Planted Clique Model

Recall : . First choose a subset K of k vertices uniformly at random from $[n]$ to form a clique.

- The remaing pairs of nodes are connected indep. w.p. $\frac{1}{2}$

In other words

$$P\{ i \sim j \} = \begin{cases} 1 & \text{if } i,j \in K \\ 0 & \text{o.w.} \end{cases} \qquad \text{for } \forall\, i \neq j$$

let $G \sim G(n, \frac{1}{2}, k)$.

Note that $G(n, \frac{1}{2}, 0) = G(n, \frac{1}{2})$ (Erdös-Rényi)

<u>Question</u>: When can we recover the planted clique ?

- If $k \geq (2+\varepsilon)\log_2 n$, then the planted clique is the unique clique of size k whp, and thus we can find it usg exhaustie search in $\binom{n}{k} \sim n^k$ time.

- <u>Intuition but not fully rigorous</u>:

  $\omega(G(n, \frac{1}{2})) < (2+\varepsilon)\log_2 n$ whp as we've shown.

- Not fully rigourous, because we may find a larger clique usg nodes in K and nodes outside K.

- rigorous proof (HW).

- If $k \le (2-\varepsilon)\log_2 n$, impossible to recover the planted clique, as there is $n^{\Omega(\log n)}$ many different cliques of size k in $G(n, \frac{1}{2})$

Question:

How to efficiently recover the planted clique?

Answer: The best known poly-time algorithms are able to recover the planted clique whp if $k = \Omega(\sqrt{n})$.

Remark:

planted clique:	statistical	computational
gap is significantly larger.	$2\log_2 n$	$\Omega(\sqrt{n})$
Max clique in $G(n, \frac{1}{2})$:	$2\log_2 n$	$\log_2 n$.

Next: We will study three efficient algorithms

- Degree test: works for $k = \Omega(\sqrt{n \log n})$
- Iterated Degree test: works for $k = C\sqrt{n}$ for some large const C in linear time.
- Spectral method: for $k \ge c\sqrt{n}$ for any c in time $n^{O(\frac{1}{c^2})}$

- Semi-definite programy (SDP): for $k \gtrsim C\sqrt{n}$.

  Additional benefits of robustness to model perturbation.

- Exercise: (update this) To find k-clique in $G(n, \frac{1}{2}, k)$, exhaustive search takes $\binom{n}{k} \approx n^k$ time. There is a $n^{O(\log n)}$-time algorithm for all $k \geq C \log_2 n$ for a sufficiently large const $C > 0$.

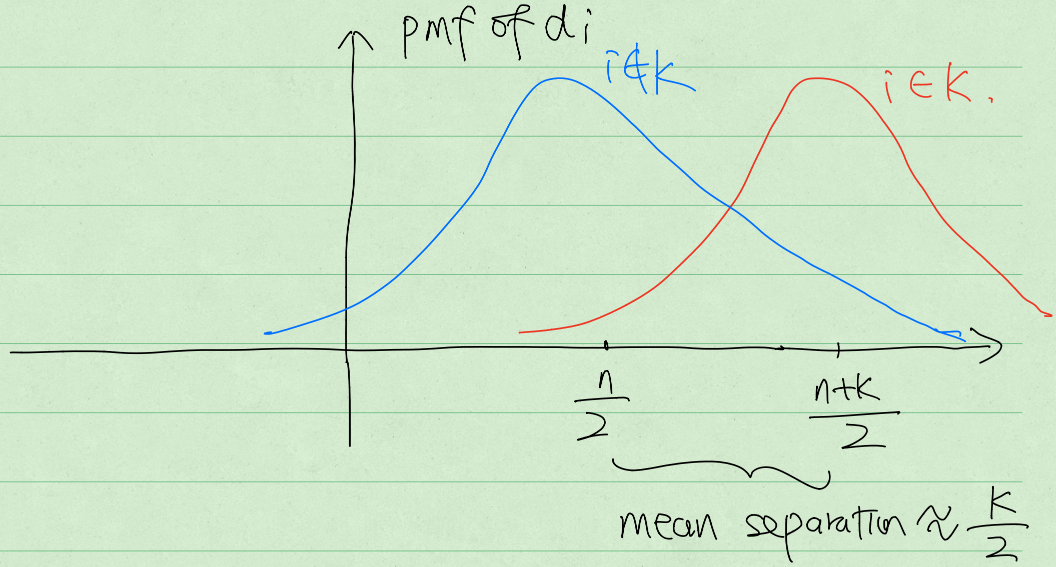
- Degree test:

- Declare the k vertices with the largest degrees as the estimated clique

- Intuition: let d_i = degree of node i in G

$$d_i \sim \begin{cases} Bin(n-1, \frac{1}{2}) & \text{if } i \notin K \\ k-1 + Bin(n-k, \frac{1}{2}) & \text{if } i \in K \end{cases}$$

$$\mathbb{E}[d_i] \approx \begin{cases} \frac{n}{2} & \text{if } i \notin K \\ \frac{n+k}{2} & \text{if } i \in K. \end{cases}$$



For the two distributions to be mostly separated, we need.

$$\underbrace{\text{mean difference}}_{\approx \frac{k}{2}} \gtrsim \underbrace{\text{std of two distribution.}}_{\approx \sqrt{n}}$$

i.e., we need $\quad k \gtrsim \sqrt{n}$.

- The extra $\sqrt{\log n}$ factor is needed to accomodate the possibility that some nodes outside K will have atypically high degrees, and some nodes in K will have atypically low degrees.

- Formally, let $\hat{K}$ = set of k vertices with the highest degree.

<u>Thm</u>: $\mathbb{P}(\hat{K}=K) \to 1$ as $n\to\infty$, if

$K \geq C\sqrt{n\log n}$ for some absolute constant $C>0$.

<u>Remark</u>: It sufficies to show whp

$$\min_{i\in K} d_i > \max_{i\notin K} d_i .$$

- preliminaries on Gaussian Approx and Concentration ineq.

<u>Lemma 1</u>: Suppose $X_1, \ldots X_n \sim N(0,1)$ (may not be independent). Then

$$X_{max} = \max_{1\le i\le n} X_i \le \sqrt{(2+\varepsilon)\log n} \text{ whp}$$

<u>Pf</u>: $P(X_{max} > t) \le n P(X_1 > t) \le n Q(t)$

$$\le n e^{-\frac{t^2}{2}}, \quad t>0$$

$$\to 0$$

if $t = \sqrt{(2+\varepsilon)\log n}$

<u>Note</u>: Lemma 1 is almost tight when X_i's are independent.

Gaussian approx of Binomial. Heuristic

For $i \notin K$,

$$d_i \sim \text{Binom}(n-1, 1/2) \approx N\left(\frac{n}{2}, \frac{n}{4}\right)$$

Lemma 1 $\Rightarrow$ $\max_{i \notin K} d_i \leq \frac{n}{2} + \sqrt{(2+\varepsilon)\log(n-k)\frac{n}{4}}$

$$\approx \frac{n}{2} + \frac{1}{2}\sqrt{(2+\varepsilon)\log n}.$$

For $i \in K$,

$$d_i \sim k-1 + \text{Bin}(n-k, 1/2)$$

$$\approx N\left(\frac{n+k}{2}, \frac{n-k}{4}\right) = \frac{n+k}{2} + N\left(0, \frac{n-k}{4}\right)$$

Lemma 1 $\Rightarrow$ $\min_{i \in K} d_i \geq \frac{n+k}{2} - \sqrt{(2+\varepsilon)\frac{n-k}{4}\log k}$

$$\geq \frac{n+k}{2} - \frac{1}{2}\sqrt{(2+\varepsilon)\, n \log n}.$$

Thus, $\max_{i \notin K} d_i < \min_{i \in K} d_i$, if

$k \geq \sqrt{C n \log n}$ for some large constant $C > 0$.

To Justify this Gaussian Approx. Heuristic,

We will use Hoeffding's inequality. (proved later in Lecture 4)

Lemma 2: (Hoeffding's inequality)

let $S = X_1 + \cdots + X_n$, where X_i's are indep.

and $a \le X_i \le b$. Then

$$\mathbb{P}\{|S - \mathbb{E}[S]| \ge t\} \le 2e^{-\frac{2t^2}{n(b-a)^2}}$$

Corollary 1:

If $S \sim \mathrm{Bin}(n,p)$, then

$$\mathbb{P}\{|S - np| \ge t\} \le 2e^{-\frac{2t^2}{n}}.$$

Let's apply Corollary 1 to analyze degree test:

- For $i \notin K$, $d_i \sim \mathrm{Bin}(n-1, \frac{1}{2})$

$$\mathbb{P}\{d_i \ge \frac{n-1}{2} + t\} \le 2e^{-\frac{2t^2}{n}}$$

$$\Rightarrow \mathbb{P}\{\max_{i \notin K} d_i \ge \frac{n-1}{2} + t\} \le (n-k)\mathbb{P}\{d_i \ge \frac{n-1}{2} + t\}$$
$$= 2(n-k)e^{-\frac{2t^2}{n-1}}$$

- For $i \in K$, $d_i \sim k-1+\mathrm{Bin}(n-k, \frac{1}{2})$

$2\tilde{t}^2$

$$\Rightarrow \mathbb{P}\{ d_i \le k-1+\frac{n-k}{2} - \tilde{t} \} \le 2e^{-\frac{\cdot}{n}}$$

$$\Rightarrow \mathbb{P}\{ \min_{i\in K} d_i \le k-1+\frac{n-k}{2} - \tilde{t} \} \le 2k\, e^{-\frac{2\tilde{t}^2}{n}}.$$

Set $t = \frac{k-1}{4} = \tilde{t}$. Then

$$\max_{i\in K} d_i < \frac{n-1}{2} + \frac{k-1}{4} = \frac{n-k}{2} + \frac{3k-1}{4} < \min_{i\in K} d_i$$

w.p. $1 - 2n e^{-\frac{k^2}{8n}} \to 1$ if $k \ge C\sqrt{n\log n}$

for large const $C>0$.

For analyzing the iterated degree test, we need to quantify the approx of Bino by Gaus in Kolmogorov – Smirnov distance.

- <u>Thm</u> [Berry – Esseen Thm]

  There exists $C = C(p)$ s.t.

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}\{ \mathrm{Bin}(n,p) \le x \} - \mathbb{P}\{ N(np, np(1-p)) \le x \} \right|$$

$$\le \frac{C(p)}{\sqrt{n}}$$

  where $C(p) = O\left(\frac{1}{\sqrt{p}}\right)$

- Iterating the degree test : [Dekel – Gurel – Gurevich – Peres] DGGP

  is able to find clique K in $O(n^2)$ times when $k = |K| = C\sqrt{n}$ for some large const C.

Idea 1. relative size $= \dfrac{k^2}{n} = C^2$

- Degree test needs $\dfrac{k^2}{n} \gtrsim \log n$

- Can we subsample the graph in a clever way so that after subsampling, relative size increases to $\log n$?

- A: • Blindly subsample each vertex w.p. τ does not work, as

$$\begin{aligned} n &\to n\tau \\ k &\to k\tau \end{aligned} \quad \Rightarrow \text{relative size} = \frac{(k\tau)^2}{n\tau} = \frac{k^2}{n}\cdot\tau$$

decreases!

• How about sampling that prefers large degree vertices.

Aim:

$$\begin{aligned} n &\to n\tau \\ k &\to \rho k \end{aligned} \quad \text{where } \rho > \sqrt{\tau}.$$

Then relative size $= \dfrac{(\rho k)^2}{n\tau} = \underbrace{\boxed{\dfrac{\rho^2}{\tau}}}_{>1} \dfrac{k^2}{n}$

needs roughly $\log\log n$ subsampling steps, so that relative size passes $\log n$ threshold.

Specifically, generate a seq. of graph

$$G = G_0 \supset G_1 \supset \cdots \supset G_T, \text{ so that}$$

each instance $G_t \sim \mathcal{G}(n_t, \tfrac{1}{2}, k_t)$

with

$$\begin{aligned} n_t &\approx n\cdot\tau^t \triangleq n_t' \\ k_t &\approx k\rho^t \triangleq k_t' \end{aligned}$$

$\left.\begin{aligned} \alpha &= 0.3728 \\ \beta &= 0.72. \\ C \geq C_0 &\approx 1.65 \end{aligned}\right\} \Rightarrow \tau \approx 0.14287$ and $\rho \approx 0.38$

where $\tau = (1-\alpha)\, Q(\beta)$ and $\rho = (1-\alpha)\, Q(\beta - c\sqrt{\alpha})$

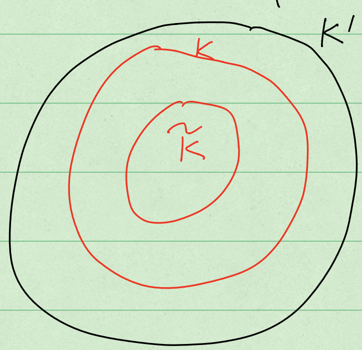
$\frac{\rho^2}{\tau} = (1-\alpha) \frac{Q^2(\beta - c\sqrt{\alpha})}{Q(\beta)} > 1$ choose c sufficiently large $\exists$ $\alpha \in (0,1), \beta > 0$ and $\frac{\rho}{\sqrt{\tau}} \approx 1.00003$

- **Idea 2**: We will end up with recoving just a subset of the hidden clique K.

Q: How to recove the entire clique?

A: use the recoved clique as a "seed set"

let $\tilde{K} \subseteq K$ be the subset of the hidden clique that is found,

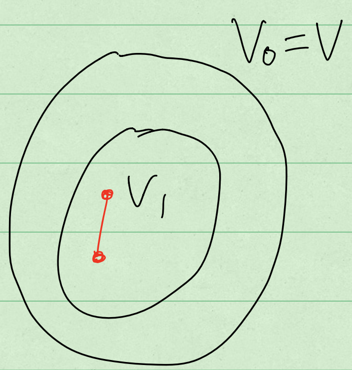


(Blow-up) $K' = \tilde{K} \cup \{\text{common neighbors of nodes in } \tilde{K}\}$

(K' must include K)

(Trim) output the k-largest dge vertices in $G[K']$.

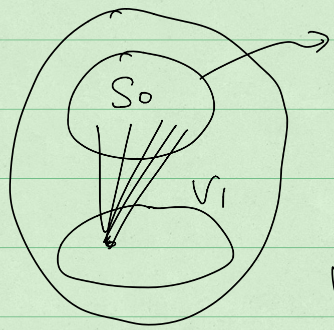
- How to generate such sequence of graphs.



- let V_1 denote the high degree vertices.
- problem: edges within V_1 are already exposed

$\Rightarrow G[V_1]$ is not distributed as a planted model.

- **Trick**:



anchor nodes ("test set") — "Beacon" set

chosen from V_0 uniformly at random.

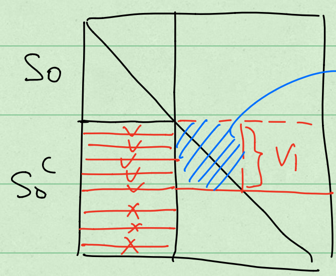
let $V_1 = \{ v \in V_0 \setminus S_0 : d_{S_0}(v) \geq \frac{1}{2}|S_0| + \frac{\beta}{2}\sqrt{|S_0|} \}$

where $d_{S_0}(v) = |\{ u \in S_0 : u \sim v \}|$

Claim: $G_1 = G[V_1] \sim G(n_1, \frac{1}{2}, k_1)$ condition on V_1

where $n_1 = |V_1|$, $k_1 = |K \cap V_1|$

Pf of claim: V_1 is chosen without exposing the edges with V_1.



G_1 (This part is never used in choosing V_1)

Claim 2: $|V_t| \triangleq n_t \approx n\tau^t \triangleq n_t'$

$|V_t \cap K| \triangleq k_t \approx k\rho^t \triangleq k_t'$

where $\tau = (1-\alpha)Q(\beta)$ $\quad \rho = (1-\alpha)Q(\beta - C_0\sqrt{\alpha})$

Pf: $d_{S_0}(v) \sim \begin{cases} \mathrm{Bin}(|S_0|, \frac{1}{2}) & \text{if } v \in K^c \setminus S_0 \\ |S_0 \cap K| + \mathrm{Bin}(|S_0 \setminus K|, \frac{1}{2}) & \text{if } v \in K \setminus S_0 \end{cases}$

$$\boxed{\begin{aligned} &\text{where } |S_0 \cap K| \sim \mathrm{Bin}(K,\alpha) \approx K\alpha = C\sqrt{n}\,\alpha \text{ whp.} \\ &|S_0| \sim \mathrm{Bin}(n,\alpha) \approx n\alpha \text{ whp.} \end{aligned}}$$

Thus

$$\mathbb{P}\left(d_{S_0}(v) \ge \tfrac{1}{2}|S_0| + \tfrac{\beta}{2}\sqrt{|S_0|} \right)$$

$$\overset{\text{B-E}}{\approx} \begin{cases} Q(\beta) \\ Q(\beta - C\sqrt{\alpha}) \end{cases} \qquad \approx N\left(C\sqrt{n}\alpha + (n\alpha - C\sqrt{n}\alpha)/2,\ \tfrac{1}{4}(n - C\sqrt{n})\alpha \right)$$

up to an error of $O\left(\frac{1}{\sqrt{n}}\right)$

Thus

$$\boxed{\begin{aligned} |V_1| &= \sum_{v \in V_0 \setminus S_0} \mathbb{1}\left\{ d_{S_0}(v) \ge \tfrac{1}{2}|S_0| + \tfrac{\beta}{2}\sqrt{|S_0|} \right\} \\ &\approx |V_0|\,(1-\alpha)\,Q(\beta) \quad \text{whp} \end{aligned}}$$

$\to$ $d_{S_0}(v)$ are i.i.d. across $v \in V_0 \setminus S_0$ condition on S_0.

Similarly.

$$\boxed{\begin{aligned} K_1 = |K \cap V_1| &= \sum_{v \in K \setminus S_0} \mathbb{1}\left\{ d_{S_0}(v) \ge \tfrac{1}{2}|S_0| + \tfrac{\beta}{2}\sqrt{|S_0|} \right\} \\ &\approx K(1-\alpha)\,Q(\beta - C\sqrt{\alpha}). \quad \text{whp.} \end{aligned}}$$

__Claim 3__: Let $\tilde{K}$ denote the set of $\frac{K_T'}{2}$ highest-degree vertices in G_T. Choose $T = C_0 \log\log n$, So that $\frac{K_T^2}{n_T} \ge \log^2 n$, say, and $n_T \approx n_T' = \rho^T n \geq \frac{n}{\mathrm{poly}(\log n)}$

$k_T \approx k_T' = k\,\tau^T \geq \frac{k}{\mathrm{polylog}(n)}$

$$e^{T\log\rho} = e^{C_0(\log\log n)\log\rho} = e^{\log[(\log n)^{C_0\log\rho}]} = (\log n)^{C_0\log\rho}$$

Then $\tilde{K} \subseteq K$ whp.

<u>Pf</u>: w.p. at least

$$1 - n_T\, e^{-\frac{k_T^2}{8 n_T}} \geq 1 - e^{-\mathrm{poly}(n)},$$

the hidden clique in $G_T \sim G(n_T, \frac{1}{2}, k_T)$ has the highest degrees.

Also, whp. $k_T \geq \frac{k_T'}{2}$.

Thus $\tilde{K} \subseteq K$ whp.

Now, we've show that $\tilde{K} \subseteq K$ whp. It remains to show we can expand it to the entire clique.

- $\tilde{K}$ is a large enough "seed set"
- <u>Caveat</u>: $\tilde{K}$ might depend on every edge in G. We get around this using union bound over all s-subsets of K, where $s = |\tilde{K}|$.

Claim4 (clean-up): whp, the follwy holds.

Let $\tilde{K}$ is an S-subset of K (which can be chosen adversarily)

let $K' = \tilde{K} \cup$ common neighbors of nodes in K.

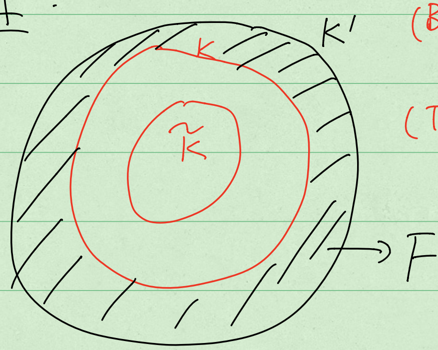
Let $\hat{K}$ denote tee highest degree vertices on $G' = G[K']$.

If $k = |K| \geq C\log n$ for sufficietly large C

nd $S \geq (1+\varepsilon)\log_2 n$ for any const $\varepsilon \in (0,1)$.

then $\hat{K} = K$ whp.

Pf:



(Blow-up) $K' = \tilde{K} \cup \{$common neighbors of nodes in $\tilde{K}\}$

(K' must in include K.)

(Trim) output the k-largest dge vertices in $G[K']$.

① $K' \supseteq K$, follows from $\tilde{K} \subseteq K$.

② write $K' = K \cup F$ where $F = K' \setminus K$

We show $|F|$ is small.

Fix a set $\tilde{K}$. For any node $u \in [n] \setminus K$

$$P(u \in F) = P\{u \sim v, \forall v \in \tilde{K}\} = \prod_{v \in \tilde{K}} P\{u \sim v\}$$
$$= 2^{-S}.$$

Moreover, $\{u \in F\}$ are independent across all $u \in [n] \setminus K$.

Thus $\mathbb{P}[\ |F| \geq \ell\ ] \leq \mathbb{P}\{\ \exists F' \subseteq [n] \setminus K : |F'| = \ell$
$F' \subseteq F\ \}$

$$\leq \sum_{\substack{F' \subseteq [n] \setminus K \\ |F'| = \ell}} \mathbb{P}\{F' \subseteq F\}$$

$$= \binom{n}{\ell} (2^{-s})^{\ell}$$

Take union bound over all possible $\tilde{K}$

$$\mathbb{P}(\exists \tilde{K} : |F| \geq \ell) \leq \binom{K}{s}\binom{n}{\ell} 2^{-s\ell}$$

$$\leq K^s n^\ell 2^{-s\ell}$$

$$= 2^{(1+\varepsilon)\log_2 n \log_2 K - \varepsilon \ell \log_2 n}$$

$$\to 0$$

if $\ell = \frac{2\log_2 K}{\varepsilon}$

so whp, for all choices of $\tilde{K}$,

$$|F| \leq \frac{2\log_2 K}{\varepsilon}$$

Now, finally in $G[K']$

for any $v \in K$, $d(v) \geq k-1$

any $\notin K$ $\quad d(v) \leq |F| + d_K(v)$

$$= \underbrace{|F|}_{O(\log k)} + \underbrace{\max_v d_K(v)}_{\frac{k}{2} + O(\sqrt{k \log n})}$$

$$= \frac{k}{2} + O(\sqrt{k \log n}) \text{ whp.}$$

$$< k-1 \quad \text{provided}$$

$$k \geq c \log n.$$

$\Rightarrow \hat{K} = K$ whp.