The Edge Density Barrier: Computational-Statistical Tradeoffs in Combinatorial Inference

Hao Lu [1] **Yuan Cao** [1] **Junwei Lu** [2] **Han Liu** [3] **Zhaoran Wang** [4]

Abstract

We study the hypothesis testing problem of inferring the existence of combinatorial structures in undirected graphical models. Although there exist extensive studies on the information-theoretic limits of this problem, it remains largely unexplored whether such limits can be attained by efficient algorithms. In this paper, we quantify the minimum computational complexity required to attain the information-theoretic limits based on an oracle computational model. We prove that, for testing common combinatorial structures, such as clique, nearest neighbor graph and perfect matching, against an empty graph, or large clique against small clique, the information-theoretic limits are provably unachievable by tractable algorithms in general. More importantly, we define structural quantities called the weak and strong edge densities, which offer deep insight into the existence of such computational-statistical tradeoffs. To the best of our knowledge, our characterization is the first to identify and explain the fundamental tradeoffs between statistics and computation for combinatorial inference problems in undirected graphical models.

1. Introduction

One of the most important goals of statistical inference is to identify dependency structures among variables. In specific, given n realizations $\{x_i\}_{i=1}^n$ of a random vector $X \in \mathbb{R}^d$, we are interested in inferring the structures of the underlying graphical model. This problem plays a fundamental role in bioinformatics (Friedman, 2004), information retrieval (Welling et al., 2005), speech recognition (Bilmes & Bartels, 2005) and image processing (Murphy et al., 2004). In this paper, we focus on a more specific inference problem: testing whether the underlying graph has a certain combinatorial structure. For instance, we consider the graph associated with the precision matrix Θ of X, we aim to test, e.g., whether it is a clique of size s or is an empty graph (see §2 for more such examples). There exists a vast body of literature on efficient algorithms and fundamental information-theoretic limits for combinatorial inference problem. See, e.g., (Arias-Castro et al., 2008; Addario-Berry et al., 2010; Chen et al., 2012; Arias-Castro et al., 2012; Verzelen & Arias-Castro, 2013; Castro et al., 2014; Arias-Castro & Verzelen, 2014; Arias-Castro et al., 2015a;b; Neykov et al., 2016) and the references therein. However, under their settings, there is generally a lack of algorithms that are both computationally efficient and information-theoretically optimal. Consequently, it gives rise to the natural question of whether or not the information-theoretic limits can be attained by any efficient algorithms, which remains largely unexplored. Moreover, it is even less clear how the formulations of testing problems, especially the combinatorial structures of graphical models, affect the achievability of the information-theoretic limits. Our goal is to understand these two questions. In particular, we aim to characterize the fundamental limits for combinatorial inference in graphical model, particularly from the computational perspective. To study the minimum computational complexity required to attain the information-theoretic limits, we use the oracle computational model developed by (Kearns, 1998; Feldman et al., 2013; 2015a;b; 2017).

Contribution: First, based on the oracle computational model, we establish the unachievability of information-theoretic limits for several common combinatorial structures. As concrete examples, for precision graphs we consider testing clique, perfectly matched block and nearest neighbor graph structures against the empty graph, as well as large clique against small clique. In these examples, we identify a significant gap between the information-theoretic limit and the minimum signal strength that allows for tractable algorithms. This gap depicts the fundamental tradeoffs between

[1]Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ, USA [2]Department of Biostatistics, Harvard University, Boston, MA, USA [3]Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA [4]Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL, USA. Correspondence to: Zhaoran Wang <zhaoranwang@gmail.com>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

computational tractability and statistical optimality. As another example, we study testing hypercube structures against the empty graph. In detail, we establish the information-theoretic lower bound for algorithms with a certain sublinear computational complexity constraint.

Second, more importantly, we identify two critical quantities — the weak and strong edge densities μ and μ' (formally defined in §3) — to characterize the computational-statistical tradeoffs. In particular, we show that, if μ and μ' are of different orders, the information-theoretic limit is not achievable by any tractable algorithm. One striking property of these two quantities is that they only depend on the topology of the combinatorial structures to be tested. Therefore, they provide new insight on how the structural properties of a combinatorial inference problem dictate its computational complexity.

Related Work: Our work is in the same nature as a recent line of work on computational-statistical tradeoffs (Berthet & Rigollet, 2013a;b; Ma & Wu, 2014; Zhang et al., 2014; Hajek et al., 2014; Chen & Xu, 2014; Wang et al., 2014; Chen, 2015; Cai et al., 2015; Krauthgamer et al., 2015; Wang et al., 2015; Kannan & Vempala, 2016). In comparison, we focus on the combinatorial inference problem in undirected graphical models, for which the computational-statistical tradeoffs remain much less well understood. In addition to establishing the unachievability of information-theoretic limits for tractable algorithms as in previous work, in the example of hypercube detection, we also characterize such unachievability for the class of algorithms with sublinear computational complexity. Moreover, from an aspect not studied in previous work, we illustrate how the achievability of information-theoretic lower bounds for tractable algorithms are governed by the underlying combinatorial structures, especially the weak and strong edge densities. To characterize the computational complexity, in this paper we focus on the oracle computational model proposed by (Kearns, 1998) and generalized by (Feldman et al., 2013; 2015a;b), which covers a broad range of algorithms, such as convex optimization algorithms, local search, Markov chain Monte Carlo, moments-based methods, and most other learning algorithms. Correspondingly, compared with existing work, our theory does not rely on any unproven computational hardness conjectures, like the planted clique hypothesis. Meanwhile, under the same computational model as in this paper, (Bresler et al., 2014) studies the computational complexity of learning antiferromagnetic Ising models. In comparison with their results, we mainly focus on continuous random variables and Gaussian graphical models, which have fundamentally different computational-statistical phase transitions and theoretical difficulties.

Notation: For a matrix A, we denote $\mathrm{A}_{j\cdot}$ and $\mathrm{A}_{\cdot j}$ to be the j^{th} row and column of A, correspondingly. Let $||\mathrm{A}_{j\cdot}||_0$ be the number of non-zero entries in the j^{th} row of A. For a set $\mathcal{D}$, we denote $|\mathcal{D}|$ as its cardinality. For any positive integer n, we use $[n]$ as an abbreviation of $\{1, 2, \ldots, n\}$. For a graph G, we denote $V(G)$ and $E(G)$ as the vertex set and edge set of G respectively. For an edge set E, we denote $V(E)$ as the vertex set of E. For two functions f and g, we say $f(x) = O(g(x))$ if and only if there exists a positive number M and a real number x_0 such that $|f(x)| \leq M|g(x)|$ for all $x \geq x_0$. We say $f(x) \asymp g(x)$ if and only if $f(x) = O(g(x))$ and $g(x) = O(f(x))$.

Organization: In §2 we introduce Gaussian graphical model and the combinatorial inference problem studied in this paper. Then we define the oracle computational model. Our main result is presented in §3, where we establish a general computational lower bound, the corresponding hypothesis tests which match upper bounds, as well as their applications to clique, perfectly matched block and nearest neighbor graph. Furthermore, we establish and discuss the intrinsic link between structural properties and computational tractability.

2. Background

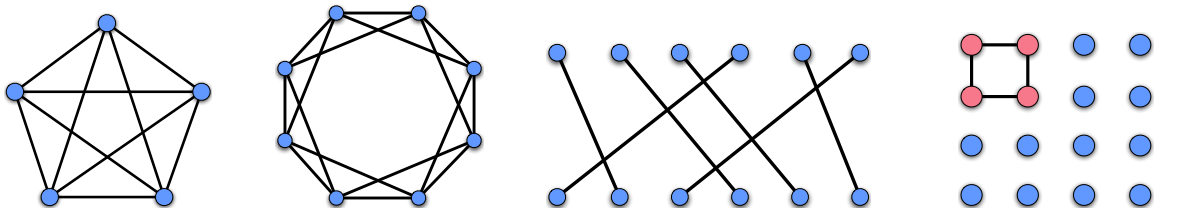
We first introduce Gaussian graphical model, then formalize the combinatorial inference problem and define the corresponding minimax testing risk. At last we introduce the oracle computational model.

2.1. Combinatorial Inference Problems

Let $X = (X_1, \ldots, X_d)^\top \in \mathbb{R}^d$ be a d-dimensional random vector following the multivariate normal distribution $N_d(0, \Theta^{-1})$, where $\Theta = (\theta_{jk})$ is the precision matrix. It is well known that the precision matrix Θ describes the conditional independence relationship among $X_1, \ldots, X_d$. More specifically, X_i and X_j are independent conditioning on all the rest variables if and only if $\theta_{jk} = 0$. We consider the undirected graph $G(\Theta) = (V, E)$, where the vertex set $V = [d]$ corresponds to the d entries of X, and $E = \{(j, k) : \theta_{jk} \neq 0\}$ is the edge set. In the sequel, we also denote $G(\Theta)$ as G when there is no ambiguity. Given n independent realizations $x_1, \ldots, x_n$ of X, the goal of combinatorial inference is to test whether the underlying graph G has certain structural properties. In specific, let $\mathcal{G}$ be the set of all possible graphs over the vertex set V. For two disjoint sets of graphs $\mathcal{G}_0 \subseteq \mathcal{G}$ and $\mathcal{G}_1 \subseteq \mathcal{G}$, we aim to test

$$\mathrm{H}_0 : G \in \mathcal{G}_0 \text{ versus } \mathrm{H}_1 : G \in \mathcal{G}_1. \tag{1}$$

Following a sequence of work on information-theoretic limits for combinatorial inference problem, e.g., (Addario-Berry et al., 2010; Arias-Castro et al., 2012; 2015a;b; Cai et al., 2015; Wang et al., 2015), in this paper we consider examples of combinatorial structures as follows.



(a) 5-clique (b) 4-nearest (c) perfect matching (d) hypercubeneighbor

Figure 1. Examples of graphs with different combinatorial structures: clique, nearest neighbor graph, perfect matching and hypercube with $d = 16, s = 4, D = 2$.

Clique: Let $\mathcal{G}_0 = \{(V, \emptyset)\}$ and $\mathcal{G}_1 = \{G_J : |J| = s \text{ and } J \subseteq V\}$. For each G_J, vertex i and vertex j are connected if and only if $i, j \in J$. Here $\mathcal{G}_1$ is the set of all cliques with size s.

Nearest Neighbor Graph: Let $\mathcal{G}_0 = \{(V, \emptyset)\}$ and $\mathcal{G}_1 = \{G_I : I = \{i_1, i_2, \ldots, i_s\} \subseteq V \text{ and } i_\ell < i_k, \ell < k\}$. For each G_I, first define a path L_I with edge set $\{(i_j, i_{j+1}) | j \in [s-1]\} \cup (i_s, i_1)$, then j and k are connected in G_I if and only if $j, k \in I$ and the distance between j and k is less or equal to $s/4$[1]. Here the distance is defined with respect to the geodesic distance in the path L.

Perfectly Matched Block: Let $\mathcal{G}_0 = \{(V, \emptyset)\}$ and $\mathcal{G}_1 = \{G : G \text{ is an } \sqrt{d}\text{-perfectly matched block}\}$. An s-perfectly matched block is a certain type of s-cliques. For each vertex j in a clique, let $j = (k-1)s + k'$ correspond to the edge between node $k \in [s]$ and $k' \in [s]$ in a bipartite graph with s nodes within each partition. If these s vertices correspond to a perfect matching of the bipartite graph, we call it a perfectly matched block.

Hypercube: Let V be a D-dimensional lattice[2] with d vertices. We consider $\mathcal{G}_0 = \{(V, \emptyset)\}$ and $\mathcal{G}_1$ contains all the s-lattice graph whose vertex sets are disjoint from each other and each of these vertex sets is a D-dimensional lattice graph with s vertices.

Small Clique vs Large Clique: Let $\mathcal{G}_0$ be the set of all cliques with size s' and $\mathcal{G}_1$ be the set of all cliques with size s. Here s' is a constant number which will not grow up with d.

All these combinatorial structures are illustrated in Figure 1. It is worth mentioning that in previous work the null hypothesis is always a simple hypothesis. But in our paper we propose a more general framework where the null hypothesis and alternative hypothesis can both be composite.

[1]Without loss of generality we assume $s/4$ is an integer.

[2]A D-dimensional lattice is a set of form $\{1, \ldots, k\}^D$ for some integer k. A D-dimensional lattice graph has the vertex set $\{1, \ldots, k\}^D$ for some k and any two vertices $u, v \in \{1, \ldots, k\}^D$ are connected if and only if their Euclidian distance $\|u - v\| = 1$.

Based on the correspondence between $G(\Theta)$ and Θ, we can reformulate the hypothesis testing problem defined in (1) as testing

$$\mathrm{H}_0 : \Theta \in \mathcal{C}_0 \text{ versus } \mathrm{H}_1 : \Theta \in \mathcal{C}_1. \tag{2}$$

Here $\mathcal{C}_0$ and $\mathcal{C}_1$ are defined as the sets of precision matrices such that $G(\Theta) \in \mathcal{G}_0, \mathcal{G}_1$ for $\Theta \in \mathcal{C}_0, \mathcal{C}_1$, respectively. In the general settings, we constrain $\mathcal{C}_0$ and $\mathcal{C}_1$ into a subspace $\mathcal{M}$, which satisfies

$$\mathcal{M} = \Big\{\Theta \in \mathbb{R}^{d\times d} : \Theta = \Theta^\top, \\ \theta_{jj} = 1 \text{ for } j \in [d], \min_{\theta_{jk}\neq 0} \theta_{jk} \geq \theta\Big\}, \tag{3}$$

where θ is the signal strength. In the sequel, we define the minimax testing risk for hypothesis testing problems. For any test $\psi : \mathbb{R}^{n\times d} \to \{0, 1\}$, let its output be zero if the null hypothesis H_0 is accepted and one otherwise. We define the minimax risk $R_n(\mathcal{C}_0, \mathcal{C}_1)$ as

$$R_n(\mathcal{C}_0, \mathcal{C}_1) \\ = \inf_\psi \Big[\sup_{\Theta\in\mathcal{C}_0} \mathbb{P}_\Theta(\psi = 1) + \sup_{\Theta\in\mathcal{C}_1} \mathbb{P}_\Theta(\psi = 0)\Big]. \tag{4}$$

Here we use $\mathbb{P}_\Theta$ to represent the distribution $N_d(0, \Theta^{-1})$. Note that if

$$\liminf_{n\to\infty} R_n(\mathcal{C}_0, \mathcal{C}_1) = 1, \tag{5}$$

then any test is asymptotically powerless, which means the problem defined in (2) is unsolvable.

2.2. Oracle Computational Model

In this section we introduce the oracle computational model (Kearns, 1998; Feldman et al., 2013; 2015a;b; Wang et al., 2015) to characterize the computational complexity. The intuition for this model is based on following facts. When solving statistical problems, any algorithm can be seen as a sequence of interactions with data. Generally, we say an algorithm is more complex if it needs more rounds of interactions with data. Therefore, the total number of rounds

of interactions between data and the algorithm can be a good proxy for the computational complexity of this algorithm.

Under this model, each round of interaction is interpreted as follows. The algorithm sends a query $q \in \mathcal{Q}$ to an oracle r, where $\mathcal{Q}$ is the query space defined in (6), and get a random response Z_q concentrated around $\mathbb{E}[q(X)]$. See Definition 1 and 3 for details.

Now we define the oracle computational model.

Definition 1 (Oracle Computational Model)**.** *An algorithm $\mathscr{A}$ under oracle computational model M is defined as a tuple $\mathscr{A} = M(\mathcal{Q}, T, q_{\text{init}}, \delta)$, where*

- *$\mathcal{Q}$ is the query space.*
- *T is the maximum number of rounds the algorithm queries an oracle.*
- *$q_{init} \in \mathcal{Q}$ is the initial query.*
- *$\delta : \mathcal{Q} \times \mathbb{R} \to (\mathcal{Q} \times \mathbb{R}) \cup \{HALT\}$ is the transition function, which means the next query can be determined by the previous query and its return. If δ returns HALT, then the model stops querying the oracle.*

Let X be the random vector of interest and $\mathcal{X}$ be the domain of X. In this paper we consider the query space $\mathcal{Q}$ as

$$\mathcal{Q} = \{q : \mathcal{X} \to \mathbb{R}, q(X) \text{ is a sub-exponential variable}\}. \tag{6}$$

By this definition, we can see that the oracle computational model covers a broad range of algorithms, such as convex optimization algorithms, local search, Markov chain Monte Carlo, moments-based methods, and most other learning algorithms. See (Wang et al., 2015) for more discussion. Next we define the computational complexity of a certain algorithm under oracle computational model.

Definition 2 (Computational Complexity)**.** *For an algorithm $\mathscr{A} = M(\mathcal{Q}, T, q_{\text{init}}, \delta)$, let $\mathcal{Q}_{\mathscr{A}}$ be the set of all queries that $\mathscr{A}$ queries the the oracle. We say $|\mathcal{Q}_{\mathscr{A}}|$ is the computational complexity of $\mathscr{A}$.*

From Definition 1 we know $|\mathcal{Q}_{\mathscr{A}}| \leq T$. Therefore the computational complexity of algorithm $\mathscr{A} = M(\mathcal{Q}, T, q_{\text{init}}, \delta)$ is upper bounded by T. Therefore, under this model any polynomial-time algorithm $\mathscr{A}$ can be written as $\mathscr{A} = M(\mathcal{Q}, T, q_{\text{init}}, \delta)$ with $T = d^a$ for some constant a.

Now given any tail probability $\xi \in [0, 1)$ and sample size n, we define the statistical query oracle as following.

Definition 3 (Statistical Query Oracle)**.** *A statistical query oracle r interacts with a query $q \in \mathcal{Q}$ at each round, and returns an output $Z_q \in \mathbb{R}$ which is a realization of $q(X)$, satisfying*

$$\mathbb{P}\left(\bigcap_{q \in \mathcal{Q}_{\mathscr{A}}} \left\{ |Z_q - \mathbb{E}[q(X)]| \leq \|q(X)\|_{\psi_1} \cdot \tau \right\} \right) \geq 1 - 2\xi,$$

and τ is defined as

$$\tau = \max \left\{ \frac{\eta(\mathcal{Q}_{\mathscr{A}}) + \log(1/\xi)}{n}, \sqrt{\frac{2[\eta(\mathcal{Q}_{\mathscr{A}}) + \log(1/\xi)]}{n}} \right\}. \tag{7}$$

Here we use $||q(X)||_{\psi_1}$ to denote the Orlicz ψ-norm of the random variable $q(X)$ and define $\eta(\mathcal{Q}_{\mathscr{A}}) = \log(|\mathcal{Q}_{\mathscr{A}}|)$ when $\mathcal{Q}_{\mathscr{A}}$ is countable. Recall the definition of query space $\mathcal{Q}$ in (6), here the condition (7) is actually a union bound version Bernstein's inequality if we take Z_q as the average of $q(x_i), i = 1, \ldots, n$, where $\{x_i\}_{i=1}^n$ are realizations of X. This implies that the deviation behavior of Z_q under the statistical query oracle is achievable. However, in most algorithmic analysis of statistical problems, like principal component analysis (Yuan & Zhang, 2013) and latent variable model estimation (Balakrishnan et al., 2017), inequality (7) is also the optimal deviation behavior we can get.

Compared with the minimax risk defined in (4), we define a new minimax risk of testing $\mathcal{C}_0$ against $\mathcal{C}_1$ with oracle r under the oracle computational model as

$$\begin{aligned} &R_n(\mathcal{C}_0, \mathcal{C}_1, \mathscr{A}, r) \\ &= \inf_{\psi \in \mathcal{H}(\mathscr{A}, r)} \left[\sup_{\Theta \in \mathcal{C}_0} \mathbb{P}_\Theta(\psi = 1) + \sup_{\Theta \in \mathcal{C}_1} \mathbb{P}_\Theta(\psi = 0) \right]. \end{aligned} \tag{8}$$

Here n is the sample size and $\mathcal{H}(\mathscr{A}, r)$ is the space of all possible tests based on algorithm $\mathscr{A}$ defined in 1 and oracle r defined in 3. Following the same idea in (5) we know that if there exists an oracle r such that

$$\liminf_{n \to \infty} R_n(\mathcal{C}_0, \mathcal{C}_1, \mathscr{A}, r) = 1, \tag{9}$$

then any hypothesis test under the oracle computational model is asymptotically powerless.

3. Main Results

In this section, we first provide a general computational lower bound. We will show that the computational lower bound for testing graph structures can be determined by two topological features of the graph: the weak edge density and the vertex cut ratio. Then by comparing the computational lower bound and the information-theoretic lower bound, we point out that the edge density is a critical structural property which determines the computational-statistical tradeoffs.

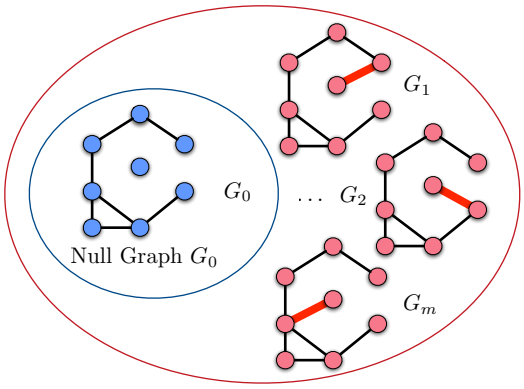


Figure 2. An example of null-alternative separator. Let $\mathcal{G}_0$ and $\mathcal{G}_1$ be the sets of disconnected and connected graphs respectively. For the null base G_0 shown in the figure, all the red thick edges form a null-alternative separator.

3.1. A General Computational Lower Bound

Before introducing the main theorem, we first define the null-alternative separator, which is originally defined in (Neykov et al., 2016).

Definition 4 (Null-Alternative Separator)**.** *Let $G_0 = (V, E_0) \in \mathcal{G}_0$ be some graph under the null. We call a collection of edge sets $\mathcal{E}$ a null-alternative separator with null base G_0 if for all edge sets $S \in \mathcal{E}$ we have $S \cap E_0 = \emptyset$ and $(V, E_0 \cup S) \in \mathcal{G}_1$.*

The reason that we consider the null-alternative separator comes from following facts. Given a $G_0 = (V, E_0) \in \mathcal{G}_0$ and a null-alternative separator $\mathcal{E}$ with null base G_0. we consider precision matrices Θ_0 and Θ_S as

$$\Theta_0 = \mathrm{I} + \theta \mathrm{A}_0, \quad \Theta_S = \mathrm{I} + \theta(\mathrm{A}_0 + \mathrm{A}_S),$$

where A_0 and A_S are the adjacency matrices of G_0 and $G_S = (V, S)$ respectively. Recall the minimax risk defined in (4) we have

$$R_n(\mathcal{C}_0, \mathcal{C}_1, \mathscr{A}, r) \geq R_n(\{\Theta_0\}, \{\Theta_S : S \in \mathcal{E}\}, \mathscr{A}, r). \tag{10}$$

However, in combinatorial inference, we know $\mathcal{C}_0$ and $\mathcal{C}_1$ have some symmetric structures. Then we can expect that $R_n(\mathcal{C}_0, \mathcal{C}_1, \mathscr{A}, r)$ and $R_n(\{\Theta_0\}, \{\Theta_S : S \in \mathcal{E}\}, \mathscr{A}, r)$ in (10) have the same order. Actually in the main theorem we can prove that they are equal.

From the above observation we find that the computational lower bound only depends on the structure of null-alternative separator $\mathcal{E}$. Here the null-alternative separator contains critical edge sets which may change a graph G_0 from the null to alternatives. Figure 2 gives an example of null-alternative separator.

With the definition of null-alternative separator, we are ready to define two important features of null-alternative separator: the weak edge density and the vertex cut ratio.

Definition 5 (Weak Edge Density)**.** *For a null-alternative separator $\mathcal{E}$, we define its edge density as*

$$\mu = \max_{S, S' \in \mathcal{E}} \frac{|S \cap S'|}{|V(S \cap S')|^2}.$$

Here we let $0/0 = 0$.

By its definition, the weak edge density of $\mathcal{E}$ characterizes how dense these critical edges which can change graphs from $\mathcal{G}_0$ to $\mathcal{G}_1$ are. Therefore, μ can reflect the level of the difference between $\mathcal{G}_0$ and $\mathcal{G}_1$. Specifically, the larger μ is, the more different $\mathcal{G}_0$ and $\mathcal{G}_1$ are, and thus it will be easier to test H_1 from H_0.

In fact, in the work of (Neykov et al., 2016), they propose another quantity to characterize the difference between $\mathcal{G}_0$ and $\mathcal{G}_1$, which is defined as

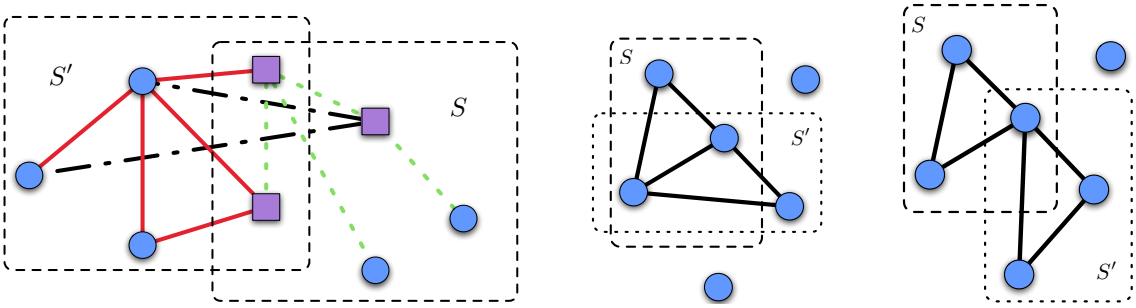
$$\mu' = \max_{S, S' \in \mathcal{E}} \frac{|S \cap S'|}{|V(S \cap S')|}.$$

In order to differentiate it from the weak edge density μ in Definition 5, we call μ' the strong edge density since we always have $\mu \leq \mu'$. In (Neykov et al., 2016), they show that μ' is an important quantity to characterize the information-theoretic lower bound of testing H_0 versus H_1 without any computational constraint. An interesting discovery in our paper is that when a polynomial query constraint is imposed as (7), instead of the strong edge density, the weak edge density μ takes the role to characterize the minimax rate. When $\mu \ll \mu'$, there exists a gap between statistical optimal rate and the rate applying computationally efficient algorithm. When $\mu \asymp \mu'$, there will be no such gap. For example, when $\mathcal{E}$ contains all possible s-chains, both μ and μ' are $O(1)$. When $\mathcal{E}$ contains all possible s-cliques, $\mu = O(1)$ but $\mu' \asymp s$. We will elaborate how μ and μ' may determine the computational-statistical tradeoffs in §3.2.

Next we introduce another concept called constrained vertex cut number, which is related to the vertex cut. In graph theory, a vertex set $\tilde{V} \subseteq V$ is called a vertex cut for nonadjacent sets $V_1, V_2 \subseteq V$ if the removal of $\tilde{V}$ from the graph separates V_1 and V_2 into distinct connected components. For edge sets S and S', we define the constrained vertex cut number of S and S' as

$$\gamma(S, S') = \min \big\{ |\tilde{V}| : \tilde{V} \subseteq V(S \cup S'), \\ \tilde{V} \text{ is a vertex cut for } V(S) \backslash \tilde{V} \text{ and } V(S') \backslash \tilde{V} \big\}. \tag{11}$$

The difference between the constrained vertex cut and ordinary vertex cut is that, we require $\tilde{V} \subseteq V(S \cup S')$, and we specify that the remaining vertices in $V(S)$ and $V(S')$ must



(a) Vertex cut (b) $\gamma(S, S') = 2$ (c) $\gamma(S, S') = 1$

Figure 3. (a) An example of the constrained vertex cut number. In this example, the solid lines, dash lines and dash-dot lines represent edges in S', S and E_0 respectively. $\tilde{V}$ consisting of the square vertices gives the constrained vertex cut number $\gamma(S, S') = 3$. (b, c) Two examples of the vertex cut ratio calculation. (b) shows an example of S' such that $\gamma(S, S') = 2$, (c) shows an example of S' such that $\gamma(S, S') = 1$. If we let $\mathcal{E}$ be all cliques with size 3, then the vertex cut ratio $\zeta = 1/9$.

be separated. Our definition is also related to the concept of vertex buffer proposed by (Neykov et al., 2016). Figure 3 (a) shows an example of the constrained vertex cut number.

We can interpret the constrained vertex cut number as the "correlation" between S and S', since it measures the number of the vertex cut of two edge sets. Using the constrained vertex cut, we can define the vertex cut ratio as follows.

Definition 6 (Vertex Cut Ratio)**.** *Give a null-alternative separator $\mathcal{E}$, let $k = \max_{S,S' \in \mathcal{E}} \gamma(S, S')$. We define the vertex cut ratio as*

$$\zeta = \inf_{0 \le j \le k-1} \frac{|\{S' \in \mathcal{E} : \max_{S \in \mathcal{E}} \gamma(S, S') = j\}|}{|\{S' \in \mathcal{E} : \max_{S \in \mathcal{E}} \gamma(S, S') = j+1\}|}. \tag{12}$$

In (12), $|\{S' \in \mathcal{E} : \max_{S \in \mathcal{E}} \gamma(S, S') = j\}|$ is the cardinality of edge sets whose maximal constrained vertex cut number is j. Since $|\mathcal{E}| = \sum_{j=0}^{k} |\{S' \in \mathcal{E} : \max_{S \in \mathcal{E}} \gamma(S, S') = j\}|$, the vertex cut ratio measures how uniform the edge sets are correlated to each other. If ζ is large, this implies many of edge sets in $\mathcal{E}$ have relative small correlation. Therefore, more queries are needed to differentiate various cases, which makes the testing harder. Figures 3(b) and (c) visualize two examples of calculating the constrained vertex cut number and the vertex cut ratio.

After introducing the weak edge density μ and the vertex cut ratio ζ, we now present our main result on the lower bound for graph property test.

Theorem 7. *Suppose that we have a null-alternative separator $\mathcal{E}$ with the null base G_0. Under the oracle computational model defined in §2.2, if we require the number of queries $T \le d^\eta$ for some constant $\eta > 0$, when $\liminf_{d \to \infty} \zeta > 1$ and the signal strength*

$$\theta \le \frac{\kappa \log \zeta}{\log d + \log \zeta} \sqrt{\frac{1}{\mu n}} \wedge \frac{\sqrt{\mu}}{8s} \wedge \frac{1}{2k\sqrt{\mu}}, \tag{13}$$

where $k = \max_{S,S' \in \mathcal{E}} \gamma(S, S')$ and κ is some sufficiently small positive constant only depends on s and η, then for any algorithm $\mathscr{A} = M(\mathcal{Q}, T, q_{\text{init}}, \delta)$, there exists an oracle r such that $\liminf_{n \to \infty} R_n(\mathcal{C}_0, \mathcal{C}_1, \mathscr{A}, r) = 1$.

Proof. Recall the inequality (10). The main idea of this proof is quantifying the number of $\Theta_{\mathcal{S}}, \mathcal{S} \in \mathcal{E}$, which can be distinguished from Θ_0 by a query q, denoted as n_q. Then for any algorithm the computational complexity T cannot be less than $|\mathcal{E}| / \sup_{q \in \mathcal{Q}} n_q$. See Appendix §B for a detailed proof. □

In practice, the first term in (13) is the leading term determining the optimal signal strength. The later two terms in (13) essentially impose two scaling conditions: $\theta = O(s^{-1})$ and $\theta = O(k^{-1})$. The signal strength rate $\theta = \log \zeta / [(\log d + \log \zeta)\sqrt{\mu n}]$ shows that if ζ is large and μ is small, then the necessary signal strength is large. This matches our heuristic discussion on the meaning of μ and ζ above.

The hypercube problem cannot be covered by Theorem 7. In fact, the hypercube problem is special since we can characterize its lower bound with sublinear computational complexity constraint.

Theorem 8. *When testing empty graph against lattice graphs with s vertices and fixed lattice dimension D, given the query complexity $T = o(d/s)$, there exists a constant κ which only depends on T such that if $\theta \le \kappa / (\sqrt{sn})$, then for any algorithm $\mathscr{A} = M(\mathcal{Q}, T, q_{\text{init}}, \delta)$, there exists an oracle r such that $\liminf_{n \to \infty} R_n(\mathcal{C}_0, \mathcal{C}_1, \mathscr{A}, r) = 1$.*

Proof. See Appendix §A.4 for a detailed proof. □

Since we allow $s = d^\alpha$ for $\alpha \in (0, 1)$, Theorem 8 gives a tradeoff between the sublinear computational complex-

Table 1. The information-theoretic lower bound θ_{Info} and the computationally efficient lower bound θ_{Comp} for s-clique, s-nearest neighbor graph, 3-clique versus s-clique and perfectly matched block, where for the first three cases, we require $s = O(d^\alpha)$ for some constant $\alpha \in (0, 1/2)$. Here all the rates in the table are up to some constants.

	s-Clique	s-Nearest Neighbor	3-Clique vs s-Clique	Perfectly Matched Block
μ	$(s-1)/(2s)$	1/4	$(s^2-s-6)/(2s^2)$	$(d^{1/2}-1)/(2d^{1/2})$
μ'	$(s-1)/2$	$s/4$	$(s^2-s-6)/(2s)$	$(d^{1/2}-1)/2$
ζ	$(d+1-2s)/s^2$	$(d+1-2s)/s^2$	$(d+1-2s)/s^2$	$(d-2d^{1/2}+1)/d$
θ_{Info}	$\sqrt{\log(d/s^2)/(ns)}$	$\sqrt{\log(d/s^2)/(ns)}$	$\sqrt{\log(d/s^2)/(ns)}$	$\sqrt{1/(nd^{1/2})}$
θ_{Comp}	$\sqrt{1/n}$	$\sqrt{1/n}$	$\sqrt{1/n}$	$1/(\sqrt{n}\log d)$

ity and the signal strength. When the signal strength θ is $O(1/\sqrt{d^\alpha n})$, all tractable algorithms must have $T = O(d/s) = O(d^{1-\alpha})$.

3.2. Computational-Statistical Tradeoffs

In this subsection, we give concrete formulations of the optimal signal strength in (13) for four examples: s-clique, s-nearest neighbor graph, small clique versus large clique and perfectly matched block. In specific, we assume that the size of small clique s' is a constant. So we give a concrete example that $s' = 3$ for small clique versus large clique example. Next we will show the computational-statistical tradeoffs for these four cases and how the weak edge density μ and the strong edge density μ' determine the tradeoffs.

When there is no computational constraint, it is shown in Theorem 4.2 of (Neykov et al., 2016) that under some scaling conditions, if $\theta = O(1/\sqrt{\mu' n})$, for clique, nearest neighbor, 3-clique versus s-clique and perfectly matched block detection problems we have $\liminf_{n\to\infty} R(\mathcal{C}_0, \mathcal{C}_1) = 1$, i.e., any test is asymptotically powerless. When there is polynomial query complexity constraint, Theorem 7 states the minimal signal strength in (13). In most cases, the rate in (13) is typically $O(1/\sqrt{\mu n})$. Our result can be summarized as follows: for some sufficiently small constants κ_1 and κ_2,

- **Information-Theoretic Bound:** $\theta \le \kappa_1/\sqrt{\mu n}$, where $\mu = \max_{S,S'\in\mathcal{E}} \frac{|S\cap S'|}{|V(S\cap S')|^2}$;
- **Computationally-Efficient Bound:** $\theta \le \kappa_2/\sqrt{\mu' n}$, where $\mu' = \max_{S,S'\in\mathcal{E}} \frac{|S\cap S'|}{|V(S\cap S')|}$.

Therefore, there will be a gap between information-theoretic lower bound and computationally efficient lower bound if $\mu \ll \mu'$. We can find the concrete optimal rates for four cases in Table 1. From Table 1, we can see that for all four examples, since $\mu/\mu' = O(1/s)$, there are all statistical-computational tradeoffs. We also remark that for perfectly matched block, we have $\zeta < 1$ which violates the condition in Theorem 7. We will generalize Theorem 7 to cover such case in Theorem A.5 in the appendix.

3.3. Upper Bounds

In this section, we construct upper bounds to match the lower bounds in §3.1 under the oracle computational model. We propose two testings methods: entrywise test and local summation test which match the computationally efficient lower bound and information-theoretic lower bound respectively.

The Entrywise Test: We consider a sequence of queries and a test ψ as following,

$$q_{jk}(X) = [(X_j + X_k)^2 - X_j^2 - X_k^2]/2, \text{ and}$$
$$\psi = \mathbb{1}\Big[\sup_{j\neq k} z_{q_{jk}} \le A_1\Big], \tag{14}$$

where $j \neq k \in [d]$ and $z_{q_{jk}}$ is the random realization of an oracle of query function $q_{jk}(\cdot)$. Here A_1 is the reject level to be specified for different structures. See Theorem 9 for details. By Definition 3, the computational complexity of the entrywise test ψ in (14) is $T = O(d^2)$. This algorithm is to calculate all the off-diagonal entries in the covariance matrix. The idea of the entrywise test is to find the strongest signal among $\{q_{jk}\}_{j\neq k}$, each of which quantifying the entrywise difference between null and alternative.

The entrywise test, as indicated by its name, only takes supreme over entrywise comparison. If each entry only has small signal, the entrywise test may fail to reject. Then we need the following local summation test.

The Local Summation Test: We consider a sequence of queries and a test ψ as following,

$$q_{\mathcal{S}}(X) = \Big(\sum_{j\in\mathcal{S}} X_j\Big)^2/s, \text{ and}$$
$$\psi = \mathbb{1}\Big[\sup_{|\mathcal{S}|=s} z_{q_{\mathcal{S}}} \le A_2\Big], \tag{15}$$

where $\mathcal{S} \subseteq [d]$, $|\mathcal{S}| = s$ and $z_{q_\mathcal{S}}$ is the random realization of an oracle of query function $q_\mathcal{S}(\cdot)$. Here A_2 is also the reject level to be specified in Theorem 9. We can check that the computational complexity $T = \binom{d}{s}$, which is no longer polynomial to d. The local summation test accumulates the signals in all s-by-s submatrices of covariance matrix. Compared with the entrywise test, this test amplifies the signal strength and thus can detect weaker signals but the price to pay is more computational cost. See Figure 3.3 for the visualization.

We have the following theorem showing two tests above match the lower bounds with or without computational complexity constraints up to some logarithmic factors.

Theorem 9. *For the empty graph versus s-clique problem defined in §2, we consider the following conditions:*

(i) For $\theta > \kappa\sqrt{\log d/n}$, we consider the entrywise test in (14) with $A_1 = 1 - \frac{\theta(1+\theta-\theta s)}{2(1-\theta)(\theta s-\theta+1)}$.

(ii) For $\theta > \kappa\sqrt{\log d/(ns)}$, we consider the local summation test in (15) with $A_2 = 1 - \frac{\theta(s-1)}{2(\theta s-\theta+1)}$. Here κ is a sufficiently large constant.

Under either (i) or (ii) above, we have

$$\liminf_{n\to\infty}\left[\sup_{\Theta\in\mathcal{C}_0}\mathbb{P}_\Theta(\psi=1)+\sup_{\Theta\in\mathcal{C}_1}\mathbb{P}_\Theta(\psi=0)\right]=0.$$

For the empty graph versus s-nearest neighbor graph problem defined in §2, we have same result if we choose $A_1 = 1 - \frac{\theta(s-1)}{2s}$ and $A_2 = 1 - \frac{\theta s}{4}$. See §B.2.1 in the appendix for details.

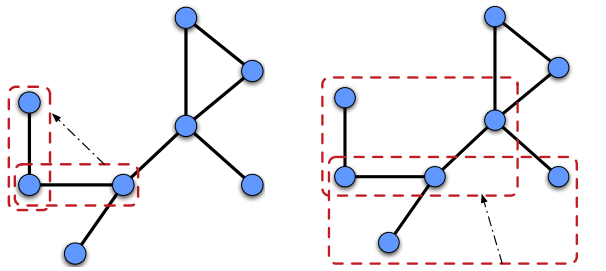
For the 3-clique versus s-clique problem defined in §2, we have the same result if we choose same A_1 and A_2 as in Theorem 9. But we need to consider a new entrywise test with the test ψ as following form,

$$\psi = \mathbb{1}\left[\inf_{j_1\neq k_1, j_2\neq k_2, j_3\neq k_3, j_4\neq k_4}\max_{1\le i\le 4} z_{q_{j_i k_i}} \le -\frac{\theta(1+\theta-\theta s)}{2(1-\theta)(\theta s-\theta+1)}\right].$$

See §B.8 in the appendix for details.

For the empty graph versus perfectly matched block problem defined in §2, notice that a perfectly matched block is also a clique, the testing algorithms for clique detection with the same values of A_1 and A_2 with $s = d^{1/2}$ works for perfectly matched block problem as in Theorem 9.

For the hypercube problem, let $\mathcal{K}$ be the class of edge sets for all disjoint s-lattice graphs. We consider a sequence of



(a) The entrywise test (b) The local summation test

Figure 4. (a) The entrywise test. This test examines the signal strength between every pair of vertices. (b) The local summation test. This test examines the sum of signal strength among all s vertices, where $s = 4$.

queries and a test as following,

$$q_S(X) = \frac{1}{|S|}\sum_{(i,j)\in S} X_i X_j, \text{ and}$$

$$\psi = \mathbb{1}\left[\sup_{S\in\mathcal{K}} z_{q_S(X)} \le -\frac{\theta}{4}\right],$$

where $S \in \mathcal{K}$. The computational complexity is $T = d/s$ here. If $\theta \ge \kappa\sqrt{\log(d/s)/(ns)}$ with some large enough constant κ, we have

$$\liminf_{n\to\infty}\left[\sup_{\Theta\in\mathcal{C}_0}\mathbb{P}_\Theta(\psi=1)+\sup_{\Theta\in\mathcal{C}_1}\mathbb{P}_\Theta(\psi=0)\right]=0.$$

We refer Theorem A.8 in the appendix for details.

These upper bound tests give us another insight of why the edge densities play an important role in computational-statistical tradeoffs. We see that actually the weak edge density μ and the strong edge density μ' characterize the global and local density of signals in the null-alternative separator respectively. If $\mu \asymp \mu'$, then the signals are "sparsely" distributed. We can find a polynomial-time algorithm, like the entrywise test, to match the information-theoretic lower bound. However, if $\mu \ll \mu'$, then the signals are locally concentrated. Only exponential-time algorithms, like the local-summation test, can amplify the signal strength and match the information-theoretic lower bound.

4. Conclusion

In this paper, we study the computational-statistical tradeoffs in some common combinatorial inference problems. Based on the oracle computational model, we build the computational lower bounds and provide the corresponding upper bounds. Theoretically, we prove that these tradeoffs are determined by two intrinsic quantities of the graph, the weak edge density and the strong edge density. In the future, we will study more computational-statistical tradeoffs problems.

References

Addario-Berry, L., Broutin, N., Devroye, L., and Lugosi, G. On combinatorial testing problems. *The Annals of Statistics*, 38(5):3063–3092, 2010.

Arias-Castro, E. and Verzelen, N. Community detection in dense random networks. *The Annals of Statistics*, 42(3): 940–969, 06 2014. doi: 10.1214/14-AOS1208.

Arias-Castro, E., Candès, E. J., Helgason, H., and Zeitouni, O. Searching for a trail of evidence in a maze. *The Annals of Statistics*, 36(4):1726–1757, 08 2008. doi: 10.1214/07-AOS526.

Arias-Castro, E., Bubeck, S., and Lugosi, G. Detection of correlations. *The Annals of Statistics*, 40(1):412–435, 2012.

Arias-Castro, E., Bubeck, S., and Lugosi, G. Detecting positive correlations in a multivariate sample. *Bernoulli*, 21(1):209–241, 2015a.

Arias-Castro, E., Bubeck, S., Lugosi, G., and Verzelen, N. Detecting markov random fields hidden in white noise. *arXiv preprint arXiv:1504.06984*, 2015b.

Balakrishnan, S., Wainwright, M. J., Yu, B., et al. Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1): 77–120, 2017.

Berthet, Q. and Rigollet, P. Computational lower bounds for sparse PCA. In *Conference on Learning Theory*, pp. 380–404, 2013a.

Berthet, Q. and Rigollet, P. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780–1815, 2013b.

Bilmes, J. A. and Bartels, C. Graphical model architectures for speech recognition. *IEEE signal processing magazine*, 22(5):89–100, 2005.

Bresler, G., Gamarnik, D., and Shah, D. Structure learning of antiferromagnetic ising models. In *Advances in Neural Information Processing Systems*, pp. 2852–2860, 2014.

Cai, T. T., Liang, T., and Rakhlin, A. Computational and statistical boundaries for submatrix localization in a large noisy matrix. *arXiv preprint arXiv:1502.01988*, 2015.

Castro, R., Lugosi, G., and Savalle, P.-A. Detection of correlations with adaptive sensing. *Information Theory, IEEE Transactions on*, 60(12):7913–7927, 2014.

Chen, S. X., Zhang, L.-X., and Zhong, P.-S. Tests for high-dimensional covariance matrices. *Journal of the American Statistical Association*, 2012.

Chen, Y. Incoherence-optimal matrix completion. *IEEE Transactions on Information Theory*, 61(5):2909–2923, May 2015.

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

Feldman, V., Grigorescu, E., Reyzin, L., Vempala, S., and Xiao, Y. Statistical algorithms and a lower bound for detecting planted cliques. In *ACM Symposium on Theory of Computing*, pp. 655–664, 2013.

Feldman, V., Guzman, C., and Vempala, S. Statistical query algorithms for stochastic convex optimization. *arXiv preprint arXiv:1512.09170*, 2015a.

Feldman, V., Perkins, W., and Vempala, S. On the complexity of random satisfiability problems with planted solutions. In *ACM Symposium on Theory of Computing*, pp. 77–86, 2015b.

Feldman, V., Guzmán, C., and Vempala, S. Statistical query algorithms for mean vector estimation and stochastic convex optimization. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1265–1277. Society for Industrial and Applied Mathematics, 2017.

Friedman, N. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805, 2004.

Hajek, B., Wu, Y., and Xu, J. Computational lower bounds for community detection on random graphs. *arXiv preprint arXiv:1406.6625*, 2014.

Kannan, R. and Vempala, S. Chi-squared amplification: Identifying hidden hubs. *arXiv preprint arXiv:1608.03643*, 2016.

Kearns, M. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM*, 45(6):983–1006, 1998.

Krauthgamer, R., Nadler, B., and Vilenchik, D. Do semidefinite relaxations solve sparse PCA up to the information limit? *The Annals of Statistics*, 43(3):1300–1322, 06 2015. doi: 10.1214/15-AOS1310.

Ma, Z. and Wu, Y. Computational barriers in minimax submatrix detection. *The Annals of Statistics*, 43(3):1089–1116, 06 2014.

Murphy, K. P., Torralba, A., and Freeman, W. T. Using the forest to see the trees: A graphical model relating features, objects, and scenes. In *Advances in neural information processing systems*, pp. 1499–1506, 2004.

Neykov, M., Lu, J., and Liu, H. Combinatorial inference for graphical models. *arXiv preprint arXiv:1608.03045*, 2016.

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

Wang, T., Berthet, Q., and Samworth, R. J. Statistical and computational trade-offs in estimation of sparse principal components. *arXiv preprint arXiv:1408.5369*, 2014.

Wang, Z., Gu, Q., and Liu, H. Sharp computational-statistical phase transitions via oracle computational model. *arXiv preprint arXiv:1512.08861*, 2015.

Welling, M., Rosen-Zvi, M., and Hinton, G. E. Exponential family harmoniums with an application to information retrieval. In *Advances in neural information processing systems*, pp. 1481–1488, 2005.

Yuan, X.-T. and Zhang, T. Truncated power method for sparse eigenvalue problems. *Journal of Machine Learning Research*, 14(1):899–925, April 2013.

Zhang, Y., Wainwright, M. J., and Jordan, M. I. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conference on Learning Theory*, pp. 323–340, 2014.