Modularity and edge sampling

Fiona Skerman
Bristol/Uppsala

Birmingham 2020

Partitioning Networks:

Network:

links between blogs on climate change

Elgesem, Steskal & Diakopoulos 2015

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

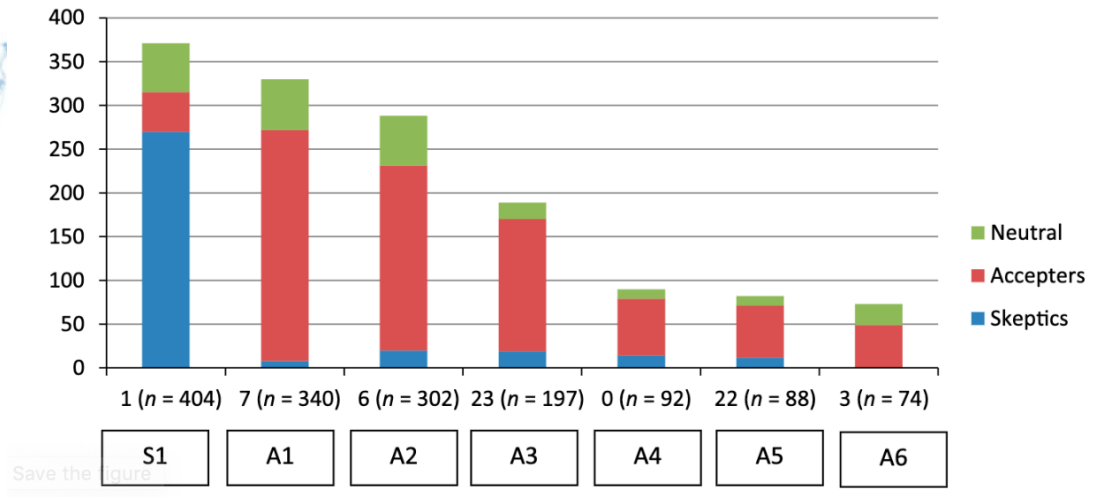


Figure 3. The distribution of skeptical, accepting, and neutral blogs in the seven largest among the central groups of blogs concerned with climate change.

Partitioning Networks:

Network:

links between blogs on climate change

Elgesem, Steskal & Diakopoulos 2015

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

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.

Definition

Motivation - configuration model Newman & Girvan 2004

Let G be a graph on $m \geq 1$ edges and $\mathcal{A}$ a vertex partition,

Write d_u for the degree of u,

$e(A)$ for the number of edges in A and $\mathrm{vol}(A) = \sum_{u \in A} d_u$.

$$q_{\mathcal{A}}(G) = \frac{1}{2m} \sum_{A \in \mathcal{A}} \sum_{u,v \in A} \left(1_{uv \in E} - \frac{d_u d_v}{2m}\right) = \sum_{A \in \mathcal{A}} \frac{e(A)}{m} - \frac{\mathrm{vol}(A)^2}{4m^2}$$

Expected number of edges between distinct u and v is $\frac{d_u d_v}{2m-1}$

Definition

Motivation - configuration model Newman & Girvan 2004

Let G be a graph on $m \geq 1$ edges and $\mathcal{A}$ a vertex partition,

Write d_u for the degree of u,

$e(A)$ for the number of edges in A and $\mathrm{vol}(A) = \sum_{u \in A} d_u$.

$$q_{\mathcal{A}}(G) = \frac{1}{2m} \sum_{A \in \mathcal{A}} \sum_{u,v \in A} \left(1_{uv \in E} - \frac{d_u d_v}{2m}\right) = \sum_{A \in \mathcal{A}} \frac{e(A)}{m} - \frac{\mathrm{vol}(A)^2}{4m^2}$$

Max. Modularity

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

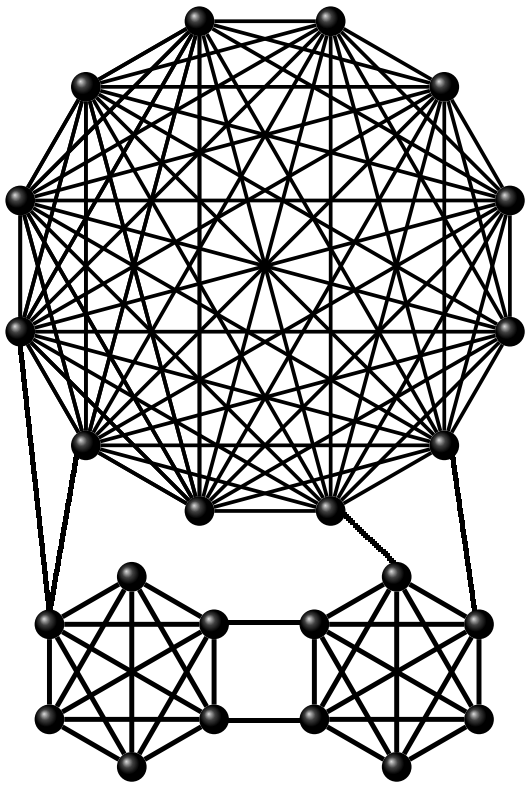
Edge contribution

$$q_{\mathcal{A}}^{E}(G) = \sum_{A \in \mathcal{A}} \frac{e(A)}{m}$$

Degree tax

$$q_{\mathcal{A}}^{D}(G) = \sum_{A \in \mathcal{A}} \frac{\mathrm{vol}(A)^2}{4m^2}$$

Example Graph



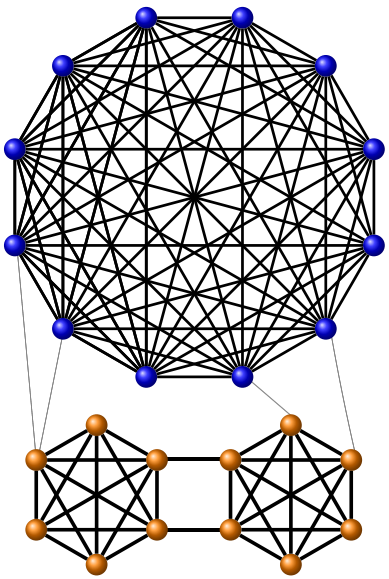
Edge contribution

$$q^E_{\mathcal{A}}(G) = \sum_{A \in \mathcal{A}} \frac{e(A)}{m}$$

Degree tax

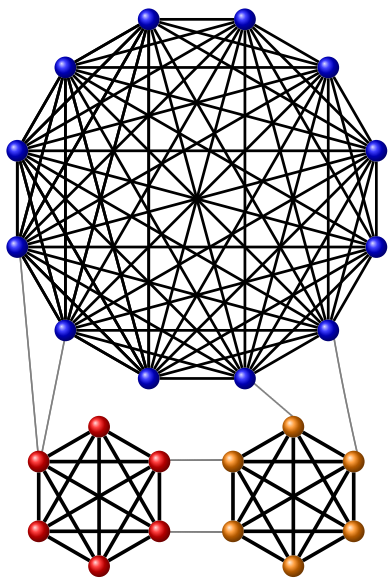
$$q^D_{\mathcal{A}}(G) = \sum_{A \in \mathcal{A}} \frac{\mathrm{vol}(A)^2}{4m^2}$$

3 Possible Partitions



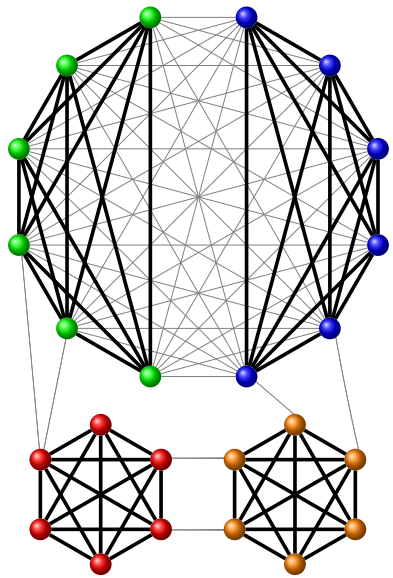
$q^E_{\mathcal{A}_1} = 0.96, \quad q^D_{\mathcal{A}_1} = 0.56$

$q_{\mathcal{A}_1} = 0.40$



$q^E_{\mathcal{A}_2} = 0.94, \quad q^D_{\mathcal{A}_2} = 0.50$

$q_{\mathcal{A}_2} = 0.44$



$q^E_{\mathcal{A}_3} = 0.59, \quad q^D_{\mathcal{A}_3} = 0.29$

$q_{\mathcal{A}_3} = 0.30$

Properties of Modularity

Edge contribution

$$q^E_{\mathcal{A}}(G) = \sum_{A \in \mathcal{A}} \frac{e(A)}{m}$$

Degree tax

$$q^D_{\mathcal{A}}(G) = \sum_{A \in \mathcal{A}} \frac{\mathrm{vol}(A)^2}{4m^2}$$

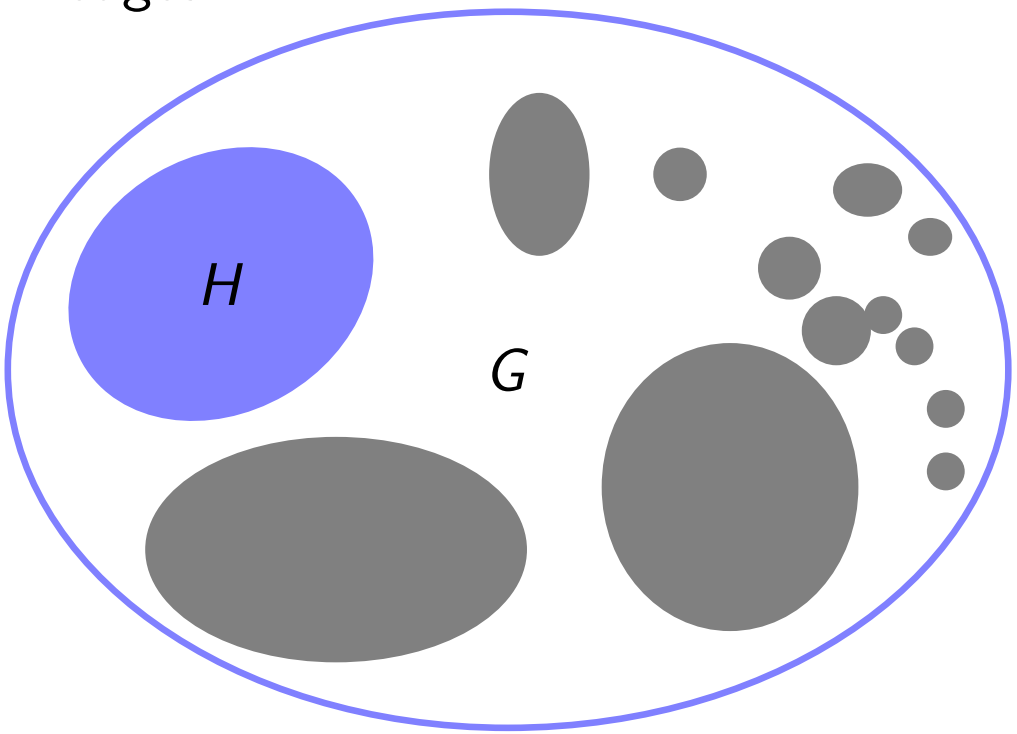
Basic Properties

- $0 \leq q^*(G) < 1$.
- 'blind' to isolated vertices. (V_0)
- parts connected in an optimal partition.
  $A \backslash V_0$ connected for all $A \in \mathcal{A}$ if $q_{\mathcal{A}}(G) = q^*(G)$.
- Resolution limit, Robustness.

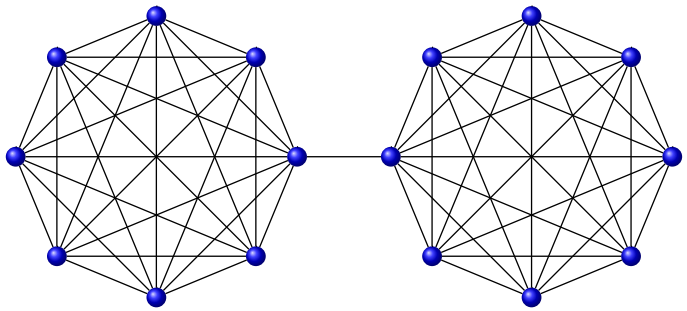
Resolution Limit

Fortunato and Barthélemy 08

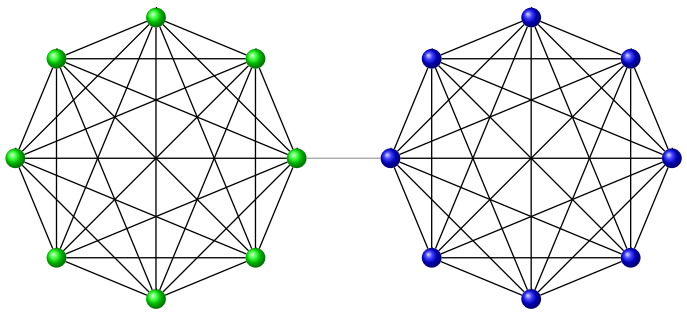
Subgraph H
h edges



Graph G, m edges

$H =$ 

If $h < \sqrt{2m}$, e.g. $m = 1625$.



If $h > \sqrt{2m}$, e.g. $m = 1624$.

Robustness

Partition structure is sensitive to noise in edges

Modularity value is robust

Proposition (McDiarmid, S.)

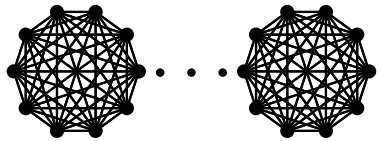
Let $G = (V, E)$ and $G' = (V, E')$ with $|E| \geq |E'|$,

$$|q^*(G) - q^*(G')| \leq \frac{2|E \backslash E'|}{|E|}$$

High

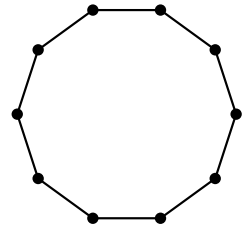
$q^* \to 1$

k cliques



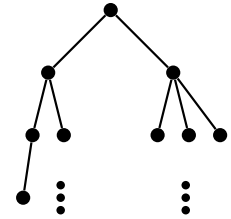
$q^* = 1 - \frac{1}{k}$

cycle C_n



$q^* \sim 1 - \frac{2}{\sqrt{n}}$

tree, max deg Δ

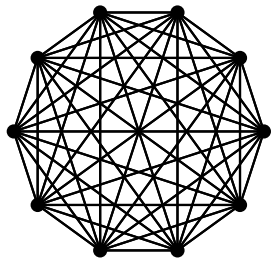


$q^* \geq 1 - 4\sqrt{\frac{\Delta}{n}}$

Low

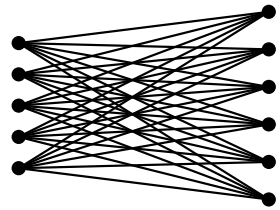
$q^* \to 0$

clique or near-clique $K_n \backslash E'$, $|E'| \leq \frac{n}{2}$



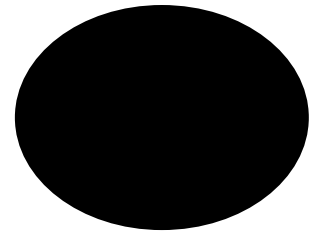
$q^* = 0$

complete multipartite



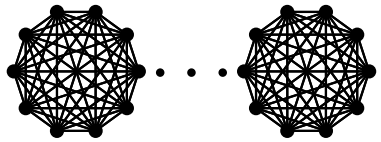
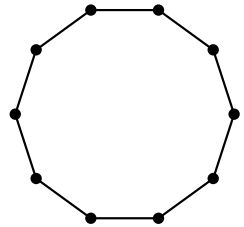
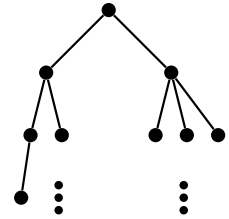
$q^* = 0$

$G(n, \frac{\log n}{n})$ random graph connectivity threshold

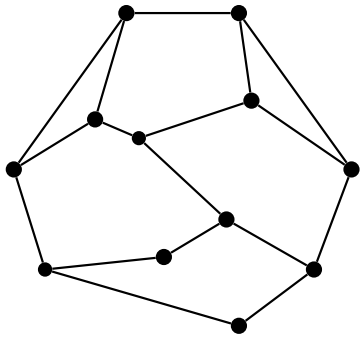
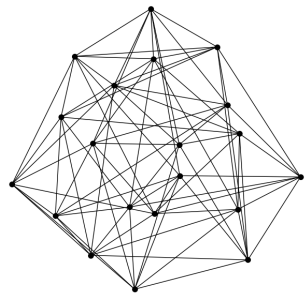
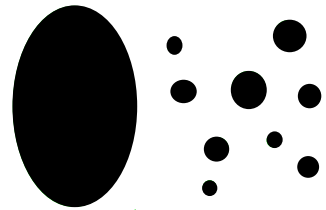


$q^* = O\left(\frac{1}{\sqrt{\log n}}\right)$

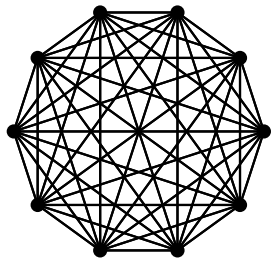
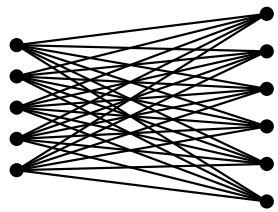
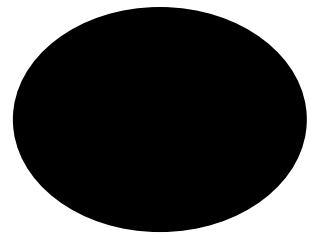
High

k cliques	cycle C_n	tree, max deg Δ
$q^* \to 1$ 		
$q^* = 1 - \frac{1}{k}$	$q^* \sim 1 - \frac{2}{\sqrt{n}}$	$q^* \geq 1 - 4\sqrt{\frac{\Delta}{n}}$

Critical

random cubic	random r-regular large r	$G(n, \frac{c}{n})$ random graph large $c > 1$
$\epsilon < q^* < 1-\epsilon$ 		
$0.66 < q^* < 0.81$	$\frac{0.76}{\sqrt{r}} < q^* < \frac{2}{\sqrt{r}}$	$\frac{1}{10\sqrt{c}} < q^* < \frac{b}{\sqrt{c}}$

Low

clique or near-clique $K_n \backslash E'$, $\lvert E' \rvert \leq \frac{n}{2}$	complete multipartite	$G(n, \frac{\log n}{n})$ random graph connectivity threshold
$q^* \to 0$ 		
$q^* = 0$	$q^* = 0$	$q^* = O\left(\frac{1}{\sqrt{\log n}}\right)$

Sampling to determine modularity

Let G be the underlying network

Let G_p be obtained by sampling each edge with probability p.

How does $q^*(G_p)$ behave? For which p do we have $q^*(G_p) \approx q^*(G)$?

Phase Transition in Erdős-Rényi $(K_n)_p$

Theorem (McDiarmid, S.)

(i) If $n^2 p \to \infty$ and $np \leq 1 + o(1)$ then whp $q^*(G_n) \to 1$. SPARSE

(ii) If $np \to c > 1$ then $\exists \epsilon = \epsilon(c) > 0$ such that whp

$$\epsilon < q^*(G_n) < 1 - \epsilon.$$

CRITICAL

(iii) If $np \to \infty$ then whp $q^*(G_n) \to 0$. DENSE

Theorem (McDiarmid, S.)

There exists $0 < a < b$ such that, if $1/n \leq p = p(n) \leq 0.99$ then

$$\frac{a}{\sqrt{np}} < q^*(G_n) < \frac{b}{\sqrt{np}} \quad \text{whp.}$$

Sampling to determine modularity

Let G be the underlying network

Let G_p be obtained by sampling each edge with probability p.

Theorem (McDiarmid, S.)

$\forall \varepsilon > 0, \exists c$ such that the following holds with probability at least $1 - \varepsilon$ if $0 < p < 1$ and graph G satisfy $e(G)p \geq c$, then

$$q^*(G_p) > q^*(G) - \varepsilon.$$

if $0 < p < 1$ and graph G satisfy $e(G)p \geq cn$, then

$$q^*(G_p) < q^*(G) + \varepsilon.$$

Sampling to determine modularity

Let G be the underlying network

Let G_p be obtained by sampling each edge with probability p.

Theorem (McDiarmid, S.)

$\forall \varepsilon > 0, \exists c$ such that the following holds with probability at least $1 - \varepsilon$ if $0 < p < 1$ and graph G satisfy $e(G)p \geq c$, then

$$q^*(G_p) > q^*(G) - \varepsilon.$$

if $0 < p < 1$ and graph G satisfy $e(G)p \geq cn$, then

$$q^*(G_p) < q^*(G) + \varepsilon.$$

Observe: if $e(G)p \leq c_0$ then $\mathbb{P}(e(G_p) = 1) \geq c_0 e^{-2c_0}$.

Earlier results: $q^*(K_n) = 0$, for $e(K_n)p \leq c_0$ whp $q^*((K_n)_p) \geq \epsilon(c_o)$.

Idea of Proof

Say vertex partition $\mathcal{A}$ is η-*fat* for G if each part has volume $\geq \eta \operatorname{vol}(G)$.

For non-empty G and vertex partition $\mathcal{A}_0$, can construct η-fat 'coarser' $\mathcal{A}$ with

$$q_{\mathcal{A}}(G) > q_{\mathcal{A}_0}(G) - 2\eta.$$

I): If $0 < p < 1$ and $e(G)p \geq c$, then w.p. $1 - \epsilon$, $q^*(G_p) > q^*(G) - \epsilon$.

Can assume $q^*(G) = q_{\mathcal{A}_0}(G) \geq \epsilon$.

From $\mathcal{A}_0$ construct 'coarser' partition $\mathcal{A}$ $\frac{\epsilon}{4}$-fat for G.

Show $q_{\mathcal{A}}(G_p) > q_{\mathcal{A}}(G) - \frac{\epsilon}{2}$.

Edges internal to $\mathcal{A}$ in G is at least $\frac{\epsilon}{2}e(G)$.

Edges internal to $\mathcal{A}$ in G_p approx $\frac{\epsilon}{2}e(G)p$. Thus $q^E_{\mathcal{A}}(G_p) > q^E_{\mathcal{A}}(G) - \frac{\epsilon}{4}$.

Idea of Proof

Say vertex partition $\mathcal{A}$ is η-*fat* for G if each part has volume $\geq \eta \operatorname{vol}(G)$.

For non-empty G and vertex partition $\mathcal{A}_0$, can construct η-fat 'coarser' $\mathcal{A}$ with

$$q_{\mathcal{A}}(G) > q_{\mathcal{A}_0}(G) - 2\eta.$$

II): If $0 < p < 1$ and $e(G)p \geq cn$, then w.p. $1 - \epsilon$, $q^*(G_p) < q^*(G) + \epsilon$.

Define bad events:

$\mathcal{B}_1$ $\exists A \subseteq V$ with $\operatorname{vol}(G) < \frac{\eta}{2}\operatorname{vol}(G)$ and $\operatorname{vol}(G_p) > \eta \operatorname{vol}(G_p)$.

i.e. $\exists A$ *not* $\frac{\eta}{2}$-fat for G but *is* η-fat for G_p.

$\mathcal{B}_2$ $\exists \mathcal{A}$ $\frac{\eta}{2}$-fat for G and $q_{\mathcal{A}}(G_p) < q_{\mathcal{A}}(G) + \epsilon/2$. .

Edge limited search

c-edge-limited search operates as follows.

initialise the sets F to be empty and $\tilde{E}$ to contain all the edges of K_n.

while $\tilde{E} \neq \emptyset$ and $|F| < c$,

sample $e \in \tilde{E}$ uniformly at random, and delete it from $\tilde{E}$

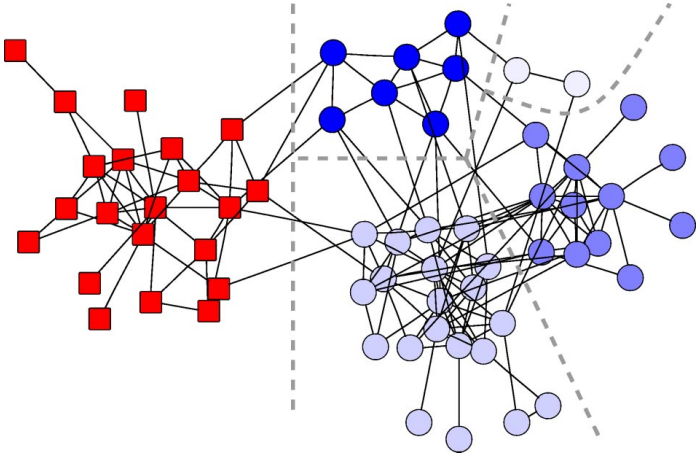
if $e \in E(G)$ **then** add e to F

return the graph $G^c = (V, F)$, where $V = [n]$.

Corollary (McDiarmid, S.)

Given $\epsilon > 0$ there exists $c = c(\epsilon)$ such that the following holds. When we run edge-limited search on G, with probability $> 1 - \epsilon$,
the random output graph G^c satisfies $q^*(G^c) > q^*(G) - \epsilon$
the random output graph G^{cn} satisfies $q^*(G^{cn}) < q^*(G) + \epsilon$

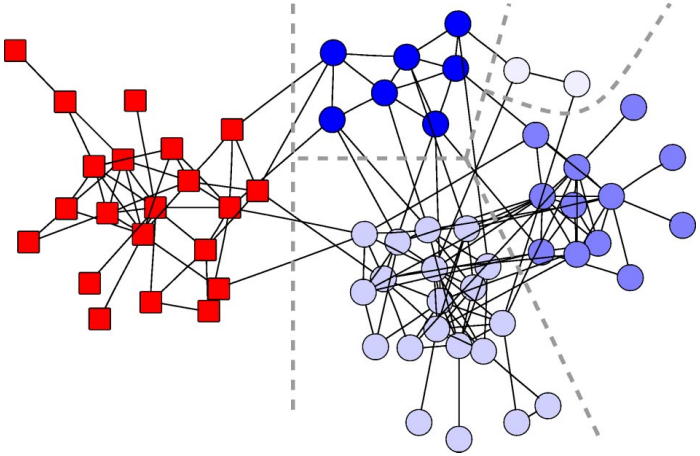
Lusseau PhD Thesis



dolphins =62
edges =159

$q^* = 0.52$

Lusseau PhD Thesis



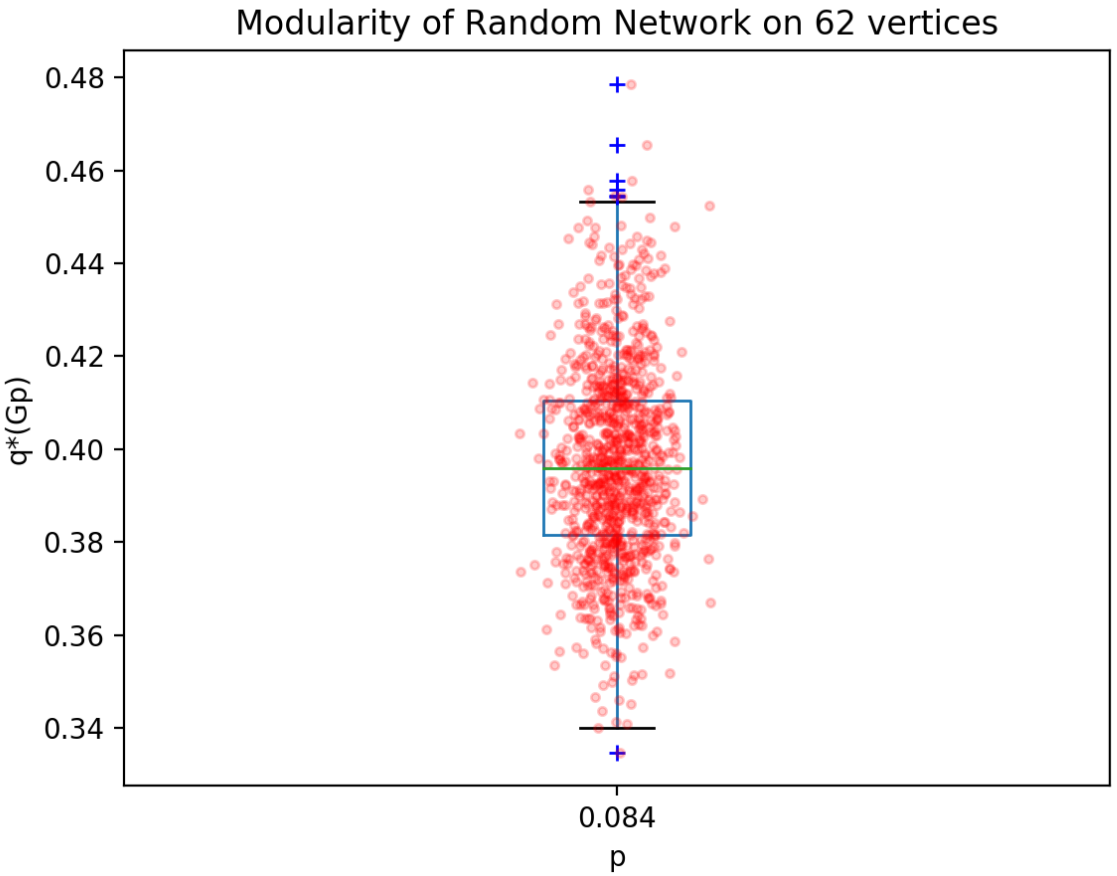
dolphins =62
edges =159

$q^* = 0.52$

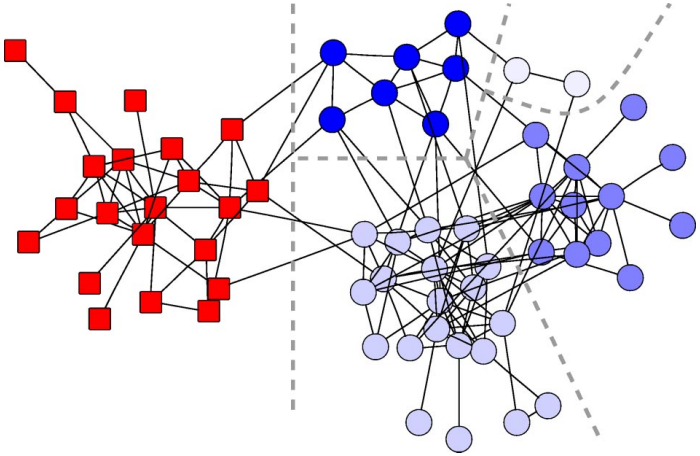
8.4% of possible edges

Random data

$q^*(\text{dolphins}) > q^*(\text{random network})$??



Lusseau PhD Thesis



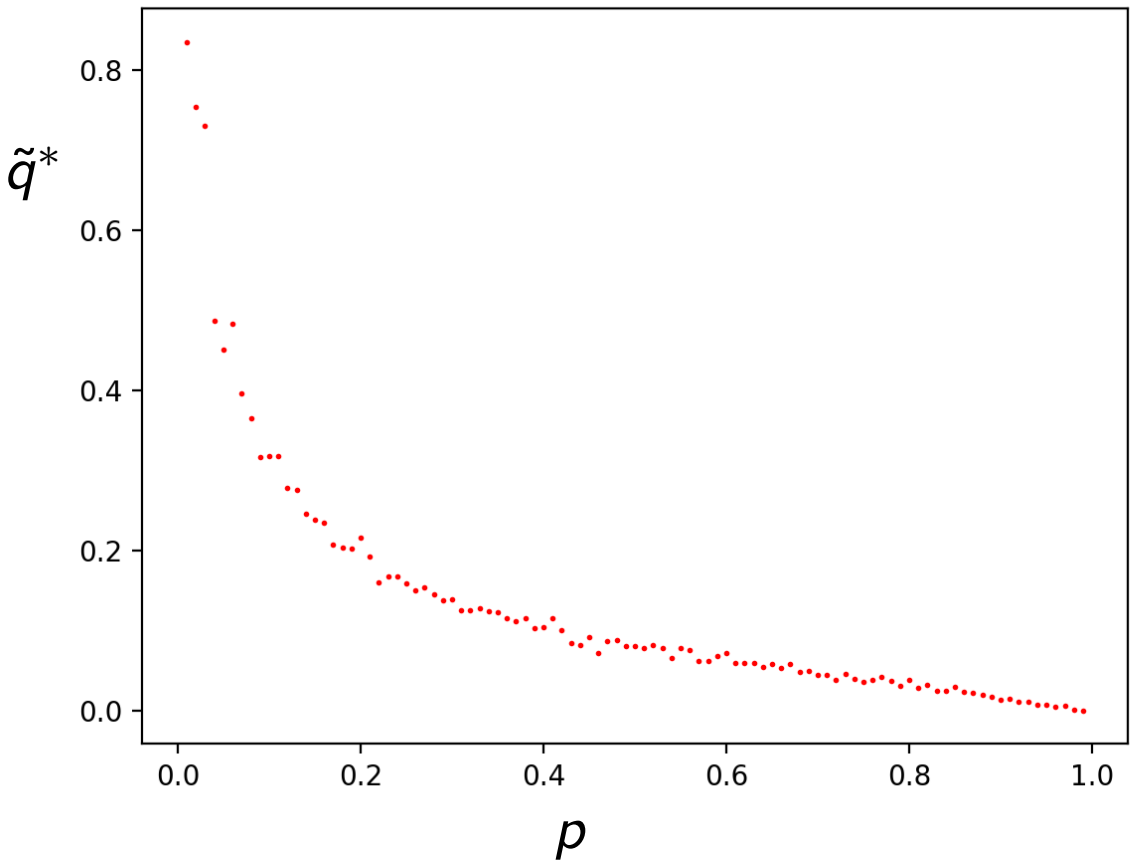
dolphins =62
edges =159

$q^* = 0.52$

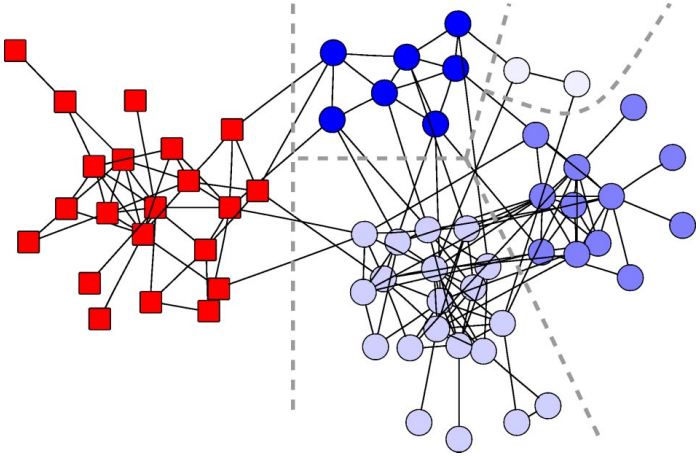
8.4% of possible edges

Random data

Simulate 62 vertices, with edge prob p.



Lusseau PhD Thesis



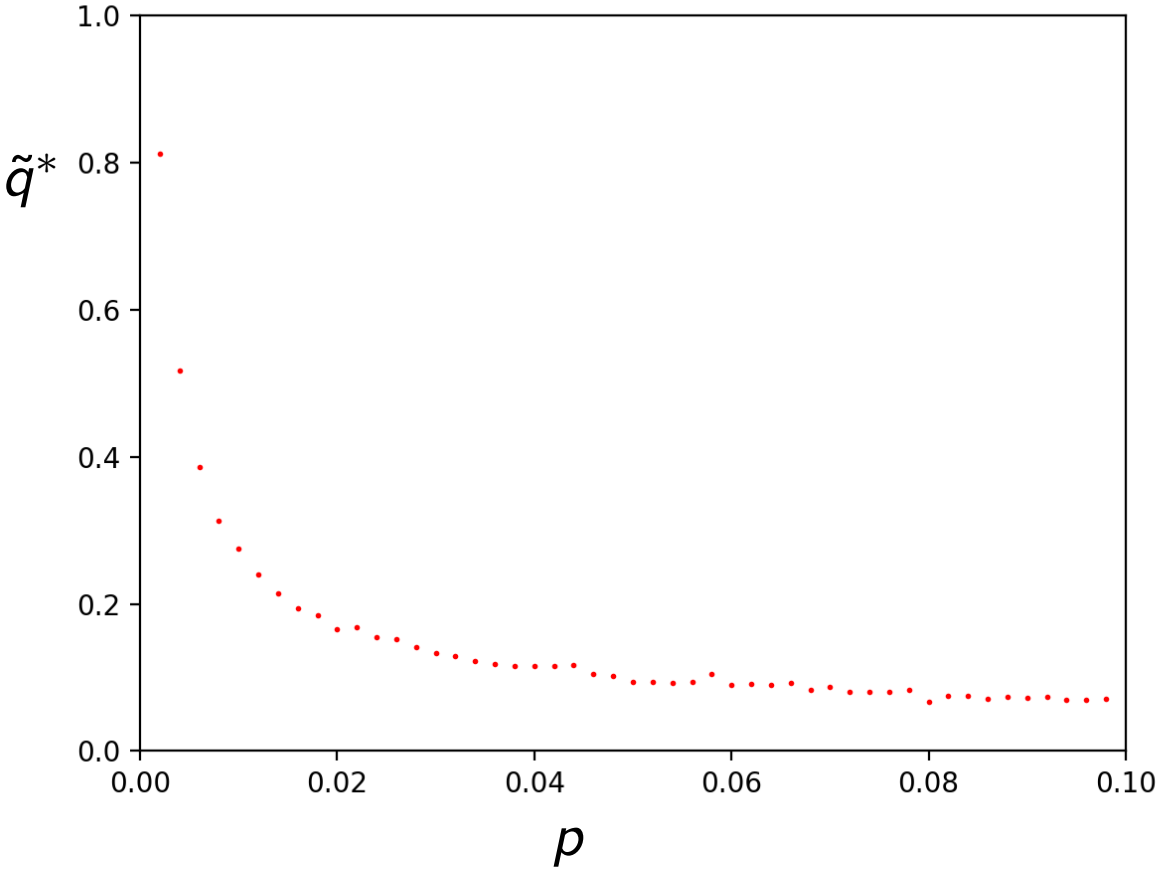
dolphins =62
edges =159

$q^* = 0.52$

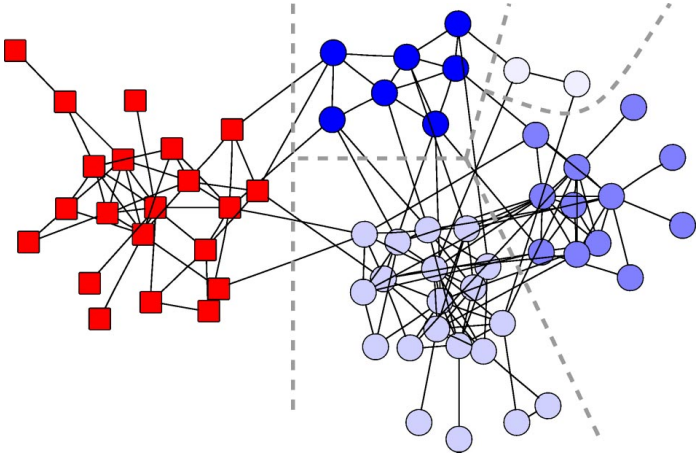
8.4% of possible edges

Random data

Simulate 1000 vertices, with edge prob p.



Lusseau PhD Thesis



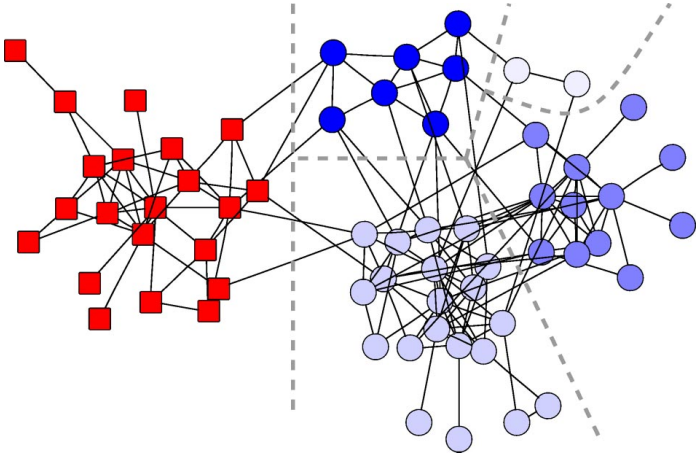
dolphins =62
edges =159

$q^* = 0.52$

Underlying network??

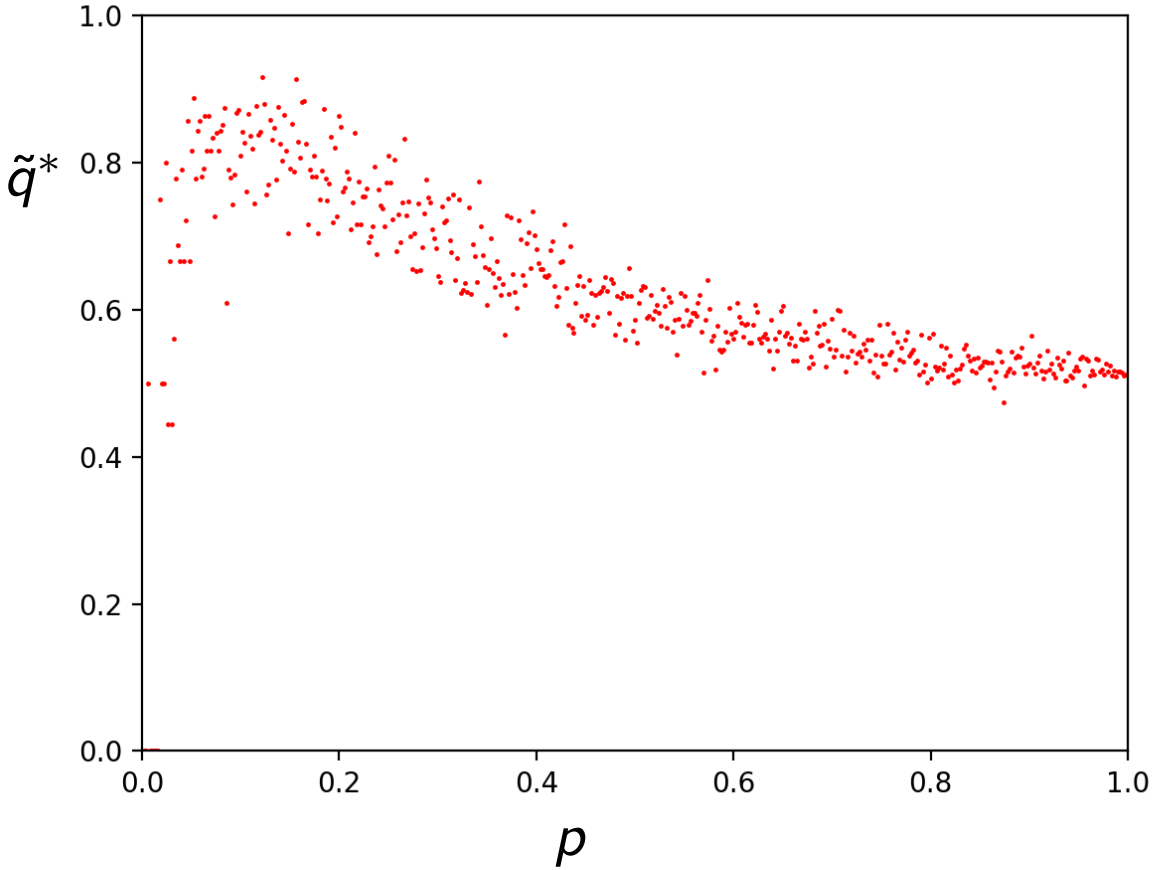
q^*(observed) vs q^*(underlying)?

Lusseau PhD Thesis

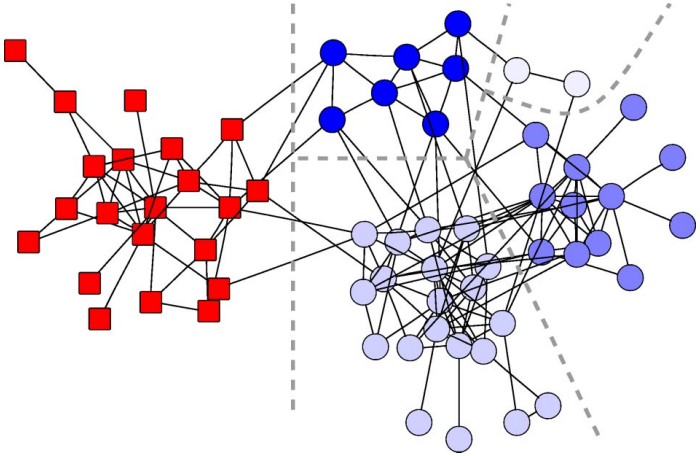


dolphins =62
edges =159

$q^* = 0.52$

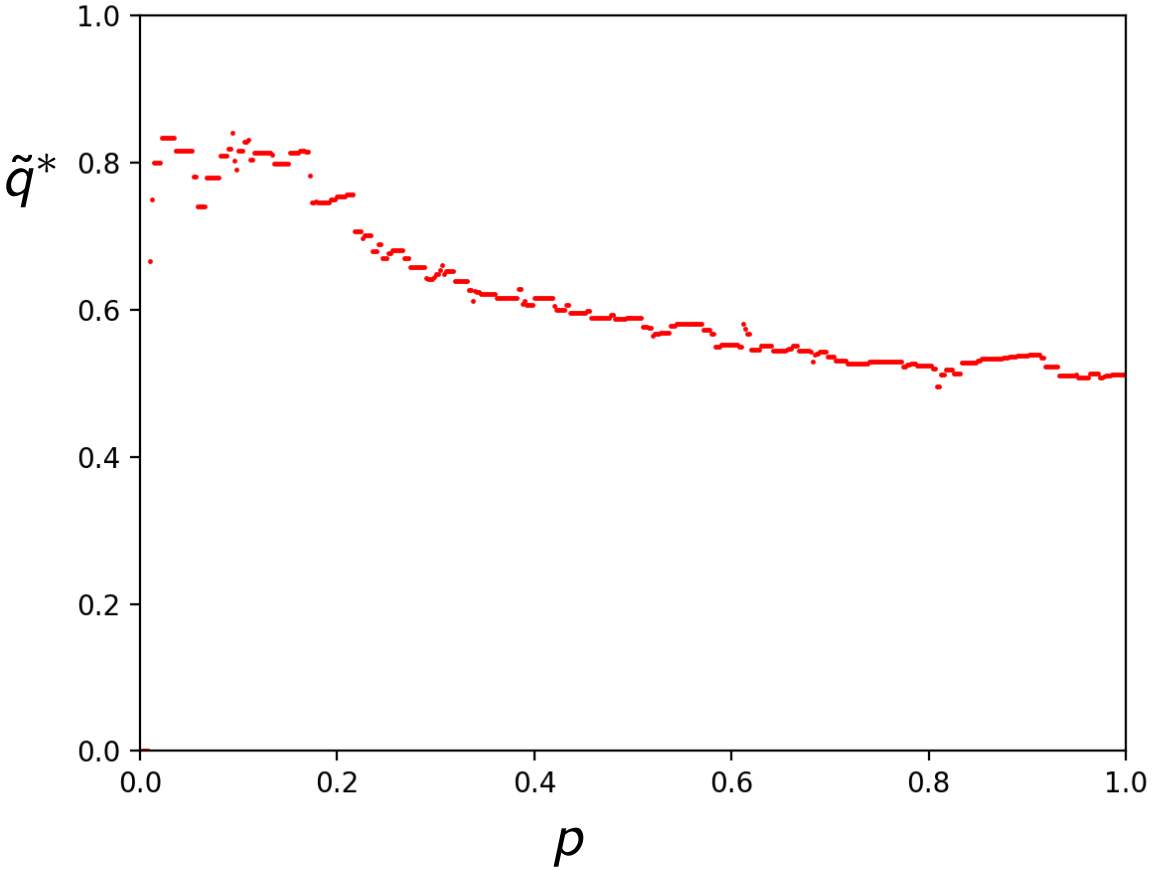


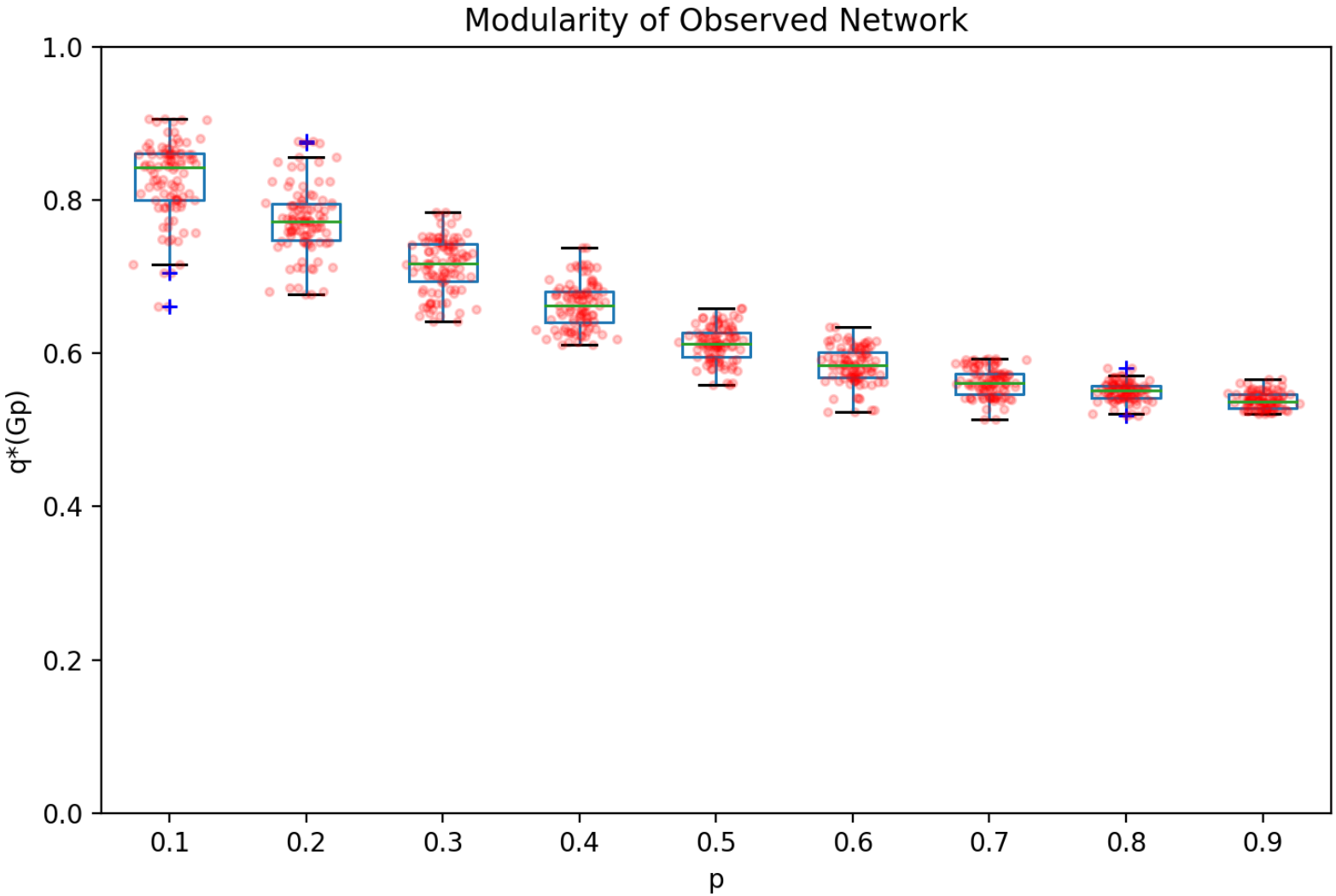
Lusseau PhD Thesis



dolphins =62
edges =159

$q^* = 0.52$




Modularity of Observed Network
1.0
0.8
0.6
0.4
0.2
0.0
q*(Gp)
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
p

Ad: Parametrised Complexity of modularity

Theorem (Meeks, S.)

Computing the modularity is w[1]-hard when parametrised by pathwidth.

Open Question:

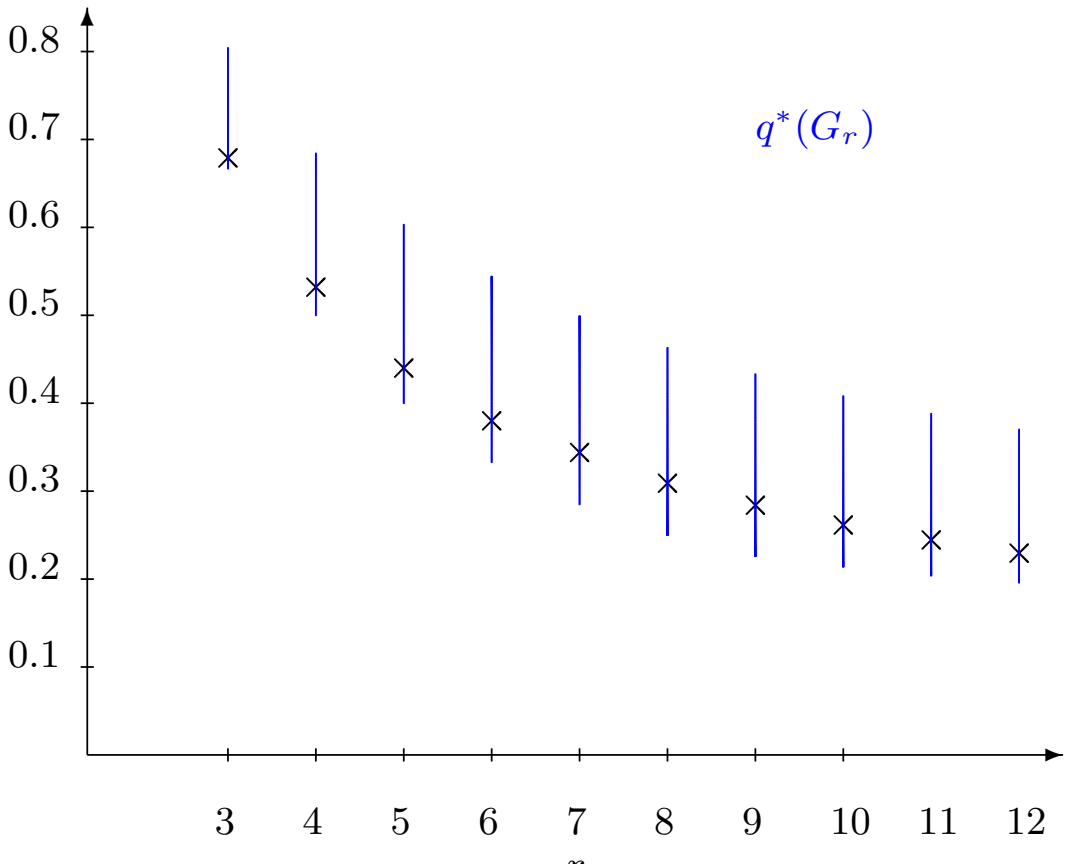
Parametrised by shrub-depth?

Idea for hardness result

Suppose only allowed partitions where each part has exactly 4 crossing-edges. Then over any possible graphs and partitions, get **best modularity** if $\mathrm{vol}(A) = 2\sqrt{2m}$ for each part. Reduces to AECP [Enciso et al. 2009]

Random r-Regular Graphs

	$r =$	3	4	5	6	7	8	9	10	11	12
bounds	$q^*(G_r) >$	0.666	0.500	0.400	0.333	0.285	0.250	0.226	0.214	0.204	0.196
	$q^*(G_r) <$	0.804	0.684	0.603	0.544	0.499	0.463	0.433	0.408	0.388	0.370
simulations	louvain	0.679	0.531	0.440	0.380	0.343	0.312	0.284	0.262	0.244	0.230
	reshuffle	0.677	0.531	0.446	0.391	0.353	0.326	0.303	0.285	0.269	0.256



McDiarmid & S.

Prokhorenkova et al

Random r-Regular Graphs

	$r =$	3	4	5	6	7	8	9	10	11	12
bounds	$q^*(G_r) >$	0.666	0.500	0.400	0.333	0.285	0.250	0.226	0.214	0.204	0.196
	$q^*(G_r) <$	0.804	0.684	0.603	0.544	0.499	0.463	0.433	0.408	0.388	0.370
simulations	louvain	0.679	0.531	0.440	0.380	0.343	0.312	0.284	0.262	0.244	0.230
	reshuffle	0.677	0.531	0.446	0.391	0.353	0.326	0.303	0.285	0.269	0.256

Open Question:

Are random graphs whp the least modular r-regular graphs?

$$q_r^-(n) = \min_{G \in \mathcal{G}_{r,n}} q^*(G)$$

Known:

$$q_2^-(n) = 1 - 2.041/\sqrt{n}$$

$$q^*(G_2) = 1 - 2/\sqrt{n} \text{ whp}$$

For $r > 2$,

$$q_r^-(n) \geq 2/r - o(1)$$

Open:

Does there exist $\delta > 0$ such that

$$q^*(G_3) > 2/3 + \delta \text{ whp ?}$$

For $r > 2$ is it true that

$$q_r^-(n) = q^*(G_r)(1 + o(1)) \text{ whp ?}$$

Open Questions

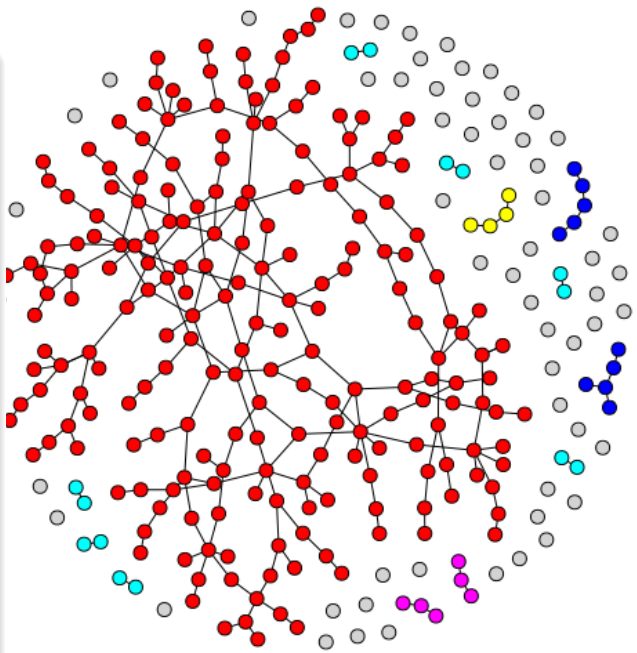
Consider $G_{n,c/n}$ for large $c > 1$.

Is it true whp $\quad q^*(G_{n,c/n}) = \frac{0.97}{\sqrt{c}} + o_c(\frac{1}{\sqrt{c}})$?

Is 5 communities best whp? (no)

Does there exist a positive integer k (perhaps $k = 5$?) with the property that, for each $\epsilon > 0$ there exists c_0 such that, if $c \geq c_0$ then whp

$$q_{\leq k}(G_{n,c/n}) \geq q^*(G_{n,c/n})(1 - \epsilon).$$



Reichardt & Bornholdt 06