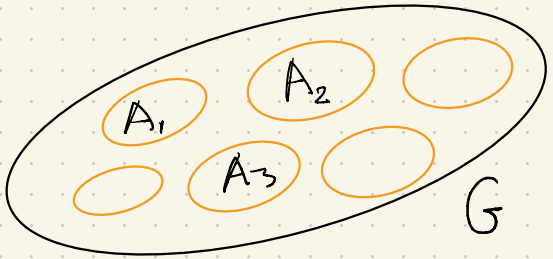
Modularity + Sampling

with Colin McDiarmid

Modularity

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



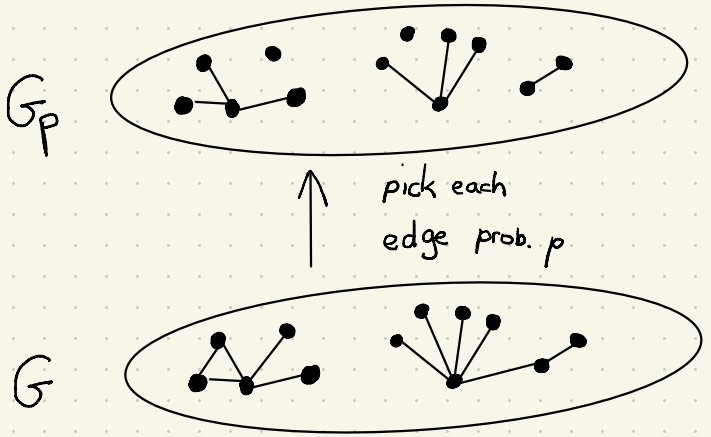
graph G, m edges. $\mathcal{A} = \{A_1, \ldots, A_k\}$ vertex partition

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"

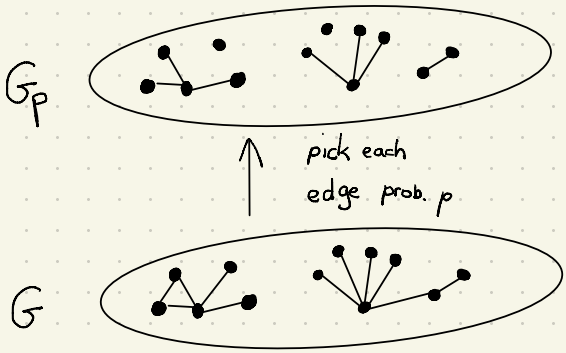
Sampling



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

$G = (V, E)$ fixed graph

Sampling

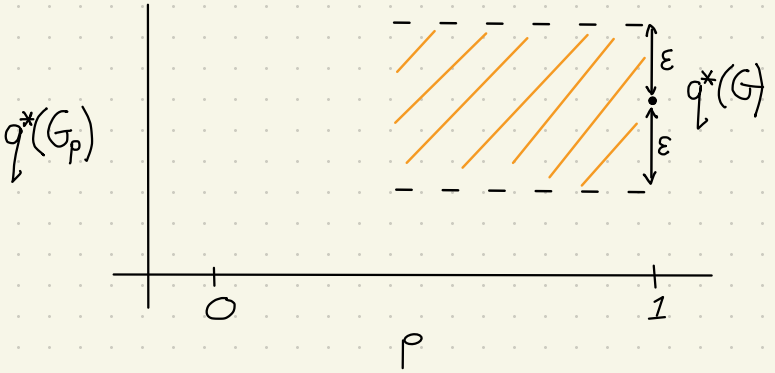


$G_p = (V, E_p)$ E_p each edge kept indep. prob. p

$G = (V, E)$ fixed graph

Q Given G, $\varepsilon > 0$ for which p is it true with prob $> 1 - \varepsilon$

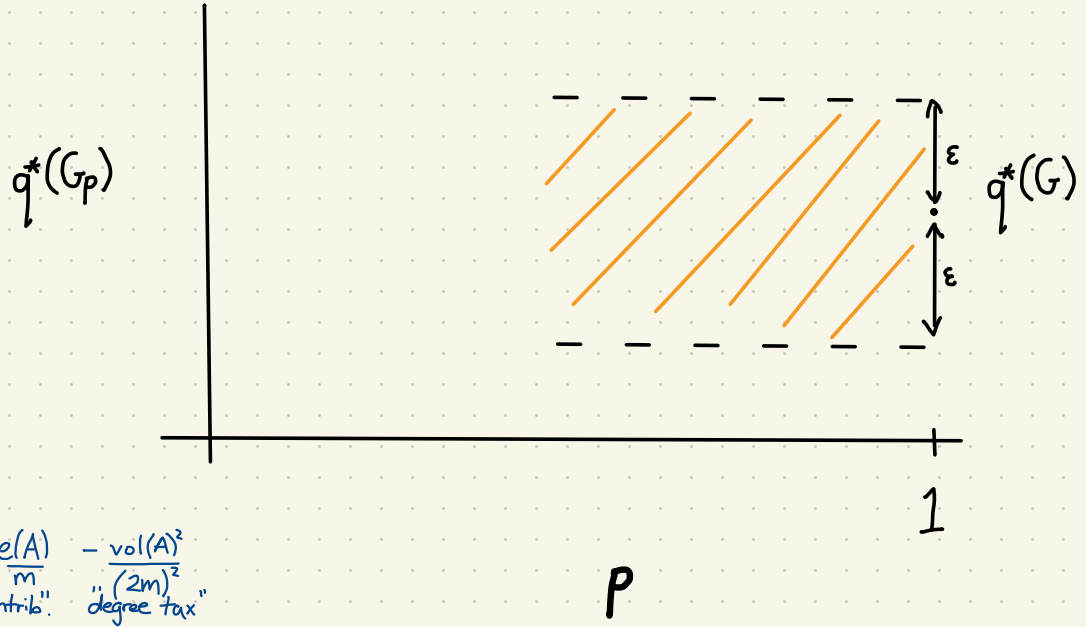
$$|q^*(G_p) - q^*(G)| < \varepsilon$$



Thm [M-D+S] $\forall \varepsilon > 0 \quad \exists c = c(\varepsilon)$

for graph G constant p

- if then w. prob $> 1-\varepsilon$ $q^*(G_p) > q^*(G) - \varepsilon$
- if " " $q^*(G_p) < q^*(G) + \varepsilon$



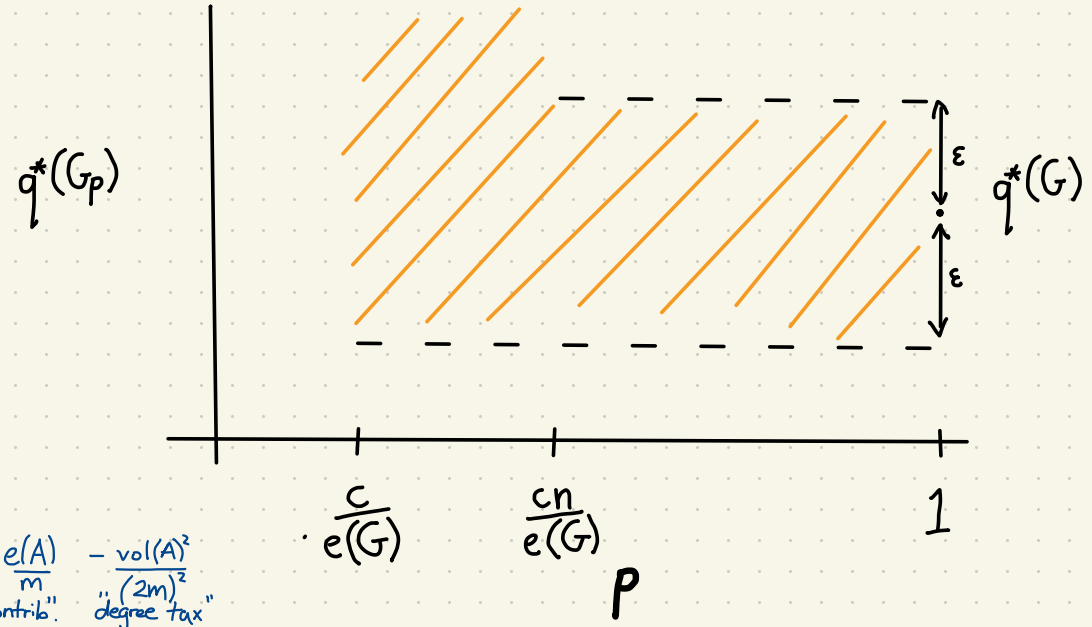
$q_A(G) = \sum_{A \in A} \frac{e(A)}{m} - \frac{\text{vol}(A)^2}{(2m)^2}$

"edge contrib." "degree tax"

Thm [M-D+S] $\forall \varepsilon > 0 \ \exists c = c(\varepsilon)$

for graph G, constant p

- if $e(G)\cdot p \geq c$ then w. prob $> 1-\varepsilon$ $\quad q^*(G_p) > q^*(G) - \varepsilon$
- if $e(G)\cdot p \geq cn$ " " $\quad q^*(G_p) < q^*(G) + \varepsilon$



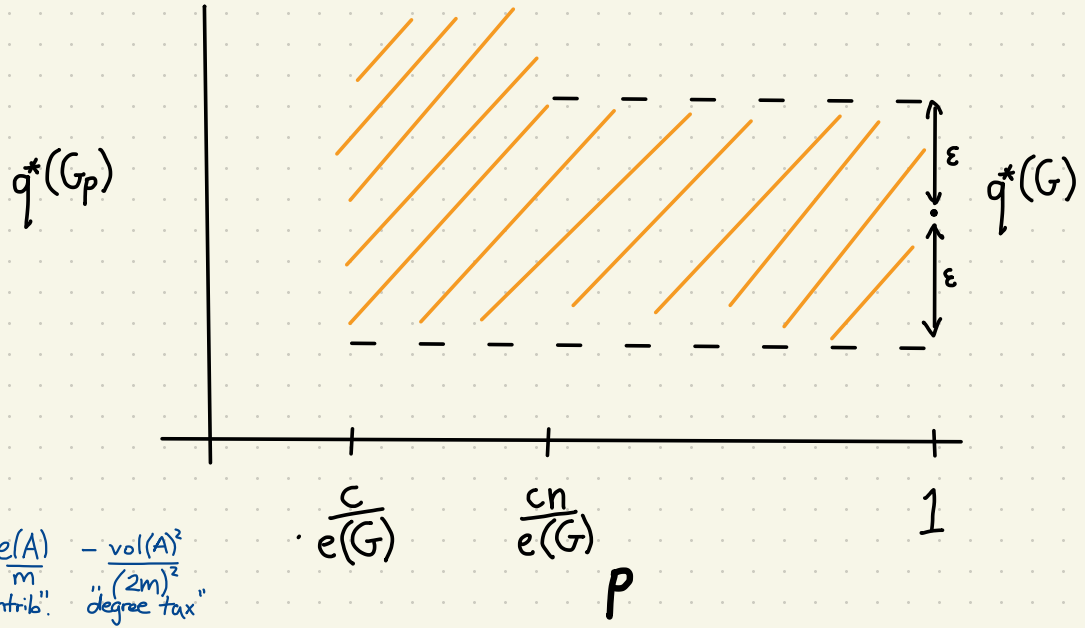
$q_A(G) = \sum_{A \in \mathcal{A}} \frac{e(A)}{m} - \frac{\mathrm{vol}(A)^2}{(2m)^2}$

"edge contrib." "degree tax"

Thm [M-D+S] $\forall \varepsilon > 0 \quad \exists c = c(\varepsilon)$

for graph G, constant p

- if $e(G)\,p \geq c$ then w. prob $> 1-\varepsilon$ $\quad q^*(G_p) > q^*(G) - \varepsilon$
- if $e(G)\,p \geq cn$ " " $\quad q^*(G_p) < q^*(G) + \varepsilon$
- if $e(G)\,p \geq cn$ given A, can construct A' w. prob $> 1-\varepsilon$ $\quad q_{A'}(G) > q_A(G_p) - \varepsilon$

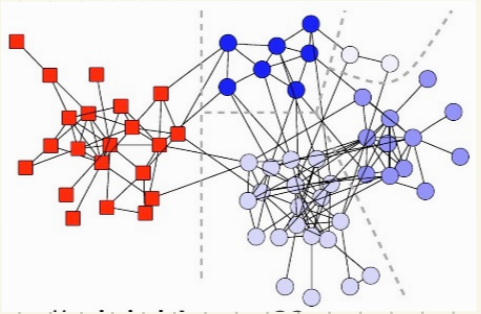


$q_A(G) = \sum_{A \in A} \frac{e(A)}{m} - \frac{\mathrm{vol}(A)^2}{(2m)^2}$

"edge contrib." "degree tax"

Simulations

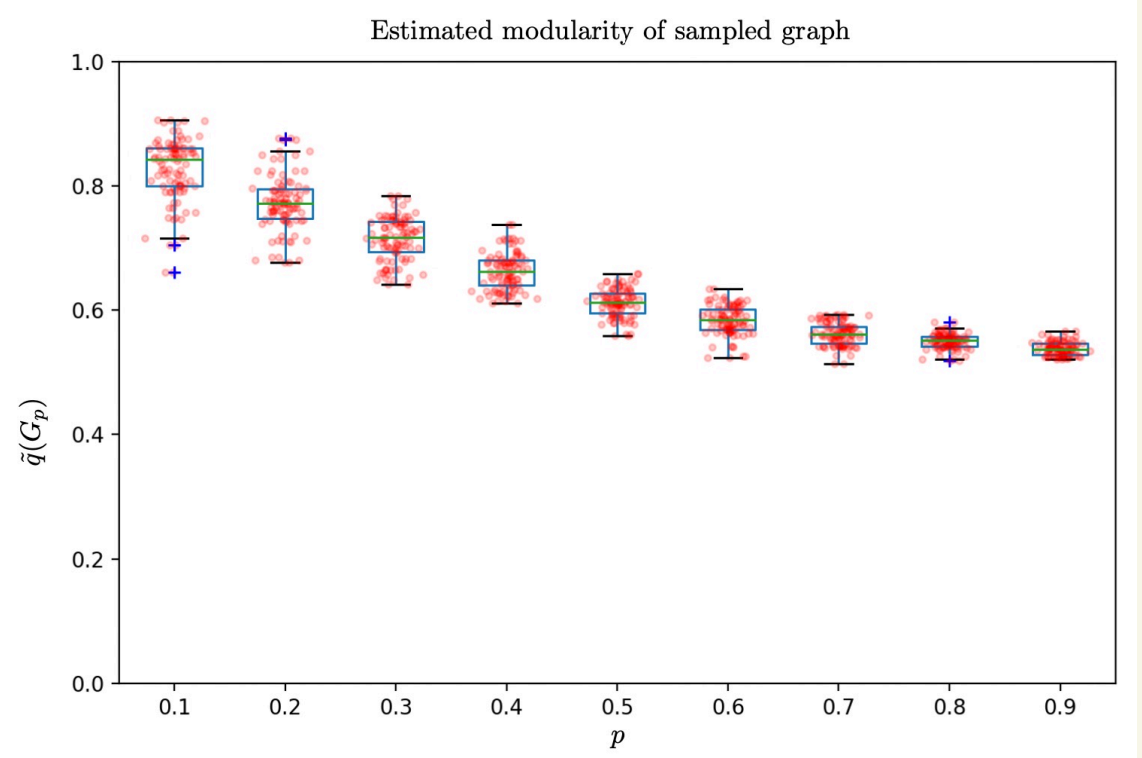
Dolphin Network [Lusseau]



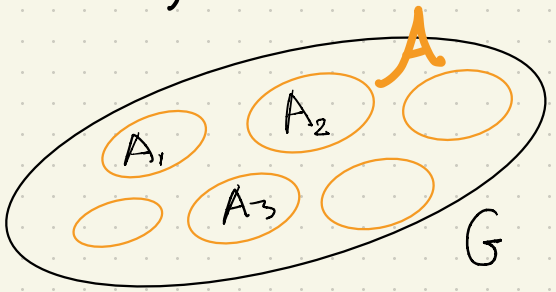
$|V| = 62 \quad |E| = 152$

$q^*(G) = 0.529\ldots$ (3 dec places)

[Brande + '08]



Modularity 'measures 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"

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

"edge contrib" "degree tax"

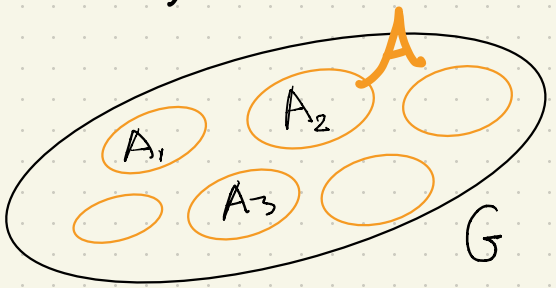
d_u = # edges incident to u

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

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

Modularity

'measures 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"

$$q_{\mathcal{A}}(G) = \sum_{A \in \mathcal{A}} \frac{e(A)}{m} - \frac{\mathrm{vol}(A)^2}{(2m)^2} = \frac{1}{2m} \sum_{A \in \mathcal{A}} \sum_{u,v \in A} \mathbb{1}[u \sim v] - \frac{d_u d_v}{2m}$$

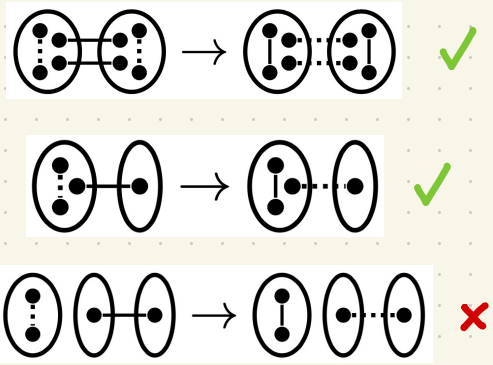
"edge contrib." "degree tax"

d_u = # edges incident to u

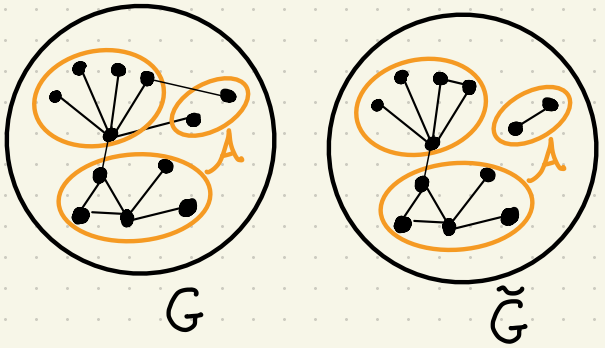
$e(A)$ = # edges in set A

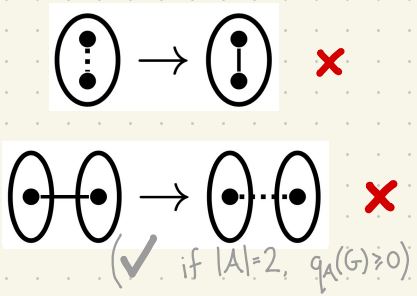
$\mathrm{vol}(A) = \sum_{u \in A} d_u$

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



✓ ✓ ✗



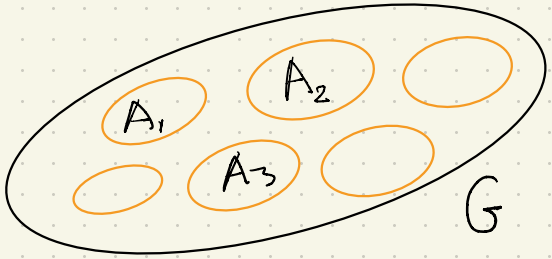


✗ ✗

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

Modularity

'measures 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"

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

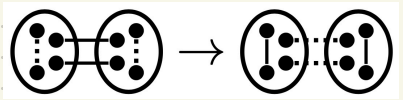
"edge contrib" "degree tax"

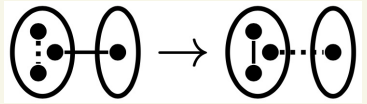
d_u = # edges incident to u

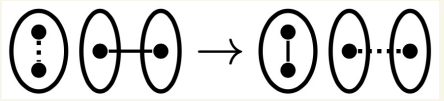
$e(A)$ = # edges in set A

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

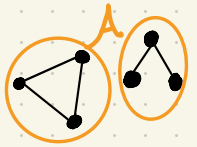
Fix A, which $G \to \tilde{G}$ ensures $q_{\mathcal{A}}(\tilde{G}) > q_{\mathcal{A}}(G)$?


✓

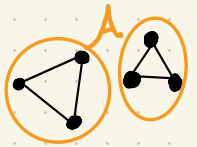

✓


✗

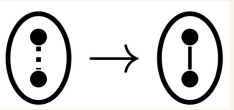
$q_{\mathcal{A}} = \frac{1}{2}$

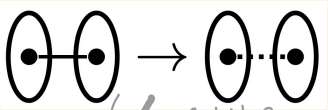


$q_{\mathcal{A}} = \frac{12}{25}$



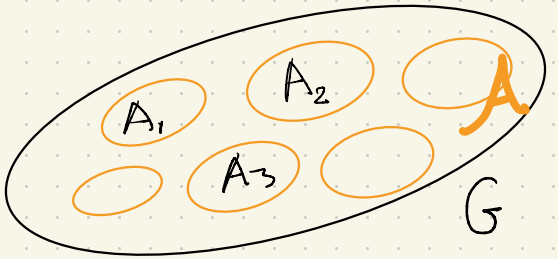
$q_{\mathcal{A}} = \frac{1}{2}$


✗


✗

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

Modularity Properties



- Robust to small perturbations in edge set

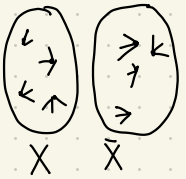
$$\left| q^*(G) - q^*(G \setminus E') \right| < \frac{2|E'|}{e(G)}$$

$q^*(G) > 1 - \varepsilon$ if any of following hold

- Connected components in G all $< \varepsilon e(G)$ edges
- $\exists \mathcal{A}$ with #edges between parts $< \frac{\varepsilon}{2} e(G)$ and $\forall A \in \mathcal{A}$ $\mathrm{vol}(A) < \frac{\varepsilon}{2} \mathrm{vol}(G)$
- no subgraph $H \subseteq G$, H δ-expander, $e(H) > \delta e(G)$ (+Louf)

$q^*(G) < \varepsilon$ if any of following hold

- $\bar{\lambda}(G) < \varepsilon$ where $\bar{\lambda}(G)$ is spectral gap of Laplacian of G
  NB: r-regular G, $\bar{\lambda}(G) = \frac{1}{r} \max_{i \neq 0} |\lambda_i(A_G)|$
- $\forall X \subseteq V(G)$, $\frac{e(X, \bar{X})}{e(G)} > 2(1-\varepsilon) \frac{\mathrm{vol}(X)}{\mathrm{vol}(G)} \frac{\mathrm{vol}(\bar{X})}{\mathrm{vol}(G)}$

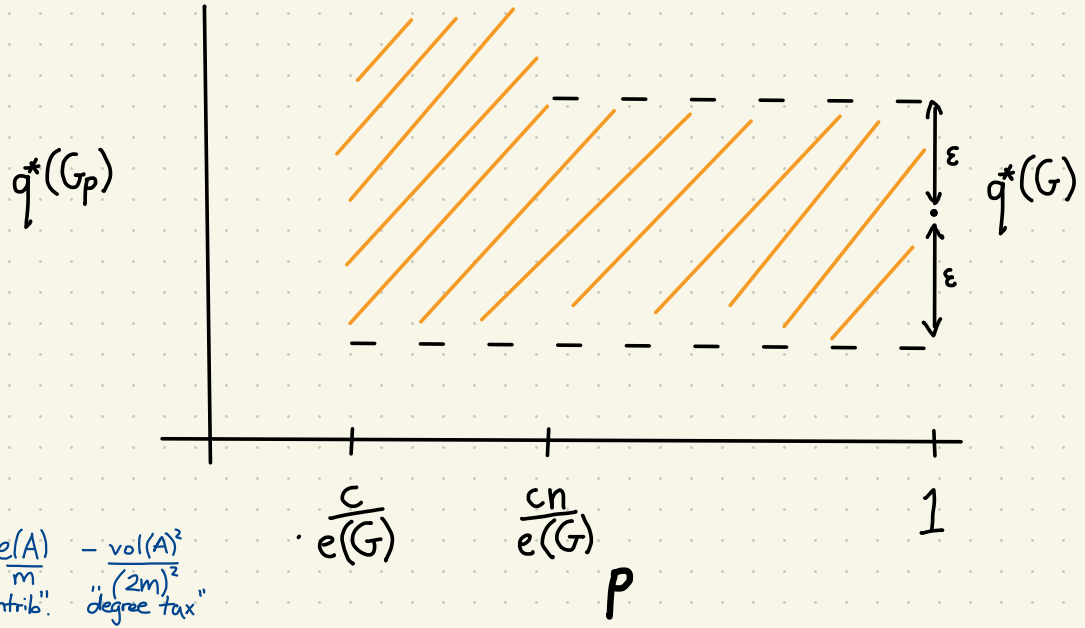


$$q_{\mathcal{A}}(G) = \sum_{A \in \mathcal{A}} \underbrace{\frac{e(A)}{m}}_{\text{"edge contrib."}} - \underbrace{\frac{\mathrm{vol}(A)^2}{(2m)^2}}_{\text{"degree tax"}}$$

Thm [M-D+S] $\forall \varepsilon > 0 \quad \exists c = c(\varepsilon)$

for graph G, constant p

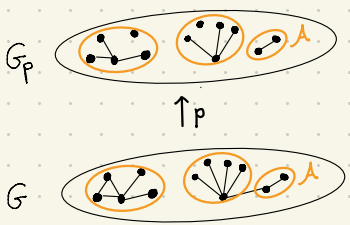
- if $e(G)\,p \geq c$ then w. prob $> 1-\varepsilon$ $\quad q^*(G_p) > q^*(G) - \varepsilon$
- if $e(G)\,p \geq cn$ " " $\quad q^*(G_p) < q^*(G) + \varepsilon$
- if $e(G)\,p \geq cn$ given A, can construct A' w. prob $> 1-\varepsilon$ $\quad q_{A'}(G) > q_A(G_p) - \varepsilon$



$q_A(G) = \sum_{A \in A} \frac{e(A)}{m} - \frac{\mathrm{vol}(A)^2}{(2m)^2}$

"edge contrib." "degree tax"

IDEA OF PROOF: approx opt partition with 'well-behaved' partition



Thm [Dinh+Thai]: for k positive integer

$$\max_{|A| \le k} q_A(G) \ge q^*(G)\left(1 - \frac{1}{k}\right)$$

A is η-fat for G if $\mathrm{vol}(A) \ge \eta\, \mathrm{vol}(G)$

Fattening Lemma: Given A partition of G can construct A' η-fat s.t. A refinement of A'

$$q_{A'}(G) \ge q_A(G) - 2\eta$$

ASIDE: LOAD BALANCING

$\vec{x} = (x_1, \dots, x_n)$ $\quad x_i \ge 0$ $\quad \sum_i x_i = 1$

let $\gamma(\vec{x}) = \max_{I \subseteq [n]} \min\left\{ \sum_{i \in I} x_i,\ 1 - \sum_{i \in I} x_i \right\}$

observe

$$\gamma(\vec{x}) \ge \frac{1}{2} - \max_i x_i$$

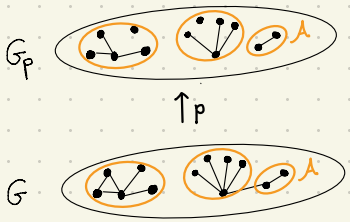
Q what is the largest α s.t.

$$\gamma(\vec{x}) \ge \alpha\left(1 - \sum_i x_i^2\right) \quad ?$$

$$q_A(G) = \sum_{A \in A} \frac{e(A)}{m} - \frac{\mathrm{vol}(A)^2}{(2m)^2}$$

"edge contrib." "degree tax"

IDEA OF PROOF: approx opt partition with 'well-behaved' partition



Thm [Dinh+Thai]: for k positive integer

$$\max_{|A| \le k} q_A(G) \ge q^*(G)\left(1 - \frac{1}{k}\right)$$

A is η-fat for G if $\mathrm{vol}(A) \ge \eta\, \mathrm{vol}(G)$

Fattening Lemma: Given A partition of G can construct A' η-fat s.t. A refinement of A'

$$q_{A'}(G) \ge q_A(G) - 2\eta$$

$$q_A(G) = \sum_{A \in A} \underbrace{\frac{e(A)}{m}}_{\text{"edge contrib"}} - \underbrace{\frac{\mathrm{vol}(A)^2}{(2m)^2}}_{\text{"degree tax"}}$$

ASIDE: LOAD BALANCING

$\vec{x} = (x_1, \dots, x_n)$ $\quad x_i \ge 0$ $\quad \sum_i x_i = 1$

let $\gamma(\vec{x}) = \max_{I \subseteq [n]} \min\left\{ \sum_{i \in I} x_i,\ 1 - \sum_{i \in I} x_i \right\}$

observe

$$\gamma(\vec{x}) \ge \frac{1}{2} - \max_i x_i$$

Q what is the largest α s.t.

$$\gamma(\vec{x}) \ge \alpha\left(1 - \sum_i x_i^2\right) \quad ?$$

$\alpha \le \frac{1}{2}$

$$\gamma\left(\left(\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}\right)\right) = \tfrac{1}{3} = \tfrac{1}{2}\left(1 - \tfrac{1}{3}\right)$$

$\alpha \ge \frac{1}{2}$

greedy alg

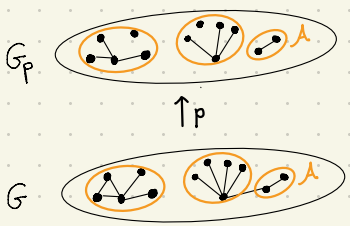
- $x_1 \ge \dots \ge x_n$
- init $A, B = \emptyset$
- step i add x_i to smaller heap

achieves $\alpha \ge \frac{1}{2}$

IDEA OF PROOF: approx opt partition with 'well-behaved' partition

Thm [Dinh+Thai]: for k positive integer

$$\max_{|A| \le k} q_A(G) \ge q^*(G)\left(1 - \frac{1}{k}\right)$$



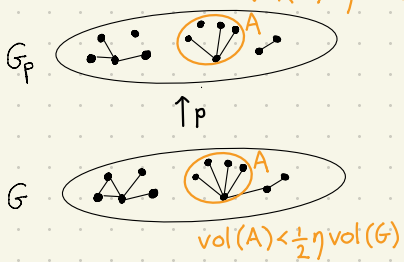
A is η-fat for G if $\mathrm{vol}(A) \ge \eta\, \mathrm{vol}(G)$

Fattening Lemma: Given A partition of G can construct A' η-fat s.t. A refinement of A'

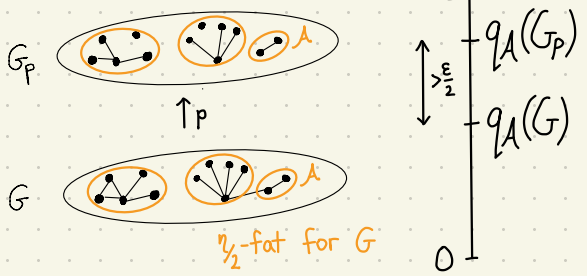
$$q_{A'}(G) \ge q_A(G) - 2\eta$$

BAD EVENTS

B_0: $\exists A \subseteq V$ s.t. $\mathrm{vol}(A) > \eta\, \mathrm{vol}(G_P)$



B_1: $\exists A$ s.t. η-fat for G_P



$$q_A(G) = \sum_{A \in \mathcal{A}} \frac{e(A)}{m} - \frac{\mathrm{vol}(A)^2}{(2m)^2}$$

"edge contrib." "degree tax"

ASIDE: LOAD BALANCING

$\vec{x} = (x_1, \dots, x_n)$ $x_i \ge 0$ $\sum_i x_i = 1$

let $\gamma(\vec{x}) = \max_{I \subseteq [n]} \min\left\{ \sum_{i \in I} x_i,\ 1 - \sum_{i \in I} x_i \right\}$

observe

$$\gamma(\vec{x}) \ge \frac{1}{2} - \max_i x_i$$

Q what is the largest α s.t.

$$\gamma(\vec{x}) \ge \alpha\left(1 - \sum_i x_i^2\right) \quad ?$$

$\alpha \le \frac{1}{2}$

$$\gamma\left(\left(\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}\right)\right) = \frac{1}{3} = \frac{1}{2}\left(1 - \frac{1}{3}\right)$$

$\alpha \ge \frac{1}{2}$

greedy alg

- $x_1 \ge \dots \ge x_n$
- init $A, B = \emptyset$
- step i add x_i to smaller heap

achieves $\alpha \ge \frac{1}{2}$

OPEN - LIMITS: let $\bar{q}(n,c) = \mathbb{E}(q^*(G_{n,\frac{c}{n}}))$

Conj [M°D+S]: $\forall c > 1$, $\bar{q}(n,c)$ tends to a limit $\bar{q}(c)$ as $n \to \infty$

If conj holds then

(i) for $0 < c \leq 1$,
$$\bar{q}(n,c) \to \bar{q}(c) = 1 \qquad \text{as } n \to \infty$$

(ii) $0 < \bar{q}(c) < 1$ for $c > 1$

(iii) $\bar{q}(c) = \Theta(c^{-\frac{1}{2}})$ as $c \to \infty$

(iv) $\bar{q}(c)$ is (uniformly) continuous for $c \in (0, \infty)$

(v) $\bar{q}(c)$ is non-increasing for $c \in (0, \infty)$

⇐ Erdös-Renyi Results

⇐ Sampling Thm

$$q_A(G) = \sum_{A \in \mathcal{A}} \frac{e(A)}{m} - \frac{\text{vol}(A)^2}{(2m)^2}$$

"edge contrib." "degree tax"

OPEN - LIMITS: let $\bar{q}(n,c) = \mathbb{E}(q^*(G_{n,\frac{c}{n}}))$

Conj [M°D+S]: $\forall c > 1$, $\bar{q}(n,c)$ tends to a limit $\bar{q}(c)$ as $n \to \infty$

If conj holds then

(i) for $0 < c \leq 1$,
$$\bar{q}(n,c) \to \bar{q}(c) = 1 \qquad \text{as } n \to \infty$$

(ii) $0 < \bar{q}(c) < 1$ for $c > 1$

(iii) $\bar{q}(c) = \Theta(c^{-\frac{1}{2}})$ as $c \to \infty$

(iv) $\bar{q}(c)$ is (uniformly) continuous for $c \in (0, \infty)$

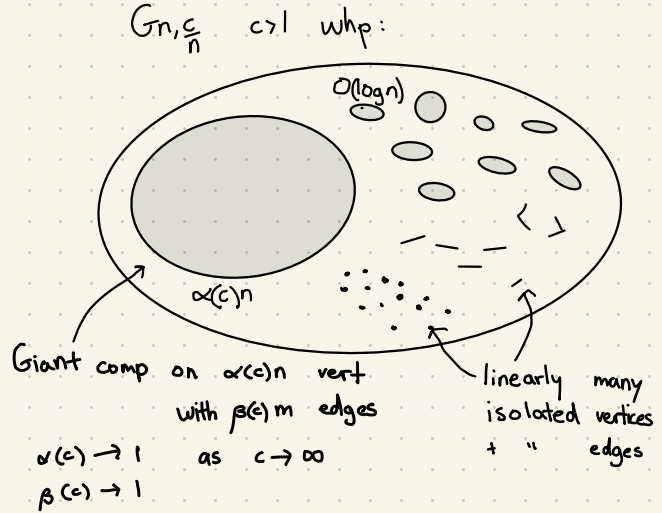
(v) $\bar{q}(c)$ is non-increasing for $c \in (0, \infty)$

(i)–(iii) $\Leftarrow$ Erdös-Renyi Results

(iv)–(v) $\Leftarrow$ Sampling Thm

OPEN - FIVE PARTS

Conj [Reichardt+Bornholdt '06]: $\max_{|A|=5} q_A(G_{n,\frac{c}{n}}) = q^*(G_{n,\frac{c}{n}})$ whp



$q_A(G) = \sum_{A \in A} \frac{e(A)}{m} - \frac{\text{vol}(A)^2}{(2m)^2}$

"edge contrib." "degree tax"

OPEN - LIMITS: let $\bar{q}(n,c) = \mathbb{E}(q^*(G_{n,\frac{c}{n}}))$

Conj [M°D+S]: $\forall c > 1$, $\bar{q}(n,c)$ tends to a limit $\bar{q}(c)$ as $n \to \infty$

If conj holds then

(i) for $0 < c \leq 1$, $\bar{q}(n,c) \to \bar{q}(c) = 1$ as $n \to \infty$

(ii) $0 < \bar{q}(c) < 1$ for $c > 1$

(iii) $\bar{q}(c) = \Theta(c^{-\frac{1}{2}})$ as $c \to \infty$

(iv) $\bar{q}(c)$ is (uniformly) continuous for $c \in (0, \infty)$

(v) $\bar{q}(c)$ is non-increasing for $c \in (0, \infty)$

} $\Leftarrow$ Erdös-Renyi Results

$\Leftarrow$ Sampling Thm

OPEN - FIVE PARTS

Conj [Reichardt+Bornholdt '06]: $\max_{|A|=5} q_A(G_{n,\frac{c}{n}}) = q^*(G_{n,\frac{c}{n}})$ whp

But! $\forall c\ \exists \delta(c)$ $\max_{|A|\leq k} q_A(G_{n,\frac{c}{n}}) \leq q^*(G_{n,\frac{c}{n}})(1 - \delta(c))$ whp

By Dinh+Thai $\max_{|A|\leq k} q_A(G_{n,\frac{c}{n}}) \geq q^*(G_{n,\frac{c}{n}})(1 - \frac{1}{k})$

Conj [M°D+S]: $\exists K$ ($k=5$?) st. $\forall \varepsilon > 0$ $\exists c_0$ and $\forall c \geq c_0$

$$\max_{|A|\leq k} q_A(G_{n,\frac{c}{n}}) \geq q^*(G_{n,\frac{c}{n}})(1 - \varepsilon)$$

$q_A(G) = \sum_{A \in A} \frac{e(A)}{m} - \frac{\mathrm{vol}(A)^2}{(2m)^2}$

"edge contrib." "degree tax"