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 are going to look at an algorithm that solves the planted clique problem. The method we are going to use involves some spectral graph theory. While this note may include many algebraic proofs, the main purpose is to let readers understand intuitively why this algorithm works. Note that this is an algorithm on a random instance, so the analysis is to show that the algorithm is effective with high probability.

3.1.1 Problem

While this problem may have a more generalized form, we are going to discuss one particular variant of this problem:

- let $G = (V, E)$ where $G \sim \mathcal{G}_{n,\frac{1}{2}}$ random graph
- $S \subseteq V$ is a set of vertices selected uniformly at random, and $|S| = k$
- Plant a clique on S, i.e. connect all edges among vertices in S

Here we propose an algorithm that can find a planted clique of size k in a graph $G \sim \mathcal{G}_{n,\frac{1}{2}}$ with high probability when $k >> \sqrt{n}$.

3.1.2 Algorithm

Input: Graph G, with a planted clique of size k

1. $M := A - \frac{1}{2}J$, where A is the adjacency matrix of G, and J is a matrix of all 1s
2. $\mathbf{x}$: eigenvector of the largest eigenvalue of M
3. I: set of k vertices what have the largest values of $|\mathbf{x}_v|$
4. C: set of vertices $v \in V$ that have at least $\frac{3}{4}k$ neighbors in I
5. Return C

The algorithm can be divided into two parts: the first two lines obtain an eigenvector from the given instance; the following lines of algorithm can be seen as a way of rounding to extract the solution from the eigenvector.

3.2 Analysis

Since the algorithm is divided 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 x is very close to solution with high probability. Second, we argue that the way we recover the solution from the eigenvector (the last three lines of algorithm) guarantees a solution with high probability.

3.2.1 Analysis of Retrieved Eigenvector

Recall that S is the set of vertices on which the clique of size k is planted. Let $\mathbf{1}_S$ be an indicator vector that is 1 at position i if vertex i is in the set S. If we can get $\mathbf{1}_S$, then the problem is solved, because that vector encodes the clique we want to find. We want to show that with high probability, the eigenvector $\mathbf{x}$ we get is close to $\mathbf{1}_S$, so that it makes sense to try to work on $\mathbf{x}$ to obtain the solution.

First we want to show that the top eigenvalue of matrix M and the top eigenvalue of $\mathbf{1}_S\mathbf{1}_S^T$, a rank one matrix that encodes the answer, are very close.

Let us look at $\mathbb{E}(A)$:

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

where J is a matrix with 1 everywhere.

- 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 $\mathbb{E}(A_{i,j}) = \frac{1}{2}$. 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 there must exist an edge between all vertices in S, $A_{i,j} = 1$ where $i, j \in S$. Note that the (i,j)th entry of $\mathbf{1}_S\mathbf{1}_S^T$ is exactly one if $i, j \in S$ 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.
- D is a diagonal matrix with $D_{i,i} = 1$ if $i \in S$ and $\frac{1}{2}$ elsewhere. This matrix is used to adjust the diagonal entry of $\mathbb{E}(A)$.

Intuitively, we assume our graph $A \approx \mathbb{E}(A)$, and we ignore the diagnal term for now, we can see $M = A - \frac{1}{2}J \approx \frac{1}{2}\mathbf{1}_S\mathbf{1}_S^T$, which is a one rank one matrix with eigenvector $\mathbf{1}_S$, which is the solution we want.

In order to show the eigenvector of the two matrices are similar, we first want to show that A and $\mathbb{E}(A)$ are "close" to each other. We can use the operator norm to measure the difference between the two matrices. In this case, since both matrices are symmetric, the operator norm is just the spectral norm. In other words, we are trying to show that the largest eigenvalue of the two matrices are close.

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.

We omit the proof here to keep the note simple. We can apply this lemma and bound the "closeness" of M and $\frac{1}{2}\mathbf{1}_S\mathbf{1}_S^T$:

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

That shows the top eigenvalue of the matrix M we defined and the matrix $\frac{1}{2}\mathbf{1}_S\mathbf{1}_S^T$ are very close.

Now, given that the top eigenvalues are close, we would like to show that the eigenvectors of matrix M and $\mathbf{1}_S\mathbf{1}_S^T$ are very close. Essentially, we want to show $\min\{\|\mathbf{x} - \mathbf{1}_S\|^2, \| - \mathbf{x} - \mathbf{1}_S\|^2\}$ is small.

Theorem 3.2 (Davis and Kahan)**.** *If M is a symmetric matrix, yy^T is a rank-one symmetric matrix, and $\mathbf{x}$ is an eigenvector of the largest eigenvalue of M, we have*

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

where $\hat{\mathbf{xy}}$ is the angle between $\mathbf{x}, \mathbf{y}$.

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

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

for sufficiently large n.

Proof. First we find the spectral norm of $\|\frac{1}{2}\mathbf{1}_S\mathbf{1}_S^T\|$. 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 and we see

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

Now we have to reason about the distance between $\mathbf{x}$ and $\mathbf{1}_S$:

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

similarly

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

Therefore we have:

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

Combining the previous results, we have

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

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 very 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 $\mathbf{x}$.

3.2.2 Recovery Step

We will now proceed to analyze the second half of the algorithm, involving I and C and their relationship to S. In particular, to show the correctness of the algorithm, we must show that C contains all the elements of S ($S \subseteq C$) and none of the elements not in S ($C \subseteq S$) with high probability, so that $C = S$ with high probability. We begin by showing the first part, that $S \subseteq C$ with high probability.

3.2.2.1 $S \subseteq C$ 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 $\|\mathbf{x}\|^2 = |S| = k$ and*

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

and if I is the set of k vertices v with largest $|\mathbf{x}_v|$, then I and S have at least $k(1 - 4\epsilon)$ vertices in common.

Proof. To prove this lemma, let's begin by defining *bad* vertices. Intuitively, since we would like I and S to be as similar as possible, bad vertices relative to I and S would be those vertices v such that $v \in I$ but $v \notin S$ or $v \notin I$ but $v \in S$, (or more concisely, $v \in I \oplus S$).

Definition 3.5. *Call a vertex v bad if $v \in S$ and $|\mathbf{x}_v| \leq \frac{1}{2}$ or if $v \notin S$ and $|\mathbf{x}_v| > \frac{1}{2}$. Let B be the set of bad vertices.*

We observe that $|B| \leq 4\epsilon k$ since each vertex of B contributes at least $\frac{1}{4}$ to both $\|\mathbf{x} - \mathbf{1}_S\|^2$ and $\| - \mathbf{x} - \mathbf{1}_S\|^2$. We can see this by considering the boundary cases for the bad vertices and their contributions to $\|\mathbf{x} - \mathbf{1}_S\|^2$ and $\| - \mathbf{x} - \mathbf{1}_S\|^2$. In the first case, where $v \in S$ and $|\mathbf{x}_v| \leq \frac{1}{2}$, the corresponding entry in the indicator

vector of the clique is $\mathbf{1}_{Sv} = 1$ since $v \in S$. Thus, v would then contribute at least $\frac{1}{4}$ to both $\|\mathbf{x} - \mathbf{1}_S\|^2$ and $\| - \mathbf{x} - \mathbf{1}_S\|^2$. In the other case, where $v \notin S$ and $|\mathbf{x}_v| > \frac{1}{2}$, the corresponding entry in the indicator vector of the clique is $\mathbf{1}_{Sv} = 0$ since $v \notin S$. Thus v would also contribute at least $\frac{1}{4}$ to both $\|\mathbf{x} - \mathbf{1}_S\|^2$ and $\| - \mathbf{x} - \mathbf{1}_S\|^2$ in this case as well. Our inequality for $|B|$ thus follows since $\frac{|B|}{4} \leq \min\{\|\mathbf{x} - \mathbf{1}_S\|^2, \| - \mathbf{x} - \mathbf{1}_S\|^2\} \leq \epsilon k$, so that $|B| \leq 4\epsilon k$.

The next observation we can make is that $|I \cap S| \geq k - |B|$, or that I and S must have at least $k - |B|$ vertices in common. To show this, we can also similarly consider 2 different cases corresponding to the 2 cases for the bad vertices. Let $t = \min_{v \in I} |x_v|$. If $t > \frac{1}{2}$, then all vertices v such that $v \notin S$ but $v \in I$ must have $|x_v| > \frac{1}{2}$ (since t is the minimum $|x_v|$ for vertices $v \in I$), and so they are bad vertices. However, I can contain at most $|B|$ of these bad vertices, and thus must contain at least $k - |B|$ vertices from S (since $|I| = |S| = k$). Similar reasoning follows for the second case as well. If $t \leq \frac{1}{2}$, then every vertex v such that $v \in S$ but $v \notin I$ must have $|x_v| \leq \frac{1}{2}$, and so is a bad vertex. However, this must mean that at least $k - |B|$ of the remaining vertices in S are also included in I (since they have $|x_v| \geq t$). Thus, $|I \cap S| \geq k - |B| \geq k(1 - 4\epsilon)$. □

Using the previous results from Corollary 3.3, we can use $\epsilon = .002$, and with Lemma 3.4, it follows that with high probability, I will contain at least $k - |B| \geq k - 4\epsilon k = .992k$ of the k vertices of S.

This takes care of the analysis of the 3rd line of the algorithm regarding I. We can now proceed to see how this property of I can be used to extract desired solution, C, recovering the planted clique. To begin, note that if I contains at least $.992k$ of the k vertices of S, then C will include of the vertices in S since all vertices in S have at least $.992k > \frac{3}{4}k$ neighbors in I. Thus, with high probability, we have that $S \subseteq C$, as we wanted to show. We now move onto showing the next part, that $C \subseteq S$ with high probability.

3.2.2.2 $C \subseteq S$ with high probability

To show this part, we will require the use of the Chernoff bound, in particular that for binomial random variables.

Theorem 3.6 (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 $0 < \epsilon < 1$*

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

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 with $\mathbb{E}[Y] = \frac{k}{2}$, and $Y = \sum_{i=1}^{k} Y_i$, where each individual $Y_i \sim \text{Bernoulli}\left(\frac{1}{2}\right)$, such that each Y_i is an indicator random variable for the corresponding edge between a vertex $i \in S$ and v. Thus we can apply the Chernoff bound for binomial random variables (Theorem 3.6) as follows.

Corollary 3.7. *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. We can apply Theorem 3.6 with $\epsilon = .1$. Letting $Y \sim \text{Binomial}\left(k, \frac{1}{2}\right)$ be a random variable denoting the number of neighbors of v that are in S as before, we get that

$$\begin{aligned}\Pr[Y \leq .6k] &= 1 - \Pr[Y > .6k] \\ &= 1 - \Pr\left[Y - \frac{k}{2} > .1k\right] \\ &= 1 - \Pr[Y - \mathbb{E}[Y] > .1k] \\ &\geq 1 - e^{-2(.1)^2 k} = 1 - e^{-.02k}\end{aligned}$$

□

We can additionally apply the union bound for an analysis of all vertices not in S.

Corollary 3.8. *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. To utilize the union bound, let's consider the negation of the event that we are interested in analyzing. This is the event that there is at least one vertex $v \notin S$ with more than $.6k$ neighbors in S, which is the union of $n - k$ events, where each individual event is the event that one of the vertices $v \notin S$ has more than $.6k$ neighbors in S. This is precisely $\Pr[Y > .6k]$ from the previous section for the proof of Corollary 3.7, which is at most $e^{-.02k}$ by Theorem 3.6. Thus, the probability of the negation of the desired event is at most $(n-k)e^{-.02k}$, by the union bound, so that the probability of the actual desired event is at least $1 - (n-k)e^{-.02k}$.

More formally, let A denote the event we are interested in, that every vertex $v \notin S$ has at most $.6k$ neighbors in S. Then, if B_i denotes the event that a vertex i has more than $.6k$ neighbors in S, we have that

$$\begin{aligned}\Pr[A] &= 1 - \Pr\left[\bigcup_{i \notin S} B_i\right] \\ &\geq 1 - \sum_{i \notin S} \Pr[B_i], && \text{by the union bound} \\ &\geq 1 - (n-k)e^{-.02k}\end{aligned}$$

□

With the results of Corollary 3.7, we can now verify that $C \subseteq S$ with high probability, and thus that $C = S$ with high probability (given the results from the previous section that $S \subseteq C$ with high probability). To do this, we first note that with probability $1 - e^{-\Omega\sqrt{n}}$ (since $k \in \Omega(\sqrt{n})$), each vertex $u \notin S$ has at most $.6k$ neighbors in S. With the results from 3.2.2.1, we know that I contains at least $.992k$ of the k vertices of S, and so I contains at most $.008k$ other vertices not in S. Thus, in the worst case, each vertex $v \notin S$ 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, as desired. As such, in addition to the results from 3.2.2.1, we have with high probability that $C \subseteq S$, and thus $C = S$ with high probability, as we wanted.