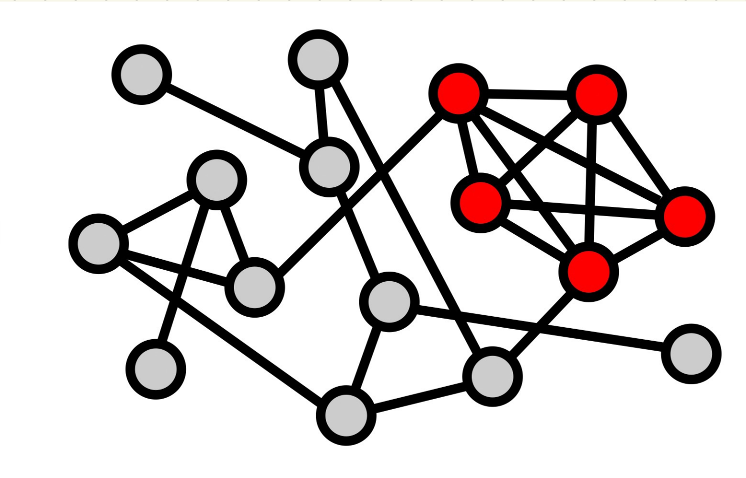
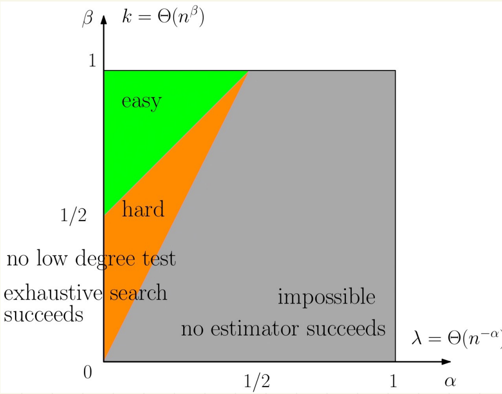
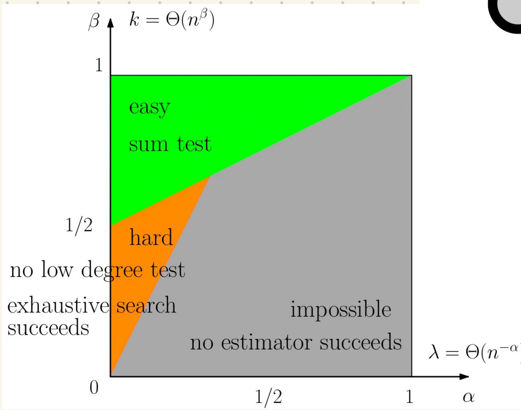
Average-case complexity + statistical inference.

FIONA SKERMAN





"recovering"



"detecting"

PLANTED DENSE SUBMATRIX

Vertex labels: $\sigma_v = \begin{cases} 1 \; \bullet & \text{w. prob } \frac{k}{n} \\ 0 & \text{w. prob } 1-\frac{k}{n} \end{cases}$

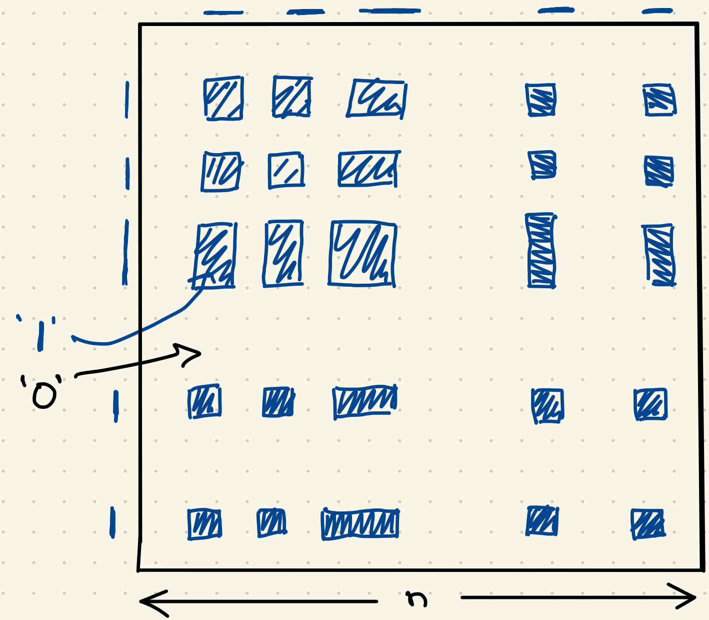
w. prob $\frac{k}{n}$

~~k of n~~, paint blue ● '1'

$$\sigma = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} \Big\} \; \cancel{k \text{ '1's}}$$

$\mathbb{E}[\#\text{'1's}] = k$

$\sigma\sigma^T =$



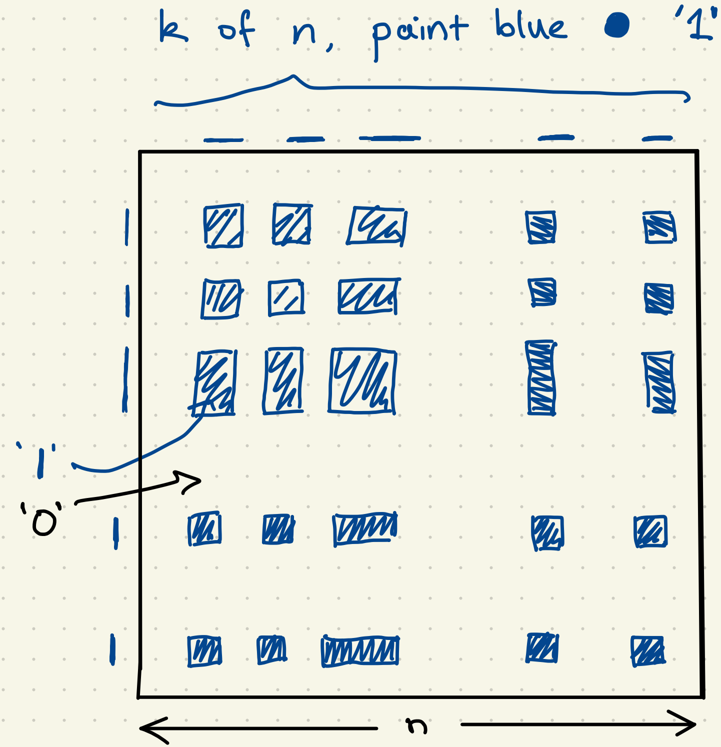
$k = n^\alpha \qquad 0 < \alpha < 1$

Planted Dense Submatrix

Vertex labels: $\sigma_v = \begin{cases} 1 & \bullet \quad \text{w. prob } \frac{k}{n} \\ 0 & \text{w. prob } 1-\frac{k}{n} \end{cases}$

$$\sigma = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} \Bigg\} \; k \text{ '1''s}$$

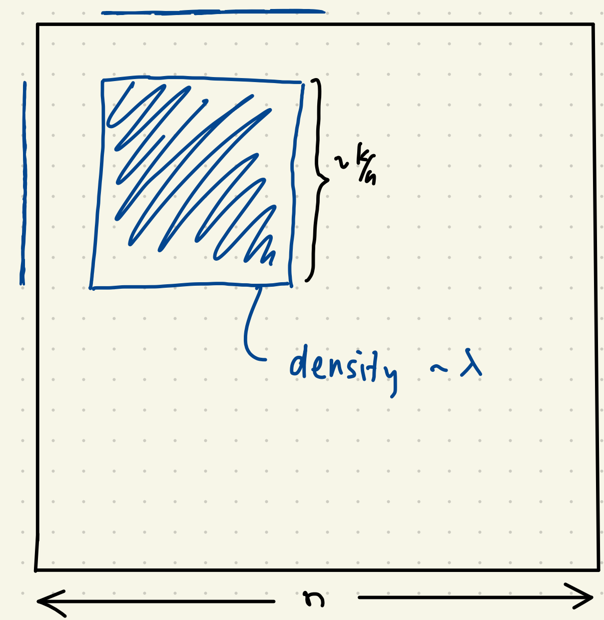
$\sigma \sigma^T =$



PLANTED DENSE SUBMATRIX

Vertex labels : $\sigma_v = \begin{cases} 1 & \text{w. prob } \frac{k}{n} \\ \emptyset & \text{w. prob } 1-\frac{k}{n} \end{cases}$

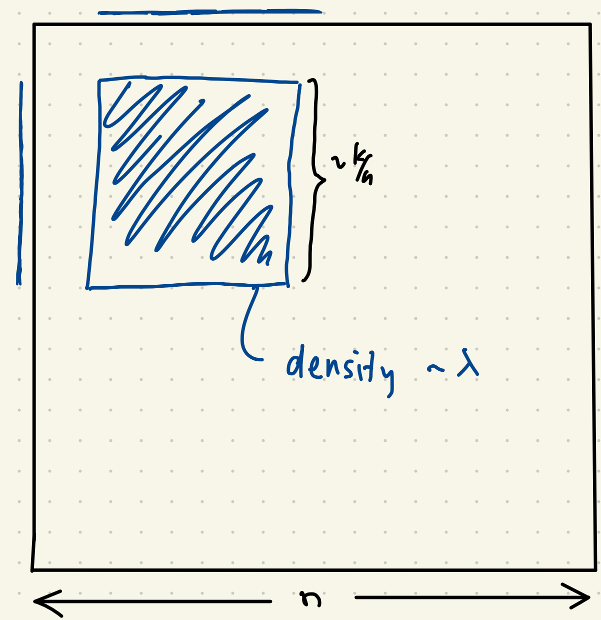
$G(n, k, \lambda)$



Planted Dense Submatrix

Vertex labels : $\sigma_v = \begin{cases} 1 & \text{w. prob } \frac{k}{n} \\ \phi & \text{w. prob } 1 - \frac{k}{n} \end{cases}$

$G(n, k, \lambda)$



Observe $Y_{uv} \sim \begin{cases} \lambda + \mathcal{N}(0,1) & \sigma_u = \sigma_v = 1 \\ \mathcal{N}(0,1) & \text{o.w.} \end{cases}$

Algorithmic Qns

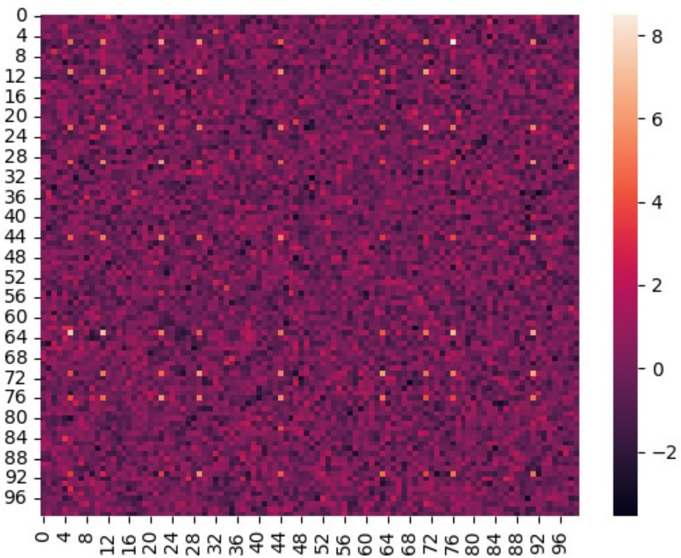
- Detection : determine if whp sample from planted model or all entries $\mathcal{N}(0,1)$
- Recovery : given sample from planted model find communities (exactly? weakly corr?)

PLANTED DENSE SUBMATRIX - SIMULATIONS

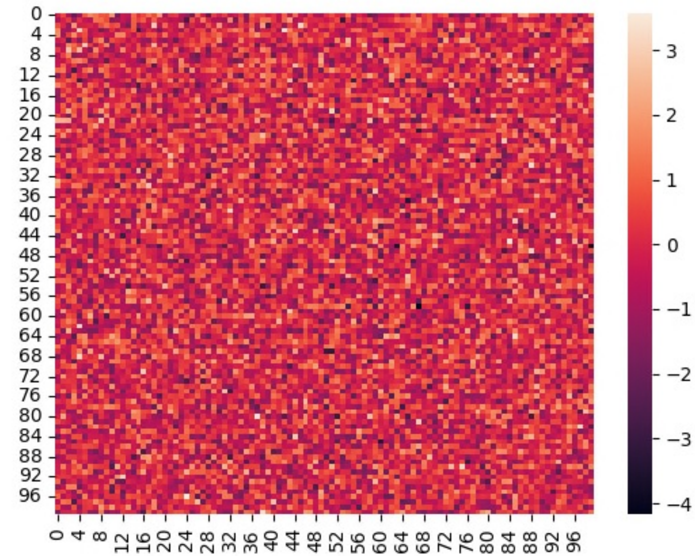
Vertex labels: $\sigma_v = \begin{cases} 1 & \text{w. prob } \frac{k}{n} \\ \emptyset & \text{w. prob } 1-\frac{k}{n} \end{cases}$

$Y_{uv} \sim \begin{cases} N(\lambda,1) & \sigma_u=\sigma_v=1 \\ N(0,1) & \text{o.w.} \end{cases}$

H_1 $n=100$, $k=15$, $\lambda=5$



H_0 $n=100$ (all $N(0,1)$)



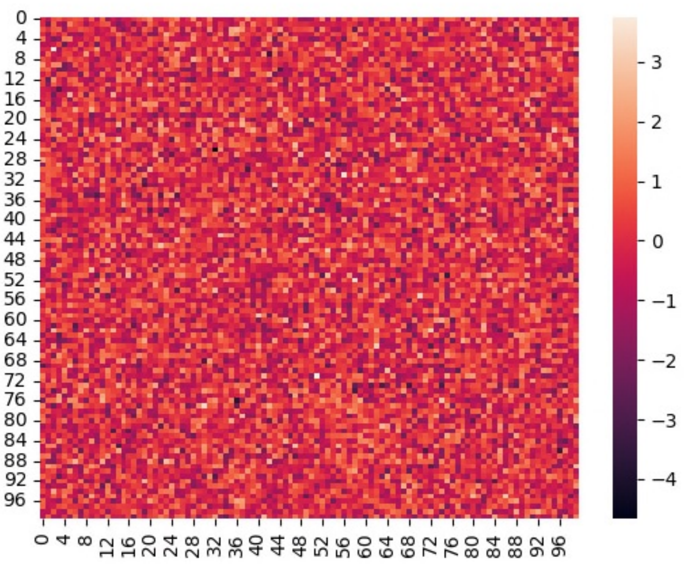
PLANTED DENSE SUBMATRIX - SIMULATIONS

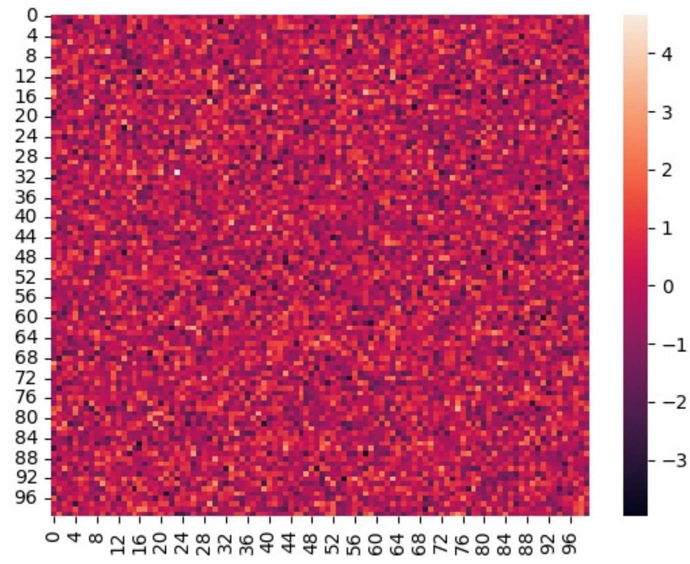
Vertex labels: $\sigma_v = \begin{cases} 1 & \text{w. prob } \frac{k}{n} \\ \emptyset & \text{w. prob } 1-\frac{k}{n} \end{cases}$

$Y_{uv} \sim \begin{cases} N(\lambda,1) & \sigma_u=\sigma_v=1 \\ N(0,1) & \text{o.w.} \end{cases}$

H_1 $n=100$, $k=15$, $\lambda=0.5$

H_0 $n=100$ (all $N(0,1)$)





PLANTED DENSE SUBMATRIX – SIMULATIONS

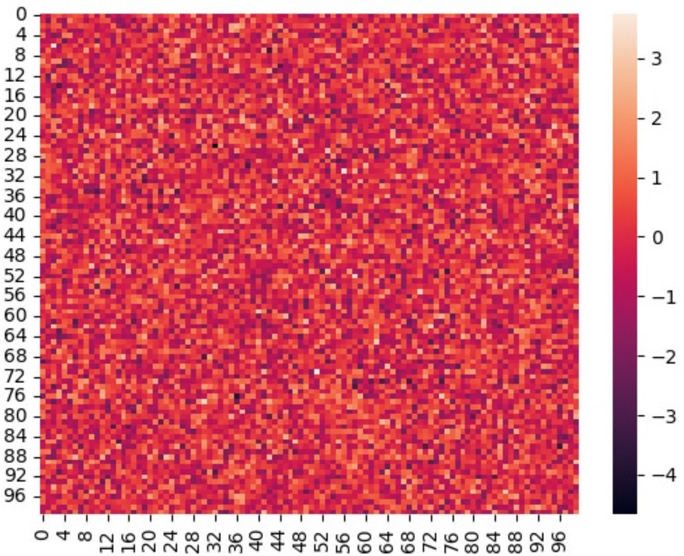
Vertex labels: $\sigma_v = \begin{cases} 1 & \text{w. prob } \frac{k}{n} \\ \emptyset & \text{w. prob } 1-\frac{k}{n} \end{cases}$

$Y_{uv} \sim \begin{cases} \mathcal{N}(\lambda, 1) & \sigma_u = \sigma_v = 1 \\ \mathcal{N}(0,1) & \text{o.w.} \end{cases}$

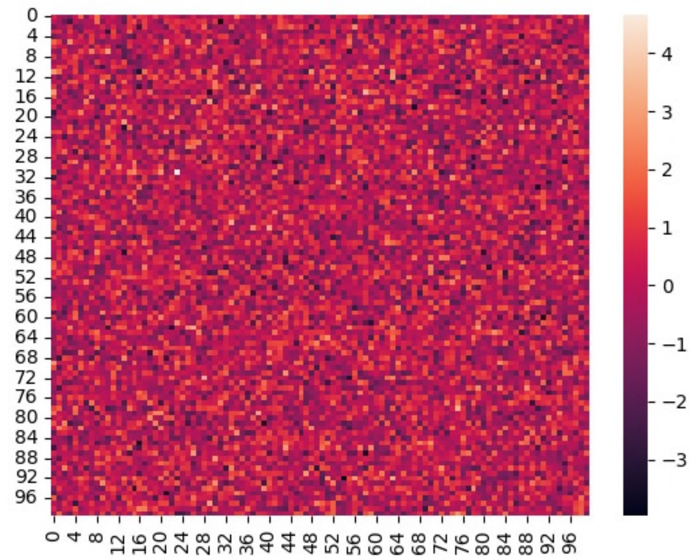
H_1 $n=100$, $k=15$, $\lambda = 0.5$

H_0 $n=100$ (all $N(0,1)$)

sum ≈ 100.35



sum ≈ 12.36



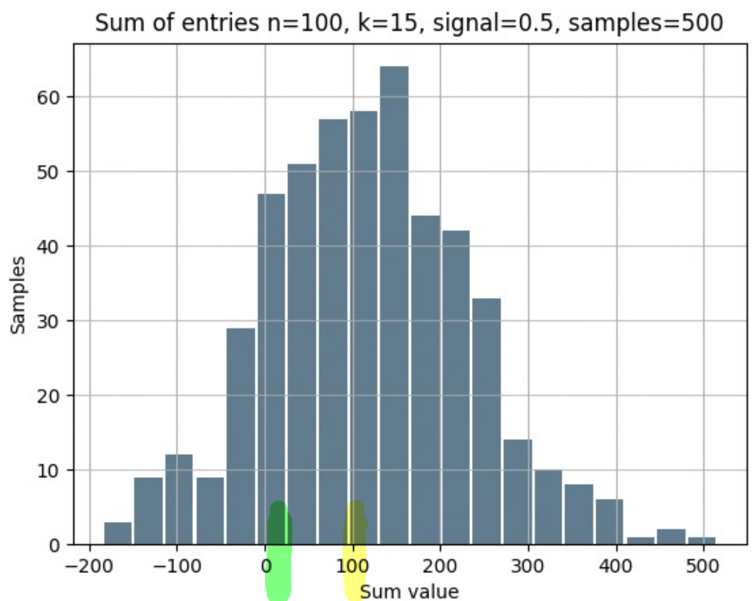
PLANTED DENSE SUBMATRIX - SIMULATIONS

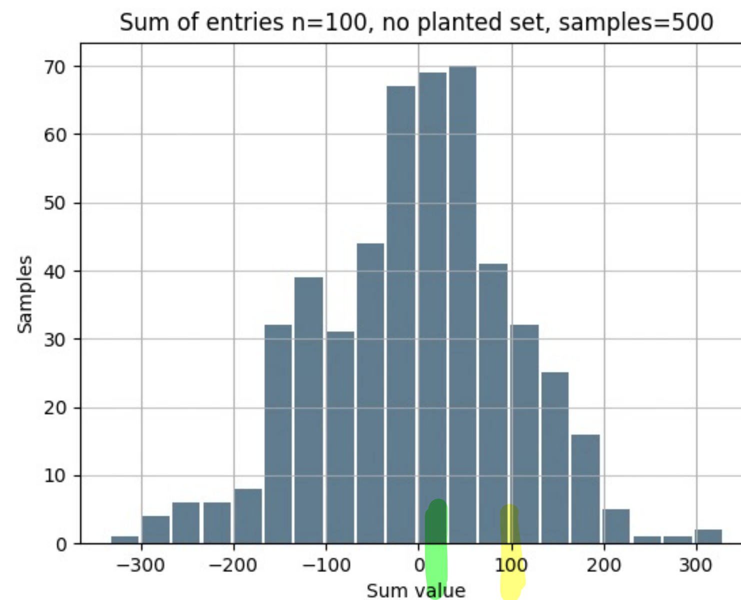
Vertex labels: $\sigma_v = \begin{cases} 1 & \text{w. prob } \frac{k}{n} \\ \emptyset & \text{w. prob } 1-\frac{k}{n} \end{cases}$

$$Y_{uv} \sim \begin{cases} \mathcal{N}(\lambda,1) & \sigma_u=\sigma_v=1 \\ \mathcal{N}(0,1) & \text{o.w.} \end{cases}$$

H_1 $n=100$, $k=15$, $\lambda=0.5$

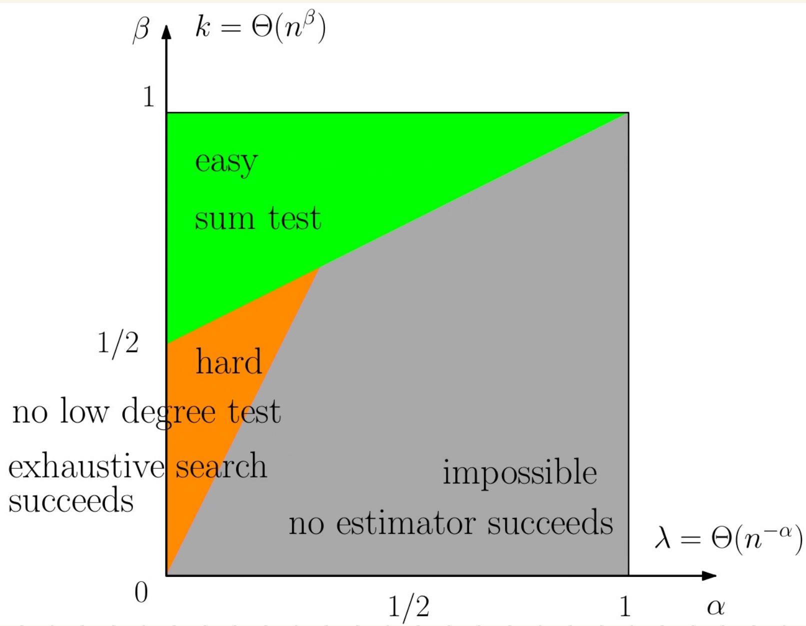
H_0 $n=100$ (all $N(0,1)$)



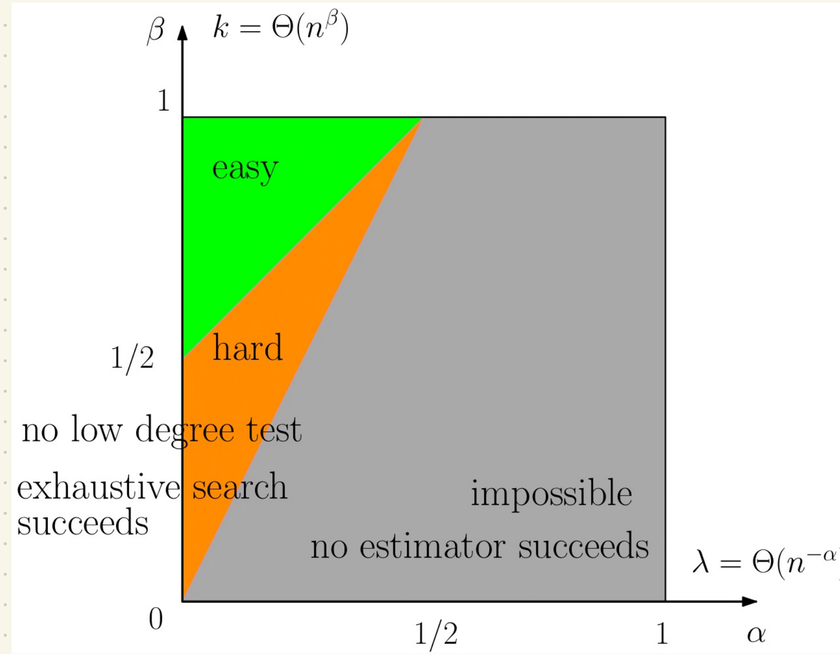


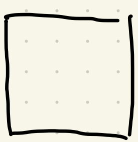
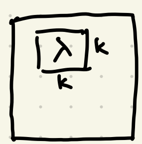
CONTEXT

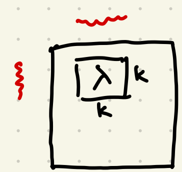
Detection



Recovery



H_0  vs. H_1 

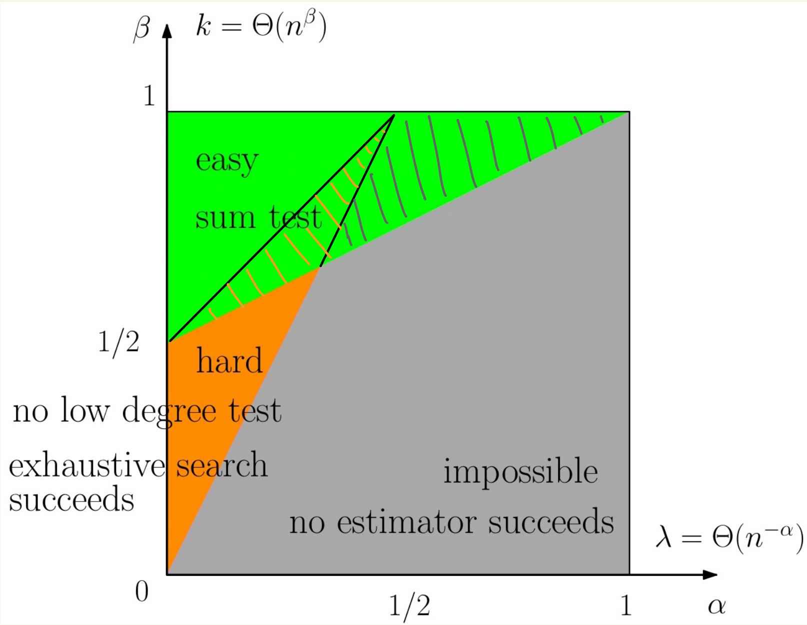
recover 

REFS: MANY AUTHORS. BI13, BIS15, MW15, CX16, DM14, CLR17, HWX17, BBH18, GJS19, BMR20, BBP05, BS06, FP07, CDF09, BGN11, SWPN09, KBRS11, BKR$^+$11, ACD11, BWZ20, SW22

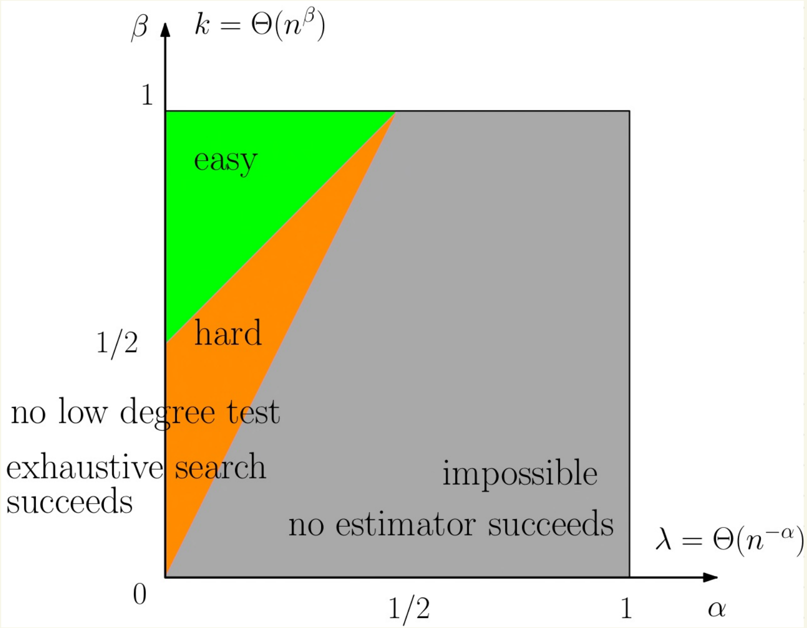
CONTEXT

'Easier to detect than recover'.

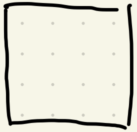
Detection



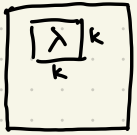
Recovery



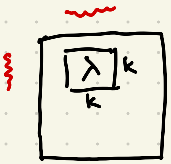
H_0

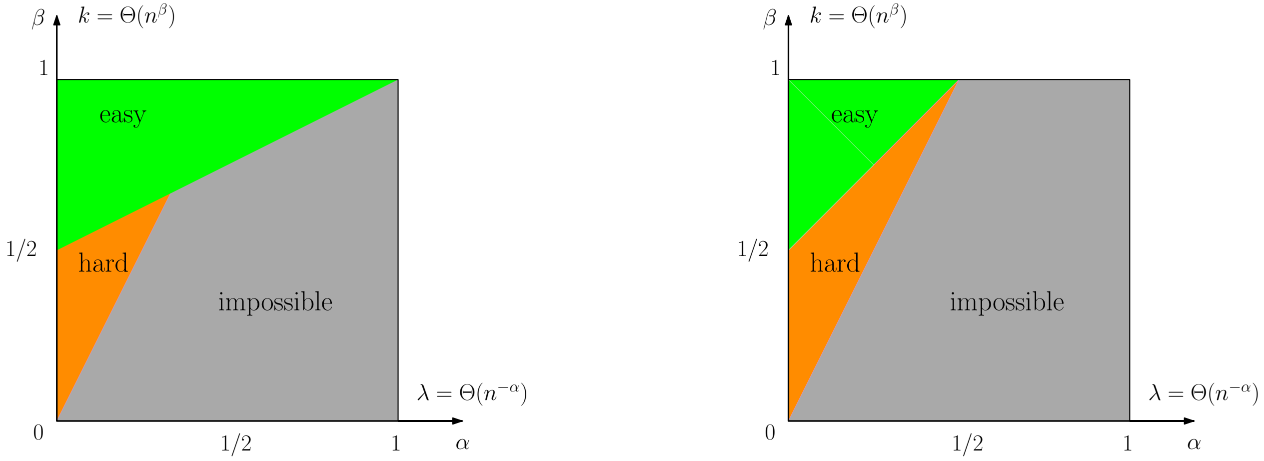


vs. H_1



recover





(a) detection

(b) recovery

Figure 1: **Spiked Matrix Model** (planted submatrix with elevated mean).
H_0: random $n \times n$ matrix with each entry independent with distribution $N(0,1)$.
H_1: $n \times n$ matrix with each index in set S independently with probability k/n. Each entry independent with distribution $N(\lambda, 1)$ if $i, j \in S$ and with distribution $N(0,1)$ otherwise.

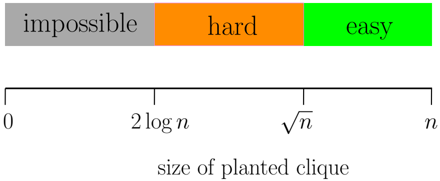


Figure 0: **Planted clique.**
H_0: $G(n, \frac{1}{2})$ random graph on n vertices where each edge is present independently with probability 1/2.
H_1: $G(n, k, \frac{1}{2})$, random graph on n vertices where each vertex is part of 'community' S independently with probability k/n. Each edge ij is present independently either with probability 1 if $i, j \in S$ or with probability 1/2 otherwise.

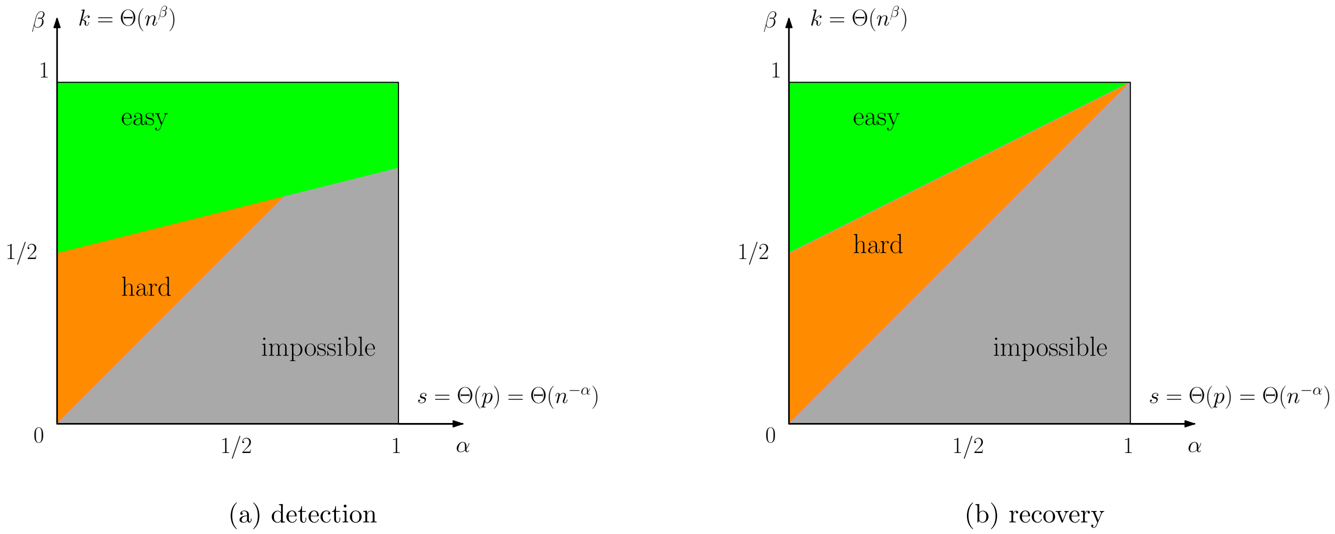


Figure 2: **Planted dense subgraph**.
H_0: $G(n, q)$ random graph on n vertices where each edge is present independently with probability q.
H_1: $G(n, k, q, s)$ with $s > 0$, random graph on n vertices where each vertex is part of 'community' S independently with probability k/n. Each edge ij is present independently either with probability $q + s$ if $i, j \in S$ or with probability q otherwise.

Planted Community $\quad G \sim G(n, \overset{\text{signal}}{p}, \overset{\text{noise}}{q}, k), \quad K \overset{u}{\in} \binom{[n]}{k}$ $\quad A_{ij} = \begin{cases} Be(p) & i,j \in K \\ Be(q) & o.w. \end{cases}$

$p > q$

n points

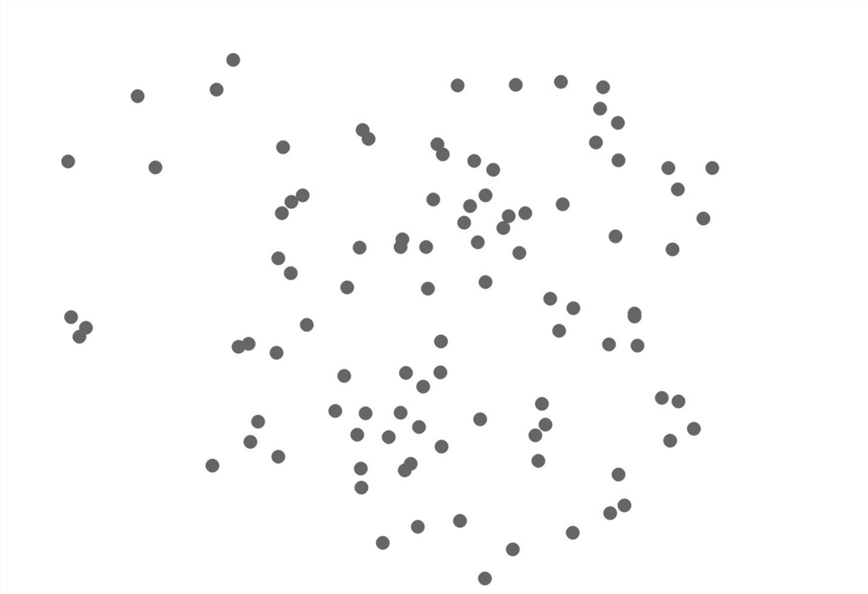


Fig: Jiaming Xu, Duke

Planted Community $G \sim G(n, \overset{\text{signal}}{p}, \overset{\text{noise}}{q}, k)$, $K \overset{u}{\in} \binom{[n]}{k}$ $A_{ij} = \begin{cases} Be(p) & i,j \in K \\ Be(q) & o/w \end{cases}$

$p > q$

n points

k 'community' nodes

n-k 'non-community' "

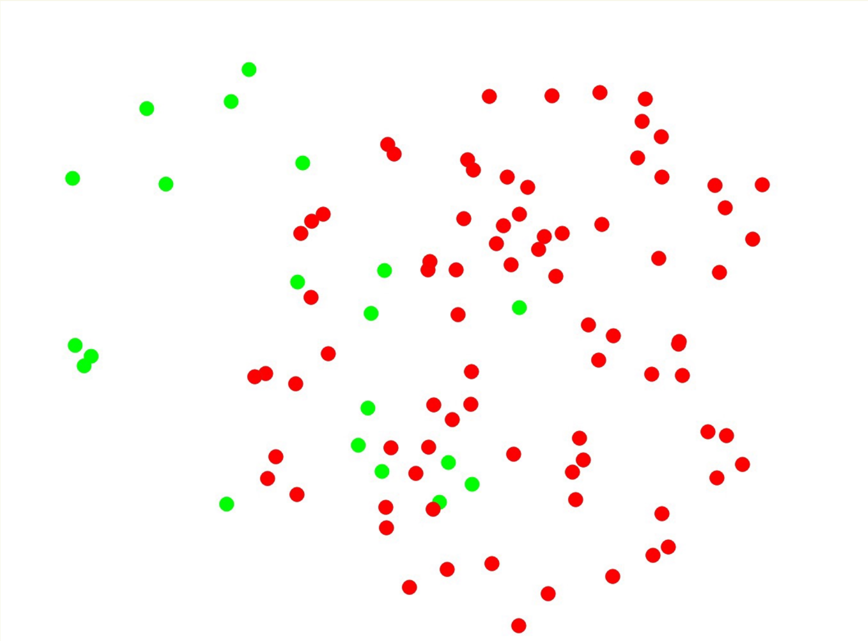


Fig: Jiaming Xu, Duke

Planted Community

$G \sim G(n, \overset{\text{signal}}{p}, \overset{\text{noise}}{q}, k)$, $\quad K \overset{u}{\in} \binom{[n]}{k}$ $\qquad A_{ij} = \begin{cases} Be(p) & i,j \in K \\ Be(q) & \text{ow} \end{cases}$

$p > q$

n points

k 'community' nodes

n-k 'non-community' "

with prob. p

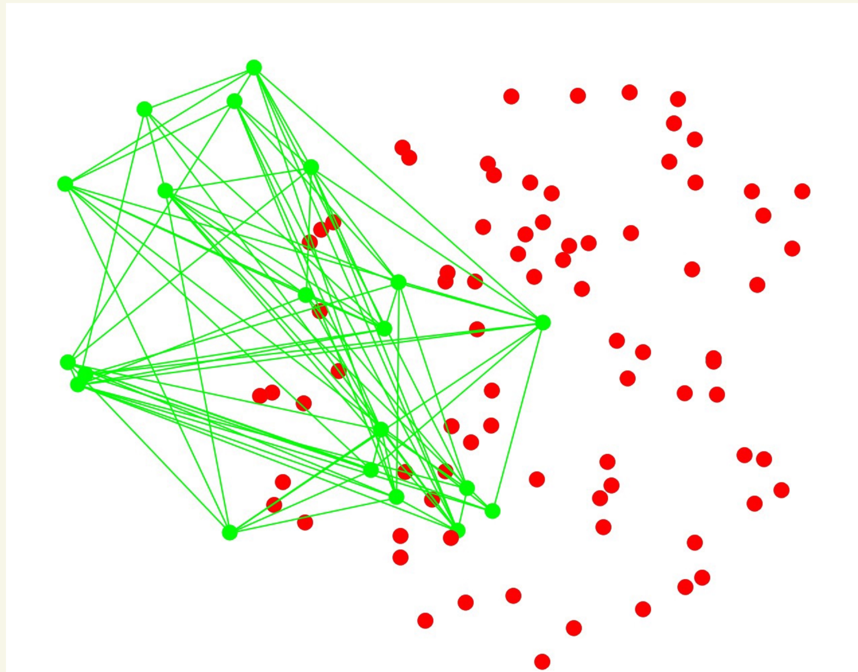


Fig: Jiaming Xu, Duke

Planted Community

$G \sim G(n, \overset{\text{signal}}{p}, \overset{\text{noise}}{q}, k)$, $\quad p > q$ $\qquad K \overset{u}{\in} \binom{[n]}{k}$ $\qquad A_{ij} = \begin{cases} Be(p) & i,j \in K \\ Be(q) & \text{o/w} \end{cases}$

n points

k 'community' nodes

$n-k$ 'non-community' "

with prob. p

" " q

" " q

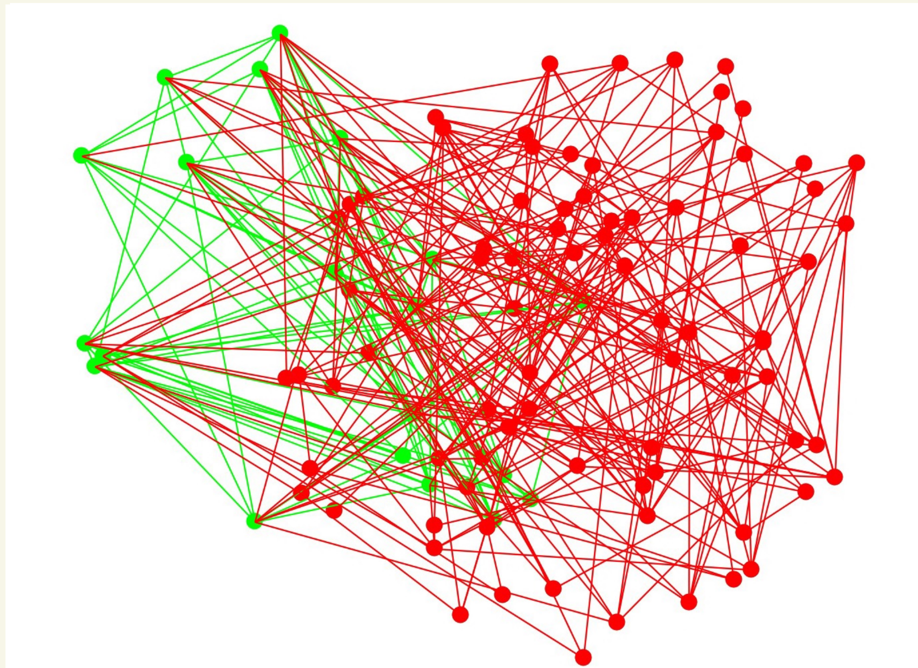


Fig: Jiaming Xu, Duke

Planted Community

$G \sim G(n, \overset{\text{signal}}{p}, \overset{\text{noise}}{q}, k)$, $\quad K \overset{u}{\in} \binom{[n]}{k}$ $\quad A_{ij} = \begin{cases} Be(p) & i,j \in K \\ Be(q) & ow \end{cases}$

$p > q$

Process

n points

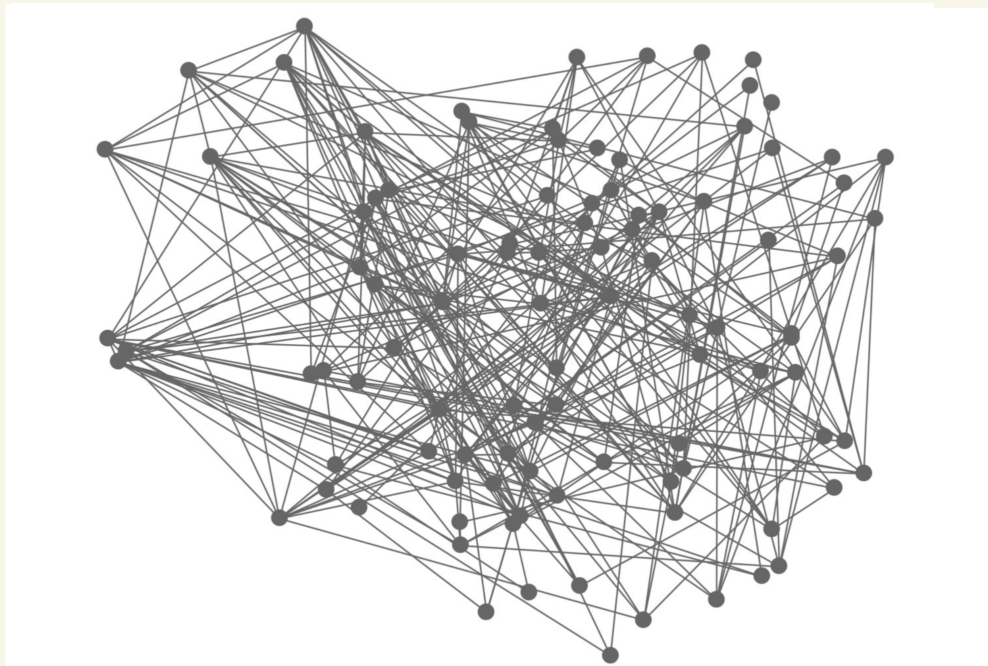
k 'community' nodes

$n-k$ 'non-community' "

with prob. p

" " q

" " q



Output

unlabelled graph

Fig: Jiaming Xu, Duke

Planted Community

$$G \sim G(n, \overset{\text{signal}}{p}, \overset{\text{noise}}{q}, k), \quad K \overset{u}{\in} \binom{[n]}{k}$$

$p > q$

$$A_{ij} = \begin{cases} Be(p) & i,j \in K \\ Be(q) & ow \end{cases}$$

Process

n points

k 'community' nodes

n-k 'non-community' "

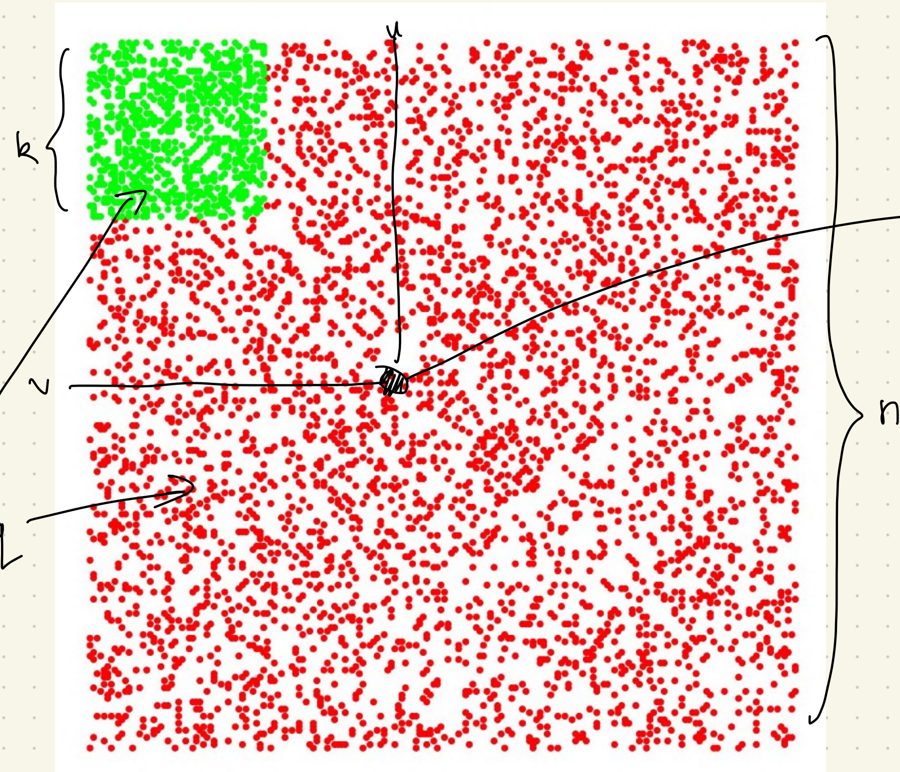
green–green: with prob. p

red–green: " " q

red–red: " " q

Output

unlabelled graph



draw dot if u — v in graph

$n = 200 \quad k = 50 \quad p = 0.3 \quad q = 0.1$

Fig: Jiaming Xu, Duke

Planted Community

signal   noise

$$G \sim G(n, p, q, k), \quad K \overset{u}{\in} \binom{[n]}{k} \qquad A_{ij} = \begin{cases} Be(p) & i,j \in K \\ Be(q) & \text{ow} \end{cases}$$

$p > q$

Process

n points

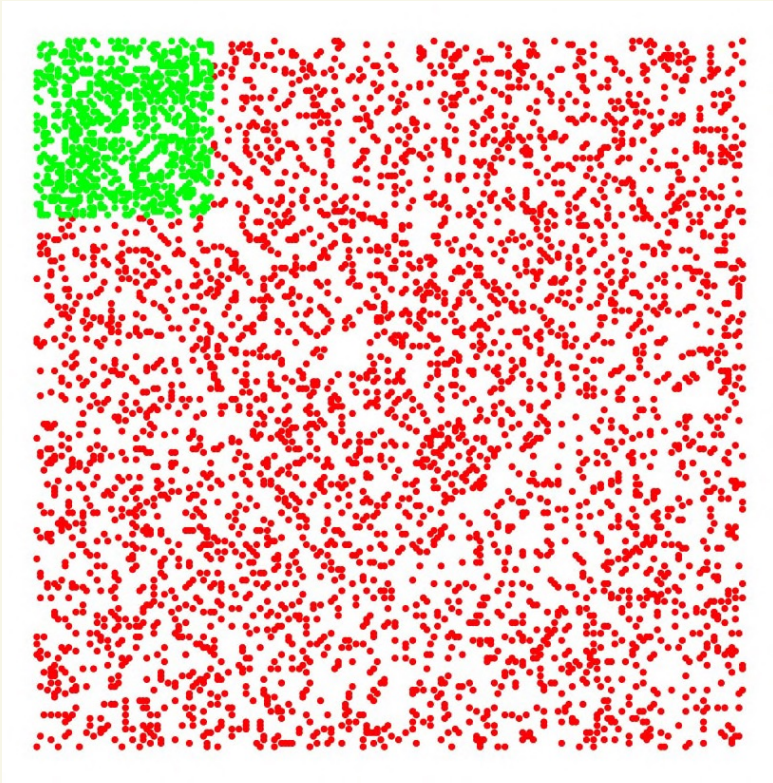
k 'community' nodes

n-k 'non-community' "

with prob. p

" " q

" " q



Output

unlabelled graph

$n = 200 \quad k = 50 \quad p = 0.3 \quad q = 0.1$

Fig: Jiaming Xu, Duke

Planted Community

$G \sim G(n, \overset{\text{signal}}{p}, \overset{\text{noise}}{q}, k)$, $\quad K \overset{u}{\in} \binom{[n]}{k}$ $\quad A_{ij} = \begin{cases} Be(p) & i,j \in K \\ Be(q) & ow \end{cases}$

$p > q$

Process

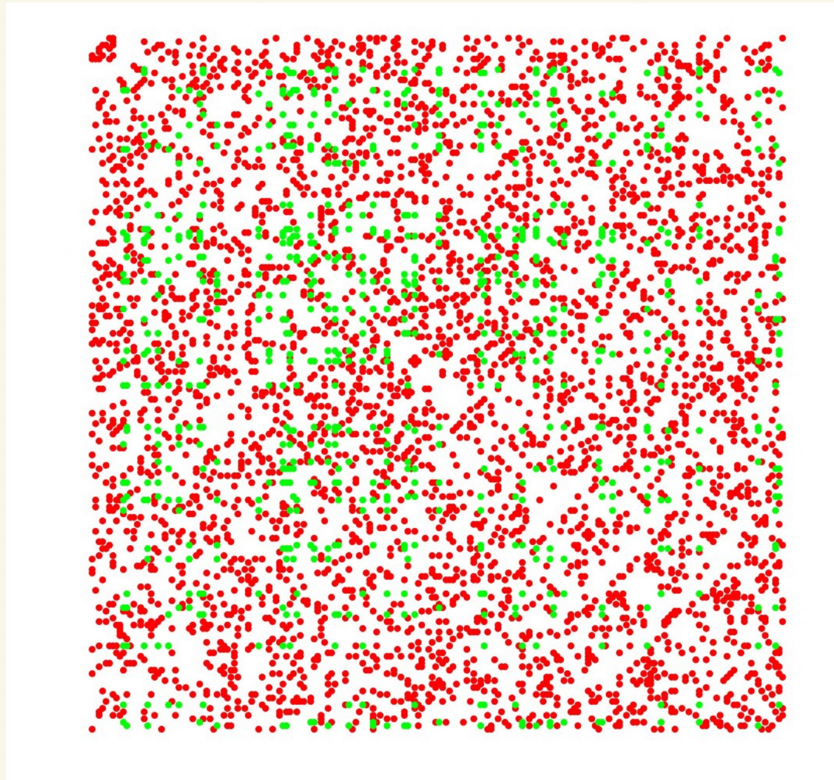
n points

- k 'community' nodes
- $n-k$ 'non-community' "

- with prob. p
- " " q
- " " q

Output

unlabelled graph



$n = 200 \quad k = 50 \quad p = 0.3 \quad q = 0.1$

Fig: Jiaming Xu, Duke

Planted Community

$G \sim G(n, \overset{\text{signal}}{p}, \overset{\text{noise}}{q}, k)$, $K \overset{u}{\in} \binom{[n]}{k}$ $\quad A_{ij} = \begin{cases} Be(p) & i,j \in K \\ Be(q) & ow \end{cases}$

$p > q$

Process

n points

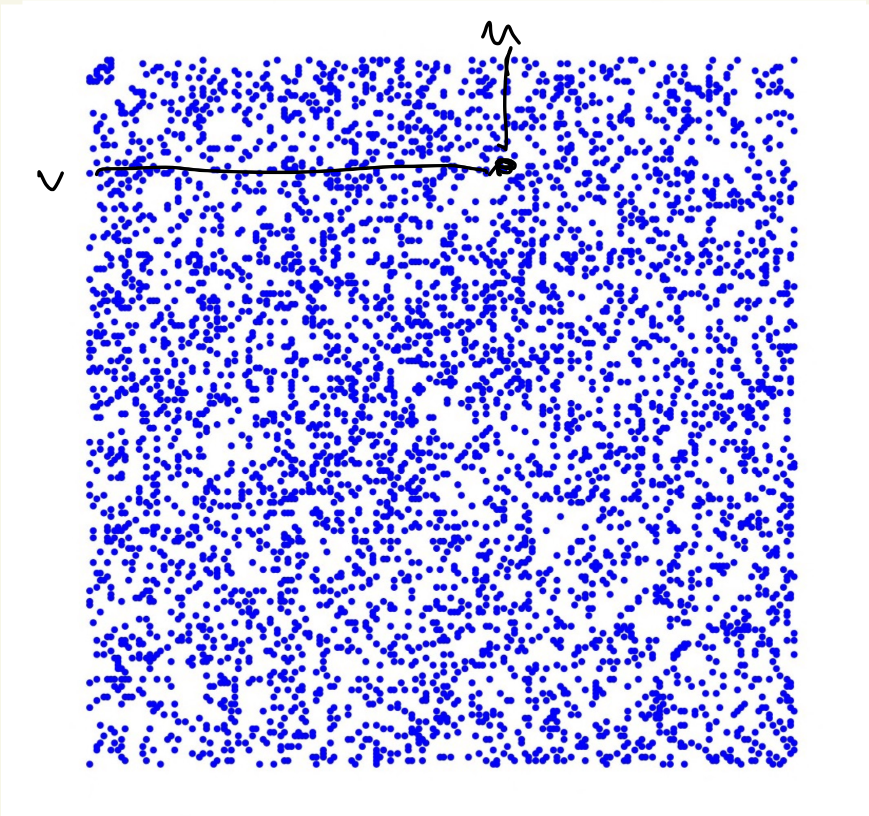
k 'community' nodes

$n-k$ 'non-community' "

with prob. p

" " q

" " q



$n = 200$ $\quad k = 50$ $\quad p = 0.3$ $\quad q = 0.1$

Output

unlabelled graph

Fig: Jiaming Xu, Duke

Planted Clique $G \sim G(n, \frac{1}{2}, k)$, $K \overset{u}{\in} \binom{[n]}{k}$ $A_{ij} = \begin{cases} 1 & i,j \in K \\ Be(\frac{1}{2}) & \text{ow} \end{cases}$

Two parameters
- size of planted structure
- size of entire network

Q: When can we find planted clique?

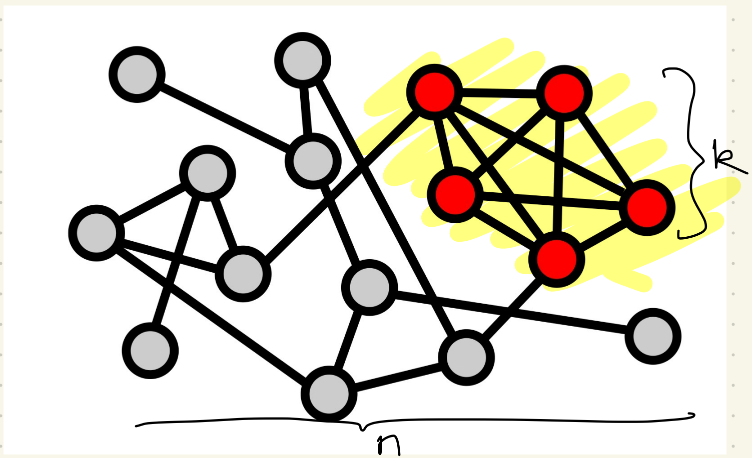


Fig: Alex Wein

Planted Clique $G \sim G(n, \frac{1}{2}, k)$, $K \overset{u}{\in} \binom{[n]}{k}$ $A_{ij} = \begin{cases} 1 & i,j \in K \\ Be(\frac{1}{2}) & ow \end{cases}$

$G' \sim G(n, \frac{1}{2})$: largest clique $2\log_2 n$. (with prob $\to 1$)

if $|K| \leq 2\log_2 n$
$\Rightarrow$ can't find "planted" one in amongst "background" one.

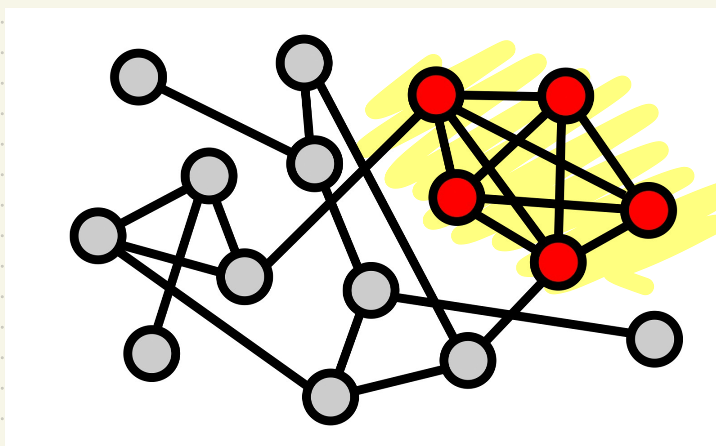
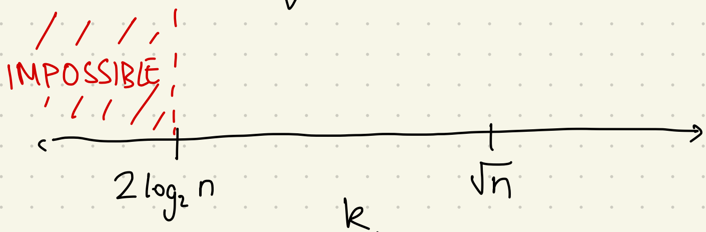


Fig: Alex Wein

Planted Clique $\quad G \sim G(n, \frac{1}{2}, k)$, $\quad K \overset{u}{\in} \binom{[n]}{k}$ $\quad A_{ij} = \begin{cases} 1 & i,j \in K \\ Be(\frac{1}{2}) & \text{ow} \end{cases}$



if $|K| \leq 2\log_2 n$

$G' \sim G(n, \frac{1}{2})$: largest clique whp $\sim 2\log_2 n$. $\Rightarrow$ can't find "planted" one in amongst "background" one.

Methods to find clique

① DEGREE TEST

$\hat{K}$ = set of k vertices of highest degree

Thm [Kuč 95] $\quad k = \Omega(\sqrt{n \log n}) \implies P(\hat{K} = K) \to 1.$

[an interative version get $\Omega(\sqrt{n})$ enough]

Planted Clique $G \sim G(n, \frac{1}{2}, k)$, $K \overset{u}{\in} \binom{[n]}{k}$ $A_{ij} = \begin{cases} 1 & i,j \in K \\ Be(\frac{1}{2}) & \text{o.w} \end{cases}$

$G' \sim G(n, \frac{1}{2})$: largest clique whp $\sim 2\log_2 n$. if $|K| \leq 2\log_2 n$

$\Rightarrow$ can't find "planted" one in amongst "background" one.

Methods to find clique

① DEGREE TEST

$\hat{K}$ = set of k vertices of highest degree

Thm [Kuč 95] $k = \Omega(\sqrt{n \log n}) \Rightarrow P(\hat{K} = K) \to 1$.

[an interative version get $\Omega(\sqrt{n})$ enough]

② SPECTRAL METHOD $W_{ij} = \begin{cases} 2A_{ij} - 1 & i \neq j \\ 0 & \text{o.w} \end{cases}$

(i) u top eigenvector of W

(ii) (threshold) $\tilde{K}$ index vector of k largest $|u_i|$

(iii) (clean-up) $\hat{K} = \{v \in V(G) : e(v, \tilde{K}) \geq \frac{3k}{4}\}$

Thm [Alon Krivelevich Sudakov '98]

$k = \Omega(\sqrt{n}) \Rightarrow P(\hat{K} = K) \to 1$

Planted Clique $\quad G \sim G(n, \frac{1}{2}, k)$, $\quad K \overset{u}{\in} \binom{[n]}{k}$ $\quad A_{ij} = \begin{cases} 1 & i,j \in K \\ Be(\frac{1}{2}) & \text{o.w} \end{cases}$

$G' \sim G(n, \frac{1}{2})$: largest clique whp $\sim 2\log_2 n$. $\quad$ if $|K| \le 2\log_2 n$ $\Rightarrow$ can't find "planted" one in amongst "background" one.

Methods to find clique

① DEGREE TEST

$\hat{K}$ = set of k vertices of highest degree

Thm [Kuč 95] $\quad k = \Omega(\sqrt{n \log n}) \Rightarrow P(\hat{K} = K) \to 1$.

[an interative version $\Omega(\sqrt{n})$ enough]

② SPECTRAL METHOD $\qquad W_{ij} = \begin{cases} 2A_{ij} - 1 & i \ne j \\ 0 & \text{o.w} \end{cases}$

(i) u top eigenvector of W

(ii) (threshold) $\quad \tilde{K}$ index vector of k largest $|u_i|$

(iii) (clean-up) $\quad \hat{K} = \{ v \in V(G) : e(v, \tilde{K}) \ge \frac{3k}{4} \}$

Thm [Alon Krivelevich Sudakov '98]

$k = \Omega(\sqrt{n}) \Rightarrow P(\hat{K} = K) \to 1$

③ SDP METHOD

Yes. If $k = \Omega(\sqrt{n})$.

Planted Clique $G \sim G(n, \frac{1}{2}, k)$, $K \overset{u}{\in} \binom{[n]}{k}$ $A_{ij} = \begin{cases} 1 & i,j \in K \\ Be(\frac{1}{2}) & \text{ow} \end{cases}$

* Rigorous Evidence (suggesting no poly-time alg).

- reductions (avg. case)
- restricted class of alg low deg poly

$G' \sim G(n, \frac{1}{2})$: largest clique whp $\sim 2\log_2 n$. $\Rightarrow$ can't find "planted" one in amongst "background" one.

Methods to find clique

① DEGREE TEST

$\hat{K}$ = set of k vertices of highest degree

Thm [Kuč 95] $k = \Omega(\sqrt{n \log n}) \implies P(\hat{K} = K) \to 1$.

[an interative version get $\Omega(\sqrt{n})$ enough]

② SPECTRAL METHOD $W_{ij} = \begin{cases} 2A_{ij} - 1 & i \neq j \\ 0 & \text{o.w} \end{cases}$

(i) u top eigenvector of W

(ii) (threshold) $\tilde{K}$ index vector of k largest $|u_i|$

(iii) (clean-up) $\hat{K} = \{v \in V(G) : e(v, \tilde{K}) \geq \frac{3k}{4}\}$

Thm [Alon Krivelevich Sudakov '98]

$k = \Omega(\sqrt{n}) \Rightarrow P(\hat{K} = K) \to 1$

③ SDP METHOD

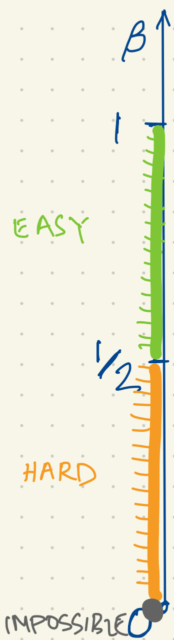
Yes. If $k = \Omega(\sqrt{n})$.

CONTEXT

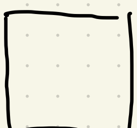
PLANTED CLIQUE

Detection

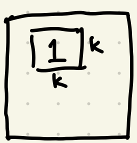
$k = \Theta(n^\beta)$



H_0

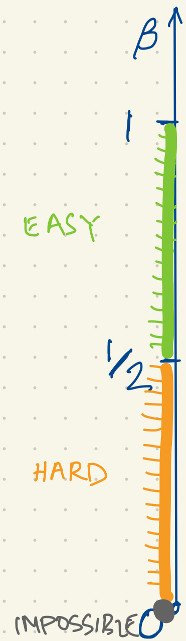


vs. H_1

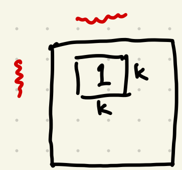


Recovery

$k = \Theta(n^\beta)$



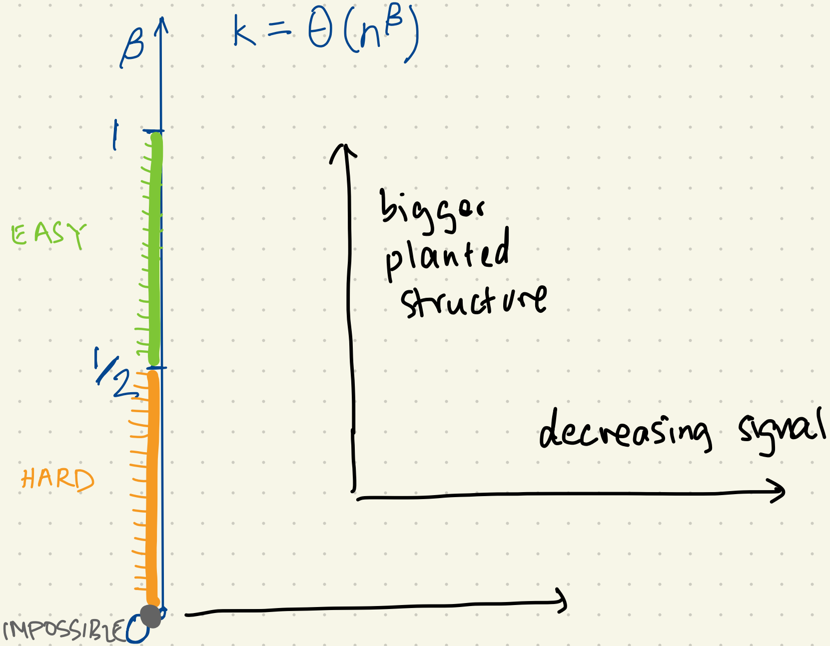
recover



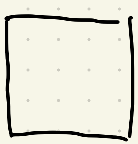
CONTEXT

PLANTED CLIQUE

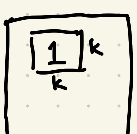
Detection



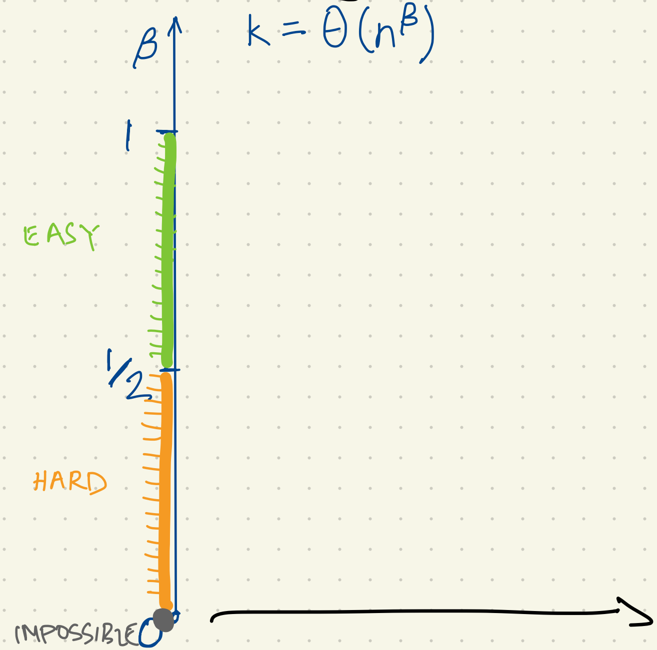
H_0



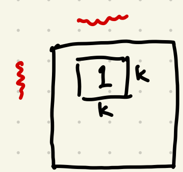
vs. H_1



Recovery



recover



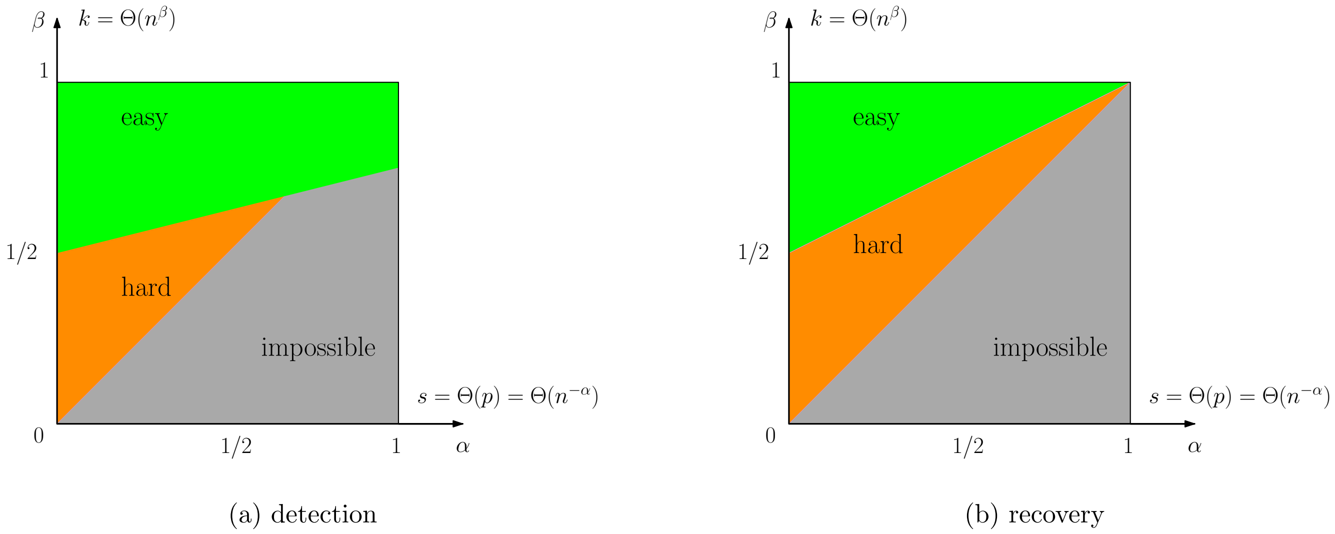
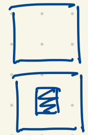


Figure 2: **Planted dense subgraph**.
H_0: $G(n, q)$ random graph on n vertices where each edge is present independently with probability q.
H_1: $G(n, k, q, s)$ with $s > 0$, random graph on n vertices where each vertex is part of 'community' S independently with probability k/n. Each edge ij is present independently either with probability $q + s$ if $i, j \in S$ or with probability q otherwise.

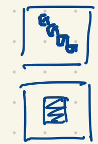
Hypothesis Testing Given Sample which model was it generated from.

H_0: $G \sim P_n : G(n, \frac{1}{2})$

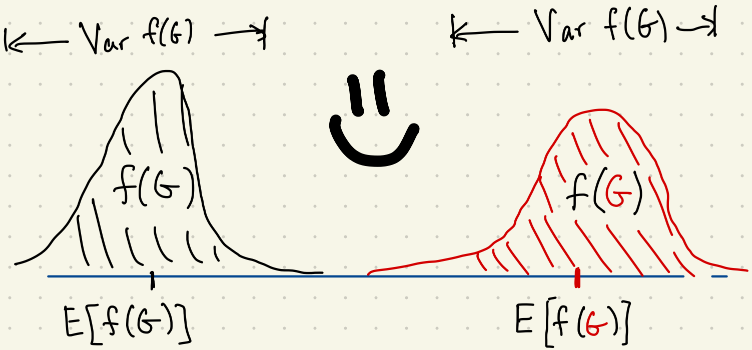
H_1: $G \sim Q_n : G(n, \frac{1}{2}, k)$



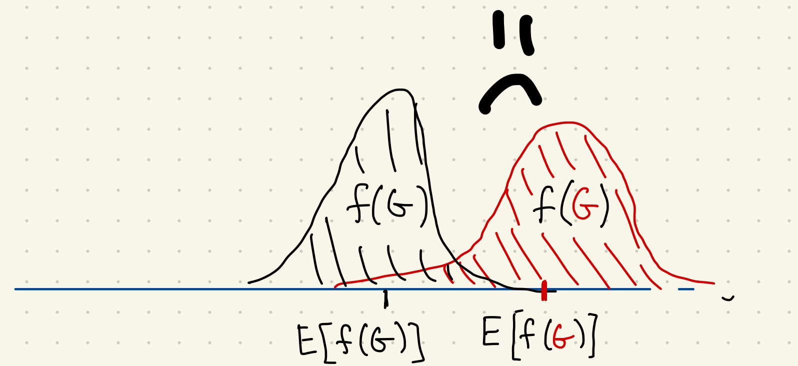
distributions on $\mathbb{R}^{\binom{n}{2}}$.



f detects



f doesn't detect



seq. of poly

A degree D test $f_n : \mathbb{R}^{n^2} \to \mathbb{R}$ deg $\leq D$. strongly seperates if

$$\mathbb{E}_{P_n}[f] - \mathbb{E}_{Q_n}[f] \gg \sqrt{\max\{\mathrm{Var}_Q[f], \mathrm{Var}_P[f]\}}$$

"difference in means" $\gg$ "fluctuations".

NB: $D \sim \log n$
consider small / fast

$D \gg \log n$
consider high deg / slow.

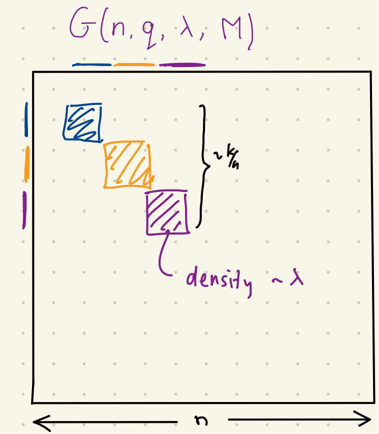
Further particulars The course will comprise ~15 lectures and ~5 problems sessions. The assessment, all of which can be done in small groups (up to 2-3), will be exercise sheets ($2\times25\%$) and 1 longer project (50%). The first exercise sheet will be out Friday 3rd and due Monday 21st February, the second will be out Friday 24th March and due 17th April.

For the longer project is to understand the proof of tractability, hardness or impossibility of a particular problem. List of suggestions will be provided (by 21st April) including some reductions in total variation from a paper by Brennan and Breser, spectral method to achieve the threshold in stochastic block from a paper by Lelarge, Bordenave and Massoulié as well as some candidate lemmas which together will prove some new results (probably a new testing problem where both H_0 and H_1 consist of different planted structures instead of planted and null: with lemmas to prove low-deg hardness, find fast algorithms, info-theoretic thresholds). Hand in either ~5-10 pages give or 25 minutes talk each person end of May / early June.

Dates (provisional) Lectures and problem sessions all in 64119 unless otherwise indicated, and will start 15min past the hour.

L1 Thu 26th Jan 3-5pm
L2 Wed 1st Feb 3-5pm
L3 Thur 9th Feb 3-5pm
L4 Wed 15th Feb 3-5pm
L5 Wed 22nd Feb 3-5pm
L6 Wed 1st Mar 3-5pm
L7 Wed 8th Mar 3-5pm

A **degree D test** $f_n : \mathbb{R}^{n^2} \to \mathbb{R}$ deg $\leq D$ **strongly separates** if



$$\mathbb{E}_{P_n}[f] - \mathbb{E}_{Q_n}[f] \gg \sqrt{\max\{\mathrm{Var}_Q[f],\ \mathrm{Var}_P[f]\}}$$

"difference in means" $\gg$ "fluctuations".

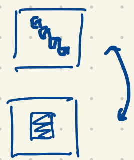
$= \Omega(\qquad)$

THM Given parameters n, k, λ, M

signal (λ), #communities (M)

$P_n \sim G(n, k, \lambda, M)$

$Q_n \sim G(n, k, \lambda, 1)$



"counting"

$D^5 \lambda^2 M^2 \left(\frac{k^2}{n} \vee 1\right) = o(1) \Rightarrow$ No deg D test weakly separates P_n, Q_n

$M^2 \lambda^2 \frac{k^2}{n} = \omega(1)$ & $k = \omega(1)$ $\Rightarrow$ Deg 1 test which strongly separates P_n, Q_n