Proceedings of Machine Learning Research vol 65:1–24, 2017

The Hidden Hubs Problem

Ravi Kannan KANNAN@MICROSOFT.COM
Microsoft Research India.

Santosh Vempala * VEMPALA@GATECH.EDU
School of Computer Science, Georgia Tech

Abstract

We introduce the following *hidden hubs* model $H(n, k, \sigma_0, \sigma_1)$: the input is an $n \times n$ random matrix A with a subset S of k special rows (hubs); entries in rows outside S are generated from the Gaussian distribution $p_0 = N(0, \sigma_0^2)$, while for each row in S, an unknown subset of k of its entries are generated from $p_1 = N(0, \sigma_1^2)$, $\sigma_1 > \sigma_0$, and the rest of the entries from p_0. The special rows with higher variance entries can be viewed as hidden higher-degree hubs. The problem we address is to identify the hubs efficiently. The planted Gaussian Submatrix Model is the special case where the higher variance entries must all lie in a $k \times k$ submatrix. If $k \geq c\sqrt{n \ln n}$, just the row sums are sufficient to find S in the general model. For the Gaussian submatrix problem (and the related planted clique problem), this can be improved by a $\sqrt{\ln n}$ factor to $k \geq c\sqrt{n}$ by spectral or combinatorial methods.

We give a polynomial-time algorithm to identify all the hidden hubs with high probability for $k \geq n^{0.5-\delta}$ for some $\delta > 0$, when $\sigma_1^2 > 2\sigma_0^2$. The algorithm extends to the setting where planted entries might have different variances, each at least σ_1^2. We also show a nearly matching lower bound: for $\sigma_1^2 \leq 2\sigma_0^2$, there is no polynomial-time Statistical Query algorithm for distinguishing between a matrix whose entries are all from $N(0, \sigma_0^2)$ and a matrix with $k = n^{0.5-\delta}$ hidden hubs for any $\delta > 0$. The lower bound as well as the algorithm are related to whether the chi-squared distance of the two distributions diverges. At the critical value $\sigma_1^2 = 2\sigma_0^2$, we show that the hidden hubs problem can be solved for $k \geq c\sqrt{n}(\ln n)^{1/4}$, improving on the naive row sum-based method.
Keywords: Planted Clique, Hidden Gaussian, Spectral Methods, Concentration, Statistical Queries.

1. Introduction

Identifying hidden structure in random graphs and matrices is a fundamental topic in unsupervised machine learning, with many application areas and deep connections to probability, information theory, linear algebra, statistical physics and other disciplines. A prototypical example is finding a large hidden clique in a random graph. A well-known extension to real-valued entries is the Gaussian hidden submatrix: each entry is drawn from $N(0, \sigma_0^2)$, except for entries from a $k \times k$ submatrix, which are drawn from $N(\mu, \sigma_1^2)$. The study of these problems has lead to interesting algorithms and analysis techniques.

Model. In this paper, we consider a more general model of hidden structure: the presence of a small number of hidden *hubs*. These hubs might represent more influential or atypical nodes of a network. Recovering such nodes is of interest in many areas (information networks, protein interaction networks, cortical networks etc.). In this model, as before, the entries of the matrix are

* Supported in part by NSF Award CCF-1563838

drawn from $N(0, \sigma_0^2)$ except for special entries that all lie in k rows, with k entries from each of these k rows. This is a substantial generalization of the above hidden submatrix problems, as the only structure is the existence of k higher "degree" rows (hubs) rather than a large submatrix. (We also consider unequal variances for the special entries and varying numbers of them for each hub.)

More precisely, we are given an $N \times n$ random matrix A with independent entries. There is some unknown subset S of special rows, with $|S| = s$. Each row in S has k special entries, each picked according to

$$p_1(x) \sim N(0, \sigma_1^2),$$

whereas, all the other $Nn - k|S|$ entries are distributed according to

$$p_0 \sim N(0, \sigma_0^2).$$

The task is to find S, given, $s = |S|$, $k, n, \sigma_0^2, \sigma_1^2$. One may also think of S rows as picking n i.i.d. samples from a mixture

$$\frac{k}{n} p_1(x) + \left(1 - \frac{k}{n}\right) p_0(x),$$

whereas, the non-S rows are picking i.i.d. samples from $p_0(x)$. This makes it clear that we cannot assume that the planted entries in the S rows are all in the same columns (while the methods in this paper can handle both variants, we focus on the version with a given number of atypical entries in each row).

If $\sigma_0^2 = \sigma_1^2$, obviously, we cannot find S. If

$$\sigma_1^2 > \sigma_0^2(1 + c),$$

for a positive constant c (independent of n, k), then it is easy to see that $k = \Omega\left(\sqrt{n \ln n}\right)$ suffices to have a polynomial time algorithm to find S: Set $B_{ij} = A_{ij}^2 - 1$. Let $\sum_j B_{ij} = \rho_i$. It is not difficult to show that if $k \geq c\sqrt{n \ln n}$, then, whp,

$$\text{Min}_{i:\text{ hub}} \quad \rho_i > 2\text{Max}_{i:\text{ non-hub}} \quad \rho_i.$$

The above algorithm is analogous to the "degree" algorithm for hidden (Gaussian) clique — take the k vertices with the highest degrees — and works with high probability for $k \geq c\sqrt{n \ln n}$. The remaining literature on upper bounds removes the $\sqrt{\ln n}$ factor, by using either a spectral approach (Alon et al., 1998) or a combinatorial approach (iteratively remove the minimum degree vertex, (Feige and Ron, 2010)). These algorithms, both spectral and combinatorial, rely on the special entries being in a $k \times k$ submatrix. This leads to our first question:

Q. Are there efficient algorithms for finding hidden hubs for $k = o(\sqrt{n \ln n})$?

1.1. Related work

Algorithms for the special cases of planted clique and hidden Gaussian submatrix are based on spectral or combinatorial methods. Information-theoretically, even a planting of size $O(\log n)$ can be found in time $n^{O(\log n)}$ by enumerating subsets of size $O(\log n)$. This raises the question of the threshold for efficient algorithms. Since the planted part has different mean or variance, it is natural to try to detect the planting using either the sums of the rows (degrees in the case of graphs)

or the spectrum of the matrix. The degree method needs $k = \sqrt{n \log n}$ for the planted clique problem or the hidden Gaussian problem with different variances. For the hidden Gaussian problem with perturbed mean, Ma and Wu (2015) give a characterization of the computational difficulty (assuming hardness of planted clique); a hidden submatrix of any size k can be detected if the mean is shifted by a sufficiently large function that grows with k as roughly $\sqrt{\log(n/k)}$.

These approaches can detect planted cliques with sufficiently large mean separation or for $k = \Omega(\sqrt{n})$ (Boppana, 1987; Kucera, 1995; Alon et al., 1998; Feige and Ron, 2010; Dekel et al., 2011; Bhaskara et al., 2010; Montanari et al., 2015; Deshpande and Montanari, 2015a). Roughly speaking, the relatively few entries of the planted part must be large enough to dominate the variance of the many entries of the rest of the matrix. Moreover, for any $\delta > 0$, finding a planted clique of size smaller than $n^{0.5-\delta}$ planted in $G_{n,\frac{1}{2}}$ is impossible by statistical algorithms (Feldman et al., 2013a) or by using convex programming hierarchies (Barak et al., 2016).

The algorithm of Bhaskara et al. (2010) for detecting dense subgraphs can be used together with thresholding to detect a hidden Gaussian clique (of different variance) of size $k = n^{0.5-\delta}$. The resulting running time grows roughly as $n^{O(1/(\epsilon-2\delta))}$ for $\sigma_1^2 = 2(1+\epsilon)\sigma_0^2$, and ϵ must be $\Omega(1)$ to be polynomial-time.

We note that all these improvements below $k = \Omega(\sqrt{n \log n})$ rely on the special entries lying in a submatrix. The use of spectral methods or finding high density regions crucially depends on this structure.

In other related work, a precise threshold for a rank-one perturbation to a random matrix to be noticeable was given by Féral and Péché (2007) and applied in a lower bound by Montanari et al. (2015) on using the spectrum to detect a planting. Tensor optimization (or higher moment optimization) rather than eigen/singular vectors can find smaller cliques (Frieze and Kannan, 2008; Brubaker and Vempala, 2009), but the technique has not yielded a polynomial-time algorithm to date. A different approach to planted clique and planted Gaussian submatrix problems is to use convex programming relaxations, which also seem unable to go below $\sqrt{n}$. Many recent papers demonstrate the limitations of these approaches (Feige and Krauthgamer, 2000; Feldman et al., 2013a; Meka et al., 2015; Hopkins et al., 2016; Barak et al., 2016; Feldman et al., 2017) (see also Jerrum (1992)).

1.2. Our results

Our main results can be summarized as follows. (For this statement, assume ε, δ are positive constants. In detailed statements later in the paper, they are allowed to depend on n.)

Theorem 1 *For the hidden hubs model with k hubs:*

1. *For $\sigma_1^2 = 2(1+\varepsilon)\sigma_0^2$, there is an efficient algorithm for $k \geq n^{0.5-\delta}$ for some $\delta > 0$, depending only on ε.*

2. *For $\sigma_1^2 \in [c\sigma_0^2, 2\sigma_0^2]$, any constant $c > 0$, no polynomial Statistical Query algorithm can detect hidden hubs for $k = n^{0.5-\delta}$, for any $\delta > 0$.*

3. *At the critical value $\sigma_1^2 = 2\sigma_0^2$, with $N = n$, $k \geq \sqrt{n}\,(\ln n)^{1/4}$ suffices.*

As far as we know, these are the first improvements for the hidden hubs problem that can recover the hubs below the threshold given by the simple row-sum (degree-based) algorithm. Previous

algorithms that go below the degree threshold need the special (higher variance) entries to span a $k \times k$ submatrix and do not extend to the hidden hubs model. At the critical threshold of $\sigma_1^2 = 2\sigma_0^2$, the size of the clique that can be recovered with our methods jumps from below $\sqrt{n}$ to $\sqrt{n}(\log n)^{1/4}$; the latter is still an improvement over the simple degree-based method. Further improvement at the threshold, even to $\sqrt{n}$, is an open problem.

Our algorithm also gives improvements for the special case of identifying hidden Gaussian cliques. For that problem, the closest upper bound in the literature is the algorithm of Bhaskara et al. (2010) for detecting dense subgraphs, which can be used to give a polynomial-time algorithm for $\epsilon = \Omega(1)$, with exponent growing with the inverse of ϵ. Their running time of the algorithm is $n^O(1/\epsilon)$ for $k = n^{0.5-\epsilon}$ (it is not stated explicitly, but is an algorithm that follows from their ideas). In contrast, our simple algorithms run in time linear in the number of entries of the matrix for $\epsilon = \Omega(1/\log n)$. All this is for the special case of the hidden Gaussian problem, where all atypical entries lie in a $k \times k$ submatrix.

Our upper bound can be extended further, to the model where each planted entry could have its own distribution $p_{ij} \sim N(0, \sigma_{ij}^2)$ with bounded σ_{ij}^2. There is a set of rows S that are hubs, with $|S| = k$. For each $i \in S$, now we assume there is some subset T_i of higher variance entries. The $|T_i|$ are not given and need not be equal. We assume that the special entries satisfy:

$$\sigma_{ij}^2 \geq \sigma_1^2, \text{ where, } \sigma_1^2 = 2(1+\varepsilon)\sigma_0^2, \varepsilon > 0.$$

Theorem 2 *Let $\tau_i = \sum_{j \in T_i} n^{-\sigma_0^2/\sigma_{ij}^2}$. Suppose, for all $i \in S$,*

$$\tau_i \geq \frac{1}{\sqrt{\varepsilon}} c(\ln N)(\ln n)^{0.5},$$

then there is a randomized algorithm to identify all of S with high probability.

As a corollary, we get that if $|T_i| = k$ for all $i \in S$, all special entries satisfy $\sigma_{ij}^2 = \sigma_1^2$, and

$$k = n^{.5-\delta}, \text{ with } \varepsilon \geq \frac{2\delta}{1-2\delta} + \frac{\ln\ln N}{\ln n} + \frac{\ln\ln n}{2\ln n},$$

then we can identify all of S.

We also have a result for values of $\varepsilon \in \Omega(1/\ln n)$, see Theorem (9).

Techniques. Our algorithm is based on a new technique to amplify the higher variance entries, which we illustrate next. Let

$$p_0(x) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{x^2}{2\sigma_0^2}\right) \qquad p_1(x) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{x^2}{2\sigma_1^2}\right)$$

be the two probability densities. The central (intuitive) idea behind our algorithm is to construct another matrix $\hat{A}$ of "likelihood ratios", defined as

$$\hat{A}_{ij} = \frac{p_1(A_{ij})}{p_0(A_{ij})} - 1.$$

Such a transformation was also described in the context of the planted clique problem Deshpande and Montanari (2015b) (although it does not give an improvement for that problem). At a high level,

one computes the row sums of $\hat{A}$ and shows that the row sums of the k rows of the planted part are all higher than all the row sums of the non-planted part. First, note that

$$\mathsf{E}_{p_0}(\hat{A}_{ij}) = \int p_1 - \int p_0 = 0 \; ; \; \mathrm{Var}_{p_0}(\hat{A}_{ij}) = \int \left(\frac{p_1}{p_0} - 1\right)^2 p_0 = \int \frac{p_1^2}{p_0} - 1 = \chi^2(p_1 \| p_0),$$

the χ-squared distance between the two distributions p_0, p_1. Also,

$$\mathsf{E}_{p_1}\left(\frac{p_1}{p_0} - 1\right) = \chi^2(p_1 \| p_0).$$

Intuitively, since the expected sum of row i, for any $i \notin S$ is 0, we expect success if the expected row sum in each row of S is greater than the standard deviation of the row sum in any row not in S times a log factor, namely, if

$$\sqrt{\chi^2(p_1 \| p_0)} \geq \tilde{\Omega}(\frac{\sqrt{n}}{k}) = \tilde{\Omega}(n^{\delta}). \tag{1}$$

$$\text{Now,} \quad \chi^2(p_1 \| p_0) = \int \frac{p_1^2}{p_0} - 1 = \frac{c\sigma_0}{\sigma_1^2} \int \exp\left(x^2 \left(\frac{1}{2\sigma_0^2} - \frac{1}{\sigma_1^2}\right)\right).$$

So, if $\sigma_1^2 \geq 2\sigma_0^2$, then, clearly, $\chi^2(p_1 \| p_0)$ is infinite and so intuitively, (1) can be made to hold. This is not a proof. Indeed substantial technical work is needed to make this succeed. The starting point of that is to truncate entries, so the integrals are finite. We also have to compute higher moments to ensure enough concentration to translate these intuitive statements into rigorous ones.

On the other hand, if $\sigma_1^2 < 2\sigma_0^2$, then $\chi^2(p_1 \| p_0)$ is finite and indeed bounded by a constant independent of $k, \sqrt{n}$. So (1) does not hold. This shows that this line of approach will not yield an algorithm. Our lower bounds show that there is no polynomial time Statistical Query algorithm at all when $\sigma_1^2 \in (0, 2\sigma_0^2]$.

The algorithms are based on the following transformation to the input matrix: truncate each entry of the matrix, i.e., set the ij'th entry to $\min\{M, A_{ij}\}$, then apply $\frac{p_1(\cdot)}{p_0(\cdot)}$ to it; then take row sums. The analysis needs nonstandard a concentration inequality via a careful estimation of higher moments; standard concentration inequalities like the Höffding inequality are not sufficient to deal with the fact that the absolute bound on p_1/p_0 is too large. More sophisticated standard inequalities such as extensions of Bernstein's inequality to higher moments also appear to be inadequate.

Our algorithms also apply directly to the following *distributional version* of the hidden hubs problem with essentially the same separation guarantees. A hidden hubs distribution is a distribution over vectors $x \in \mathbb{R}^n$ defined by a subset $S \subset [n]$ and parameters μ, σ_1, σ_0 as follows: $x_i \sim N(0, \sigma_0^2)$ for $i \notin S$, and for $i \in S$,

$$x_i \sim \begin{cases} N(\mu, \sigma_1^2) & \text{with probability } \frac{k}{n} \\ N(0, \sigma_0^2) & \text{with probability } 1 - \frac{k}{n}. \end{cases}$$

The problem is to identify S.

For almost all known distributional problems[1], the best-known algorithms are *statistical* or can be made statistical, i.e., they only need to compute expectations of functions on random samples

1. The only known exception where a nonstatistical algorithm solves a distributional problem efficiently is learning parities with no noise using Gaussian elimination.

rather than requiring direct access to the samples. This characterization of algorithms, introduced by Kearns (1993, 1998), has been insightful in part because it is possible to prove lower bounds on the complexity of statistical query algorithms. For example, Feldman et al. (2013a) have shown that the bipartite planted clique problem cannot be solved efficiently by such algorithms when the clique size is $k \leq n^{0.5-\delta}$ for any $\delta > 0$. A statistical query algorithm can query the input distribution via a statistical oracle. Three natural oracles are STAT, VSTAT and 1-STAT. Roughly speaking, STAT(τ) returns the expectation of any bounded function on a random sample to within additive tolerance τ; VSTAT(t) returns the expectation of a $0/1$-valued function to within error no more than the standard deviation of t random samples; and 1-STAT returns the value of a $0/1$ function on a random sample.

For the hidden hubs problem, our algorithmic results show that one can go below $\sqrt{n}$ hubs (size of clique for the special case of hidden Gaussian clique). Under the conditions of the algorithmic bounds, for $\sigma_1^2 \geq 2(1+\epsilon)\sigma_0^2$, there is a $\delta > 0$ s.t., a planting can be detected using a single statistical query whose tolerance is at most the standard deviation of the average of $O(n/k)$ independent samples. We complement the algorithmic results with a lower bound on the separation between parameters that is *necessary* for statistical query algorithms to be efficient (Theorem 18). Our application of statistical query lower bounds to problems over continuous distributions might be of independent interest. Our matching upper and lower bounds can be viewed in terms of a single function, namely the χ-squared divergence of the planted Gaussian and the base Gaussian.

The model and results raise several interesting open questions, including: (1) Can the upper bounds be extended to more general distributions on the entries, assuming independent entries? (2) Does the χ-squared divergence condition suffice for general distributions? (3) Can we recover $k = O(\sqrt{n})$ hidden hubs when $\sigma_1^2 = 2\sigma_0^2$? (our current upper bound is $k = \sqrt{n}(\ln n)^{1/4}$ and our lower bounds do not apply above $\sqrt{n}$) (4) Are there reductions between planted clique problems with $1/-1$ entries and the hidden hubs problem addressed here?

Summary of algorithms. Our basic algorithm for all cases is the same:

Define an M (which is $\sigma_0\sqrt{\ln n}(1+o(1))$.) The exact value of M differs from case to case. Define matrix B by $B_{ij} = \exp(\gamma \text{Min}(x^2, M^2))$, where, γ is always $= \frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2}$. Then, we prove that with high probability, the maximum $|S|$ row sums of B occur precisely in the S rows. However, the bounds are delicate and so we present the proofs in each case separately.

2. At the threshold: $\sigma_1^2 = 2\sigma_0^2$

In this section, we assume

$$\sigma_1^2 = 2\sigma_0^2 \quad \text{and } N = n.$$

$$\frac{p_1}{p_0} = ce^{\gamma x^2},$$

where, $\gamma > 0$ is given by:

$$\gamma = \frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2} = \frac{1}{4\sigma_0^2}. \tag{2}$$

Define L, M by:

$$L = \sqrt{2\left(\ln n - \ln\ln n\right)} \quad ; \quad M = L\sigma_0. \tag{3}$$

$$B_{ij} = \exp\left(\gamma \text{Min}(M^2, A_{ij}^2)\right). \tag{4}$$

Theorem 3 *If*

$$k \geq c\sqrt{n}(\ln n)^{1/4},$$

then with probability $1 - o(1)$*, the top* s *row sums of the matrix* B *occur precisely in the* S *rows.*

Proposition 4 *Suppose* X *is a non-negative real-valued random variable and* l *is a positive integer.*

$$\mathsf{E}\left(|X - E(X)|^l\right) \leq 2\mathsf{E}(X^l).$$

Proof

$$\begin{aligned}\mathsf{E}\left(|X - \mathsf{E}(X)|^l\right) &\leq \int_{x=0}^{\mathsf{E}(X)} (\mathsf{E}X)^l \mathsf{Pr}(X = x)\, dx + \int_{x=\mathsf{E}(X)}^{\infty} x^l \mathsf{Pr}(X = x)\, dx \\ &\leq (\mathsf{E}X)^l + \mathsf{E}(X^l) \leq 2\mathsf{E}(X^l),\end{aligned}$$

the last, since, $\mathsf{E}(X) \leq (\mathsf{E}(X^l))^{1/l}$. ■

2.1. Non-planted entries are small

Let

$$\mu_0 = \mathsf{E}_{p_0}(B_{ij}) = \frac{1}{\sqrt{2\pi}\sigma_0} \int_{-\infty}^{\infty} \exp\left(\gamma \mathsf{Min}(M^2, x^2)\right) \exp(-x^2/2\sigma_0^2). \tag{5}$$

$$\mu_0 \leq \frac{1}{\sqrt{2\pi}\sigma_0} \int_{-\infty}^{\infty} \exp(\gamma x^2) p_0(x)\, dx = \frac{1}{\sqrt{2\pi}\sigma_0} \int_{-\infty}^{\infty} \exp(-x^2/2\sigma_1^2) = \sqrt{2}. \tag{6}$$

.

$$\begin{aligned}&\mathsf{E}_{p_0}((B_{ij} - \mu_0)^2) \\ &\leq \mathsf{E}_{p_0}(B_{ij}^2) \\ &\leq \frac{2}{\sqrt{2\pi}\sigma_0} \int_0^M \exp(2\gamma x^2) \exp(-x^2/2\sigma_0^2) + \frac{2\exp(2\gamma M^2)}{\sqrt{2\pi}\sigma_0} \int_M^{\infty} \frac{x}{M} \exp(-x^2/2\sigma_0^2) dx \\ &\leq \frac{2}{\sigma_0} \int_0^M dx \ + \ \frac{2\sigma_0}{M} \exp\left(M^2\left(2\gamma - \frac{1}{2\sigma_0^2}\right)\right) \leq \quad cL.\end{aligned} \tag{7}$$

For $l \geq 4$, even, we have $\gamma l - (1/2\sigma_0^2) > 0$ and using Proposition (4), we get

$$\begin{aligned}&\mathsf{E}_{p_0}((B_{ij} - \mu_0)^l) \\ &\leq 2\mathsf{E}_{p_0}(B_{ij}^l) \\ &\leq \frac{4}{\sqrt{2\pi}\sigma_0} \int_0^M \exp(\gamma l x^2) \exp(-x^2/2\sigma_0^2) + \frac{4\exp(\gamma l M^2)}{\sqrt{2\pi}\sigma_0} \int_M^{\infty} \frac{x}{M} \exp(-x^2/2\sigma_0^2) dx \\ &\leq \frac{2}{\sigma_0} \int_0^M \exp\left(Mx\left(\gamma l - \frac{1}{2\sigma_0^2}\right)\right) dx \ + \ \frac{2\sigma_0}{M} \exp\left(M^2\left(\gamma l - \frac{1}{2\sigma_0^2}\right)\right) \\ &\leq \frac{c}{L} \exp\left(\frac{L^2(l-2)}{4}\right),\end{aligned} \tag{8}$$

We will use a concentration result from (Kannan (2009), Theorem 1) which specialized to our case states

Theorem 5 *If $X_1, X_2, \ldots, X_n$ are i.i.d. mean 0 random variables, for any even positive integer m, we have*

$$\mathsf{E}\left(\left(\sum_{j=1}^{n} X_j\right)^m\right) \leq (cm)^m \left[\sum_{l=1}^{m/2} \frac{1}{l^2}\left(\frac{n\mathsf{E}(X_1^{2l})}{m}\right)^{1/l}\right]^{m/2}.$$

With $X_j = B_{ij} - \mu_0$, in Theorem (5), we plug in the bounds of (7) and (8) to get:

Lemma 6

$$\forall m \text{ even}, m \leq c \ln n, \qquad \mathsf{E}_{p_0}\left(\sum_{j=1}^{n} (B_{ij} - \mu_0)\right)^m \leq (cmnL)^{m/2}$$

Proof For all even m,

$$\mathsf{E}_{p_0}\left(\sum_{j=1}^{n} (B_{ij} - \mu_0)\right)^m \leq (cm)^m \left[\frac{nL}{m} + \exp(L^2/2)\sum_{l=2}^{m/2} \frac{1}{l^2}\left(\frac{n}{mL}\exp(-L^2/2)\right)^{1/l}\right]^{m/2}.$$

Now, it is easy to check that

$$\frac{cnL}{m} \geq \exp(L^2/2)\left(n\exp(-L^2/2)/(mL)\right)^{1/l} \forall l \geq 2.$$

Hence the Lemma folows, noting that $\sum_l (1/l^2) \leq c$. ■

Lemma 7 *Let*

$$t = c\sqrt{n}\,(\ln n)^{3/4}.$$

for c a suitable constant. For $i \notin S$,

$$\mathsf{Pr}\left(\left|\sum_{j=1}^{n}(B_{ij} - \mu_0)\right| \geq t\right) \leq \frac{1}{n^2}.$$

Thus, we have

$$\mathsf{Pr}\left(\exists i \notin S : \left|\sum_{j=1}^{n}(B_{ij} - \mu_0)\right| \geq t\right) \leq \frac{1}{n}.$$

Proof We use Markov's inequality on the random variable $\left|\sum_{j=1}^{n}(B_{ij} - \mu_0)\right|^m$ and Lemma (6) with m set to $4\ln n$ to get

$$\mathsf{Pr}\left(\left|\sum_{j=1}^{n}(B_{ij} - \mu_0)\right| \geq t\right) \leq e^{-m} \leq \frac{1}{n^2},$$

giving us the first inequality. The second follows by union bound. ■

2.2. Planted entries are large

Now focus on $i \in S$. Let T_i be the set of k special entries in row i. We will use arguments similar to (8) to prove an upper bound on the l th moment of $B_{ij} - \mu_1$ for planted entries and use that to prove that $\sum_{T_i} B_{ij}$ is concentrated about its mean.

We first need to get a lower bound on $\mu_1 = \mathsf{E}_{p_1}(B_{ij})$:

$$\mu_1 \geq \frac{c}{\sigma_1} \int_0^M e^{x^2/4\sigma_0^2} e^{-x^2/4\sigma_0^2}\, dx = \frac{c}{\sigma_1} \int_0^M dx = cL.$$

Let $l \geq 2$ be an integer. Using Proposition (4), we get

$$\begin{aligned}
&\mathsf{E}_{p_1}((B_{ij} - \mu_1)^l) \\
&\leq 2\mathsf{E}_{p_1}(B_{ij}^l) \\
&\leq \frac{4}{\sqrt{2\pi}\sigma_1} \int_0^M \exp(\gamma l x^2) \exp(-x^2/2\sigma_1^2) + \frac{4\exp(\gamma l M^2)}{\sqrt{2\pi}\sigma_1} \int_M^\infty \frac{x}{M} \exp(-x^2/2\sigma_1^2) dx \\
&\leq \frac{2}{\sigma_1} \int_0^M \exp\left(Mx\left(\gamma l - \frac{1}{2\sigma_1^2}\right)\right)\, dx \;+\; \frac{2\sigma_1}{M} \exp\left(M^2\left(\gamma l - \frac{1}{2\sigma_1^2}\right)\right) \\
&\leq \frac{4}{\sigma_1 M(2\gamma - (1/2\sigma_1^2))} \exp\left(M^2\left(\gamma l - \frac{1}{2\sigma_1^2}\right)\right) \leq \frac{c}{L} \exp\left(\frac{L^2(l-1)}{4}\right). \qquad (9)
\end{aligned}$$

Lemma 8 *Let t be as in Lemma (7). Let*

$$t_2 = c\left(\ln n \exp(L^2/4) \;+\; \frac{\sqrt{k \ln n}}{\sqrt{L}} \exp(L^2/8)\right).$$

$$\mathsf{Pr}\left(\exists i \in S : \sum_{j \in T_i} (B_{ij} - \mu_1) < \; - t_2\right) \leq \frac{1}{n}.$$

$$\mathsf{Pr}\left(\exists i \in S : \sum_{j=1}^{n} (B_{ij} - \mu_0) < 100t\right) < \frac{1}{n}.$$

Proof First, fix attention on one $i \in S$. We use Theorem (5) with $X_j = B_{ij} - \mu_1$ for $j \in T_i$. We plug in (9) for $E(X_j^{2l})$ to get, with $m = 4 \ln N$:

$$\begin{aligned}
\mathsf{E}\left(\sum_{j \in T_i} (B_{ij} - \mu_1)\right)^m &\leq (cm \exp(L^2/4))^m \left[\sum_{l=1}^{m/2} \frac{1}{l^2} \left(\frac{k}{mL} \exp(-L^2/4)\right)^{1/l}\right]^{m/2} \\
&\leq (cm \exp(L^2/4))^m \left(\frac{k}{mL} \exp(-\,L^2/4) + 1\right)^{m/2},
\end{aligned}$$

the last using $x^{1/l} \leq x + 1$ for all $x > 0$. Now, we get that for a single $i \in S$, probability that $\sum_{j \in T_i} (B_{ij} - \mu_1) < \; - t_2$ is at most $1/n^2$ by using Markov inequality on $\left|\sum_{j \in T_i} (B_{ij} - \mu_1)\right|^m$. We get the first statement of the Lemma by a union bound over all $i \in S$.

For the second statement we have, using the same argument as in Lemma (7), with high probability,

$$\forall i \in S, \sum_{j \notin T_i} (B_{ij} - \mu_0) \geq \ -t. \tag{10}$$

We now claim that

$$kL > 100(t + t_2).$$

From the definition of t, t_2, it suffices to prove the following three inequalities to show this:

$$kL > c\sqrt{n}(\ln n)^{3/4} \ ; \ kL > c \ln n e^{L^2/4} \ ; \ kL > \frac{\sqrt{k \ln n}}{\sqrt{L}} e^{L^2/8}.$$

Each is proved by a straightforward (but tedious) calculation.

From the first assertion of the Lemma and (10), we now get that with high probability:

$$\sum_{j=1}^{n} (B_{ij} - \mu_0) \geq k(\mu_1 - \mu_0) - t_2 - t \geq 100(t + t_2),$$

proving Lemma (8).

■

3. Above the threshold: $\sigma_1^2 > 2\sigma_0^2$

Recall that all planted entries are $N(0, \sigma_1^2)$. There are k planted entries in each row of S. Assume (only) $\varepsilon > \frac{c}{\ln n}$. Define:

$$M^2 = 2\sigma_0^2(\ln n - \ln \varepsilon - \ln \ln N - \frac{1}{2} \ln \ln n) \quad \text{and} \quad B_{ij} = \exp\left(\gamma \text{Min}(M^2, A_{ij}^2)\right).$$

Theorem 9 *If $\varepsilon > c/\ln n$ and*

$$k > (\varepsilon \ln N \sqrt{\ln n})^{1 - \frac{1}{2(1+\varepsilon)}} n^{1/(2(1+\varepsilon))},$$

then with probability $1 - o(1)$, the top s row sums of B occur precisely in the S rows.

Corollary 10 *If $\varepsilon > c/\ln n$ and $k \in \tilde{\Omega}\left(n^{0.5 - \frac{\varepsilon}{2(1+\varepsilon)}}\right)$, then, with high probability, the top s row sums of B occur precisely in the S rows.*

3.1. Non-planted entries are small

Let

$$\begin{aligned} \mu_0 = \mathsf{E}_{p_0}(B_{ij}) &= \frac{1}{\sqrt{2\pi}\sigma_0} \int_{-\infty}^{\infty} \exp\left(\gamma \text{Min}(M^2, x^2)\right) \exp(-x^2/2\sigma_0^2) \\ &\leq \frac{1}{\sqrt{2\pi}\sigma_0} \int_{-\infty}^{\infty} \exp(-x^2/2\sigma_1^2) = \sqrt{2(1+\varepsilon)}. \end{aligned} \tag{11}$$

Let $l \geq 2$ be an integer. We note that $\gamma l - (1/2\sigma_0^2) > 0$ for $l \geq 2$. Using Proposition (4), we get (recall $i \notin S$)

$$\begin{aligned}
\mathsf{E}_{p_0}((B_{ij}-\mu_0)^l) &\leq 2\mathsf{E}_{p_0}(B_{ij}^l) \\
&\leq \frac{4}{\sqrt{2\pi}\sigma_0}\int_0^M \exp(\gamma l x^2)\exp(-x^2/2\sigma_0^2) + \frac{4\exp(\gamma l M^2)}{\sqrt{2\pi}\sigma_0}\int_M^\infty \frac{x}{M}\exp(-x^2/2\sigma_0^2)dx \\
&\leq \frac{2}{\sigma_0}\int_0^M \exp\left(Mx\left(\gamma l - \frac{1}{2\sigma_0^2}\right)\right)\,dx \;+\; \frac{2\sigma_0}{M}\exp\left(M^2\left(\gamma l - \frac{1}{2\sigma_0^2}\right)\right) \\
&\leq \frac{c\sigma_0}{M\varepsilon}\exp\left(M^2\left(\gamma l - \frac{1}{2\sigma_0^2}\right)\right), \qquad (12)
\end{aligned}$$

using $2\gamma - (1/2\sigma_0^2) = \frac{\varepsilon}{2\sigma_0^2(1+\varepsilon)} \geq \frac{\varepsilon}{4\sigma_0^2}$.

With $X_j = B_{ij} - \mu_0$, in Theorem (5), we plug in the bounds of (12) to get:

Lemma 11 *For all even m,*

$$\mathsf{E}_{p_0}\left(\sum_{j=1}^n (B_{ij}-\mu_0)\right)^m \leq (cm)^m e^{\gamma m M^2}\left[\sum_{l=1}^{m/2}\frac{1}{l^2}\left(\frac{cn\sigma_0}{mM\varepsilon}\exp(-M^2/(2\sigma_0^2))\right)^{1/l}\right]^{m/2}$$

$$\Longrightarrow$$

$$\mathsf{E}_{p_0}\left|\sum_{j=1}^n (B_{ij}-\mu_0)\right|^m \leq \left(cm\exp(\gamma M^2)\right)^m\left(1+\frac{cn\sigma_0}{mM\varepsilon}\exp(-M^2/2\sigma_0^2)\right)^{m/2} \qquad (13)$$

with $m = 4\ln N$.

Here, the last inequality is because $x^{1/l} \leq x+1$ for all real x and further $\sum_l (1/l^2)$ is a convergent series.

Lemma 12 *Let*

$$t = c(\ln N)\exp(\gamma M^2)\left(1+\frac{\sqrt{cn\sigma_0}}{\sqrt{mM\varepsilon}}\exp\left(-\frac{M^2}{4\sigma_0^2}\right)\right),$$

for c a suitable constant. For $i \notin S$,

$$\mathsf{Pr}\left(\left|\sum_{j=1}^n (B_{ij}-\mu_0)\right| \geq t\right) \leq \frac{1}{N^2}.$$

Thus, we have

$$\mathsf{Pr}\left(\exists i \notin S : \left|\sum_{j=1}^n (B_{ij}-\mu_0)\right| \geq t\right) \leq \frac{1}{N}.$$

Proof We use Markov's inequality on the random variable $\left|\sum_{j=1}^n (B_{ij}-\mu_0)\right|^m$ and (13) with m set to $4\ln N$ to get

$$\mathsf{Pr}\left(\left|\sum_{j=1}^n (B_{ij}-\mu_0)\right| \geq t\right) \leq e^{-m} \leq \frac{1}{N^2},$$

giving us the first inequality. The second follows by union bound. ■

3.2. Planted Entries are large

Now focus on $i \in S$. We will use arguments similar to (12) to prove an upper bound on the l th moment of $B_{ij} - \mu_1$ for planted entries and use that to prove that $\sum_{T_i} B_{ij}$ is concentrated about its mean. Let $l \geq 2$ be an integer. Using Proposition (4), we get

$$\begin{aligned}
&\mathsf{E}_{p_1}((B_{ij} - \mu_1)^l) \\
&\leq 2\mathsf{E}_{p_1}(B_{ij}^l) \\
&\leq \frac{4}{\sqrt{2\pi}\sigma_1}\int_0^M \exp(\gamma l x^2)\exp(-x^2/2\sigma_1^2) + \frac{4\exp(\gamma l M^2)}{\sqrt{2\pi}\sigma_1}\int_M^\infty \frac{x}{M}\exp(-x^2/2\sigma_1^2)dx \\
&\leq \frac{2}{\sigma_1}\int_0^M \exp\left(Mx\left(\gamma l - \frac{1}{2\sigma_1^2}\right)\right)\,dx \;+\; \frac{2\sigma_1}{M}\exp\left(M^2\left(\gamma l - \frac{1}{2\sigma_{ij}^2}\right)\right) \\
&\leq \frac{4}{\sigma_0 M(2\gamma - (1/2\sigma_1^2))}\exp\left(M^2\left(\gamma l - \frac{1}{2\sigma_1^2}\right)\right) \leq \frac{c\sigma_0}{M}\exp\left(M^2(\gamma l - (1/2\sigma_1^2))\right).
\end{aligned}$$

Now, applying Theorem (5), we get:

$$\mathsf{E}_{p_1}\left(\sum_{j\in T_i}(B_{ij}-\mu_1)^m\right) \leq (cm\exp(\gamma M^2))^m\left[\sum_{l=1}^{m/2}\frac{1}{l^2}\left(\frac{k\sigma_0}{mM}\exp(-M^2/2\sigma_1^2)\right)^{1/l}\right]^{m/2} \tag{14}$$

Lemma 13 *Let*

$$t_2 = c\ln N\exp(\gamma M^2)\left[1 + \frac{c\sqrt{k}}{\sqrt{\ln N}(\ln n)^{1/4}}\exp(-M^2/4\sigma_1^2)\right].$$

$$\mathsf{Pr}\left(\exists i \in S : \left|\sum_{j\in T_i}(B_{ij}-\mu_1)\right| \geq t_2\right) \leq \frac{1}{N}$$

$$\mathsf{Pr}\left(\exists i \in S : \sum_{j=1}^{n}(B_{ij}-\mu_0) < 50t\right) \leq \frac{1}{N}.$$

Proof The first statement of the Lemma follows from (14) with $m = 4\ln N$ by applying Markov inequality to $|\sum_{j\in T_i}(B_{ij}-\mu_1)|$ and then union bound over all $i \in S$ (using $\sum_l \frac{1}{l^2}x^{1/l} \leq \sum_l (1/l^2)(1+x) \leq c(1+x)$.)

For the second statement, we start with a lower bound on μ_1.

$$\mu_1 \geq \frac{c}{\sigma_1}\int_0^M \exp(\gamma x^2 - x^2/2\sigma_1^2) \geq \frac{c\sigma_0}{\varepsilon M}\exp(\gamma M^2 - (M^2/2\sigma_1^2)), \tag{15}$$

the last using: for $\lambda > 0$, $\int_0^M e^{\lambda x^2} \geq \int_{M-(1/\lambda M)}^M \exp(\lambda(M-(1/\lambda M))^2)dx \geq c\exp(\lambda M^2)/\lambda M$. [Note: We also needed: $M \geq 1/\varepsilon M$ which holds because $M \in O(\sqrt{\ln n})$ and $\varepsilon > c/\ln n$.] We assert that

$$k\mu_1 > ct, t_2.$$

This is proved by checcking three inequalities:

$$\frac{kc\sigma_0}{\varepsilon M}\exp(\gamma M^2 - (M^2/2\sigma_1^2)) > c\ln N\exp(\gamma M^2)$$
$$\frac{kc\sigma_0}{\varepsilon M}\exp(\gamma M^2 - (M^2/2\sigma_1^2)) > c\ln N\exp(\gamma M^2)\frac{\sqrt{n\sigma_0}}{\sqrt{mM\varepsilon}}\exp(-M^2/4\sigma_0^2)$$
$$\frac{kc\sigma_0}{\varepsilon M}\exp(\gamma M^2 - (M^2/2\sigma_1^2)) > \frac{c\ln N\exp(\gamma M^2)\sqrt{k}}{(\ln N)^{1/2}(\ln n)^{1/4}}\exp(-M^2/4\sigma_1^2).$$

These all hold as can be checked by doing simple calculations.

Now, we have

$$\sum_{j=1}^n (B_{ij}-\mu_0) = k(\mu_1-\mu_0) + \sum_{j\in T_i}(B_{ij}-\mu_1) + \sum_{j\notin T_i}(B_{ij}-\mu_0).$$

The last term is at least $-t$ with high probability (the proof is exactly as for the non-planted entries). The second term is at least $-t_2$ (whp). We have already shown that $\mu_0 \leq \sqrt{2}$ and that $k\mu_1 > 100(t+t_2+\mu_0)$. This proves the second statement of the Lemma. ■

Lemmas (13) and (12) together prove Theorem (9).

Noise Tolerance This algorithm can tolerate (adversarial) noise which can perturb $\tilde{\Omega}(e^{1/2\varepsilon})$ (which is, for example, a power of n when $\varepsilon = c/\ln n$) of the planted entries in each row of S. Here is a sketch of the argument for this: Note that the crucial lower bound on planted row sums in B comes from the lower bound on $k\mu_1$, the expected row sum in S rows. The lower bound of L on μ_1 involves the integral (15). It is easy to see that we only loose a constant factor if the integral is taken from 0 to $M - \frac{\sigma_0^2}{\varepsilon M}$ (instead of to M). Thus, corruption of all $x \in \left[M - \frac{\sigma_0^2}{\varepsilon M}\ ,\ M\right]$ would only cost a constant factor. It is easy to see that (i) there are $\tilde{\Omega}(e^{1/2\varepsilon})$ points in this interval and (ii) these are the worst possible points to be corrupted.

4. Generalization to unequal variances of planted entries

We assume the non-planted entries of an $N \times n$ matrix are drawn from $N(0,\sigma_0^2)$. There is again a set S of "planted" rows, with $|S| = k$. For each $i \in S$, now we assume there is some subset T_i of "planted entries". [But $|T_i|$ are not equal and we are not given $|T_i|$.] Planted entry (i,j) has distribution $p_{ij} \sim N(0,\sigma_{ij}^2)$. We assume each planted

$$\sigma_{ij}^2 \geq \sigma_1^2, \text{ where, } \sigma_1^2 = 2(1+\varepsilon)\sigma_0^2, \varepsilon > 0.$$

$$\text{Let } \tau_i = \sum_{j\in T_i} n^{-\sigma_0^2/\sigma_{ij}^2}. \tag{16}$$

$$\text{Let } \gamma = \frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2}. \tag{17}$$

Define M by:

$$M = \sqrt{2}\sigma_0\sqrt{\ln n}. \tag{18}$$

$$B_{ij} = \exp\left(\gamma \text{Min}(M^2, A_{ij}^2)\right). \tag{19}$$

Theorem 14 *With the above notation, if, for all $i \in S$,*

$$\tau_i \geq \frac{1}{\sqrt{\varepsilon}} c(\ln N)(\ln n)^{0.5},$$

then, with high probability, the set of k rows of B with the largest row sums is precisley S.

Corollary 15 *If $|T_i| = k$ for all $i \in S$ and all planted $\sigma_{ij}^2 = \sigma_1^2$, and*

$$k = n^{.5-\delta}, \text{ with } \varepsilon \geq \frac{2\delta}{1-2\delta} + \frac{\ln\ln N}{\ln n} + \frac{\ln\ln n}{2\ln n},$$

then, with high probability, the largest k row sums of B occur in the S rows.

The analysis for the non-planted entries is the same as before.

4.1. Planted Entries are large

Now focus on $i \in S$. We will use arguments similar to (12) to prove an upper bound on the l th moment of $B_{ij} - \mu_{ij}$ ($\mu_{ij} = E_{p_{ij}}(B_{ij})$) for planted entries and use that to prove that $\sum_{T_i} B_{ij}$ is concentrated about its mean. Let $l \geq 2$ be an integer. Using Proposition (4), we get

$$\begin{aligned}
&\mathsf{E}_{p_{ij}}((B_{ij} - \mu_{ij})^l) \\
&\leq 2\mathsf{E}_{p_1}(B_{ij}^l) \\
&\leq \frac{4}{\sqrt{2\pi}\sigma_{ij}} \int_0^M \exp(\gamma l x^2)\exp(-x^2/2\sigma_{ij}^2) + \frac{4\exp(\gamma l M^2)}{\sqrt{2\pi}\sigma_{ij}} \int_M^\infty \frac{x}{M}\exp(-x^2/2\sigma_{ij}^2)dx \\
&\leq \frac{2}{\sigma_{ij}} \int_0^M \exp\left(Mx\left(\gamma l - \frac{1}{2\sigma_{ij}^2}\right)\right) dx + \frac{2\sigma_{ij}}{M}\exp\left(M^2\left(\gamma l - \frac{1}{2\sigma_{ij}^2}\right)\right) \\
&\leq \frac{4}{\sigma_0 M(2\gamma - (1/2\sigma_{ij}^2))}\exp\left(M^2\left(\gamma l - \frac{1}{2\sigma_{ij}^2}\right)\right) \leq \frac{c\sigma_0}{M}\exp\left(M^2(\gamma l - (1/2\sigma_{ij}^2))\right).
\end{aligned} \tag{20}$$

Lemma 16 *For $i \in S$, let $t_i = c\ln N \exp(\gamma M^2)\left(1 + \frac{\sqrt{\tau_i}}{\sqrt{\ln N}(\ln n)^{1/4}}\right)$.*

$$\mathsf{Pr}\left(\exists i \in S : \sum_{j\in T_i}(B_{ij} - \mu_{ij}) < -t_i\right) \leq \frac{1}{N}.$$

$$\mathsf{Pr}\left(\exists i \in S : \sum_{j=1}^{n}(B_{ij} - \mu_0) < 100t\right) < \frac{1}{N}.$$

Proof First, fix attention on one $i \in S$. We use a more general version of Theorem (5) also from (Kannan (2009)):

Theorem 17 *If $X_1, X_2, \ldots, X_n$ are independent (not necessarily identical) mean 0 random variables, for any even positive integer m, we have*

$$\mathsf{E}\left(\left(\sum_{j=1}^{n} X_j\right)^m\right) \leq (cm)^m \left[\sum_{l=1}^{m/2} \frac{1}{l^2} \left(\sum_{j=1}^{n} \frac{\mathsf{E}(X_j^{2l})}{m}\right)^{1/l}\right]^{m/2}.$$

We apply this with $X_j = B_{ij} - \mu_{ij}$ for $j \in T_i$. We plug in (20) for $\mathsf{E}(X_j^{2l})$ to get, with $m = 4 \ln N$:

$$\begin{aligned}
\mathsf{E}\left(\sum_{j\in T_i}(B_{ij}-\mu_{ij})\right)^m &\leq (cm\exp(\gamma M^2))^m \left[\sum_{l=1}^{m/2}\frac{1}{l^2}\left(\sum_{j\in T_i}\frac{1}{mM}\exp(-M^2/2\sigma_{ij}^2)\right)^{1/l}\right]^{m/2} \\
&\leq (cm)^m \exp(\gamma m M^2)\left[\sum_{l=1}^{m/2}\frac{1}{l^2}\left(\sum_{j\in T_i}\frac{1}{mM}n^{-\sigma_0^2/\sigma_{ij}^2}\right)^{1/l}\right]^{m/2} \\
&\leq (cm)^m \exp(\gamma m M^2)\left(1+\frac{\tau_i^{m/2}}{(mM)^{m/2}}\right),
\end{aligned}$$

the last using $x^{1/l} \leq x+1$ for all $x > 0$.

Now, with $m = 4 \ln N$, we get that for a single $i \in S$, probability that $\sum_{j\in T_i}(B_{ij}-\mu_1) < -t_i$ is at most $1/N^2$ by using Markov inequality on $\left|\sum_{j\in T_i}(B_{ij}-\mu_1)\right|^m$ (noting: $M \geq c\sqrt{\ln n}$). We get the first statement of the Lemma by a union bound over all $i \in S$.

For the second statement, we first need to get a lower bound on μ_{ij}:

$$\mu_{ij} \geq \int_{x=0}^{M} \frac{c}{\sigma_{ij}} \exp(\gamma x^2 - x^2/2\sigma_{ij}^2)\, dx \geq \frac{c\sigma_0}{M}\exp(\gamma M^2 - M^2/2\sigma_{ij}^2),$$

the last using: for $\lambda > 0$, $\int_0^M e^{\lambda x^2} \geq \int_{M-(1/\lambda M)}^{M} \exp(\lambda(M-(1/\lambda M))^2)dx \geq c\exp(\lambda M^2)/\lambda M$. So,

$$\sum_{j\in T_i}\mu_{ij} \geq \frac{c\sigma_0}{M}\exp(\gamma M^2)\tau_i. \tag{21}$$

We have, using the same argument as in Lemma (12), with high probability,

$$\forall i \in S, \sum_{j\notin T_i}(B_{ij}-\mu_0) \geq -t. \tag{22}$$

Thus, from (22), (21) and the first assertion of the current Lemma,

$$\begin{aligned}
\sum_{j=1}^{n}(B_{ij}-\mu_0) &= \sum_{j\in T_i}(B_{ij}-\mu_{ij}) + \sum_{j\in T_i}(\mu_{ij}-\mu_0) + \sum_{j\notin T_i}(B_{ij}-\mu_{ij}) \\
&\geq -t_i + \frac{c\sigma_0}{M}\exp(\gamma M^2) - t.
\end{aligned}$$

We would like to assert the follwing inequalities, which together prove the second assertion of the Lemma.

$$\begin{aligned}\frac{c\sigma_0}{M}\exp(\gamma M^2)\tau_i &> c\ln N\exp(\gamma M^2)\\ &> c(\ln N)\exp(\gamma M^2)\frac{\sqrt{\tau_i}}{\sqrt{\ln N}(\ln n)^{1/4}}\\ &> c\ln N\exp(\gamma M^2)\left(\frac{\sqrt{cn\sigma_0}}{\sqrt{mM\varepsilon}}\exp(-M^2/4\sigma_0^2)\right).\end{aligned}$$

Each follows by a simple calculation.

■

5. Statistical algorithms and lower bounds

For problems over distributions, the input is a distribution which can typically be accessed via a sampling oracle that provide iid samples from the unknown distribution. *Statistical* algorithms are a restricted class of algorithms that are only allowed to query functions of the distribution rather than directly access samples. We consider three types of statistical query oracles from the literature. Let X be the domain over which the input distribution D is defined (e.g., $\{-1,1\}^n$ or $\mathbb{R}^n$).

1. STAT(τ): For any bounded function $f : X \to [-1,1]$, and any $\tau \in [0,1]$, STAT(τ) returns a number $p \in [\mathsf{E}_D(f(x)) - \tau, \mathsf{E}_D(f(x)) + \tau]$.

2. VSTAT(t): For any function $f : X \to \{0,1\}$, and any integer $t > 0$, VSTAT(t) returns a number $p \in [\mathsf{E}_D(f(x)) - \gamma, \mathsf{E}_D(f(x)) + \gamma]$ where $\gamma = \text{Max}\left\{\frac{1}{t}, \sqrt{\frac{\text{Var}_D(f)}{t}}\right\}$. Note that in the second term, $\text{Var}_D(f) = \mathsf{E}_D(f)(1 - \mathsf{E}_D(f))$.

3. 1-STAT: For any $f : X \to \{0,1\}$, returns $f(x)$ on a single random sample from D.

The first oracle was defined by Kearns in his seminal paper Kearns (1993, 1998) showing a lower bound for learning parities using statistical queries and analyzed more generally by Blum et al. Blum et al. (1994). The second oracle was introduced in Feldman et al. (2013a) to get stronger lower bounds, including for the planted clique problem. For relationships between these oracles (and simulations of one by another), the reader is referred to Feldman et al. (2013a,b).

Our algorithm for the hidden hubs problem can be made statistical. We focus on the detection problem $\mathcal{P}$: determine with probability at least $3/4$ whether the input distribution is $N(0,\sigma_0^2)$ for every entry with no planting, or if it is a hidden hubs instance, i.e., on a fixed k-subset of coordinates, the distribution is a mixture of $N(0,\sigma_0^2)$ and $N(\mu,\sigma_1^2)$ where the latter distribution is used with mixing weight k/n. To get a statistical version of our algorithm (p_1/p_0), consider the following query function f: For a random sample (column) x, truncate each entry, apply $p_1/p_0 - \mu_0$, add all the entries and output 1 if the sum exceeds t_0; else output 0.

By Lemmas 12 and 16, with $T_0 = 100t$ and the threshold t as in Lemma 12, we have the following consequence: if there is no planting, the probability that this query is 1 is at most $1/N$, while if there is a planting it is one with probability at least $\frac{k}{n}(1-\frac{1}{N})$. Thus it suffices to approximate

the expectation to within relative error $1/2$. To do this with VSTAT(t), we set $t = Cn/k$ for a large enough constant C. Thus, a planted Gaussian of size $n^{0.5-\delta}$ can be detected with a single query to VSTAT($O(n/k)$), provided $\sigma_1^2 \geq 2(1+\epsilon)\sigma_0^2$.

We will now prove that this upper bound is essentially tight. For $c\sigma_0^2 \leq \sigma_1^2 \leq 2\sigma_0^2$, for any $c > 0$, and $k = n^{0.5-\delta}$ for any $\delta > 0$, any statistical algorithm that detects hidden hubs must have superpolynomial complexity. For the lower bounds we assume the planted entries are drawn from $N(\mu, \sigma_1^2)$. The cases of most interest are (a) $\mu = 0$ and (b)$\sigma_1 = \sigma_2$. In both cases, the lower bounds will nearly match algorithmic upper bounds.

Theorem 18 *For a planting of size $k = n^{\frac{1}{2}-\delta}$,*

1. *For $\mu = 0$ and $c\sigma_0^2 \leq \sigma_1^2 \leq 2\sigma_0^2(1-\epsilon)$, any constant $c > 0$, any statistical algorithm that solves $\mathcal{P}$ with probability at least $3/4$ needs $n^{\Omega(\log n)}$ calls to VSTAT($n^{1+\delta}$).*
2. *For $\mu = 0$ and $\sigma_1^2 = 2\sigma_0^2$, any statistical algorithm that solves $\mathcal{P}$ with probability at least $3/4$ needs $n^{\Omega(\log n/\log\log n)}$ calls to VSTAT($n^{1+\delta}$).*
3. *For $\mu = 0$ and $\sigma_1^2 \leq (2+o(\delta))\sigma_0^2$, any statistical algorithm that solves $\mathcal{P}$ with probability at least $3/4$ needs $n^{\omega(1)}$ calls to VSTAT($n^{1+\delta}$).*
4. *For $\sigma_1 = \sigma_0$, if $\mu^2 = o(\sigma^2 \ln(\sqrt{n}/k))$, any statistical algorithm that solves $\mathcal{P}$ with probability at least $3/4$ needs $n^{\omega(1)}$ calls to VSTAT($n^{1+\delta}$).*

Moreover, the number of queries to 1-STAT for any of the above settings is $\Omega(n^{1+\delta})$.

The proof of the theorem is based on the notion of *Statistical Dimension with Average Correlation* defined in Feldman et al. (2013a). It is a generalization of statistical dimension as defined by Blum et al. (1994) for learning problems. We first need to define the correlation of two distributions A, B and a reference distribution U, all over a domain X,

$$\rho_U(A,B) = \mathsf{E}_X\left(\left(\frac{A(x)}{U(x)} - 1\right)\left(\frac{B(x)}{U(x)} - 1\right)\right).$$

The average correlation of a set of distributions $\mathcal{D}$ with respect to reference distribution U is

$$\rho_U(\mathcal{D}) = \frac{1}{|\mathcal{D}|^2}\sum_{A,B\in\mathcal{D}} \rho_U(A,B).$$

Definition 19 *For $\bar{\gamma} > 0$, domain X, a set of distributions $\mathcal{D}$ over X and a reference distribution U over X the statistical dimension of $\mathcal{D}$ relative to U with average correlation $\bar{\gamma}$ is denoted by SDA($\mathcal{D}, U, \bar{\gamma}$) and defined to be the largest integer d such that for any subset $\mathcal{D}' \subset \mathcal{D}$, $|\mathcal{D}'| > |\mathcal{D}|/d \Rightarrow \rho_U(\mathcal{D}') \leq \bar{\gamma}$.*

The main application of this definition is captured in the following theorem.

Theorem 20 *(Feldman et al., 2013a) For any decision problem $\mathcal{P}$ with reference distribution U, let $\mathcal{D}$ be a set of distributions such that $d = SDA(\mathcal{D}, U, \bar{\gamma})$. Then any randomized algorithm that solves $\mathcal{P}$ with probability at least $\nu > \frac{1}{2}$ must make at least $(2\nu-1)d$ queries to $VSTAT(1/3\bar{\gamma})$. Moreover, any algorithm that solves $\mathcal{P}$ with probability at least $3/4$ needs $\Omega(1)\min\{d, \frac{1}{\bar{\gamma}}\}$ calls to 1-STAT.*

5.1. Average correlation

For two subsets S, T, each of size k, the correlation of their corresponding distributions F_S, F_T is

$$\rho(F_S, F_T) = \left\langle \frac{F_S(x)}{F(x)} - 1, \frac{F_T(x)}{F(x)} - 1 \right\rangle_F = \mathsf{E}_F \left(\left(\frac{F_S(x)}{F(x)} - 1 \right) \left(\frac{F_T(x)}{F(x)} - 1 \right) \right)$$

where F is the distribution with no planting, i.e., $N(0, \sigma_0^2)^n$. For proving the lower bound at the threshold $\sigma_1^2 = 2\sigma_0^2$, it will be useful to define $\bar{F}_S$ as F_S with each coordinate restricted to the interval $[-M, M]$. We will set $M = \sigma_1 \sqrt{C \ln k}$. As before, we focus on the range $\sigma_1^2 \in [c\sigma_0^2, (2 + o(1))\sigma_0^2]$.

Lemma 21 *For $\sigma_1^2 < 2\sigma_0^2$*

$$\rho(F_S, F_T) = \frac{k^2}{n^2} \left(\left(\frac{\sigma_0^2}{\sigma_1 \sqrt{2\sigma_0^2 - \sigma_1^2}} \right)^{|S \cap T|} \exp\left(\frac{\mu^2}{2\sigma_0^2 - \sigma_1^2} \cdot |S \cap T| \right) - 1 \right).$$

For $\sigma_1^2 = 2\sigma_0^2$,

$$\rho(\bar{F}_S, \bar{F}_T) \leq \frac{k^2 (C \ln k)^{|S \cap T|/2}}{n^2}.$$

For $\sigma_1^2 = (2 + \alpha)\sigma_0^2$ and $\alpha = o(1)$,

$$\rho(\bar{F}_S, \bar{F}_T) \leq \frac{k^2}{n^2} k^{C\alpha |S \cap T|/4}.$$

Proof

$$\begin{aligned}
\rho(F_S, F_T) &= \left\langle \frac{F_S(x)}{F(x)} - 1, \frac{F_T(x)}{F(x)} - 1 \right\rangle_F \\
&= \int \frac{dF_S(x)\, dF_T(x)}{dF(x)} - 1 \\
&= \frac{k^2}{n^2} \left(\Pi_{i \in S \cap T} \frac{\sigma_0}{\sqrt{2\pi}\sigma_1^2} \int \exp\left(-\frac{(x_i - \mu)^2}{2\sigma_1^2} - \frac{(x_i - \mu)^2}{2\sigma_1^2} + \frac{x_i^2}{2\sigma_0^2} \right) - 1 \right) \\
&= \frac{k^2}{n^2} \left(\Pi_{i \in S \cap T} \frac{\sigma_0}{\sqrt{2\pi}\sigma_1^2} \int \exp\left(-x_i^2 \cdot \frac{2\sigma_0^2 - \sigma_1^2}{2\sigma_1^2 \sigma_0^2} - \frac{2\mu^2 - 4x_i\mu}{2\sigma_1^2} \right) - 1 \right)
\end{aligned}$$

Setting $z = \frac{\sigma_1 \sigma_0}{\sqrt{2\sigma_0^2 - \sigma_1^2}}$,

$$\rho(F_S, F_T) = \frac{k^2}{n^2} \left(\prod_{i \in S \cap T} \frac{\sigma_0}{\sqrt{2\pi}\sigma_1^2} \int \exp\left(-\frac{(x_i - 2\mu z^2/\sigma_1^2)^2}{2z^2} + \mu^2 \left(\frac{2z^2}{\sigma_1^4} - \frac{1}{\sigma_1^2} \right) \right) - 1 \right).$$

We note that if $z^2 \leq 0$, then the integral diverges. Assuming that $z^2 > 0$.

$$\begin{aligned}
&\rho(F_S, F_T) \\
&= \frac{k^2}{n^2}\left(\prod_{i\in S\cap T}\frac{\sigma_0}{\sqrt{2\pi}\sigma_1^2}\int \exp\left(-\frac{(x_i - 2\mu z^2/\sigma_1^2)^2}{2z^2} + \mu^2\left(\frac{2\sigma_0^2}{\sigma_1^2(2\sigma_0^2-\sigma_1^2)} - \frac{1}{\sigma_1^2}\right)\right) - 1\right) \\
&= \frac{k^2}{n^2}\left(\exp\left(\frac{\mu^2|S\cap T|}{2\sigma_0^2-\sigma_1^2}\right)\prod_{i\in S\cap T}\frac{\sigma_0}{\sqrt{2\pi}\sigma_1^2}\int \exp\left(-\frac{(x_i - 2\mu z^2/\sigma_1^2)^2}{2z^2}\right) - 1\right) \\
&= \frac{k^2}{n^2}\left(\left(\exp\left(\frac{\mu^2}{2\sigma_0^2-\sigma_1^2}\right)\frac{\sigma_0 z}{\sigma_1^2}\right)^{|S\cap T|} - 1\right) \\
&= \frac{k^2}{n^2}\left(\left(\frac{\sigma_0^2}{\sigma_1\sqrt{2\sigma_0^2-\sigma_1^2}}\exp\left(\frac{\mu^2}{2\sigma_0^2-\sigma_1^2}\right)\right)^{|S\cap T|} - 1\right)
\end{aligned}$$

Note that $\sigma_0^2 \geq \sigma_1\sqrt{2\sigma_0^2-\sigma_1^2}$, so the above bound is of the form $\alpha\beta^{|S\cap T|}$, where $\beta > 1$. For the second part, we have

$$\begin{aligned}
\rho(\bar{F}_S, \bar{F}_T) &\leq \frac{k^2}{n^2}\left(\frac{\sigma_0}{\sqrt{2\pi}\sigma_1^2}\int_{-M}^{M} 1\,dx\right)^{|S\cap T|} \\
&\leq \frac{k^2}{n^2}\left(\frac{C\ln k}{2}\right)^{|S\cap T|/2}.
\end{aligned}$$

The last part is similar. With $\sigma_1^2 = (2+\alpha)\sigma_0^2$,

$$\begin{aligned}
\rho(\bar{F}_S, \bar{F}_T) &\leq \frac{k^2}{n^2}\left(\frac{\sigma_0}{\sqrt{2\pi}\sigma_1^2}\int_{-M}^{M} e^{\frac{\alpha x^2}{2\sigma_1^2}}\,dx\right)^{|S\cap T|} \\
&\leq \frac{k^2}{n^2}\left(k^{C\alpha/2}\right)^{|S\cap T|/2}.
\end{aligned}$$

■

5.2. Statistical dimension of planted Gaussian

Lemma 22 *Let $\sigma_1^2 < 2\sigma_0^2$ and D be set of distributions induced by every possible subset of $[n]$ of size k. Assume $\rho(F_S, F_T) \leq \alpha\beta^{|S\cap T|}$ for some $\beta > 1$. Then, for any subset $A \subset D$ with*

$$|A| \geq \frac{2\binom{n}{k}}{\ell!(n/2k^2)^\ell},$$

the average correlation of A with any subset S is at most

$$\rho(A, S) = \frac{1}{|A|}\sum_{T\in A}\rho(F_T, F_S) \leq 2\alpha\beta^\ell.$$

Proof This proof is similar to Feldman et al. (2013a). Define $T_r = \{T \in A : |T \cap S| = r\}$. Then,

$$\sum_{T \in A} \rho(F_S, F_T) \leq \alpha \sum_{T \in A} \beta^{|S \cap T|} = \alpha \sum_{r=r_0}^{k} |T_r \cap A| \beta^r.$$

To maximize the bound, we would include in A sets that intersect S in $k-1$ indices, then $k-2$ indices and so on. Taking this extremal choice of A gives us a lower bound on the minimum intersection size r_0 as follows. Note that for $0 \leq j \leq k-1$,

$$\begin{aligned}
\frac{|T_{j+1}|}{|T_j|} &= \frac{\binom{k}{j+1}\binom{n-k}{k-j-1}}{\binom{k}{j}\binom{n-k}{k-j}} \\
&= \frac{(k-j)^2}{(j+1)(n-2k+j+1)} \\
&\leq \frac{k^2}{jn}
\end{aligned}$$

where the last step assumes $2k^2 < n$. Therefore,

$$|T_j| \leq \frac{1}{j!}\left(\frac{k^2}{n}\right)^j |T_0| \leq \frac{\binom{n}{k}}{j!(n/k^2)^j}.$$

This gives a bound on the minimum intersection size since

$$\sum_{j=r_0}^{k} |T_j| < \frac{2\binom{n}{k}}{r_0!(n/k^2)^{r_0}}$$

Therefore under the assumption on $|A|$, we get that $r_0 < \ell$. Using this,

$$\begin{aligned}
\sum_{T \in A} \rho(F_S, F_T) &\leq \alpha \sum_{r=r_0}^{k} |T_r \cap A| \beta^r \\
&\leq \alpha \left(|T_{r_0} \cap A| \beta^{r_0} + \sum_{r=r_0+1}^{k} |T_r| \beta^r \right) \\
&\leq \alpha \left(|T_{r_0} \cap A| \beta^{r_0} + 2|T_{r_0+1}| \frac{\beta^{r_0+1} - 1}{(r_0+1)(\beta - 1)} \right) \\
&\leq 2\alpha |A| \beta^{r_0+1} \leq 2\alpha \beta^{\ell} |A|.
\end{aligned}$$

■

Theorem 23 *For the planted Gaussian problem $\mathcal{P}$, with (a) $\sigma_1^2 < 2\sigma_0^2$, and average correlation at most*

$$\bar{\gamma} = 2\frac{k^2}{n^2}\left(\frac{\sigma_0^2}{\sigma_1 \sqrt{2\sigma_0^2 - \sigma_1^2}} \exp\left(\frac{\mu^2}{2\sigma_0^2 - \sigma_1^2} \right) \right)^{\ell}$$

or (b) $\sigma_1^2 = 2\sigma_0^2$*, and average correlation*

$$\bar{\gamma} = 2\frac{k^2}{n^2}\left(\frac{C\ln k}{2}\right)^{\ell/2}$$

or (c) $\sigma_1^2 = (2+\alpha)\sigma_0^2$ *for* $\alpha = o(1)$*, and average correlation*

$$\bar{\gamma} = 2\frac{k^2}{n^2}k^{C\alpha\ell/4}$$

the statistical dimension of $\mathcal{P}$ *is at least* $\ell!(n/k^2)^\ell/2$*.*

We now state explicitly the three main corollaries of this theorem. This completes the proof of Theorem 18.

Corollary 24 *With* $\mu = 0$*, and* $\sigma_1^2 = 2\sigma_0^2(1-\epsilon)$*, we have*

$$\bar{\gamma} = 2\frac{k^2}{n^2}\left(\frac{1}{4\epsilon(1-\epsilon)}\right)^{\ell/2}$$

and for any $\delta > 0$*, with* $k = n^{0.5-\delta}$*,* $\ell = c\log n/\log(1/\epsilon(1-\epsilon))$*, we have* $\bar{\gamma} = 2n^{c-2\delta-1}$ *and*

$$SDA(\mathcal{P}, \bar{\gamma}) = \Omega(n^{2\delta\log_{\frac{1}{\epsilon(1-\epsilon)}} n}).$$

Hence with $c = \delta$*, any statistical algorithm that solves* $\mathcal{P}$ *with probability at least* $3/4$ *needs* $n^{\Omega(\log n)}$ *calls to VSTAT(*$n^{1+\delta}$*).*

We note that the above corollary applies for any $0 < \sigma_1^2 < 2\sigma_0^2$, with the bounds depending mildly on how close σ_1^2 is to the ends of this range. This is quantified by the dependence on $\epsilon(1-\epsilon)$ above.

Our lower bound extends slightly above the threshold $\sigma_1^2 = 2\sigma_0^2$. For this, we need to observe that with respect to any n^C samples, the distributions F_S and $\hat{F}_S$ are indistinguishable with high probability $(1 - n^{-C})$. Therefore, proving a lower bound on the statistical dimension of $\mathcal{P}$ with distributions $\hat{F}_S$ is effectively a lower bound for the original problem $\mathcal{P}$ with distributions F_S.

Corollary 25 *With* $\mu = 0$*, and* $\sigma_1^2 = 2\sigma_0^2$*, we have*

$$\bar{\gamma} = 2\frac{k^2}{n^2}\left(\frac{C\ln k}{2}\right)^{\ell/2}$$

and for any $\delta > 0$*, with* $k = n^{0.5-\delta}$*,* $\ell = c\log n/2\log\log k$*, we have* $\bar{\gamma} = 2n^{c-2\delta-1}$ *and*

$$SDA(\mathcal{P}, \bar{\gamma}) = \Omega(n^{\delta\log n/\log\log n}).$$

Hence with $c = \delta$*, any statistical algorithm that solves* $\mathcal{P}$ *with probability at least* $3/4$ *needs* $n^{\Omega(\log n/\log\log n)}$ *calls to VSTAT(*$n^{1+\delta}$*). Moreover, for* $\sigma_1^2 = (2+\alpha)\sigma_0^2$*,* $\alpha = o(\delta)$*, we have*

$$\bar{\gamma} = 2\frac{k^2}{n^2}k^{C\alpha\ell/4}$$

and for any $\delta > 0$, with $k = n^{0.5-\delta}$, $\ell = 8\delta/C\alpha$, we have $\bar{\gamma} = 2n^{-\delta-1}$ and

$$SDA(\mathcal{P}, \bar{\gamma}) \geq n^{\delta\ell}.$$

Hence any statistical algorithm that solves $\mathcal{P}$ with probability at least $3/4$ needs $n^{\omega(1)}$ calls to VSTAT($n^{1+\delta}$).

Corollary 26 *For $\sigma_1 = \sigma_0$,*

$$\bar{\gamma} = 2\frac{k^2}{n^2}\exp\left(\frac{\mu^2\ell}{\sigma^2}\right).$$

and for any $\delta > 0$, with $k = n^{0.5-\delta}$, $\mu^2 = c\sigma^2\ln(\sqrt{n}/k)$, we have $\bar{\gamma} = 2n^{c\delta\ell-2\delta-1}$ and

$$SDA(\mathcal{P}, \bar{\gamma}) = \Omega(n^{2\delta\ell}).$$

If $\mu^2 = o(\sigma^2\ln(\sqrt{n}/k))$, any statistical algorithm that solves $\mathcal{P}$ with probability at least $3/4$ needs $n^{\omega(1)}$ calls to VSTAT($n^{1+\delta}$).

References

N. Alon, M. Krivelevich, and B. Sudakov. Finding a large hidden clique in a random graph. *Random Structures and Algorithms*, 13:457–466, 1998.

Boaz Barak, Samuel B. Hopkins, Jonathan A. Kelner, Pravesh Kothari, Ankur Moitra, and Aaron Potechin. A nearly tight sum-of-squares lower bound for the planted clique problem. *CoRR*, abs/1604.03084, 2016. URL http://arxiv.org/abs/1604.03084.

Aditya Bhaskara, Moses Charikar, Eden Chlamtac, Uriel Feige, and Aravindan Vijayaraghavan. Detecting high log-densities: an $O(n^{1/4})$ approximation for densest k-subgraph. In *Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC 2010, Cambridge, Massachusetts, USA, 5-8 June 2010*, pages 201–210, 2010. doi: 10.1145/1806689.1806718. URL http://doi.acm.org/10.1145/1806689.1806718.

Avrim Blum, Merrick Furst, Jeffrey Jackson, Michael Kearns, Yishay Mansour, and Steven Rudich. Weakly learning DNF and characterizing statistical query learning using Fourier analysis. In *STOC*, pages 253–262, 1994.

R. Boppana. Eigenvalues and graph bisection: An average-case analysis. *Proceedings of the 28th IEEE Symposium on Foundations of Computer Science*, pages 280–285, 1987.

S. Charles Brubaker and Santosh Vempala. Random tensors and planted cliques. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, 12th International Workshop, APPROX 2009, and 13th International Workshop, RANDOM 2009, Berkeley, CA, USA, August 21-23, 2009. Proceedings*, pages 406–419, 2009. doi: 10.1007/978-3-642-03685-9_31. URL http://dx.doi.org/10.1007/978-3-642-03685-9_31.

Yael Dekel, Ori Gurel-Gurevich, and Yuval Peres. Finding hidden cliques in linear time with high probability. In *Proceedings of the Meeting on Analytic Algorithmics and Combinatorics*, ANALCO '11, pages 67–75, Philadelphia, PA, USA, 2011. Society for Industrial and Applied Mathematics. URL http://dl.acm.org/citation.cfm?id=2790409.2790417.

Yash Deshpande and Andrea Montanari. Improved sum-of-squares lower bounds for hidden clique and hidden submatrix problems. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, pages 523–562, 2015a. URL http://jmlr.org/proceedings/papers/v40/Deshpande15.html.

Yash Deshpande and Andrea Montanari. Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time. *Found. Comput. Math.*, 15(4):1069–1128, August 2015b. ISSN 1615-3375. doi: 10.1007/s10208-014-9215-y. URL http://dx.doi.org/10.1007/s10208-014-9215-y.

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

Uriel Feige and Dorit Ron. Finding hidden cliques in linear time. In *21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA'10)*, pages 189–204. Discrete Mathematics and Theoretical Computer Science, 2010.

Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh Vempala, and Ying Xiao. Statistical algorithms and a lower bound for planted clique. In *Proceedings of the 45th annual ACM symposium on Symposium on theory of computing*, pages 655–664. ACM, 2013a.

Vitaly Feldman, Will Perkins, and Santosh Vempala. On the complexity of random satisfiability problems with planted solutions. *CoRR*, abs/1311.4821, 2013b. Extended abstract in STOC 2015.

Vitaly Feldman, Cristobal Guzman, and Santosh Vempala. Statistical query algorithms for mean estimation and stochastic convex optimization. In *SIAM Symposium on Discrete Algorithms*, 2017. URL http://arxiv.org/abs/1512.09170.

Delphine Féral and Sandrine Péché. The largest eigenvalue of rank one deformation of large wigner matrices. *Communications in Mathematical Physics*, 272(1):185–228, 2007. ISSN 1432-0916. doi: 10.1007/s00220-007-0209-3. URL http://dx.doi.org/10.1007/s00220-007-0209-3.

Alan M. Frieze and Ravi Kannan. A new approach to the planted clique problem. In *IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science, FSTTCS 2008, December 9-11, 2008, Bangalore, India*, pages 187–198, 2008. doi: 10.4230/LIPIcs.FSTTCS.2008.1752. URL http://dx.doi.org/10.4230/LIPIcs.FSTTCS.2008.1752.

Samuel B. Hopkins, Pravesh Kothari, Aaron Henry Potechin, Prasad Raghavendra, and Tselil Schramm. On the integrality gap of degree-4 sum of squares for planted clique. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, pages 1079–1095, 2016. doi: 10.1137/1.9781611974331.ch76. URL http://dx.doi.org/10.1137/1.9781611974331.ch76.

M. Jerrum. Large cliques elude the metropolis process. *Random Structures and Algorithms*, 3(4):347–360, 1992.

Ravindran Kannan. A new probability inequality using typical moments and concentration results. In *50th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2009, October 25-27, 2009, Atlanta, Georgia, USA*, pages 211–220, 2009. doi: 10.1109/FOCS.2009.20. URL http://dx.doi.org/10.1109/FOCS.2009.20.

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

Michael J. Kearns. Efficient noise-tolerant learning from statistical queries. In *Proceedings of the Twenty-Fifth Annual ACM Symposium on Theory of Computing, May 16-18, 1993, San Diego, CA, USA*, pages 392–401, 1993. doi: 10.1145/167088.167200. URL http://doi.acm.org/10.1145/167088.167200.

L. Kucera. Expected complexity of graph partitioning problems. *Discrete Applied Mathematics*, 57:193–212, 1995.

Zongming Ma and Yihong Wu. Computational barriers in minimax submatrix detection. *Ann. Statist.*, 43(3):1089–1116, 06 2015. doi: 10.1214/14-AOS1300. URL http://dx.doi.org/10.1214/14-AOS1300.

Raghu Meka, Aaron Potechin, and Avi Wigderson. Sum-of-squares lower bounds for planted clique. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pages 87–96, 2015. doi: 10.1145/2746539.2746600. URL http://doi.acm.org/10.1145/2746539.2746600.

Andrea Montanari, Daniel Reichman, and Ofer Zeitouni. On the limitation of spectral methods: From the gaussian hidden clique problem to rank-one perturbations of gaussian tensors. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 217–225. Curran Associates, Inc., 2015.