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

Hao Lu [1] **Yuan Cao** [1] **Zhuoran Yang** [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 and nearest neighbor graph 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).

Our 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 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 computational tractability and statistical optimality.

[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: Hao Lu <haolu@princeton.edu>.

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

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; Yi et al., 2016; Kannan & Vempala, 2016; Fan et al., 2018). In comparison, we focus on the combinatorial inference problem in undirected graphical models, for which most existing work studies the information-theoretic limits and the computational-statistical tradeoffs remain much less well understood. See, e.g., (Addario-Berry et al., 2010; Arias-Castro et al., 2012; 2015a;b; Cai et al., 2015) and the references therein. Furthremore, from an aspect not studied in previous work, we illustrate how the achievability of the information-theoretic lower bounds for tractable algorithms are governed by the underlying combinatorial structures, especially the weak edge density and the constrained vertex cut number, which are formally defined in §3.

To characterize the computational complexity, in this paper, we adopt the oracle computational model proposed by (Kearns, 1998), which is further generalized by (Feldman et al., 2013; 2015a;b; Wang et al., 2015; Fan et al., 2018). This model is able to capture the computational aspect of a broad range of statistical algorithms, including stochastic convex optimization methods, 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 by $\mathrm{A}_{j\cdot}$ and $\mathrm{A}_{\cdot j}$ the j^{th} row and column of A, respectively. 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 the combinatorial inference problems 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 the lower bounds, as well as their applications to concrete examples. Furthermore, we establish and discuss the intrinsic link between structural properties and computational tractability.

2. Background

We first introduce the combinatorial inference problem on the Gaussian graphical model. Then we introduce the framework of oracle computational model, which is used to formally study the performance of statistical algorithms under computational budgets.

2.1. Combinatorial Inference Problems

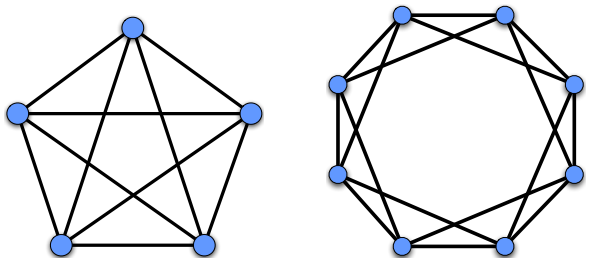
The Gaussian graphical model assumes that a d-dimensional random vector $X = (X_1, \ldots, X_d)^\top \in \mathbb{R}^d$ follows a multivariate normal distribution $N_d(0, \Theta^{-1})$. Here $\Theta = (\theta_{jk})_{j,k\in[d]}$ is known as the precision matrix, which encodes the pairwise conditional independence among $X_1, \ldots, X_d$. More specifically, for any $j, k \in [d]$, $j \neq k$, X_j and X_k are independent conditioning on all the remaining variables if and only if $\theta_{jk} = 0$. In addition, we consider an 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. By definition, to infer the conditional dependency of X, it suffices to learn the structural properties of $G(\Theta)$. In the sequel, we denote $G(\Theta)$ by G when no ambiguity arises.

Given n independent realizations $x_1, \ldots, x_n$ of X, the goal of combinatorial inference is to test whether the underlying graph G possesses 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 are interested in the hypothesis testing problem

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

Here the graphs in $\mathcal{G}_1$ share the combinatorial structure of interest. To illustrate the notion of combinatorial inference, we consider the following three concrete instances.

Clique Detection. Let $\mathcal{G}_0 = \{(V, \emptyset)\}$ and $\mathcal{G}_1 = \{G_J : |J| = s \text{ and } J \subseteq V\}$. In each G_J, vertices i and j are



(a) A clique of size 5 (b) A nearest neighbor graph

Figure 1. We plot a clique of size 5 and a nearest neighbor graph with $s = 8$ in (a) and (b), respectively.

connected if and only if $i, j \in J$. In other words, J is a clique of G_J. For this problem, our goal is to test if the graph contains a clique of size s. For illustration, we plot a clique of size 5 in Figure 1-(a).

Test of 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 node subset I of size s, the corresponding nearest neighbor graph G_I is defined as follows. We 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_I. In the sequel, we call such a graph as s-nearest neighbor graph. We plot a nearest neighbor graph with $s = 8$ in Figure 1-(b), where each node is connected to 4 of its neighbors.

Small Clique vs Large Clique. It is also interesting to distinguish graphs with small cliques from those with large cliques. In this case, we 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, where s' is a constant number and $s > s'$ may grow with d. Without loss of generality, we focus on the case where $s' = 3$ hereafter.

It is worth mentioning that the null hypothesis in previous work is always a simple hypothesis, i.e., $\mathcal{G}_0$ contains only one element. Whereas we propose a more general framework where both the null and alternative hypotheses are allowed to be composite, as in the last example given above.

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

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

were $\mathcal{C}_0$ and $\mathcal{C}_1$ are the sets of precision matrices corresponding to $G(\Theta)$ in $\mathcal{G}_0$ and $\mathcal{G}_1$, respectively. In addition, we constrain $\mathcal{C}_0$ and $\mathcal{C}_1$ into a subspace $\mathcal{M}$ defined as

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

where θ is the signal strength. We define the minimax testing risk $R_n(\mathcal{C}_0, \mathcal{C}_1)$ as

$$R_n(\mathcal{C}_0, \mathcal{C}_1) \\ = \inf_{\psi} \Big[\sup_{\Theta\in\mathcal{C}_0} \mathbb{P}_\Theta(\psi = 1) + \sup_{\Theta\in\mathcal{C}_1} \mathbb{P}_\Theta(\psi = 0)\Big], \tag{3}$$

where the infimum is taken over all test functions ψ based on $\{x_i\}_{i\in[n]}$. Here we reject the null hypothesis H_0 if ψ takes value one, and we accept if ψ takes value zero. Besides, in (3) we use $\mathbb{P}_\Theta$ to represent the distribution $N_d(0, \Theta^{-1})$. Note that any test is asymptotically powerless if

$$\liminf_{n\to\infty} R_n(\mathcal{C}_0, \mathcal{C}_1) = 1. \tag{4}$$

In this case, the problem defined in (2) is unsolvable.

2.2. Oracle Computational Model

When solving statistical problems, any algorithm can be seen as a sequence of interactions with data. Intuitively, the computational complexity of an algorithm can be qualitatively measured by the number of interactions with the data. The framework of oracle computational model (Kearns, 1998) captures such a fact by assuming that any algorithm for a statistical problem interacts with an oracle r. More specifically, in each round, the algorithm sends a query $q \in \mathcal{Q}$ to r and obtain a random response $Z_q \in \mathbb{R}$, where $\mathcal{Q}$, named the query space, is the set of all possible queries. Upon receiving a realization of Z_q, the algorithm determine its next query based on all the past responses. We formally define algorithms under the oracle computational model as follows.

Definition 1. *Under an oracle computational model M, an algorithm $\mathscr{A}$ defined as a tuple $M(\mathcal{Q}, T, q_{\text{init}}, \{\delta_t\}_{t\in[T]})$, where $\mathcal{Q}$ is the query space, T is the maximum number of rounds the algorithm is allowed to query the oracle, $q_{init} \in \mathcal{Q}$ is the initial query, and $\delta_t : (\mathcal{Q}\times\mathbb{R})^{\otimes t} \to \mathcal{Q}\cup\{\texttt{HALT}\}$ is the transition function after the t^{th} round. Here δ_t decides the $(t+1)^{\text{th}}$ query based on all the previous t queries and their returns. In addition, if δ_t returns* `HALT`*, then the algorithm terminates.*

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 (Feldman et al., 2013; 2015a;b) for more discussions. Under such a computational model, the number

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

of rounds that the algorithm interacts with the oracle r serves as the the computational complexity, which is rigorously defined as follows.

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 define $|\mathcal{Q}_{\mathscr{A}}|$ as the computational complexity of $\mathscr{A}$.*

In this work, we base upon this framework to study the computational hardness in combinatorial inference. In this case, the query is a function of $X \in \mathbb{R}^d$, the random vector of interest. Throughout this paper, we consider queries that are almost surely bounded, i.e.,

$$\mathcal{Q} = \{q : \mathbb{R}^d \to \mathbb{R}, q(X) \in [-B, B] \text{ almost surely}\} \quad (5)$$

for some $B > 0$. By Definitions 1 and 2, the computational complexity of an algorithm $\mathscr{A} = M(\mathcal{Q}, T, q_{\text{init}}, \delta)$ is upper bounded by T. We say $\mathscr{A}$ is an polynomial-time algorithm if it can be written as $\mathscr{A} = M(\mathcal{Q}, T, q_{\text{init}}, \delta)$ with $T = d^\eta$ for some constant η, i.e., the number of queries made by $\mathscr{A}$ is bounded by a polynomial of d.

Furthermore, the performance of statistical methods rely on the error of estimating population quantities based on n samples. We capture this fact by defining a statistical query oracle as follows, which, for any query $q \in \mathcal{Q}$, returns a noisy estimate of $\mathbb{E}[q(X)]$.

Definition 3 (Statistical Query Oracle)**.** *Let $\mathcal{Q}_{\mathscr{A}}$ be the set of queries that an algorithm makes. A statistical query oracle r interacts with a query $q \in \mathcal{Q}_{\mathscr{A}}$ at each round, and returns an output $Z_q \in \mathbb{R}$. Moreover, for a fixed $\xi \in (0, 1)$, we assume that $\{Z_q\}_{q \in \mathcal{Q}_{\mathscr{A}}}$ satisfy*

$$\mathbb{P}\bigg(\bigcap_{q \in \mathcal{Q}_{\mathscr{A}}} \Big\{ |Z_q - \mathbb{E}[q(X)]| \le \tau_q \Big\} \bigg) \ge 1 - 2\xi, \quad (6)$$

where τ_q is defined as

$$\tau_q = \max\Big\{ \big[\eta(\mathcal{Q}_{\mathscr{A}}) + \log(1/\xi)\big] \cdot B/n, \sqrt{2\big[\eta(\mathcal{Q}_{\mathscr{A}}) + \log(1/\xi)\big] \cdot \{B^2 - \mathbb{E}^2[q(X)]\}/n} \Big\}. \quad (7)$$

Here we define $\eta(\mathcal{Q}_{\mathscr{A}}) = \log(|\mathcal{Q}_{\mathscr{A}}|)$ when $\mathcal{Q}_{\mathscr{A}}$ is countable, and let $\eta(\mathcal{Q}_{\mathscr{A}})$ be other capacity measures such as the Vapnik-Chervonenkis dimension and metric entropy when $\mathcal{Q}_{\mathscr{A}}$ is uncountable.

The intuition of Definition 3 is that, (6) and (7) characterizes the concentration effect of the empirical measure. Specifically, if we define Z_q^* as the average of $\{q(x_i)\}_{i=1}^n$, where $\{x_i\}_{i=1}^n$ are n i.i.d. realizations of X, Bernstein's inequality yields that

$$\mathbb{P}\big\{\big|Z_q^* - \mathbb{E}\big[q(X)\big]\big| > t\big\} \le 2\exp\bigg\{\frac{-n \cdot t^2}{2 \cdot \mathrm{Var}\big[q(X)\big] + 2B/3 \cdot t}\bigg\}. \quad (8)$$

Moreover, since $q(X)$ is bounded in $[-B, B]$, we have

$$\mathrm{Var}\big[q(X)\big] = \mathbb{E}\big[q^2(X)\big] - \mathbb{E}^2\big[q(X)\big] \le M^2 - \mathbb{E}^2\big[q(X)\big].$$

Thus, in (7) we replace the unknown $\mathrm{Var}[q(X)]$ by its upper bound. Moreover, we obtain uniform concentration over $\mathcal{Q}_{\mathscr{A}}$ by bounding the suprema of empirical processes based on (8), which implies that

$$\mathbb{P}\Big(\sup_{q \in \mathcal{Q}_{\mathscr{A}}} \big\{\big|Z_q^* - \mathbb{E}\big[q(X)\big]\big| \le c\tau_q\big\}\Big) \ge 1 - 2\xi$$

for an absolute constant c, where $\eta(\mathcal{Q}_{\mathscr{A}})$ in τ_q measures the capacity of $\mathcal{Q}_{\mathscr{A}}$ in the logarithmic scale. Therefore, the oracle that returns Z_q^* for any $q \in \mathcal{Q}$ satisfies Definition (3), which implies that the deviation behavior of Z_q under the statistical query oracle is achievable. Furthermore, in most algorithmic analysis of statistical problems, like principal component analysis (Yuan & Zhang, 2013) and latent variable model estimation (Balakrishnan et al., 2017), the deviation behavior in (7) is optimal in terms of order.

Compared with the minimax risk defined in (3), we define a new minimax risk of testing $\mathcal{C}_0$ against $\mathcal{C}_1$ under the oracle computational model with a statistical query oracle r as

$$R_n(\mathcal{C}_0, \mathcal{C}_1, \mathscr{A}, r) = \inf_{\psi \in \mathcal{H}(\mathscr{A}, r)} \Big[\sup_{\Theta \in \mathcal{C}_0} \mathbb{P}_\Theta(\psi = 1) + \sup_{\Theta \in \mathcal{C}_1} \mathbb{P}_\Theta(\psi = 0)\Big]. \quad (9)$$

Here n is the sample size and $\mathcal{H}(\mathscr{A}, r)$ is the space of all possible tests based on an algorithm $\mathscr{A}$ and the oracle r, as specified in Definitions 1 and 3, respectively. Following the same idea as in (4), if there exists an oracle r such that

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

then any hypothesis test computed by an algorithm based on at most T queries 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.

3.1. A General Computational Lower Bound

Before introducing the main theorem, we first define the notion of null-alternative separator, which is originally proposed in (Neykov et al., 2016) in order to study information-theoretical limits of combinatorial inference.

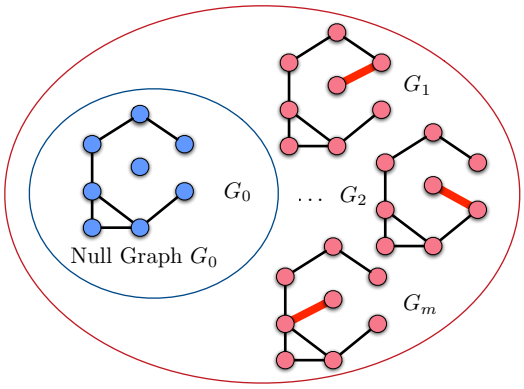


Figure 2. An example of null-alternative separator. Let $\mathcal{G}_0$ and $\mathcal{G}_1$ be the sets of disconnected and connected graphs respectively. For the null base G_0 shown in the figure, all the red thick edges form a null-alternative separator.

Definition 4 (Null-Alternative Separator)**.** *Let $G_0 = (V, E_0) \in \mathcal{G}_0$ be some graph under the null hypothesis. 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 $(V, E_0 \cup S) \in \mathcal{G}_1$.*

In Figure 2 we illustrate an example of the null-alternative separator, whose intuition can be understood as follows. Given $G_0 = (V, E_0) \in \mathcal{G}_0$ and a null-alternative separator $\mathcal{E}$ with null base G_0, we consider precision matrices

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

where A_0 and A_S are the adjacency matrices of G_0 and $G_S = (V, S)$, respectively. By the definition of $\mathcal{E}$, we have $\Theta_0 \in \mathcal{C}_0$ and $\{\Theta_S\}_{S\in\mathcal{E}} \subseteq \mathcal{C}_1$. Thus, the null-alternative separator contains critical edge sets which may change a graph G_0 from the the null to the alternative.

In addition, for the minimax risk defined in (9), we have

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

In combinatorial inference, since $\mathcal{C}_0$ and $\mathcal{C}_1$ have some symmetric structures, we expect that $R_n(\mathcal{C}_0, \mathcal{C}_1, \mathscr{A}, r)$ and $R_n(\{\Theta_0\}, \{\Theta_S\}_{S\in\mathcal{E}}, \mathscr{A}, r)$ to have the same order. As a result, the restricted parameter spaces $\{\Theta_0\}$ and $\{\Theta_S\}_{S\in\mathcal{E}}$ capture the fundamental hardness of the testing problem in (2), and thus the computational lower bound only depends on the structure of $\mathcal{E}$.

Based on the null-alternative separator, now we are ready to introduce two pivotal notions that determine the computational lower bound, namely the weak edge density and the vertex cut ratio.

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

$$\mu = \max_{S,S'\in\mathcal{E}} \frac{|S \cap S'|}{|V(S \cap S')|^2}. \quad (11)$$

Here we follow the convention that $0/0 = 0$.

Intuitively, the weak edge density of $\mathcal{E}$ measures the concentration of the critical edges that can change graphs from $\mathcal{G}_0$ to $\mathcal{G}_1$. Specifically, since a clique with n vertices has $n(n-1)/2$ edges, μ essentially compares the subgraph with edges $S \cap S'$ against the clique with vertex set $V(S \cap S')$. Thus, when μ is large, it would be easier to find an element in $\mathcal{E}$, which distinguishes $\mathcal{G}_1$ from $\mathcal{G}_0$. Hence, μ captures the level of the difference between $\mathcal{G}_0$ and $\mathcal{G}_1$ and is critical to our computational lower bound.

Similar to μ in Definition 5, (Neykov et al., 2016) propose a similar quantity to characterize the difference between $\mathcal{G}_0$ and $\mathcal{G}_1$, which is defined as

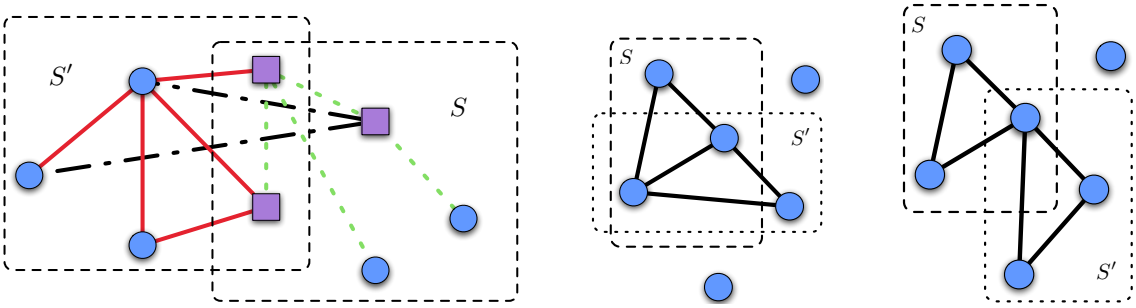
$$\mu' = \max_{S,S'\in\mathcal{E}} \frac{|S \cap S'|}{|V(S \cap S')|}. \quad (12)$$

This notion plays a significant role in the information-theoretic lower bound for combinatorial inference. In order to differentiate it from μ in (11), we call μ' the strong edge density since we always have $\mu \leq \mu'$ by definition. 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. As a result, μ and μ' together determine the computational-statistical tradeoffs. Specifically, when $\mu \ll \mu'$, there exists a gap between statistical optimal rate and the rate applying computationally efficient algorithm. Whereas 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 $\mu' \asymp s$.

Furthermore, we define the vertex cut ratio as follows, which builds upon a restricted version of the vertex cut. Recall that, in graph theory, a vertex set $\tilde{V} \subseteq V$ is called a vertex cut of 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 two edge sets S and S', we define the constrained vertex cut number of S and S' as

$$\gamma(S, S') = \min\big\{|\tilde{V}| : \tilde{V} \subseteq V(S \cup S'), \\ \tilde{V} \text{ is a vertex cut for } V(S)\backslash\tilde{V} \text{ and } V(S')\backslash\tilde{V}\big\}. \quad (13)$$

That is, $\gamma(S, S')$ is the size of the minimum vertex cut of $V(S)$ and $V(S')$, when restricting the whole graph to $V(S \cup S')$. We remark that this definition is also related to the concept of vertex buffer proposed in (Neykov et al., 2016). Figure 3-(a) shows an example of the constrained vertex cut number.



(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 $\zeta = 1/9$.

By definition, $\gamma(S, S')$ reflects the connectivity of the graph restricted to $V(S \cup S')$. When S and S' are more overlapping, we expected that the restricted graph is more connected. Thus, we may interpret the constrained vertex cut number as the "correlation" between S and S', which is larger when S and S' are more similar. Based on the constrained vertex cut number, we define the vertex cut ratio as follows.

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

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

Here, $\max_{S \in \mathcal{E}} \gamma(S, S')$ is the maximal value of the constrained vertex cut number of S' and the edge set in $\mathcal{E}$. Notice that we could partition $\mathcal{E}$ according to the the value of the maximal constrained vertex cut number. That is, we have $|\mathcal{E}| = \sum_{j=0}^{k} |\{S' \in \mathcal{E} : \max_{S \in \mathcal{E}} \gamma(S, S') = j\}|$. Thus, the vertex cut ratio characterizes the growth of sequence

$$\left\{ |\{S' \in \mathcal{E} : \max_{S \in \mathcal{E}} \gamma(S, S') = j\}| \right\}_{0 \le j \le k}$$

by comparing it with a geometric sequence. Since the constrained vertex cut number can be viewed as the correlation of edge sets, when ζ is large, many of the edge sets in $\mathcal{E}$ have relative small correlation. Under this scenario, more queries are needed to differentiate various cases, which makes the testing harder. In Figures 3-(b) and (c) we 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, now we are ready to present the our general result on the computational lower bound for combinatorial inference.

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 θ satisfies*

$$\theta \le \frac{\kappa \log \zeta}{\log d + \log \zeta} \sqrt{\frac{\log(1/\xi)}{\mu n}} \wedge \frac{\sqrt{\mu}}{8s} \wedge \frac{1}{2k\sqrt{\mu}}, \qquad (14)$$

where $k = \max_{S,S' \in \mathcal{E}} \gamma(S, S')$ and κ is some sufficiently small positive constant only depends on 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). We consider the restricted parameter spaces $\{\Theta_0\}$ and $\{\Theta_S\}_{S \in \mathcal{E}}$. The main idea of this proof is quantifying the number of $\Theta_S, \mathcal{S} \in \mathcal{E}$, 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 (14) is the leading term determining the optimal signal strength. The later two terms in (14) essentially impose two scaling conditions: $\theta = O(s^{-1})$ and $\theta = O(k^{-1})$. When the tail probability ξ is a small constant, efficient algorithms for combinatorial inference requires a signal strength rate $\theta \asymp \log \zeta / [(\log d + \log \zeta)\sqrt{\mu n}]$, which shows that if ζ is large and μ is small, the necessary signal strength is large. This matches our heuristic discussion on the meaning of μ and ζ above.

In the following, we apply the above theorems to the three concrete examples given in §2. We state below a corollary for clique detection; the results for the other two instances are stated in §A.

Corollary 8 (Computational Lower Bound for Clique Detection)**.** *For testing if the graph has a s-clique, we define*

$\mathcal{C}_0 = \{I_d\}$ *and*

$$\mathcal{C}_1 = \{\Theta : \exists \mathcal{S} \subset [d] \text{ and } |\mathcal{S}| = s, \\ \theta_{j,k} = \mathbb{1}(j = k) + \theta \cdot \mathbb{1}(j, k \in \mathcal{S}, j \neq k)\}.$$

Let η be a positive constant. If $s = O(d^\alpha)$ for some $\alpha \in (0, 1/2)$, and $\theta \leq \min\{\kappa/\sqrt{n},\ 1/(8s)\}$, then for any algorithm that queries at most d^η rounds, there exists a statistical oracle r such that

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

Here κ is a small enough constant that only depends on α and η,

Proof. Let $G_0 = (V, \emptyset)$. Clearly, $\mathcal{C} = \{E[G(\Theta)] : \Theta \in \mathcal{C}_1\}$ is a null-alternative separator with base G_0. To apply Theorem 7, we need to compute the weak edge density μ and the vertex cut ratio ζ. By setting $S' = S$ in Definition 5, we have

$$\mu \geq \frac{|S|}{|V(S)|^2} = \frac{|V(S)| \cdot (|V(S)| - 1)}{2|V(S)|^2} \\ = \frac{1}{2} - \frac{1}{2|V(S)|} \geq \frac{1}{4},$$

where we use $|V(S)| \geq 2$ in the last inequality. Here the first equality follows from the fact that the subgraph with vertex set $V(S)$ and edge set S is a clique of size s. Moreover, since the the weak edge density of any graph is bounded by $1/2$, we conclude that μ is of constant order.

Furthermore, note that $\max_{S,S' \in \mathcal{E}} \gamma(S, S') = s$. For $j = 0, \ldots, s$, we define

$$m_j = |\{S' \in \mathcal{E} : \max_{S \in \mathcal{E}} \gamma(S, S') = k - j\}|.$$

Then by definition, we have $m_j = \binom{s}{s-j}\binom{d-s}{j}$. By direct computation, for all $j \in \{0, \ldots, s-1\}$, it can be shown that $m_{j+1}/m_j \geq d/s^2$, which implies that

$$\zeta \geq d/s^2 \geq C d^{1-2\alpha}$$

for some absolute constant C. Where the last inequality follow from $s = O(d^\alpha)$. Finally, we conclude the proof by applying Theorem 7. □

This corollary shows that $\theta = O(1/\sqrt{n})$ is the critical threshold for the existence of an asymptotically powerful hypothesis test that runs in polynomial time. Moreover, (Neykov et al., 2016) show that the information-theoretical bound $\theta = O(1/\sqrt{ns})$ is tight. Thus, we observe a computational-statistical tradeoff for the detection of cliques. That is, a statistical price of $\sqrt{1/s}$ is paid in order to achieve computational efficiency.

Furthermore, the computational-statistical tradeoffs appear widely in combinatorial inference problems, which are determined by the weak and strong edge densities together. Specifically, when there is no computational constraint, it is shown in Theorem 4.2 of (Neykov et al., 2016) that under some scaling conditions, if $\theta = O(1/\sqrt{\mu' n})$, we have $\liminf_{n\to\infty} R(\mathcal{C}_0, \mathcal{C}_1) = 1$ for all these three problems, where μ' is the strong edge density defined in (12). That is, $1/\sqrt{\mu' n}$ is the critical threshold for the existence of an asymptotically powerful test. In addition, when there is a polynomial query complexity constraint, Theorem 7 states the minimal signal strength in (14), which is typically $O(1/\sqrt{\mu n})$. Our result can be summarized as follows. For some sufficiently small constants κ_1 and κ_2, we have:

- **Information-Theoretic Bound:** If $\theta \leq \kappa_1/\sqrt{\mu' n}$, any hypothesis test is asymptotically powerless;
- **Computationally-Efficient Bound:** $\theta \leq \kappa_2/\sqrt{\mu n}$, any hypothesis test computed by a polynormial-time algorithm is asymptotically powerless.

Therefore, there will be a gap between information-theoretic lower bound and computationally efficient lower bound if $\mu \ll \mu'$. For clique detection, we have $\mu/\mu' = O(1/s)$. As shown in §A, this is also the case for detecting s-nearest neighbor graphs and 3-cliques against s-cliques. As a result, statistical-computational tradeoffs appear in all the three examples given in §2. To our best knowledge, these interesting tradeoffs are first established in the literature.

3.2. 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 and information-theoretic lower bounds respectively.

The Entrywise Test. For any $j, k \in [d]$ with $j \neq k$, we define $q^*_{jk} \colon \mathbb{R}^d \to \mathbb{R}$ by $q^*_{jk}(x) = [(x_j + x_k)^2 - x_j^2 - x_k^2]/2$. We consider a sequence of queries and a test ψ as following,

$$q_{jk}(X) = q^*_{jk}(X) \cdot \inf\{|q^*_{jk}(X)| \leq R \cdot \log n\} \text{ and} \\ \psi = \mathbb{1}\left[\sup_{j \neq k} z_{q_{jk}} \leq A_1\right], \tag{15}$$

where R is an absolute constant, A_1 is the reject level to be specified for different combinatorial structures, and $z_{q_{jk}}$ is the response returned by a statistical query for query $q_{jk}(X)$. Here we apply truncation to ensure boundedness of the queries, as required by the statistical query oracle. We note that such a truncation is unnecessary when using real data, and we set R sufficiently large such that $q^*_{jk}(X)$ and $q_{jk}(X)$ is close in expectation. By Definition 3, the

computational complexity of the entrywise test ψ in (15) 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. For any $\mathcal{S} \subseteq [d]$ with $|\mathcal{S}| = s$, we define $q^*_{\mathcal{S}}(x) = (\sum_{j\in\mathcal{S}} x_j)^2/s$ We consider a sequence of queries and a test ψ as following,

$$q_{\mathcal{S}}(X) = q^*_{\mathcal{S}}(X) \cdot \mathbb{1}\{|q^*_{\mathcal{S}}(X)| \leq R \cdot \log n\}, \text{ and}$$
$$\psi = \mathbb{1}\bigg[\sup_{|\mathcal{S}|=s} z_{q_{\mathcal{S}}} \leq A_2\bigg], \tag{16}$$

where $z_{q_{\mathcal{S}}}$ is the response of an oracle for query $q_{\mathcal{S}}(X)$, A_2 is the reject level, and R is an absolute constant to ensure boundedness. In this the computational complexity is $T = \binom{d}{s}$, which superpolynomial in d. Intuitively, 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 3.2 for a visualization.

The following theorem establish the performances of the hypothesis tests defined above, which shows that our computational lower bounds are tight 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 $\theta > \kappa \log n \cdot \sqrt{\log(d/\xi)/n}$, we consider the entrywise test in (15) with $A_1 = 1 - \frac{\theta(1+\theta-\theta\cdot s)}{4(1-\theta)(\theta\cdot s-\theta+1)}$.

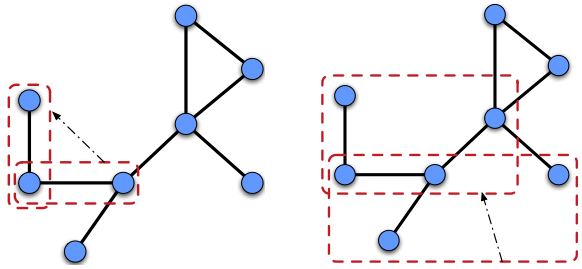
(ii) For $\theta > \kappa \log n \cdot \sqrt{\log(d/\xi)/(ns)}$, we consider the local summation test in (16) with $A_2 = 1 - \frac{\theta(s-1)}{4(\theta s-\theta+1)}$.

Here κ is a sufficiently large constant. Under either (i) or (ii) above, we have

$$\sup_{\Theta\in\mathcal{C}_0} \mathbb{P}_\Theta(\psi = 1) + \sup_{\Theta\in\mathcal{C}_1} \mathbb{P}_\Theta(\psi = 0) \leq 2\xi.$$

Since $\xi > 0$ in Definition 3 is arbitrary, the above theorem implies that both the hypothesis tests in (15) and (16) are asymptotically powerful. Moreover, when setting $\xi = 1/d$, we conclude that the signal strengths in conditions (i) and (ii) match the corresponding lower bounds up to some logarithmic terms.

Moreover, similar upper bounds can be established for the other two instances. Specifically, for the empty graph versus



(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 $s = 4$.

s-nearest neighbor graph problem defined in §2, we have same result if we choose $A_1 = 1 - \frac{\theta(s-1)}{4s}$ and $A_2 = 1 - \frac{\theta s}{8}$. In addition, for the problem of testing 3-clique versus s-clique defined in §2, we have the same result if we choose same A_1 and A_2 as in Theorem 9. However, in this case, we need to consider a new entrywise test, whose test function is defined as

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

See §B.3 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 on the Gaussian graphical model. Based on the oracle computational model, we build the computational lower bounds and provide matching upper bounds. Interestingly, our results characterize the statistical price paid to achieve computational efficiency, which is shown to be determined by two intrinsic quantities of the graph, namely, the weak and strong edge densities.

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.

Fan, J., Liu, H., Wang, Z., and Yang, Z. Curse of heterogeneity: Computational barriers in sparse mixture models and phase retrieval. *Manuscipt*, 2018.

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.

Yi, X., Wang, Z., Yang, Z., Caramanis, C., and Liu, H. More supervision, less computation: statistical-computational tradeoffs in weakly supervised learning. In *Advances in Neural Information Processing Systems*, pp. 4482–4490, 2016.

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.