Graph Theory

Finding a planted clique

Instructor: *Nicolò Cesa-Bianchi* version of May 5, 2024

The material in this handout is taken from: Luca Trevisan, Handout 7 of Beyond Worst-Case Analysis, 2017.

Social networks are naturally modeled as undrected graphs, where the presence of an edge between two individuals denotes friendship or shared interests. In social network analysis, one often wants to find communities, defined as subsets of individuals who form dense clusters in the graph. A stylized version of this task is the problem of finding a large clique in a random graph. Take $G_n \sim \mathcal{G}(n, \frac{1}{2})$ and suppose a clique of size $k \gg 2\log_2 n$ is "planted" in G_n by adding all missing edges in an arbitrary subset S of $k < n$ vertices, where the subset S does not depend on the realization of the random edges of G_n. Let $G = (V, E)$ be the resulting graph. We want to know whether there exists an efficient algorithm to find this clique. In summary, we assume G is built according to the following process:

1. A subset $S \subset V$ of size k is selected
2. $G_n \sim \mathcal{G}(n, \frac{1}{2})$ is drawn
3. G is builts by adding to G_n all missing edges between pairs of vertices in S.

We know that $\omega(G_n) \xrightarrow{p} 2\log_2 n$. Therefore, the problem of finding the planted clique is meaningful when $k > 3\log_2 n$ because—with high probability—there are no random cliques of size larger than $3\log_2 n$. The simplest algorithm to find the planted clique checks every subset of V of size $k > 3\log_2 n$ until a clique is found. The running time of this algorithm is $\binom{n}{k}k^2 \geq \left(\frac{n}{k}\right)^k k^2$, and thus exponential in $k = \Omega(\log n)$. A better, but still inefficient algorithm is the following.

Algorithm 1 (Inefficient Clique Finder)

Input: Graph $G = (V, E)$, integer $k > 3\log_2 n$.

for all $S \subset V$ with $|S| = 3\log_2 n$ **do**
  if S is a clique **then**
    Let $T \subseteq V \setminus S$ be the set of vertices adjacent to all vertices in S
    if $T \cup S$ is a clique larger than k **then**
      Return T
    end if
  end if
end for
Return the empty set

Output: A clique of size at least k or the empty set.

The running time of this algorithm is $\mathcal{O}\big(n^{\log n}(\log n)^2\big)$ for any $k = \Omega(\log n)$, which is quasi-polynomial in n.

Algorithm 2 (Clique Finder)

Input: Graph $G = (V, E)$.

Let $v_1, \ldots, n_n$ be an ordering of the vertices in V such that $d(v_1) \geq d(v_2) \geq \cdots \geq d(v_n)$.
for $k = n, n-1, \ldots, 3\log_2 n$ **do**
    if $\{v_1, \ldots, v_k\}$ form a clique **then**
        Return $\{v_1, \ldots, v_k\}$
    end if
end for
Return the empty set

Output: A clique of size at least $3\log_2 n$ or the empty set.

A simple and efficient algorithm is Algorithm 2.

In order to analyze this algorithm, we look at the distribution of the degree in $\mathcal{G}(n, \frac{1}{2})$. Recall the following special case of Chernoff-Höffding bounds.

Lemma 1 *Let N have a binomial distribution of parameters n, p. Then, for any $0 < \delta < 1$,*

$$N \leq np + \sqrt{\frac{n}{2}\ln\frac{1}{\delta}} \qquad \text{and} \qquad N \geq np - \sqrt{\frac{n}{2}\ln\frac{1}{\delta}}$$

hold with probability at least $1 - \delta$.

Now note that if $G_n \sim \mathcal{G}(n, \frac{1}{2})$, then $d_{G_n}(v)$ is a binomial random variable of parameters $n-1, \frac{1}{2}$ for each vertex v. Therefore, using a union bound over the n vertices, we get that with probability at least $1 - \frac{\delta}{2}$

$$\max_{v \in V} d_{G_n}(v) \leq \frac{n-1}{2} + \sqrt{\frac{n-1}{2}\ln\frac{2n}{\delta}} \tag{1}$$

Let S with $|S| = k$ be the subset of vertices where the clique is planted. Each vertex in S in G_n receives an average of $\frac{k-1}{2}$ random edges from the other vertices in S. By using Chernoff-Höffding bounds again, we see that with probability at least $1 - \frac{\delta}{2}$ the number $\left|N_{G_n}(v) \cap S\right|$ of these random edges from S in G_n satisfies

$$\max_{v \in S} \left|N_{G_n}(v) \cap S\right| \leq \frac{k-1}{2} + \sqrt{\frac{k-1}{2}\ln\frac{2k}{\delta}}$$

Since each $v \in S$ has actually $k-1$ edges in G, this implies that, with high probability, at least

$$\frac{k-1}{2} - \sqrt{\frac{k-1}{2}\ln\frac{2k}{\delta}}$$

edges are added to each $v \in S$. Hence, with probability at least $1 - \delta$,

$$\begin{aligned}\min_{v \in S} d_G(v) &\geq \underbrace{\frac{n-1}{2} - \sqrt{\frac{n-1}{2}\ln\frac{2n}{\delta}}}_{\text{random edges}} + \underbrace{\frac{k-1}{2} - \sqrt{\frac{k-1}{2}\ln\frac{2k}{\delta}}}_{\text{added edges}} \\ &\geq \frac{n-1}{2} + \frac{k-1}{2} - 2\sqrt{\frac{n-1}{2}\ln\frac{2n}{\delta}}\end{aligned}$$

Hence, if

$$\frac{k-1}{2} = 4\sqrt{\frac{n-1}{2}\ln\frac{2n}{\delta}}$$

then with probability at least $1-\delta$

$$\min_{v\in S} d_G(v) \geq \frac{n-1}{2} + 2\sqrt{\frac{n-1}{2}\ln\frac{2n}{\delta}}$$

$$\max_{v\in V\setminus S} d_G(v) \leq \max_{v\in V} d_{G_n}(v) \leq \frac{n-1}{2} + \sqrt{\frac{n-1}{2}\ln\frac{2n}{\delta}} \qquad \text{(using (1))}$$

implying that, with the same probability, the k highest degree vertices in G belong to S.

We just saw that it is easy to find a hidden clique of size $k = \Omega\big(\sqrt{n\ln n}\big)$. We now look at the case $k = \Omega(\sqrt{n})$. Here we hide the dependence on $\ln\frac{1}{\delta}$ and simple say that any result holds with high probability (w.h.p.).

We need the following result (stated without proof). Let $J_n = \mathbf{1}^\top\mathbf{1}$ the all-one $n\times n$ matrix and let $\lambda_{\max}(M)$ be the largest eigenvalue of a symmetric matrix M. Note that $\mathbf{1}^\top M\mathbf{1} = \sum_{i,j} M_{i,j}$.

Lemma 2 *If A_n is the adiancency matrix of $G_n \sim \mathcal{G}\big(n,\frac{1}{2}\big)$, then $\lambda_{\max}\big(A_n - \frac{1}{2}J_n\big) \leq 2\sqrt{n}$ with high probability.*

The proof of Lemma 2 shows that $\mathbf{1}$ is close to the eigenvector of the largest eigenvalue. Now note that for all $1\leq i,j\leq n$, the random variables $X_{i,j} = A_n(i,j) - \frac{1}{2}$ are i.i.d. with zero expected value and variance equal to $\frac{1}{4}$. Then, by Chernoff-Hoeffding bounds, $(A_n\mathbf{1})_i = \sum_j X_{i,j} = \mathcal{O}\big(\sqrt{n}\big)$ w.h.p. for any i. The lemma then shows that $A\mathbf{1} \leq \big(2\sqrt{n}\big)\mathbf{1}$ w.h.p.

If A is now the adiacency matrix of G with a planted clique of size k in S and $\mathbf{1}_S\in\{0,1\}^n$ has nonzero components only on coordinates corresponding to elements of S, we have that

$$\begin{aligned}
\lambda_{\max}\left(A - \frac{J_n}{2}\right) &= \max_{\boldsymbol{x}\,:\,\boldsymbol{x}\neq\mathbf{0}} \frac{\boldsymbol{x}^\top\big(A-\frac{1}{2}J_n\big)\boldsymbol{x}}{\boldsymbol{x}^\top\boldsymbol{x}} \\
&\geq \frac{\mathbf{1}_S^\top\big(A-\frac{1}{2}J_n\big)\mathbf{1}_S}{\mathbf{1}_S^\top\mathbf{1}_S} \\
&\geq \frac{\mathbf{1}_S^\top A\mathbf{1}_S - \frac{1}{2}\mathbf{1}_S^\top J_n\mathbf{1}_S}{k} \\
&= \frac{k(k-1)-\frac{1}{2}k^2}{k} \\
&= \frac{k}{2} - 1
\end{aligned}$$

where we used

$$\mathbf{1}_S^\top A\mathbf{1}_S = \sum_{i\in S}\sum_{j\in S} A_{i,j} = k(k-1) \qquad \mathbf{1}_S^\top J_n\mathbf{1}_S = \sum_{i\in S}\sum_{j\in S} 1 = k^2$$

Therefore, if we pick $\frac{k}{2}-1 \geq 4\sqrt{n}$, which is equivalent to $|S| = \Omega\big(\sqrt{n}\big)$, then Lemma 2 tells us that we can distinguish G with a planted clique from $G_n\sim\mathcal{G}\big(n,\frac{1}{2}\big)$. However, we do not know yet how to find S.

The intuition behind Algorithm 3, which finds S with high probability, is the following. Let L be the top k components (in absolute value) of the eigenvector of the largest eigenvalue of $A - \frac{1}{2}J_n$. Then we show that L contains at least $\frac{3}{4}$ of the elements in S. Hence, any element of S has at least $\frac{3k}{4}$ components in L while Lemma 1 implies that all $v \notin S$ have at most $\frac{k}{2} + \mathcal{O}\big(\sqrt{k \ln n}\big)$ neighbors in L with high probability.

Now let $A = A_n + A_S$ where A_n is the adiacency matrix of $G_n \sim \mathcal{G}\big(n, \frac{1}{2}\big)$ and A_S is the $n \times n$ adiacency matrix only containing the edges missing from G_n to form a clique on S. Note that the probability that A_S contains an edge between any two vertices in S is $\frac{1}{2}$, which is the probability that edge is missing from G_n. Therefore, we can think of A_S as the adiacency matrix of a graph $G_k \sim \mathcal{G}\big(k, \frac{1}{2}\big)$ where, clearly, G_k and G_n are dependent. Using Lemma 2 and the union bound, we conclude that w.h.p.,

$$\lambda_{\max}\left(A_n - \frac{J_n}{2}\right) \leq 2\sqrt{n} \tag{2}$$

$$\lambda_{\max}\left(A_S - \frac{J_S}{2}\right) \leq 2\sqrt{k} \tag{3}$$

where $J_S = \mathbf{1}_S^\top \mathbf{1}_S$. Therefore, if $\boldsymbol{x}$ is any eigenvector for the largest eigenvalue of $A - \frac{J_n}{2}$,

$$\begin{aligned}
\frac{k}{2} - 1 &\leq \boldsymbol{x}^\top \left(A - \frac{J_n}{2}\right) \boldsymbol{x} \\
&= \boldsymbol{x}^\top A_S \boldsymbol{x} + \boldsymbol{x}^\top \left(A_n - \frac{J_n}{2}\right) \boldsymbol{x} \\
&\leq \boldsymbol{x}^\top A_S \boldsymbol{x} + 2\sqrt{n} && \text{(holds w.h.p. by (2))}
\end{aligned}$$

implying

$$\boldsymbol{x}^\top A_S \boldsymbol{x} \geq \frac{k}{2} - 1 - 2\sqrt{n}$$

Now,

$$\begin{aligned}
\frac{k}{2} - 1 - 2\sqrt{n} &\leq \boldsymbol{x}^\top \left(A_S - \frac{J_S}{2} + \frac{J_S}{2}\right) \boldsymbol{x} \\
&= \boldsymbol{x}^\top \left(A_S - \frac{J_S}{2}\right) \boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^\top \big(\mathbf{1}_S \mathbf{1}_S^\top\big)\boldsymbol{x} && \text{(since } J_S = \mathbf{1}_S \mathbf{1}_S^\top\text{)} \\
&\leq \lambda_{\max}\left(A_S - \frac{J_S}{2}\right) + \frac{1}{2}\big(\mathbf{1}_S^\top \boldsymbol{x}\big)^2 \\
&\leq 2\sqrt{k} + \frac{1}{2}\big(\mathbf{1}_S^\top \boldsymbol{x}\big)^2 && \text{(holds w.h.p. by (3))}
\end{aligned}$$

Therefore,

$$\big(\mathbf{1}_S^\top \boldsymbol{x}\big)^2 \geq k - 2 - 4\sqrt{n} - 4\sqrt{k}$$

Recall that $\|\boldsymbol{x}\| = 1$ because $\boldsymbol{x}$ is an eigenvalue and let $\boldsymbol{y}$ be such that $y_i = |x_i|$ for all $i = 1, \ldots, n$ (so that also $\|\boldsymbol{y}\| = 1$). By choosing $k \geq 10\sqrt{n}$ we can ensure $\mathbf{1}_S^\top \boldsymbol{y} \geq \big|\mathbf{1}_S^\top \boldsymbol{x}\big| \geq \frac{15}{16}\sqrt{k}$ for $n \geq 256$. this gives

$$\Big\|\mathbf{1}_S - \sqrt{k}\boldsymbol{y}\Big\|^2 = k + k\,\|\boldsymbol{y}\|^2 - 2\sqrt{k}\mathbf{1}_S^\top \boldsymbol{y} \leq 2k - 2\frac{15k}{16} = \frac{2k}{16} \tag{4}$$

This shows that $\sqrt{k}\boldsymbol{y}$ is close to $\mathbf{1}_S$. We can use this insight to devise an algorithm (Algorithm 3) that, with high probability, finds the hidden clique given G and k as input.

Algorithm 3 (Improved Clique Finder)

Input: Graph $G = (V, E)$, $k \in \mathbb{N}$.

1: Compute the adiacency matrix A of G
2: Let $\boldsymbol{x}$ the eigenvector of largest eigenvalue for $A - \frac{1}{2}J_n$
3: Let L be the set of k vertices i with largest $|x_i|$

Output: The set of k vertices in V with the largest number of neighbors in L.

Let L be the set of vertices computed in line 3 of Algorithm 3. Since $\|\boldsymbol{y}\| = 1$, the smallest y_i such that $i \in L$ with $|L| = k$ must satisfy $y_i \leq \frac{1}{\sqrt{k}}$, which implies $\min_{j \in L} \sqrt{k} y_j \leq 1$. Therefore, there exists $0 < t \leq 1$ such that $i \in L$ if and only if $\sqrt{k} y_i \geq t$. Let $m = |S \setminus L|$. Therefore, m elements of S are missing from L. Since $|L| = k$, L also contains m elements that are not in S. Therefore,

$$\begin{aligned}
\left\| \mathbf{1}_S - \sqrt{k}\boldsymbol{y} \right\|^2 &= \sum_{i \in S} \left(1 - \sqrt{k} y_i\right)^2 + \sum_{j \notin S} k y_j^2 \\
&\geq \sum_{i \in S \setminus L} \left(1 - \sqrt{k} y_i\right)^2 + \sum_{j \in L \setminus S} k y_j^2 \\
&\geq m(1-t)^2 + mt^2 \qquad \text{(because } 0 \leq \sqrt{k} y_i < t \leq 1 \text{ and } \sqrt{k} y_j > t > 0\text{)} \\
&\geq \frac{m}{2}
\end{aligned}$$

Combining with (4) we get $m \leq \frac{4k}{16}$ and so L contains at least $\frac{3k}{4}$ elements of S. Hence, any $v \in S$ has at least $\frac{3k}{4}$ neighbours in L. On the other hand, Lemma 1 implies that

$$\max_{v \in V \setminus S} \left| N(v) \cap L \right| \leq \frac{k}{2} + \sqrt{\frac{k}{4} \ln n}$$

with high probability.