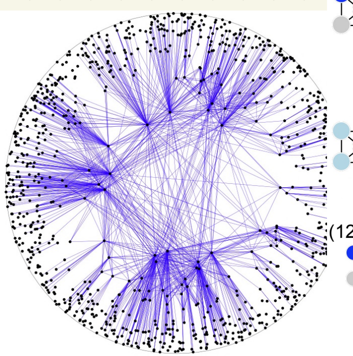
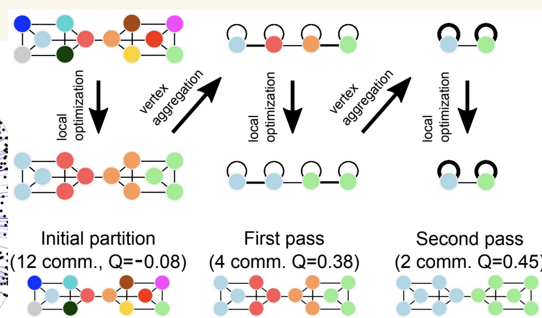
STRUCTURE IN NOISY NETWORKS

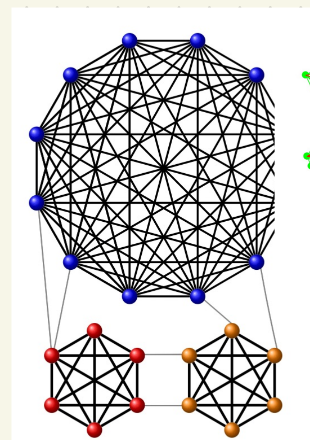
Fiona Skerman

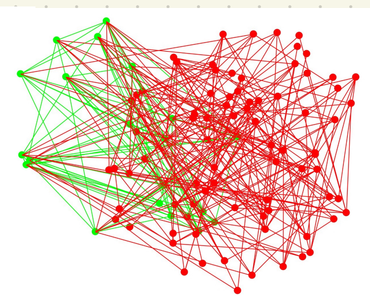
I: Modularity based clustering

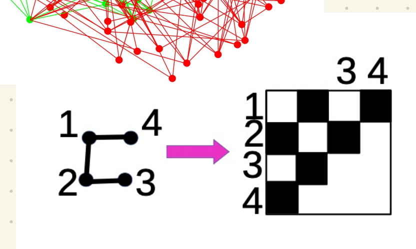
II: Fundamental limits of learning











I: Modularity-based clustering

- louvain: most commonly used method of community detection
- modularity: definitions + properties.

Example : Linguistics

$V \approx 3000$ climate change blogs

$E \sim$ based on links between blogs

Figure 1. The network of climate change blogs, colored by community.

Elgesam D, Steskal L. + Diakopoulos

"Structure and content of the discourse on climate change in the blogosphere"

Environmental Communication '2015.

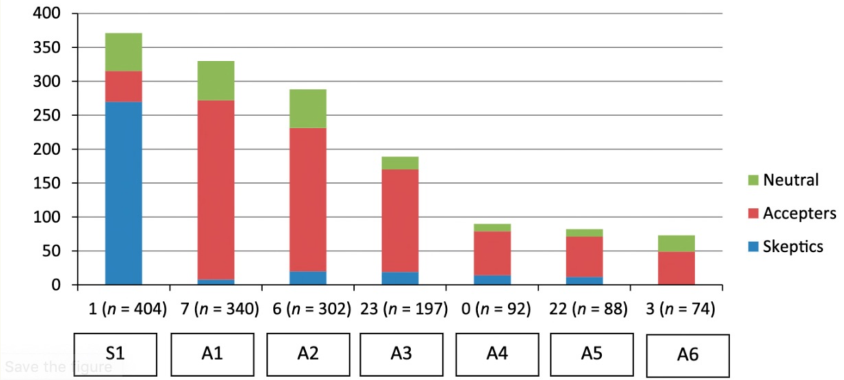


Figure 3. The distribution of skeptical, accepting, and neutral blogs in the seven largest among the central groups of blogs concerned with climate change.

Table 5. The top 15 collocates around "climate" in communities 1 (skeptic), 23 (accepter), and 7 (accepter) computed with the point-wise mutual information metric.

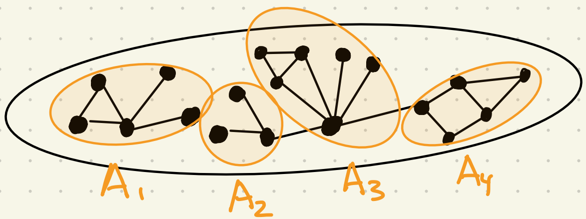
Top collocates of "CLIMATE" in the skeptical community S1	Top collocates of "CLIMATE" in the accepter community A3	Top collocates of "CLIMATE" in the accepter community A1
1 CLIMATE	1 DENIERS	1 POPPIN
2 SKEPTICS	2 SKEPTICS	2 DENIERS
3 ALARMISM	3 CLIMAT	3 SKEPTICS
4 DENIERS	4 DECADAL	4 OBAMA
5 IPCC	5 CONTRARIANS	5 WWW
6 DECADAL	6 OBAMA	6 EU'S
7 ALARMISTS	7 NOAA'S	7 CLIMATE
8 CLIMAT	8 AGW	8 YVO
9 CHANGE	9 WWW	9 NOAA'S
10 INTERGOVERNMENTAL	10 DENIER	10 WILDFIRES
11 OBAMA	11 CLIMATE	11 CHANGE'S
12 ANTHROPOGENIC	12 VAPOR	12 IPCC
13 AGW	13 ANTHROPOGENIC	13 ALARMISM
14 IPCC'S	14 ALARMISM	14 PACHAURI
15 WARMING	15 CONTRARIAN	15 DENIER

Reference corpus: The British National Corpus, approximately 100 million words.

Modularity

'meas. of how well a graph can be clustered'

G



$\mathcal{A} = \{A_1, \dots, A_4\}$

graph G, m edges. $\mathcal{A} = \{A_1, \dots, A_k\}$ vertex partition

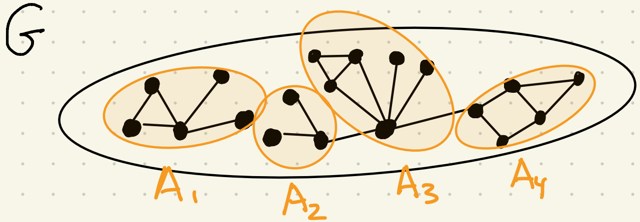
score of partition $\mathcal{A}$, $q_{\mathcal{A}}(G) =$

modularity of G $\quad q^*(G) = \max_{\mathcal{A}} q_{\mathcal{A}}(G)$

'higher values taken to indicate more community structure'

Modularity

'meas. of how well a graph can be clustered'



$\mathcal{A} = \{A_1, \dots, A_4\}$

graph G, m edges. $\mathcal{A} = \{A_1, \dots, A_k\}$ vertex partition

score of partition $\mathcal{A}$, $\quad q_{\mathcal{A}}(G) =$

modularity of G $\quad q^*(G) = \max_{\mathcal{A}} q_{\mathcal{A}}(G)$

'higher values taken to indicate more community structure'

Community Detection

input graph $G = (V, E)$

V: vertices / nodes

E: edges (weighted)

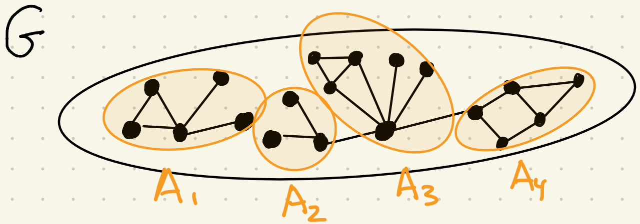
output vertex partition $\mathcal{A}$

'community division'

- modularity score NP-hard to opt.
- Louvain & Leiden ~ modularity based
  ~ iteratively build a partition
  local choices – maximise mod.
- most popular methods use modularity

Modularity

'meas. of how well a graph can be clustered'



$\mathcal{A} = \{A_1, \dots, A_4\}$

graph G, m edges. $\mathcal{A} = \{A_1, \dots, A_k\}$ vertex partition

score of partition $\mathcal{A}$, $q_{\mathcal{A}}(G) =$

modularity of G $\quad q^*(G) = \max_{\mathcal{A}} q_{\mathcal{A}}(G)$

$$q_{\mathcal{A}}(G) = \sum_{A \in \mathcal{A}} \underbrace{\frac{e(A)}{m}}_{\text{"edge contrib"}} - \underbrace{\frac{\mathrm{vol}(A)^2}{(2m)^2}}_{\text{"degree tax"}} = \frac{1}{2m} \sum_{A \in \mathcal{A}} \sum_{u,v \in A} \mathbb{1}[u \sim v] - \frac{d_u d_v}{2m}$$

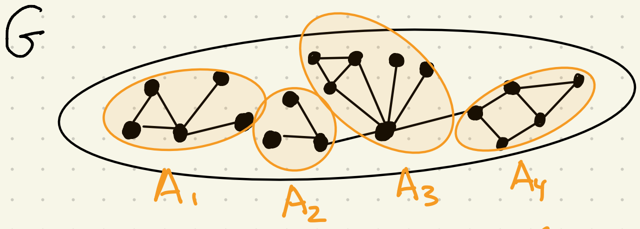
$\rightarrow \frac{|A|^2}{n^2}$ for regular graphs

$d_u = \#$ edges incident to u

$e(A) = \#$ edges in set A

$\mathrm{vol}(A) = \sum_{u \in A} d_u$

Modularity 'meas. of how well a graph can be clustered'



$\mathcal{A} = \{A_1, \dots, A_4\}$

graph G, m edges. $\mathcal{A} = \{A_1, \dots, A_k\}$ vertex partition

score of partition $\mathcal{A}$, $q_{\mathcal{A}}(G) =$

modularity of G $\quad q^*(G) = \max_{\mathcal{A}} q_{\mathcal{A}}(G)$

$$q_{\mathcal{A}}(G) = \underbrace{\sum_{A \in \mathcal{A}} \frac{e(A)}{m}}_{\text{"edge contrib."}} - \underbrace{\frac{\mathrm{vol}(A)^2}{(2m)^2}}_{\text{"degree tax"}} = \frac{1}{2m} \sum_{A \in \mathcal{A}} \sum_{u,v \in A} \mathbb{1}[u \sim v] - \frac{d_u d_v}{2m}$$

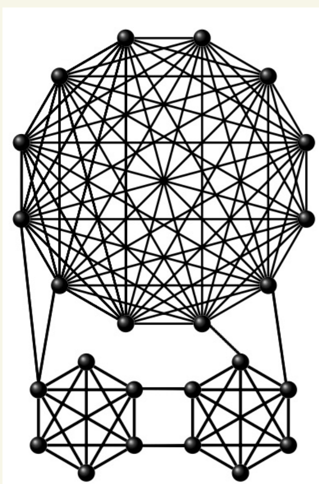
$\rightarrow \frac{|A|^2}{n^2}$ for regular graphs

d_u = # edges incident to u

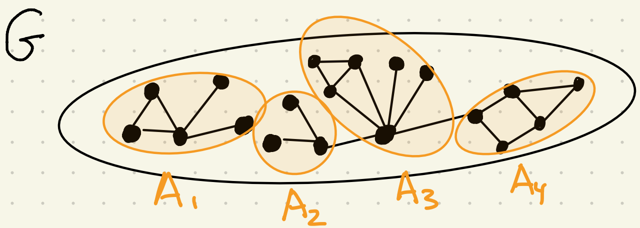
$e(A)$ = # edges in set A

$\mathrm{vol}(A) = \sum_{u \in A} d_u$

Example



Modularity
'meas. of how well a graph can be clustered'



$\mathcal{A} = \{A_1, \dots, A_4\}$

graph G, m edges. $\mathcal{A} = \{A_1, \dots, A_k\}$ vertex partition

score of partition $\mathcal{A}$, $q_{\mathcal{A}}(G) =$

modularity of G $\quad q^*(G) = \max_{\mathcal{A}} q_{\mathcal{A}}(G)$

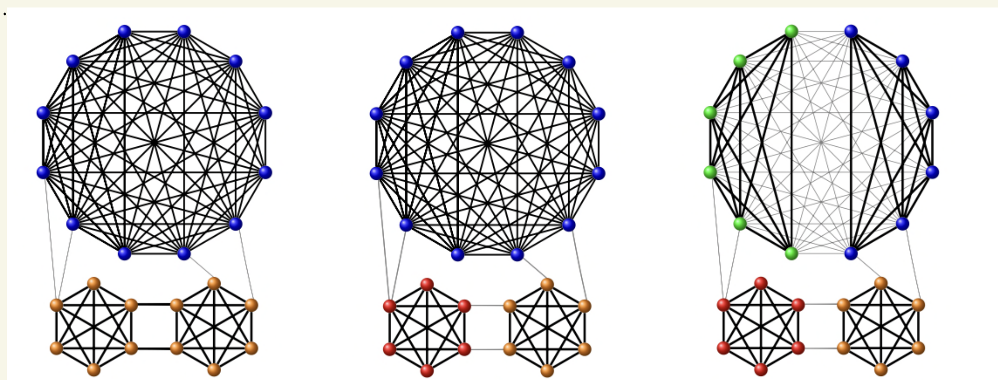
$$q_{\mathcal{A}}(G) = \sum_{A \in \mathcal{A}} \underbrace{\frac{e(A)}{m}}_{\text{"edge contrib"}} - \underbrace{\frac{\mathrm{vol}(A)^2}{(2m)^2}}_{\text{"degree tax"}} = \frac{1}{2m} \sum_{A \in \mathcal{A}} \sum_{u,v \in A} \mathbb{1}[u \sim v] - \frac{d_u d_v}{2m}$$

$d_u = \#$edges incident to u

$e(A) = \#$edges in set A

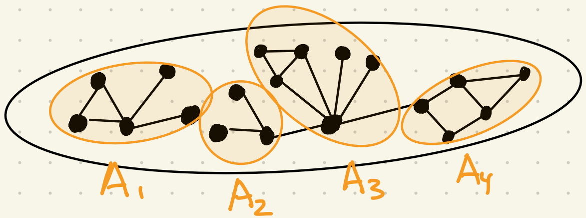
$\mathrm{vol}(A) = \sum_{u \in A} d_u$

Example



Modularity 'meas. of how well a graph can be clustered'

G



$\mathcal{A} = \{A_1, \dots, A_4\}$

graph G, m edges. $\mathcal{A} = \{A_1, \dots, A_k\}$ vertex partition

score of partition $\mathcal{A}$, $q_{\mathcal{A}}(G) =$

modularity of G $\quad q^*(G) = \max_{\mathcal{A}} q_{\mathcal{A}}(G)$

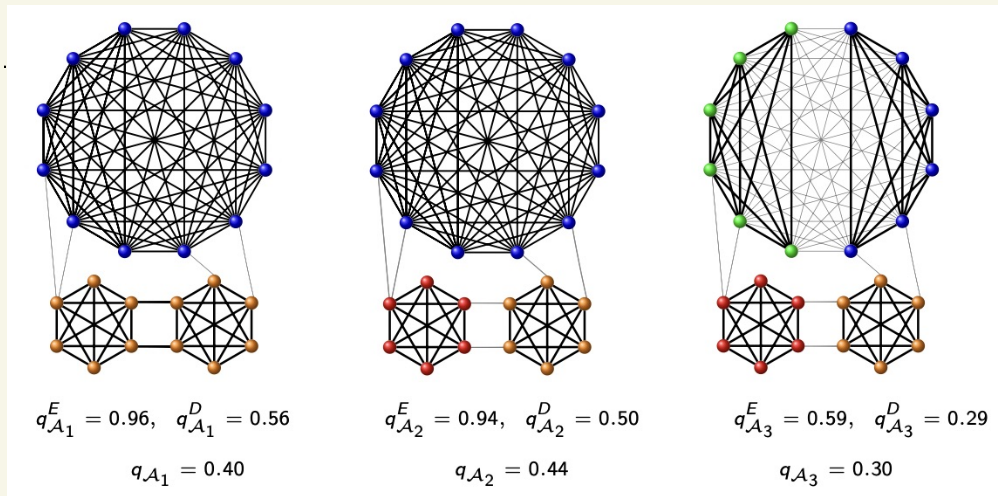
$$q_{\mathcal{A}}(G) = \underbrace{\sum_{A \in \mathcal{A}} \frac{e(A)}{m}}_{\text{"edge contrib."}} - \underbrace{\frac{\mathrm{vol}(A)^2}{(2m)^2}}_{\text{"degree tax"}} = \frac{1}{2m} \sum_{A \in \mathcal{A}} \sum_{u,v \in A} \mathbb{1}[u \sim v] - \frac{d_u d_v}{2m}$$

$d_u = \#$ edges incident to u

$e(A) = \#$ edges in set A

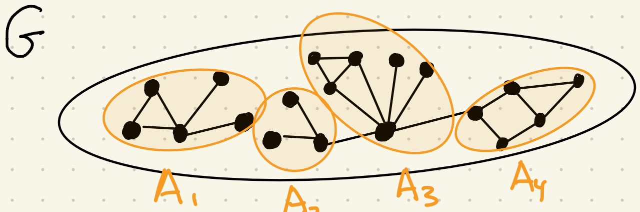
$\mathrm{vol}(A) = \sum_{u \in A} d_u$

Example



Modularity

'meas. of how well a graph can be clustered'



graph G, m edges. $\mathcal{A}=\{A_1,\dots,A_k\}$ vertex partition

score of partition $\mathcal{A}$, $q_{\mathcal{A}}(G)=$

modularity of G $\quad q^*(G)=\max_{\mathcal{A}} q_{\mathcal{A}}(G)$

"high vals taken to indicate more community structure"

$\mathcal{A}=\{A_1,\dots,A_4\}$

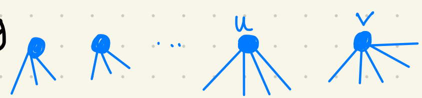
$$q_{\mathcal{A}}(G)=\sum_{A\in\mathcal{A}}\underbrace{\frac{e(A)}{m}}_{\text{"edge contrib"}}-\underbrace{\frac{vol(A)^2}{(2m)^2}}_{\text{"degree tax"}}=\frac{1}{2m}\sum_{A\in\mathcal{A}}\sum_{u,v\in A}\mathbb{1}[u\sim v]-\frac{d_u d_v}{2m}$$

d_u = # edges incident to u

$e(A)$ = # edges in set A

$vol(A)=\sum_{u\in A} d_u$

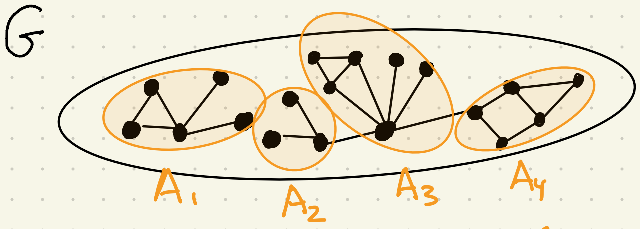
$R\sim$ config model



$$q_{\mathcal{A}}(G)\approx\frac{1}{m}\left(e^{int}_G(\mathcal{A})-\mathbb{E}\left[e^{int}_R(\mathcal{A})\right]\right)$$

Modularity

'meas. of how well a graph can be clustered'



$\mathcal{A} = \{A_1, \dots, A_4\}$

graph G, m edges. $\mathcal{A} = \{A_1, \dots, A_k\}$ vertex partition

score of partition $\mathcal{A}$, $q_{\mathcal{A}}(G) =$

modularity of G $\quad q^*(G) = \max_{\mathcal{A}} q_{\mathcal{A}}(G)$

"high vals taken to indicate more community structure"

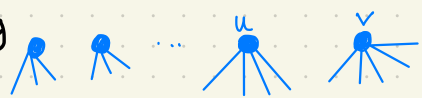
$$q_{\mathcal{A}}(G) = \sum_{A \in \mathcal{A}} \underbrace{\frac{e(A)}{m}}_{\text{"edge contrib"}} - \underbrace{\frac{\mathrm{vol}(A)^2}{(2m)^2}}_{\text{"degree tax"}} = \frac{1}{2m} \sum_{A \in \mathcal{A}} \sum_{u,v \in A} \mathbb{1}[u \sim v] - \frac{d_u d_v}{2m}$$

$d_u = \#$ edges incident to u

$e(A) = \#$ edges in set A

$\mathrm{vol}(A) = \sum_{u \in A} d_u$

$R \sim$ config model



$$q_{\mathcal{A}}(G) \approx \frac{1}{m}\left(e^{int}_G(\mathcal{A}) - \mathbb{E}[\, e^{int}_R(\mathcal{A})\,]\right)$$

- $u \neq v$ $\quad \mathbb{E}[\#\text{edges } u \sim v \text{ in } R] = \dfrac{d_u d_v}{2m-1}$
- $\mathbb{E}[\#\text{edges within parts of } \mathcal{A} \text{ in } R] = \sum_{A \in \mathcal{A}} \dfrac{\mathrm{vol}(A)(\mathrm{vol}(A)-1)}{2m(2m-1)}$

Fast unfolding of communities in large networks

VD **Blondel**, JL Guillaume, R Lambiotte… - Journal of statistical …, 2008 - iopscience.iop.org

We propose a simple method to extract the community structure of large networks. Our method is a heuristic method that is based on modularity optimization. It is shown to outperform all …

☆ Save 99 Cite Cited by 25562 Related articles All 35 versions

RESEARCH PAPER

Hub disruption in patients with chronic neck pain: a graph analytical approach

De Pauw, Robby[a,*]; Aerts, Hannelore[b]; Siugzdaite, Roma[c]; Meeus, Mira[a,d,e]; Coppieters, Iris[a,f,d]; Caeyenberghs, Karen[a,g]; Cagnie, Barbara[a]

forest mice

About 410 results (**0.07** sec)

Fast unfolding of communities in large networks

☑ Search within citing articles

[HTML] Molecular logic of cellular diversification in the **mouse** cerebral cortex

DJ Di Bella, E Habibi, RR Stickels, G Scalia, J Brown… - Nature, 2021 - nature.com

… To define the closest identity of these cells, we applied a multi-class random **forest** classifier trained on the wild-type cell types (Extended Data Fig. 11f). Most of the knockout cells were …

☆ Save 99 Cite Cited by 360 Related articles All 11 versions

[PDF] Identification of protein functions in **mouse** with a label space partition method

X Li, L Lu, L Chen - Math. Biosci. Eng, 2022 - aimspress.com

… This study presented a new multi-label classifier for identifying functions of **mouse** proteins. Due to the number of functional types, which were termed as labels in the classification …

☆ Save 99 Cite Cited by 31 Related articles All 5 versions »

[HTML] Molecular architecture of the **mouse** nervous system

A Zeisel, H Hochgerner, P Lönnerberg, A Johnsson… - Cell, 2018 - cell.com

… To assess the robustness of the clusters, we trained a random **forest** classifier to recognize cluster labels and then assessed its performance on held-out data (80% training set, 20% test …

☆ Save 99 Cite Cited by 2519 Related articles All 22 versions

[HTML] Decoding molecular and cellular heterogeneity of **mouse** nucleus accumbens

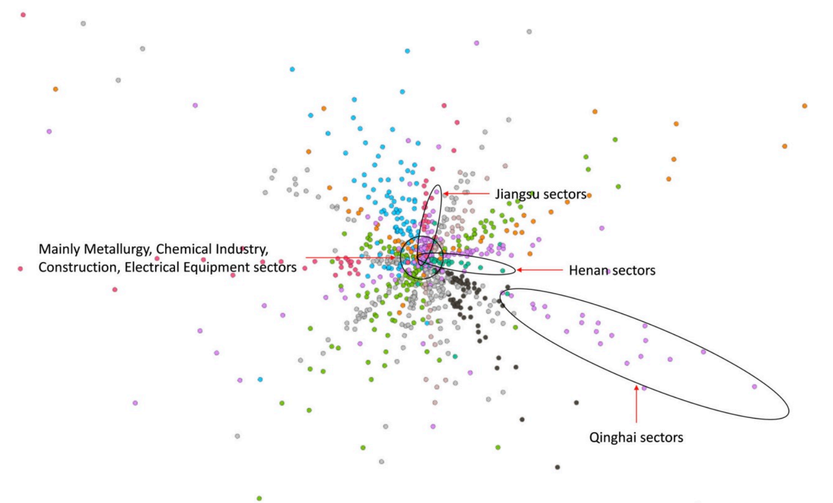
R Chen, TR Blosser, MN Djekidel, J Hao… - Nature …, 2021 - nature.com

… In this study, we generated a cell census of the **mouse** NAc using single-cell RNA sequencing and multiplexed error-robust fluorescence in situ hybridization, revealing a high level of …

THE IMPORTANCE OF SOCIAL NETWORKS AMONGST REFUGEES RESETTLED THROUGH THE COMMUNITY SPONSORSHIP SCHEME AND THE VULNERABLE PERSONS RESETTLEMENT SCHEME

Critical transmission sectors in embodied atmospheric mercury emission network in China

Kehan He[1] | Zhifu Mi[1] | Long Chen[2] | D'Maris Coffman[1] | Sai Liang[3]



Fast unfolding of communities in large networks

VD **Blondel**, JL Guillaume, R Lambiotte… - Journal of statistical …, 2008 - iopsci

We propose a simple method to extract the community structure of large network
is a heuristic method that is based on modularity optimization. It is shown to outp

OPEN

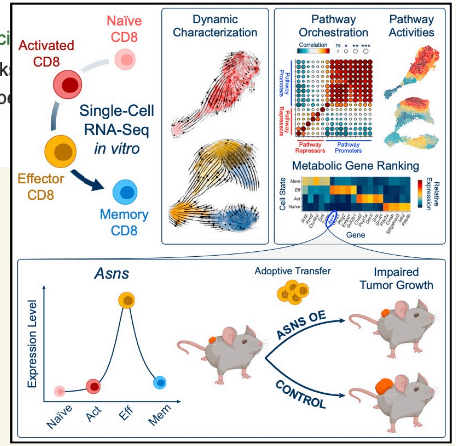
Inferring strategies from observations in long iterated Prisoner's dilemma experiments

Eladio Montero-Porras[1,5]✉, Jelena Grujić[1,5], Elías Fernández Domingos[1,2] & Tom Lenaerts[1,2,3,4]

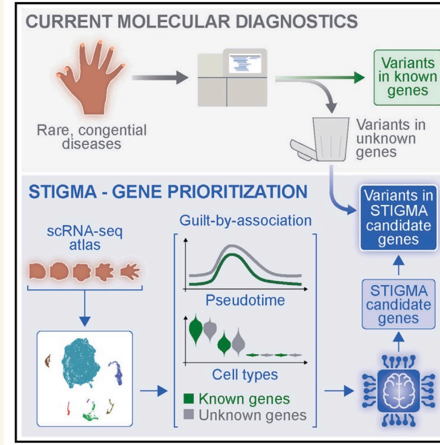
Cell Reports Article

CD8$^+$ T cell metabolic rewiring defined by scRNA-seq identifies a critical role of ASNS expression dynamics in T cell differentiation

Graphical abstract



Graphical abstract



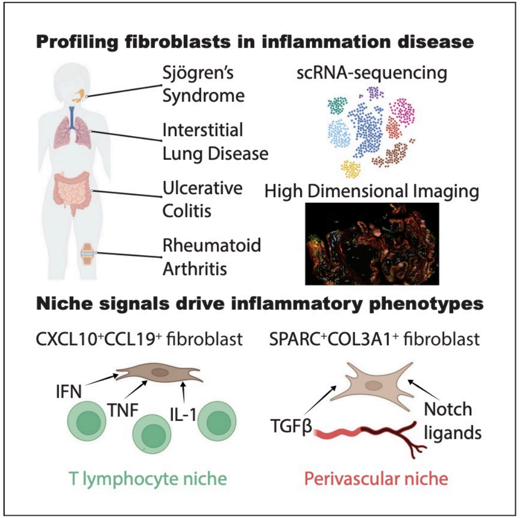
Untangling the

STRUCTURE *and* DYNAMICS

of ecological networks

Bernat Bramon Mora

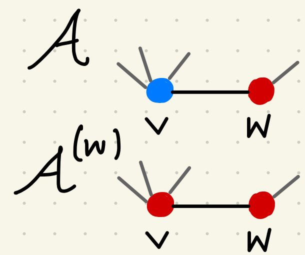
June 1, 2019



Modularity — louvain alg.

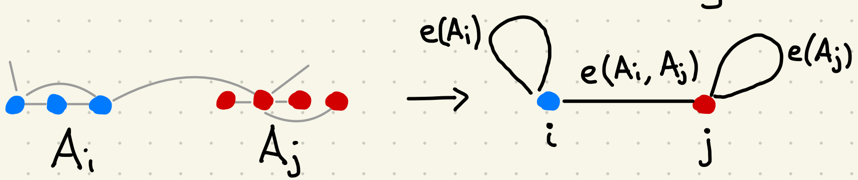
- set $\mathcal{A} = \{ \{v_1\}, \{v_2\}, \dots, \{v_n\} \}$

(I)
- pick a unif. random labelling of vertices $1, \dots, n$
- for v $1, \dots, n$
  - for each w nhbr of v
    - construct $\mathcal{A}^{(w)}$ re-colour v with colour of w.
    - if $q_{\mathcal{A}^{(w)}} > q_{\mathcal{A}}$ $\quad \mathcal{A} \to \mathcal{A}^{(w)}$



- if no change → (II)

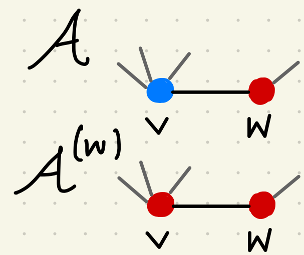
(II)
- construct G' by shrinking each colour class to a single vertex
- if $G' = G$ output $\mathcal{A}$
  else → (I)



Modularity — Louvain alg.

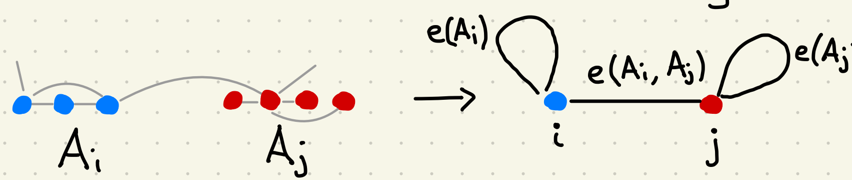
- set $\mathcal{A} = \{ \{v_1\}, \{v_2\}, \dots, \{v_n\} \}$

I
- pick a unif. random labelling of vertices $1, \dots, n$
- for v $1, \dots, n$
  - for each w nhbr of v
    - construct $\mathcal{A}^{(w)}$ re-colour v with colour of w.
    - if $q_{\mathcal{A}^{(w)}} > q_{\mathcal{A}}$ $\quad \mathcal{A} \to \mathcal{A}^{(w)}$



- if no change $\to$ II

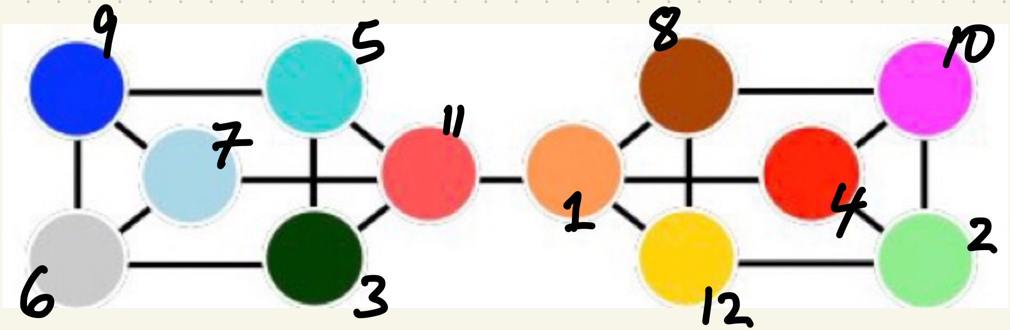
II
- construct G' by shrinking each colour class to a single vertex
- if $G' = G$ output $\mathcal{A}$
  else $\to$ I

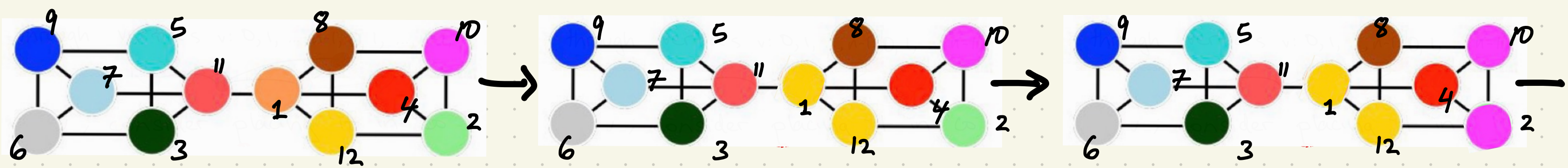


weighted graph, with loops.

Modularity — louvain alg.

- pick a unif. random labelling of vertices 1, ..., n





etc; until no local move increases modularity ...

Fig. LAMBIOTTE

Modularity louvain alg.

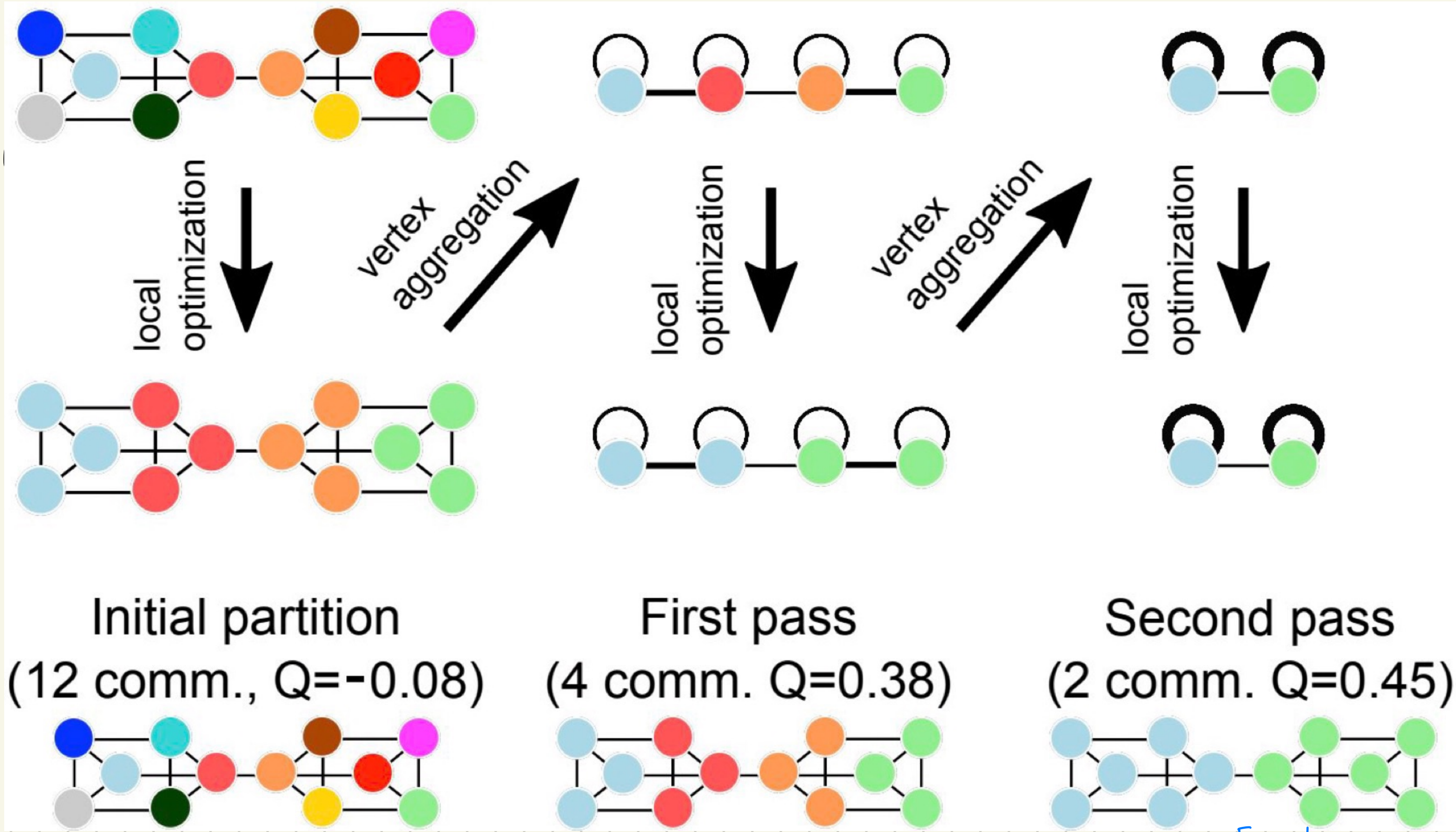
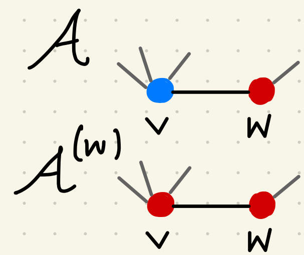


Fig. Lambiotte

Modularity $\boxed{\text{louvain alg.}}$

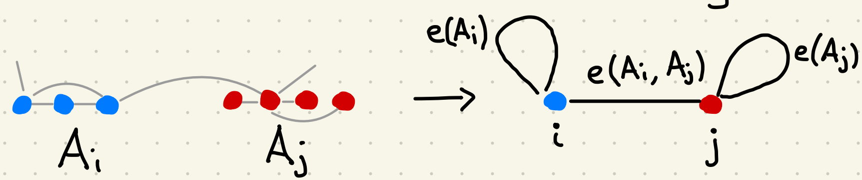
- set $\mathcal{A} = \{\{v_1\}, \{v_2\}, \dots, \{v_n\}\}$

Ⅰ
- pick a unif. random labelling of vertices $1, \dots, n$
- for v $1, \dots, n$
  - for each w nhbr of v
    - construct $\mathcal{A}^{(w)}$ re-colour v with colour of w.
    - if $q_{\mathcal{A}^{(w)}} > q_{\mathcal{A}}$ $\quad \mathcal{A} \to \mathcal{A}^{(w)}$
- if no change $\to$ Ⅱ



Ⅱ
- construct G' by shrinking each colour class to a single vertex
- if $G' = G$ output $\mathcal{A}$

  else $\to$ Ⅰ



weighted graph, with loops.

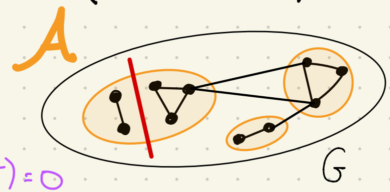
Modularity Properties

$\mathcal{A} \in OPT(G)$ i.e $q_{\mathcal{A}}(G) = q^*(G)$

$\Rightarrow \forall A \in \mathcal{A}$ $G[A]$ conn. (+ isolated vert)

$\Rightarrow$ pendant vertex in same part 

$\Rightarrow q^*(\text{star}) = 0$



Modularity value:

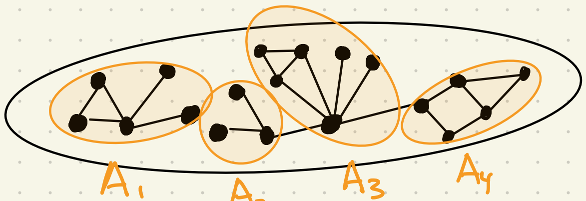
Robust to small perturbations in edge set

$$\left| q^*(G) - q^*(G \setminus E') \right| < \frac{2|E'|}{e(G)}$$

$$\forall \mathcal{A}: \quad \left| q_{\mathcal{A}}(G) - q_{\mathcal{A}}(G \setminus E') \right| < \frac{2|E'|}{e(G)}$$

Modularity 'meas. of how well a graph can be clustered'

G



$\mathcal{A} = \{A_1, \dots, A_4\}$

graph G, m edges. $\mathcal{A} = \{A_1, \dots, A_k\}$ vertex partition

score of partition $\mathcal{A}$, $q_{\mathcal{A}}(G) =$

modularity of G $\quad q^*(G) = \max_{\mathcal{A}} q_{\mathcal{A}}(G)$

"high vals taken to indicate more community structure"

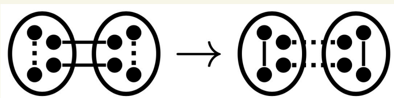
$$q_{\mathcal{A}}(G) = \underbrace{\sum_{A \in \mathcal{A}} \frac{e(A)}{m}}_{\text{"edge contrib"}} - \underbrace{\frac{\mathrm{vol}(A)^2}{(2m)^2}}_{\text{"degree tax"}} = \frac{1}{2m} \sum_{A \in \mathcal{A}} \sum_{u,v \in A} \mathbb{1}[u \sim v] - \frac{d_u d_v}{2m}$$

d_u = # edges incident to u

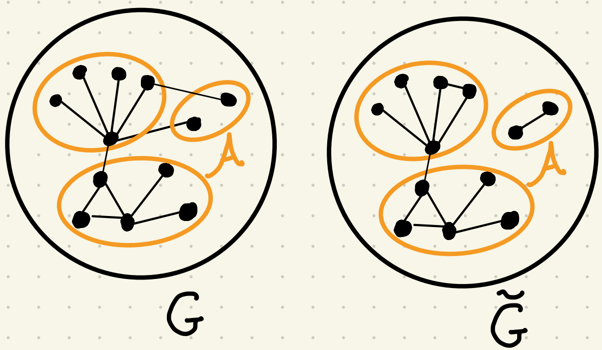
$e(A)$ = # edges in set A

$\mathrm{vol}(A) = \sum_{u \in A} d_u$

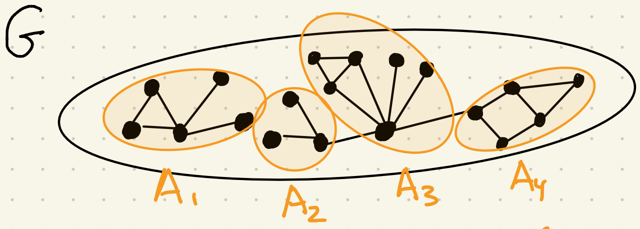
Fix $\mathcal{A}$, which $G \to \tilde{G}$ ensures $q_{\mathcal{A}}(\tilde{G}) > q_{\mathcal{A}}(G)$?



✓



Modularity 'meas. of how well a graph can be clustered'



$\mathcal{A} = \{A_1, \dots, A_4\}$

graph G, m edges. $\mathcal{A} = \{A_1, \dots, A_k\}$ vertex partition

score of partition $\mathcal{A}$, $q_{\mathcal{A}}(G) =$

modularity of G $\quad q^*(G) = \max_{\mathcal{A}} q_{\mathcal{A}}(G)$

"high vals taken to indicate more community structure"

$$q_{\mathcal{A}}(G) = \sum_{A \in \mathcal{A}} \underbrace{\frac{e(A)}{m}}_{\text{"edge contrib"}} - \underbrace{\frac{\mathrm{vol}(A)^2}{(2m)^2}}_{\text{"degree tax"}} = \frac{1}{2m} \sum_{A \in \mathcal{A}} \sum_{u,v \in A} \mathbb{1}[u \sim v] - \frac{d_u d_v}{2m}$$

d_u = # edges incident to u

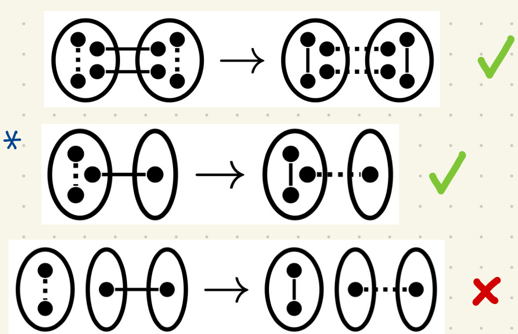
$e(A)$ = # edges in set A

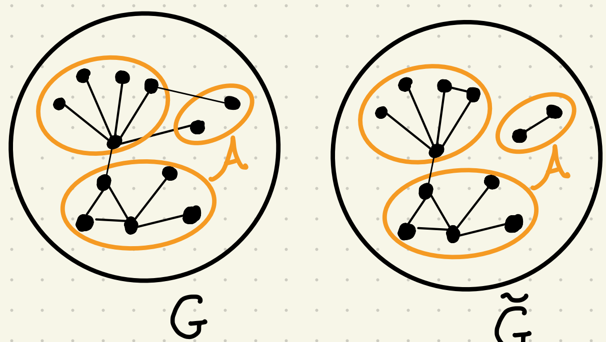
$\mathrm{vol}(A) = \sum_{u \in A} d_u$

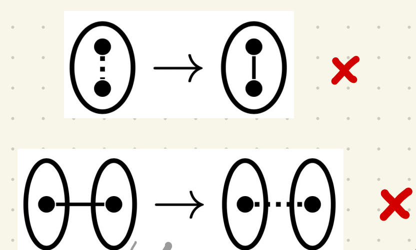
Fix $\mathcal{A}$, which $G \to \tilde{G}$ ensures $q_{\mathcal{A}}(\tilde{G}) > q_{\mathcal{A}}(G)$?

*

*





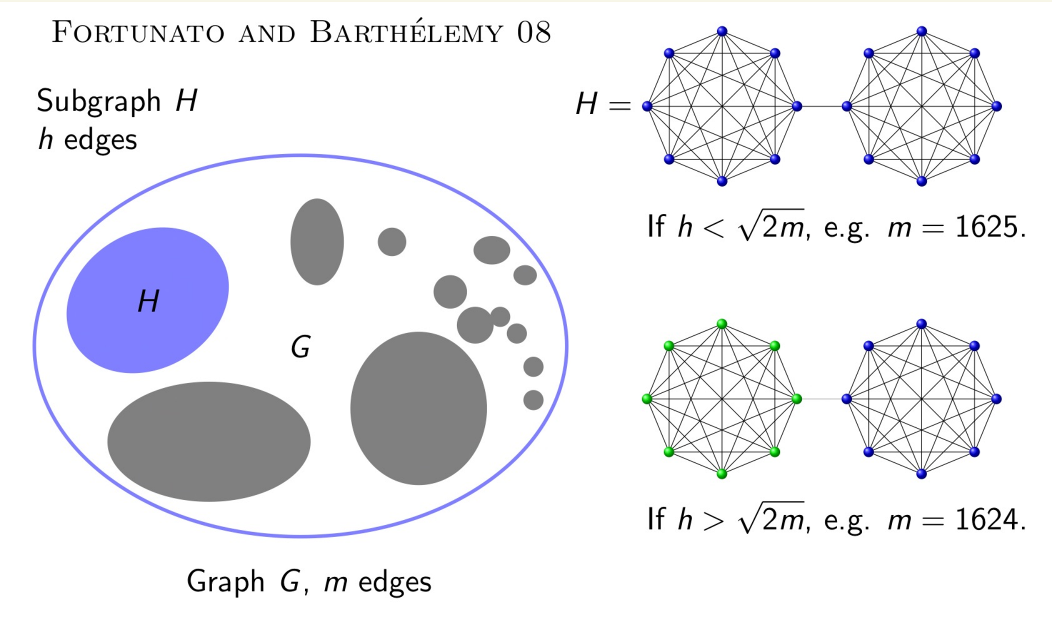


(✓ if $|\mathcal{A}| = 2$, $q_{\mathcal{A}}(G) \geq 0$)

*Trajanovski, Wang, Van Mieghem '12.

Modularity Properties

- Resolution limit   OPT partition fragile

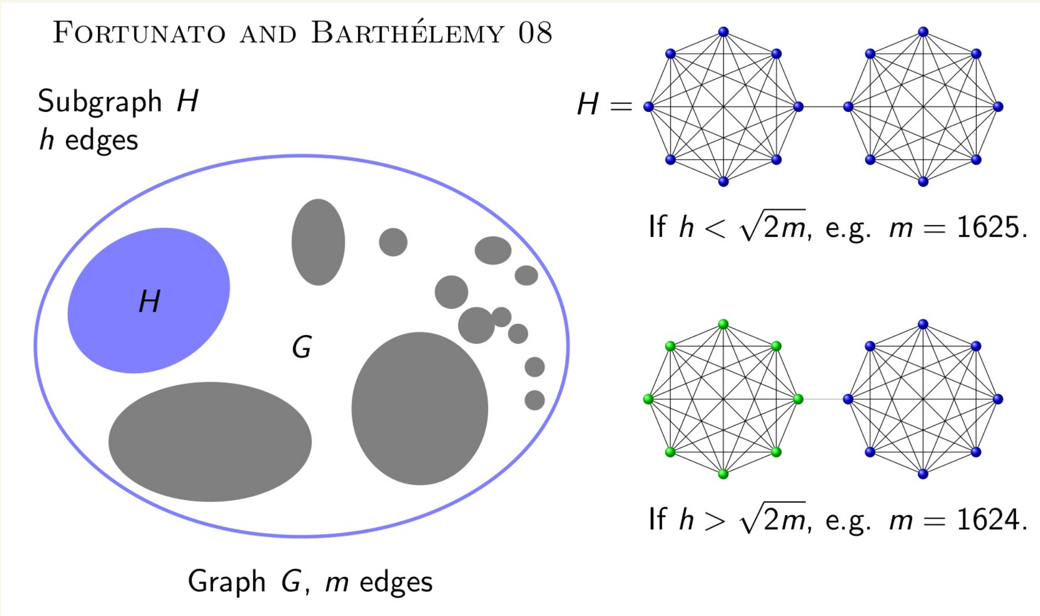


- Modularity value: Robust to small perturbations in edge set

$$\left| q^*(G) - q^*(G \setminus E') \right| < \frac{2|E'|}{e(G)}$$

Modularity Properties

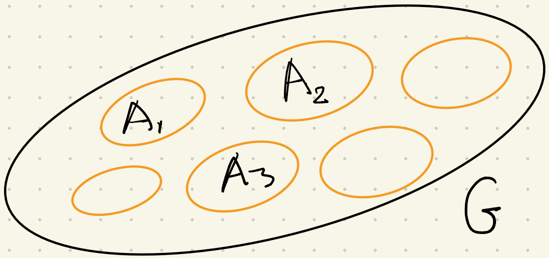
- Resolution limit — OPT partition fragile



- Modularity value: Robust to small perturbations in edge set

$$\forall \mathcal{A}: \quad \left| q_{\mathcal{A}}(G) - q_{\mathcal{A}}(G \setminus E') \right|, \ \left| q^*(G) - q^*(G \setminus E') \right| < \frac{2|E'|}{e(G)}$$

Modularity 'meas. of how well a graph can be clustered'



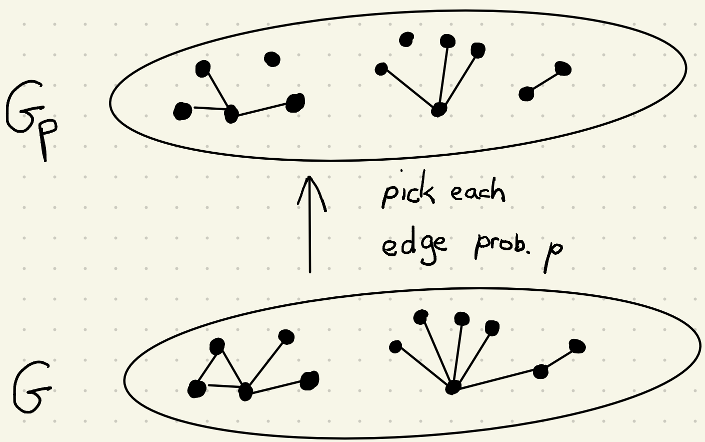
graph G, m edges. $\mathcal{A} = \{A_1, \ldots, A_k\}$ vertex partition

score of partition $\mathcal{A}$, $q_{\mathcal{A}}(G) =$

modularity of G $\quad q^*(G) = \max_{\mathcal{A}} q_{\mathcal{A}}(G)$

"high vals taken to indicate more community structure"

Sampling

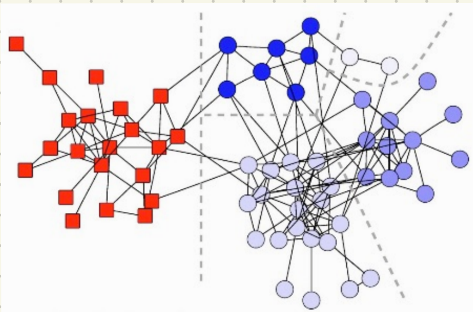


$G_p = (V, E_p)$ $\quad E_p$ each edge kept indep. prob. p

$G = (V, E)$ fixed graph

Simulations

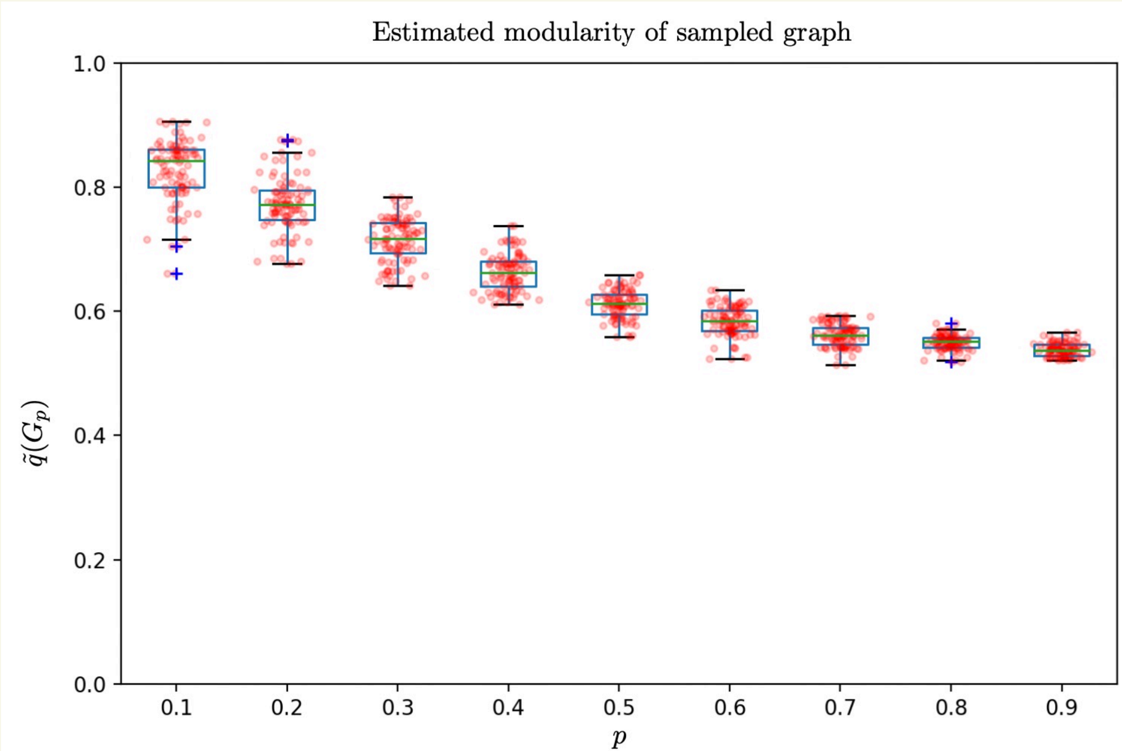
Dolphin Network [LUSSEAU]



$|V| = 62 \qquad |E| = 152$

$q^*(G) = 0.\underline{529}\ldots$ (3 dec places)

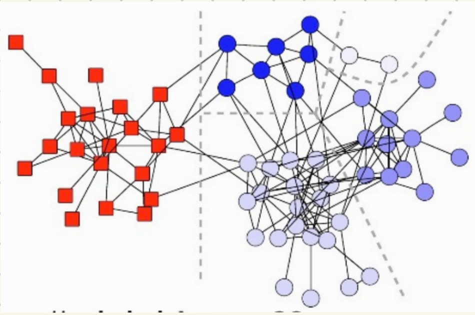
[BRANDES + '08]



To estimate modularity take max of 200 runs of Lovain and Leiden algs.

Simulations

Dolphin Network [LUSSEAU]

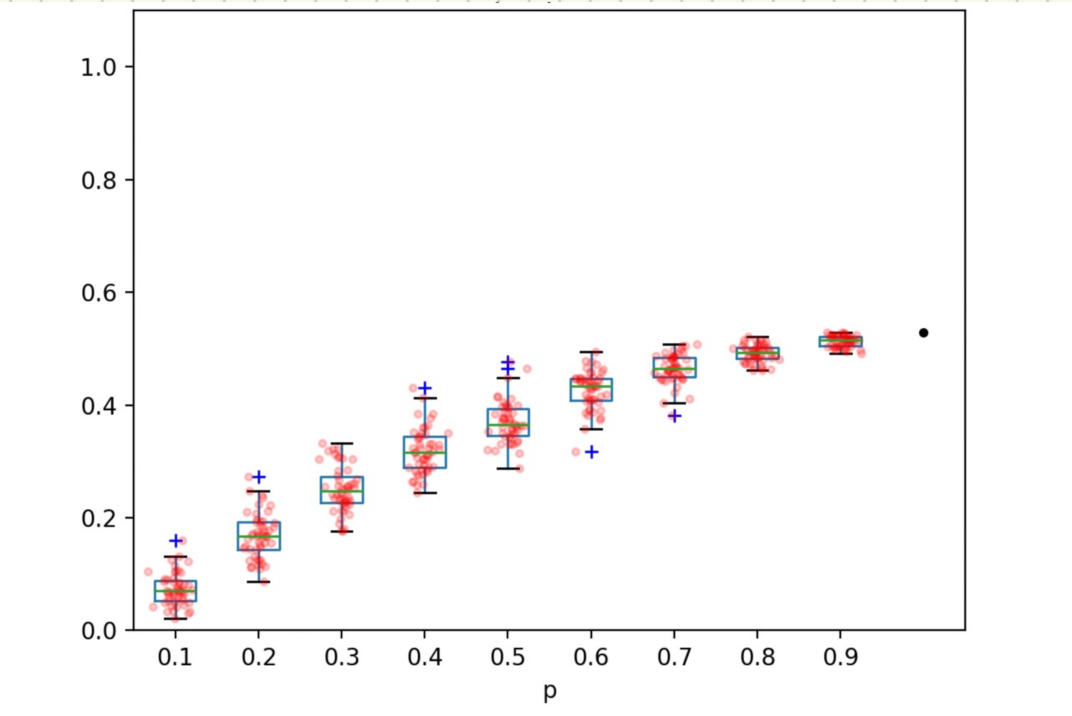


$|V| = 62 \quad |E| = 152$

$q^*(G) = 0.\underline{529}\ldots$ (3 dec places)

[BRANDES + '08]

Modularity score of underlying dolphin network using partition ~ OPT of sampled network.

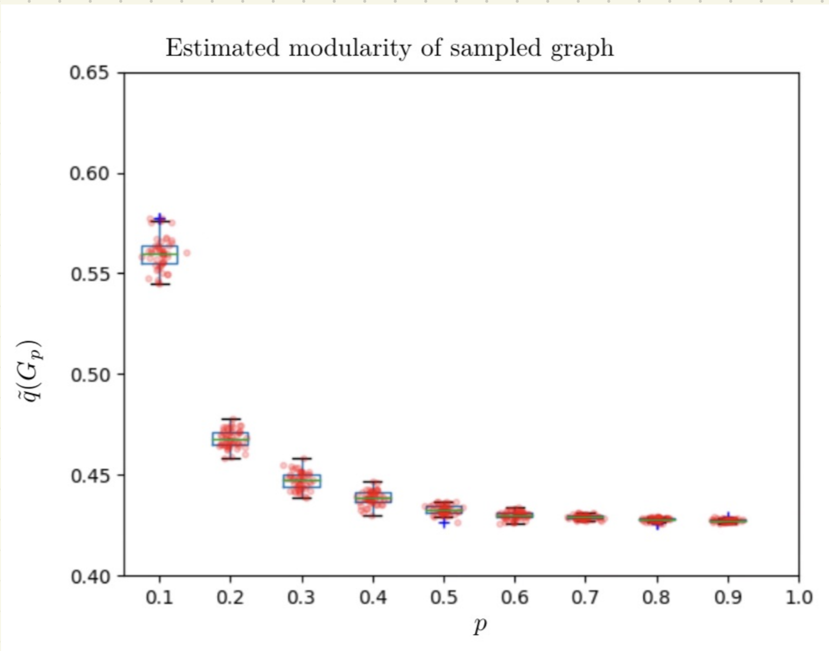


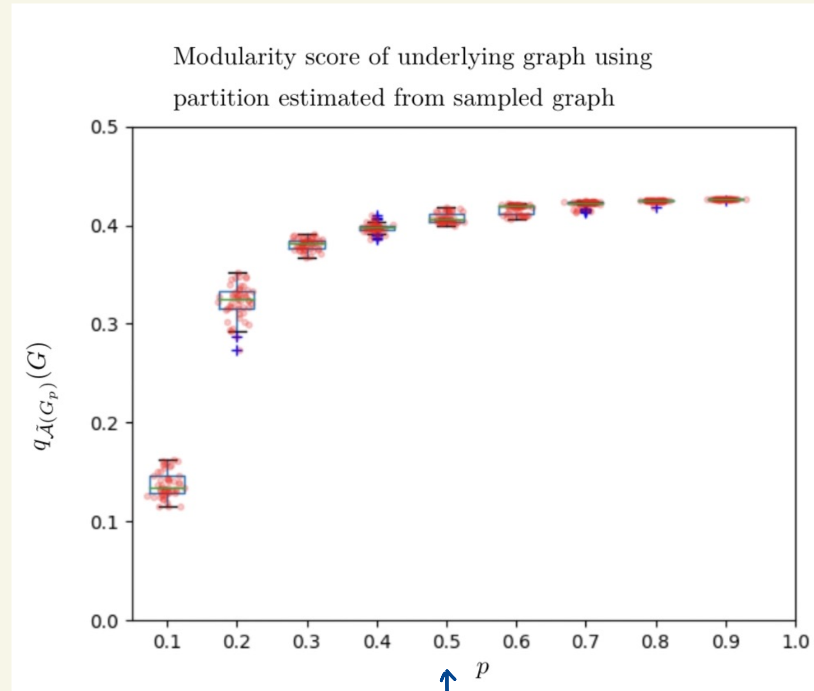
To estimate OPT of sampled G_p take max of 200 runs of Lovain and Leiden algs.

↑ on G_p

US Political Blogs [Adamic, Glance '2005]

Graph $V \sim 1500$ $E \sim 16000$





seeing half the edges ['how good the partition is']

≈ " all " "

II: Fundamental limits of learning

- when can we detect / recover planted communities ?
- when can we do this fast?

Planted Community

$G \sim G(n, \underbrace{p}_{\text{signal}}, \underbrace{q}_{\text{noise}}, k)$, $\quad i \in K$ w.prob $\frac{k}{n}$ $\quad A_{ij} = \begin{cases} Be(p) & i,j \in K \\ Be(q) & \text{ow} \end{cases}$

$p > q$

n points

~k 'community' nodes

~n-k 'non-community' "

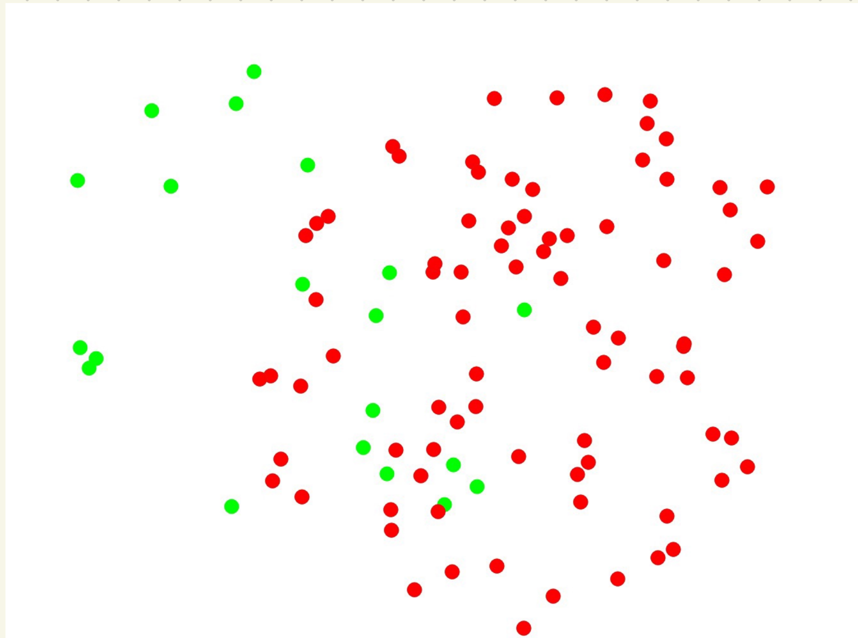


Fig: Jiaming Xu, Duke

Planted Community

$G \sim G(n, \underset{\text{signal}}{p}, \underset{\text{noise}}{q}, k)$, $\quad p > q \quad$ $i \in K$ w.prob $\frac{k}{n}$

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

n points

k 'community' nodes

n-k 'non-community' "

with prob. p

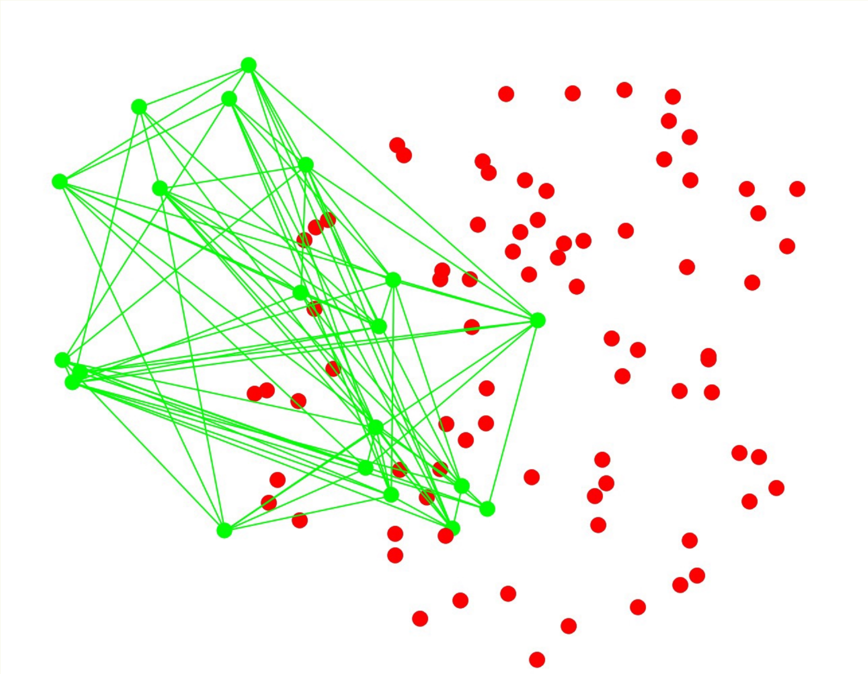


Fig: Jiaming Xu, Duke

Planted Community

$G \sim G(\overset{\text{signal}}{p},\ \overset{\text{noise}}{q}$ — $G(n, p, q, k)$, $\quad p > q$

$i \in K$ w.prob $\frac{k}{n}$

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

n points

k 'community' nodes

n-k 'non-community' "

with prob. p

" " q

" " q

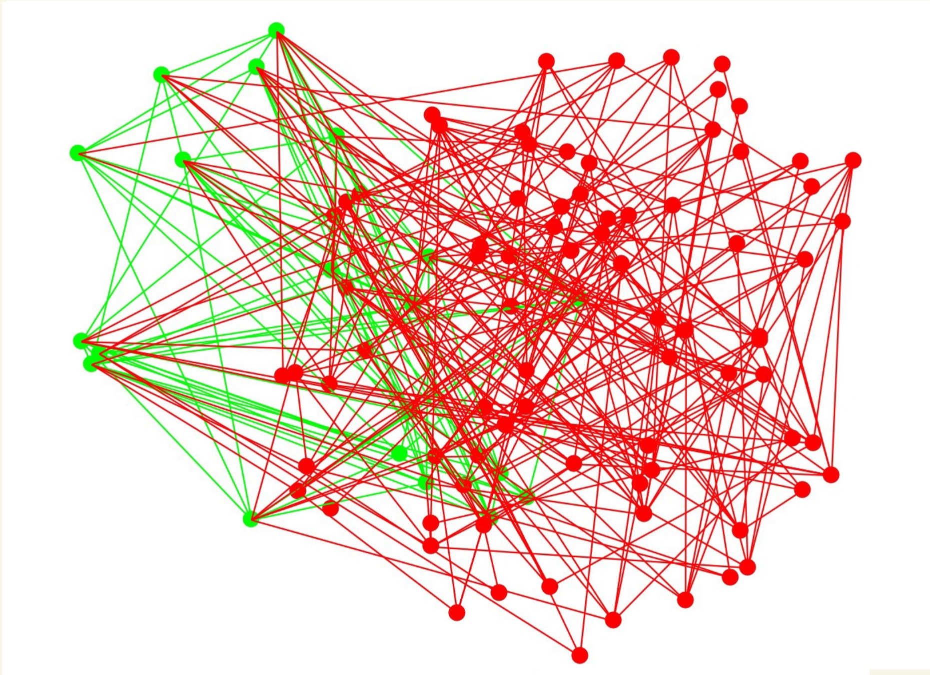


Fig: Jiaming Xu, Duke

Planted Community

$G \sim G(n, \underset{\text{signal}}{p}, \underset{\text{noise}}{q}, k)$, $i \in K$ w.prob $\frac{k}{n}$ $A_{ij} = \begin{cases} Be(p) & i,j \in K \\ Be(q) & \text{ow} \end{cases}$

$p > q$

Process

n points

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

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

Output

unlabelled graph

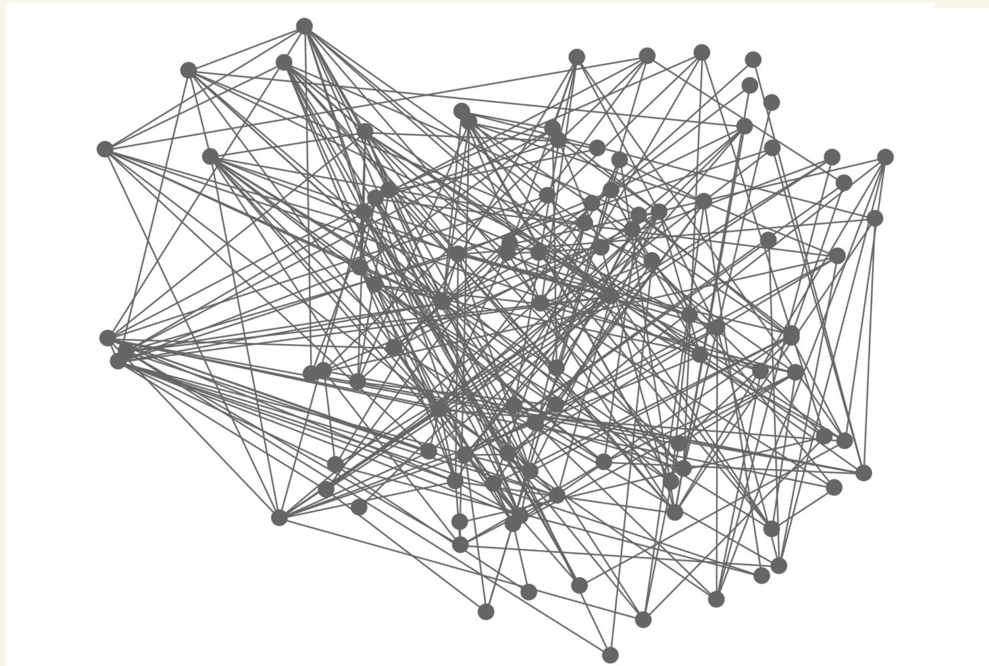


Fig: Jiaming Xu, Duke

Planted Community

$$G \sim G(\underbrace{n, p}_{\text{signal}}, \underbrace{q}_{\text{noise}}, k), \quad i \in K \text{ w.prob } \frac{k}{n} \qquad A_{ij} = \begin{cases} Be(p) & i,j \in K \\ Be(q) & o.w \end{cases}$$

$p > q$

<u>Process</u>

n points

k 'community' nodes

n-k 'non-community' "

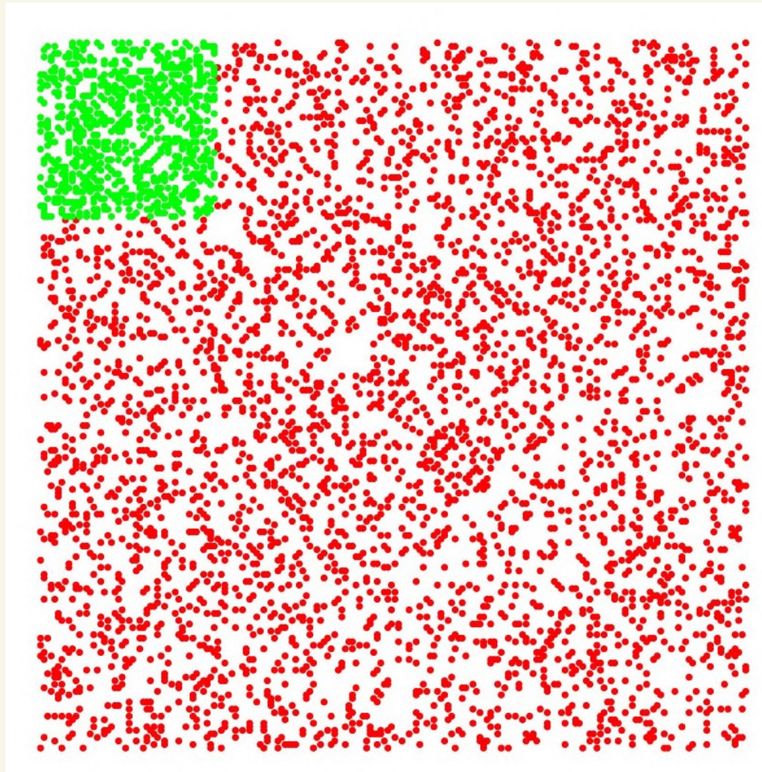
with prob. p

" " q

" " q

<u>Output</u>

unlabelled graph



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

Fig: Jiaming Xu, Duke

Planted Community

$G \sim G(\underset{\text{signal}}{n, p}, \underset{\text{noise}}{q}, k)$, $i \in K$ w.prob $\frac{k}{n}$

$p > q$

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

Process

n points

k 'community' nodes

n-k 'non-community' "

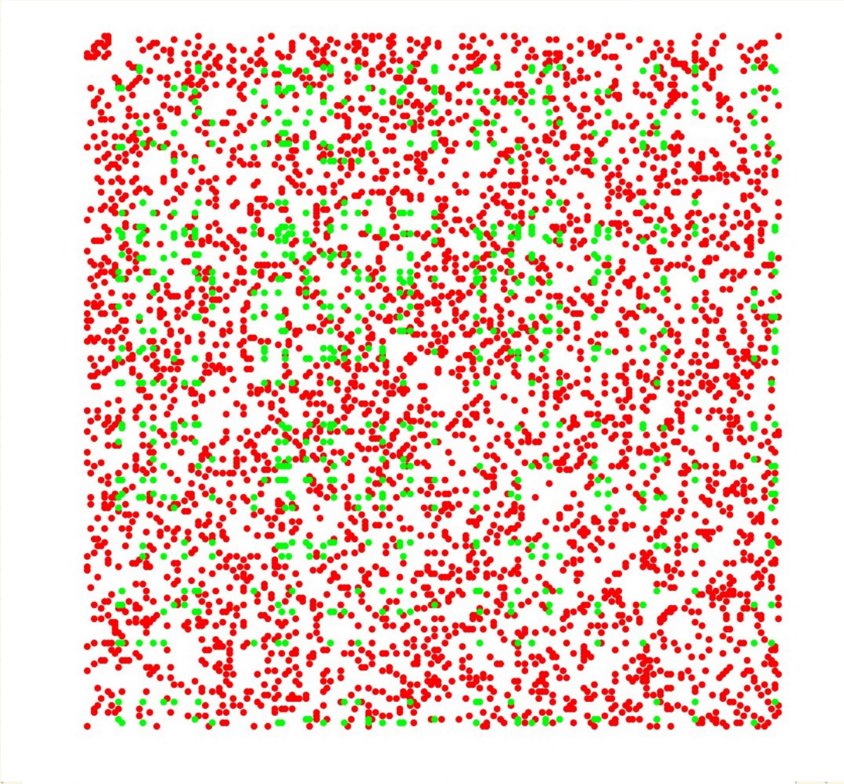
with prob. p

" " q

" " q

Output

unlabelled graph



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

Fig: Jiaming Xu, Duke

Planted Community

$$G \sim G(n, \underbrace{p}_{\text{signal}}, \underbrace{q}_{\text{noise}}, k), \quad i \in K \text{ w.prob } \frac{k}{n}$$

$$p > q$$

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

Process

n points

k 'community' nodes

n-k 'non-community' "

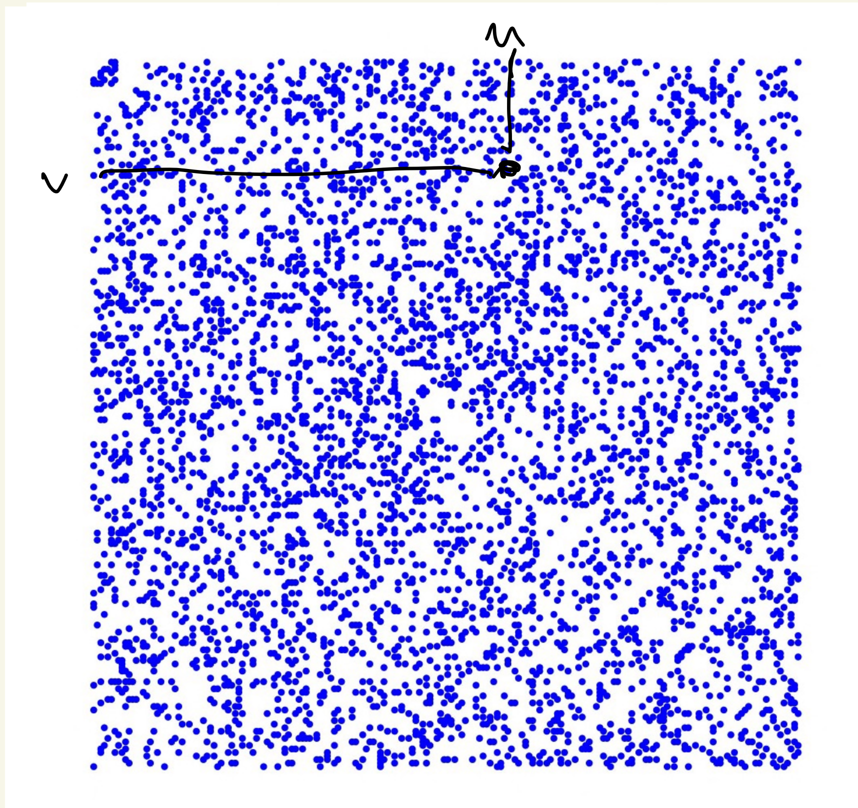
with prob. p

" " q

" " q

Output

unlabelled graph



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

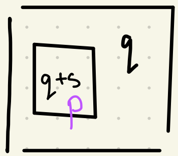
Fig: Jiaming Xu, Duke

Planted Dense subgraph / Submatrix

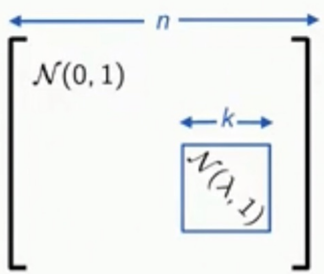
Hypothesis Testing : asymptotics

$H_0: G(n, q)$

$H_1: G(n, k, s, q)$



$\lambda = \frac{s}{\sqrt{q}}$



H_0 : i.i.d. $\mathcal{N}(0,1)$

H_1 : submatrix of $\mathcal{N}(\lambda, 1)$

- $n \to \infty$
- for what
  - $k(n)$ size of planted structure
  - $\lambda(n)$ strength of signal
- can we find test ϕ

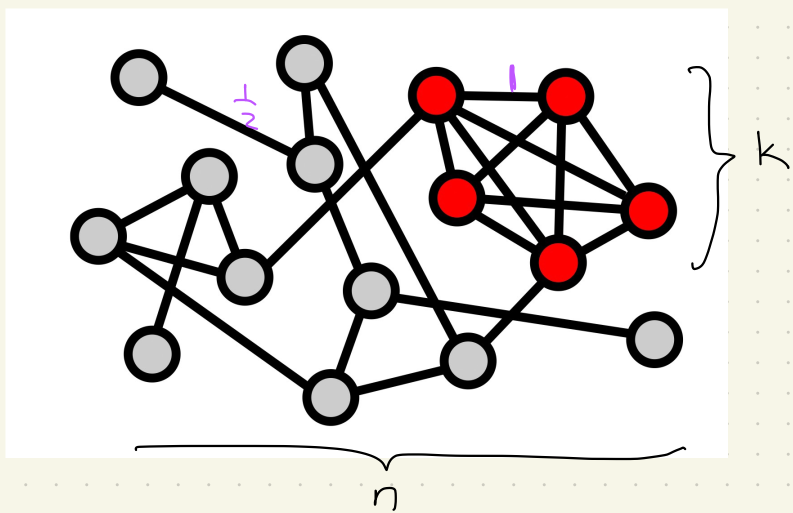
$$\mathbb{P}_0(\phi(Y) = 1) + \mathbb{P}_1(\phi(Y) = 0) \to 0$$

- when is there a fast test ?

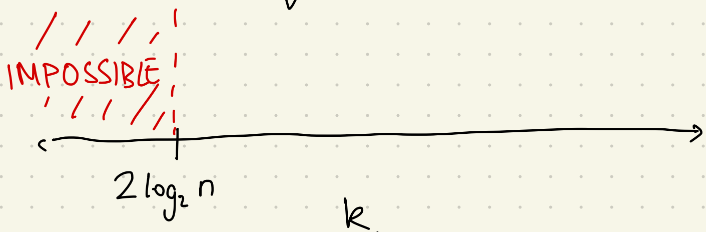
Planted Clique $G \sim G(n, \frac{1}{2}, k)$, $v \in K$ with prob. $\frac{k}{n}$ $\quad A_{uv} = \begin{cases} 1 & u, v \in K \\ Be(\frac{1}{2}) & \text{ow} \end{cases}$

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

Q: When can we find planted clique?

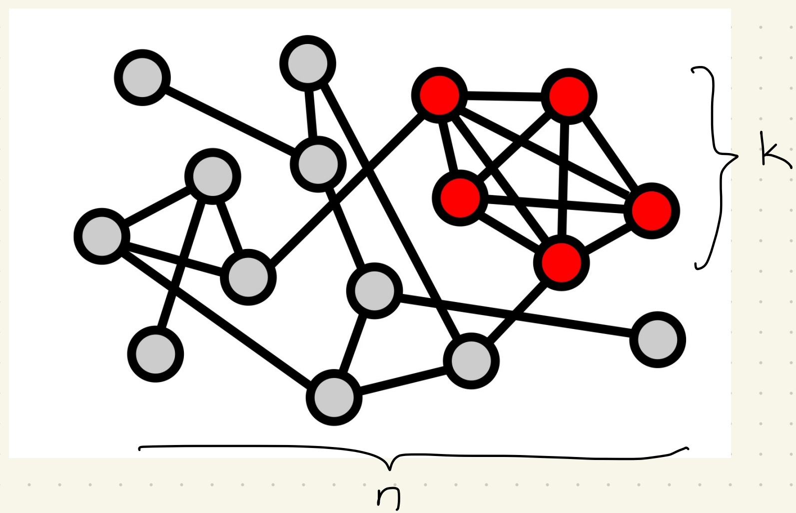


Planted Clique $G \sim G(n, \frac{1}{2}, k)$, $v \in K$ with prob. $\frac{k}{n}$ $\quad A_{uv} = \begin{cases} 1 & u, v \in K \\ Be(\frac{1}{2}) & \text{ow} \end{cases}$



if $k \leq 2\log_2 n$

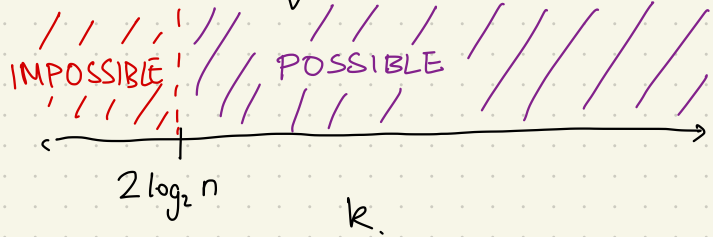
$G' \sim G(n, \frac{1}{2})$: largest clique whp $\sim 2\log_2 n$. $\Rightarrow$ can't find 'planted' one in amongst 'background' one.



Planted Clique $G \sim G(n, \frac{1}{2}, k)$,

$v \in K$ with prob. $\frac{k}{n}$

$$A_{uv} = \begin{cases} 1 & u, v \in K \\ Be(\frac{1}{2}) & \text{ow} \end{cases}$$



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

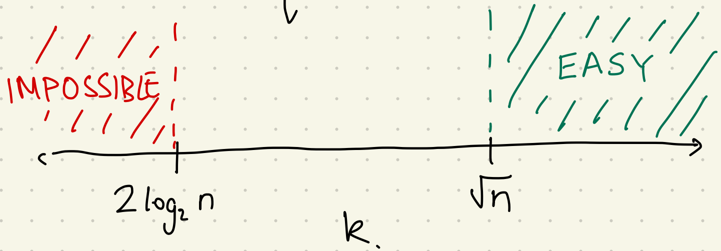
① BRUTE-FORCE

Search all k-vertex subsets

$\hat{K}$ first clique found.

if $k \geqslant (2+\varepsilon)\log n$ $\quad P(\hat{K} = K) \to 1$

Planted Clique $G \sim G(n, \frac{1}{2}, k)$, $v \in K$ with prob. $\frac{k}{n}$ $A_{uv} = \begin{cases} 1 & u,v \in K \\ Be(\frac{1}{2}) & \text{ow} \end{cases}$



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

$$k \geq \sqrt{n}\log n \quad \Rightarrow \quad P(\hat{K} = K) \to 1.$$

Planted Clique $G \sim G(n, \tfrac{1}{2}, k)$, $v \in K$ with prob. $\frac{k}{n}$ $\quad A_{uv} = \begin{cases} 1 & u, v \in K \\ Be(\tfrac{1}{2}) & \text{ow} \end{cases}$

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

Methods to find clique

$$K \overset{u}{\in} \binom{[n]}{k}$$

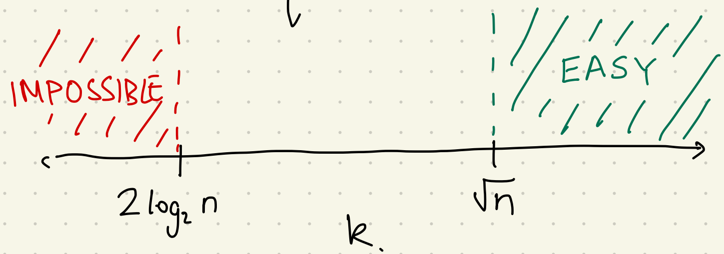
① DEGREE TEST

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

$$k \geq \sqrt{n}\log n \implies P(\hat{K} = K) \to 1.$$

Planted Clique $G \sim G(n, \frac{1}{2}, k)$,

$v \in K$ with prob. $\frac{k}{n}$ $\quad A_{uv} = \begin{cases} 1 & u, v \in K \\ Be(\frac{1}{2}) & \text{ow} \end{cases}$



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

$$k \geq \sqrt{n}\log n \quad \Longrightarrow \quad P(\hat{K} = K) \to 1.$$

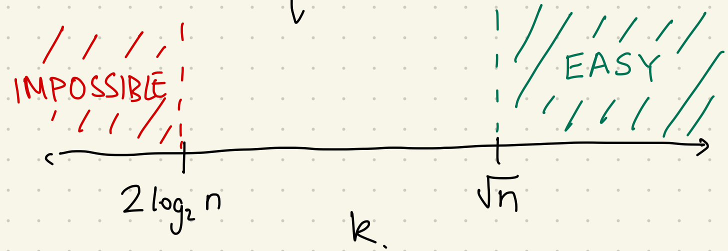
② SPECTRAL METHOD $\quad W_{ij} = \begin{cases} 2A_{ij} - 1 & i \neq j \\ 0 & \text{o.w} \end{cases}$

(i) u top eigenvector of W

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

Planted Clique

$G \sim G(n, \frac{1}{2}, k)$, $\quad v \in K$ with prob. $\frac{k}{n}$ $\quad A_{uv} = \begin{cases} 1 & u, v \in K \\ Be(\frac{1}{2}) & \text{o.w} \end{cases}$



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

$$k \geq \sqrt{n} \log n \implies P(\hat{K} = K) \to 1.$$

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

(i) u top eigenvector of W

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

$$k \geq \sqrt{n} \log n \implies P(\hat{K} = K) \to 1$$

Planted Clique

$G \sim G(n, \frac{1}{2}, k)$, $v \in K$ with prob. $\frac{k}{n}$ $A_{uv} = \begin{cases} 1 & u, v \in K \\ Be(\frac{1}{2}) & \text{o.w} \end{cases}$

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

$$k \geq \sqrt{n}\log n \implies P(\hat{K} = K) \to 1.$$

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) $\hat{K}$ index vector of k largest $|u_i|$

$$k \geq \sqrt{n}\log n \implies P(\hat{K} = K) \to 1$$

③ SDP METHOD   Yes. If $k = \Omega(\sqrt{n})$.

Planted Clique $G \sim G(n, \frac{1}{2}, k)$, $v \in K$ with prob. $\frac{k}{n}$ $\quad A_{uv} = \begin{cases} 1 & u, v \in K \\ Be(\frac{1}{2}) & \text{ow} \end{cases}$

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

$k \geq \sqrt{n} \log n \implies P(\hat{K} = K) \to 1.$

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

(i) u top eigenvector of W

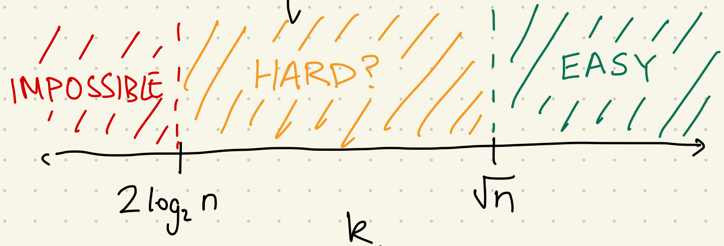
(ii) (threshold) $\hat{K}$ index vector of k largest $|u_i|$

$k \geq \sqrt{n} \log n \implies P(\hat{K} = K) \to 1$

③ SDP METHOD Yes. If $k = \Omega(\sqrt{n})$.

Planted Clique $G \sim G(n, \frac{1}{2}, k)$,

$v \in K$ with prob. $\frac{k}{n}$ $\qquad A_{uv} = \begin{cases} 1 & u, v \in K \\ Be(\frac{1}{2}) & \text{o.w} \end{cases}$



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

Methods to find clique

① DEGREE TEST

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

$$k \geq \sqrt{n} \log n \quad \Rightarrow \quad P(\hat{K} = K) \to 1.$$

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

(i) u top eigenvector of W

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

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

③ SDP METHOD $\quad$ Yes. If $k = \Omega(\sqrt{n})$.

* Rigorous Evidence suggesting no poly-time alg

- reductions (avg case)
- restricted class of alg
  - low deg poly
    - subgraph tests
      - # edges, #△'s, ...
    - spectral methods.

Planted Clique $G \sim G(n, \frac{1}{2}, k)$,

$v \in K$ with prob. $\frac{k}{n}$

$$A_{uv} = \begin{cases} 1 & u, v \in K \\ Be(\frac{1}{2}) & \text{o.w} \end{cases}$$

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

Methods to find clique

① DEGREE TEST

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

$$k \geq \sqrt{n} \log n \implies P(\hat{K} = K) \to 1.$$

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

(i) u top eigenvector of W

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

$$k \geq \sqrt{n} \log n \implies P(\hat{K} = K) \to 1$$

③ SDP METHOD Yes. If $k = \Omega(\sqrt{n})$.

* Rigorous Evidence suggesting no poly-time alg

- reductions (avg case)
- restricted class of alg

  low deg poly

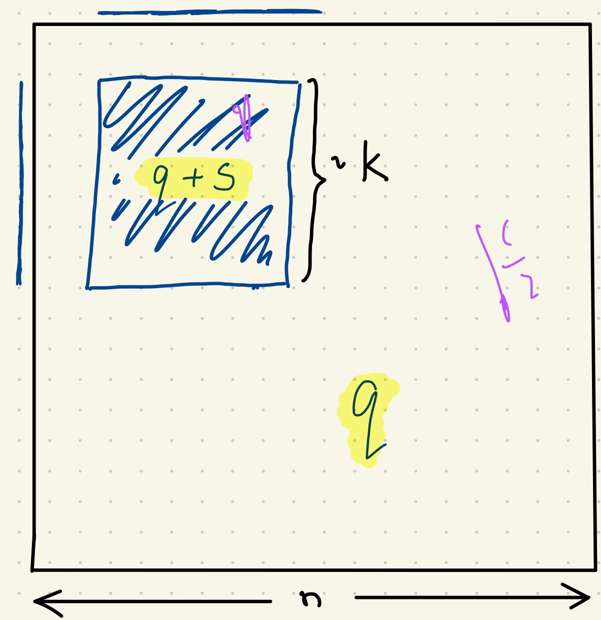
  - subgraph tests
    - # edges, # $\triangle$'s, ...
  - spectral methods.

Planted Dense Subgraph

$G(n, k, q, s)$

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

Observe $Y_{uv} \sim \begin{cases} \mathrm{Ber}(\overset{p}{q+s}) & \sigma_u = \sigma_v = 1 \\ \mathrm{Ber}(q) & \text{o.w.} \end{cases}$



Algorithmic Qns

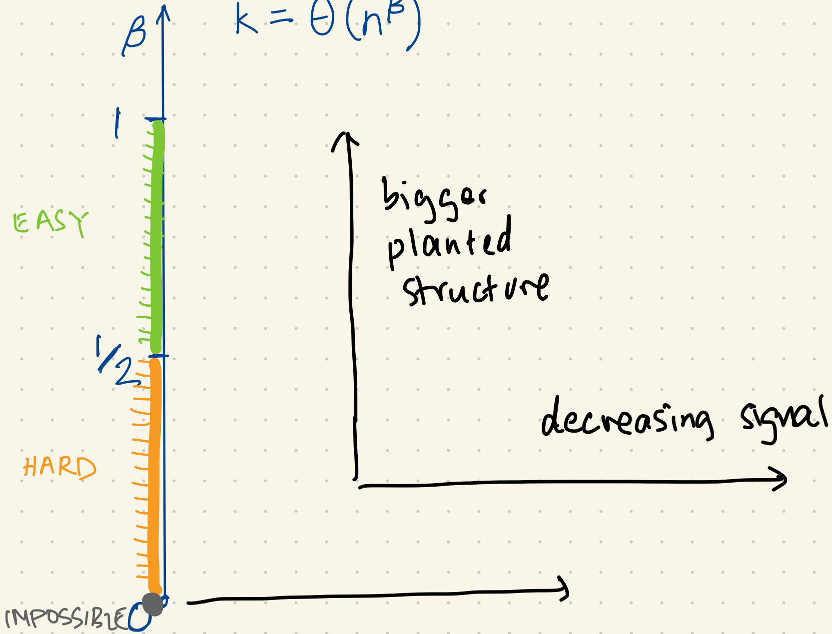
- Detection : determine if whp sample from planted model or not
- Recovery : given sample from planted model find community (exactly? weakly corr?)
- "Counting" ....?

CONTEXT

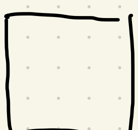
PLANTED DENSE SUBGRAPH

Detection

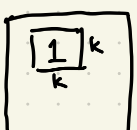
$k = \Theta(n^\beta)$



H_0

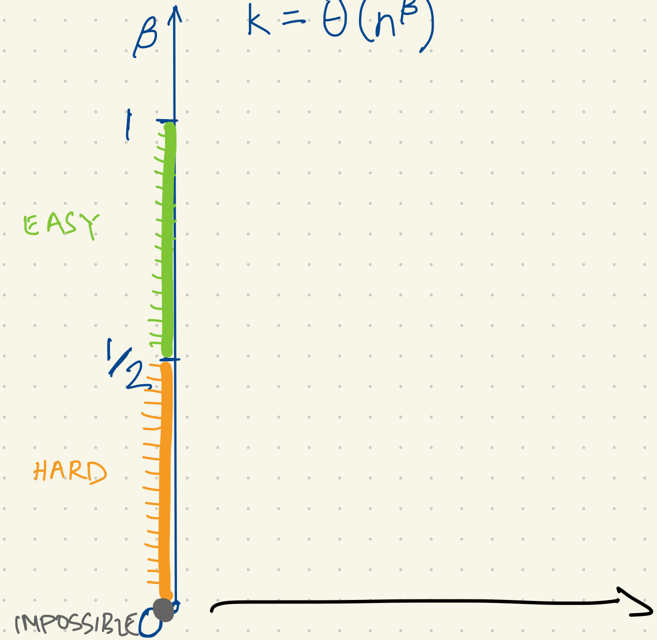


vs. H_1

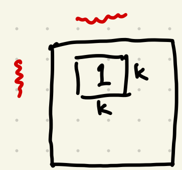


Recovery

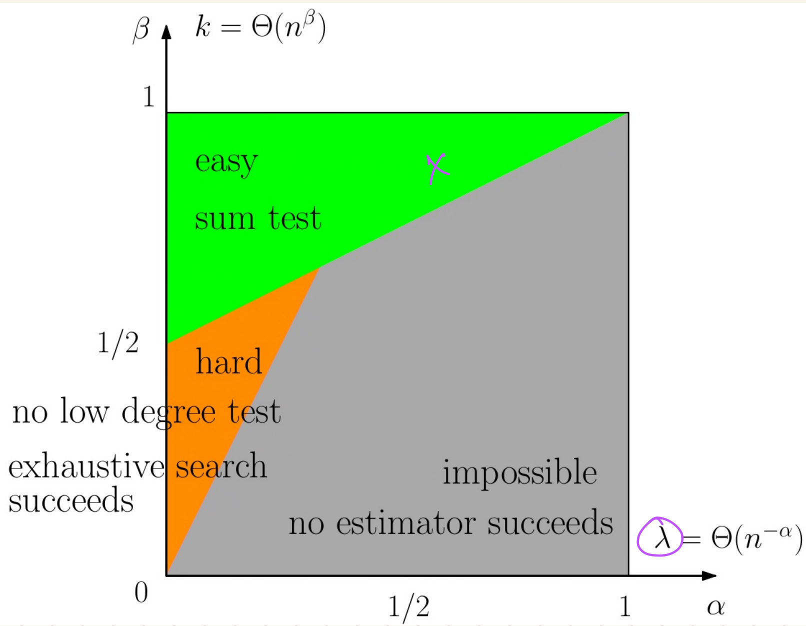
$k = \Theta(n^\beta)$



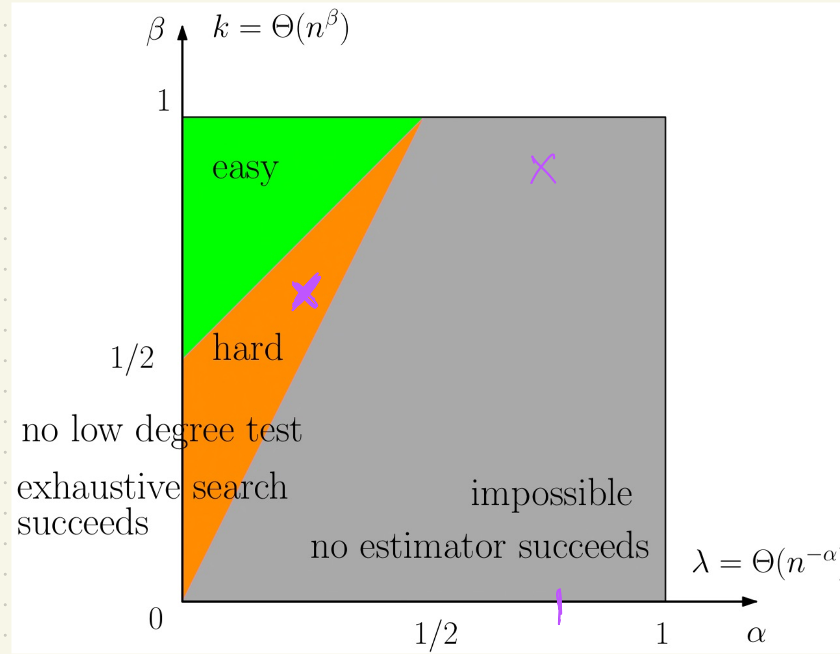
recover



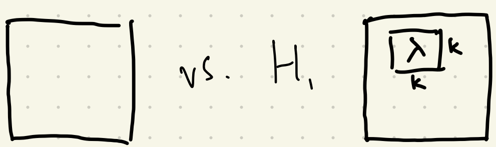
Detection



Recovery



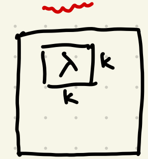
H_0 vs. H_1



$N(\lambda, 1)$

$\lambda = \frac{s}{\sqrt{q}}$

recover

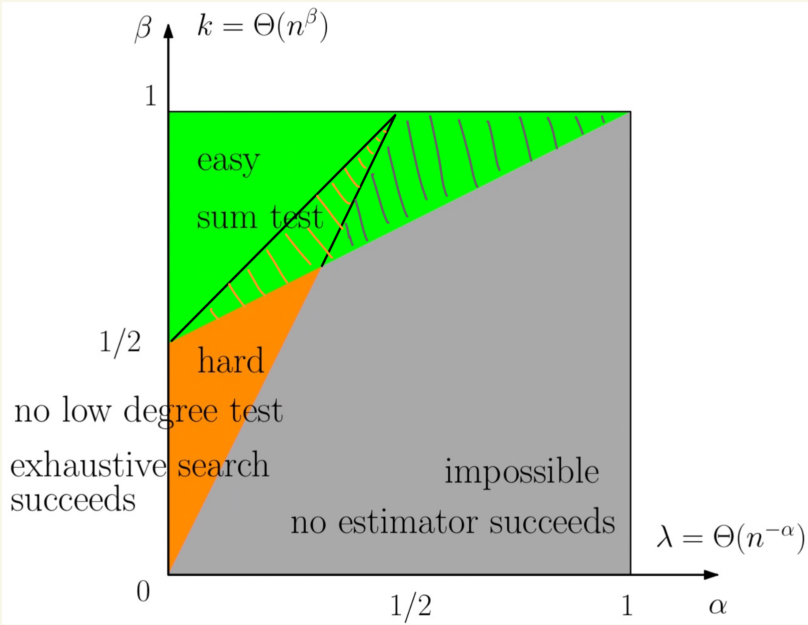


REFS: MANY AUTHORS. BI13, BIS15, MW15, CX16, DM14, CLR17, HWX17, BBH18, GJS19, BMR20, BBP05, BS06, FP07, CDF09, BGN11, SWPN09, KBRS11, BKR^{+}11, ACD11, BWZ20, SW22

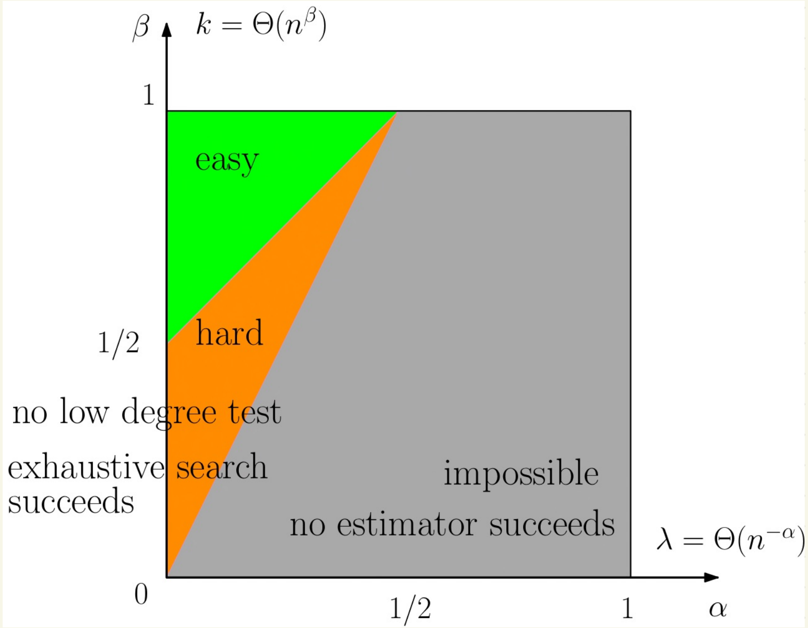
CONTEXT|

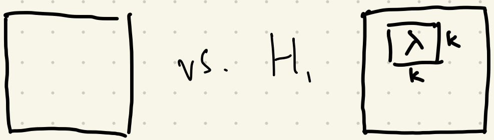
'Easier to detect than recover'.

Detection

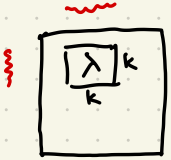


Recovery



H_0  vs. H_1

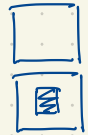
recover



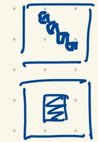
Hypothesis Testing

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

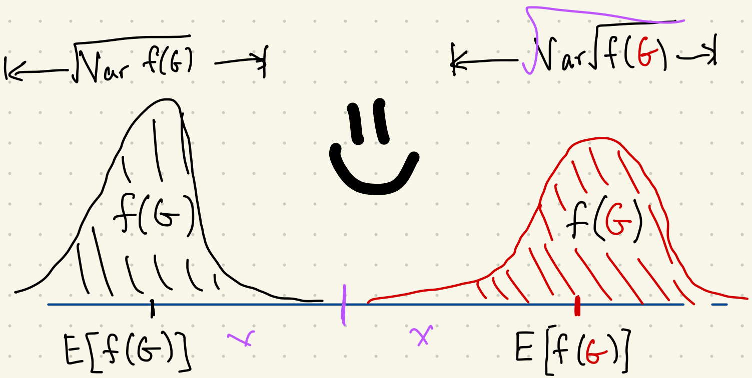
$H_1: \; G \sim Q_n : G(n, \frac{1}{2}, k)$



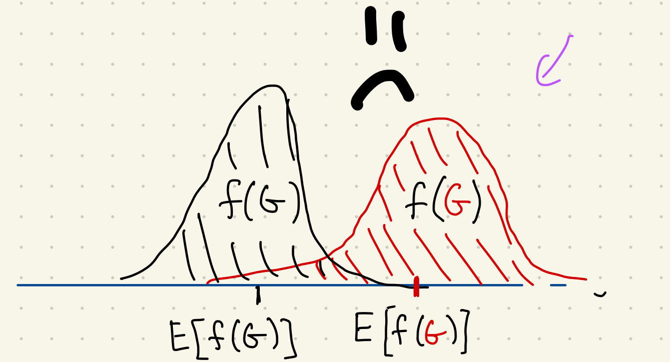
distributions on $\{0,1\}^{\binom{n}{2}}$ or $\mathbb{R}^{\binom{n}{2}}$.



f detects



f doesn't detect



A 'degree D test' $f_n : \mathbb{R}^{\binom{n}{2}} \to \mathbb{R}$ 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" >> "fluctuations".

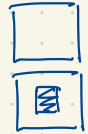
LOW DEG POLY $\boxed{\text{fail}}$ if

no $\Theta(\log(n))$-degree test.

Hypothesis Testing

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

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



distributions on $\{0,1\}^{\binom{n}{2}}$ or $\mathbb{R}^{\binom{n}{2}}$.

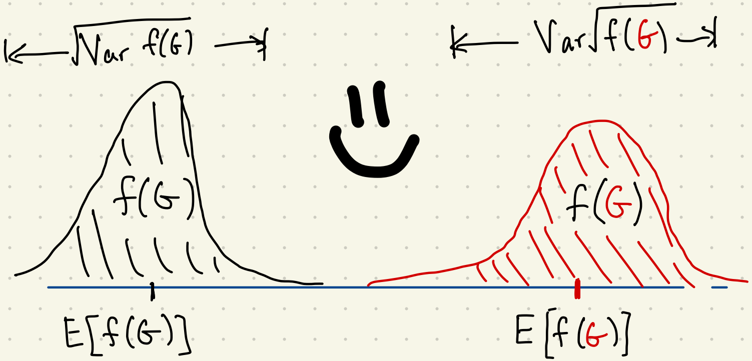
$$\boxed{\mathbb{E}_{P_n}[f] - \mathbb{E}_{Q_n}[f] \gg \sqrt{\max\{\mathrm{Var}_Q[f], \mathrm{Var}_P[f]\}}}$$

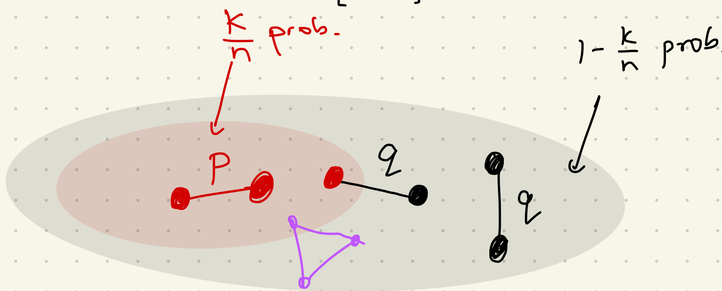
$\underline{Ex}$: $\quad f = \sum_{u,v,w} \mathbb{1}[uv \in E]\, \mathbb{1}[uw \in E]\, \mathbb{1}[vw \in E]$

$$= \sum_{u,v,w} A_{uv} A_{uw} A_{vw}$$

$f(G) = 3! \; \# \Delta\text{'s}$

f detects





$\underline{\text{calc}}$

$$\mathbb{E}_{Q_n}[f] = \sum_{u,v,w} \mathbb{P}[uv, uw, vw \in E] \qquad q^3 \cdot \left(1 - \tfrac{k}{n}\right)^3$$

$$= \sum_{u,v,w} \mathbb{P}\left[uv, uw, vw \in E \,\middle|\, \begin{smallmatrix} u & \bullet \\ v & \bullet \\ w & \bullet \end{smallmatrix}\right] \cdot \mathbb{P}\left[\begin{smallmatrix} u & \bullet \\ v & \bullet \\ w & \bullet \end{smallmatrix}\right]$$

$$+ \mathbb{P}\left[uv, uw, vw \in E \,\middle|\, \begin{smallmatrix} u & \color{red}\bullet \\ v & \bullet \\ w & \bullet \end{smallmatrix}\right] \cdot \mathbb{P}\left[\begin{smallmatrix} u & \color{red}\bullet \\ v & \bullet \\ w & \bullet \end{smallmatrix}\right]$$

$$+ \mathbb{P}\left[uv, uw, vw \in E \,\middle|\, \begin{smallmatrix} u & \bullet \\ v & \color{red}\bullet \\ w & \bullet \end{smallmatrix}\right] \cdot \mathbb{P}\left[\begin{smallmatrix} u & \bullet \\ v & \color{red}\bullet \\ w & \bullet \end{smallmatrix}\right]$$

$$\vdots$$

$$+ \mathbb{P}\left[uv, uw, vw \in E \,\middle|\, \begin{smallmatrix} u & \color{red}\bullet \\ v & \color{red}\bullet \\ w & \color{red}\bullet \end{smallmatrix}\right] \cdot \mathbb{P}\left[\begin{smallmatrix} u & \color{red}\bullet \\ v & \color{red}\bullet \\ w & \color{red}\bullet \end{smallmatrix}\right]$$

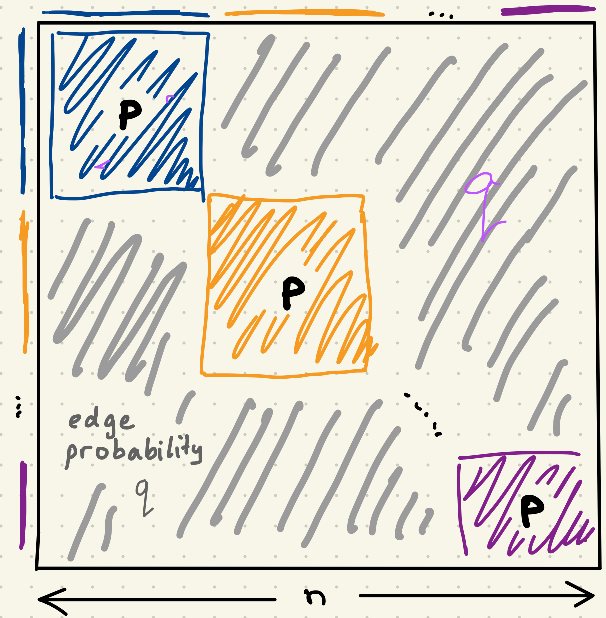
$$p^3 \cdot \left(\tfrac{k}{n}\right)^3$$

PLANTED PARTITION / STOCHASTIC Block model : M COMMUNITIES

Vertex labels : $\sigma_v = \begin{cases} 1 & \text{w. prob } \frac{1}{M} \\ \vdots & \vdots \\ M & " \end{cases}$

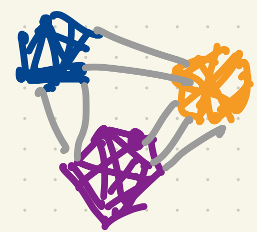
Edges

for $u < v$: $A_{uv} \sim \begin{cases} \mathrm{Ber}(p) & \text{if } \sigma_u = \sigma_v \\ \mathrm{Ber}(q) & \text{o.w.} \end{cases}$



ALGORITHMIC QNS

- Detection : determine if whp sample from planted model or not
- Recovery : given sample from planted model find communities (exactly? weakly corr?)
- "Counting" .... ?



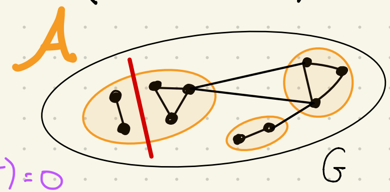
Modularity Properties

$\mathcal{A} \in \mathrm{OPT}(G)$ i.e $q_{\mathcal{A}}(G) = q^*(G)$

$\Rightarrow \forall A \in \mathcal{A}$ $G[A]$ conn. (+ isolated vert)

$\Rightarrow$ pendant vertex in same part 

$\Rightarrow$ $q^*(\text{star}) = 0$



Modularity value:

Robust to small perturbations in edge set

$$\left| q^*(G) - q^*(G \setminus E') \right| < \frac{2|E'|}{e(G)}$$

$$\forall \mathcal{A}: \quad \left| q_{\mathcal{A}}(G) - q_{\mathcal{A}}(G \setminus E') \right| < \frac{2|E'|}{e(G)}$$

SBM

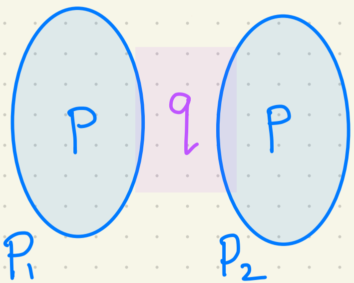
$p = \frac{a}{n} \cdot w(n)$

$q = \frac{b}{n} \cdot w(n)$,

$a > b$

$\mathcal{P} = \{P_1, P_2\}$

'planted partition'



Thm [BICKEL CHEN] $G \sim G_{n,p,q}$, let $\mathcal{A} \in \mathrm{OPT}(G)$

- $w(n) \to \infty \Rightarrow$ whp $\mathcal{A}$ is $o(n)$ away $\mathcal{P}$
- $\frac{w(n)}{\log n} \to \infty \Rightarrow$ " $\mathcal{A} = \mathcal{P}$

Percolated random graph G_p

G, G_p similar mod values + partitions

if $\frac{e(G)p}{n} \to \infty$ whp

$$\left| q^*(G) - q^*(G_p) \right| = o(1)$$

$\forall \mathcal{A}\ \exists \mathcal{A}'$: (similar)

$$\left| q_{\mathcal{A}'}(G) - q_{\mathcal{A}}(G_p) \right| = o(1)$$

III Group Exercises

- modularity-based + planted structure
- graph theory, random graphs, theory of algs, simulations