Abstract

The planted models assume that a graph is generated from a set of 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 and planted coloring. This paper studies the statistical-computational tradeoffs of these models. Our focus is the high-dimensional setting, where the number of clusters is allowed to grow with the number of nodes. We show that the complexities of cluster recovery exhibit phase transitions. In particular, the space of model parameters can be partitioned into four regions with decreasing statistical and computational complexities: (1) the impossible regime, where all algorithms fail; (2) the hard regime, where the exponential-time Maximum Likelihood Estimator (MLE) succeeds; (3) the easy regime, where a polynomial-time convexified MLE succeeds; (4) the simple regime, where a simple algorithm based on counting degrees and common neighbors succeeds. Moreover, each of these algorithms is likely to fail in the harder regime.

1 Introduction

We consider the following planted clustering problem, defined by five parameters $n, n_1, K, p$ and $q$.

**Definition 1 (Planted Clustering).** Suppose $n$ nodes are divided into two subsets $V_1$ and $V_2$ with $|V_1| = n_1$ and $|V_2| = n - n_1$. The nodes in $V_1$ are partitioned into $r = n_1/K$ disjoint clusters $C_1^*, \ldots, C_r^*$ (called true clusters), where $|C_m^*| = K$ for all $1 \leq m \leq r$ and $\bigcup_{m=1}^{r} C_m^* = V_1$. Nodes in $V_2$ do not belong to any clusters and are called the isolated nodes. A random graph is generated based on the cluster structure: 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 $\{C_m^*\}_{m=1}^{r}$ given the random graph.

This general form of the models generalizes several classical planted models, including Planted Clique and $r$-Clique, Planted Coloring, Planted Partition and Stochastic Blockmodel [1][2][3]. These models are widely used and studied, and are important in both theory and practice.

Cluster recovery under the planted models are both a statistical problem and a computational problem. Statistically, it is an inference problem where one is given as data a graph generated stochastically, and would like to infer the underlying structures. The problem becomes noisier and statistically harder with a small $p$, a large $q$, a smaller $K$, a large $r$ and a small $n_1$ (all allowed to scale with $n$). Computationally, the problem has a discrete and exponentially large solution space containing all possible ways of assigning $n$ nodes into $r$ clusters. An algorithm exhaustively searching for the best-fitting clustering will have a time-complexity growing exponentially fast with $n$ and $r$.
A deeper understanding of the problem requires one to study its statistical and computational aspects jointly. It is important to understand the tradeoffs between the two aspects: How do algorithms with different computational complexity achieve different statistical performance, and what are the statistical and computational limits of the problem. We focus particularly on the so-called high-dimensional regime [4], where $r$ is allowed to grow unbounded with $n$.

Our results highlight the following: As the model parameters change, the statistical and computational complexities of the recovery problem exhibit phase transitions. In particular, the parameter space can be partitioned into four regions; each region corresponds to statistically easier instances of the planted models than the previous region, and recovery can be achieved by simpler algorithms with lower computational requirements.

### 1.1 The Four Regimes

We first consider the setting with $p > q$ and $p/q = Θ(1)$. This covers the standard planted partition and planted $r$-clique models. The statistical hardness of cluster recovery can be summarized by the quantity $(p−q)^2/p(1−q)$. Our main theorems identify four regimes defined by the values of this quantity.

- **The Impossible Regime**: $(p−q)^2/p(1−q) ≳ 1/K^2$. There is no algorithm, regardless of its computational complexity, that can recover the clusters with reasonable probability in this regime.

- **The Hard Regime**: $1/K ≲ (p−q)^2/p(1−q) ≲ n^p/(p^2)$. There exists an exponential-time algorithm – specifically the Maximum Likelihood Estimator (MLE) – that recovers the clusters in this regime (as well as in the next two easier regimes; we omit such implications in the sequel).

- **The Easy Regime**: $n/K^2 ≳ (p−q)^2/p(1−q) ≲ \sqrt{n}/K$. There exists a polynomial-time algorithm – a convex relaxation of MLE – that recovers the clusters with high probability in this regime.

- **The Simple Regime**: $(p−q)^2/p(1−q) ≳ \sqrt{n}/K$. A simple algorithm based on counting degrees and common neighbors recovers the clusters with high probability in this regime.

We illustrate these four regimes in Fig. 1 assuming $p = 2q = Θ(n^{-α})$ and $K = Θ(n^β)$. Here cluster recovery becomes 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. This means using a computationally more complicated algorithm leads to a significant (order-wise) enhancement in statistical power.

Importantly, the transitions between the four regimes are likely to be tight. The MLE is statistically optimal as it achieves the statistical limit that no algorithm can break. The impossible regime is separated from the hard regime by an information barrier: there the graph does not carry enough information about the clusters. Similarly, we prove that the counting algorithm fails outside the simple regime, due to a variance barrier associated with the node degrees and the numbers of common neighbors.

Between the easy and hard regimes, we conjecture the following: no polynomial-time algorithm succeeds in the hard regime, i.e., the convexified MLE achieves the computational limit. While proving this conjecture is considered hard, there are evidences supporting it. The hard regime contains the planted clique problem with clique size $K = o(\sqrt{n})$, which is widely believed to be intractable. Moreover, there is a spectral barrier, determined by the spectral norm of a noise matrix, that prevents spectral algorithms, and possibly all polynomial-time algorithms as well, from succeeding in the hard regime.

**General results** Our main theorems apply to general values of $p$, $q$, $K$, $r$ and $n_1$. We summarize some of these results in Table 1 (cf. Section 1.2 for definitions of the models).

---

1Here $\gtrapprox$ and $\lesssim$ ignore constant and $\log n$ factors.
Table 1: The four regimes for the planted partition, coloring and r-clique models

<table>
<thead>
<tr>
<th>Simple</th>
<th>Planted Partition</th>
<th>Planted Coloring</th>
<th>Planted r-Clique</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{(p-q)^r}{p} \gtrsim \frac{\sqrt{n} \log n}{n}$</td>
<td>$q \gtrsim \frac{\sqrt{n} \log n}{n}$</td>
<td>$K \gtrsim \sqrt{n} \log n$</td>
<td></td>
</tr>
<tr>
<td>Easy</td>
<td>$\frac{n+K \log n}{K^2} \lesssim \frac{(p-q)^r}{p} \lesssim \frac{\sqrt{n}}{K}$</td>
<td>$\frac{n+K \log n}{K^2} \lesssim q \lesssim \frac{\sqrt{n}}{K}$</td>
<td>$\sqrt{n} \lesssim K \lesssim n \log n$</td>
</tr>
<tr>
<td>Hard</td>
<td>$\frac{\log n}{K} \lesssim \frac{(p-q)^r}{p} \lesssim \frac{n+K \log n}{K^2}$</td>
<td>$\frac{\log n}{K} \lesssim q \lesssim \frac{n+K \log n}{K^2}$</td>
<td>$\log n \lesssim K \lesssim \sqrt{n}$</td>
</tr>
<tr>
<td>Impossible</td>
<td>$\frac{(p-q)^r}{p} \lesssim \frac{\log n}{K}$</td>
<td>$q \lesssim \frac{\log n}{K}$</td>
<td>$K \lesssim \log n$</td>
</tr>
</tbody>
</table>

Summary These results in this paper highlight the interaction between the statistical and computational aspects in the planted models. Our theorems characterize how the model parameters govern the hardness of the problem, and we demonstrate a series algorithms from a statistically powerful but computationally expensive one to a simple and less powerful one.

Our work parallels a recent line of work that takes a joint statistical-computational view on inference problems [5-6]. While we focus on the specific problem of planted clustering, this model is not just a random example. To the best of our knowledge, existing work on proving average-case computational complexity is based on the hardness assumption of planted clique (e.g. the seminal work [5] on Sparse PCA). The planted model is thus in certain sense archetypical – it has rich structures that capture many aspects of the problem of computational-statistical tradeoff, and is thus used as a starting point for studying the problem. Our work extends previous ones that focus on the rank-1 case (e.g. a single clique), and we expect that the principles in this paper are relevant more broadly.

Our results also demonstrate new phenomena not observed in previous work on Sparse PCA and submatrix detection. In these problems, the statistical limit of polynomial-time algorithms is usually achieved by naive algorithms based on simple thresholding/averaging, and more algorithms do not provide additional statistical power. In contrast, we show that for planted clustering, convex optimization is significantly more powerful than the simple counting algorithm. This points to the existence of a finer spectrum of statistical power among all polynomial-time algorithms.

1.2 Related Work

Planted Clique For the planted clique model ($r = 1, p = 1, q = 1/2$), it is known that if $K = \Omega(\sqrt{n} \log n)$, the clique nodes will have the highest degrees and can be easily identified; if $K = \Omega(\sqrt{n})$, various polynomial-time algorithms work [8]; if $K = \Omega(\log n)$, an exhaustive search using super-polynomial time succeeds [8]; if $K = o(\log n)$, recovery is fundamentally impossible. It is an open problem to find polynomial-time algorithms for the $K = o(\sqrt{n})$ regime, which is widely considered intractable [9]. See [10] for extensions. The four regimes above can be considered as the $r = 1$ special case of our results for the general planted clustering model.

Planted r-Cliques, Partition, and Coloring Subsequent works consider the planted clique case with $r \geq 1$ cliques, and the planted partition setting with general values of $p$ and $q$ (aka stochastic blockmodel). The special case with $r = 2$ is known as planted bisection. Existing work focus on the statistical performance of polynomial-time algorithms. The state-of-art results are given in [11][12][13] for planted r-clique, and in [12][13][14] for planted partition. The setting with $p < q$ includes planted noisy coloring or planted $r$-cut [13]. The most important special case is the planted coloring model ($p = 0$) [14]. It is previous unknown that if any super-polynomial time algorithm performs better statistically then known polynomial-time algorithms. This paper provides an affirmative answer.

Complementary to achievability results that prove when an algorithm succeeds, another line of work considers converse results, i.e., when recovery is impossible by any algorithm in a class [2][12][13].

1.3 Preliminaries

We represent the true clusters $\{C_i\}_{i=1}^m$ by a cluster matrix $Y^* \in \{0, 1\}^{n \times n}$, where $Y^*_i = 1$ for $i \in V_1$, $Y^*_i = 0$ for $i \in V_2$, and $Y^*_{ij} = 1$ if and only if nodes $i$ and $j$ are in the same cluster. Let the adjacency matrix of the graph be $A$. Under the planted clustering model, we have $P(A_{ij} = 1) = p$ if $Y^*_{ij} = 1$ and $P(A_{ij} = 1) = q$ if $Y^*_{ij} = 0$ for all $i \neq j$. The problem reduces to recovering $Y^*$
given $A$. We assume that the values of $p$, $q$, $n$, $r$ and $K$ are known. We use $c_1$, $c_2$ etc. to denote absolute constants. With high probability (w.h.p.) means with probability at least $1 - c_1 n^{-c_2}$.

2 The Impossible Regime

We characterize for what values of $p$, $q$, $n$, $r$ and $K$ all algorithms fail. We assume that the true cluster matrix $Y^*$ is sampled uniformly from $\mathcal{Y}$, the set of admissible cluster matrices:

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

**Theorem 1.** Any algorithm fails to output the true clusters $Y^*$ with probability more than $1/2$, if

$$K(p - q)^2 \leq \frac{1}{4} \beta (1 - \beta) \log \frac{n}{K},$$

where $\alpha := \frac{n_1(K-1)}{n(n-1)}$ and $\beta := \alpha p + (1 - \alpha)q$.

The theorem shows that it is statistically impossible to recover the clusters with reasonable probability in the regime for which (1) holds, hence the name the impossible regime. See Table 1 for special cases of the impossible regime for various planted problems.

Our analysis reveals that the RHS of (1) corresponds to the entropy of $Y^*$ randomly chosen from $\mathcal{Y}$, and the LHS corresponds to the mutual information between $A$ and $Y^*$. Therefore, the impossible regime is a consequence of the information (statistical) barrier: the observation $A$ does not carry enough information to distinguish between different possible values for $Y^*$.

3 The Hard Regime and Optimal Algorithms

We show that the statistical limit in Theorem 1 is achievable. One such algorithm is the Maximum Likelihood Estimator. For any clustering $Y \in \mathcal{Y}$, the log likelihood $\log P(A|Y)$ of observing the graph $A$ given the clusters $Y$ is

$$C_1(p,q)\sum_{i<j} A_{ij}Y_{ij} + C_2(p,q,r,K,A),$$

where $C_i(\cdot)$ collects the terms that are independent of $Y$. Given $A$, the MLE (Algorithm 1) maximizes the above log likelihood function over the set $\mathcal{Y}$. Enumerating all elements in the set $\mathcal{Y}$ takes exponential-time.

**Algorithm 1 Maximum Likelihood Estimator**

$$\hat{Y} = \arg\max_Y \sum_{i,j} A_{ij}Y_{ij} \quad \text{s.t. } Y \in \mathcal{Y}. \quad (2) \quad (3)$$

**Theorem 2.** With high probability, $\hat{Y} = Y^*$ is the unique optimum of the problem (2)–(3) provided

$$(p - q)^2 K \geq c_3 \max\{p(1-q), q(1-p)\} \log n. \quad (4)$$

We refer to the regime for which the condition (4) holds but (1) fails as the hard regime, since cluster recovery is statistically possible but computationally hard. When $p > q$, the condition (4) simplifies to

$$(p - q)^2 \geq \frac{\log \frac{n}{K}}{p(1-q)}.$$ 

By comparing with the results in Section 2, we conclude that Theorems 1 and 2 are both tight (up to a log factor), and the condition (4) is sufficient and necessary for cluster recovery. Therefore, the MLE is an statistically optimal algorithm for the planted clustering model.

When there are a growing number of cliques/clusters ($r = \omega(1)$), Theorem 2 provides the first achievability result that matches the information-theoretic limit. In particular, it shows that even if $r$ grows, possibly at a near-linear rate $r = O(n/\log n)$, MLE still recovers all the clusters.

4 The Easy Regime and Polynomial-Time Algorithms

In this section, we present a tractable polynomial-time algorithm, which is based on relaxing the MLE in Algorithm 1. Note that the objective (2) is linear, so the complexity of the MLE is due to
With high probability, Algorithm 3 correctly finds the isolated nodes and the clusters if Theorem 4.
The algorithm has been considered for the special cases of planted clique and planted bisection.
We consider a simple algorithm (Algorithm 3) based on counting degrees and common neighbors.

### 5 The Simple Regime and A Counting Algorithm

We consider a simple algorithm (Algorithm 3) based on counting degrees and common neighbors. The algorithm has been considered for the special cases of planted clique and planted bisection.

**Theorem 4.** With high probability, Algorithm 3 correctly finds the isolated nodes and the clusters if

\[
K^2(p-q)^2 \geq c_3 \left[ K \max \{p(1-q), q(1-p)\} + q(n-1) \right] n \log n, \tag{9}
\]

\[
K^2(p-q)^2 \geq c_4 \left[ K \max \{p^2(1-q^2), q^2(1-p^2)\} + q^2(n-1) \right] n \log n. \tag{10}
\]
**Algorithm 3** A Simple Counting Algorithm

1. For each node $i$, compute its degree $d_i$. Declare $i$ as isolated if $d_i < \frac{(p-q)K}{2} + qn$. Remove all isolated nodes.
2. (a) Among the remaining nodes, each node $i$ computes the number of common neighbors $S_{ij} := \sum_{k \neq i,j} A_{ik} A_{jk}$ for all nodes $j \neq i$.
(b) Assign into the same cluster nodes $i, j$ with $S_{ij} > \frac{(p-q)^2K}{4} + 2Kpq + q^2(n - 2K)$. Declare error if inconsistency is found.

We refer to the regime for which the conditions (9) and (10) hold as the simple regime. When $1 - c_0 \geq p > q$ and $p/q = \Theta(1)$, these conditions simplify to $\frac{(p-q)^2K}{4} \geq \sqrt{\frac{\pi \log n}{K}}$. Compared with (7) for the convexified MLE, the counting algorithm requires an additional $K \log n/\sqrt{n}$ on the RHS.

**Limits of the counting algorithm** We have a converse to Theorem 4.

**Theorem 5.** Assume $1 - c_3 \geq p > q$, $c_1 \leq \frac{p}{q} \leq c_2$ and $Kp^2 + nq^2 \geq c_4 \log n$. With positive probability, Algorithm 3 fails to identify the isolated nodes and clusters if

$$K^2(p-q)^4 < c_6 (Kp^2 + nq^2) \log n.$$ 

Apart from some mild technical conditions, Theorems 4 and 5 show that the conditions (9) and (10) are both sufficient and necessary for the counting algorithm, so the simple regime characterizes the limit of the counting algorithm. Our analysis reveals that the RHS of (9) and (10) correspond to the variance of the node degrees and the number of common neighbors, respectively. Therefore, there is a “variance barrier” that prevents the counting algorithm from succeeding beyond the easy regime.

**References**