Computational Lower Bounds for Community Detection on Random Graphs

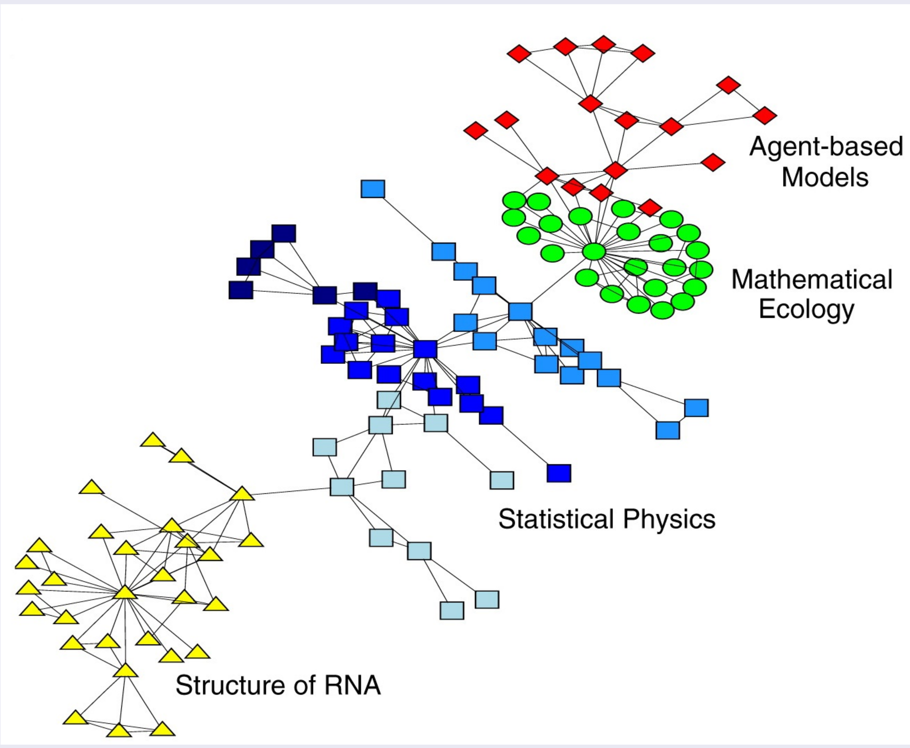
Bruce Hajek, Yihong Wu, and Jiaming Xu

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

Wharton School of Statistics
University of Pennsylvannia

Community detection in networks

- Networks with community structures arise in many applications



Collaboration network: [Girvan-Newman '02]

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

Cluster recovery under stochastic blockmodel

Vast literature on stochastic blockmodel [Holland et al. '83] and planted partition model [Condon-Karp '01]:

- [Bickel-Chen '09] [Rohe et al. '10] [Jin '12] [Mossel et al. '12] [Mossel et al. '13] [Mossel et al. '14] [Cai-Li' 14] [Guédon-Vershynin '14] [Arias-Castro-Verzelen '14] [Lei-Rinaldo '14] [Le-Levina-Vershynin '15] ...
- [Karrer-Newman '11] [Decelle et al. '11] [Nadakuditi-Newman '12] [Krzakala et al. '13] [Saade et al. '15] ...
- [McSherry '01] [Coja-Oghlan '10] [Chaudhuri et al. '12] [Ames '12] [Chen-Sanghavi-Xu '12] [Heimlicher et al. '12] [Anandkumar et al. '13] [Lelarge et al. '13] [Massoulié '13] [Vinayak-Oymak-Hassibi '14] [Abbe et al. '14] [Yun-Proutiere '14] [Abbe-Sandon '15] [Chin-Rao-Vu '15] ...

This paper focuses on a single community

p	q
q	q

- 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

H_0 : Bern(γ) vs H_1 : Bern(γ) with a $K \times K$ block Bern(1)

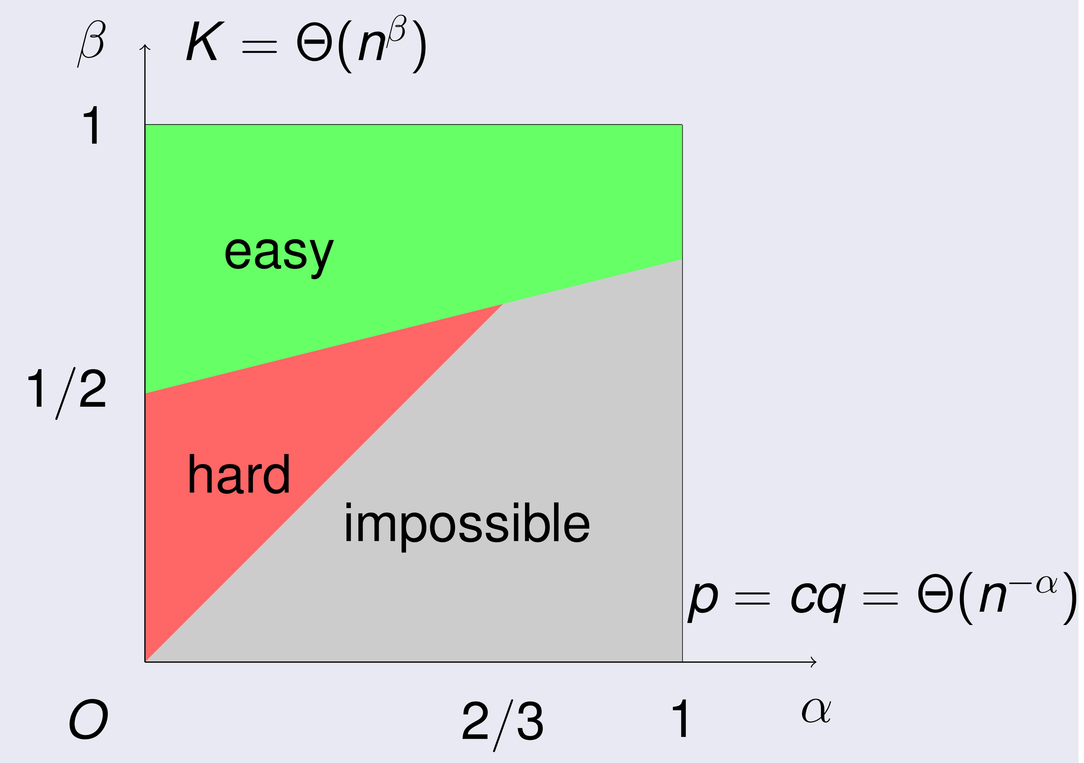
[Alon et al. '98] [Dekel et al. '10] [Deshpande-Montanari '13]...

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]: $\gamma = 2^{-\log^{0.99} n}$

Hardness for *detecting* a single cluster

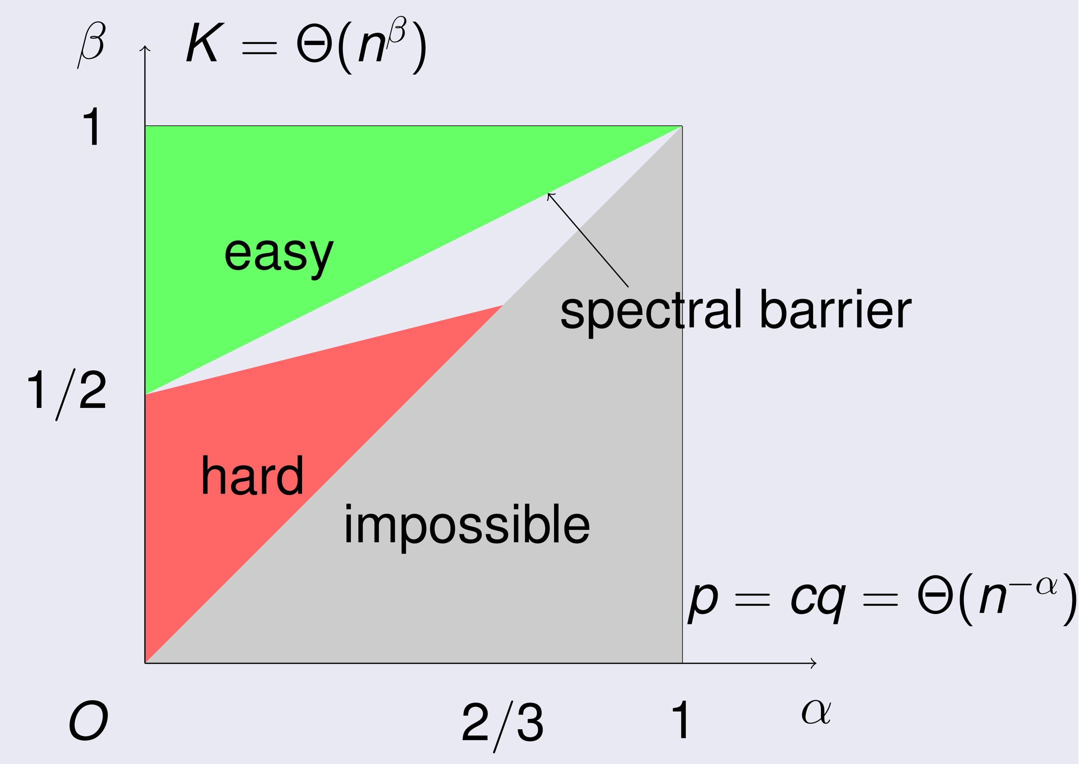
Assuming Planted Clique hardness for any constant $\gamma > 0$:



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

Hardness for *recovering* a single cluster

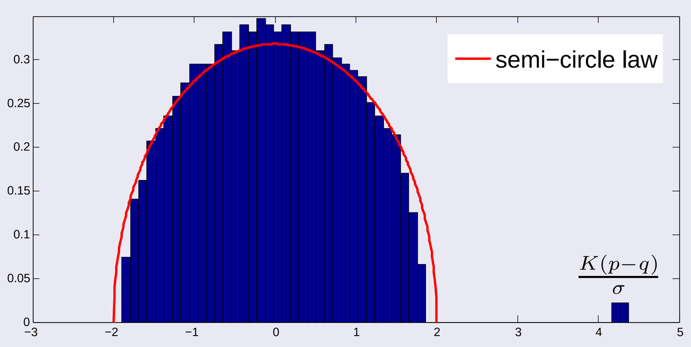
Can show: Hardness of detection implies hardness or recovery, so:
Assuming Planted Clique hardness for any constant $\gamma > 0$:



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

About the spectral barrier [Nadakuditi-Newman '12]

$$A = \begin{array}{cc} & K \\ K & \begin{array}{|c|c|} \hline p & q \\ \hline q & q \\ \hline \end{array} \end{array} + A - \mathbb{E}[A]$$



Eigenvalue distribution of $\frac{A - q\mathbf{1}\mathbf{1}^\top}{\sigma}$ for $\sigma = \sqrt{Kp + (n-K)q}$

Conjecture [Chen-Xu '14]: no polynomial-time algorithm can recover beyond the spectral barrier. (Our corollary partially resolves this conjecture.)

Formal statement of hardness of detecting a cluster

γ: edge probability in Planted Clique

Theorem

Assume Planted Clique Hypothesis holds for all $0 < \gamma \le 1/2$. *Let* $\alpha > 0$ *and* $0 < \beta < 1$ *be such that*

$$\alpha < \beta < \frac{1}{2} + \frac{\alpha}{4}.$$

Then there exists a sequence $\{(N_\ell, K_\ell, q_\ell)\}_{\ell \in \mathbb{N}}$ *satisfying* $\lim_{\ell\to\infty} \frac{-\log q_\ell}{\log N_\ell} = \alpha$ *and* $\lim_{\ell\to\infty} \frac{\log K_\ell}{\log N_\ell} = \beta$ *such that for any sequence of randomized polynomial-time tests* ϕ_ℓ *for the* PDS$(N_\ell, K_\ell, 2q_\ell, q_\ell)$ *problem, the Type-I+II error probability is lower bounded by* 1.

Proof requires a polynomial time reduction

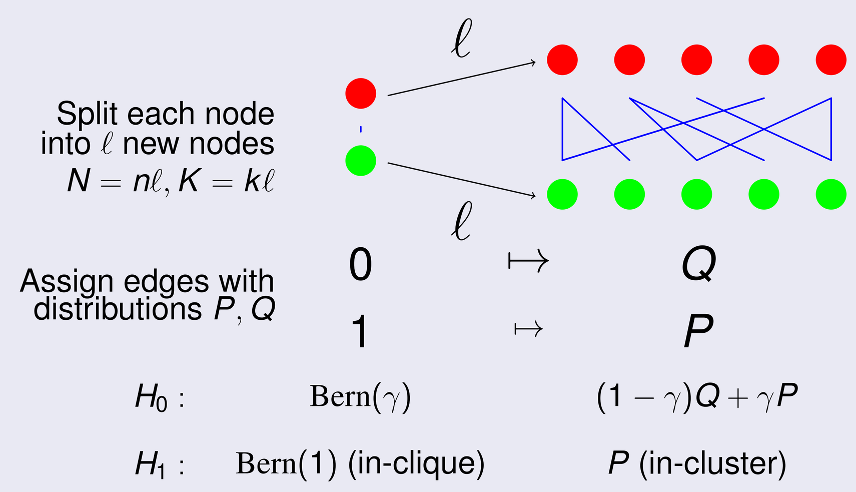
$$h: A_{n\times n} \mapsto \widetilde{A}_{N\times N}$$

H_0 : Bern(γ) $\mapsto$ Bern(q)

vs vs

H_1 : clique K $\mapsto$ Bern(p) K

Need $h: A \mapsto \widetilde{A}$ agnostic to the clique and computable in polynomial time.
Given an integer ℓ, two probability distributions P, Q on $\{0, 1, \dots, \ell^2\}$



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

Lemma (Bound the total variation distance)

Let $\ell, n \in \mathbb{N}$, $k \in [n]$ *and* $\gamma \in (0, \frac{1}{2}]$. *Let* $N = \ell n$, $K = k\ell$, $p = 2q$ *and* $m_0 = \lfloor \log_2(1/\gamma) \rfloor$. *Assume that* $16q\ell^2 \le 1$ *and* $k \ge 6e\ell$. *If* $G \sim \mathcal{G}(n, \gamma)$, *then* $\widetilde{G} \sim \mathcal{G}(N, q)$. *If* $G \sim \mathcal{G}(n, k, 1, \gamma)$, *then*

$$d_{\mathrm{TV}}\left(P_{\widetilde{G}}, \mathcal{G}(N, K, p, q)\right) \lesssim e^{-K} + ke^{-\ell} + k^2 (q\ell^2)^{m_0+1} + \sqrt{e^{q\ell^2} - 1}$$

Please see paper for more information and references

Thanks!