Does robustness imply tractability? A lower bound for planted clique in the semi-random model

Jacob Steinhardt[*]
Stanford University
`jsteinhardt@cs.stanford.edu`

Abstract

We consider a robust analog of the planted clique problem. In this analog, a set S of vertices is chosen and all edges in S are included; then, edges between S and the rest of the graph are included with probability $\frac{1}{2}$, while edges not touching S are allowed to vary arbitrarily. For this *semi-random* model, we show that the information-theoretic threshold for recovery is $\tilde{\Theta}(\sqrt{n})$, in sharp contrast to the classical information-theoretic threshold of $\Theta(\log(n))$. This matches the conjectured computational threshold for the classical planted clique problem, and thus raises the intriguing possibility that, once we require robustness, there is no computational-statistical gap for planted clique.

---

1 Introduction

The planted clique problem is perhaps the most famous example of a *computational-statistical gap* — while it is information-theoretically possible to recover planted cliques even of size $2\log_2(n)$, the best efficient recovery algorithms require the clique to have size $\Omega(\sqrt{n})$. It has long been conjectured that no polynomial-time algorithm can find cliques of size $n^{1/2-\epsilon}$, with recent breakthrough work by Barak et al. (2016) establishing this for the class of sum-of-squares algorithms. There thus appears to be an exponential gap between what is possible statistically and what is tractable computationally.

In this paper we revisit this gap, and question whether recovering cliques of size $k \ll \sqrt{n}$ is actually meaningful, or if it can only be done by over-exploiting the particular details of the planted clique model. Recall that in the planted clique model, a set S of vertices is chosen at random and all vertices in S are connected; the remaining edges are then each included independently with probability $\frac{1}{2}$. While this model is convenient in its simplicity, it also has a number of peculiarities — for instance, simply returning the highest-degree nodes already performs nearly as well at recovering S as sophisticated spectral algorithms.

Feige and Kilian (2001) argue that it is more realistic to consider a *semi-random* model, in which edges that do not touch any vertices in the clique are allowed to vary arbitrarily. This forces recovery algorithms to be more robust by not relying on simple heuristics such as maximum degree to identify the planted clique. It is then natural to ask — once we require such robustness, how large must a clique be to be identifiable? To this end, we establish a strong information-theoretic lower bound:

Theorem. *In the semi-random model, it is information-theoretically impossible to even approximately recover planted cliques of size $o(\sqrt{n})$. Moreover, it is information-theoretically possible to exactly recover cliques of size $\omega(\sqrt{n\log(n)})$.*

It is striking that the information-theoretic threshold in the semi-random model essentially matches the computational threshold of $\sqrt{n}$ in the standard model. We conjecture that, in the semi-random model, the computational and statistical thresholds in fact coincide — i.e., there is an efficient algorithm for recovering cliques of size $\sqrt{n\log(n)}$. In such a case, the previous exponential gap between the statistical and computational limits would vanish entirely.

The Model

We now explain and justify the semi-random model in more detail. We consider a graph on n nodes with a planted clique of size k. Label the nodes $1,\dots,n$, and let $S \subseteq \{1,\dots,n\}$ be the set of vertices in the clique. We represent the graph by its adjacency matrix $A \in \{0,1\}^{n\times n}$, which must be symmetric and satisfy $A_{ii} = 0$ for all i. Beyond these constraints, A is generated as follows:

$$A_{ij} = \left\{ \begin{array}{ccl} 1 & : & i,j \in S \\ \mathrm{Ber}(\frac{1}{2}) & : & i \in S, j \notin S \text{ or } j \in S, i \notin S \\ \text{arbitrary} & : & \text{else} \end{array} \right. \tag{1}$$

Here $\mathrm{Ber}(p)$ denotes the Bernoulli distribution with parameter p (i.e., a coin toss with probability p of being 1). In words, A is generated by planting a clique S and connecting all pairs of vertices in S; then connecting each clique vertex to each non-clique vertex independently with probability $\frac{1}{2}$; and finally filling in edges between non-clique vertices arbitrarily. This is essentially the model considered by Feige and Kilian (2001), except they additionally allow any number of edges between S and $[n]\backslash S$ to be deleted. Note that lower bounds in our model imply lower bounds in Feige and Kilian's model.

The motivation for this model comes from thinking of planted clique recovery as a community detection problem, where there is a community of vertices in a large random graph which are connected to each other with high probability (in this case, probability 1). It would be strange if whether a set of vertices S is considered a community could be affected by edges having nothing to do with S. The semi-random model above essentially asserts this.

From another perspective, even if we don't believe that the edges outside of S are truly arbitrary, it also seems unlikely that they are completely random (as in the classical planted clique problem) or a disjoint union of other planted cliques (as in the stochastic block model). It would be ideal to design recovery algorithms that are not tied to the specifics of these toy models, and allowing arbitrary edge connections outside of the planted structure is one way of getting at this.

Recovery criterion. Because edges not touching S can be chosen arbitrarily, the set S need not be the unique clique in the graph, even when $k \gg \log(n)$. Indeed, even if $k = \frac{n}{2}$, we could observe a graph with two cliques of size $\frac{n}{2}$ (with all edges between them occurring with probability $\frac{1}{2}$) and it would be impossible to tell which of the two was the planted clique S.

To overcome this ambiguity, one is given a vertex v sampled uniformly from S, and asked to output an estimate $\hat{S}$ of the entire set S. Intuitively, given a vertex that belongs to S, one should be able to tell which of its neighbors are also part of S, and which are due to random chance; this is a baseline pre-requisite for more difficult tasks such as predicting the probability of occurrence of a held-out edge.

Our error metric is $|S \triangle \hat{S}|$, the symmetric difference between S and $\hat{S}$. We will show that when $k \ll \sqrt{n}$, even partial recovery is impossible under this metric; in fact, we will show that it is impossible to distinguish between the case where a subset of k vertices belongs to a unique planted clique, and the case where every vertex belongs to exactly two planted cliques. In contrast, when $k \gg \sqrt{n \log(n)}$, *exact* recovery is possible—$|S \triangle \hat{S}| = 0$ with high probability.

Results

Our main result is an information-theoretic lower bound on recovery in the semi-random model:

Theorem 1.1. *Fix $\delta < 1$. Then for sufficiently large n, there is a distribution over instances of the semi-random planted clique model, each with planted clique size at least $k = \frac{1}{6} \delta^{1/3} \sqrt{n}$, such that given the adjacency matrix A and v sampled uniformly from S, any candidate $\hat{S}(A, v)$ for S must satisfy*

$$\mathbb{E}_{A,v}\left[|S \triangle \hat{S}(A, v)|\right] \geq (1 - \delta)\mathbb{E}_A[|S|]. \tag{2}$$

Note that we can trivially obtain symmetric difference $|S| - 1$ by taking $\hat{S}(A, v) = \{v\}$. Theorem 1.1 thus says that, when $k \ll \sqrt{n}$, it is impossible to improve upon this trivial bound by even a small amount.

The proof idea is to construct a distribution P_0 over matrices A where every vertex belongs to exactly two identically distributed cliques, and show that it has small total variational distance to an instance P_1 of the semi-random model. Then, given v, it is impossible to tell which of the two cliques is S, and so one cannot do better than simply outputting $\{v\}$ (note that guessing one of the cliques at random would lead to $|S \triangle \hat{S}| = 2|S| - 2$ half the time, and 0 the other half, which is still $|S| - 1$ on average).

A key challenge is that P_0 has a number of dependencies in it which do not exist in P_1. To address this, we establish a technical lemma – Lemma 2.1 – which shows that these dependencies are "sufficiently hidden" by the randomness in the edges between S and $[n]\backslash S$.

In addition to our lower bound, we establish an upper bound showing that exact recovery is information-theoretically possible for sets of size $\sqrt{n \log(n)}$, and computationally possible for sets of size $n^{2/3} \log^{1/3}(n)$:

Theorem 1.2. *Under the semi-random model, for a planted clique of size $k = \omega(\sqrt{n \log(n)})$, exact recovery is possible with high probability— there is an $\hat{S}(A, v)$ such that $\mathbb{P}_{A,v}[S = \hat{S}(A, v)] = 1 - o(1)$.*

Moreover, there exists a constant C such that for $k \geq Cn^{2/3} \log^{1/3}(n)$, efficient exact recovery is possible—there is a polynomial-time algorithm $\hat{S}(A, v)$ such that $\mathbb{P}_{A,v}[S = \hat{S}(A, v)] = 1 - o(1)$.

The information-theoretic bound uses the fact that the true planted clique must have small intersection with any other large clique, and holds in Feige and Kilian's model as well. The computational bound is essentially Corollary 9.3 of Charikar et al. (2017).

Open Problems

The results in this paper show that in the semi-random model, the information-theoretic threshold for planted clique is between $\sqrt{n}$ and $\sqrt{n \log(n)}$, while the computational threshold is at most $\tilde{\mathcal{O}}(n^{2/3})$. We believe that both thresholds should be exactly on the order of $\sqrt{n \log(n)}$:

Conjecture 1.3. *If $k = o(\sqrt{n \log(n)})$, no algorithm can even partially recover S in the semi-random model.*

Conjecture 1.4. *If $k = \omega(\sqrt{n \log(n)})$, there is an algorithm that exactly recovers S in polynomial time in the semi-random model.*

In our lower bound construction below, we consider distributions over cliques that overlap in a single element. Proving Conjecture 1.3 would require considering cliques that overlap in $\log(n)$ elements, which seems possible but more challenging.

We can generalize the planted clique model to the more general *planted dense subgraph* model, in which edges within S are formed with probability p, and edges between S and $[n] \backslash S$ are formed with probability q. The planted clique model corresponds to $p = 1$, $q = \frac{1}{2}$. We can then ask for the information-theoretic recovery threshold in this more general model, as a function of p and q.

Conjecture 1.5. *The information-theoretic threshold for recovery in the semi-random planted dense subgraph model is $\tilde{\Theta}(\sqrt{\frac{n(p+q)}{(p-q)^2}})$.*

The threshold in Conjecture 1.5 is the *Kesten-Stigum threshold* (Kesten and Stigum, 1966b;a), which is believed to be the threshold for efficient recoverability in the stochastic block model (Decelle et al., 2011). As semi-random dense subgraph is a natural robust analog of the stochastic block model, a proof of Conjecture 1.5 would be another data point suggesting a relationship between robustness and computation. We remark that at least the upper bound can be established with similar ideas to Theorem 1.2.

Related Work

The type of semi-random model studied here was first proposed by Feige and Kilian (2001), who provide algorithms for finding maximum independent sets and k-colorings that were later extended by Coja-Oghlan (2007). While maximum independent set is equivalent to maximum clique, the regime they consider corresponds to cliques of size $\Theta(n)$, with connection probabilities of $1 - \tilde{\mathcal{O}}(1/n)$ between the clique and non-clique vertices; their results are thus inapplicable in our setting. In another direction, Makarychev et al. (2012) proposed a similar (but more general) model for semi-random graph partitioning, and provided algorithms for a number of problems including sparsest cut and small-set expansion.

Another popular semi-random model is the monotone adversaries model introduced by Blum and Spencer (1995) and since considered by a number of authors (Feige and Krauthgamer, 2000; Coja-Oghlan, 2004;

Chen et al., 2014; Guédon and Vershynin, 2014; Moitra et al., 2015; Agarwal et al., 2015). In this model, all non-clique edges are generated with probability $\frac{1}{2}$, and then an adversary can remove any of the non-clique edges. The monotone adversaries model rules out interesting phenomena such as the possibility of other cliques appearing in the graph, and is thus arguably too restrictive. Moitra et al. (2015) study information-theoretic thresholds for planted recovery problems in this model, but show only a constant factor gap between the robust and non-robust thresholds, in contrast to the exponential $\log(n)$ vs. $\sqrt{n}$ gap shown here.

Finally, there has recently been work on robust learning when a large fraction of the data can be arbitrarily corrupted. We can embed our model in this model by treating each row of the adjacency matrix as a data point, so that the k rows from the clique are "clean" while the remaining $n - k$ are corrupted. This setting, where the majority of points are corrupted, was studied by Steinhardt et al. (2016); Charikar et al. (2017); and Steinhardt et al. (2017). In particular, Charikar et al. (2017) provide a polynomial-time algorithm which translates to a $\tilde{\mathcal{O}}(n^{2/3})$ upper bound for planted clique, while Steinhardt et al. (2017) provide an information-theoretic upper bound criterion for a general class of problems.

2 Lower Bound

In this section we will prove Theorem 1.1. The idea is to construct a distribution P_0 which does not lie in the semi-random model, and from which S cannot be recovered, and show that it is close to a distribution P_1 lying in the semi-random model.

The base distribution P_0. We will start by describing the distribution P_0, which is parameterized by integers n and m; it will induce cliques of size roughly $\frac{n}{m}$. It is constructed so that samples from P_0 have the following properties:

- The graph contains $2m$ cliques, each of size approximately $\frac{n}{m}$.

- Every vertex lies in two cliques.

- Every pair of cliques is either disjoint or has intersection 1.

We do this as follows: split the cliques into two groups each of size m. For each vertex $i \in [n]$, sample (without replacement) a pair $(a_i, b_i) \in [m] \times [m]$. Vertex i will belong to the a_ith clique in the first group, and the b_ith clique in the second group. Note that sampling without replacement ensures that no two cliques intersect in more than one element.

Given $a_{1:n}$ and $b_{1:n}$, we generate the adjacency matrix A for the graph as follows:

$$A_{ij} = \left\{ \begin{array}{ccc} 1 & : & a_i = a_j \text{ or } b_i = b_j, \\ \mathrm{Ber}(q) & : & \text{else.} \end{array} \right. \tag{3}$$

Here q will be chosen to be $\frac{1}{2} - \frac{1}{2m-2}$; it is slightly less than $\frac{1}{2}$ in order to correct for the extra edges created by the two cliques, so that the expected degree is the same as in the semi-random model.

The semi-random instance P_1. We want to construct a semi-random instance P_1 that is close to P_0. We will do this via a coupling argument: we express a sample from P_0 as a sequence of local decisions, and show that each decision can either be exactly imitated under P_1, or approximately imitated with small KL divergence.

First let us re-express P_0 in a way that makes it look more like a planted clique instance, where we will think of the clique corresponding to $a_i = 1$ as the "planted" clique S. Let $\mathrm{HG}(n, K, N)$ denote the hypergeometric distribution, which samples n items without replacement from $[N]$, and counts the number of sampled items lying in $\{1, \ldots, K\}$. We can generate $A \sim P_0$ sequentially as follows:
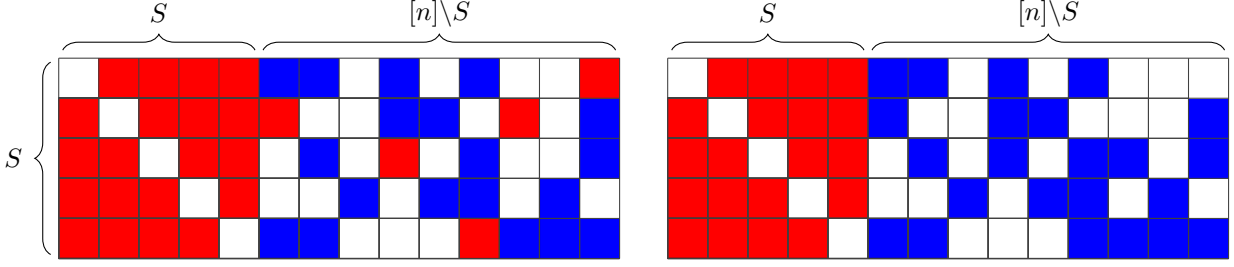
Figure 1: Illustration of step 2 of P_0 (left) vs. step 2' of P_1 (right). Each grid is the top $s \times n$ portion of the adjacency matrix. Red edges are from cliques while blue edges are random. In the left matrix, $b_6 = 2$, $b_9 = 3$, $b_{11} = 5$, $b_{12} = 2$, and $b_{14} = 1$ (the remaining elements of $b_{6:14}$ are all larger than 5). The extra red edges introduce anti-correlations in P_0 which could potentially distinguish it from P_1.

1. Sample $s \sim \mathrm{HG}(n, m, m^2)$ and let S be a uniformly random subset of $[n]$ of size s. Set $a_i = 1$ for each $i \in S$ and sample the corresponding b_i from $[m]$ without replacement. Connect all vertices in S.

2. For each $i \notin S$, determine which edges should exist between S and $[n] \backslash S$. This involves sampling (a_i, b_i) and determining if $b_i = b_j$ for some $j \in S$ (in which case $A_{ji} = 1$), and including the remaining edges with probability q.

3. Finally, fill in all of the remaining edges (i.e., the edges between elements of $[n] \backslash S$) conditioned on the decisions made in steps 1 and 2.

Steps 1 and 3 can be exactly mimicked under the semi-random model — step 1 because it involves planting a clique at random, and step 3 because it only involves decisions in $[n] \backslash S$, which we are allowed to choose arbitrarily. For step 2, however, we cannot exactly mimic P_0 in the semi-random model — our hands are tied because all edges between S and $[n] \backslash S$ must be generated at random. Under P_1, then, steps 1 and 3 are the same as P_0, but we use the alternate step 2':

2'. For each $i \notin S$, determine which edges should exist between S and $[n] \backslash S$. This involves setting A_{ji} to be 1 with probability $\frac{1}{2}$ independently for all $j \in S$.

It is also necessary to sample (a_i, b_i) in step 2', since these are used in step 3. Since a_i and b_i do not affect the edges generated in 2', we can sample them however we want without violating the semi-random model. For each i, we will thus mimic P_0 by sampling (a_i, b_i) from the conditional distribution under P_0 given the decisions made so far (e.g. conditioned on S, on the previous $(a_{i'}, b_{i'})$, and on A_{ji}).

Comparing P_0 and P_1. We next want to show that P_0 and P_1 are close. For this, it helps to more concretely represent the difference between P_0 and P_1 in step 2. For convenience, number the vertices in S as $1, \ldots, s$ and assume that $b_i = i$ for $i = 1, \ldots, s$. Suppose that we have already filled in the edges from S to vertices $s + 1, \ldots, i - 1$ (and also sampled $b_{s+1}, \ldots, b_{i-1}$), and want to fill in the edges from S to vertex i. Conditioned on $b_{s+1:i-1}$, these edges have the following distribution under P_0:

- For each $j = 1, \ldots, s$, set $A_{ji} = 1$ with probability q.

- Additionally, sample $j_0 \sim \pi$, where $\pi_j = \frac{m-1-|\{b_l=j|s+1 \le l < i\}|}{m^2-m-(i-s-1)}$. If $j_0 \in [s]$ then set $A_{j_0 i} = 1$ as well.

5

The expression for π_j comes from the fact that there are $m - 1$ pairs (a, b) with $a \neq 1$ and $b = j$, but some might have been used up already from the samples $b_{s+1}, \ldots, b_{i-1}$.

Under P_1, the A_{ji} are instead all generated independently with probability $\frac{1}{2}$. The following technical lemma shows that these two distributions are close together:

Lemma 2.1. *Fix non-negative reals $\tau_1, \ldots, \tau_s$ and $\pi_1, \ldots, \pi_s$ such that $\tau_j \leq \frac{1}{5}$ and $\sum_j \tau_j \leq 1$, and let $q \in [4/9, 1]$. Consider random variables $X_{1:s}$ and $X'_{1:s}$ defined as follows:*

- *$X_{1:s} \sim \text{Ber}(q)$ initially, and then an index j_0 is sampled with probability π_{j_0} and X_{j_0} is set to 1. (With probability $1 - \sum_{j=1}^{s} \pi_j$, no X_{j_0} is set to 1 in this latter step.)*

- *$X'_j \sim \text{Ber}(q + \tau_j(1 - q))$ for all j.*

Then, we have $D_{KL}(P_X, P_{X'}) \leq 10 \left(\sum_{j=1}^{s} \tau_j^2 \right)^2 + 10 \sum_{j=1}^{s} (\tau_j - \pi_j)^2$, where P_X and $P_{X'}$ are the distributions over X and X'.

We can apply Lemma 2.1 with $\tau_i = \frac{1}{m}$ (since $q + (1 - q)/m = \frac{1}{2}$), which yields

$$D_{KL}(P_0(A_{1:s,i}), P_1(A_{1:s,i}) \mid b_{1:i-1}, s) \leq 10 \frac{s^2}{m^4} + 10 \sum_{j=1}^{s} (\pi_j(b_{1:i-1}) - 1/m)^2. \tag{4}$$

Chaining the KL divergence. Having bounded the KL divergence of each local decision, we would like to obtain a global bound on the difference between P_0 and P_1. We can do this with the following inequality, which follows from the chain rule for KL divergence (all expectations are with respect to P_0):

$$D_{KL}(P_0(A), P_1(A)) \leq \mathbb{E}_s \left[\sum_{i=s+1}^{n} \mathbb{E}_{b_{1:i-1}} \left[D_{KL}(P_0(A_{1:s,i}), P_1(A_{1:s,i}) \mid b_{1:i-1}, s) \right] \right]. \tag{5}$$

Plugging (4) into (5), we obtain

$$D_{KL}(P_0(A), P_1(A)) \leq \mathbb{E}_s \left[\frac{10s^2 n}{m^4} + 10 \sum_{i=s+1}^{n} \sum_{j=1}^{s} \mathbb{E}_{b_{1:i-1}} \left[(\pi_j(b_{1:i-1}) - 1/m)^2 \right] \right]. \tag{6}$$

To bound (6) we need to analyze the mean and variance of π_j. By symmetry, $\mathbb{E}[\pi_j(b_{1:i-1})] = \pi_j(b_{1:s}) = \frac{1}{m}$. Also, $\text{Var}[\pi_j(b_{1:i-1})] = \frac{1}{(m^2 - m - (i-s-1))^2} \text{Var}[\sum_{l=s+1}^{i-1} \mathbb{I}[b_l = j]]$. We have $(m^2 - m - (i - s - 1))^2 \geq (m^2 - 2n)^2 \geq \frac{1}{2} m^4$ assuming $m^2 \geq 7n$. Furthermore, the events $\mathbb{I}[b_l = j]$ are negatively correlated (because we sample without replacement). Therefore, $\text{Var}[\pi_j]$ is bounded above as

$$\text{Var}[\pi_j(b_{1:i-1})] \leq \frac{2}{m^4} \text{Var} \left[\sum_{l=s+1}^{i-1} \mathbb{I}[b_l = j] \right] \tag{7}$$

$$\leq \frac{2}{m^4} \sum_{l=s+1}^{i-1} \text{Var}[\mathbb{I}[b_l = j]] \tag{8}$$

$$\leq \frac{2(i - s - 1)}{m^5} \leq \frac{2n}{m^5}. \tag{9}$$

We thus have $\mathbb{E}_{b_{1:i-1}}[(\pi_j(b_{1:i-1}) - 1/m)^2] = \text{Var}_{b_{1:i-1}}[\pi_j(b_{1:i-1})] \leq \frac{2n}{m^5}$.

Now, plugging back into (6) and using the fact that $\mathbb{E}[s] = n/m$, $\mathrm{Var}[s] \leq n/m$, we obtain

$$D_{KL}(P_0(A), P_1(A)) \leq \mathbb{E}_s \left[\frac{10s^2 n}{m^4} + \frac{20sn^2}{m^5} \right] \tag{10}$$

$$\leq \frac{10n}{m^4}((n/m)^2 + n/m) + \frac{20n^3}{m^6} \leq \frac{40n^3}{m^6}. \tag{11}$$

By Pinsker's inequality, this implies that $D_{TV}(P_0(A), P_1(A)) \leq \sqrt{20n^3/m^6} \leq 5(n/m^2)^{1.5}$.

Proving Theorem 1.1. Now that we have a bound on $D_{TV}(P_0(A), P_1(A))$, proving Theorem 1.1 is straightforward. First note that under P_0, each vertex v belongs to two cliques and it is impossible to tell which one is S. Therefore, we have $\mathbb{E}_{P_0}[|S \triangle \hat{S}(A, v)|] \geq \mathbb{E}_{P_0}[|S| - 1] = \frac{n}{m} - 1$.

We note that $P_0[|S| \leq \frac{n}{2m}] \leq e^{-n/8m}$ and $P_0[|S| \geq \frac{3n}{2m}] \leq e^{-n/8m}$ (see Hoeffding (1963)). Therefore, at the cost of an additional TV distance of $2e^{-n/8m}$, we can modify P_1 to ensure that $|S| \in [\frac{n}{2m}, \frac{3n}{2m}]$ almost surely, while also preserving $\mathbb{E}[|S|]$ for convenience.

Now, let $\hat{S}(A, v)$ be the optimal estimator under P_1, which satisfies $|\hat{S}(A, v)| \leq \frac{3n}{m}$ since $|S| \leq \frac{3n}{2m}$. The difference of $\mathbb{E}[|S \triangle \hat{S}(A, v)|]$ between P_0 and P_1 can be bounded by $\mathbb{E}_{P_0}[|S \triangle \hat{S}(A, v)||E]$, where E is an event of probability $p = D_{TV}(P_0, P_1)$. Then

$$\mathbb{E}_{P_0}[|S \triangle \hat{S}(A, v)|] - \mathbb{E}_{P_1}[|S \triangle \hat{S}(A, v)|] \leq \mathbb{E}_{P_0}[|S \triangle \hat{S}(A, v)||E] \tag{12}$$

$$\leq \mathbb{E}_{P_0}[(|S| + |\hat{S}(A, v)|)E] \tag{13}$$

$$\leq p \left(\frac{4n}{m} + \sqrt{(2n/m)\log(e/p)} \right), \tag{14}$$

where the final line exploits the boundedness of $|\hat{S}|$ and the sub-Gaussianity of $|S|$.

In sum, we have $\mathbb{E}_{P_1}[|S \triangle \hat{S}(A, v)|] \geq \mathbb{E}_{P_0}[|S \triangle \hat{S}(A, v)|] - p \left(\frac{4n}{m} + \sqrt{(2n/m)\log(e/p)} \right)$, where $p = D_{TV}(P_0, P_1)$. Using our previous bound $p \leq 5(n/m^2)^{1.5} + 2e^{-n/8m}$ and letting $n \gg m \log(m)$ to bound lower-order terms, we obtain:

Proposition 2.2. *If $24m \log(m) \leq n \leq \frac{m^2}{7}$, there is a distribution over instances of the semi-random planted clique model such that S always has size at least $\frac{n}{2m}$, and $\mathbb{E}[|S \triangle \hat{S}(A, v)|] \geq \mathbb{E}[|S|](1 - \frac{m}{n} - 25(n/m^2)^{1.5})$.*

We will set $m = 3(1/\delta)^{1/3}\sqrt{n}$, in which case we get cliques of size at least $k = \frac{1}{6}\delta^{1/3}\sqrt{n}$, and we have $25(n/m^2)^{1.5} + \frac{m}{n} \leq 0.93\delta + \frac{3}{\delta^{1/3}\sqrt{n}}$. For sufficiently large n, this is at most δ, which yields Theorem 1.1.

3 Upper Bound

We next turn to the upper bound (Theorem 1.2). The crux is the following lemma showing that the planted clique is nearly disjoint from all other large cliques:

Proposition 3.1. *Let S be a planted clique of size k in a subgraph of size n under the semi-random model. Then, with probability at least $1 - \frac{2k}{n^2}$, any other clique S' of size at least k satisfies $|S \cap S'| < 3\log_2(n)$.*

Now, call a clique *good* if it has size at least k, and its intersection with any other clique of size at least k is at most $3\log_2(n)$. We have just seen that the planted clique S is good with probability $1 - o(1)$. Moreover, if $k \geq 3\sqrt{n}\log_2(n)$, then there must be less than $\frac{2n}{k}$ good cliques. To see this, note that by the principle of

inclusion-exclusion, the union of m good cliques has size at least $mk - 3\binom{m}{2}\log_2(n) > mk - \frac{3}{2}m^2\log_2(n)$, and so we must have $m(k - \frac{3}{2}m\log_2(n)) < n$. If we take $m = \frac{2n}{k}$ then we obtain $\frac{3}{2}m\log_2(n) < \frac{k}{2}$, and hence $m(k - \frac{3}{2}m\log_2(n)) \geq mk/2 = n$, which is a contradiction. This shows that we indeed have $m < \frac{2n}{k}$.

Now, since there are at most $\frac{2n}{k}$ good cliques, the total number of vertices in S that intersect with any other good clique is at most $\frac{6n\log_2(n)}{k}$, and so the fraction of such vertices in S is at most $\frac{6n\log_2(n)}{k^2}$. This yields the following recovery algorithm: given A and v, if v lies in a unique good clique then output that clique as S; otherwise, output the empty set. With probability $1 - \frac{2k}{n^2} - \frac{6n\log_2(n)}{k^2}$, this gives us exact recovery of the planted clique S, which completes the first part of Theorem 1.2.

For the second part, we invoke Corollary 9.3 of Charikar et al. (2017). While their result is more general, in our context it specializes to the following:

Theorem 3.2 (Charikar et al. (2017)). *Let A be a graph drawn from the semi-random model with a planted clique of size k. Then there is a polynomial time algorithm which, with probability $1 - \exp(-\Omega(k))$, outputs sets $\hat{S}_1, \ldots, \hat{S}_m$ with $m \leq \frac{4n}{k}$, such that $\min_{j=1}^{m} |S \triangle \hat{S}_j| = \mathcal{O}\left((n/k)^2 \log(n)\right)$.*

Now for a large enough constant C, if $k \geq C \cdot n^{2/3}\log^{1/3}(n)$ then the bound in Theorem 3.2 translates to $|S \triangle \hat{S}_j| \leq \frac{k}{8}$. Then, any vertex $i \in S$ will be connected to at least $\frac{7k}{8}$ elements in $\hat{S}_j$, while with probability $1 - n\exp(-\Omega(k))$, no vertex not in S will be connected to more than $\frac{3k}{4}$ elements in $\hat{S}_j$. We can thus define $\tilde{S}_j$ to be the set of vertices that are connected to at least $\frac{7k}{8}$ elements in $\hat{S}_j$, and with high probability one of the $\tilde{S}_j$ will be the planted clique S.

To finish, we remove any $\tilde{S}_j$ that is not a clique of size at least k, and then also any $\tilde{S}_j$ that has intersection greater than $3\log_2(n)$ with any of the other remaining $\tilde{S}_{j'}$. Given a vertex v, we return the $\tilde{S}_j$ that it belongs to (if it exists and is unique) and otherwise return the empty set. By the same logic as before, this outputs S with probability at least $1 - \frac{2k}{n^2} - \frac{12n\log_2(n)}{k^2} - (n+1)\exp(-\Omega(k)) = 1 - o(1)$.

This completes the proof of Theorem 1.2. We remark that Proposition 3.1 remains true under Feige and Kilian's model, and hence the information-theoretic part of Theorem 1.2 holds in that model as well.

A   Proof of Lemma 2.1

Proof. Note that $X'_{1:s}$ are independent Bernoulli variables, and hence

$$P_{X'}(x_{1:s}) = \prod_{j=1}^{s}(q + \tau_j(1-q))^{x_j}((1-q)(1-\tau_j))^{1-x_j} \tag{15}$$

$$= \left(\prod_{j=1}^{s} q^{x_j}(1-q)^{1-x_j}\right)\prod_{j=1}^{s}(1 + \tau_j(1/q - 1))^{x_j}(1-\tau_j)^{1-x_j} \tag{16}$$

$$= \left(\prod_{j=1}^{s} q^{x_j}(1-q)^{1-x_j}\right)\left(\prod_{j=1}^{s} 1 + \tau_j(-1 + \mathbb{I}[x_j = 1]/q)\right). \tag{17}$$

On the other hand, by summing over the different possible samples from π, we can calculate P_X as

$$P_X(x_{1:s}) = (1 - \sum_{j=1}^{s}\pi_j)\prod_{j=1}^{s} q^{x_j}(1-q)^{1-x_j} + \sum_{j=1}^{s}\pi_j\mathbb{I}[x_j = 1]\prod_{j'\neq j} q^{x_{j'}}(1-q)^{1-x_{j'}} \tag{18}$$

$$= \left(\prod_{j=1}^{s} q^{x_j}(1-q)^{1-x_j}\right)\left(1 + \sum_{j=1}^{s}\pi_j(-1 + \mathbb{I}[x_j = 1]/q)\right). \tag{19}$$

8

Note the similarity between (17) and (19). Motivated by this, we define $y_j(x) = \pi_j(-1 + \mathbb{I}[x_j = 1]/q)$ and $z_j(x) = \tau_j(-1 + \mathbb{I}[x_j = 1]/q)$. Since $q \geq 4/9$, we have $|z_j(x)| \leq 1.25\tau_j \leq \frac{1}{4}$, and $\sum_j |z_j(x)| \leq 1.25$.

By Lemma 2.7 of Tsybakov (2009), we can upper bound KL divergence by χ^2-divergence:

$$D_{KL}(P_X, P_{X'}) \leq D_{\chi^2}(P_X, P_{X'}) \tag{20}$$

$$= \sum_{x \in \{0,1\}^n} \frac{(P_X(x) - P_{X'}(x))^2}{P_{X'}(x)} \tag{21}$$

$$= \sum_{x \in \{0,1\}^n} \left(\prod_{j=1}^s q^{x_j}(1-q)^{1-x_j} \right) \frac{\left(\prod_{j=1}^s (1 + z_j(x)) - \sum_{j=1}^s y_j(x) - 1 \right)^2}{\prod_{j=1}^s (1 + z_j(x))}. \tag{22}$$

Using the fact that $|z_j(x)| \leq \frac{1}{4}$ and hence $1 + z_j(x) \geq \exp(-1.2|z_j(x)|)$, we can bound the denominator as $\prod_{j=1}^s (1 + z_j(x)) \geq \exp(-1.2 \sum_{j=1}^s |z_j(x)|) \geq \exp(-1.5) > \frac{1}{5}$. We can also re-write the term in the numerator as $\sum_{|J| \geq 2} \prod_{j \in J} z_j(x) + \sum_{j=1}^s (z_j(x) - y_j(x))$, and treat the sum over $x \in \{0,1\}^n$ as an expectation with respect to a $\mathrm{Ber}(q)$ distribution. Together, these yield

$$D_{KL}(P_X, P_{X'}) \leq 5 \cdot \mathbb{E}_{x_{1:s} \sim \mathrm{Ber}(q)} \left[\left(\sum_{|J| \geq 2} \prod_{j \in J} z_j(x) + \sum_{j=1}^s (z_j(x) - y_j(x)) \right)^2 \right]. \tag{23}$$

Now note that $\mathbb{E}_{x \sim \mathrm{Ber}(q)}[z_j(x)] = \mathbb{E}_{x \sim \mathrm{Ber}(q)}[y_j(x)] = 0$ for all j, and furthermore that the y_j, z_j are independent across different j. Together this implies that most of the terms in (23) are 0; indeed, a term will have expectation 0 unless each index j occurs twice, which implies

$$\mathbb{E}_x \left[\left(\sum_{|J| \geq 2} \prod_{j \in J} z_j(x) + \sum_{j=1}^s (z_j(x) - y_j(x)) \right)^2 \right] = \mathbb{E}_x \left[\sum_{|J| \geq 2} \prod_{j \in J} z_j(x)^2 + \sum_{j=1}^s (z_j(x) - y_j(x))^2 \right]. \tag{24}$$

Now, since $q \geq \frac{4}{9}$ we have $z_j(x)^2 \leq 1.6\tau_j^2$ and $(z_j(x) - y_j(x))^2 \leq 1.6(\tau_j - \pi_j)^2$. We thus obtain

$$D_{KL}(P_X, P_{X'}) \leq 5 \sum_{|J| \geq 2} \prod_{j \in J} 1.6\tau_j^2 + 8 \sum_{j=1}^s (\tau_j - \pi_j)^2. \tag{25}$$

To finish, we make use of the following lemma:

Lemma A.1. *Suppose that $c_j \geq 0$ and $\sum_{j=1}^s c_j < 3$. Then,*

$$\sum_{|J| \geq 2} \prod_{j \in J} c_j \leq \frac{\sum_{1 \leq j < j' \leq s} c_j c_{j'}}{1 - \frac{1}{3} \sum_{j=1}^s c_j}. \tag{26}$$

Since $\sum_{j=1}^s 1.6\tau_j^2 \leq \frac{1.6}{5} < \frac{1}{3}$, Lemma A.1 yields the bound

$$D_{KL}(P_X, P_{X'}) \leq \frac{5 \cdot 1.6^2}{8/9} \sum_{j < j'} \tau_j^2 \tau_{j'}^2 + 8 \sum_{j=1}^s (\tau_j - \pi_j)^2 \tag{27}$$

$$\leq 8 \left(\sum_{j=1}^s \tau_j^2 \right)^2 + 8 \sum_{j=1}^s (\tau_j - \pi_j)^2, \tag{28}$$

as was to be shown. □

9

Proof of Lemma A.1. For a given size r, let

$$T_r \overset{\text{def}}{=} \sum_{|J|=r} \prod_{j \in J} c_j. \tag{29}$$

We have $(\sum_{j=1}^s c_j)T_r \geq (r+1)T_{r+1}$, since the left-hand-side contains $r+1$ instances of each term in T_{r+1}, as well as additional terms which will always be non-negative since the c_j are non-negative. In particular, for all $r \geq 2$ we have $T_{r+1} \leq \frac{1}{3}(\sum_{j=1}^s c_j)T_r$. Then $\sum_{r \geq 2} T_r \leq \frac{T_2}{1-\frac{1}{3}\sum_{j=1}^s c_j}$, as was to be shown. $\qquad\square$

B   Proof of Equation (5)

We make use of the chain rule for KL divergence, which says that given distributions $P_0(X_{1:N})$ and $P_1(X_{1:N})$, we have $D_{KL}(P_0(X_{1:N}), P_1(X_{1:N})) = \sum_{i=1}^N \mathbb{E}[D_{KL}(P_0(X_i), P_1(X_i)) \mid X_{1:i-1}]$ (all expectations are with respect to P_0).

In our case, we want to bound $D_{KL}(P_0(A), P_1(A))$. We will add in the auxiliary variables $b_{1:n}$ and s, which will only increase the KL divergence, and then apply the chain rule. For short-hand, we use $D_{KL}(X \mid Y)$ to denote $D_{KL}(P_0(X \mid Y), P_1(X \mid Y))$. We have:

$$D_{KL}(P_0(A), P_1(A)) \leq D_{KL}(P_0(A, b, s), P_1(A, b, s)) \tag{30}$$

$$= D_{KL}(s, A_{1:s,1:s}, b_{1:s}) + \mathbb{E}_s\left[\sum_{i=s+1}^n \mathbb{E}[D_{KL}(A_{1:s,i}, b_i \mid s, A_{1:s,1:i-1}, b_{1:i-1})]\right]$$

$$+ \mathbb{E}[D_{KL}(A_{s+1:n,s+1:n} \mid s, A_{1:s,1:n}, b_{1:n})] \tag{31}$$

$$= \mathbb{E}_s\left[\sum_{i=s+1}^n \mathbb{E}[D_{KL}(A_{1:s,i}, b_i \mid s, A_{1:s,1:i-1}, b_{1:i-1})]\right], \tag{32}$$

where the final equality is because all of the other conditional distributions are identical under P_0 and P_1.

Furthermore, $A_{1:s,i}, b_i$ are independent of $A_{1:s,1:i-1}$ conditioned on s and $b_{1:i-1}$, so

$$D_{KL}(A_{1:s,i}, b_i \mid s, A_{1:s,1:i-1}, b_{1:i-1}) = D_{KL}(A_{1:s,i}, b_i \mid s, b_{1:i-1}) \tag{33}$$

$$= D_{KL}(A_{1:s,i} \mid s, b_{1:i-1}) + \mathbb{E}[D_{KL}(b_i \mid s, b_{1:i-1}, A_{1:s,i})] \tag{34}$$

$$= D_{KL}(A_{1:s,i} \mid s, b_{1:i-1}). \tag{35}$$

Here again the final equality is because b_i has an identical conditional distribution under P_0 and P_1.

Plugging (35) into (32), we obtain

$$D_{KL}(P_0(A), P_1(A)) \leq \mathbb{E}_s\left[\sum_{i=s+1}^n \mathbb{E}_{b_{1:i-1}}[D_{KL}(A_{1:s,i} \mid s, b_{1:i-1})]\right], \tag{36}$$

which is exactly the statement of (5).

C   Proof of Proposition 3.1

Take any candidate clique S', which we can assume has size exactly k (since any larger S'' would contain a clique S' of size k). For any such S', all of the edges between S and S' must be present, which occurs

with probability $(1/2)^{l(k-l)}$, where $l = |S \cap S'|$. On the other hand, there are $\binom{k}{l}\binom{n-k}{k-l}$ sets of size k with intersection l. Union bounding over all $l \geq l_0$, the probability that there is any clique with intersection greater than $l_0 = 3\log_2(n)$ is at most

$$\sum_{l_0 \leq l < k} \binom{k}{l}\binom{n-k}{k-l} 2^{-l(k-l)} \leq \sum_{l_0 \leq l < k} \binom{k}{l}\left(\frac{n-k}{2^l}\right)^{k-l} \tag{37}$$

$$\leq \sum_{l_0 \leq l < k} \binom{k}{l}(1/n^2)^{k-l} \tag{38}$$

$$\leq (1+1/n^2)^k - 1 \leq 1 + \frac{2k}{n^2}, \tag{39}$$

where the final inequality holds because $(1+1/n^2)^k \leq \exp(k/n^2) \leq 1 + 2k/n^2$ since $k/n^2 \leq 1$.

References

N. Agarwal, A. S. Bandeira, K. Koiliaris, and A. Kolla. Multisection in the stochastic block model using semidefinite programming. *arXiv*, 2015.

B. Barak, S. B. Hopkins, J. Kelner, P. Kothari, A. Moitra, and A. Potechin. A nearly tight sum-of-squares lower bound for the planted clique problem. In *Foundations of Computer Science (FOCS)*, pages 428–437, 2016.

A. Blum and J. Spencer. Coloring random and semi-random k-colorable graphs. *Journal of Algorithms*, 19 (2):204–234, 1995.

M. Charikar, J. Steinhardt, and G. Valiant. Learning from untrusted data. In *Symposium on Theory of Computing (STOC)*, 2017.

Y. Chen, S. Sanghavi, and H. Xu. Improved graph clustering. *IEEE Transactions on Information Theory*, 60 (10):6440–6455, 2014.

A. Coja-Oghlan. Coloring semirandom graphs optimally. *Automata, Languages and Programming*, pages 71–100, 2004.

A. Coja-Oghlan. Solving NP-hard semirandom graph problems in polynomial expected time. *Journal of Algorithms*, 62(1):19–46, 2007.

A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6), 2011.

U. Feige and J. Kilian. Heuristics for semirandom graph problems. *Journal of Computer and System Sciences*, 63(4):639–671, 2001.

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.

O. Guédon and R. Vershynin. Community detection in sparse networks via Grothendieck's inequality. *arXiv*, 2014.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

H. Kesten and B. P. Stigum. Additional limit theorems for indecomposable multidimensional Galton-Watson processes. *The Annals of Mathematical Statistics*, 37(6):1463–1481, 1966a.

H. Kesten and B. P. Stigum. A limit theorem for multidimensional Galton-Watson processes. *The Annals of Mathematical Statistics*, 37(5):1211–1223, 1966b.

K. Makarychev, Y. Makarychev, and A. Vijayaraghavan. Approximation algorithms for semi-random partitioning problems. In *Symposium on Theory of Computing (STOC)*, pages 367–384, 2012.

A. Moitra, W. Perry, and A. S. Wein. How robust are reconstruction thresholds for community detection? *arXiv*, 2015.

J. Steinhardt, G. Valiant, and M. Charikar. Avoiding imposters and delinquents: Adversarial crowdsourcing and peer prediction. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

J. Steinhardt, M. Charikar, and G. Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. *arXiv*, 2017.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.