

## Finding a planted clique

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

Social networks are naturally modeled as undirected 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. This task can be modeled as 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 built by adding to  $G_n$  all missing edges between pairs of vertices in  $S$ .

We know that  $\omega(G_n) = 2 \log_2 n$  asymptotically holds for almost all  $G_n \sim \mathcal{G}(n, \frac{1}{2})$ . Therefore, the problem of finding the planted clique is meaningful when  $k \gg 2 \log_2 n$ , implying that—with high probability—there are no random cliques of size  $k$ . The simplest algorithm to find the planted clique of size  $k$  checks every subset of  $V$  of size  $k$  until a clique is found. The running time of this algorithm is  $\binom{n}{3 \log_2 n} k^2$ , which is of order  $\left(\frac{n}{k}\right)^k k^2$  and thus exponential in  $k = \Omega(\log n)$ . A better, but still inefficient algorithm is the following.

---

**Algorithm 1** (Simple Clique Finder)

---

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

- 1: **for** all  $S \subset V$  with  $|S| = 3 \log_2 n$  **do**
- 2:     **if**  $S$  is a clique **then**
- 3:         Let  $T \subseteq V \setminus S$  be the set of vertices adjacent to all vertices in  $S$
- 4:         **if**  $T \cup S$  is a clique larger than  $k$  **then**
- 5:             Return  $T$
- 6:         **end if**
- 7:     **end if**
- 8: **end for**
- 9: Return the empty set

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

---

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

A simple and efficient algorithm ranks the vertices in descending order of their degree, and picks the  $k$  highest-degree vertices, where  $k$  is the largest integer such that the first  $k$  vertices in the ranking form a clique. 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-Höfding bounds.

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

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

*each 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  $1 - \frac{\delta}{2}$

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

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-Höfding bounds again, we see that with probability at least  $1 - \frac{\delta}{2}$  the number  $|N_{G_n}(v) \cap S|$  of these random edges from  $S$  in  $G_n$  satisfies

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

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  $1 - \delta$

$$\begin{aligned} \min_{v \in S} d_G(v) &\geq \frac{n-1}{2} + 2\sqrt{\frac{n-1}{2} \ln \frac{2n}{\delta}} \\ \max_{v \in V \setminus S} d_G(v) &= \max_{v \in V} d_{G_n}(v) \leq \frac{n-1}{2} + \sqrt{\frac{n-1}{2} \ln \frac{2n}{\delta}} \quad (\text{using (1)}) \end{aligned}$$

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  $k = \Omega(\sqrt{n \ln n})$ . We now look at the case  $k = \Omega(\sqrt{n})$ . Here we hide the dependence on  $\ln \frac{1}{\delta}$  and simply 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  $A$  is the adjacency matrix of  $G \sim \mathcal{G}(n, \frac{1}{2})$ , then  $\lambda_{\max}(A - \frac{1}{2}J_n) \leq 2\sqrt{n}$  with high probability.*

If  $A$  is now the adjacency 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

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

where we used

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

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

Now let  $A = A_n + A_S$  where  $A_n$  is the adjacency matrix of  $G_n \sim \mathcal{G}(n, \frac{1}{2})$  and  $A_S$  is the  $n \times n$  adjacency 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 adjacency matrix of a graph  $G_k \sim \mathcal{G}(k, \frac{1}{2})$  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  $\mathbf{x}$  is any eigenvector for the largest eigenvalue of  $A - \frac{J_n}{2}$ ,

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

implying

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

Now,

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

Therefore,

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

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

$$\left\| \mathbf{1}_S - \sqrt{k} \mathbf{y} \right\|^2 = k + k \|\mathbf{y}\|^2 - 2\sqrt{k} \mathbf{1}_S^\top \mathbf{y} \leq 2k - 2\frac{15k}{16} = \frac{2k}{16} \quad (4)$$

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

---

**Algorithm 2** (Clique Finder)

---

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

- 1: Compute the adjacency matrix  $A$  of  $G$
- 2: Let  $\mathbf{x}$  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 vertices in  $V$  with at least  $\left(\frac{3}{4} - \frac{1}{8}\right)k$  neighbors in  $L$ .

---

Let  $t$  be such that  $i \in L$  if and only if  $\sqrt{k}y_i \geq t$  (note that  $t$  can always be chosen such that  $t \in [0, 1]$ ; indeed, since  $\|\mathbf{y}\| = 1$ , the smallest  $y_i$  such that  $i \in L$  with  $|L| = k$  must satisfy  $y_i \leq \frac{1}{\sqrt{k}}$ , which implies  $\sqrt{k}y_i \leq 1$ ). 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,

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

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 any  $v \notin S$  with high probability has at most  $\frac{k}{2} + \sqrt{k \ln n}$  neighbors in  $L$ .