Statistical-Computational Phase Transitions in Planted Models: The High-Dimensional Setting

Yudong Chen YUDONG.CHEN@EECS.BERKELEY.EDU
Department of EECS, University of California, Berkeley, Berkeley, CA 94704, USA

Jiaming Xu JXU18@ILLINOIS.EDU
Department of ECE, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Abstract

The planted models assume that a graph is generated from some unknown clusters by randomly placing edges between nodes according to their cluster memberships; the task is to recover the clusters given the graph. Special cases include planted clique, planted partition, planted densest subgraph and planted coloring. Of particular interest is the *high-dimensional* setting where the number of clusters is allowed to grow with the number of nodes. We show that the space of model parameters can be partitioned into four disjoint regions: (1) the *impossible* regime, where all algorithms fail; (2) the *hard* regime, where the computationally intractable Maximum Likelihood Estimator (MLE) succeeds, and no polynomial-time method is known; (3) the *easy* regime, where the polynomial-time convexified MLE succeeds; (4) the *simple* regime, where a simple counting/thresholding procedure succeeds. Moreover, each of these algorithms provably fails in the previous harder regimes. Our theorems establish the first minimax recovery results for the high-dimensional setting, and provide the best known guarantees for polynomial-time algorithms. These results demonstrate the tradeoffs between statistical and computational considerations.

1. Introduction

The planted models are standard models for generating a random graph from the underlying clustering structure. We consider a general setup called the *planted clustering* model. The model assumes that rK out of a total n nodes are grouped into r *clusters* of equal size K, while the other $n-rK$ nodes (called *isolated nodes*) are not in any clusters; each pair of nodes are connected by an edge independently with probability p if they are in the same cluster, and with probability q otherwise. The goal is to recover the underlying unknown clusters given the graph.

We are particularly interested in the so-called **high-dimenional** setting (Rohe et al., 2011) where the number r of clusters may grow unbounded with the problem dimensions n. This setting is important in many empirical networks (Leskovec et al., 2008), and more challenging to analyze than the $r = \Theta(1)$ setting. The parameters p, q and K can scale with n as well.

The formulation above covers many classical planted problems including planted clique/r-clique (Alon et al., 1998; McSherry, 2001), planted coloring (Alon & Kahale, 1997), planted densest subgraph (Arias-Castro & Verzelen, 2013), planted partition and the stochastic blockmodel (Holland et al., 1983; Condon & Karp, 2001). These models have a broad range of applications: They are used as generative models for approximating real world networks with natural cluster/community structures (Fortunato, 2010), and serve as benchmarks in the evaluation of clustering and community detection algorithms (Newman & Girvan, 2004); they also provide a standard venue for studying the average-case behaviors of NP-hard graph theoretic problems including max-clique, max-cut, graph partitioning and coloring (Bollobás & Scott, 2004).

The planted clustering problems pose themselves as both statistical and computational problems. Statistically, the parameters n, r, K, p, q govern the "noisiness" of the problems: The problems become statistically harder with smaller values of $p - q$, K and larger r, as the observations are noisier and the cluster structures are more complicated and weakly expressed in the data. A statistically powerful algorithm is one that can recover the clusters for a large range of model parameters.

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

Computationally, we are concerned with the running-time of different algorithms. An exhaustive search over all possible clusterings might make for a statistically powerful algorithm but has high time-complexity. A simpler algorithm with polynomial or even linear running time is computationally more desirable, but might succeed for a smaller region of the parameter space and thus have less statistical power.

Here we take a joint statistical-computational view to planted clustering, and try to understand the *tradeoffs* between these two aspects: How do algorithms with different computational time achieve different statistical performance? For what regions of the parameter space is recovery infeasible, either for any algorithm, or for an algorithm with specific computational time?

Our results highlight the following: The parameter space can be partitioned into four regions, where each region corresponds to successively easier instances of the problem than the previous region statistically, and recovery can be achieved by simpler algorithms with lower time complexities. Significantly, there are large gaps between the statistical performance of computationally expensive algorithms and computationally efficient algorithms. We elaborate below.

1.1. The Four Regimes for Planted Clustering

For concreteness, we first consider the setting with $r \geq 2$, $p > q$ and $p/q = \Theta(1)$. This covers the standard planted partition and planted r-clique models. The statistical hardness of the cluster recovery problem can be summarized by the quantity $\frac{(p-q)^2}{q(1-q)}$, essentially a measure of the Signal-to-Noise Ratio. Our main theorems identify the following four regimes of the problem defined by the values of this quantity.

- **The Impossible Regime:** $\frac{(p-q)^2}{q(1-q)} \dot{\lesssim} \frac{1}{K}$.[1] In this regime, there is no algorithm, regardless of its computational complexity, that can recover the clusters with reasonable probability.
- **The Hard Regime:** $\frac{1}{K} \dot{\lesssim} \frac{(p-q)^2}{q(1-q)} \dot{\lesssim} \frac{n}{K^2}$. The computationally expensive Maximum Likelihood Estimator (MLE) recovers the clusters with high probability in this regime (as well as in the next two easier regimes; we omit such implications in the sequel). No polynomial-time algorithm is known for this regime.
- **The Easy Regime:** $\frac{n}{K^2} \dot{\lesssim} \frac{(p-q)^2}{q(1-q)} \dot{\lesssim} \frac{\sqrt{n}}{K}$. There exists a polynomial-time algorithm – a convex relaxation of the MLE – which recovers the clusters with high probability here. Moreover, this algorithm provably fails in the hard regime above.
- **The Simple Regime:** $\frac{(p-q)^2}{q(1-q)} \dot{\gtrsim} \frac{\sqrt{n}}{K}$. A simple algorithm based on counting degrees and common neighbors recovers the clusters with high probability in this regime, and provably fails outside this regime (i.e., in the hard and easy regimes).

We illustrate these four regimes in Figure 1 assuming $p = 2q = \Theta(n^{-\alpha})$ and $K = \Theta(n^{\beta})$. Here cluster recovery is harder with larger α and smaller β. In this setting, the four regimes correspond to four disjoint and non-empty regions of the parameter space. Therefore, a computationally more complicated algorithm leads to a *significant* (order-wise) enhancement in statistical power. For example, when $p = 2q = n^{-1/4}$, the cluster sizes K that can be handled by the simple, polynomial-time and computationally expensive algorithms are $\Omega(n^{0.75})$, $\Omega(n^{0.625})$ and $\Omega(n^{0.25})$, respectively.

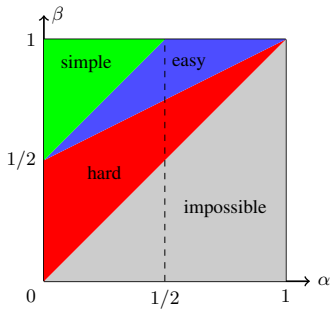

Figure 1. Illustration of the four regimes. It applies to the planted clustering problem with $p = 2q = \Theta(n^{-\alpha})$ and $K = \Theta(n^{\beta})$, as well as the submatrix localization problem with $n_L = n_R = n$, $\mu^2 = \Theta(n^{-\alpha})$ and $K_L = K_R = \Theta(n^{\beta})$.

The results in the impossible and hard regimes together establish the *minimax recovery boundary* of the planted clustering problem, and show that MLE is statistically optimal. These two regimes are separated by an "information barrier": In the impossible regime the graph does not carry enough information about the clusters, so recovery is statistically impossible.

We conjecture that no polynomial-time algorithm succeeds in the hard regime. This will mean the convex relaxation of MLE achieves the *computational limit*. While rigorously proving the conjecture is difficult, there are many evidences supporting it. The hard regime contains the standard Planted Clique problem with clique size $K = o(\sqrt{n})$, which has no polynomial-time algorithms so far despite decades of effort and is widely believed to be computationally intractable (Juels & Peinado, 2000). Moreover, there is a "spectral barrier", determined by the spectral norm of an appropriately defined noise matrix, that prevents the convexified MLE and spectral algorithms, and possibly many

[1] The notations $\dot{\gtrsim}$ and $\dot{\lesssim}$ ignore constant and $\log n$ factors; $\gtrsim$ and $\lesssim$ ignore constant factors.

other polynomial-time algorithms as well, from succeeding in the hard regime.

The simple counting algorithm fails outside the simple regime due to a *variance barrier* associated with the node degrees and the numbers of common neighbors. Therefore, the simple algorithm is order-wise weaker statistically than the convexified MLE.

General results: Our main theorems apply to general values of p, q, K and r. The four regimes and the statistical-computational tradeoffs can be observed for a broad spectrum of planted models. We discuss them in the main result section.

1.2. Extensions to Submatrix Localization

Similar results hold for the related *submatrix localization* problem, a.k.a. *bi-clustering* (Kolar et al., 2011). Here we assume $A \in \mathbb{R}^{n_L \times n_R}$ is a random matrix with i.i.d. Gaussian entries with unit variance, where there are r submatrices of size $K_L \times K_R$ with disjoint row and column supports, such that the entries inside the submatrices have mean $\mu > 0$, and the entries outside have mean zero. The goal is to locate these submatrices given A. We allow r to grow unbounded. This generalizes the single-submatrix model previously studied (Balakrishnan et al.; Arias-Castro et al., 2011).

The quantity μ^2 measures the Signal-to-Noise Ratio (SNR). Most interesting is the low SNR setting with $\mu^2 = O(\log n)$. Suppose $n_L = n_R = n$ and $K_L = K_R = K$; the problem has the following four regimes, with the same meaning as before:

- **The Impossible Regime:** $\mu^2 \lesssim \frac{1}{K}$.
- **The Hard Regime:** $\frac{1}{K} \lesssim \mu^2 \lesssim \frac{n}{K^2}$.
- **The Easy Regime:** $\frac{n}{K^2} \lesssim \mu^2 \lesssim \frac{\sqrt{n}}{K}$.
- **The Simple Regime:** $\frac{\sqrt{n}}{K} \lesssim \mu^2 \lesssim 1$.

We illustrate these regimes in Figure 1 assuming $\mu^2 = \Theta(n^{-\alpha})$ and $K = \Theta(n^{\beta})$. Complete results will be provided in a forthcoming full paper.

1.3. Discussions

The results above highlight the interaction between the statistical and computational considerations in planted clustering and submatrix localization. Our study parallels a recent line of work that takes a joint statistical and computational view on learning problems (e.g., Berthet & Rigollet (2013); Chandrasekaran & Jordan (2013)). While we investigate two specific problems, we expect that the phenomena and principles in this paper are relevant more generally. Below we provide additional discussions on our innovation compared to previous work.

The high dimensional setting: Several recent works investigate the statistical-computational tradeoffs in submatrix detection/localization (Kolar et al., 2011; Ma & Wu, 2013), planted densest subgraph detection (Arias-Castro & Verzelen, 2013) and sparse PCA (Berthet & Rigollet, 2013; Krauthgamer et al., 2013). Even earlier is the extensive study of the Planted Clique problem. The majority of these previous works focus on the setting with a single clique, cluster, submatrix or principal component (i.e., $r = 1$). In this paper, we study the more general *high-dimensional* setting with a growing of clusters/submatrices, which is more difficult and poses significant challenge to the analysis. Moeover, there are qualitative differences between these two settings, where are discussed in the next paragraph.

The power of convex relaxations: In previous work on the $r = 1$ setting of submatrix localization (Kolar et al., 2011) and sparse PCA (Krauthgamer et al., 2013), it is shown that very simple algorithms based on thresholding/averaging have the order-wise similar statistical performance as more sophisticated convex optimization approaches. In contrast, for planted clustering and submatrix localization with multiple clusters/submatrices, we show that convex relaxation approaches are order-wise statistically more powerful than simple counting/thresholding. Our analysis reveals that the power of convex relaxations lies in *separating different clusters/submatrices*, but not in identifying a single cluster/submatrix. This demonstrates a finer spectrum of computational-statistical tradeoffs.

Detection vs. estimation: Several recent works on planted densest subgraph (Arias-Castro & Verzelen, 2013), submatrix detection (Ma & Wu, 2013) and sparse PCA (Berthet & Rigollet, 2013) have focused on the *detection* or *hypothesis testing* version of the problems, i.e., deciding whether or not there is a structured cluster/submatrix/principal component. In this paper, we study the *estimation* problem, i.e., to estimate the locations of the clusters/submatrices. If we compare Figure 1 in this paper with Figure 1 in (Ma & Wu, 2013), we see that submatrix localization is strictly harder than its detection counterpart. We have a similar observation for planted densest subgraph by comparing with (Arias-Castro & Verzelen, 2013). Conditional computational hardness results have been obtained for the detection of submatrix and sparse principal component (Ma & Wu, 2013; Berthet & Rigollet, 2013), and it is interesting to see if similar results can be obtained for the estimation problems here.

1.4. Main Technical Contributions

We consider the general planted clustering model, which allows for a growing number of clusters and covers many existing models including planted clique, planted partition

and planted coloring.

- We obtain minimax lower bounds for planted clustering, and then prove that the computationally expensive Maximum-Likelihood Estimator achieves the lower bounds. This establishes the minimax recovery boundary for various planted models, and is the first such result for the setting with an arbitrarily growing number of clusters.
- We consider a polynomial-time algorithm based on a convex relaxation of the MLE, and obtain nearly matching sufficient and necessary conditions for the success of the algorithm. It shows that the algorithm does no achieve the statistical limit. Our performance guarantee improves upon all existing ones for polynomial-time algorithms.
- We analyze a simple algorithm based on counting node degrees and common neighbors, for which we obtain nearly matching sufficient and necessary conditions for success. Our necessary condition is the first such result for general planted clustering, and it reveals that the counting algorithm is less powerful in separating different clusters compared to the more sophisticated convexified MLE.

2. Related Work

There is a vast literature on graph clustering and their extensions. Here we focus on planted clustering and its special cases, and primarily on theoretical work that studies exact cluster recovery. Detailed comparisons are provided after each of our main theorems.

Planted Clique/Densest Subgraph: Planted Clique is the most widely studied planted model. It is known that a clique with size $K = \Omega(\sqrt{n \log n})$ can be easily identified by counting degrees (Kučera, 1995); if $K = \Omega(\sqrt{n})$, various polynomial-time algorithms work (Dekel et al., 2010); if $K = \Omega(\log n)$, an exhaustive search in super-polynomial time succeeds (Alon et al., 1998); if $K = o(\log n)$, recovery is statistically impossible. It is an open problem to find polynomial-time algorithms for the $K = o(\sqrt{n})$ regime, which is widely believed to be intractable (Juels & Peinado, 2000). The four regimes here can be considered the $r = 1$ special case of our results for general planted clustering. Extension to general values of p and q, namely planted densest subgraph, has also been considered (Arias-Castro & Verzelen, 2013).

Planted r-Cliques, Partition and Coloring: Subsequent works consider $r \geq 1$ planted cliques, and the planted partition setting (Condon & Karp, 2001) with general values of r, p and q. Existing work focus on the statistical performance of polynomial-time algorithms. The state-of-art results are given in (McSherry, 2001; Ames & Vavasis, 2014; Chen et al., 2012) for planted r-clique and in (Chen et al., 2012; Anandkumar et al., 2013) for planted partition. The $p < q$ setting is called the heterophily case, with planted coloring ($p = 0$) as an important special case (Alon & Kahale, 1997).

Converse Results for Planted Problems: Complementary to the *achievability* results above, another line of work studies *converse* results, i.e., when recovery is impossible, either by any algorithm, or by any algorithm in a specific class. The $K = \Theta(n)$ case is considered by Chaudhuri et al. (2012) and Chen et al. (2012), who establish that $p - q \gtrsim \sqrt{p/n}$ is necessary for any algorithm. For spectral clustering algorithms and convex optimization approaches, more stringent conditions are needed (Nadakuditi & Newman, 2012; Vinayak et al., 2014). We generalize and improve upon these existing converse results.

Sparse PCA: A similar gap between the statistical power of computationally expensive algorithms and known polynomial-time algorithms is observed for the sparse PCA problem. The computational hardness for detecting a single sparse principal component is proved in the seminal work (Berthet & Rigollet, 2013) conditioned on the hardness of planted clique detection.

3. Main Results

We now define the *planted clustering* problem, which has five parameters n, r, K, p and q.

Definition 1 (Planted Clustering)**.** *Suppose n nodes are divided into two subsets V_1 and V_2 with $|V_1| = rK$ and $|V_2| = n - rK$. The nodes in V_1 are partitioned into r clusters $C_1^*, \ldots, C_r^*$ (called true clusters), where $|C_m^*| = K, \forall m$. The nodes in V_2 do not belong to any clusters and are called isolated nodes. A random graph is generated as follows: for each pair of nodes and independently of all others, we connect them by an edge with probability p if they are in the same cluster, and with probability q otherwise.*

The goal is to exactly recover the true clusters $\{C_m^*\}$ given the graph. We emphasize that K, r, p and q are allowed to scale with n. We assume the values of (p, q, r, K) are known to the algorithms.

To facilitate subsequent discussion, we introduce a matrix representation of the problem. We represent the true clusters $\{C_m^*\}_{m=1}^r$ by a *cluster matrix* $Y^* \in \{0,1\}^{n \times n}$, where $Y_{ii}^* = 1$ if and only if $i \in V_1$, and $Y_{ij}^* = 1$ if and only if nodes i and j are in the same cluster. Note that the rank of Y^* equals r. The adjacency matrix of the graph is denoted as A, with the convention that $A_{ii} = 0$ for all i. Under the planted clustering model, we have $\mathbb{P}(A_{ij} = 1) = p$ if $Y_{ij}^* = 1$ and $\mathbb{P}(A_{ij} = 1) = q$ if $Y_{ij}^* = 0$ for all $i \neq j$. The problem reduces to recovering Y^* given A.

The above formulation covers many classical models.

- **Planted r-Disjoint-Clique**: Here $p = 1$ and $0 < q < 1$, so there are r cliques of size K "planted" into an Erdős-Rényi random graph $G(n, q)$. The special case with $r = 1$ is known as *planted clique*.
- **Planted Densest Subgraph**: Here $0 < q < p < 1$ and $r = 1$, so there is a dense subgraph of size K planted into a $G(n, q)$ graph.
- **Planted Partition/stochastic blockmodel**: Here $n = rK$ and $1 > p > q > 0$. The special case with $r = 2$ is called *planted bisection*.
- **Planted r-Coloring**: Here $n = rK$ and $0 = p < q < 1$, so each cluster corresponds to a group of disconnected nodes assigned with the same color.

In the next four subsections, we present our main theorems for the four regimes of the planted clustering problem. For clarity of the presentation, we shall focus on the $p > q$ setting in the sequel, as the theorems and proofs for the $p < q$ setting are very much similar. We use c_1, c_2 etc to denote universal constants independent of (n, r, K, p, q). *With high probability* (w.h.p.) means with probability at least $1 - c_1 n^{-c_2}$.

3.1. Impossible Regime: Minimax Lower Bounds

We first characterize the statistical limit of any algorithm regardless of its computational complexity. Let $\mathcal{Y}$ be the set of admissible cluster matrices, given by

$$\mathcal{Y} = \{Y | \text{there exist clusters } \{C_m\}_{m=1}^r \text{ with } |C_m| = K, \text{ and } Y \text{ is the corresponding cluster matrix}\}.$$

We use $\hat{Y} \equiv \hat{Y}(A)$ to denote an estimator which outputs an element of $\mathcal{Y}$ as an estimate of the true Y^*. We have the following lower bound on the minimax error probability of recovering Y^*.

Theorem 1. *Suppose that $8 \le K \le n/2$, $n \ge 128$ and $p > q$. If any one of the following conditions holds:*

$$4K(p-q)^2 \le q(1-q)\log(n/K), \quad (1)$$

$$12K(p-q)^2 \le \min\{p(1-p), q(1-q)\}\log(n-K), \quad (2)$$

$$16Kp \le 1, \quad (3)$$

then $\inf_{\hat{Y}} \sup_{Y^* \in \mathcal{Y}} \mathbb{P}\left[\hat{Y} \ne Y^*\right] \ge \frac{1}{2}$.

The theorem shows that it is fundamentally impossible to recover the clusters with reasonable probability in the regime where (1), (2) or (3) holds, which is thus called the *impossible regime*. If $p/q = \Theta(1)$, then the condition (2) is the least restrictive when p is bounded away from 1 and the condition (1) is the least restrictive otherwise. They imply the following impossible regimes for various standard models:

- Planted r-clique: $K(1-q) \lesssim \log(n/K)$.
- Planted r-coloring: $Kq \lesssim \log r$.
- Planted partition/densest subgraph with p bounded away from 1: $K(p-q)^2 \lesssim q \log n$.

If $q = o(p)$, then Condition (3) is the least restrictive.

Theorem 1 is proved using an information-theoretic argument. The ratio of the RHS and LHS of (1) corresponds to the ratio of the entropy of Y^* randomly chosen from $\mathcal{Y}$ and the mutual information between A and Y^*. Therefore, the impossible regime is due to an *information/statistical barrier*: the graph A does not carry enough information about the clusters Y^*.

Comparison to previous work: For $r = 1$ and q is bounded away from 1, our results recover the well-known $K = \Theta(\log n)$ threshold for planted clique; we show that the same is true for $r \to \infty$. For planted partition with $p > q$ and $p/q = \Theta(1)$, previous work (Chaudhuri et al., 2012; Chen et al., 2012) considers the $r = O(1)$ and $K = \Theta(n)$ case; our results are tighter and apply to the general setting.

3.2. Hard Regime: Optimal Algorithms

We show that the statistical limit in Theorem 1 is achieved by the Maximum Likelihood Estimator given in Algorithm 1.

Algorithm 1 Maximum Likelihood Estimator $(p > q)$

$$\hat{Y} = \arg\max_{Y \in \mathcal{Y}} \sum_{i,j} A_{ij} Y_{ij} \quad (4)$$

Enumerating over the set $\mathcal{Y}$ is computationally expensive in general since $|\mathcal{Y}| = \Omega(e^{rK})$. The following theorem provides success condition for the MLE.

Theorem 2. *Suppose $K \ge 8$ and $p > q$. With high probability, the optimal solution $\hat{Y}$ to (4) is unique and equals to Y^* provided that for some constant $c_1 > 0$*

$$K(p-q)^2 \ge c_3 p(1-q)\log n. \quad (5)$$

We refer to the regime for which the condition (5) holds but (9) fails as the *hard regime*. When $p/q = \Theta(1)$, Condition (5) reduces to $\frac{(p-q)^2}{q(1-q)} \gtrsim \frac{\log n}{K}$. This implies the following success conditions for the MLE:

- Planted r-clique: $K(1-q) \gtrsim \log n$.
- Planted r-coloring: $Kq \gtrsim \log n$.
- Planted partition/densest subgraph: $K(p-q)^2 \gtrsim q(1-q)\log n$.

If $q = o(p)$, Condition (5) reduces to $Kp \gtrsim \log n$. By comparing with Theorem 1, we see that the MLE achieves the statistical limit up to at most a log factor and is thus statistically optimal. In particular, if $p/q = \Theta(1)$ and p, q are bounded away from 1, the conditions (2) and (5) match each other up to a constant, thus establishing the minimax recovery boundary $\frac{(p-q)^2}{q(1-q)} \asymp \frac{\log n}{K}$ for planted clustering.

Comparison to previous work: Theorem 2 provides the first achievability result that is information-theoretic optimal when the number of clusters grows. It shows that for a fixed cluster size K, even if r grows, possibly at a *nearly linear* rate $r = O(n/\log n)$, MLE still succeeds under the same condition (5). When $r = p = 1$, $q = 1/2$, our result recovers the $K \asymp \log n$ boundary for planted clique (Alon et al., 1998).

3.3. Easy Regime: Polynomial-Time Methods

We present a polynomial-time algorithm that succeeds in the easy regime described in the introduction. Our algorithm is based on taking the convex relaxation of the MLE in Algorithm 1. Note that the objective function in the MLE (4) is linear, so complications come from the non-convex combinatorial constraint $Y \in \mathcal{Y}$. To obtain a computationally tractable algorithm, we replace this constraint with a convex *trace norm* constraint and a set of linear constraints. The resulting convexified MLE is given in Algorithm 2. Here the trace norm $\|Y\|_*$ (also known as the nuclear norm) is the sum of the singular values of Y. Note that the true Y^* is a feasible solution as $\|Y^*\|_* = \text{trace}(Y^*) = rK$.

Algorithm 2 Convexified Max Likelihood Estimator

$$\hat{Y} = \arg\max_{Y} \textstyle\sum_{i,j} A_{ij} Y_{ij} \quad (6)$$

$$\text{s.t. } \|Y\|_* \le rK, \quad (7)$$

$$\textstyle\sum_{i,j} Y_{ij} = rK^2,\ 0 \le Y_{ij} \le 1, \forall i,j. \quad (8)$$

The optimization problem in Algorithm 2 can be cast as a semidefinite program (SDP) and solved in polynomial time. Fast specialized algorithms have also been developed (Jalali & Srebro, 2012; Chen et al., 2012).

The following theorem provides a sufficient condition for the success of the convexified MLE.

Theorem 3 (Easy). *Suppose $p > q$. With high probability, the optimal solution to the problem (6)–(8) is unique and equals to Y^* provided*

$$(p-q)^2 K^2 \ge c_3 \left(p(1-q)K \log n + q(1-q)n\right). \quad (9)$$

Remark 1. *Theorem 3 immediately implies guarantees for other tighter convex relaxations. Define the sets $\mathcal{B} := \{Y | Eq.(8) \text{ holds}\}$ and*

$$\mathcal{S}_1 := \{Y \mid \|Y\|_* \le n_1\},\ \mathcal{S}_2 := \mathcal{S}_1 \cap \{Y \mid \|Y\|_{\max} \le 1\},$$
$$\mathcal{S}_3 := \{Y \mid Y \succeq 0; \textit{Trace}(Y) = n_1\},$$

where $\|Y\|_{\max} := \min_{K=LR^\top} \|L\|_{\infty,2}\|R\|_{\infty,2}$ is the max norm of Y and $\|\cdot\|_{\infty,2}$ is the maximum of the ℓ_2 norms of the rows. The constraint (7) is equivalent to $Y \in \mathcal{S}_1 \cap \mathcal{B}$. Observe that (Jalali & Srebro, 2012)

$$(\mathcal{S}_1 \cap \mathcal{B}) \supseteq (\mathcal{S}_2 \cap \mathcal{B}) \supseteq (\mathcal{S}_3 \cap \mathcal{B}) \supseteq \mathcal{Y}.$$

Therefore, if we replace the constraint (7) with $Y \in \mathcal{S}_2$ or $Y \in \mathcal{S}_3$, we obtain tighter *convex relaxations of the MLE. Theorem 3 immediately implies that these relaxations also recover Y^* w.h.p. under (9).*

When $r = 1$, the *easy* regime is where the condition (9) holds and (12) fails. When $r > 1$, the easy regime is where (9) holds and (13) fails. When $p/q = \Theta(1)$, the condition (9) reduces to $\frac{(p-q)^2}{q(1-q)} \gtrsim \frac{K \log n + n}{K^2}$. This implies the following success conditions for the convexified MLE under standard models:

- Planted r-clique: $K(1-q) \gtrsim \log n + \frac{n}{K}$.
- Planted r-coloring: $Kq \gtrsim \log n + \frac{n}{K}$.
- Planted partition and densest subgraph with p bounded away from 1: $K(p-q)^2 \gtrsim q(\log n + \frac{n}{K})$.

In all these cases, the smallest possible cluster size is $K = \Theta(\sqrt{n})$ and the largest possible number of clusters is $r = \Theta(\sqrt{n})$, both achieved when $p, q, |p-q| = \Theta(1)$. This generalizes the tractability threshold $K = \Omega(\sqrt{n})$ of Planted Clique to the growing r setting. In the high SNR case with $q = o(p)$, the condition (9) reduces to $Kp \gtrsim \max\{\log n, \sqrt{qn}\}$. In this case it is possible to go beyond the $K = \Omega(\sqrt{n})$ limit as the smallest possible cluster size is $K = \Theta(\max\{\log n, \sqrt{qn}\})$, achieved when $p = \Theta(1)$.

Converse for trace norm relaxation We have a converse to the achievability results in Theorem 3. The following theorem characterizes when the trace norm relaxation (6)–(8) fails.

Theorem 4 (Easy, Converse). *Suppose $p > q$. For any constant $1 > \epsilon_0 > 0$, there exist positive constants c_1, c_2 for which the following holds. Suppose $c_1 \log n \le K \le n/2$, $q \ge c_1 \log(n)/n$ and $p \le 1 - \epsilon_0$. If*

$$K^2(p-q)^2 \le c_2(Kp + qn),$$

then w.h.p. Y^* is not optimal to the program (6)–(8).

Theorems 3 and 4 together establish that under the assumptions of both theorems and ignoring logarithmic factors,

the sufficient and necessary condition for the success of the convexified MLE is

$$\frac{pK}{K^2(p-q)^2} + \frac{qn}{K^2(p-q)^2} \dot{\lesssim} 1. \tag{10}$$

Comparing this with the condition (5) for the MLE, we see that the convexified MLE is statistically sub-optimal due to the extra second term in (10). This term thus represents the statistical price of computational tractability. It has an interesting interpretation. Let $\widetilde{A} := A - q\mathbf{1}\mathbf{1}^\top + qI$ be the centered adjacency matrix. The matrix $E := (Y - \mathbf{1}\mathbf{1}^\top) \circ (\widetilde{A} - \mathbb{E}\widetilde{A})$, i.e., the deviation $\widetilde{A} - \mathbb{E}\widetilde{A}$ projected onto the cross-cluster node pairs, can be viewed as the "cross-cluster noise matrix"[2]. Note that the squared largest singular values of the matrix $\mathbb{E}\widetilde{A} = (p-q)Y^*$ is $K^2(p-q)^2$, and the squared largest singular value of E is $\Theta(qn)$ w.h.p. by standard results. Therefore, the extra second term in (10) is the "Spectral Noise-to-Signal Ratio". In fact, our proofs for Theorems 3 and 4 build on this intuition.

We note that when $K = \Theta(n)$, the conditions (5) and (9) coincide up to constant factors, and the performance of MLE and its relaxation matches. In this case the hard regime disappears.

Comparison to previous work: We refer to (Chen et al., 2012) for a survey of the statistical performance of the state-of-art polynomial-time algorithms for various planted models. Theorem 3 order-wise matches and in many cases improves upon these existing results. For planted partition, the previous best result is $(p-q)^2 \gtrsim p(K\log^4 n + n)/K^2$ in (Chen et al., 2012). Our results remove a $\log^3 n$ factor, and is also sharper for small q. For planted r-clique, existing results require $1-q$ to be $\Omega((rn+rK\log n)/K^2)$ (McSherry, 2001), $\Omega(\sqrt{n}/K)$ (Ames & Vavasis, 2014) or $\Omega((n + K\log^4)/K^2)$ (Chen et al., 2012). We improve them to $\Omega((n + K\log n)/K^2)$. Our converse result in Theorem 4 improves on the recent work by Vinayak et al. (2014). There they focus on the special case $p > 1/2 > q$, and show that a convex relaxation that is equivalent to our formulation (6)–(8) without the equality constraint in (8) fails when $K^2(p-1/2)^2 \lesssim qn$. Our result is stronger, as it applies to a tighter convex relaxation and a larger region of the parameter space.

LIMITS OF POLYNOMIAL-TIME ALGORITHMS

We conjecture that no polynomial-time algorithm has order-wised better statistical performance than the convexified MLE and succeeds beyond Condition (9).

Conjecture 1. *For any constant $\epsilon > 0$, there is no algorithm with running time polynomial in n that, for all n and with probability at least $1/2$, outputs the true Y^* of the planted clustering problem with*

$$(p-q)^2K^2 \le n^{-\epsilon}\left(Kp(1-p) + q(1-q)n\right). \tag{11}$$

If the conjecture is true, then the boundary of the easy regime will characterize the computational limit of planted clustering. This will mean there exists a significant gap between the statistical performance of intractable and polynomial-time algorithms.

A rigorous proof of Conjecture 1 seems difficult with current techniques. There are however several evidences which support the conjecture:

- The special case with $p = 1$ and $q = \frac{1}{2}$ corresponds to the hard $K = o(\sqrt{n})$ regime for Planted Clique, which is widely believed to be computationally hard, and used as an assumption for proving other hardness results (Juels & Peinado, 2000). Conjecture 1 can be considered as *a generalization of the Planted Clique conjecture* to the setting with multiple clusters and general values of p and q.
- As mentioned earlier, if (11) holds, then the spectrum of the observed graph is dominated by noise and thus fails to reveal the underlying cluster structure. The condition (11) therefore represents a "spectral barrier" for cluster recovery. A large class of algorithms that rely on the graph spectrum is proved to fail using this spectral barrier argument (Nadakuditi & Newman, 2012). The convexified MLE fails for a similar reason.
- In the sparse graph case with $p, q = O(1/n)$, Decelle et al. (2011) use non-rigorous but deep arguments from statistical physics to argue that all polynomial-time algorithms fail under (11).

3.4. Simple Regime: A Counting Algorithm

We consider a simple procedure in Algorithm 3 based on counting node degrees and common neighbors.

Algorithm 3 A Simple Counting Algorithm

1. (Identify isolated nodes) For each node i, declare it as isolated iff its degree $d_i < \frac{(p-q)K}{2} + qn$.
2. (Identify clusters when $r > 1$) Assign each pair of non-isolated nodes i, j to the same cluster iff their number of common neighbors $S_{ij} := \sum_{k\neq i,j} A_{ik}A_{jk}$ satisfies $S_{ij} > \frac{(p-q)^2K}{3} + 2Kpq + q^2(n-2K)$. Terminate if inconsistency found.

The two steps in Algorithm 3 are considered in (Kučera, 1995; Dyer & Frieze, 1989) for the special cases of finding a single planted clique or planted bisection. Let E be the set of edges. The first step runs in time $O(|E|)$, and the second

[2] $\circ$ denotes the element-wise product.

step runs in $O(n|E|)$ since each node only needs to look up its local neighborhood up to distance two to compute S_{ij}.

The following theorem provides sufficient conditions for the simple counting algorithm to succeed.

Theorem 5 (Simple). *Suppose $p > q$. W.h.p. Algorithm 3 correctly identifies the isolated nodes if*

$$K^2(p-q)^2 \geq c_3[Kp(1-q)+nq(1-q)]\log n, \quad (12)$$

and finds the clusters if further

$$K^2(p-q)^4 \geq c_4[Kp^2(1-q^2)+nq^2(1-q^2)]\log n. \quad (13)$$

When there is a single clusters $r = 1$, the *simple regime* is where the condition (12) holds; if $r > 1$, the simple regime is where both conditions (12) and (13) holds. When $p/q = \Theta(1)$, these conditions simplify to

$$r = 1 : \frac{(p-q)^2}{q(1-q)} \gtrsim \frac{n\log n}{K^2};\ r > 1 : \frac{(p-q)^2}{q(1-q)} \gtrsim \frac{\sqrt{n}\log n}{K}.$$

This implies the following success conditions for the counting algorithm under various standard models:

- Planted clique and densest subgraph: $K^2(p-q)^2 \gtrsim q(1-q)n\log n$.
- Planted r-clique ($r > 1$): $K(1-q) \gtrsim \sqrt{n}\log n$.
- Planted r-coloring ($r > 1$): $Kq \gtrsim \sqrt{n}\log n$.
- Planted partition with p bounded away from 1: $K(p-q)^2 \gtrsim q\sqrt{n}\log n$.

Comparing these conditions with (9) for the convexified MLE, we see that the counting algorithm requires an additional $\log n$ factor on the R.H.S when $r = 1$, and an additional $K\log n/\sqrt{n}$ factor when $r > 1$.

The last discussion shows that in the $r = 1$ case where the task is to separate isolated and non-isolated nodes, the counting algorithm has similar (up to a log factor) statistical performance as the more sophisticated convexified MLE, which is the best known polynomial-time algorithm. However, when $r > 1$, the convexified MLE is much more powerful. In particular, its power lies in separating different clusters, as can be seen by comparing the conditions (9) and (13).

Converse for the counting algorithm We have a partial converse to Theorem 5. The following theorem shows that the conditions (12) and (13) are also nearly necessary for the counting algorithm to succeed.

Theorem 6 (Simple, Converse). *Suppose $p > q$. For any constant $c_0 < 1$, there exist constants c_1, c_2 for which the following holds. Suppose $K \leq n/2$, $p \leq 1-c_0$ and $Kp^2 + nq^2 \geq c_1\log n$. Algorithm 3 fails to identify all the isolated nodes with probability at least $1/4$ if*

$$K^2(p-q)^2 < c_2[(Kp+nq)\log(rK) + nq\log(n-rK)], \quad (14)$$

and fails to correctly recover all the clusters with probability at least $1/4$ if

$$K^2(p-q)^4 < c_2(Kp^2+nq^2)\log(rK). \quad (15)$$

Remark 2. *Theorem 6 requires a technical condition $Kp^2 + nq^2 \geq c_1\log n$, which is not too restrictive. If $Kp^2 + nq^2 = o(\log n)$, then two nodes from the same cluster will have no common neighbor with probability $(1-p^2)^K(1-q^2)^{n-K} \geq \exp[-c(p^2K + q^2(n-K))] = \exp[-o(\log n)]$, so Algorithm 3 cannot succeed w.h.p.*

Apart from the technical condition discussed above and the assumption $p < 1 - c_0$, Theorems 5 and 6 show that the conditions (12) and (13) are sufficient and necessary for the counting algorithm. In particular, the counting algorithm is indeed strictly weaker in separating different clusters as compared to the convexified MLE. Our proof reveals that the R.H.S. of (12) and (13) are associated with the variance of the node degrees and common neighbors, respectively. If (12) does not hold, the difference between the expected degrees of isolated and non-isolated nodes will be outweighed by their deviations; a similar argument holds for the number of common neighbors. Therefore, there is an *variance barrier* that prevents the counting algorithm from succeeding outside the simple regime.

Acknowledgements: Research supported in part by NSF ECCS 10-28464.

References

Alon, N. and Kahale, N. A spectral technique for coloring random 3-colorable graphs. *SIAM Journal on Computing*, 26(6):1733–1748, 1997.

Alon, N., Krivelevich, M., and Sudakov, B. Finding a large hidden clique in a random graph. *Random Structures and Algorithms*, 13(3-4):457–466, 1998.

Ames, B. P.W. and Vavasis, S. A. Convex optimization for the planted k-disjoint-clique problem. *Mathematical Programming*, 143(1-2), 2014.

Anandkumar, A., Ge, R., Hsu, D., and Kakade, S. A tensor spectral approach to learning mixed membership community models. *arXiv:1302.2684*, 2013.

Arias-Castro, E. and Verzelen, N. Community detection in random networks. *arXiv:1302.7099*, 2013.

Arias-Castro, E., Candès, E. J, and Durand, A. Detection of an anomalous cluster in a network. *The Annals of Statistics*, 39(1):278–304, 2011.

Balakrishnan, S., Kolar, M., Rinaldo, A., Singh, A., and Wasserman, L. Statistical and computational tradeoffs in biclustering. In *NIPS 2011 Workshop on Computational Trade-offs in Statistical Learning*.

Berthet, Q. and Rigollet, P. Complexity theoretic lower bounds for sparse principal component detection. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 30:1–21, 2013.

Bollobás, B. and Scott, AD. Max cut for random graphs with a planted partition. *Combinatorics, Probability and Computing*, 13(4-5):451–474, 2004.

Chandrasekaran, V. and Jordan, M. I. Computational and statistical tradeoffs via convex relaxation. *PNAS*, 110 (13):E1181–E1190, 2013.

Chaudhuri, K., Graham, F. C., and Tsiatas, A. Spectral clustering of graphs with general degrees in the extended planted partition model. *JMLR*, 23, 2012.

Chen, Y., Sanghavi, S., and Xu, H. Clustering sparse graphs. *arXiv:1210.3335*, 2012.

Condon, A. and Karp, R. M. Algorithms for graph partitioning on the planted partition model. *Random Struct. Algorithms*, 18(2), 2001.

Decelle, A., Krzakala, F., Moore, C., and Zdeborova, L. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physics Review E*, 84:066106, 2011.

Dekel, Y., Gurel-Gurevich, O., and Peres, Y. Finding hidden cliques in linear time with high probability. *arxiv:1010.2997*, 2010.

Dyer, M.E and Frieze, A.M. The solution of some random NP-hard problems in polynomial expected time. *Journal of Algorithms*, 10(4):451 – 489, 1989.

Fortunato, S. Community detection in graphs. *arXiv:0906.0612*, 2010.

Holland, P. W., Laskey, K. B., and Leinhardt, S. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.

Jalali, A. and Srebro, N. Clustering using max-norm constrained optimization. *ICML*, 2012.

Juels, A. and Peinado, M. Hiding cliques for cryptographic security. *Designs, Codes & Crypto.*, 2000.

Kolar, M., Balakrishnan, S., Rinaldo, A., and Singh, A. Minimax localization of structural information in large noisy matrices. In *NIPS*, 2011.

Krauthgamer, R., Nadler, B., and Vilenchik, D. Do semidefinite relaxations really solve sparse PCA? *arXiv:1306.3690*, 2013.

Kučera, L. Expected complexity of graph partitioning problems. *Discrete Appl. Math.*, 57(2-3), 1995.

Leskovec, J., Lang, K., Dasgupta, A., and Mahoney, M. Statistical properties of community structure in large social and information networks. In *WWW*, 2008.

Ma, Z. and Wu, Y. Computational barriers in minimax submatrix detection. *arXiv:1309.5914*, 2013.

McSherry, F. Spectral partitioning of random graphs. In *FOCS*, pp. 529 – 537, 2001.

Nadakuditi, R. R. and Newman, M.E.J. Graph spectra and the detectability of community structure in networks. *Physical Review Letters*, 108(18), 2012.

Newman, M. E. J. and Girvan, M. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69: 026113, Feb 2004.

Rohe, K., Chatterjee, S., and Yu, B. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4), 2011.

Vinayak, R. K., Oymak, S., and Hassibi, B. Sharp performance bounds for graph clustering via convex optimization. In *ICASSP*, 2014.