratio of  $S$  is at least  $\frac{\gamma}{\alpha} + \delta$ . For  $y \in \{0, 1\}^M$ , we define a distribution  $\mathcal{P}'$  by

$$\mathcal{P}'(Y = y) := \frac{\mathcal{P}(Y = y) \mathbb{1}\{y \in \mathcal{E}\}}{\mathcal{P}(\mathcal{E})}.$$

We will show  $\mathcal{P}(\mathcal{E}) = 1 - o(1)$ . To assist with the bound, we start with the following lemma.

**Lemma 3.2.** *Under model  $\mathcal{P}$ , we have*

$$\frac{n^\gamma}{2} \leq |Z| \leq \frac{3n^\gamma}{2},$$

with probability at least  $1 - 2\exp(-cn^\gamma)$  for some absolute constant  $c > 0$ .

*Proof.* Note that  $|Z| = \sum_{i=1}^n z_i$  is the sum of  $n$  i.i.d.  $\text{Ber}(\rho)$  variables, that is,  $|Z| \sim \text{Bin}(n, \rho)$ . It follows from the Chernoff bound [38, Exercise 2.3.5] that

$$\mathcal{P}\left(|Z| \notin \left(1 \pm \frac{1}{2}\right) \mathbb{E}_{\mathcal{P}}[|Z|]\right) \leq 2 \exp\left(-C \frac{\mathbb{E}_{\mathcal{P}}[|Z|]}{4}\right),$$

for some absolute constant  $C > 0$ . The lemma follows as  $\mathbb{E}_{\mathcal{P}}[|Z|] = n\rho = n^\gamma$ .  $\square$

**Lemma 3.3.** *We have  $\mathcal{P}(\mathcal{E}) = 1 - o(1)$ .*

*Proof.* For any events  $A$  and  $B$ , we have  $\mathbb{P}[A] \leq \mathbb{P}[A|B] + \mathbb{P}[B^c]$ . We let  $A$  be  $\mathcal{E}^c$  and  $B$  be the event that  $|Z|$  is bounded as in Lemma 3.2. As a result,

$$\mathcal{P}(\mathcal{E}^c) \leq \mathcal{P}\left(\mathcal{E}^c \mid \frac{n^\gamma}{2} \leq |Z| \leq \frac{3n^\gamma}{2}\right) + 2 \exp(-c n^\gamma) \quad (12)$$

We now condition on a realization of  $Z$  such that  $\frac{n^\gamma}{2} \leq |Z| \leq \frac{3n^\gamma}{2}$ . We define the following set to assist with our proof:

$$S_{\ell,m}(Z) := \{S \in S_{\ell,m} : V(S) \subseteq Z\}.$$

An identical argument as that in (9) shows that

$$|S_{\ell,m}(Z)| \leq \left(\frac{3}{2}n^\gamma\right)^\ell (eD)^{rm}.$$

With this in hand, we have

$$\begin{aligned} \mathcal{P}(\mathcal{E}_{\ell,m}^c | Z) &\leq \sum_{S \in S_{\ell,m}(Z)} \mathcal{P}(S \subseteq E(C) | Z) \\ &\leq \left(\frac{3}{2}n^\gamma\right)^\ell (eD)^{rm} \max_{S \in S_{\ell,m}(Z)} \mathcal{P}(S \subseteq E(C) | Z). \end{aligned}$$

For  $S \in S_{\ell,m}(Z)$ , as  $Y_e$  and  $Y_f$  are conditionally independent given  $Z$ , we have

$$\mathcal{P}(S \subseteq E(C) | Z) = p^m = n^{-\alpha m}.$$

Plugging in this value above, we have

$$\mathcal{P}(\mathcal{E}_{\ell,m}^c | Z) \leq \left(\frac{3}{2}n^\gamma\right)^\ell (eD)^{rm} n^{-\alpha m} \leq n^{\gamma\ell - (\alpha - o(1))m}.$$

Let us now bound  $\mathcal{P}(\mathcal{E}^c | Z) \leq \sum_{(\ell,m) \in I} \mathcal{P}(\mathcal{E}_{\ell,m}^c | Z)$ . We have

$$\begin{aligned} \mathcal{P}(\mathcal{E}^c | Z) &\leq \sum_{\ell=r}^{rD} \sum_{m=m_\ell}^D n^{\gamma\ell - (\alpha - o(1))m} = \sum_{\ell=r}^{rD} n^{\gamma\ell} \sum_{m=m_\ell}^D n^{-(\alpha - o(1))m} \\ &\leq 2 \sum_{\ell=r}^{rD} n^{\gamma\ell} n^{-(\alpha - o(1))m_\ell}. \end{aligned}$$

By (10), we have  $m_\ell \geq \ell(\frac{\gamma}{\alpha} + \delta)$ , so

$$\mathcal{P}(\mathcal{E}^c | Z) \leq 2 \sum_{\ell=r}^{rD} n^{-(\delta\alpha - o(1))\ell} \leq 4n^{-\delta\alpha r/2}.$$

This combined with (12) completes the proof.  $\square$

### 3.2.2 | Bounding the Low-Degree Norm

The main result of this subsection consists in controlling the norm of the degree- $D$  likelihood ratio between  $\mathcal{P}'$  and  $\mathcal{Q}$  defined in (4).

**Proposition 3.2.** *Suppose  $p = n^{-\alpha}$ ,  $q = n^{-\beta}$ , and  $\rho = n^{\gamma-1}$  with fixed  $r \geq 2$ ,  $0 < \alpha < \beta < r - 1$ , and  $\gamma \in (0, 1)$  such that*

$$\gamma < \frac{1}{2}, \quad \alpha > \beta\gamma. \quad (13)$$

*If  $D = n^{o(1)}$ , then we have*

$$\|L'_{\leq D}\|^2 = 1 + o(1).$$

To prove the proposition, we start from (4). Note the following as a result of §3.2.1:

$$\begin{aligned} \mathbb{E}_{\mathcal{P}'}[\phi_S(Y)] &= \frac{1}{\mathcal{P}(\mathcal{E})} \mathbb{E}_{\mathcal{P}}[\phi_S(Y) \mathbb{1}\{Y \in \mathcal{E}\}] \\ &= (1 + o(1)) \mathbb{E}_{\mathcal{P}}[\phi_S(Y) \mathbb{1}\{Y \in \mathcal{E}\}] \end{aligned} \quad (14)$$

It is now enough to consider the final term. We first condition on the outcome  $Z$  to get

$$\mathbb{E}_{\mathcal{P}}[\phi_S(Y) \mathbb{1}\{Y \in \mathcal{E}\}] = \mathbb{E}_Z[\mathbb{E}_{\mathcal{P}}[\phi_S(Y) \mathbb{1}\{Y \in \mathcal{E}\} | Z]].$$

Consider a realization  $Z$ . Suppose there exists  $e \in S$  such that  $e \notin Z$ . Since  $\mathcal{E}$  is determined by the edges in  $C$ ,  $Y_e$  is conditionally independent of  $\phi_{S \setminus \{e\}}(Y) \mathbb{1}\{Y \in \mathcal{E}\}$ , given  $Z$ . In particular, we have

$$\begin{aligned} \mathbb{E}_{\mathcal{P}}[\phi_S(Y) \mathbb{1}\{Y \in \mathcal{E}\} | Z] \\ = \mathbb{E}_{\mathcal{P}}[\tilde{Y}_e | Z] \mathbb{E}_{\mathcal{P}}[\phi_{S \setminus \{e\}}(Y) \mathbb{1}\{Y \in \mathcal{E}\} | Z] = 0, \end{aligned}$$

as  $\mathbb{E}_{\mathcal{P}}[\tilde{Y}_e | Z] = 0$  for  $e \notin Z$ . It follows that

$$\begin{aligned} \mathbb{E}_{\mathcal{P}}[\phi_S(Y) \mathbb{1}\{Y \in \mathcal{E}\}] \\ = \mathbb{E}_Z[\mathbb{E}_{\mathcal{P}}[\phi_S(Y) \mathbb{1}\{Y \in \mathcal{E}\} | Z] \mathbb{1}\{V(S) \subseteq Z\}] \end{aligned} \quad (15)$$

We split the set of  $S \subseteq K_n^r$  with  $|S| \leq D$  into a set of “bad” subgraphs and a set of “good” subgraphs, defined respectively as

$$B := \bigcup_{(\ell,m) \in I} S_{\ell,m}, \quad \mathcal{G} := \{S \subseteq K_n^r : |S| \leq D\} \setminus B.$$

Let us first consider a good subgraph  $S$ . We prove the following lemma.

**Lemma 3.4.** *Let  $S \in \mathcal{G}$  be a good subgraph on  $\ell$  vertices and  $m$  edges. We have*

$$|\mathbb{E}_{\mathcal{P}}[\phi_S(Y) \mathbb{1}\{Y \in \mathcal{E}\}]| \leq \rho^\ell \left(\frac{2p}{\sigma}\right)^m.$$

*Proof.* Note the following as a result of (15) and Jensen’s inequality:

$$\begin{aligned} |\mathbb{E}_{\mathcal{P}}[\phi_S(Y) \mathbb{1}\{Y \in \mathcal{E}\}]| \\ \leq \mathbb{E}_Z[\mathbb{E}_{\mathcal{P}}[|\phi_S(Y)| \mathbb{1}\{Y \in \mathcal{E}\} | Z] \mathbb{1}\{V(S) \subseteq Z\}] \end{aligned} \quad (16)$$

Let us consider  $\mathbb{E}_{\mathcal{P}}[|\phi_S(Y)| \mathbb{1}\{Y \in \mathcal{E}\} | Z]$  for  $V(S) \subseteq Z$ . Since  $|\phi_S(Y)| \geq 0$  we have

$$\begin{aligned}\mathbb{E}_p[|\phi_S(Y)|\mathbb{1}\{Y \in \mathcal{E}\}|Z] &\leq \mathbb{E}_p[|\phi_S(Y)||Z] \\ &= \left(\frac{p(1-q) + (1-p)q}{\sigma}\right)^m \leq \left(\frac{2p}{\sigma}\right)^m,\end{aligned}$$

where the last step follows since  $q < p$ . We now have:

$$\begin{aligned}|\mathbb{E}_p[\phi_S(Y)\mathbb{1}\{Y \in \mathcal{E}\}]| \\ \leq \left(\frac{2p}{\sigma}\right)^m \mathbb{E}_Z[\mathbb{1}\{V(S) \subseteq Z\}] = \rho^\ell \left(\frac{2p}{\sigma}\right)^m,\end{aligned}$$

as desired.  $\square$

Next, let us consider a bad subgraph  $S$ . We will prove the following lemma.

**Lemma 3.5.** *Let  $S \in \mathcal{B}$  be a bad subgraph on  $\ell$  vertices and  $m$  edges. We have*

$$|\mathbb{E}_p[\phi_S(Y)\mathbb{1}\{Y \in \mathcal{E}\}]| \leq \rho^\ell \binom{m}{m_\ell - 1} \frac{q^{m-m_\ell+1} (2p)^{m_\ell-1}}{\sigma^m}.$$

*Proof.* As in the proof of Lemma 3.4, we will first bound  $\mathbb{E}_p[|\phi_S(Y)|\mathbb{1}\{Y \in \mathcal{E}\}|Z]$  for  $V(S) \subseteq Z$ . By the definition of  $\mathcal{E}$  in (11),  $Y \in \mathcal{E}$  implies that  $C$  contains at most  $m_\ell - 1$  edges in  $S \in \mathcal{B}$  which has  $m$  edges. Therefore, for  $s := m - m_\ell + 1 \geq 1$ , there exist  $e_1, \dots, e_s \in S$  such that  $Y_{e_i} = 0$  for each  $i \in [s]$ . With this in mind, we have

$$\begin{aligned}\mathbb{E}_p[|\phi_S(Y)|\mathbb{1}\{Y \in \mathcal{E}\}|Z] \\ \leq \sum_{e_1, \dots, e_s \in S} \mathbb{E}_p[|\phi_S(Y)|\mathbb{1}\{Y_{e_1} = \dots = Y_{e_s} = 0\}|Z] \\ \leq \sum_{e_1, \dots, e_s \in S} \left(\frac{q}{\sigma}\right)^s \left(\frac{p(1-q) + (1-p)q}{\sigma}\right)^{m-s} \leq \binom{m}{s} \frac{q^s (2p)^{m-s}}{\sigma^m},\end{aligned}$$

where the last step follows since  $q < p$ . By (16) again, we obtain

$$\begin{aligned}|\mathbb{E}_p[\phi_S(Y)\mathbb{1}\{Y \in \mathcal{E}\}]| \\ \leq \binom{m}{s} \frac{q^s (2p)^{m-s}}{\sigma^m} \mathbb{E}_Z[\mathbb{1}\{V(S) \subseteq Z\}] = \rho^\ell \binom{m}{s} \frac{q^s (2p)^{m-s}}{\sigma^m},\end{aligned}$$

as desired.  $\square$

Recall our original goal of bounding (4). By (14), we can simplify (4) to

$$\begin{aligned}\|L'_{\leq D}\|^2 \\ = (1 + o(1)) \sum_{\substack{S \subseteq [M], \\ |S| \leq D}} \mathbb{E}_p[\phi_S(Y)\mathbb{1}\{Y \in \mathcal{E}\}]^2 \\ = (1 + o(1)) \left( \sum_{S \in \mathcal{G}} \mathbb{E}_p[\phi_S(Y)\mathbb{1}\{Y \in \mathcal{E}\}]^2 \right. \\ \left. + \sum_{S \in \mathcal{B}} \mathbb{E}_p[\phi_S(Y)\mathbb{1}\{Y \in \mathcal{E}\}]^2 \right)\end{aligned}\quad (17)$$

Let us consider the first sum in (17).

**Lemma 3.6.** *We have  $\sum_{S \in \mathcal{G}} \mathbb{E}_p[\phi_S(Y)\mathbb{1}\{Y \in \mathcal{E}\}]^2 = 1 + o(1)$ .*

*Proof.* If  $S$  is the empty graph, then  $\mathbb{E}_p[\phi_S(Y)] = 1$ . For nonempty  $S \in \mathcal{S}_{\ell, m} \subseteq \mathcal{G}$ , we have

$$\frac{\ell}{r} \leq m \leq m_\ell - 1,$$

since  $S$  has no isolated vertices. From (9) and Lemma 3.4, we obtain

$$\begin{aligned}\sum_{S \in \mathcal{G}} \mathbb{E}_p[\phi_S(Y)\mathbb{1}\{Y \in \mathcal{E}\}]^2 \\ \leq 1 + \sum_{\ell=r}^{rD} \sum_{m=\ell/r}^{\min(m_\ell-1, D)} \sum_{S \in \mathcal{S}_{\ell, m}} \rho^{2\ell} \left(\frac{2p}{\sigma}\right)^{2m} \\ \leq 1 + \sum_{\ell=r}^{rD} \sum_{m=\ell/r}^{\min(m_\ell-1, D)} n^\ell (eD)^{rm} \rho^{2\ell} \left(\frac{2p}{\sigma}\right)^{2m} \\ \leq 1 + \sum_{\ell=r}^{rD} \sum_{m=\ell/r}^{\min(m_\ell-1, D)} (n\rho^2)^\ell \left(\frac{e^{r+3} D^r p^2}{q}\right)^m \\ = 1 + \sum_{\ell=r}^{rD} n^{(2\gamma-1)\ell} \sum_{m=\ell/r}^{\min(m_\ell-1, D)} n^{(\beta-2\alpha+o(1))m},\end{aligned}$$

where we use (8). We now consider two cases. First, suppose  $\beta < 2\alpha$ . Then, we have for some small enough constant  $\delta_1 = \delta_1(\alpha, \beta) > 0$  and  $n$  large enough:

$$\begin{aligned}\sum_{\ell=r}^{rD} n^{(2\gamma-1)\ell} \sum_{m=\ell/r}^{\min(m_\ell-1, D)} n^{(\beta-2\alpha+o(1))m} \leq \sum_{\ell=r}^{rD} n^{(2\gamma-1)\ell} \\ \sum_{m=\ell/r}^{\min(m_\ell-1, D)} n^{-\delta_1 m} \leq 2 \sum_{\ell=r}^{rD} n^{(2\gamma-1)\ell - \delta_1 \ell/r}.\end{aligned}$$

Since  $\gamma < 1/2$ , we can simplify the above further to get

$$2 \sum_{\ell=r}^{rD} n^{(2\gamma-1)\ell - \delta_1 \ell/r} < 2 \sum_{\ell=r}^{rD} n^{-\delta_1 \ell/r} \leq 4n^{-\delta_1},$$

as desired. Now, let us assume  $\beta \geq 2\alpha$ . We have

$$\begin{aligned}\sum_{\ell=r}^{rD} n^{(2\gamma-1)\ell} \sum_{m=\ell/r}^{\min(m_\ell-1, D)} n^{(\beta-2\alpha+o(1))m} \\ \leq \sum_{\ell=r}^{rD} n^{(2\gamma-1)\ell} \left(D - \frac{\ell}{r} + 1\right) n^{(\beta-2\alpha+o(1))(m_\ell-1)} \\ \leq \sum_{\ell=r}^{rD} n^{(2\gamma-1)\ell + (\beta-2\alpha+o(1))(m_\ell-1) + o(1)}\end{aligned}\quad (18)$$

Let us now consider the coefficient of  $m_\ell - 1$  in the exponent of  $n$ . By the definition of  $m_\ell$  in (10), we have

$$m_\ell - 1 < \ell \left(\frac{\gamma}{\alpha} + \delta\right).$$

With this in hand, we have

$$\begin{aligned}(\beta - 2\alpha + o(1))(m_\ell - 1) \\ < (\beta - 2\alpha + o(1)) \left(\frac{\gamma}{\alpha} + \delta\right) \ell \\ < \left(\frac{\beta\gamma}{\alpha} - 2\gamma + r\delta\right) \ell.\end{aligned}$$



From here it follows that the exponent in (18) can be bounded as

$$(2\gamma - 1)\ell + (\beta - 2\alpha + o(1))(m_\ell - 1) + o(1) \leq \left(\frac{\beta\gamma}{\alpha} - 1 + 2r\delta\right)\ell \leq -\delta\ell \quad (19)$$

where the last step holds in view of the condition  $\alpha > \beta\gamma$  in (13), once we choose  $\delta = \delta(\alpha, \beta, \gamma) > 0$  to be sufficiently small. Then (18) can be further bounded as

$$\sum_{\ell=r}^{rD} n^{(2\gamma-1)\ell + (\beta-2\alpha+o(1))(m_\ell-1) + o(1)} \leq \sum_{\ell=r}^{rD} n^{-\delta\ell} \leq 2n^{-\delta r},$$

as desired.  $\square$

Now, let us consider the bad subgraphs and bound the second sum in (17).

**Lemma 3.7.** *We have  $\sum_{S \in B} \mathbb{E}_p[\phi_S(Y)\mathbb{1}\{Y \in \mathcal{E}\}]^2 = o(1)$ .*

*Proof.* Note that for  $S \in S_{\ell,m} \subseteq B$ , we have

$$m_\ell \leq m \leq D.$$

From (9) and Lemma 3.5, we have

$$\begin{aligned} \sum_{S \in B} \mathbb{E}_p[\phi_S(Y)\mathbb{1}\{Y \in \mathcal{E}\}]^2 &\leq \sum_{\ell=r}^{rD} \sum_{m=m_\ell}^D \sum_{S \in S_{\ell,m}} \rho^{2\ell} \binom{m}{m_\ell - 1}^2 \frac{q^{2(m-m_\ell+1)} (2p)^{2(m_\ell-1)}}{\sigma^{2m}} \\ &\leq \sum_{\ell=r}^{rD} \sum_{m=m_\ell}^D n^\ell (eD)^{rm} \rho^{2\ell} \left(\frac{q}{\sigma}\right)^{2(m-m_\ell+1)} \left(\frac{2m}{\sigma}\right)^{2(m_\ell-1)} \\ &\leq \sum_{\ell=r}^{rD} \sum_{m=m_\ell}^D (n\rho^2)^\ell \left(\frac{e^r D^r q^2}{\sigma^2}\right)^m \left(\frac{2m}{q}\right)^{2(m_\ell-1)} \\ &= \sum_{\ell=r}^{rD} n^{(2\gamma-1)\ell} n^{2(\beta-\alpha+o(1))(m_\ell-1)} \sum_{m=m_\ell}^D n^{(-\beta+o(1))m}, \end{aligned}$$

where we use (8). Note that the coefficient of  $m$  in the exponent is negative for large enough  $n$ . Thus, we can further simplify the above to

$$\begin{aligned} \sum_{S \in B} \mathbb{E}_p[\phi_S(Y)\mathbb{1}\{Y \in \mathcal{E}\}]^2 &\leq 2 \sum_{\ell=r}^{rD} n^{(2\gamma-1)\ell} n^{2(\beta-\alpha+o(1))(m_\ell-1)} n^{(-\beta+o(1))m_\ell} \\ &= 2 \sum_{\ell=r}^{rD} n^{(2\gamma-1)\ell + (\beta-2\alpha+o(1))(m_\ell-1) - \beta + o(1)}. \end{aligned}$$

Once again, we consider two cases. First, let us assume  $\beta < 2\alpha$ . As  $\gamma < 1/2$ , we have for  $n$  large enough:

$$\begin{aligned} \sum_{S \in B} \mathbb{E}_p[\phi_S(Y)\mathbb{1}\{Y \in \mathcal{E}\}]^2 &\leq 2 \sum_{\ell=r}^{rD} n^{-\beta+o(1)} \\ &\leq 2rDn^{-\beta+o(1)} \leq n^{-\beta/2}, \end{aligned}$$

as desired. Now, let us assume  $\beta \geq 2\alpha$ . An identical argument as in the proof of (19) shows that

$$(2\gamma - 1)\ell + (\beta - 2\alpha + o(1))(m_\ell - 1) - \beta + o(1) \leq -\delta\ell - \beta.$$

In particular, we have

$$\sum_{S \in B} \mathbb{E}_p[\phi_S(Y)\mathbb{1}\{Y \in \mathcal{E}\}]^2 \leq 2 \sum_{\ell=r}^{rD} n^{-\delta\ell - \beta} \leq 4n^{-\beta - \delta r},$$

as desired.  $\square$

*Proof of Proposition 3.2.* It suffices to combine (17) with Lemmas 3.6 and 3.7.  $\square$

## 4 | Detection Upper Bound

As in the previous section, we split the analysis into two cases based on the value  $\gamma$ . For  $\gamma \geq 1/2$ , we consider the signed count of the total number of hyperedges in  $H$ , as defined in (20). In particular, it is a linear function in the entries of the adjacency tensor  $Y$ . For  $\gamma < 1/2$ , our statistic is the (unsigned) count of the occurrences of a certain *balanced* subhypergraph in  $H$ , as defined in (25). The size of the subhypergraph depends only on the fixed parameters  $(\alpha, \beta, \gamma)$  as given in Proposition 4.2. Therefore, the statistic in (25) is a constant-degree polynomial in the entries of  $Y$  and can be computed in polynomial time.

*Proof of Theorem 2.1. (upper bound).* If  $\gamma \geq 1/2$  and  $\alpha < \beta/2 + r(\gamma - 1/2)$ , Proposition 4.1 shows that the statistic  $\tilde{T}$  defined in (20) strongly separates  $\mathcal{P}$  and  $\mathcal{Q}$ . If  $\gamma < 1/2$  and  $\alpha < \beta\gamma$ , Proposition 4.3 shows that the statistic  $T$  defined in (25) strongly separates  $\mathcal{P}$  and  $\mathcal{Q}$ .

### 4.1 | Upper Bound for Large $\gamma$

In the regime where  $\gamma \geq 1/2$ , we will use the signed count of hyperedges in  $H$  as our test statistic. Recall  $M = \binom{r}{r}$  as defined in §2, and  $[M]$  denotes the set of all possible hyperedges. We define

$$\tilde{T} := \sum_{e \in [M]} \tilde{Y}_e \quad (20)$$

where  $\tilde{Y}_e$  is defined in (1). Trivially,  $\tilde{T}$  is a degree-1 polynomial of the entries of  $Y$ .

**Proposition 4.1.** *Suppose  $p = n^{-\alpha}$ ,  $q = n^{-\beta}$ , and  $\rho = n^{\gamma-1}$  with fixed  $r \geq 2$ ,  $0 < \alpha < \beta < r - 1$ , and  $\gamma \in (0, 1)$  such that*

$$\gamma \geq \frac{1}{2}, \quad \alpha < \frac{\beta}{2} + r\left(\gamma - \frac{1}{2}\right). \quad (21)$$

*For the statistic  $\tilde{T}$  defined in (20), we have  $\sqrt{\text{Var}_{\mathcal{P}}(\tilde{T}) \vee \text{Var}_{\mathcal{Q}}(\tilde{T})} = o(|\mathbb{E}_{\mathcal{P}}[\tilde{T}] - \mathbb{E}_{\mathcal{Q}}[\tilde{T}]|)$ .*

Let us first compute the expectation and variance under each distribution.

**Lemma 4.1.** We have

$$\mathbb{E}_Q[\tilde{T}] = 0, \quad \text{Var}_Q(\tilde{T}) = M, \quad \mathbb{E}_P[\tilde{T}] = M \rho^r \left( \frac{p-q}{\sigma} \right),$$

$$\text{Var}_P(\tilde{T}) \leq M + \frac{2M\rho^r p}{\sigma^2} + \frac{2M rn^{r-1} \rho^{2r-1} p^2}{\sigma^2}.$$

*Proof.* Note that  $\tilde{Y}_e$  is defined by standardizing  $Y_e$  under the distribution  $Q$ , and  $Y_e$  and  $Y_f$  are independent under  $Q$  for  $e \neq f$ . The results on  $\mathbb{E}_Q[\tilde{T}]$  and  $\text{Var}_Q(t)$  follow immediately.

Let us now consider an edge  $e \in [M]$ . We have

$$\mathbb{E}_P[\tilde{Y}_e] = \rho^r \left( \frac{p-q}{\sigma} \right) + (1 - \rho^r) \cdot 0 = \rho^r \left( \frac{p-q}{\sigma} \right),$$

which leads to the desired formula for  $\mathbb{E}_P[\tilde{T}]$ .

Now, we have the following for the variance under  $P$ :

$$\text{Var}_P(\tilde{T}) = \sum_{e, f \in [M]} \text{Cov}_P(\tilde{Y}_e, \tilde{Y}_f).$$

Notice that  $\tilde{Y}_e$  and  $\tilde{Y}_f$  are independent for  $e \cap f = \emptyset$ . It follows that

$$\begin{aligned} \text{Var}_P(\tilde{T}) &= \sum_{\substack{e, f \in [M], \\ e \cap f \neq \emptyset}} \text{Cov}_P(\tilde{Y}_e, \tilde{Y}_f) \\ &= \sum_{\substack{e, f \in [M], \\ e \cap f \neq \emptyset}} \left( \mathbb{E}_P[\tilde{Y}_e \tilde{Y}_f] - \rho^{2r} \left( \frac{p-q}{\sigma} \right)^2 \right) \\ &\leq \sum_{e \in [M]} \mathbb{E}_P[\tilde{Y}_e^2] + \sum_{e \in [M]} \sum_{\substack{f \in [M] \setminus e, \\ e \cap f \neq \emptyset}} \mathbb{E}_P[\tilde{Y}_e \tilde{Y}_f] \end{aligned} \quad (22)$$

Let us consider the first term. We have

$$\begin{aligned} \mathbb{E}_P[\tilde{Y}_e^2] &= \rho^r \left( p \left( \frac{1-q}{\sigma} \right)^2 + (1-p) \left( \frac{q}{\sigma} \right)^2 \right) \\ &\quad + (1 - \rho^r) \left( q \left( \frac{1-q}{\sigma} \right)^2 + (1-q) \left( \frac{q}{\sigma} \right)^2 \right) \\ &\leq \rho^r \left( \frac{p(1-q)^2 + (1-p)q^2}{\sigma^2} \right) + 1 \\ &\leq \frac{2\rho^r p}{\sigma^2} + 1, \end{aligned}$$

as  $q^2 < q < p$ . For the second term in (22), let us first compute  $\mathbb{E}_P[\tilde{Y}_e \tilde{Y}_f]$ . As  $\tilde{Y}_e$  and  $\tilde{Y}_f$  are independent given  $Z$ , we have

$$\mathbb{E}_P[\tilde{Y}_e \tilde{Y}_f] = \mathbb{E}_Z[\mathbb{E}_P[\tilde{Y}_e \tilde{Y}_f | Z]] = \mathbb{E}_Z[\mathbb{E}_P[\tilde{Y}_e | Z] \mathbb{E}_P[\tilde{Y}_f | Z]].$$

We note the following:

$$\begin{aligned} \mathbb{E}_P[\tilde{Y}_e | Z] &= \left( \frac{p-q}{\sigma} \right) \mathbb{1}_{\{e \subseteq Z\}} + 0 \cdot \mathbb{1}_{\{e \not\subseteq Z\}} \\ &= \left( \frac{p-q}{\sigma} \right) \mathbb{1}_{\{e \subseteq Z\}}. \end{aligned}$$

In particular, we have

$$\begin{aligned} \mathbb{E}_P[\tilde{Y}_e \tilde{Y}_f] &= \mathbb{E}_Z \left[ \left( \frac{p-q}{\sigma} \right)^2 \mathbb{1}_{\{e \cup f \subseteq Z\}} \right] = \rho^{2r-|e \cap f|} \left( \frac{p-q}{\sigma} \right)^2 \\ &\leq \frac{\rho^{2r-|e \cap f|} p^2}{\sigma^2}, \end{aligned}$$

where the last step follows since  $p > q$ . Putting these bounds together with (22), we now have

$$\text{Var}_P(\tilde{T}) \leq M + \frac{2M\rho^r p}{\sigma^2} + \frac{\rho^{2r} p^2}{\sigma^2} \sum_{e \in [M]} \sum_{\substack{f \in [M] \setminus e, \\ e \cap f \neq \emptyset}} \rho^{-|e \cap f|}.$$

Let us fix  $e$ . We will consider bounding

$$\sum_{\substack{f \in [M] \setminus e, \\ e \cap f \neq \emptyset}} \rho^{-|e \cap f|} \leq \sum_{s=1}^{r-1} \sum_{v_1, \dots, v_s \in e} \sum_{\substack{f \in [M], \\ v_1, \dots, v_s \in f}} \rho^{-s}.$$

For a fixed  $v_1, \dots, v_s$ , there are at most  $\binom{n}{r-s} \leq n^{r-s}$  choices for  $f$ . Similarly, there are at most  $\binom{r}{s} \leq r^s$  choices for  $v_1, \dots, v_s$ . It follows that

$$\sum_{\substack{f \in [M] \setminus e, \\ e \cap f \neq \emptyset}} \rho^{-|e \cap f|} \leq \sum_{s=1}^{r-1} r^s n^{r-s} \rho^{-s} = n^r \sum_{s=1}^{r-1} \left( \frac{r}{n\rho} \right)^s.$$

Note that  $n\rho = n^r \gg r$  and so we have

$$\sum_{\substack{f \in [M] \setminus e, \\ e \cap f \neq \emptyset}} \rho^{-|e \cap f|} \leq 2rn^{r-1} \rho^{-1}.$$

Plugging this value in for our variance computation, we get

$$\text{Var}_P(\tilde{T}) \leq M + \frac{2M\rho^r p}{\sigma^2} + \frac{2M rn^{r-1} \rho^{2r-1} p^2}{\sigma^2},$$

as desired.  $\square$

*Proof of Proposition 4.1.* Lemma 4.1 implies that it is enough to show

$$\begin{aligned} M^2 \rho^{2r} \left( \frac{p-q}{\sigma} \right)^2 &\gg M + \frac{2M\rho^r p}{\sigma^2} + \frac{2M rn^{r-1} \rho^{2r-1} p^2}{\sigma^2} \\ \Leftrightarrow \rho^{2r} \left( \frac{p-q}{\sigma} \right)^2 &\gg \frac{1}{M} + \frac{2\rho^r p}{M\sigma^2} + \frac{2 rn^{r-1} \rho^{2r-1} p^2}{M\sigma^2}. \end{aligned}$$

Let us first lower bound the left-hand side. From (8), we have

$$LHS \geq \frac{\rho^{2r} p^2}{\sigma^2 q} = e^{-2} n^{2r(\gamma-1)+\beta-2\alpha}.$$

Next, let us upper bound the right-hand side. Note that  $M = \binom{n}{r} \geq \left( \frac{n}{r} \right)^r$  so that  $\frac{1}{M} = n^{-r+o(1)}$ . From (8), we have

$$RHS \leq n^{-r+o(1)} + n^{r(\gamma-2)+\beta-\alpha+o(1)} + n^{-1+(2r-1)(\gamma-1)+\beta-2\alpha+o(1)}.$$

It remains to verify that

$$\begin{aligned} 2r(\gamma-1) + \beta - 2\alpha &> -r + o(1), \\ 2r(\gamma-1) + \beta - 2\alpha &> r(\gamma-2) + \beta - \alpha + o(1), \\ 2r(\gamma-1) + \beta - 2\alpha &> -1 + (2r-1)(\gamma-1) + \beta - 2\alpha + o(1), \end{aligned}$$

using the assumption (21), which is straightforward.  $\square$

## 4.2 | Upper Bound for Small $\gamma$

In this subsection, we will assume

$$\gamma < \frac{1}{2}, \quad \alpha < \beta \gamma. \quad (23)$$

As in the previous subsection, we will define a test statistic  $T$  to distinguish between  $\mathcal{P}$  and  $\mathcal{Q}$ .

### 4.2.1 | Balanced Hypergraphs

The statistic  $T$  we choose will count the occurrences of a specific hypergraph  $\tilde{H}$  in  $H$ . We first define balanced hypergraphs.

**Definition 4.1.** A hypergraph  $H$  is balanced if for every nonempty  $H' \subseteq H$ , we have

$$\frac{|E(H')|}{|V(H')|} \leq \frac{|E(H)|}{|V(H)|}.$$

Let us now define our hypergraph of interest  $\tilde{H}$ , whose existence is guaranteed by [39] for  $r = 2$  and by [40] in the general case.

**Proposition 4.2.** Assuming (23), there exists a balanced hypergraph  $\tilde{H}$  with  $\ell$  vertices and  $m$  edges satisfying

$$\frac{1}{\beta} < \frac{m}{\ell} < \frac{\gamma}{\alpha} \quad (24)$$

*Proof.* Since  $\alpha < \beta\gamma$  by assumption, there is a rational number  $\lambda$  such that  $\frac{1}{\beta} < \lambda < \frac{\gamma}{\alpha}$ . Furthermore, note that  $\lambda > \frac{1}{\beta} \geq \frac{1}{r-1}$ . By [40, Theorem 2], there exists a balanced hypergraph with  $\ell$  vertices and  $m$  edges such that  $\frac{m}{\ell} = \lambda$ , completing the proof.  $\square$

Similar to (2), we define

$$\psi_S(Y) := \prod_{e \in S} Y_e, \quad \forall S \subseteq [M].$$

With  $\tilde{H}$  in hand, we are ready to define  $T$ :

$$T := \sum_{S \in \mathcal{S}} \psi_S(Y), \quad \text{where } \mathcal{S} := \{S \subseteq [M] : K_n^r[S] \cong \tilde{H}\} \quad (25)$$

In particular,  $T$  counts the number of edge-induced subgraphs in  $H$  isomorphic to  $\tilde{H}$ . Since the number of edges  $m$  in  $\tilde{H}$  is a constant depending only on  $(\alpha, \beta, \gamma)$ , the statistic  $T$  is a constant-degree polynomial in the entries of  $Y$ . We remark that  $T$  is taken to be the (unsigned) count rather than the signed count mainly for technical convenience. The fact that  $\psi_S(Y)$  takes value only in  $\{0, 1\}$  simplifies the computation of the variance of  $T$ .

Before turning to the main result of this section, we note a simple fact about balanced hypergraphs that will be useful in the proof later.

**Claim 4.1.** Let  $H$  be a balanced hypergraph and let  $H'$  be a proper subhypergraph of  $H$  with  $V(H') \subsetneq V(H)$ . Then,

$$\frac{|E(H)| - |E(H')|}{|V(H)| - |V(H')|} \geq \frac{|E(H)|}{|V(H)|}.$$

*Proof.* If  $H' = \emptyset$ , the inequality holds trivially. Therefore, we may assume  $H'$  is nonempty. As  $H$  is balanced and  $H'$  is nonempty, by Definition 4.1 we have

$$\begin{aligned} \frac{|E(H)| - |E(H')|}{|V(H)| - |V(H')|} &\geq \frac{|E(H)| - |V(H')| \frac{|E(H)|}{|V(H)|}}{|V(H)| - |V(H')|} \\ &= \frac{|E(H)|(|V(H)| - |V(H')|)}{|V(H)|(|V(H)| - |V(H')|)} = \frac{|E(H)|}{|V(H)|}, \end{aligned}$$

as desired.  $\square$

### 4.2.2 | Strong Separation

We now show that the statistic  $T$  strongly separates  $\mathcal{P}$  and  $\mathcal{Q}$  in the regime of interest.

**Proposition 4.3.** Suppose  $p = n^{-\alpha}$ ,  $q = n^{-\beta}$ , and  $\rho = n^{\gamma-1}$  with fixed  $r \geq 2$ ,  $0 < \alpha < \beta < r-1$ , and  $\gamma \in (0, 1)$  such that (23) holds. The statistic  $T$  defined in (25) satisfies  $\sqrt{\text{Var}_{\mathcal{P}}(T) \vee \text{Var}_{\mathcal{Q}}(T)} = o(|\mathbb{E}_{\mathcal{P}}[T] - \mathbb{E}_{\mathcal{Q}}[T]|)$ .

Let us first consider  $T$  under  $\mathcal{Q}$ . We have the following lemma.

**Lemma 4.2.** With  $T$  as defined in (25), we have

$$\begin{aligned} \mathbb{E}_{\mathcal{Q}}[T] &= N q^m, \quad \text{Var}_{\mathcal{Q}}(T) \\ &\leq \max \{m^2 N n^{\ell(1-1/m)} \ell^{r(m-1)} q^{2m-1}, m^{m+1} \mathbb{E}_{\mathcal{Q}}[T]\}, \end{aligned}$$

where  $N = |S|$ .

*Proof.* As  $Y_e$  and  $Y_f$  are independent for  $e \neq f$  under  $\mathcal{Q}$ , we note that for any  $S \subseteq [M]$ , we have

$$\psi_S(Y) \sim \text{Ber}(q^m),$$

from where the result on  $\mathbb{E}_{\mathcal{Q}}[T]$  follows. For the variance, we have

$$\text{Var}_{\mathcal{Q}}(T) = \sum_{S_1, S_2 \in \mathcal{S}} \text{Cov}_{\mathcal{Q}}(\psi_{S_1}(Y), \psi_{S_2}(Y)).$$

Note that  $\psi_{S_1}(Y)$  and  $\psi_{S_2}(Y)$  are independent if  $S_1 \cap S_2 = \emptyset$ . Furthermore, we have

$$\begin{aligned} \psi_{S_1}(Y) \psi_{S_2}(Y) &= \prod_{e \in S_1} Y_e \prod_{e \in S_2} Y_e = \prod_{e \in S_1 \cap S_2} Y_e^2 \prod_{e \in S_1 \Delta S_2} Y_e \\ Y_e &= \prod_{e \in S_1 \cap S_2} Y_e \prod_{e \in S_1 \Delta S_2} Y_e = \prod_{e \in S_1 \cup S_2} Y_e, \end{aligned}$$

as  $Y_e^2 = Y_e$ . In particular, we can bound the variance by

$$\begin{aligned} \text{Var}_{\mathcal{Q}}(T) &= \sum_{\substack{S_1, S_2 \in \mathcal{S}, \\ S_1 \cap S_2 \neq \emptyset}} \text{Cov}_{\mathcal{Q}}(\psi_{S_1}(Y), \psi_{S_2}(Y)) \\ &= \sum_{\substack{S_1, S_2 \in \mathcal{S}, \\ S_1 \cap S_2 \neq \emptyset}} (\mathbb{E}_{\mathcal{Q}}[\psi_{S_1}(Y) \psi_{S_2}(Y)] - \mathbb{E}_{\mathcal{Q}}[\psi_{S_1}(Y)] \mathbb{E}_{\mathcal{Q}}[\psi_{S_2}(Y)]) \\ &\leq \sum_{\substack{S_1, S_2 \in \mathcal{S}, \\ S_1 \cap S_2 \neq \emptyset}} \mathbb{E}_{\mathcal{Q}}[\psi_{S_1 \cup S_2}(Y)] \\ &= \sum_{\substack{S_1, S_2 \in \mathcal{S}, \\ S_1 \cap S_2 \neq \emptyset}} q^{|S_1 \cup S_2|} = q^{2m} \sum_{\substack{S_1, S_2 \in \mathcal{S}, \\ S_1 \cap S_2 \neq \emptyset}} q^{-|S_1 \cap S_2|} \end{aligned} \quad (26)$$

Let us fix  $S_1$ . We will consider bounding

$$\sum_{\substack{S_2 \in S, \\ S_1 \cap S_2 \neq \emptyset}} q^{-|S_1 \cap S_2|} \leq \sum_{m'=1}^m \sum_{\substack{S' \subseteq S_1, \\ |S'|=m'}} \sum_{\substack{S_2 \in S, \\ S' \subseteq S_2}} q^{-m'}. \quad (27)$$

Let us first consider a fixed  $S'$  and bound the number of choices for  $S_2$  in the inner-most sum. There are at most  $n^{\ell-|V(S')|}$  choices for the remaining vertices of  $S_2$  and at most  $\ell^{r(m-m')}$  choices for the remaining edges  $S_2 \setminus S'$ . We note that, as  $S' \subseteq S_1$  and  $K_n[S_1]$  is isomorphic to a balanced hypergraph  $\tilde{H}$ , we have

$$\frac{|S'|}{|V(S')|} \leq \frac{m}{\ell} \Rightarrow |V(S')| \geq \frac{\ell|S'|}{m}.$$

In particular, we can bound (27) by

$$\begin{aligned} & \sum_{\substack{S_2 \in S, \\ S_1 \cap S_2 \neq \emptyset}} q^{-|S_1 \cap S_2|} \\ & \leq \sum_{m'=1}^m \sum_{\substack{S' \subseteq S_1, \\ |S'|=m'}} n^{\ell-|V(S')|} \ell^{r(m-m')} q^{-m'} \\ & \leq n^{\ell} \ell^{rm} \sum_{m'=1}^m \left( \frac{m}{\ell^r n^{\ell/m} q} \right)^{m'}, \end{aligned}$$

where we use the fact that there are at most  $m^{m'}$  choices for  $S'$ , given  $S_1$ . Note that the sum above is dominated by either the first or last term. Therefore, we get

$$\sum_{\substack{S_2 \in S, \\ S_1 \cap S_2 \neq \emptyset}} q^{-|S_1 \cap S_2|} \leq mn^{\ell} \ell^{rm} \max \left\{ \frac{m}{\ell^r n^{\ell/m} q}, \left( \frac{m}{\ell^r n^{\ell/m} q} \right)^m \right\}.$$

With this in hand, we return to our original computation of the variance to get

$$\begin{aligned} \text{Var}_Q(T) & \leq m N q^{2m} n^{\ell} \ell^{rm} \max \left\{ \frac{m}{\ell^r n^{\ell/m} q}, \left( \frac{m}{\ell^r n^{\ell/m} q} \right)^m \right\} \\ & = \max \{ m^2 N n^{\ell(1-1/m)} \ell^{r(m-1)} q^{2m-1}, m^{m+1} N q^m \}, \end{aligned}$$

as desired.  $\square$

Before studying the expectation and variance of  $T$  under  $\mathcal{P}$ , we establish the following lemma.

**Lemma 4.3.** Fix  $S_1, S_2 \in S$  as defined in (25). Assume (23) and (24). Then we have

$$\mathbb{E}_{\mathcal{P}}[\psi_{S_1 \cup S_2}(Y)] \leq 2^{2\ell} \rho^{|V(S_1 \cup S_2)|} p^{|S_1 \cup S_2|}.$$

*Proof.* By considering all possible realizations of  $V(S_1 \cup S_2) \cap Z$ , we have

$$\begin{aligned} & \mathbb{E}_{\mathcal{P}}[\psi_{S_1 \cup S_2}(Y)] \\ & = \sum_{V' \subseteq V(S_1 \cup S_2)} \mathcal{P}(V(S_1 \cup S_2) \cap Z = V') \mathcal{P}(\psi_{S_1 \cup S_2}(Y)) \\ & = 1|V(S_1 \cup S_2) \cap Z = V'| \\ & \leq \sum_{V' \subseteq V(S_1 \cup S_2)} \rho^{|V'|} p^{|E(V')|} q^{|S_1 \cup S_2| - |E(V')|}, \end{aligned}$$

where  $E(V')$  consists of the edges in  $S_1 \cup S_2$  whose vertices are contained entirely in  $V'$ . We claim that the summand above is

maximized when  $V' = V(S_1 \cup S_2)$ . Given the claim, each summand is bounded by  $\rho^{|V(S_1 \cup S_2)|} p^{|S_1 \cup S_2|}$ , and there are at most  $2^{2\ell}$  choices of  $V'$ . Thus, the result follows.

To prove the claim, we first note that if  $V'$  maximizes  $\rho^{|V'|} p^{|E(V')|} q^{|S_1 \cup S_2| - |E(V')|}$ , then the subgraph  $(V', E(V'))$  has no isolated vertices. This is because removing isolated vertices from  $V'$  only decreases the exponent in  $\rho$ , which then increases the overall value as  $\rho < 1$ . As a result, we may instead consider the problem

$$\max_{S' \subseteq S_1 \cup S_2} \rho^{|V(S')|} p^{|S'|} q^{|S_1 \cup S_2| - |S'|},$$

where  $V(S')$  is the set of vertices induced by  $S'$ . We will show that the maximizer is  $S' = S_1 \cup S_2$ . We make the following definitions:

$$S := S_1 \cup S_2, \quad S'_1 := S' \cap S_1, \quad S'_2 := S' \setminus S'_1.$$

In particular,  $S'_i \subseteq S_i$  and  $S'_2 \cap S'_1 = S'_2 \cap S_1 = \emptyset$ . The goal is to show the following:

$$\rho^{|V(S)|} p^{|S|} \geq \rho^{|V(S')|} p^{|S'|} q^{|S| - |S'|},$$

which is equivalent to showing

$$\rho^{|V(S)| - |V(S')|} \left( \frac{p}{q} \right)^{|S| - |S'|} \geq 1. \quad (28)$$

Let us first consider the exponent of  $p/q$ . We have

$$\begin{aligned} |S| - |S'| & = |S_1| + |S_2| - |S_1 \cap S_2| - |S'_1| - |S'_2| \\ & = |S_1| - |S'_1| + |S_2| - |(S_1 \cap S_2) \cup S'_2|, \end{aligned}$$

as  $S'_2 \cap S_1 \cap S_2 = \emptyset$ . Next, let us consider the exponent of  $\rho$ . We have

$$\begin{aligned} |V(S)| - |V(S')| & = |V(S_1)| + |V(S_2)| - |V(S_1 \cap S_2)| \\ & \quad - |V(S'_1)| - |V(S'_2)| + |V(S'_1) \cap V(S'_2)| \\ & = |V(S_1)| - |V(S'_1)| + |V(S_2)| - |V(S_1 \cap S_2)| \\ & \quad - |V(S'_2)| + |V(S'_1) \cap V(S'_2)|. \end{aligned}$$

We note the following:

$$V(S'_1) \cap V(S'_2) \subseteq V(S_1) \cap V(S_2) \cap V(S'_2),$$

as  $S'_i \subseteq S_i$ . Therefore,

$$\begin{aligned} & |V(S)| - |V(S')| \\ & \leq |V(S_1)| - |V(S'_1)| + |V(S_2)| - |V(S_1 \cap S_2)| - |V(S'_2)| \\ & \quad + |V(S_1 \cap S_2) \cap V(S'_2)| \\ & \leq |V(S_1)| - |V(S'_1)| + |V(S_2)| - |(V(S_1 \cap S_2) \cup V(S'_2))| \\ & \leq |V(S_1)| - |V(S'_1)| + |V(S_2)| - |V((S_1 \cap S_2) \cup S'_2)|, \end{aligned}$$

where the last inequality follows from the fact that

$$\begin{aligned} V((S_1 \cap S_2) \cup S'_2) & = V(S_1 \cap S_2) \cup V(S'_2) \\ & \subseteq (V(S_1) \cap V(S_2)) \cup V(S'_2). \end{aligned}$$



We can now simplify the LHS of (28) to get:

$$\begin{aligned} & \rho^{|V(S)|-|V(S')|} \left(\frac{p}{q}\right)^{|S|-|S'|} \\ & \geq \rho^{|V(S_1)|-|V(S'_1)|+|V(S_2)|-|V((S_1 \cap S_2) \cup S'_2)|} \\ & \quad \left(\frac{p}{q}\right)^{|S_1|-|S'_1|+|S_2|-|(S_1 \cap S_2) \cup S'_2|} \\ & = \rho^{|V(S_1)|-|V(S'_1)|} \left(\frac{p}{q}\right)^{|S_1|-|S'_1|} \rho^{|V(S_2)|-|V((S_1 \cap S_2) \cup S'_2)|} \\ & \quad \left(\frac{p}{q}\right)^{|S_2|-|(S_1 \cap S_2) \cup S'_2|}. \end{aligned}$$

Hence, it suffices to show that

$$\rho^{|V(S_1)|-|V(S'_1)|} \left(\frac{p}{q}\right)^{|S_1|-|S'_1|} \geq 1 \quad (29a)$$

$$\rho^{|V(S_2)|-|V((S_1 \cap S_2) \cup S'_2)|} \left(\frac{p}{q}\right)^{|S_2|-|(S_1 \cap S_2) \cup S'_2|} \geq 1 \quad (29b)$$

We now prove (29a). The proof of (29b) is the same once  $S_1$  is replaced by  $S_2$  and  $S'_1$  is replaced by  $(S_1 \cap S_2) \cup S'_2$ . First, if  $V(S'_1) = V(S_1)$ , then the LHS of (29a) can be simplified to  $(\frac{p}{q})^{|S_1|-|S'_1|}$ , which is at least 1 as  $p > q$ . Therefore, we may assume that  $V(S'_1) \subsetneq V(S_1)$ . Recall that  $K_n^r[S_1]$  is isomorphic to  $\tilde{H}$ , a balanced hypergraph with  $\ell$  vertices and  $m$  edges. By Claim 4.1, we have

$$\begin{aligned} & \rho^{|V(S_1)|-|V(S'_1)|} \left(\frac{p}{q}\right)^{|S_1|-|S'_1|} \\ & = \left( \rho \left(\frac{p}{q}\right)^{\frac{|S_1|-|S'_1|}{|V(S_1)|-|V(S'_1)|}} \right)^{|V(S_1)|-|V(S'_1)|} \\ & \geq \left( \rho \left(\frac{p}{q}\right)^{m/\ell} \right)^{|V(S_1)|-|V(S'_1)|}. \end{aligned}$$

Note that

$$\rho \left(\frac{p}{q}\right)^{m/\ell} = n^{\gamma-1+(\beta-\alpha)m/\ell}.$$

It is enough to show the exponent above is nonnegative, which holds if  $(\beta - \alpha)m \geq (1 - \gamma)\ell$ . From (23) and (24), we have

$$(\beta - \alpha)m > (\beta - \beta\gamma)m = (1 - \gamma)\beta m > (1 - \gamma)\ell,$$

as desired.  $\square$

Let us now bound expectation and variance of  $T$  under  $\mathcal{P}$ .

**Lemma 4.4.** *With  $T$  as defined in (25), we have*

$$\mathbb{E}_{\mathcal{P}}[T] \geq N \rho^\ell p^m, \quad \text{Var}_{\mathcal{P}}(T) \leq 8^\ell N n^{\ell-1} \ell^{1+rm} \rho^{2\ell-1} p^{2m-m/\ell},$$

where  $N = |S|$ .

*Proof.* As  $Y_e$  and  $Y_f$  are conditionally independent for  $e \neq f$  given  $Z$ , we note that for a specific  $S \in \mathcal{S}$ , we have

$$\mathcal{P}(\psi_S(Y) = 1 | V(S) \subseteq Z) = p^m.$$

In addition,  $\psi_S(Y) \in \{0, 1\}$ , so we have

$$\mathbb{E}_{\mathcal{P}}[T] \geq \sum_{S \in \mathcal{S}} \mathcal{P}(V(S) \subseteq Z) \mathcal{P}(\psi_S(Y) = 1 | V(S) \subseteq Z) = N \rho^\ell p^m.$$

For the variance, we note that  $z_i$  and  $z_j$  are independent for  $i \neq j$  and so  $\psi_{S_1}(Y)$  and  $\psi_{S_2}(Y)$  are independent if  $V(S_1) \cap V(S_2) = \emptyset$ . Using again the argument that gives (26), we obtain

$$\begin{aligned} \text{Var}_{\mathcal{P}}(T) & = \sum_{\substack{S_1, S_2 \in \mathcal{S}, \\ V(S_1) \cap V(S_2) \neq \emptyset}} \text{Cov}_{\mathcal{P}}(\psi_{S_1}(Y), \psi_{S_2}(Y)) \\ & \leq \sum_{\substack{S_1, S_2 \in \mathcal{S}, \\ V(S_1) \cap V(S_2) \neq \emptyset}} \mathbb{E}_{\mathcal{P}}[\psi_{S_1 \cup S_2}(Y)]. \end{aligned}$$

By Lemma 4.3, we then get

$$\begin{aligned} \text{Var}_{\mathcal{P}}(T) & \leq \sum_{\substack{S_1, S_2 \in \mathcal{S}, \\ V(S_1) \cap V(S_2) \neq \emptyset}} 2^{2\ell} \rho^{|V(S_1 \cup S_2)|} p^{|S_1 \cup S_2|} \\ & = 2^{2\ell} \rho^{2\ell} p^{2m} \sum_{\substack{S_1, S_2 \in \mathcal{S}, \\ V(S_1) \cap V(S_2) \neq \emptyset}} \rho^{-|V(S_1) \cap V(S_2)|} p^{-|S_1 \cap S_2|} \end{aligned} \quad (30)$$

where we use the facts

$$\begin{aligned} |S_1 \cup S_2| & = 2m - |S_1 \cap S_2|, \quad |V(S_1 \cup S_2)| = \\ |V(S_1) \cup V(S_2)| & = 2\ell - |V(S_1) \cap V(S_2)|. \end{aligned}$$

As in the proof of Lemma 4.2, let us fix  $S_1$  and bound the following:

$$\begin{aligned} & \sum_{\substack{S_2 \in \mathcal{S}, \\ V(S_1) \cap V(S_2) \neq \emptyset}} \rho^{-|V(S_1) \cap V(S_2)|} p^{-|S_1 \cap S_2|} \\ & \leq \sum_{\ell'=1}^{\ell} \sum_{\substack{V' \subseteq V(S_1), \\ |V'|=\ell'}} \sum_{\substack{S_2 \in \mathcal{S}, \\ V' \subseteq V(S_2)}} \rho^{-\ell'} p^{-|S_1 \cap S_2|} \end{aligned} \quad (31)$$

Let us first consider a fixed  $V'$  and bound the number of choices for  $S_2$  in the inner-most sum. There are at most  $n^{\ell-\ell'}$  choices for the remaining vertices in  $S_2$  and at most  $\ell^{rm}$  possible choices for the edges in  $S_2$ . Note that  $V(S_1 \cap S_2) \subseteq V(S_1) \cap V(S_2)$ . Furthermore, as  $S_1 \cap S_2 \subseteq S_2$  and  $K_n^r[S_2]$  is a balanced hypergraph, we have

$$\frac{|S_1 \cap S_2|}{|V(S_1) \cap V(S_2)|} \leq \frac{|S_1 \cap S_2|}{|V(S_1 \cap S_2)|} \leq \frac{m}{\ell}.$$

In particular, we can bound (31) by

$$\begin{aligned} & \sum_{\substack{S_2 \in \mathcal{S}, \\ V(S_1) \cap V(S_2) \neq \emptyset}} \rho^{-|V(S_1) \cap V(S_2)|} p^{-|S_1 \cap S_2|} \\ & \leq \sum_{\ell'=1}^{\ell} \sum_{\substack{V' \subseteq V(S_1), \\ |V'|=\ell'}} n^{\ell-\ell'} \ell^{rm} \rho^{-\ell'} p^{-m\ell'/\ell} \\ & \leq n^{\ell} \ell^{rm} \sum_{\ell'=1}^{\ell} \left( \frac{\ell}{n \rho p^{m/\ell}} \right)^{\ell'}, \end{aligned}$$

where we use the fact that there are at most  $\ell^{\ell'}$  choices for  $V'$  given  $S_1$ . Let us consider the term in the sum. We have

$$\frac{\ell}{n \rho p^{m/\ell}} = \ell n^{am/\ell-\gamma}.$$

Note that as a result of (24), we have

$$\alpha \frac{m}{\ell} - \gamma < 0.$$

In particular, we have for some small constant  $\delta = \delta(\alpha, \beta, \gamma) > 0$ :

$$\frac{\ell}{n \rho p^{m/\ell}} \leq \ell n^{-\delta} < \frac{1}{2}.$$

Finally, we can bound (31) as follows:

$$\sum_{\substack{S_2 \in \mathcal{S}_2 \\ V(S_1) \cap V(S_2) \neq \emptyset}} \rho^{-|V(S_1) \cap V(S_2)|} p^{-|S_1 \cap S_2|} \leq 2n^\ell \ell^{rm} \left( \frac{\ell}{n \rho p^{m/\ell}} \right).$$

With this in hand, we return to our original computation for the variance in (30) to get

$$\begin{aligned} \text{Var}_P(T') &\leq 2^{2\ell+1} N n^\ell \ell^{rm} \rho^{2\ell} p^{2m} \left( \frac{\ell}{n \rho p^{m/\ell}} \right) \\ &\leq 8^\ell N n^{\ell-1} \ell^{1+rm} \rho^{2\ell-1} p^{2m-m/\ell}, \end{aligned}$$

as desired.  $\square$

**Proof of Proposition 4.3.** Let  $\lambda := N \rho^\ell p^m$ . By Lemma 4.4, we have  $\mathbb{E}_P[T] \geq \lambda$ . Therefore, it suffices to show that

1.  $\mathbb{E}_Q[T] = o(\lambda)$ ,
2.  $\text{Var}_Q(T) = o(\lambda^2)$ ,
3.  $\text{Var}_P(T) = o(\lambda^2)$ .

Let us first consider  $\mathbb{E}_Q[T]$ . By Lemma 4.2, we have

$$\begin{aligned} \frac{\mathbb{E}_Q[T]}{\lambda} &= \frac{N q^m}{N \rho^\ell p^m} = n^{(1-\gamma)\ell - (\beta-\alpha)m} \\ &< n^{(1-\gamma)\ell - \beta(1-\gamma)m} = n^{(1-\gamma)(\ell - \beta m)} = o(1), \end{aligned}$$

where the bounds follow from the conditions  $\alpha < \beta\gamma$  in (23),  $\ell < m\beta$  in (24), and  $1 - \gamma > 0$ . This completes the proof of Condition 1.

Now, let us consider the bound on  $\text{Var}_Q(T)$  from Lemma 4.2. We note the following

$$N \geq \binom{n}{\ell} \geq \left( \frac{n}{\ell} \right)^\ell.$$

With this and the above bound on  $\mathbb{E}_Q[T]$ , we have

$$\frac{m^{m+1} \mathbb{E}_Q[T]}{\lambda^2} \ll \frac{m^{m+1}}{\lambda} \leq \frac{m^{m+1} \ell^\ell}{n^\ell \rho^\ell p^m} = n^{\alpha m - \gamma \ell + o(1)} = o(1),$$

because  $\alpha m < \gamma \ell$  by (24). Now, in order to prove Condition 2, it remains to bound the following:

$$\begin{aligned} \frac{m^2 N n^{\ell(1-1/m)} \ell^{r(m-1)} q^{2m-1}}{\lambda^2} &= \frac{m^2 n^{\ell(1-1/m)} \ell^{r(m-1)} q^{2m-1}}{N \rho^{2\ell} p^{2m}} \\ &\leq \frac{m^2 n^{-\ell/m} \ell^{\ell+r(m-1)} q^{2m-1}}{\rho^{2\ell} p^{2m}} \quad (32) \\ &= n^{-\ell/m + \beta + 2\ell(1-\gamma) - 2m(\beta-\alpha) + o(1)} \end{aligned}$$

Let us consider the exponent in (32). As a result of (23), we have

$$\begin{aligned} &-\ell/m + \beta + 2\ell(1-\gamma) - 2m(\beta-\alpha) \\ &= \left(1-\gamma - \frac{1}{2m}\right) 2\ell - \left(\beta \left(1 - \frac{1}{2m}\right) - \alpha\right) 2m \\ &< \left(1-\gamma - \frac{1}{2m}\right) 2\ell - 2m\beta \left(1-\gamma - \frac{1}{2m}\right) \\ &= 2 \left(1-\gamma - \frac{1}{2m}\right) (\ell - m\beta). \end{aligned}$$

Since  $\gamma < 1/2$  and  $m \geq 1$ , we have  $1 - \gamma - 1/(2m) > 0$ . In addition,  $\ell < m\beta$  as a result of (24). Hence, the quantity above is negative, and so (32) is  $o(1)$ . Putting it all together, we have proved  $\frac{\text{Var}_Q(T)}{\lambda^2} = o(1)$ , which is Condition 2.

Finally, let us consider the bound on  $\text{Var}_P(T)$  in Lemma 4.4. We have

$$\begin{aligned} \frac{\text{Var}_P(T)}{\lambda^2} &\leq \frac{8^\ell N n^{\ell-1} \ell^{1+rm} \rho^{2\ell-1} p^{2m-m/\ell}}{N^2 \rho^{2\ell} p^{2m}} \\ &\leq \frac{8^\ell \ell^{1+rm+\ell}}{n \rho p^{m/\ell}} = n^{\alpha m/\ell - \gamma + o(1)} = o(1), \end{aligned}$$

where we again use  $N \geq \left(\frac{n}{\ell}\right)^\ell$  and  $\alpha m/\ell - \gamma < 0$  by (24). This proves Condition 3.  $\square$

## Acknowledgments

We thank Guy Bresler and Aaron Potechin for helpful discussions. We are also grateful to the anonymous referees for carefully reading the manuscript and providing helpful comments and suggestions.

## Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

## References

1. M. Jerrum, "Large Cliques Elude the Metropolis Process," *Random Structures & Algorithms* 3, no. 4 (1992): 347–359.
2. R. Karp, "Probabilistic Analysis of Some Combinatorial Search Problems. Algorithms and Complexity: New Directions and Recent Results," (1976).
3. 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* (New York, NY: ACM, 2010).
4. U. Feige and M. Seltser, "On the Densest k-Subgraph Problem," 1997.
5. E. Arias-Castro and N. Verzelen, "Community Detection in Dense Random Networks," *Annals of Statistics* 42, no. 3 (2014): 940–969.
6. S. Fortunato, "Community Detection in Graphs," *Physics Reports* 486, no. 3–5 (2010): 75–174.
7. E. Chlamtác, M. Dinitz, C. Konrad, G. Kortsarz, and G. Rabanca, "The Densest k-Subhypergraph Problem," *SIAM Journal on Discrete Mathematics* 32, no. 2 (2018): 1458–1477.
8. L. Corinzia, P. Penna, W. Szpankowski, and J. Buhmann, "Statistical and Computational Thresholds for the Planted k-Densest Sub-Hypergraph Problem," in *International Conference on Artificial Intelligence and Statistics* (New York, NY: PMLR, 2022), 11615–11640.

9. Y. Luo and A. R. Zhang, "Tensor Clustering With Planted Structures: Statistical Optimality and Computational Limits," *Annals of Statistics* 50, no. 1 (2022): 584–613.
10. M. Brennan and G. Bresler, "Reducibility and Statistical-Computational Gaps From Secret Leakage," in *Conference on Learning Theory* (New York, NY: PMLR, 2020), 648–847.
11. S. Hopkins, *Statistical Inference and the Sum of Squares Method* (PhD thesis, Cornell University, 2018).
12. D. Kunisky, A. S. Wein, and A. S. Bandeira, "Notes on Computational Hardness of Hypothesis Testing: Predictions Using the Low-Degree Likelihood Ratio," in *Mathematical Analysis, Its Applications and Computation: ISAAC 2019* (New York, NY: Springer, 2022), 1–50.
13. C. Butucea and Y. I. Ingster, "Detection of a Sparse Submatrix of a High-Dimensional Noisy Matrix," *Bernoulli* 19, no. 5B (2013): 2652–2688.
14. B. Hajek, W. Yihong, and X. Jiaming, "Computational Lower Bounds for Community Detection on Random Graphs," in *Conference on Learning Theory* (New York, NY: PMLR, 2015), 899–928.
15. N. Verzelen and E. Arias-Castro, "Community Detection in Sparse Random Networks," *Annals of Applied Probability* 25, no. 6 (2015): 3465–3510.
16. M. Brennan, G. Bresler, and W. Huleihel, "Reducibility and Computational Lower Bounds for Problems With Planted Sparse Structure," in *Conference on Learning Theory* (New York, NY: PMLR, 2018), 48–166.
17. M. Brennan, G. Bresler, and W. Huleihel, "Universality of Computational Lower Bounds for Submatrix Detection," in *Conference on Learning Theory* (New York, NY: PMLR, 2019), 417–468.
18. B. P. W. Ames, "Guaranteed Recovery of Planted Cliques and Dense Subgraphs by Convex Relaxation," *Journal of Optimization Theory and Applications* 167 (2015): 653–675.
19. Y. Chen and X. Jiaming, "Statistical-Computational Tradeoffs in Planted Problems and Submatrix Localization With a Growing Number of Clusters and Submatrices," *Journal of Machine Learning Research* 17, no. 1 (2016): 882–938.
20. B. Hajek, W. Yihong, and X. Jiaming, "Information Limits for Recovering a Hidden Community," *IEEE Transactions on Information Theory* 63, no. 8 (2017): 4729–4745.
21. A. Montanari, "Finding One Community in a Sparse Graph," *Journal of Statistical Physics* 161 (2015): 273–299.
22. Z. Ma and W. Yihong, "Computational Barriers in Minimax Submatrix Detection," *Annals of Statistics* 43, no. 3 (2015): 1089–1116.
23. T. Schramm and A. S. Wein, "Computational Barriers to Estimation From Low-Degree Polynomials," *Annals of Statistics* 50, no. 3 (2022): 1833–1858.
24. B. Barak, S. Hopkins, J. Kelner, P. K. Kothari, A. Moitra, and A. Potechin, "A Nearly Tight Sum-Of-Squares Lower Bound for the Planted Clique Problem," *SIAM Journal on Computing* 48, no. 2 (2019): 687–735.
25. S. B. Hopkins, P. K. Kothari, A. Potechin, P. Raghavendra, T. Schramm, and D. Steurer, "The Power of Sum-Of-Squares for Detecting Hidden Structures," in *58th Annual Symposium on Foundations of Computer Science (FOCS)* (New York, NY: IEEE, 2017), 720–731.
26. S. B. Hopkins and D. Steurer, "Efficient Bayesian Estimation From Few Samples: Community Detection and Related Problems," in *58th Annual Symposium on Foundations of Computer Science (FOCS)* (New York, NY: IEEE, 2017), 379–390.
27. A. S. Bandeira, A. El Alaoui, S. Hopkins, T. Schramm, A. S. Wein, and I. Zadik, "The Franz-Parisi Criterion and Computational Trade-Offs in High Dimensional Statistics," *Advances in Neural Information Processing Systems* 35 (2022): 33831–33844.
28. A. Coja-Oghlan, O. Gebhard, M. Hahn-Klimroth, A. S. Wein, and I. Zadik, "Statistical and Computational Phase Transitions in Group Testing," in *Conference on Learning Theory* (New York, NY: PMLR, 2022), 4764–4781.
29. A. Bhaskara, M. Charikar, V. Guruswami, A. Vijayaraghavan, and Y. Zhou, "Polynomial Integrality Gaps for Strong SDP Relaxations of Densest k-Subgraph," in *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms* (New York, NY: SIAM, 2012), 388–405.
30. E. Chlamtác and P. Manurangsi, "Sherali-Adams Integrality Gaps Matching the Log-Density Threshold," arXiv preprint, arXiv:1804.07842 (2018).
31. C. Jones, A. Potechin, G. Rajendran, and J. Xu, "Sum-of-Squares Lower Bounds for Densest k-Subgraph," arXiv preprint, arXiv:2303.17506 (2023).
32. B. Applebaum, "Pseudorandom Generators With Long Stretch and Low Locality From Random Local One-Way Functions," in *Proceedings of the Forty-Fourth Annual ACM Symposium on Theory of Computing* (New York, NY: ACM, 2012), 805–816.
33. M. Yuan and Z. Shang, "Heterogeneous Dense Subhypergraph Detection," arXiv preprint, arXiv:2104.04047 (2021).
34. M. Yuan and Z. Shang, "Information Limits for Detecting a Subhypergraph," *Stat* 10, no. 1 (2021): e407.
35. M. Yuan and Z. Shang, "Sharp Detection Boundaries on Testing Dense Subhypergraph," *Bernoulli* 28, no. 4 (2022): 2459–2491.
36. L. Corinzia, P. Penna, L. Mondada, and J. M. Buhmann, "Exact Recovery for a Family of Community-Detection Generative Models," in *2019 IEEE International Symposium on Information Theory (ISIT)* (New York, NY: IEEE, 2019), 415–419.
37. T. Lanciano, A. Miyauchi, A. Fazzone, and F. Bonchi, "A Survey on the Densest Subgraph Problem and Its Variants," arXiv preprint, arXiv:2303.14467 (2023).
38. R. Vershynin, *High-Dimensional Probability: An Introduction With Applications in Data Science*, vol. 47 (Cambridge, England: Cambridge University Press, 2018).
39. A. Ruciński and A. Vince, "Strongly Balanced Graphs and Random Graphs," *Journal of Graph Theory* 10, no. 2 (1986): 251–264.
40. A. D. Matushkin and S. N. Popova, "Strictly Balanced Uniform Hypergraphs and Generalizations of Zero-One Law," *Discrete Mathematics* 345, no. 6 (2022): 112835.
41. A. S. Bandeira, J. Banks, D. Kunisky, C. Moore, and A. S. Wein, "Spectral Planting and the Hardness of Refuting Cuts, Colorability, and Communities in Random Graphs," in *Conference on Learning Theory* (New York, NY: PMLR, 2021), 410–473.
42. A. S. Bandeira, D. Kunisky, and A. S. Wein, "Computational Hardness of Certifying Bounds on Constrained PCA Problems," in *11th Innovations in Theoretical Computer Science Conference (ITCS 2020)*, vol. 151 (Wadern, Germany: Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2020), 78.
43. A. Bandeira, D. Kunisky, and A. Wein, "Average-Case Integrality Gap for Non-Negative Principal Component Analysis," in *Mathematical and Scientific Machine Learning* (New York, NY: PMLR, 2022), 153–171.
44. J. Banks, S. Mohanty, and P. Raghavendra, "Local Statistics, Semidefinite Programming, and Community Detection," in *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)* (Philadelphia, PA: SIAM, 2021), 1298–1316.
45. P. K. Kothari, S. S. Vempala, A. S. Wein, and J. Xu, "Is Planted Coloring Easier than Planted Clique?," arXiv preprint, arXiv:2303.00252 (2023).

46. T. Wang, Q. Berthet, and Y. Plan, “Average-Case Hardness of RIP Certification,” *Advances in Neural Information Processing Systems* 29 (2016): 3826–3834.

## Appendix A

### Detection-Refutation Gap

Here we justify the claim that for  $\gamma > 1/2$  we expect an inherent detection-refutation gap, implying that tests based on checking feasibility of a convex relaxation (e.g., the sum-of-squares program of [31]) are strictly suboptimal for detection. We focus throughout on the graph case  $r = 2$ .

We first formally define the refutation task. We say that a graph contains a “dense subgraph” if there is a subgraph on  $(1 \pm o(1))\rho n$  vertices with  $(1 \pm o(1))p(\rho n)^2/2$  edges, where each  $o(1)$  stands for some particular  $o(1)$  quantity chosen so that the planted distribution  $\mathcal{P}$  (defined in Section 2) will contain a dense subgraph with high probability (i.e., probability  $1 - o(1)$ ). Refutation is the following algorithmic task. Given a graph  $G$ , the goal is to output NO or MAYBE with the following two guarantees: (1) if  $G$  contains a dense subgraph then the output must be MAYBE, and (2) if  $G$  is drawn from the null distribution  $\mathcal{Q} = G(n, q)$ , the output must be NO with probability  $1 - o(1)$ . For intuition, it is crucial to note that the algorithm is only allowed to output NO if it has proven with absolute certainty that there is no dense subgraph. Refutation (or certification) tasks of this flavor have been considered in prior work, e.g., [41–46].

A natural approach to solve the refutation problem is to check feasibility of a convex relaxation for existence of a dense subgraph. If the relaxation is infeasible w.h.p. over  $\mathcal{Q}$  (for some parameters  $p, q, \rho$ ), we have a successful refutation algorithm: output NO if the relaxation is infeasible, and MAYBE otherwise. (Note that by virtue of being a relaxation, if the relaxation is infeasible then this proves there is no dense subgraph.) This also gives a successful detection algorithm: output “Q” if the relaxation is infeasible, and “P” otherwise. (Note that by design, the relaxation will be feasible w.h.p. over  $\mathcal{P}$ , because a dense subgraph exists.)

Our goal in this section is to argue that for  $\gamma > 1/2$ , the refutation problem is computationally hard whenever  $\alpha > \beta/2 + \gamma - 1/2$ , making it strictly harder than detection (because detection is easy whenever  $\alpha < \beta/2 + 2(\gamma - 1/2)$ ). As a result, we expect that *any* test based on checking feasibility of a convex relaxation (in the sense described above) cannot be optimal for detection, or else this would imply a too-good-to-be-true refutation algorithm. Indeed, if the relaxation succeeds at detection, it must be infeasible w.h.p. over  $\mathcal{Q}$ , meaning it also succeeds at refutation.

We will argue hardness of refutation in a manner similar to [42]: we will construct a *different* planted distribution  $\tilde{\mathcal{P}}$  that, like  $\mathcal{P}$ , contains a dense subgraph w.h.p.; we will then show low-degree hardness of distinguishing  $\tilde{\mathcal{P}}$  from  $\mathcal{Q}$ , leading us to conjecture that no polynomial-time algorithm can distinguish  $\tilde{\mathcal{P}}$  and  $\mathcal{Q}$ . This conjecture, if true, formally implies hardness of refutation because a successful refutation algorithm could be used to distinguish  $\tilde{\mathcal{P}}$  and  $\mathcal{Q}$  as discussed above.

To summarize, our goal for the rest of this section is to construct a distribution  $\tilde{\mathcal{P}}$  over graphs such that a dense subgraph exists w.h.p., and then prove that if  $\gamma > 1/2$  and  $\alpha > \beta/2 + \gamma - 1/2$  then no degree- $n^{o(1)}$  polynomial weakly separates  $\tilde{\mathcal{P}}$  and  $\mathcal{Q}$ .

### Construction of $\tilde{\mathcal{P}}$

We now construct an auxiliary planted distribution  $\tilde{\mathcal{P}}$  that has a planted dense subgraph but is more difficult to distinguish from  $\mathcal{Q} = G(n, q)$  than  $\mathcal{P}$  is. As in the main text, we fix parameters  $0 < \alpha < \beta < 1$  and  $\gamma \in (0, 1)$ , and consider the scaling  $p = n^{-\alpha}$ ,  $q = n^{-\beta}$ , and  $\rho = n^{\gamma-1}$ . It will be convenient to parametrize the observed graph in an unusual way: for  $i < j$ , let  $Y_{ij} = a := \sqrt{(1-q)/q}$  if edge  $(i, j)$  is present and  $Y_{ij} = b := -\sqrt{q/(1-q)}$  otherwise. (This ensures  $\mathbb{E}_{\mathcal{Q}}[Y_{ij}] = 0$  and  $\mathbb{E}_{\mathcal{Q}}[Y_{ij}^2] = 1$ .) A graph is sampled from  $\tilde{\mathcal{P}}$  as follows. As before, sample a set of planted vertices  $Z \subseteq [n]$  where each vertex is included independently with probability  $\rho$ . Define  $u \in \mathbb{R}^n$  by  $u_i = \sqrt{(1-\rho)/\rho}$  if  $i \in Z$  and  $u_i = -\sqrt{\rho/(1-\rho)}$  if  $i \notin Z$ . (This

ensures  $\mathbb{E}[u_i] = 0$  and  $\mathbb{E}[u_i^2] = 1$ .) Now, conditioned on  $u$ , let  $Y = (Y_{ij})_{i < j}$  have independent entries  $Y_{ij} \in \{a, b\}$  such that  $\mathbb{E}_{\mathcal{P}}[Y_{ij}|u] = \lambda u_i u_j$ . Choose the scalar  $\lambda$  so that planted edges have the desired probability, that is, edge  $(i, j)$  occurs with probability  $p$  when  $i, j \in Z$ ; this gives  $\lambda = (1 + o(1))\rho/\sqrt{q}$ .

### Proof of Low-Degree Hardness

We now show that if  $\gamma > 1/2$ ,  $\alpha > \beta/2 + \gamma - 1/2$ , and  $D = n^{o(1)}$ , then  $\|\tilde{L}_{\leq D}\| = 1 + o(1)$ , which rules out weak separation as discussed in Section 3. By [41, Proposition B.1] along with the fact that revealing extra observations  $(Y_{ij})_{i \geq j}$  can only increase  $\|\tilde{L}_{\leq D}\|$ ,

$$\|\tilde{L}_{\leq D}\|^2 \leq \sum_{d=0}^D \frac{1}{d!} \mathbb{E} \langle \lambda u u^\top, \lambda v v^\top \rangle^d = \sum_{d=0}^D \frac{\lambda^{2d}}{d!} \mathbb{E} \langle u, v \rangle^{2d},$$

where  $u$  is defined above and  $v$  is an independent copy of  $u$ . Note that  $\langle u, v \rangle$  is a sum of  $n$  i.i.d. bounded random variables and following the proof of [23, Lemma B.3], we may conclude

$$\mathbb{E} \langle u, v \rangle^{2d} \leq \sqrt{2\pi} \rho^{-2d} [(4d\rho^2 n)^d + (8d/3)^{2d}].$$

Combining this with the above,  $\|\tilde{L}_{\leq D}\| = 1 + o(1)$  provided  $\lambda \ll \rho$  and  $\lambda^2 n \ll 1$ . Since  $\gamma > 1/2$ , this reduces to the condition  $\alpha > \beta/2 + \gamma - 1/2$  as desired.