Beyond Worst Case Analysis **Fall 2018**

Lecture 3: Planted Clique in Random Graph

Scribe: Jiazheng Zhao, William Yang Sep 24 2018

3.1 Introduction

In this note, we introduce the planted clique problem, where instead of finding a large clique in a random graph, we are tasked with finding a large clique that is "planted" within a random graph. We discuss the AKS spectral algorithm for finding planted cliques of size $k \geq \Omega(\sqrt{n})$ and demonstrate a flavor of algorithmic analysis that apply to a broad class of spectral algorithms.

3.2 The Planted Clique Problem

The planted clique problem can be formulated as a search problem over random graphs selected from the *planted clique distribution* which is defined as follows. Given n, k as parameters, the planted clique distribution is a distribution over simple graphs constructed by the process

1. Sample $G' = (V, E)$ from the Erdős-Rényi distribution with $p = \frac{1}{2}$ (i.e. $G' \sim \mathcal{G}_{n,1/2}$).
2. Choose $S \subseteq V$ a subset of k vertices uniformly at random.
3. Construct G by planting a clique on S (i.e. connect all edges among vertices in S)

With G sampled from this process, the *planted clique problem* is to then recover S from G without knowledge of where the clique is planted. We should expect this problem to be easy to solve if the cliques in G' are much smaller than k. Last week, we demonstrated that w.h.p. the size of the maximum clique in $G' \sim \mathcal{G}_{n,1/2}$ is at most $(2 \pm o(1)) \log n$. It is information theoretically impossible to recover S if $k \leq 2 \log n$. However, the max-clique size of G' also implies a straight-forward quasi-polynomial time algorithm: simply brute force search for a clique over subsets of $3 \log n$ vertices.

This $n^{O(\log(n))}$-time algorithm will likely succeed as w.h.p. if G' contains a $3 \log n$ sized clique, it will be the only clique of that size and thus must be a subset of the planted k-clique. A natural question is to ask for what k is it possible to recover S in polynomial time? The AKS (Alon, Krevich and Sudakov) spectral algorithm we present recovers S if $k \geq \Omega(\sqrt{n})$. Many people believe this to be the best achievable in polynomial time. Indeed, the planted clique conjecture states for any $\epsilon \geq \Omega(1)$, it is NP-Hard to recover S from G where $k \leq n^{1/2-\epsilon}$. There have additionally been a number of hardness results using this conjecture in computing nash equilibria, property testing, and solving online learning problems.

3.2.1 Algorithm

The AKS spectral algorithm utilizes the spectrum of A, the adjacency matrix of sampled graph G, to recover S. Since A is a random matrix, a natural starting point for building an algorithm is to analyze the spectrum of $\mathbb{E}[A]$. Without loss of generality, let's relabel the vertices in G such that the planted clique comes first (i.e. vertices $1, \ldots, k$ contain the planted clique). The matrix $\mathbb{E}[A]$ looks like

$$\mathbb{E}[A] = \left(\begin{array}{c:c} 1 & 1/2 \\ \hdashline 1/2 & 1/2 \end{array}\right) - D$$

where $D_{ii} = 1$ if $i \in S$ otherwise $D_{ii} = \frac{1}{2}$. Written another way, $\mathbb{E}[A]$ is

$$\mathbb{E}[A] = \frac{1}{2}J + \frac{1}{2}\mathbf{1}_S\mathbf{1}_S^T - D$$

where J is the all-ones matrix and $\mathbf{1}_S$ is the 0-1 indicator vector on S. Why is this the expectation?

- First ignore the clique we planted. Since A is the adjacency matrix of $G \sim \mathcal{G}_{n,\frac{1}{2}}$, each edge exists with probability $\frac{1}{2}$, so each entry of A is 1 with probability 0.5 and 0 with probability 0.5. Therefore $\mathbb{E}[A_{ij}] = \frac{1}{2}$. We use a matrix where all entries are $\frac{1}{2}$ to represent that.
- S is the set of vertices that forms the clique. Because we deterministically plant an edge between all vertices in S, $A_{ij} = 1$ where $i, j \in S$. Note that the (i, j)-th entry of $\mathbf{1}_S\mathbf{1}_S^T$ is exactly one if $i, j \in S$ and 0 elsewhere. Therefore adding $\frac{1}{2}\mathbf{1}_S\mathbf{1}_S^T$ to the all $\frac{1}{2}$ will adjust the expected value of entries that represent edges in clique.

This says something rather useful! The diagonal matrix $-D$ can be split into a sum of rank-one matrices weighted by the elements of $-D$. By the spectral theorem, the top eigenvector of $\mathbb{E}[A] - \frac{1}{2}J$ is $\mathbf{1}_S$ since $-D$ only contains negative entries. The vector $\mathbf{1}_S$ is the indicator vector for the clique we wish to recover. For G sampled from the planted clique distribution, if the top eigenvector of $A - \frac{1}{2}J$ is close to that of $\mathbb{E}[A] - \frac{1}{2}J$, then we can hope to recover the planted clique S via some rounding procedure with high probability. This is precisely the intuition behind the AKS algorithm.

AKS Spectral Algorithm for Planted Clique
Given G sampled from the planted clique distribution, do the following:

1. Let $M = A - \frac{1}{2}J$ where A is the adjacency matrix of G.
2. Compute $\mathbf{x}$ the eigenvector corresponding to the largest eigenvalue of M.
3. Let I be the set of k vertices v corresponding to the largest values of $|\mathbf{x}_v|$.
4. Return $C = \{v \in V : v \text{ has at least } \frac{3}{4}k \text{ neighbors in } I\}$

The algorithm can be divided into two parts: steps (1) and (2) obtain an eigenvector from sampled G; steps (3) and (4) round the eigenvector to extract the planted clique. We note that spectral algorithms for "planted" problems like planted clique have a similar flavor to AKS. Supposing one can show on expectation, that

a witness for the planted structure manifests itself as (perhaps) a large eigenvector and if the behavior of sampled instances concentrates around the expectation, then with an appropriate (but sometimes non-trivial) rounding procedure, one can (at least) approximately recover the planted solution.

3.3 Analysis

Since the algorithm splits into two parts, the analysis of this algorithm is divided accordingly: first, we focus on the first two lines of the algorithm and show that the eigenvector $\mathbf{x}$ is very close to $\mathbf{1}_S$, the top eigenvector of $\mathbb{E}[A] - \frac{1}{2}J$ and planted solution, with high probability. Second, we argue that the way we recover the solution from the eigenvector selects the planted clique with high probability.

3.3.1 Analysis of Retrieved Eigenvector

If we can get that $\mathbf{x}$ is close to $\mathbf{1}_S$, then we have good reason to believe that rounding $\mathbf{x}$ will yield the planted solution. To show this, we will begin by demonstrating that M and $\frac{1}{2}\mathbf{1}_S\mathbf{1}_S^\top$ are "close" under the appropriate notion of distance. Once we show that these two matrices are "close", we can then show that their top eigenvectors (i.e. $\mathbf{x}$ and $\mathbf{1}_S$) are also "close".

How do we quantify "close" for matrices? Since we have only worked with real, symmetric matrices so far, we can use the spectral norm to measure distance between the two matrices. Let $\lambda_1, \ldots, \lambda_n$ denote the eigenvalues of any real, symmetric matrix M'. The spectral norm of M' is defined as

$$\|M'\| = \max_i |\lambda_i|$$

It will be helpful to know that A and $\mathbb{E}[A]$ are "close" in the spectral norm – their largest eigenvalues are not too far away from each other. This is what the following lemma states.

Lemma 3.1. *Fix a set S of size k. With high probability the following is true:*

$$\|A - \mathbb{E}[A]\| \leq (1 + o(1)) \cdot \sqrt{n}$$

where $\|\cdot\|$ is the spectral norm.

There are a number of ways to show this, so we omit the proof for a later set of notes to keep this exposition simple. We can apply this lemma and bound the "closeness" of M and $\frac{1}{2}\mathbf{1}_S\mathbf{1}_S^T$:

$$\left\|M - \frac{1}{2}\mathbf{1}_S\mathbf{1}_S^T\right\| = \left\|A - \frac{1}{2}J - \frac{1}{2}\mathbf{1}_S\mathbf{1}_S^T\right\| \leq \|A - \mathbb{E}[A] + D\| \leq \|A - \mathbb{E}[A]\| + \|D\| \leq (1 + o(1)) \cdot \sqrt{n}$$

The last two inequalities follow by triangle inequality and lemma 3.1 respectively. We have that the top eigenvalues of M and $\frac{1}{2}\mathbf{1}_S\mathbf{1}_S^T$ are close, but does that translate to their top eigenvectors? By the Davis-Kahan theorem, yes!

Theorem 3.2 (Davis and Kahan)**.** *If M is a symmetric matrix, $\mathbf{y}$ is a vector, $\mathbf{x}$ is an eigenvector of the largest eigenvalue of M, and $\hat{\mathbf{xy}}$ is the angle between $\mathbf{x}, \mathbf{y}$*

$$|\sin(\hat{\mathbf{xy}})| \leq \frac{\|M - \mathbf{y}\mathbf{y}^T\|}{\|\mathbf{y}\mathbf{y}^T\| - \|M - \mathbf{y}\mathbf{y}^T\|}$$

Again we leave the proof to a later set of notes. We quantify vector distance via $\|\cdot\|_2$, the ℓ_2-norm.

Corollary 3.3. *If $\|M - \frac{1}{2}\mathbf{1}_S\mathbf{1}_S^T\| \leq (1+o(1))\sqrt{n}$, $k > 100\sqrt{n}$, and $\mathbf{x}$ is an eigenvector corresponding to the largest eigenvalue of M scaled so that $\|\mathbf{x}\|_2^2 = k$, then*

$$\min\left\{\|\mathbf{x} - \mathbf{1}_S\|_2^2, \|-\mathbf{x} - \mathbf{1}_S\|_2^2\right\} \leq 0.002k$$

for sufficiently large n.

Proof. First we find $\|\frac{1}{2}\mathbf{1}_S\mathbf{1}_S^T\|$. Notice that the spectral norm of this rank one matrix is just the maximized Rayleigh quotient of the matrix. Since the matrix is rank one and the only eigenvector is $\mathbf{1}_S$, we can get

$$\left\|\frac{1}{2}\mathbf{1}_S\mathbf{1}_S^T\right\| = \frac{1}{2}\|\mathbf{1}_S\mathbf{1}_S^T\| = \frac{1}{2}\frac{\mathbf{1}_S^T\mathbf{1}_S\mathbf{1}_S^T\mathbf{1}_S}{\mathbf{1}_S^T\mathbf{1}_S} = \frac{1}{2}\mathbf{1}_S^T\mathbf{1}_S = \frac{k}{2}$$

Then we apply the Davis-Kahan theorem to derive the following.

$$|\sin(\mathbf{x}\hat{\mathbf{1}}_S)| \leq \frac{(1+o(1))\sqrt{n}}{\frac{k}{2} - (1+o(1))\sqrt{n}} = \frac{1}{49} + o(1)$$

Observe that by law of cosines, we have the following regarding ℓ_2-distance between $\mathbf{x}$ and $\mathbf{1}_S$.

$$\|\mathbf{x} - \mathbf{1}_S\|_2^2 = \|\mathbf{x}\|_2^2 + \|\mathbf{1}_S\|_2^2 - 2\langle\mathbf{x}, \mathbf{1}_S\rangle = 2k - 2\langle\mathbf{x}, \mathbf{1}_S\rangle$$

Similarly

$$\|-\mathbf{x} - \mathbf{1}_S\|_2^2 = 2k + 2\langle\mathbf{x}, \mathbf{1}_S\rangle$$

Therefore we have

$$\min\{\|\mathbf{x} - \mathbf{1}_S\|_2^2, \|-\mathbf{x} - \mathbf{1}_S\|_2^2\} = 2k - 2\cdot|\langle\mathbf{x}, \mathbf{1}_S\rangle|$$

Combining the previous results, we have

$$|\langle\mathbf{x}, \mathbf{1}_S\rangle| = \|\mathbf{x}\|_2 \cdot \|\mathbf{1}_S\|_2 \cdot \cos(\mathbf{x}\hat{\mathbf{1}}_S) = k\cos(\mathbf{x}\hat{\mathbf{1}}_S) = k\cdot\sqrt{1 - \sin^2(\mathbf{x}\hat{\mathbf{1}}_S)} \geq k\sqrt{\frac{49^2-1}{49^2} - o(1)} > 0.999\cdot k$$

Therefore $\min\{\|\mathbf{x} - \mathbf{1}_S\|^2, \|-\mathbf{x} - \mathbf{1}_S\|^2\} \leq 2k - 2\cdot 0.999\cdot k = 0.002k$. □

We have finally shown that the eigenvector $\mathbf{x}$ is close to the solution vector $\mathbf{1}_S$. But since $\mathbf{x}$ is a real valued vector, the solution to this problem is still not clear yet. The next step is to perform some deterministic rounding to retrieve the solution from the vector $\mathbf{x}$.

3.3.2 Analysis of the Rounding Step

To analyze the second half of the algorithm, we want to show that our rounding procedures recovers S exactly by demonstrating that, with high probability, C contains all the elements of S ($S \subseteq C$) and no elements outside of S ($C \subseteq S$).

3.3.2.1 $S \subseteq C$ with high probability

The first step towards showing this is that if I is defined to be the set of k vertices v for which $|\mathbf{x}_v|$ is largest, then the sets I and S are almost the same. This results from the following lemma:

Lemma 3.4. *If* $\mathbf{x}$ *is a vector such that* $\|\mathbf{x}\|_2^2 = |S| = k$ *where* $\min\{\|\mathbf{x} - \mathbf{1}_S\|_2^2, \| - \mathbf{x} - \mathbf{1}_S\|_2^2\} \leq \epsilon k$*, then*

$$|I \cap S| \geq k(1 - 4\epsilon)$$

Proof. Let's begin by defining bad vertices. Since we would like I and S to be as similar as possible, bad vertices relative to I and S should be v such that $v \in I$ but $v \notin S$ or $v \notin I$ but $v \in S$ (more concisely, $v \in I \oplus S$). Call $v \in V$ is *bad* if $v \in S$ and $|\mathbf{x}_v| \leq \frac{1}{2}$ or $v \notin S$ and $|\mathbf{x}_v| > \frac{1}{2}$. Let B be the set of bad vertices.

This definition matches our intuitive notion of bad vertices, since vertices v with higher values of $|\mathbf{x}_v|$ (say, $|\mathbf{x}_v| > \frac{1}{2}$) are more likely to be contained in I (as I is the set of k vertices for which $|\mathbf{x}_v|$ is the largest) and thus are bad if $v \notin S$, while vertices with lower values of $|\mathbf{x}_v|$ (say, $|\mathbf{x}_v| \leq \frac{1}{2}$) are less likely to be contained in I and are thus bad vertices if $v \in S$.

Observe that $|B| \leq 4\epsilon k$ since each vertex of B contributes at least $\frac{1}{4}$ to both $\|\mathbf{x} - \mathbf{1}_S\|_2^2$ and $\| - \mathbf{x} - \mathbf{1}_S\|_2^2$. Consider a bad vertex v and its contribution to $\|\mathbf{x} - \mathbf{1}_S\|_2^2$ and $\| - \mathbf{x} - \mathbf{1}_S\|_2^2$. There are two cases:

1. Suppose $v \in S$ and $|\mathbf{x}_v| \leq \frac{1}{2}$. The corresponding entry in the clique indicator vector is $\mathbf{1}_{S_v} = 1$ since $v \in S$. Vertex v would then contribute at least $\frac{1}{4}$ to both $\|\mathbf{x} - \mathbf{1}_S\|_2^2$ and $\| - \mathbf{x} - \mathbf{1}_S\|_2^2$ as both $(\mathbf{x}_v - 1)^2 \geq \frac{1}{4}$ and $(-\mathbf{x}_v - 1)^2 \geq \frac{1}{4}$ (terms in the expansion of the ℓ_2-norm).
2. Suppose $v \notin S$ and $|\mathbf{x}_v| > \frac{1}{2}$. The corresponding entry in the indicator vector of the clique is $\mathbf{1}_{S_v} = 0$. Thus v would also contribute at least $\frac{1}{4}$ to both $\|\mathbf{x} - \mathbf{1}_S\|_2^2$ and $\| - \mathbf{x} - \mathbf{1}_S\|_2^2$.

Our inequality for $|B|$ thus follows since

$$\frac{|B|}{4} \leq \min\{\|\mathbf{x} - \mathbf{1}_S\|_2^2, \| - \mathbf{x} - \mathbf{1}_S\|_2^2\} \leq \epsilon k$$

Concluding $|B| \leq 4\epsilon k$. The next observation we make is that $|I \cap S| \geq k - |B|$. Let $t = \min_{v \in I} |\mathbf{x}_v|$ and consider the two cases:

1. If $t > \frac{1}{2}$, then any bad vertices $v \in I$ but $v \notin S$ must have $|\mathbf{x}_v| > \frac{1}{2}$ by definition of t. However, I can contain at most $|B|$ of these bad vertices, and thus must contain at least $k - |B|$ vertices from S (since $|I| = |S| = k$).
2. If $t \leq \frac{1}{2}$ then any bad vertex v such that $v \notin I$ but $v \in S$ must have $|\mathbf{x}_v| \leq \frac{1}{2}$ otherwise v would have been added to I. However, this must mean that at least $k - |B|$ of the remaining vertices in S are also included in I since they have $|\mathbf{x}_v| \geq t$.

In both cases we have $|I \cap S| \geq k - |B|$. By the inequality above $|I \cap S| \geq k(1 - 4\epsilon)$ as required. □

From corollary 3.3, we can apply lemma 3.4 with $\epsilon = .002$ to derive with high probability, $I \cap S$ will contain at least $k - |B| \geq k - 4\epsilon k = .992k$ vertices. Immediately, we have that $S \subseteq C$ since all vertices in S will have at least $.992k > \frac{3}{4}k$ neighbors in I as S contains a clique.

3.3.2.2 $C \subseteq S$ with high probability

The final portion of our analysis is to show $C \subseteq S$ with high probability which will require Chernoff bounds.

Theorem 3.5 (Chernoff bound for a Binomial Random Variable)**.** *If $X_1, \ldots, X_k$ are i.i.d* Bernoulli(p) *random variables, and $X \coloneqq \sum_{i=1}^{k} X_i$, so that $X \sim$ Binomial(k, p), then, for every $0 < \epsilon < 1$*

$$\Pr[X - \mathbb{E}[X] \geq \epsilon k] \leq e^{-2\epsilon^2 k}$$
$$\Pr[X - \mathbb{E}[X] \leq -\epsilon k] \leq e^{-2\epsilon^2 k}$$

To relate this to the situation we have at hand, note that for a vertex $v \notin S$, the number of neighbors of v that are in S is a binomial random variable $Y = \sum_{i=1}^{k} Y_i$ where Y_i is an indicator random variable for the corresponding edge between a vertex $i \in S$ and v (i.e. $Y_i \sim$ Bernoulli $\left(\frac{1}{2}\right)$). With $\mathbb{E}[Y] = \frac{|S|}{2} = \frac{k}{2}$ in mind, we can apply a Chernoff bound as follows.

Corollary 3.6. *Fix a vertex $v \notin S$. With probability at least $1 - e^{-.02k}$ over the choice of G, vertex v has at most $.6k$ neighbors in S.*

Proof. Apply a Chernoff bound with $\epsilon = 0.1$. Letting $Y \sim$ Binomial $\left(k, \frac{1}{2}\right)$ be a random variable denoting the number of neighbors of v that are in S as before, we get that

$$\Pr[Y \leq .6k] = 1 - \Pr[Y > .6k] = 1 - \Pr\left[Y - \frac{k}{2} > .1k\right] = 1 - \Pr[Y - \mathbb{E}[Y] > .1k]$$

The last line is greater than or equal to $1 - e^{-2(.1)^2 k} = 1 - e^{-.02k}$ by Chernoff's bound. □

This is for one $v \notin S$ but we want to make sure this holds for *all* $v \notin S$. Let's apply a union bound to complete our analysis.

Corollary 3.7. *With probability at least $1 - (n-k)e^{-.02k}$ over the choice of G, every vertex $v \notin S$ has at most $.6k$ neighbors in S.*

Proof. Let A denote the event that every vertex $v \notin S$ has at most $.6k$ neighbors in S. Additionally define B_i as events where $i \in V$ has more than $.6k$ neighbors in S, we have that. The negation of A is the event where there is at least one vertex $v \notin S$ with more than $.6k$ neighbors in S. This is the union of $n - k$ events B_i thus by the union bound

$$\Pr[A] = 1 - \Pr\left[\bigcup_{i \notin S} B_i\right] \geq 1 - \sum_{i \notin S} \Pr[B_i]$$

However, $\Pr[B_i] = \Pr[Y > .6k]$ from the proof of Corollary 3.6, which is at most $e^{-.02k}$ by Chernoff. Since there are $n - k$ vertices not in S, we have $\Pr[A] \geq 1 - (n-k)e^{-.02k}$ as required. □

With the results of corollary 3.6, we can now verify that $C \subseteq S$ with high probability, and thus $C = S$. Note with probability $1 - e^{-\Omega(\sqrt{n})}$ (since $k \geq \Omega(\sqrt{n})$), each vertex $v \notin S$ has at most $.6k$ neighbors in S. The results from section 3.3.2.1 implies I contains at least $.992k$ of the k vertices of S, thus I contains at most $.008k$ other vertices not in S. Thus, in the worst case, each vertex $u \notin S$ has at most $.608k$ neighbors in I. Since the algorithm sets C to be the set of vertices $v \in V$ that have at least $\frac{3}{4}k$ neighbors in I, the vertices $u \notin S$ will thus not be included in C, completing our analysis.