Scribed by Jeff Xu

Lecture 7

In which we discussed planted clique distribution, specifically, we talked about how to find a planted clique in a random graph. We heavily relied upon our material back in lecture 2 and lecture 3 in which we covered the upper bound certificate for max clique in $G_{n,\frac{1}{2}}$. At the end of this class, we wrapped up this topic and started the topic of k-SAT.

1 Planted Clique

To start with, we describe a distribution of graphs with a planted clique. Suppose that we sample G from $G_{n,\frac{1}{2}}$ and we want to modify G s.t. it has a size k clique, i.e., we have a clique $S \subseteq V$ with $|S| = k$. The following code describes a sampler for the distribution.

- $G \leftarrow G_{n,\frac{1}{2}}$
- Pick a subset of vertices S from V s.t. $|S| = k$
- Sample Edges:

$$\text{edge } (u,v) \text{ exists w.p. } \begin{cases} 1 & \text{if } u,v \in S \\ \frac{1}{2} & \text{otherwise} \end{cases}$$

Note: We are only interested in the case $k \geq 2\log n$, which is the case in which the planted clique is, with high probability, larger than any pre-existing clique

1.1 Finding the planted clique when $k \gg \sqrt{n \log n}$

When $k \gg \sqrt{n \log n}$, finding the planted clique is easy because the k vertices in the planted clique are precisely the k vertices of higher degree.

Lemma 1 *In $G_{n,\frac{1}{2}}$, w.h.p., for every vertex v, $\frac{n-1}{2} - \sqrt{n-1}\sqrt{\ln n} \leq deg(v) \leq \frac{n-1}{2} + \sqrt{n-1}\sqrt{\ln n}$.*

PROOF: For each vertex v in a graph $G \sim G_{n,\frac{1}{2}}$, we have $\deg(v)$= sum of $n-1$ random bits, which is simply a Binomial distribution. By Chernoff bound,

$$\mathbb{P}\left[\left|\sum_i x_i - \frac{n-1}{2}\right| > t\right] \leq 2e^{-\frac{2t^2}{n-1}}$$

For this probability to be upper bounded, say by $\frac{1}{n^2}$, we can fix $t = \sqrt{(n-1)\ln n}$ s.t. $t^2 \geq (n-1)\ln n$ and this completes the proof that with high probability, a vertex v in $G_{n,\frac{1}{2}}$ random graph has degree $\frac{n-1}{2} \pm \sqrt{n-1}\sqrt{\ln n}$. $\square$

Now we consider a vertex v in the planted clique S.

Claim 2 *In a graph with a planted clique coming from a $G_{n,\frac{1}{2}}$ random graph to which we add all the edges necessary to make S a clique, each node in S will receive $\geq 0.4k$ added edges w.h.p. over the sampling of graph.*

PROOF: Again, we regard the number of neighbors of a vertex v in S as a sum of Bernoulli distribution and denote it by X. By Chernoff bound, we obtain an upper bound on the probability that a vertex $v \in S$ has more than $0.6k$ neighbors in S in a random $G_{n,\frac{1}{2}}$ graph.

$$\mathbb{P}\left[X > 0.6k\right] = \mathbb{P}\left[X - 0.5k > 0.1k\right] \leq e^{-\frac{0.2^2 * k}{2}} = e^{-O(k)}$$

Since the probability is exponentially small in k, we can conclude that a node in S, with high probability, has less than $0.6k$ edges in the original $G_{n,\frac{1}{2}}$ random graph, and thus at least $0.4k$ edges will be added to each vertex in S. $\square$

Corollary 3 *In a graph with a planted clique, a vertex in S will have degree $\geq \frac{n-1}{2} + 36\sqrt{n\log n}$. (Note: we have $k = 100\sqrt{n\log n}$ in this example)*

Therefore, we show that in graph with a large planted clique, we can distinguish it from $G_{n,\frac{1}{2}}$ distribution by the existence of node with large degree, i.e., degree over $\frac{n-1}{2} + 4\sqrt{n\log n}$.

1.2 Distinguish Planted Clique Distribution with $k \gg \sqrt{n}$

Moving on to the case in which k is of the order of $\sqrt{n}$, we first show how to distinguish graphs sampled from the planted clique distribution from $G_{n,\frac{1}{2}}$ random graphs.

Say that $k = 100\sqrt{n}$, and let A be the adjacency matrix of a graph from the planted clique distribution. Then

$$\begin{aligned}
\left|\left|A - \frac{1}{2}J\right|\right| &\geq \lambda_{max}\left(A - \frac{1}{2}J\right) \\
&= \max_x \frac{x^T\left(A - \frac{1}{2}J\right)x}{x^T x} \\
&\geq \frac{\mathbf{1}_s^T\left(A - \frac{1}{2}J\right)\mathbf{1}_s}{||\mathbf{1}_s||^2} \\
&= \frac{\mathbf{1}_s^T A\mathbf{1}_s - \frac{1}{2}\mathbf{1}_s^t J\mathbf{1}_s}{k} \\
&= \frac{1}{k}\left(k^2 - k - \frac{1}{2}k^2\right) \\
&= \frac{k}{2} - 1
\end{aligned}$$

Now, recall he following theorem from Lecture 2.

Theorem 4 *If A is the adjacency matrix of a $G_{n,\frac{1}{2}}$ graph, w.h.p., $\left|\left|A - \frac{1}{2}J\right|\right| \leq 2\sqrt{n}$.*

Therefore, we can identify graph with a planted clique from a random $G_{n,\frac{1}{2}}$ distribution using this method.

1.3 Uniqueness of Maximum Clique in Planted Clique Distribution

In order to show that we can find the planted clique from a graph, we want to prove first that the maximum clique in planted clique Distribution is unique. In other words, we want to prove that the planted clique is the maximum clique in planted clique distribution. We first prove the following lemmas:

Lemma 5 *For each vertex not in the planted clique, i.e., $v \in V - S$, #of v's neighbors in $S \leq .5k + \sqrt{k-1}\sqrt{\ln n} \leq .55k$.*

PROOF: This is largely similar to Lemma 1. We see this as a sum of $k-1$ random $0-1$ bits and by Chernoff bound, we have:

$$\mathbb{P}\left[\left|\sum_i x_i - \frac{k-1}{2}\right| > t\right] \leq e^{-\frac{2t^2}{k-1}}$$

For this probability to be upper bounded, say by $\frac{1}{n}$, we can choose a t s.t. $t^2 \geq (k-1)\ln n$. Therefore, we pick $t = \sqrt{(k-1)\ln n}$ and this completes the proof that, with high probability, each vertex not in the planted clique has less no more than $.55k$ neighbors in the planted clique. $\square$

Lemma 6 *G sampled from $G_{n,\frac{1}{2}}$, w.h.p., has a largest clique of size $2\log n$. (proved in lecture 2)*

Claim 7 *Under the above assumptions, S is the unique clique of size k in G.*

PROOF: Suppose, for the sake of contradiction, we find a clique T s.t. $T \neq S, |T| \geq k$. Since T, S are both cliques by assumption, $T - S$ is also clique. $G_{n,\frac{1}{2}}$ has a largest clique of size $\leq 2\log n$ w.h.p., so $|T - S| \leq 2\log n$ since it is a clique in $G_{n,\frac{1}{2}}$. Consider a vertex $t \in T - S$: the number of t's neighbors in $S \geq |T| \geq |T \cap S|$ is at least $k - 2\log n > 0.55k$, but this contradicts Lemma 5, which states that t should have no more than $.55k$ neighbors in S. □

1.4 Finding the Planted Clique

Now that we have shown the uniqueness of maximum clique, we want to proceed and show that we can find the planted clique. Let A be the adjacency matrix of a $G_{n,\frac{1}{2}}$ random graph with a planted clique of size $k = 100\sqrt{n}$, and let $\mathbf{x}$ be the maximizer of

$$\frac{\mathbf{x}^T \left(A - \frac{1}{2}J\right)\mathbf{x}}{||\mathbf{x}||^2} \geq \frac{k}{2} - 1 = 50\sqrt{n} - 1$$

We will show below that $\mathbf{x}$ is close to the indicator vector of S.

First, we need to note that we are no longer using the sampling method described earlier to attain a planted clique distribution. Alternatively, we sample our graph G from random $G_{n,\frac{1}{2}}$ distribution, pick a subset of vertices S from G and add to it the necessary edges to make S a clique. From this point of view, we have $G = G_{n,\frac{1}{2}} + G_{clique}$, where G is the graph with a planted clique and G_{clique} is a distribution of edges that we need to add. We can then represent the adjacency matrix of G as:

$$A = A_{random} + A_{clique}$$

where $A_{random} \sim G_{n,\frac{1}{2}}$ and $A_{clique} \sim G_{k,\frac{1}{2}}$.

By the theorem shown in lecture 2 (which we just recapped above), we have the following equations with high probability:

$$\left|\left|A_{random} - \frac{1}{2}J\right|\right| \leq 2\sqrt{n}$$

$$\left|\left|A_{clique} - \frac{1}{2}\mathbf{1}_s\mathbf{1}_s^T\right|\right| \leq 2\sqrt{k} = o(k)$$

Now we combine the equations listed above, and wlog, let $||\mathbf{x}|| = 1$.

$$\begin{aligned} 50\sqrt{n} - 1 &= \frac{k}{2} - 1 \\ &\leq \mathbf{x}^T \left(A - \frac{1}{2} J \right) \mathbf{x} \\ &= \mathbf{x}^T \left(A_{clique} + A_{random} - \frac{1}{2} J \right) \mathbf{x} \end{aligned}$$

With $||A_{random} - \frac{1}{2}J|| \leq 2\sqrt{n}$ from above, we have

$$\mathbf{x}^T A_{clique} \mathbf{x} \geq 48\sqrt{n} = 0.48k.$$

Therefore, $\mathbf{x}^T \left(A_{clique} - \frac{1}{2}\mathbf{1}_s\mathbf{1}_s^T + \frac{1}{2}\mathbf{1}_s\mathbf{1}_s^T \right) \mathbf{x} \geq 0.48k$. With $\left|\left| A_{clique} - \frac{1}{2}\mathbf{1}_s\mathbf{1}_s^T \right|\right| = o(k)$ shown above, we can conclude that

$$\mathbf{x}^T \left(\frac{1}{2}\mathbf{1}_s\mathbf{1}_s^T \right) \mathbf{x} \geq (0.48 - o(1))k$$

That is,

$$\langle \mathbf{x}, \mathbf{1}_S \rangle^2 \geq (0.96 - o(1))k$$

and, up to passing to $-\mathbf{x}$, which has the same set of largest k entries in absolute value, we have $\langle \mathbf{x}, \mathbf{1}_S \rangle \geq \sqrt{.96k - o(k)} \geq .979\sqrt{k}$, for sufficiently large k.

This means that, up to scaling, $\mathbf{x}$ and $\mathbf{1}_S$ are nearly identical.

$$\begin{aligned} ||\sqrt{k}\mathbf{x} - \mathbf{1}_S||^2 &= k||\mathbf{x}||^2 + ||\mathbf{1}_S|| - 2\sqrt{k}\langle \mathbf{x}, \mathbf{1}_S \rangle \\ &\leq 2k \cdot (1 - .979) \\ &\leq .042k \end{aligned}$$

Let L be the set of k largest entries of $\sqrt{k}\mathbf{x}$, and hence of $\mathbf{x}$, breaking ties arbitrarily, and let t be the threshold value for membership in L (that is, $\sqrt{k}x_i \geq t$ for all $i \in L$ and $\sqrt{k}x_i \leq t$ for all $i \notin L$). Suppose that there are B elements of S that are not in L, and hence B elements not in S that are in L. Then

$$\begin{aligned} ||\sqrt{k}\mathbf{x} - \mathbf{1}_S||^2 &= \sum_{i \in S} (\sqrt{k}x_i - 1)^2 + \sum_{i \notin S} k x_i^2 \\ &\geq B \cdot (1 - t)^2 + Bt^2 \\ &\geq \frac{1}{2}B \end{aligned}$$

And we conclude that $B \leq .084k$, that is, L contains at least $.9k$ of the k elements of S.

We now present an algorithm to find the planted clique. Let L be the set of k vertices that has largest $|x_i|$. We then consider each vertex $v \in L$. If $v \in S$, by our proof above, it should have ≥ 0.9 neighbors in L. If $v \notin S$, v should have $\leq 0.55k + 0.2k = 0.75k$ neighbors in L. Therefore, we can easily verify a vertex is in S by looking at its number of neighbors in L, and this gives us an algorithm.

- Algorithm(G):
- $A \leftarrow$ adjacency matrix of G
- $\mathbf{x} \leftarrow$ eigenvector of largest eigenvalue to $A - \frac{1}{2}J$
- $L \leftarrow$ set of k vertices with largest $|x_i|$
- clique $\leftarrow$ set of vertices with at least $0.8k$ neighbors in L

It is still an open problem to find a planted clique of size $o(\sqrt{n})$

2 Random k-SAT and Proof of Unsatisfiability

We start the topic on random k-SAT formulas. In k-SAT problem, we are trying to decide whether a formula in CNF with each clause containing up to k literals is satisfiable. We note in class that checking satisfiability for randomly generated equations is hard even in average case. Similar to the $G_{n,p}$ model, we generate a k-SAT formula on n variable with parameter p s.t. each of the $\binom{n}{k}2^k$ clauses exists with probability p.

Besides the model above, we also briefly mention another model, in which we randomly pick m of the $\binom{n}{k}2^k$ possible clauses. (Note: these two models are closely related when we have $m = p\binom{n}{k}2^k$).

To gain more insights, we discussed the example of 3-SAT problem. It has an expected number of satisfying assignments $2^n(\frac{7}{8})^m$. We also observe that for any 3-SAT instance f, we have $\mathbb{P}[f \text{ is satisfiable}] \leq \mathbb{E}[\#\text{of satisfying assignments to variables}] = 2^n(\frac{7}{8})^m$ which goes to 0 if $m > \log_{\frac{8}{7}} 2n$.