When is there modularity from fluctuations in random graphs?

Fiona Skerman
Uppsala University

7th UU-TT Symposium 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.

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.

Introduction

Modularity of graph G — Newman & Girvan 2004

- *measures how well graph G can be clustered into communities*
- $0 \leq q^*(G) < 1$.
- higher value indicates *more community structure*
- *'can be clustered'* $\sim$ maximum of $q_{\mathcal{A}}(G)$ over vertex partitions $\mathcal{A}$
- most popular community detection algorithms use modularity to compare partitions
  - *Louvain, Leiden, Clauset-Newman-Moore*

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 number of edges incident to u,

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

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 number of edges incident to u,

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

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 number of edges incident to u,

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

Max. Modularity

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

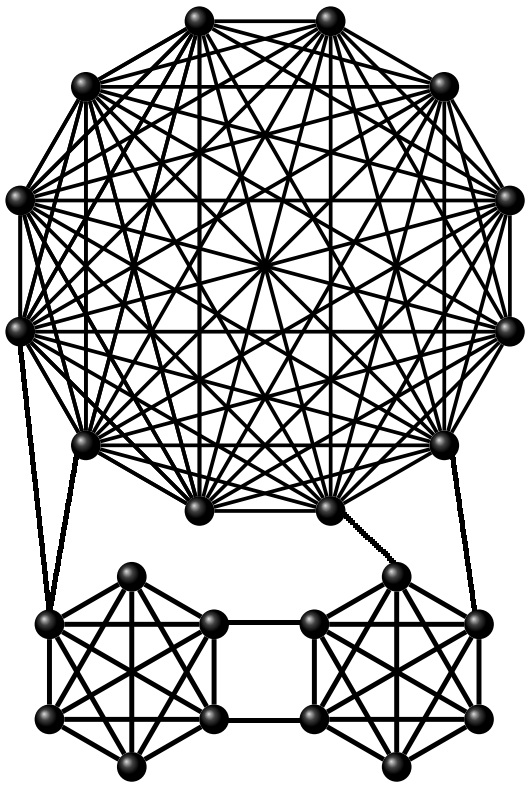
Edge contribution

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

Degree tax

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

Example Graph



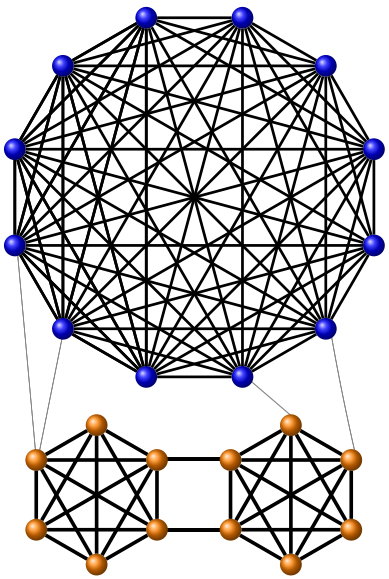
Edge contribution

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

Degree tax

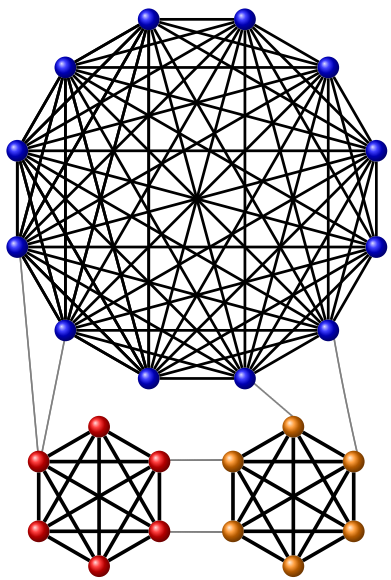
$$q^D_{\mathcal{A}}(G) = \sum_{A \in \mathcal{A}} \frac{\mathrm{vol}(A)^2}{4m^2}$$

3 Possible Partitions



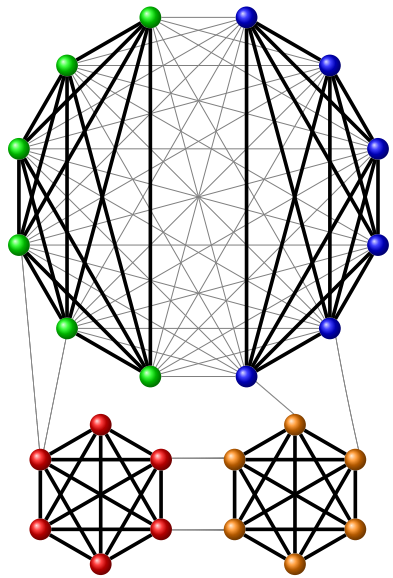
$q^E_{\mathcal{A}_1} = 0.96, \quad q^D_{\mathcal{A}_1} = 0.56$

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



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

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



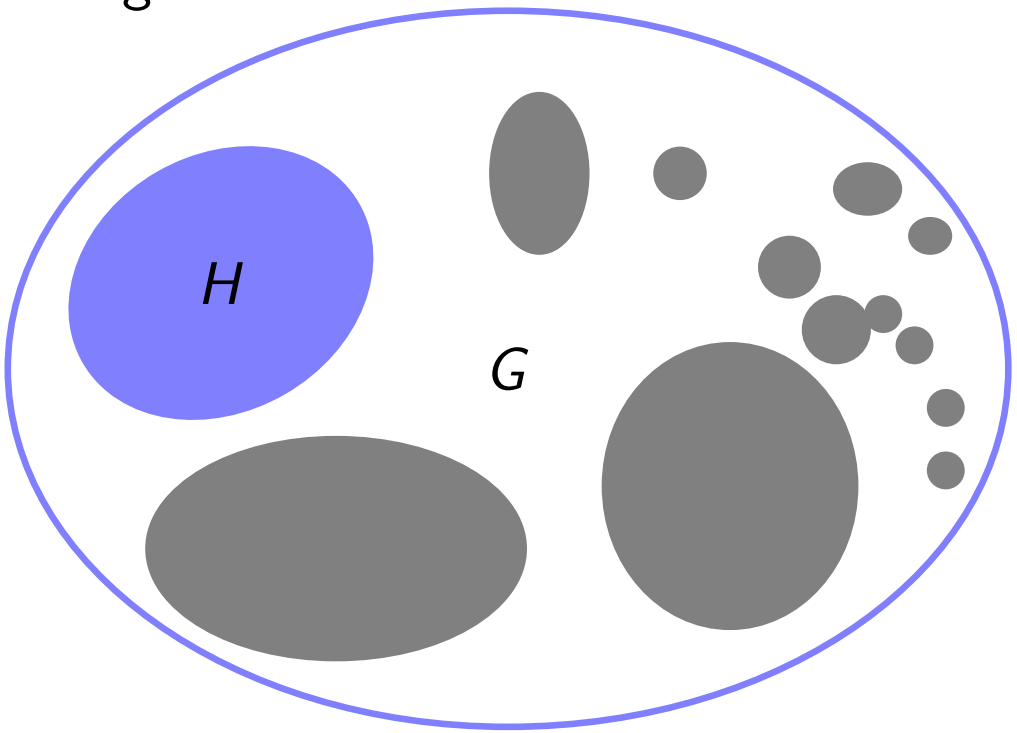
$q^E_{\mathcal{A}_3} = 0.59, \quad q^D_{\mathcal{A}_3} = 0.29$

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

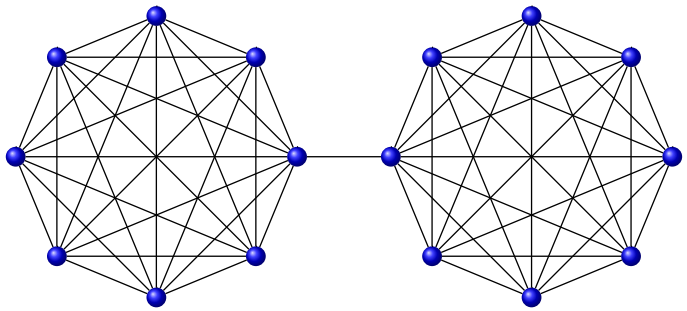
Resolution Limit

Fortunato and Barthélemy 08

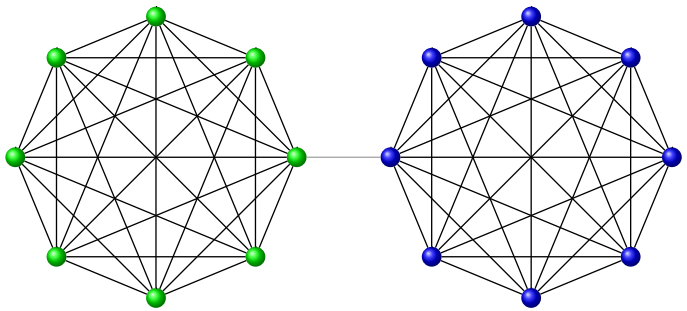
Subgraph H
h edges



Graph G, m edges

$H =$ 

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

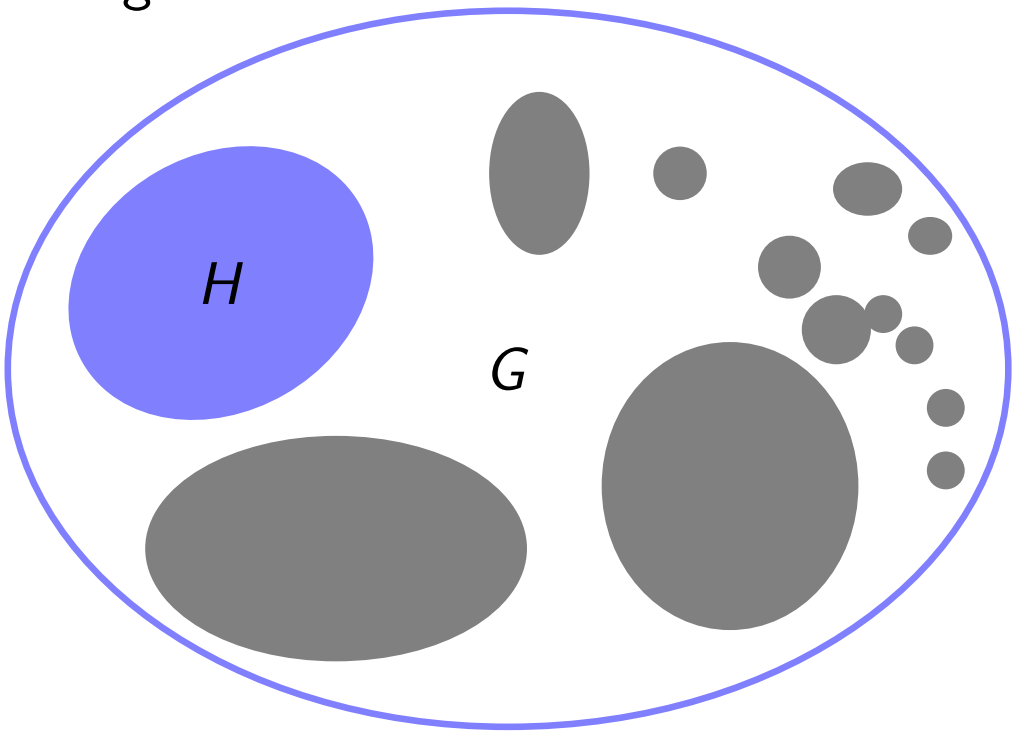


If $h > \sqrt{2m}$, e.g. $m = 1624$.

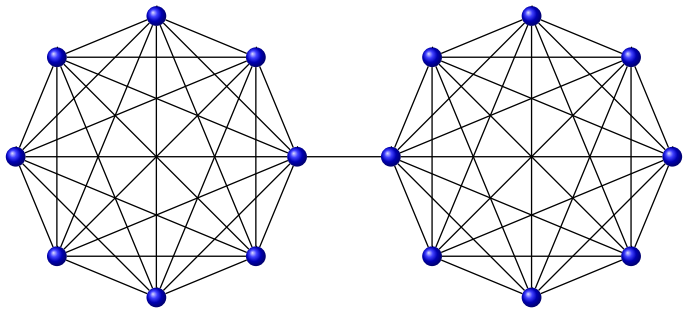
Resolution Limit

Fortunato and Barthélemy 08

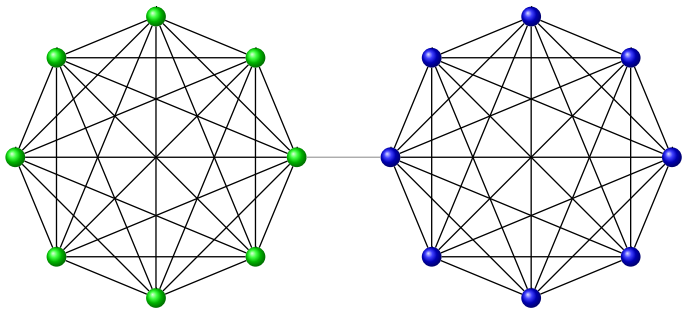
Subgraph H
h edges



Graph G, m edges

$H =$ 

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



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

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

Properties of Modularity

Partition structure is sensitive to noise in edges

Modularity value is **robust** to small perturbations in the edge set

$$|q^*(G) - q^*(G \backslash E')| \leq 2|E'|/e(G)$$

$q^*(G) > 1 - \varepsilon$ if

- connected components in G each have $< \varepsilon e(G)$ edges
- or $\exists$ partition with #edges between parts $< \frac{\varepsilon}{2} e(G)$ and volume of each part $< \varepsilon e(G)$

$q^*(G) < \varepsilon$ if

- $\bar{\lambda}(G) < \varepsilon$ where $\bar{\lambda}(G) = \max_i |1 - \lambda_i|$ is *spectral gap* of Laplacian of G [G r-regular, $\bar{\lambda}(G) = \frac{1}{r} \max_{i \neq 0} |\lambda_i(A_G)|$]
- or $\forall X \subset V(G)$, G regular $\frac{e(X, \bar{X})}{e(G)} > (2 - \epsilon) \frac{|X|}{|G|} \frac{|\bar{X}|}{|G|}$

Modularity at Criticality

Predictions from Statistical Physics for $G_n \sim \mathcal{G}(n, c/n)$, large $c > 1$

Guiméra et al. 04

$$q^*(G_n) \sim \frac{1}{c^{2/3}}$$

Reichardt & Bornholdt 06

$$q^*(G_n) \sim \frac{0.97}{c^{1/2}}$$

A differentiation between graphs which are truly modular and those which are not can ... only be made if we gain an understanding of the intrinsic modularity of random graphs – Reichardt and Bornholdt 06

Modularity at Criticality

Predictions from Statistical Physics for $G_n \sim \mathcal{G}(n, c/n)$, large $c > 1$

Guiméra et al. 04 ×

$$q^*(G_n) \sim \frac{1}{c^{2/3}}$$

Reichardt & Bornholdt 06 **?**

$$q^*(G_n) \sim \frac{0.97}{c^{1/2}}$$

Theorem (McDiarmid, S.)

Let $G_n \sim \mathcal{G}(n, p)$.
There exist constants a, b such that if $p = c/n$ with $c > 1$ then whp

$$\frac{a}{c^{1/2}} < q^*(G_n) < \frac{b}{c^{1/2}}.$$

Open Questions

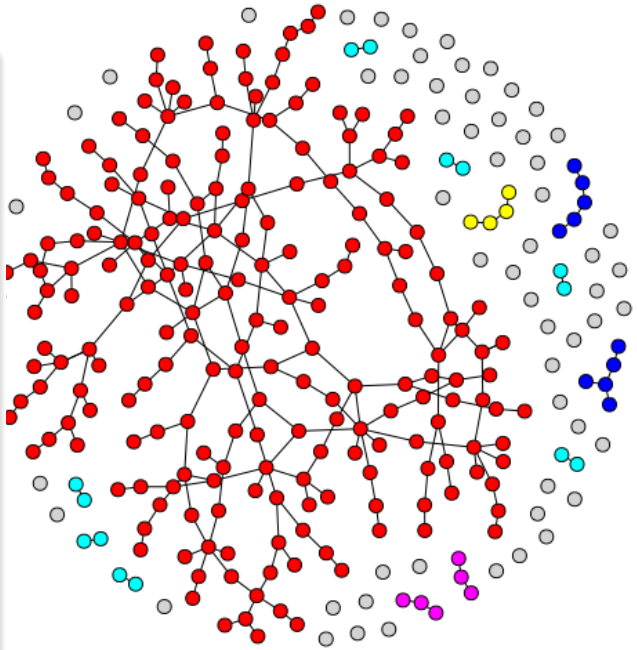
Consider $G_{n,c/n}$ for large $c > 1$.

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)

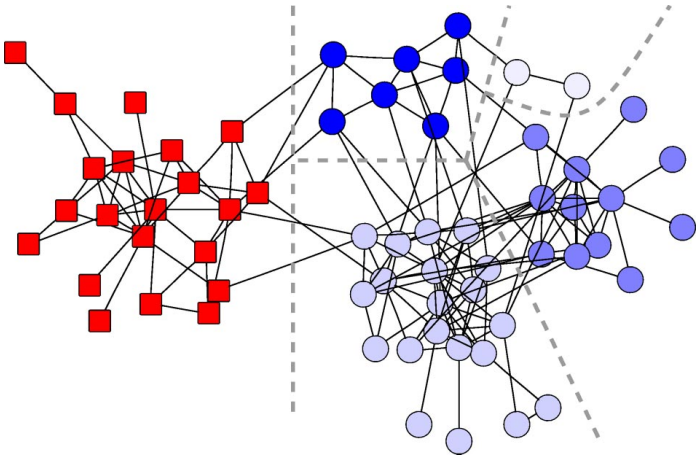
Does there exist a positive integer k (perhaps $k = 5$?) with the property that, for each $\epsilon > 0$ there exists c_0 such that, if $c \geq c_0$ then whp

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



Reichardt & Bornholdt 06

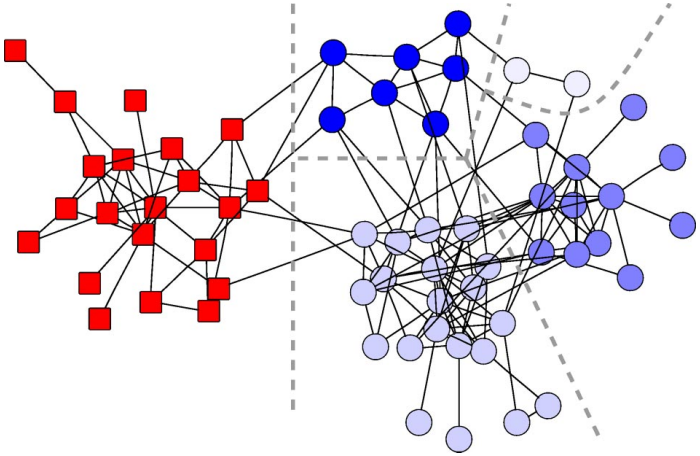
Lusseau PhD Thesis



dolphins =62
edges =159

$q^* = 0.52$

Lusseau PhD Thesis



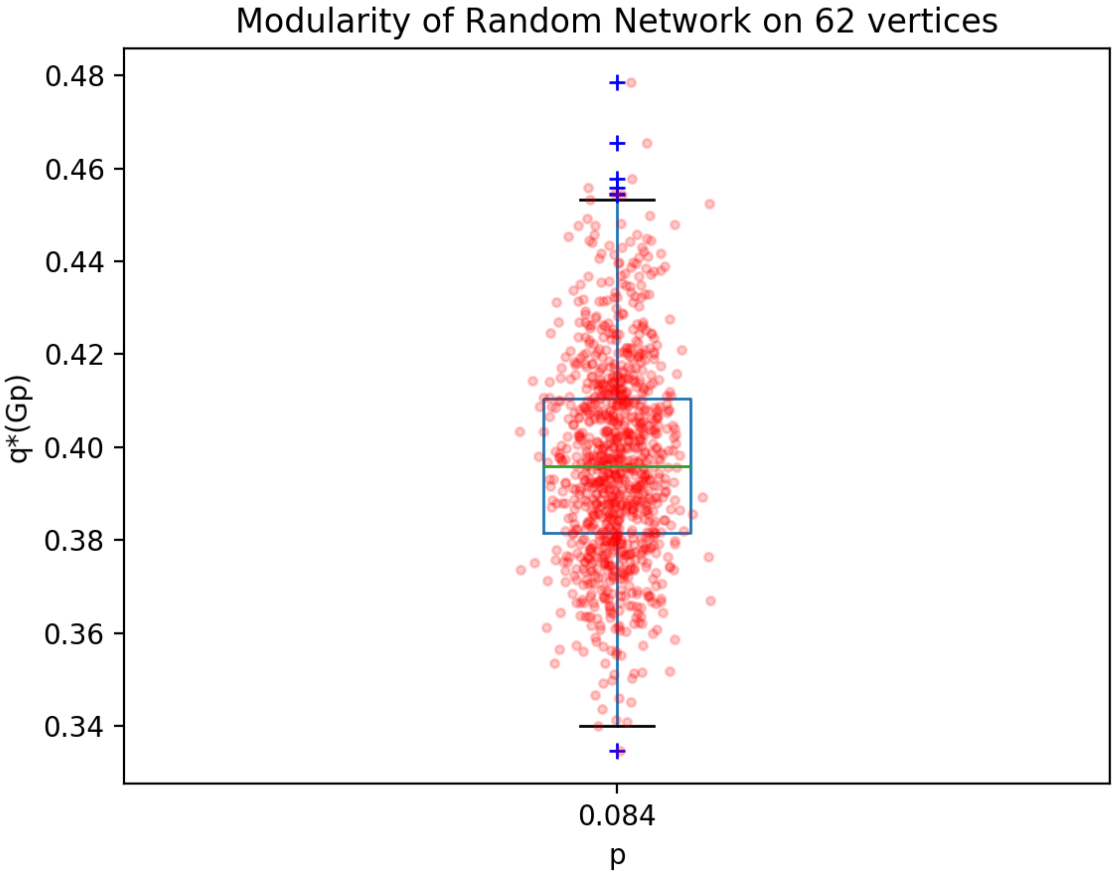
dolphins =62
edges =159

$q^* = 0.52$

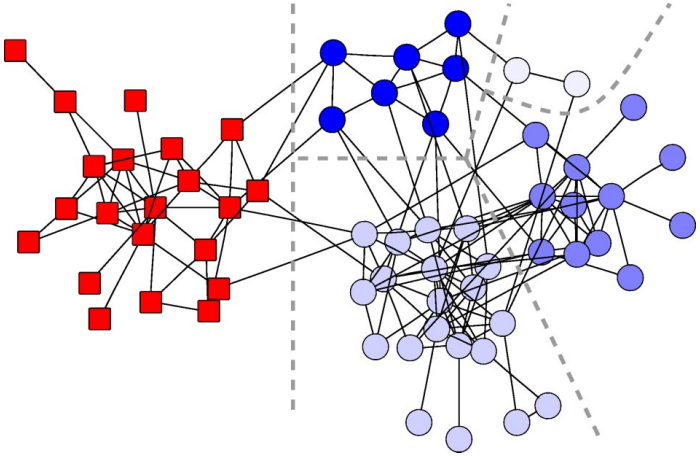
8.4% of possible edges

Random data

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



Lusseau PhD Thesis

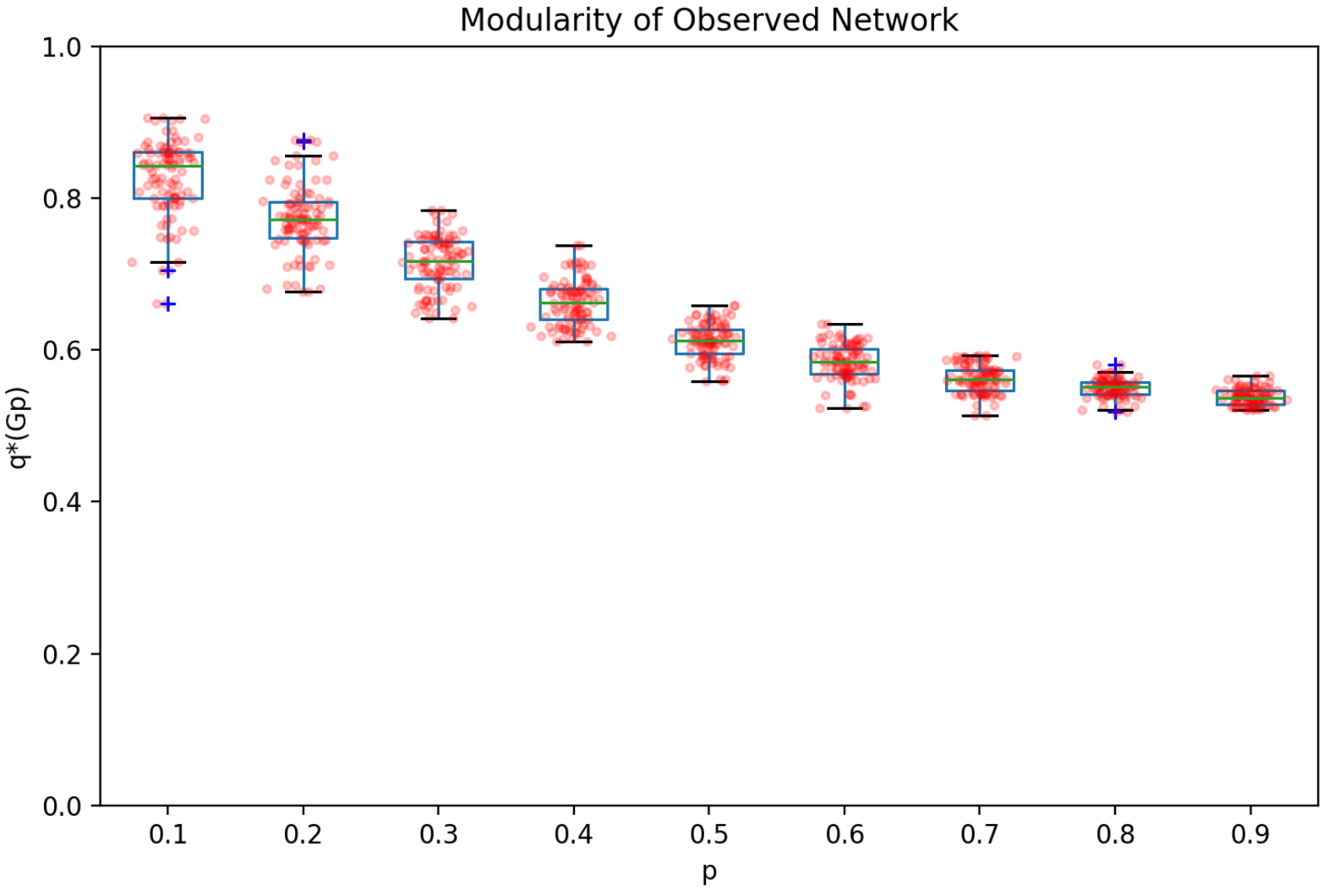


dolphins =62
edges =159

$q^* = 0.52$

Underlying network??

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


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

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

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