Computational Lower Bounds for Community Detection on Random Graphs

Bruce Hajek, Yihong Wu, Jiaming Xu

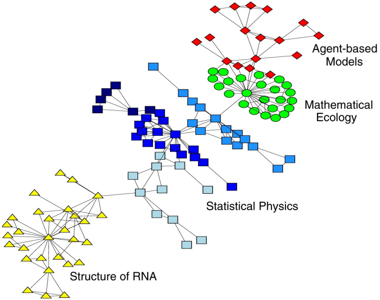
Department of Electrical and Computer Engineering
Coordinated Science Laboratory
University of Illinois at Urbana-Champaign

Wharton School of Statistics
University of Pennsylvannia

COLT, July 3-6, 2015

Community detection in networks

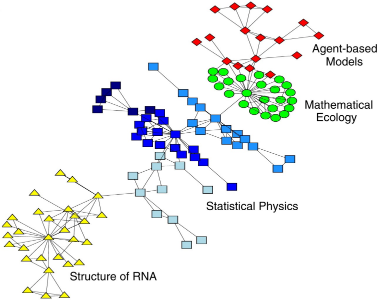
- Networks with community structures arise in many applications



Collaboration network: 118 scientists [Girvan-Newman ’02]

Community detection in networks

- Networks with community structures arise in many applications

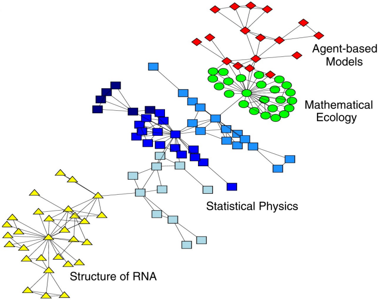


Collaboration network: 118 scientists [Girvan-Newman '02]

- Task: Find underlying communities based on the network topology

Community detection in networks

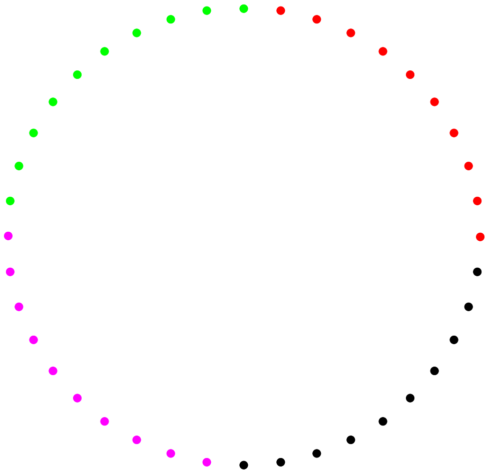
- Networks with community structures arise in many applications



Collaboration network: 118 scientists [Girvan-Newman ’02]

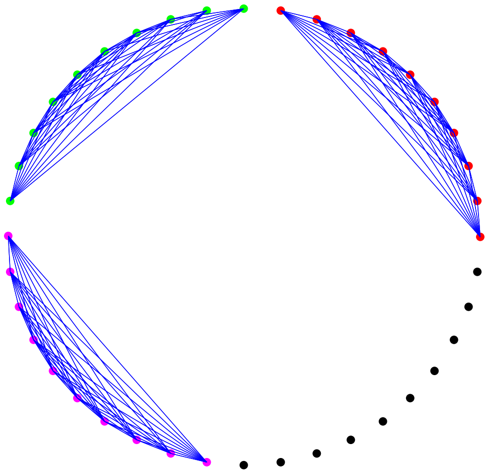
- Task: Find underlying communities based on the network topology
- Applications: Friend or movie recommendation in online social networks

Stochastic block model (Planted partition model)



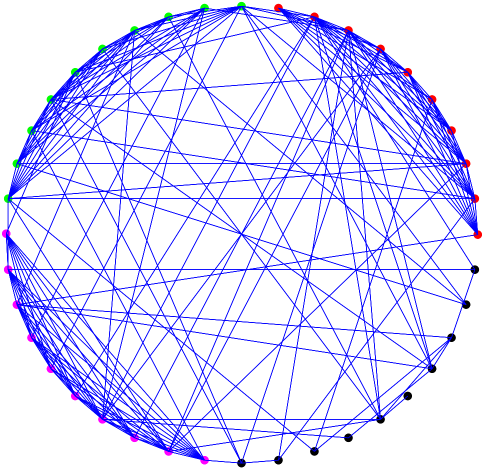
$n = 40, K = 10, r = 3$

Stochastic block model (Planted partition model)



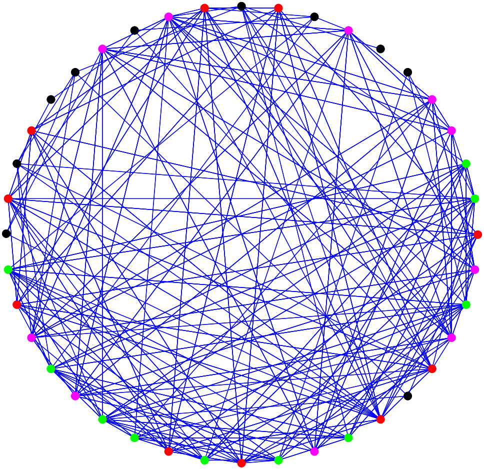
$p = 0.9$

Stochastic block model (Planted partition model)



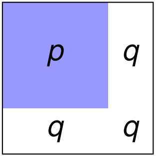
$p = 0.9 \quad q = 0.1$

Stochastic block model (Planted partition model)



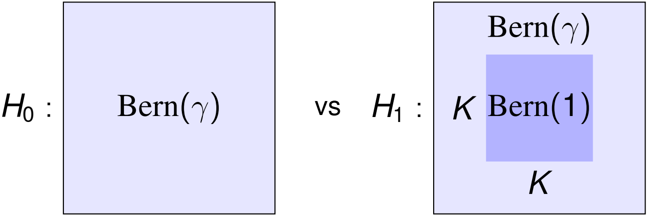
$p = 0.9 \quad q = 0.1$

This paper focuses on $r = 1$: a single community



- One cluster of size K plus $n - K$ outliers
- Connectivity p within cluster and q otherwise
- Also known as *Planted Dense Subgraph* model
- $p = 1$, $q = \gamma$ corresponds to *Planted Clique* model

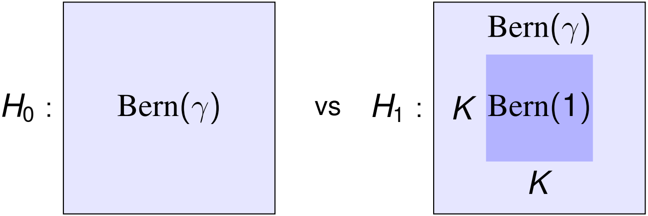
Planted clique hardness hypothesis



[Alon et al. '98] [Dekel et al. '10] [Deshpande-Montanari '13]...
Intermediate regime: $\log n \ll K \ll \sqrt{n}$, $\gamma = \Theta(1)$

- detection is possible but believed to have high computational complexity: [Alon et al. '11] [Feldman et al. '13]...

Planted clique hardness hypothesis

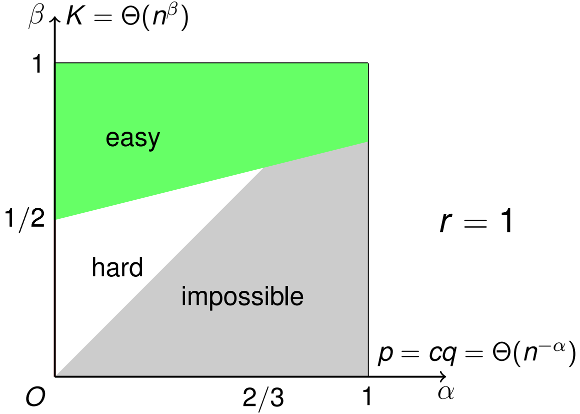


[Alon et al. '98] [Dekel et al. '10] [Deshpande-Montanari '13]...
Intermediate regime: $\log n \ll K \ll \sqrt{n}$, $\gamma = \Theta(1)$

- detection is possible but believed to have high computational complexity: [Alon et al. '11] [Feldman et al. '13]...
- many (worst-case) hardness results assuming Planted Clique hardness with $\gamma = \frac{1}{2}$
  - detecting sparse principal component [Berthet-Rigollet '13]
  - detecting sparse submatrix [Ma-Wu '13]
  - cryptography [Applebaum et al. '10]: $\gamma = 2^{-\log^{0.99} n}$

Main result: Hardness for *detecting* a single cluster

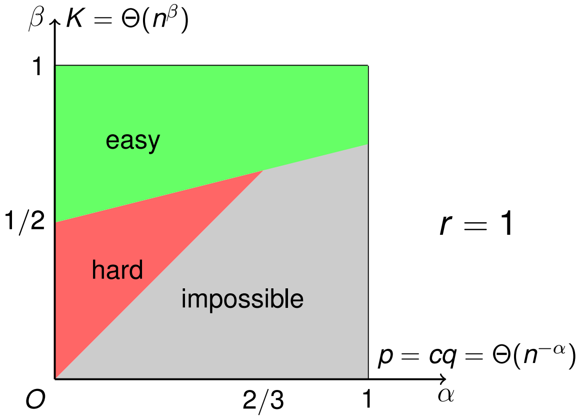
Assuming Planted Clique hardness for any constant $\gamma > 0$



$r = 1$

Main result: Hardness for *detecting* a single cluster

Assuming Planted Clique hardness for any constant $\gamma > 0$

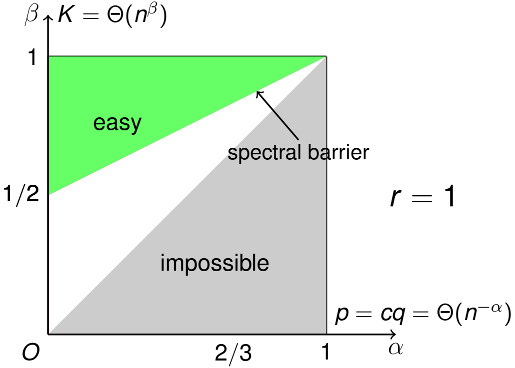


$r = 1$

Detecting a single cluster in the red regime is at least as hard as detecting a clique of size $K = o(\sqrt{n})$

Corollary: Hardness for *recovering* a single cluster

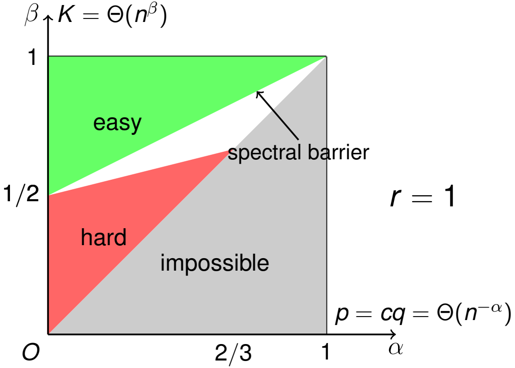
Can show: Hardness of detection implies hardness or recovery, so:
Assuming Planted Clique hardness for any constant $\gamma > 0$



$r = 1$

Corollary: Hardness for *recovering* a single cluster

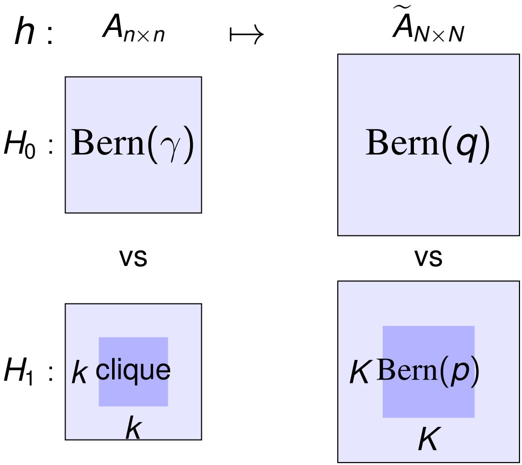
Can show: Hardness of detection implies hardness or recovery, so:
Assuming Planted Clique hardness for any constant $\gamma > 0$



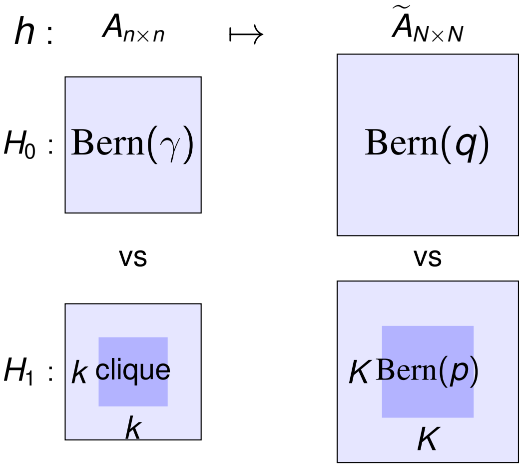
$r = 1$

Recovering a single cluster in the red regime is at least as hard as detecting a clique of size $K = o(\sqrt{n})$

Proof requires a polynomial time reduction



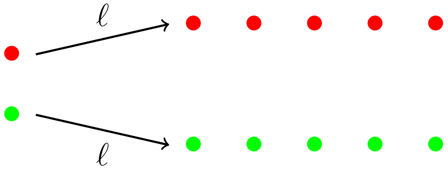
Proof requires a polynomial time reduction



$h : A \mapsto \widetilde{A}$ is agnostic to the clique and can be computed in P-time

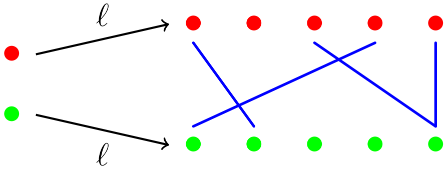
Given an integer ℓ, two probability distributions P, Q on $\{0, 1, \ldots, \ell^2\}$

Split each node
into ℓ new nodes
$N = n\ell, K = k\ell$



Given an integer ℓ, two probability distributions P, Q on $\{0, 1, \ldots, \ell^2\}$

Split each node into ℓ new nodes
$N = n\ell, K = k\ell$



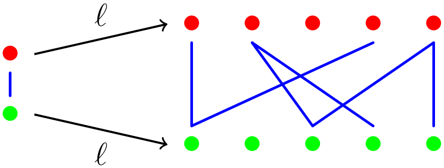
Assign edges with distributions P, Q

$0 \mapsto Q$

Given an integer ℓ, two probability distributions P, Q on $\{0, 1, \dots, \ell^2\}$

Split each node into ℓ new nodes

$N = n\ell, K = k\ell$



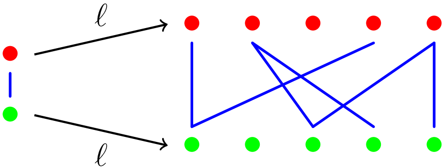
Assign edges with distributions P, Q

$0 \mapsto Q$

$1 \mapsto P$

Given an integer ℓ, two probability distributions P, Q on $\{0, 1, \ldots, \ell^2\}$

Split each node into ℓ new nodes
$N = n\ell, K = k\ell$



Assign edges with distributions P, Q

$$0 \mapsto Q$$

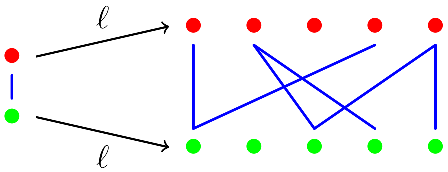
$$1 \mapsto P$$

$$H_0: \quad \mathrm{Bern}(\gamma) \qquad (1-\gamma)Q + \gamma P$$

$$H_1: \quad \mathrm{Bern}(1) \text{ (in-clique)} \qquad P \text{ (in-cluster)}$$

Given an integer ℓ, two probability distributions P, Q on $\{0, 1, \ldots, \ell^2\}$

Split each node into ℓ new nodes
$N = n\ell, K = k\ell$



Assign edges with distributions P, Q

$$0 \mapsto Q$$
$$1 \mapsto P$$

$$H_0 : \quad \mathrm{Bern}(\gamma) \qquad (1-\gamma)Q + \gamma P$$

$$H_1 : \quad \mathrm{Bern}(1) \text{ (in-clique)} \qquad P \text{ (in-cluster)}$$

How to choose P, Q?

Matching H_0: $(1-\gamma)Q + \gamma P = \mathrm{Binom}(\ell^2, q)$
Matching H_1 approximately: $P \approx \mathrm{Binom}(\ell^2, p)$ in total variation distance

Please see paper for more information and references

Thanks!