Community Detection in Networks: SDP relaxation and Computational Gaps

Yihong Wu

Department of ECE
University of Illinois at Urbana-Champaign

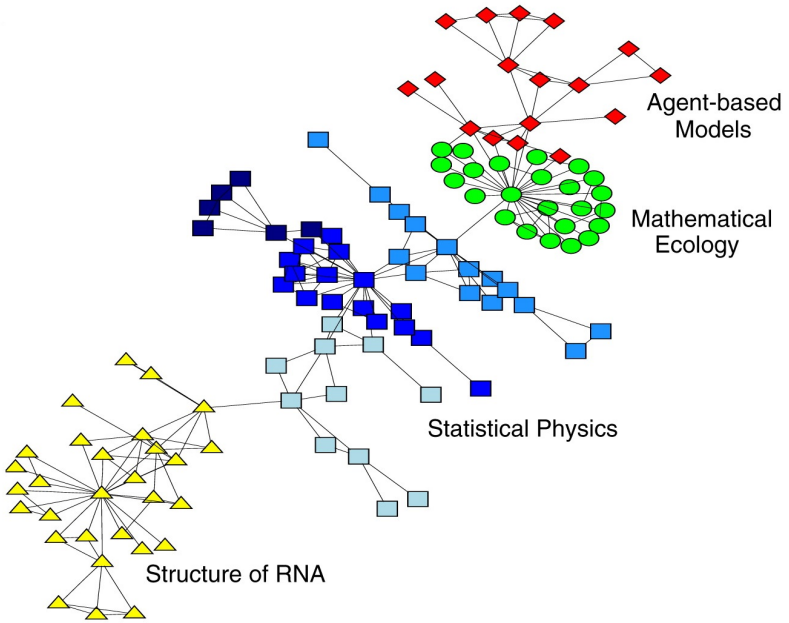
yihongwu@illinois.edu

Joint work with Bruce Hajek (Illinois) and Jiaming Xu (Wharton)

May 20, 2015

Community detection in networks

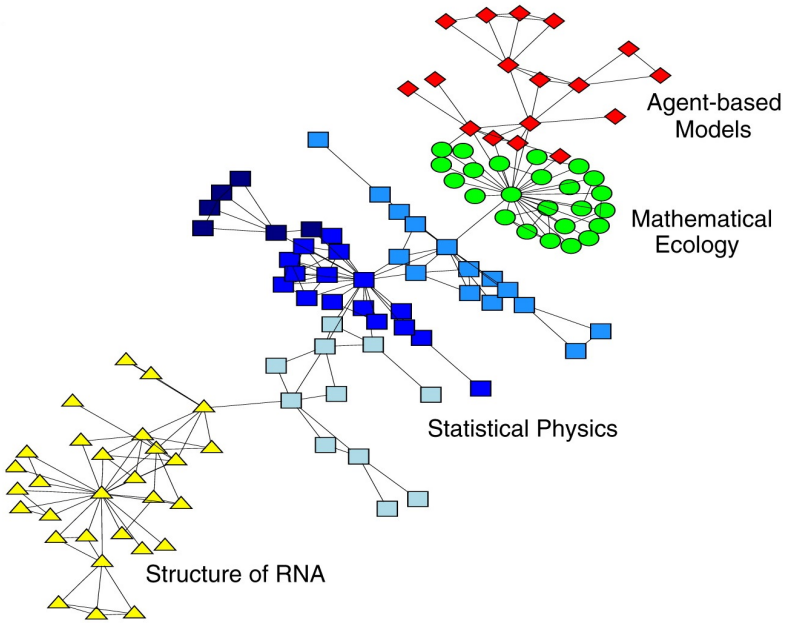
- Networks with community structures arise in many applications



Santa Fe Institute Collaboration network [Girvan-Newman '02]

Community detection in networks

- Networks with community structures arise in many applications

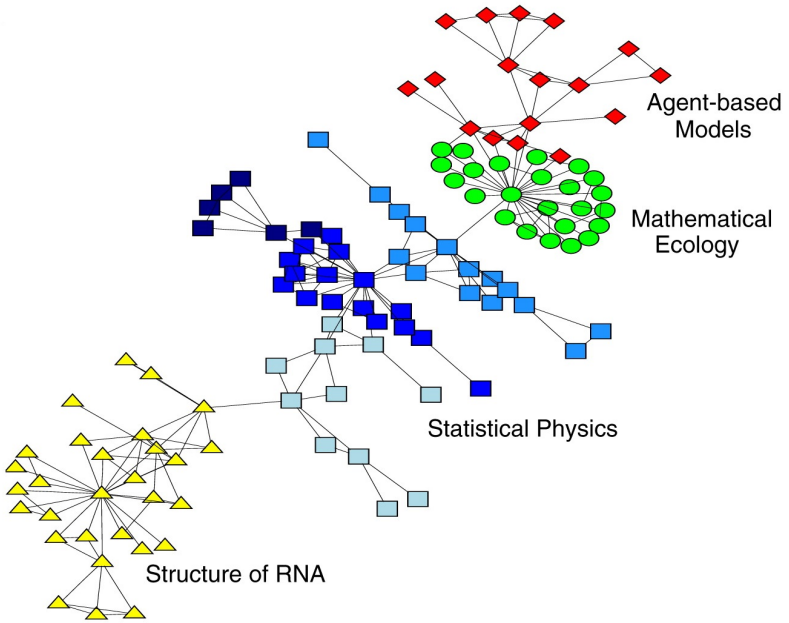


Santa Fe Institute Collaboration network [Girvan-Newman '02]

- Task: Discover underlying communities based on the network topology

Community detection in networks

- Networks with community structures arise in many applications



Santa Fe Institute Collaboration network [Girvan-Newman '02]

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

Statistical and computational challenges

- The observed network is sparse
- Large solution space

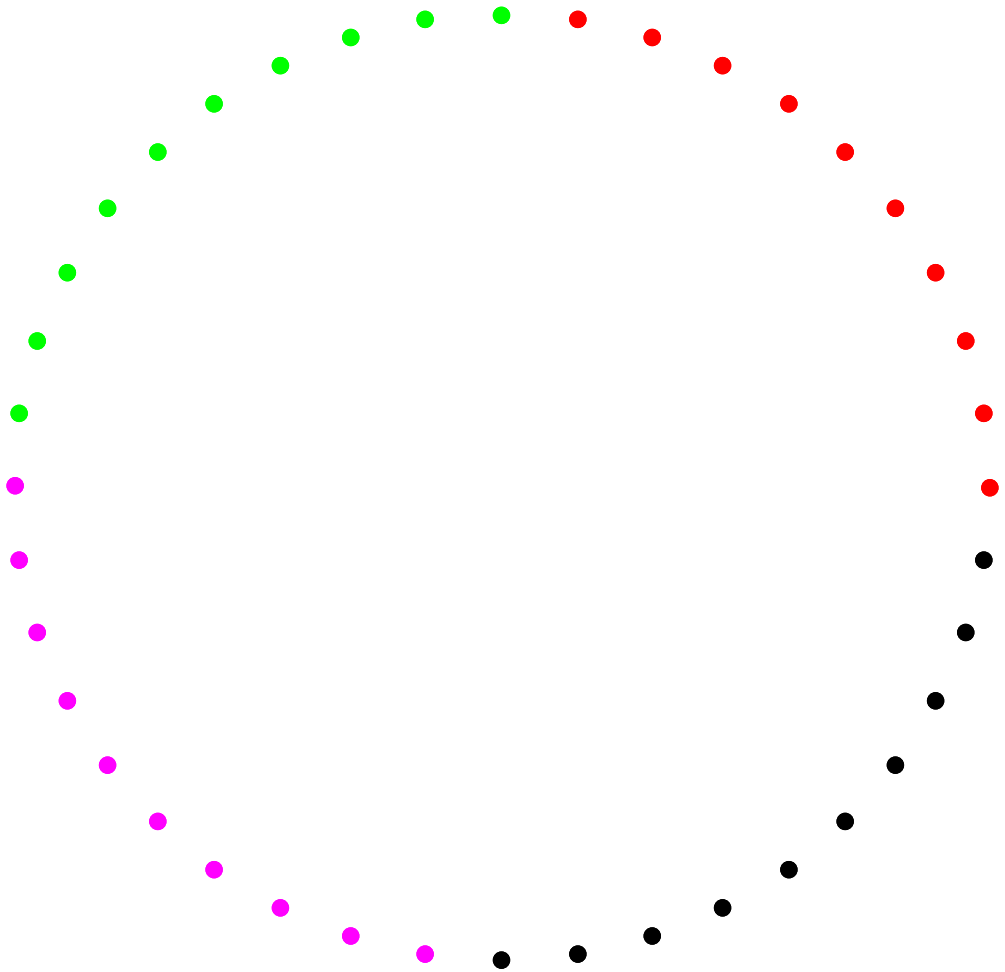
Statistical and computational challenges

- The observed network is sparse
- Large solution space

Question

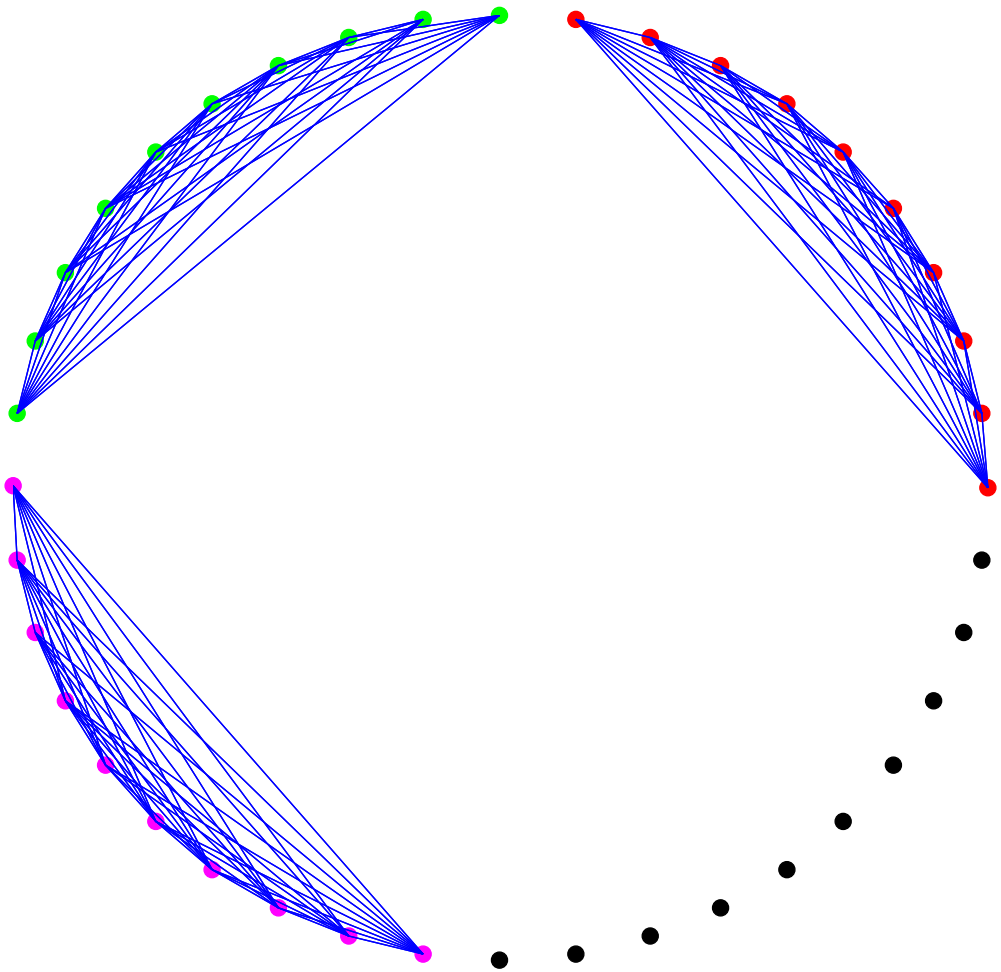
- Is there a computationally efficient and statistically optimal community detection algorithm?

Stochastic block model [Holland et al. '83]
Planted partition model [Condon-Karp 01]



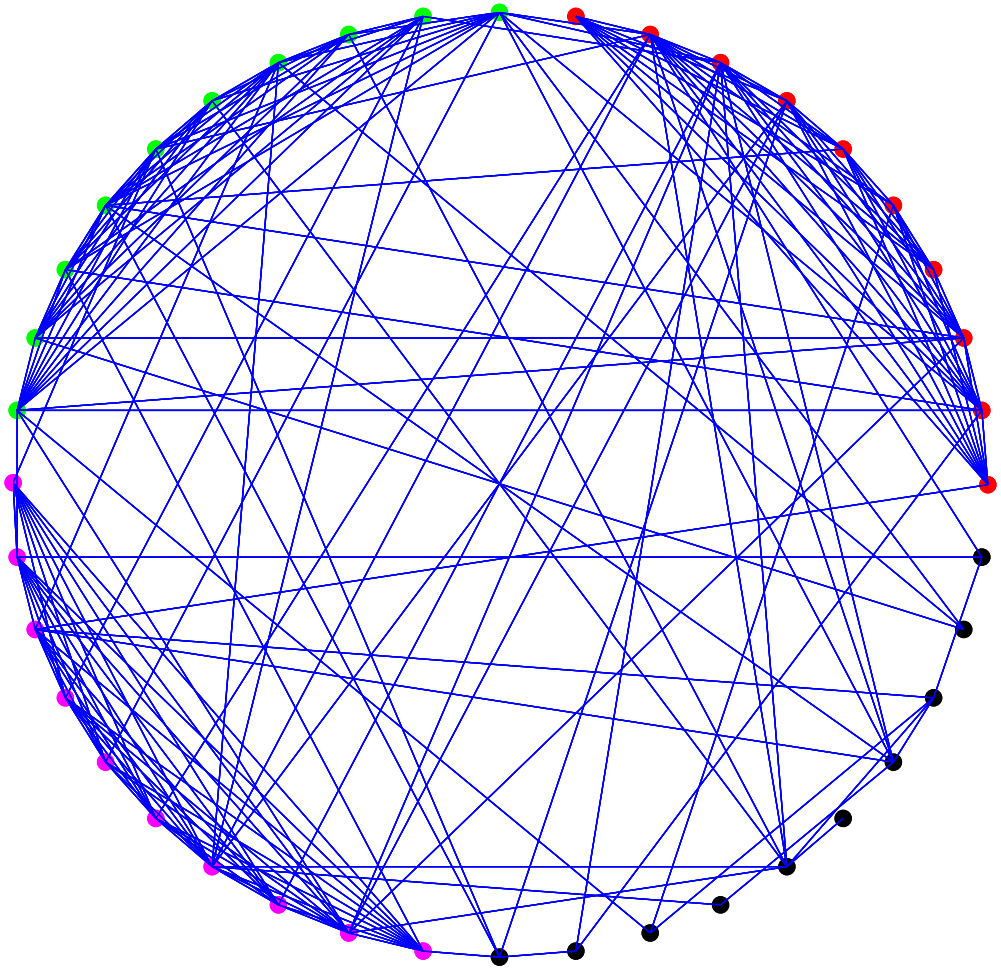
$n = 40,\ K = 10,\ r = 3$

Stochastic block model [Holland et al. '83]
Planted partition model [Condon-Karp 01]



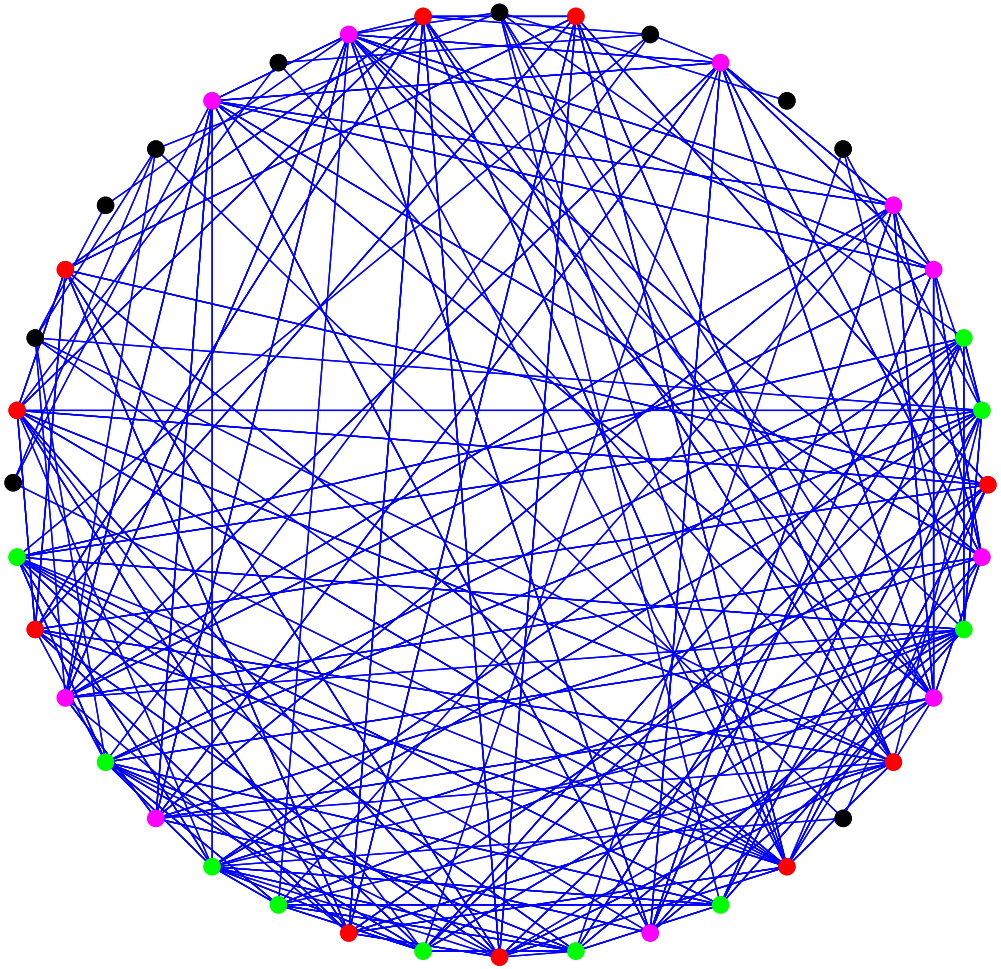
$p = 0.9$

Stochastic block model [Holland et al. '83]
Planted partition model [Condon-Karp 01]



$p = 0.9 \quad q = 0.1$

Stochastic block model [Holland et al. '83]
Planted partition model [Condon-Karp 01]



$p = 0.9 \qquad q = 0.1$

Exact recovery

- True cluster: C^*
- Estimated cluster: $\widehat{C}$
- Goal: exact recovery (strong consistency)

$$\mathbb{P}\{\widehat{C} = C^*\} \xrightarrow{n\to\infty} 1$$

Exact recovery

- True cluster: C^*
- Estimated cluster: $\widehat{C}$
- Goal: exact recovery (strong consistency)

$$\mathbb{P}\{\widehat{C} = C^*\} \xrightarrow{n\to\infty} 1$$

- Alternatives
  - almost exact recovery (weak consistency): [Mossel-Neeman-Sly '13, Abbe-Sandon '15, Montanari '15]...
  - correlated recovery: [Decelle-Krzakala-Moore-Zdeborova '11, Mossel-Neeman-Sly '12, Massoulie '13]...

Objectives of this talk

- Statistical limit: When is exact recovery possible (impossible)?
- Computational limit: When is exact recovery computationally easy (hard)?

remainder of the talk

1. Linear community size: Sharp recovery via semidefinite programming
2. Sublinear community size: Computational lower bounds

Two equal-sized communities

Binary symmetric SBM

Model:

- n nodes partitioned into two communities of size $\frac{n}{2}$ ($\sigma_i = \pm 1$).
- $i \sim j$ independently w.p. $\begin{cases} p = \frac{a \log n}{n} & \sigma_i = \sigma_j \\ q = \frac{b \log n}{n} & \sigma_i \neq \sigma_j \end{cases}$

MLE $\Leftrightarrow$ MIN BISECTION

Assuming $p > q$

- Maximum likelihood estimator (MLE)

$$\begin{aligned} \max_{\sigma} \ & \langle A, \sigma\sigma^\top \rangle \\ \text{s.t.} \ \ & \sigma_i \in \{\pm 1\}, \quad i \in [n] \\ & \sigma^\top \mathbf{1} = 0, \end{aligned}$$

MLE $\Leftrightarrow$ MIN BISECTION

Assuming $p > q$

- Maximum likelihood estimator (MLE)

$$\begin{aligned} \max_{\sigma} \ & \langle A, \sigma\sigma^\top \rangle \\ \text{s.t.} \ & \sigma_i \in \{\pm 1\}, \quad i \in [n] \\ & \sigma^\top \mathbf{1} = 0, \end{aligned}$$

$\xLeftrightarrow{\text{lift}}$

$$\begin{aligned} \max_{Y} \ & \langle A, Y \rangle \\ \text{s.t.} \ & \mathsf{rank}(Y) = 1 \\ & Y_{ii} = 1, \quad i \in [n] \\ & \langle \mathbf{J}, Y \rangle = 0. \end{aligned}$$

where $\mathbf{J}$ = all-one matrix

SDP relaxation

- Semidefinite programming (SDP) relaxation of MLE

$$\begin{aligned} \widehat{Y}_{\mathrm{SDP}} = \arg\max_{Y} \ & \langle A, Y \rangle \\ \text{s.t.} \ \ & Y \succeq 0 \\ & Y_{ii} = 1, \quad i \in [n] \\ & \langle \mathbf{J}, Y \rangle = 0. \end{aligned}$$

- similar SDP as in [Frieze-Jerrum '95] for $\mathrm{MAX\ BISECTION}$
- average-case analysis on generative model (SBM)
- focus on $\arg\max$ rather than approximating $\max$
- goal: $\mathbb{P}\left\{ \widehat{Y}_{\mathrm{SDP}} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \right\} \to 1$

Optimal recovery via SDP

Theorem (Abbe-Bandeira-Hall '14, Mossel-Neeman-Sly '14)

- *If* $(\sqrt{a}-\sqrt{b})^2 > 2$, *recovery is achievable in polynomial-time.*
- *If* $(\sqrt{a}-\sqrt{b})^2 < 2$, *recovery is impossible.*

Optimal recovery via SDP

Theorem (Abbe-Bandeira-Hall '14, Mossel-Neeman-Sly '14)

- *If* $(\sqrt{a}-\sqrt{b})^2 > 2$, *recovery is achievable in polynomial-time.*
- *If* $(\sqrt{a}-\sqrt{b})^2 < 2$, *recovery is impossible.*

Theorem (Hajek-W.-Xu '14)

SDP achieves the optimal recovery threshold $(\sqrt{a}-\sqrt{b})^2 > 2$.

Optimal recovery via SDP

Theorem (Abbe-Bandeira-Hall '14, Mossel-Neeman-Sly '14)

- *If* $(\sqrt{a}-\sqrt{b})^2 > 2$, *recovery is achievable in polynomial-time.*
- *If* $(\sqrt{a}-\sqrt{b})^2 < 2$, *recovery is impossible.*

Theorem (Hajek-W.-Xu '14)

SDP achieves the optimal recovery threshold $(\sqrt{a}-\sqrt{b})^2 > 2$.

Remarks

- originally conjectured in [Abbe-Bandeira-Hall '14]
- independently proved by [Bandeira '15]
- $\mathbb{P}\left\{\widehat{Y}_{\mathrm{SDP}} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\right\} = 1 - n^{-\Omega(1)}$

Dual certificate

$$\begin{aligned} \max_{Y} \ & \langle A, Y \rangle \\ \text{s.t.} \ & Y \succeq 0 \\ & Y_{ii} = 1 \\ & \langle \mathbf{J}, Y \rangle = 0 \end{aligned}$$

Dual certificate

$$\max_{Y} \langle A, Y \rangle \qquad \text{dual variables}$$

$$\text{s.t.} \quad Y \succeq 0 \qquad S \succeq 0$$

$$Y_{ii} = 1 \qquad D = \text{diag}\{d_i\}$$

$$\langle \mathbf{J}, Y \rangle = 0 \qquad \lambda \in \mathbb{R}$$

Lemma

$Y^* = \sigma^*(\sigma^*)^\top$ is unique solution if $\exists D, \lambda$ s.t. $S = \lambda \mathbf{J} + D - A$ satisfies

$$S\sigma = 0 \quad \textit{and} \quad \lambda_2(S) > 0.$$

Dual certificate

$$\max_{Y} \langle A, Y \rangle \qquad \text{dual variables}$$

$$\text{s.t.} \quad Y \succeq 0 \qquad S \succeq 0$$

$$Y_{ii} = 1 \qquad D = \text{diag}\{d_i\}$$

$$\langle \mathbf{J}, Y \rangle = 0 \qquad \lambda \in \mathbb{R}$$

Lemma

$Y^* = \sigma^*(\sigma^*)^\top$ is unique solution if $\exists D, \lambda$ s.t. $S = \lambda \mathbf{J} + D - A$ satisfies

$$S\sigma = 0 \quad \textit{and} \quad \lambda_2(S) > 0.$$

$$\Rightarrow d_i = (\#\text{ of nbrs in own cluster}) - (\#\text{ of nbrs in other cluster})$$

$$= \begin{cases} e(i, C_1) - e(i, C_2) & i \in C_1 \\ e(i, C_2) - e(i, C_1) & i \in C_2 \end{cases}$$

Verify PSD

- Mean adj matrix: $\mathbb{E}[A] = \frac{p+q}{2}\mathbf{J} + \frac{p-q}{2}\sigma^*(\sigma^*)^\top - p\mathbf{I}$

Verify PSD

- Mean adj matrix: $\mathbb{E}[A] = \frac{p+q}{2}\mathbf{J} + \frac{p-q}{2}\sigma^*(\sigma^*)^\top - p\mathbf{I}$
-

$$\begin{aligned} S &= \lambda\mathbf{J} - A + D \\ &= \underbrace{\left(\lambda - \frac{p+q}{2}\right)\mathbf{J}} - \frac{p-q}{2}\sigma^*(\sigma^*)^\top + p\mathbf{I} + \underbrace{D - (A - \mathbb{E}[A])} \end{aligned}$$

Verify PSD

- Mean adj matrix: $\mathbb{E}[A] = \frac{p+q}{2}\mathbf{J} + \frac{p-q}{2}\sigma^*(\sigma^*)^\top - p\mathbf{I}$
-

$$\begin{aligned} S &= \lambda\mathbf{J} - A + D \\ &= \underbrace{\left(\lambda - \frac{p+q}{2}\right)\mathbf{J}} - \frac{p-q}{2}\sigma^*(\sigma^*)^\top + p\mathbf{I} + \underbrace{D - (A - \mathbb{E}[A])} \end{aligned}$$

- $\lambda_2(S) = \inf_{x\perp\sigma^*, \|x\|_2=1} x^\top S x > 0$ if $\min d_i \geq \|A - \mathbb{E}[A]\|$ and $\lambda \geq (p+q)/2$

Verify PSD

- Mean adj matrix: $\mathbb{E}[A] = \frac{p+q}{2}\mathbf{J} + \frac{p-q}{2}\sigma^*(\sigma^*)^\top - p\mathbf{I}$
-

$$\begin{aligned} S &= \lambda\mathbf{J} - A + D \\ &= \underbrace{\left(\lambda - \frac{p+q}{2}\right)\mathbf{J}} - \frac{p-q}{2}\sigma^*(\sigma^*)^\top + p\mathbf{I} + \underbrace{D - (A - \mathbb{E}[A])} \end{aligned}$$

- $\lambda_2(S) = \inf_{x\perp\sigma^*, \|x\|_2=1} x^\top S x > 0$ if $\min d_i \geq \|A - \mathbb{E}[A]\|$ and $\lambda \geq (p+q)/2$
- To finish the proof:

Verify PSD

- Mean adj matrix: $\mathbb{E}[A] = \frac{p+q}{2}\mathbf{J} + \frac{p-q}{2}\sigma^*(\sigma^*)^\top - p\mathbf{I}$
-

$$\begin{aligned} S &= \lambda\mathbf{J} - A + D \\ &= \underbrace{\left(\lambda - \frac{p+q}{2}\right)\mathbf{J}} - \frac{p-q}{2}\sigma^*(\sigma^*)^\top + p\mathbf{I} + \underbrace{D - (A - \mathbb{E}[A])} \end{aligned}$$

- $\lambda_2(S) = \inf_{x\perp\sigma^*, \|x\|_2=1} x^\top S x > 0$ if $\min d_i \geq \|A - \mathbb{E}[A]\|$ and $\lambda \geq (p+q)/2$
- To finish the proof:
  1. $\min d_i = \Omega_P(\log n)$ if $\sqrt{a} - \sqrt{b} > \sqrt{2}$

Verify PSD

- Mean adj matrix: $\mathbb{E}[A] = \frac{p+q}{2}\mathbf{J} + \frac{p-q}{2}\sigma^*(\sigma^*)^\top - p\mathbf{I}$
-

$$\begin{aligned} S &= \lambda\mathbf{J} - A + D \\ &= \underbrace{\left(\lambda - \frac{p+q}{2}\right)\mathbf{J}} - \frac{p-q}{2}\sigma^*(\sigma^*)^\top + p\mathbf{I} + \underbrace{D - (A - \mathbb{E}[A])} \end{aligned}$$

- $\lambda_2(S) = \inf_{x\perp\sigma^*, \|x\|_2=1} x^\top S x > 0$ if $\min d_i \geq \|A - \mathbb{E}[A]\|$ and $\lambda \geq (p+q)/2$
- To finish the proof:
  1. $\min d_i = \Omega_P(\log n)$ if $\sqrt{a} - \sqrt{b} > \sqrt{2}$
  2. $\|A - \mathbb{E}[A]\| = O_P(\sqrt{\log n})$

Remarks

1. Necessity

$$\sqrt{a} - \sqrt{b} < \sqrt{2}$$

$\Rightarrow \min d_i < 0$ w.h.p.

$\Rightarrow \exists i$: # of nbrs in own cluster $<$ # of nbrs in other cluster

$\Rightarrow$ MLE fails

Remarks

1. Necessity

$$\sqrt{a} - \sqrt{b} < \sqrt{2}$$

$\Rightarrow \min d_i < 0$ w.h.p.

$\Rightarrow \exists i$: # of nbrs in own cluster $<$ # of nbrs in other cluster

$\Rightarrow$ MLE fails

2. Proof of $\|A - \mathbb{E}[A]\| = O_P(\sqrt{\log n})$
   - 2nd-order stochastic dominance argument [Tomozei-Massoulié '14] + result for iid matrix [Seginer '00]
   - [Feige-Ofek '05]: $\mathcal{G}(n, \frac{C \log n}{n})$ for sufficiently large C
   - [Bandeira-van Handel '14]: comparison argument

Multiple equal-sized communities

r equal-sized clusters

- $\{0,1\}$-cluster matrix:

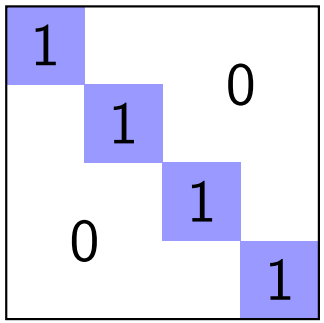
$$Y^* = \sum_{k=1}^{r} \xi_k (\xi_k)^\top \quad = \quad \begin{bmatrix} 1 & & & 0 \\ & 1 & & \\ & & 1 & \\ 0 & & & 1 \end{bmatrix}$$

where $\xi_k =$ indicator of the k^{th} cluster of size $K = n/r$

r equal-sized clusters

- $\{0,1\}$-cluster matrix:

$$Y^* = \sum_{k=1}^{r} \xi_k (\xi_k)^\top \quad = \quad$$



where $\xi_k =$ indicator of the k^{th} cluster of size $K = n/r$

- SDP relaxation of MLE:

$$\begin{aligned} \max_{Y} \ & \langle A, Y \rangle \\ \text{s.t.} \ & Y \succeq 0 \\ & Y_{ii} = 1 \\ & Y_{ij} \geq 0 \\ & \sum_j Y_{ij} = K \end{aligned}$$

Optimality of SDP

Theorem ([Hajek-W.-Xu '15])

SDP achieves optimal threshold $(\sqrt{a} - \sqrt{b})^2 > r$.

Optimality of SDP

Theorem ([Hajek-W.-Xu '15])

SDP achieves optimal threshold $(\sqrt{a}-\sqrt{b})^2 > r$.

Proof of correctness:

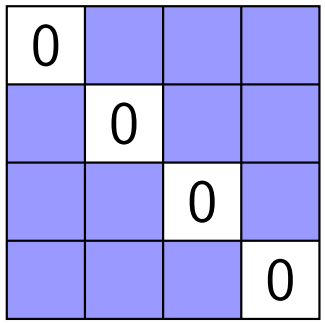
$$\begin{aligned} \max_{Y} \ & \langle A, Y \rangle \\ \text{s.t.} \ & Y \succeq 0 \quad S \succeq 0 \\ & Y_{ii} = 1 \quad d_i \\ & Y_{ij} \geq 0 \quad B \geq 0 \\ & \sum_j Y_{ij} = K \quad \lambda_i \end{aligned}$$

Construction of the dual witness

- For node $i \in C_k$,

$$\lambda_i = \frac{1}{K}\Big(\max_{\ell \neq k} e(i, C_\ell) - Kq/2 + \sqrt{\log n}/2 \Big)$$

$$d_i = e(i, C_k) - \max_{\ell \neq k} e(i, C_\ell) - \frac{1}{K} \sum_{j \in C_k} \max_{\ell \neq k} e(j, C_\ell) + Kq - \sqrt{\log n}$$

- $B =$ 

, where each ■ is rank-2, specified by

$$B_{C_k \times C_{k'}}(i,j) = \frac{1}{K}\Big(\max_{\ell \neq k} e(i, C_{k'}) - e(i, C_{k'}) + \max_{\ell \neq k'} e(j, C_k) - e(j, C_k) + \frac{e(C_k, C_{k'})}{K} - Kq + \sqrt{\log n} \Big)$$

- $S = D - A - B + \lambda \mathbf{1}^\top + \mathbf{1}\lambda^\top$

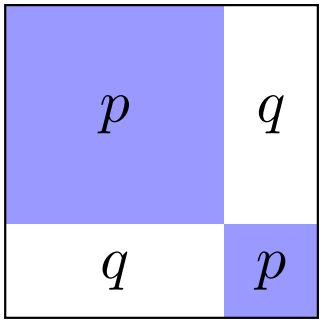
Construction of the dual witness

$$\begin{aligned}
\max_{Y} \ & \langle A, Y \rangle \\
\text{s.t.} \ & Y \succeq 0 \quad S \succeq 0 \\
& Y_{ii} = 1 \quad d_i \\
& Y_{ij} \geq 0 \quad B \geq 0 \\
& \sum_j Y_{ij} = K \quad \lambda_i
\end{aligned}$$

- $S\xi_k = 0$ for $k = 1, \ldots, r$.
- $\lambda_{r+1}(S) > 0$ if $\min d_i \geq \|A - \mathbb{E}[A]\| = O_P(\sqrt{\log n})$
- $d_i =$ (# of nbrs in own cluster)
  $-$maximal (# of nbrs in other clusters) $+ O_P(\sqrt{\log n})$.
- Sharp threshold
  - $\sqrt{a} - \sqrt{b} > \sqrt{r} \Rightarrow \min d_i = \Omega(\log n) \Rightarrow$ SDP succeeds
  - $\sqrt{a} - \sqrt{b} < \sqrt{r} \Rightarrow \min d_i = -\Omega(\log n) \Rightarrow$ MLE fails

Unequal-sized clusters

Two unequal-sized clusters: known size

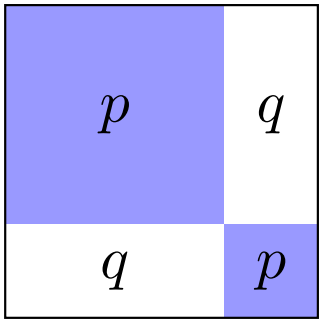


Two clusters of size K and $n-K$ ($K=\rho n$):

$$\begin{aligned}\widehat{Y}_{\mathrm{SDP}} = \arg\max_{Y} \ & \langle A, Y\rangle \\ \text{s.t.} \ & Y \succeq 0 \\ & Y_{ii} = 1, \quad i \in [n] \\ & \langle \mathbf{J}, Y\rangle = (2K-n)^2\end{aligned}$$

achieves optimal threshold $\eta(\rho, a, b) > 1$.

Two unequal-sized clusters: known size



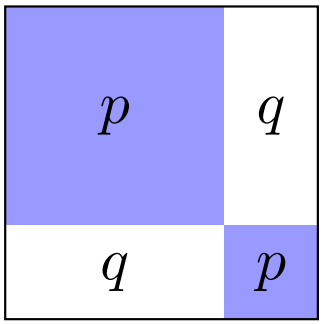
Two clusters of size K and $n-K$ ($K=\rho n$):

$$\widehat{Y}_{\mathrm{SDP}} = \arg\max_{Y} \langle A, Y\rangle$$
$$\text{s.t.} \quad Y \succeq 0$$
$$Y_{ii} = 1, \quad i \in [n]$$
$$\langle \mathbf{J}, Y\rangle = (2K-n)^2$$

achieves optimal threshold $\eta(\rho, a, b) > 1$.

Note: $\rho \mapsto \eta(\rho, a, b)$ is minimized at $\eta(1/2, a, b) = \frac{1}{2}(\sqrt{a}-\sqrt{b})^2 \Rightarrow$ "suggests" equal-sized case is the hardest for two communities

Two unequal-sized clusters: unknown size

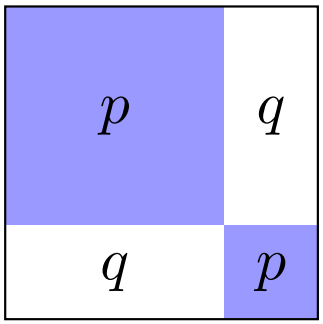


Two clusters of size K and $n-K$ $(K=0,1,\ldots,n)$:

$$\widehat{Y}_{\mathrm{SDP}} = \arg\max_{Y} \langle A, Y\rangle - \lambda\langle \mathbf{J}, Y\rangle$$
$$\text{s.t.} \quad Y \succeq 0$$
$$Y_{ii} = 1, \quad i \in [n]$$

with $\lambda = \frac{a-b}{\log a - \log b}\frac{\log n}{n}$ achieves optimal threshold $(\sqrt{a}-\sqrt{b})^2 > 2$.

Two unequal-sized clusters: unknown size



Two clusters of size K and $n-K$ ($K=0,1,\ldots,n$):

$$\widehat{Y}_{\mathrm{SDP}} = \arg\max_{Y} \langle A, Y\rangle - \lambda\langle \mathbf{J}, Y\rangle$$
$$\text{s.t.} \quad Y \succeq 0$$
$$Y_{ii} = 1, \quad i \in [n]$$

with $\lambda = \frac{a-b}{\log a - \log b}\frac{\log n}{n}$ achieves optimal threshold $(\sqrt{a}-\sqrt{b})^2 > 2$.

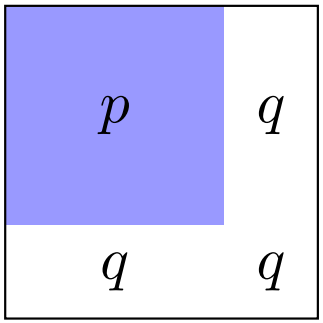
Note: If $K = \Omega(n)$, there exists a data-driven choice of λ.

More generally...

- Binary censored block model: $\mathcal{G}(n, \frac{a \log n}{n})$ observe edge label flipped w.p. ϵ
  - SDP achieves sharp threshold $a\ (\sqrt{1-\epsilon} - \sqrt{\epsilon})^2 > 1$
  - Closes the gap in [Abbe-Bandeira-Bracher-Singer '14]
- General SBM:
  - Optimality of SDP relaxation remains open (but within a factor of 4)
  - Sharp threshold is found in [Abbe-Sandon '15] via a two-stage procedure.

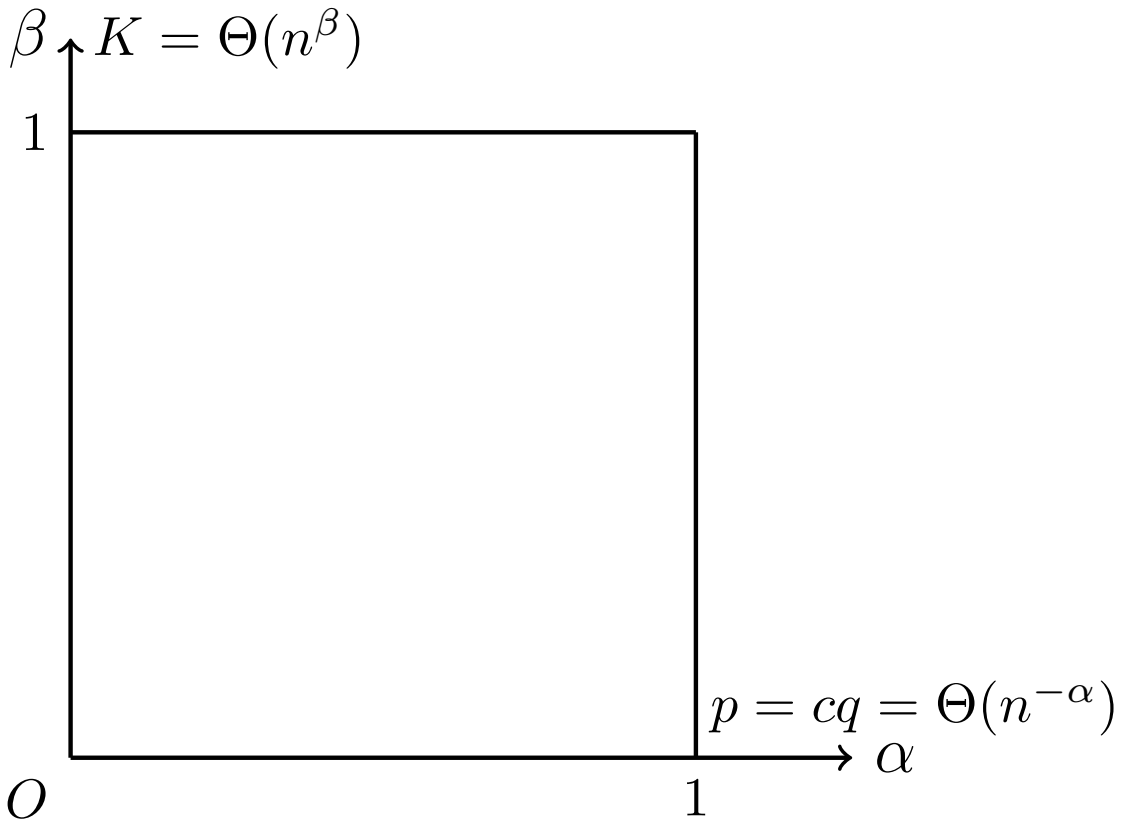
Detecting a single cluster

Finding a single community

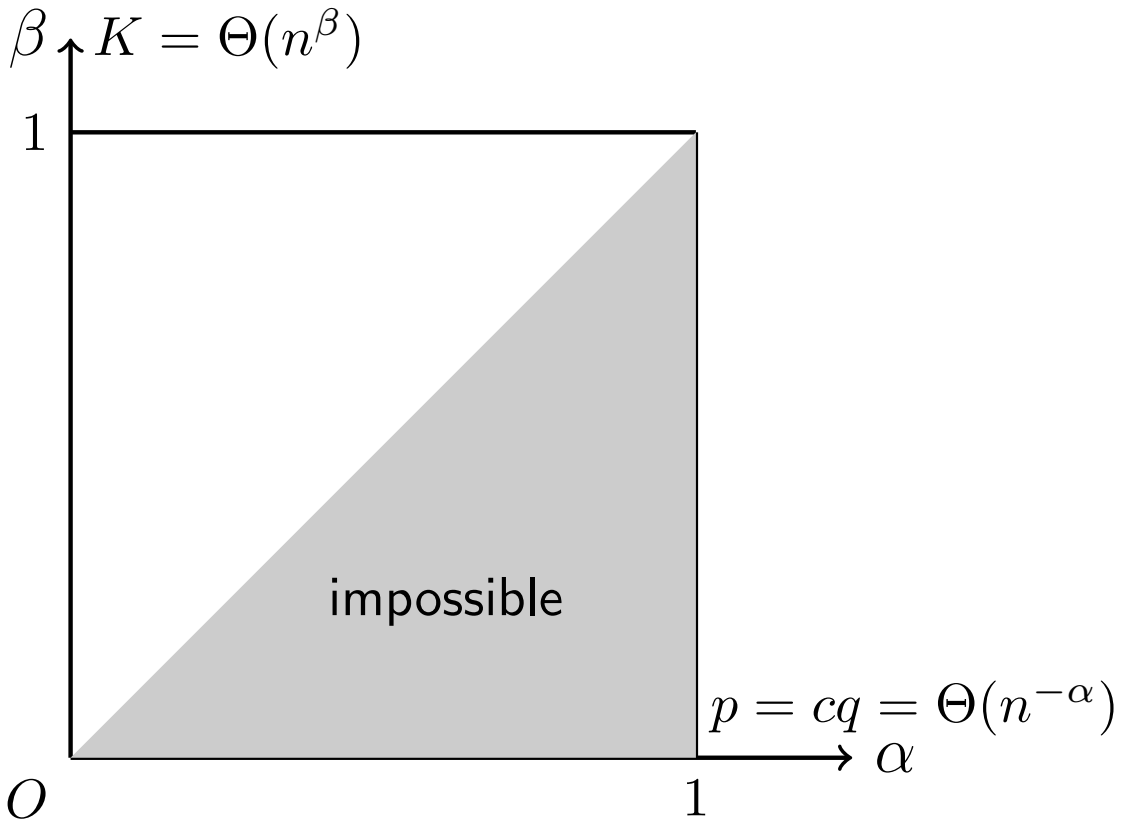


- One cluster of size K plus $n - K$ outliers
- Connectivity p within cluster and q otherwise
- Also known as Planted Dense Subgraph model
- Linear community size: $K = \rho n$ and SDP achieves sharp threshold
- Next focus on $K = \Theta(n^\beta)$.

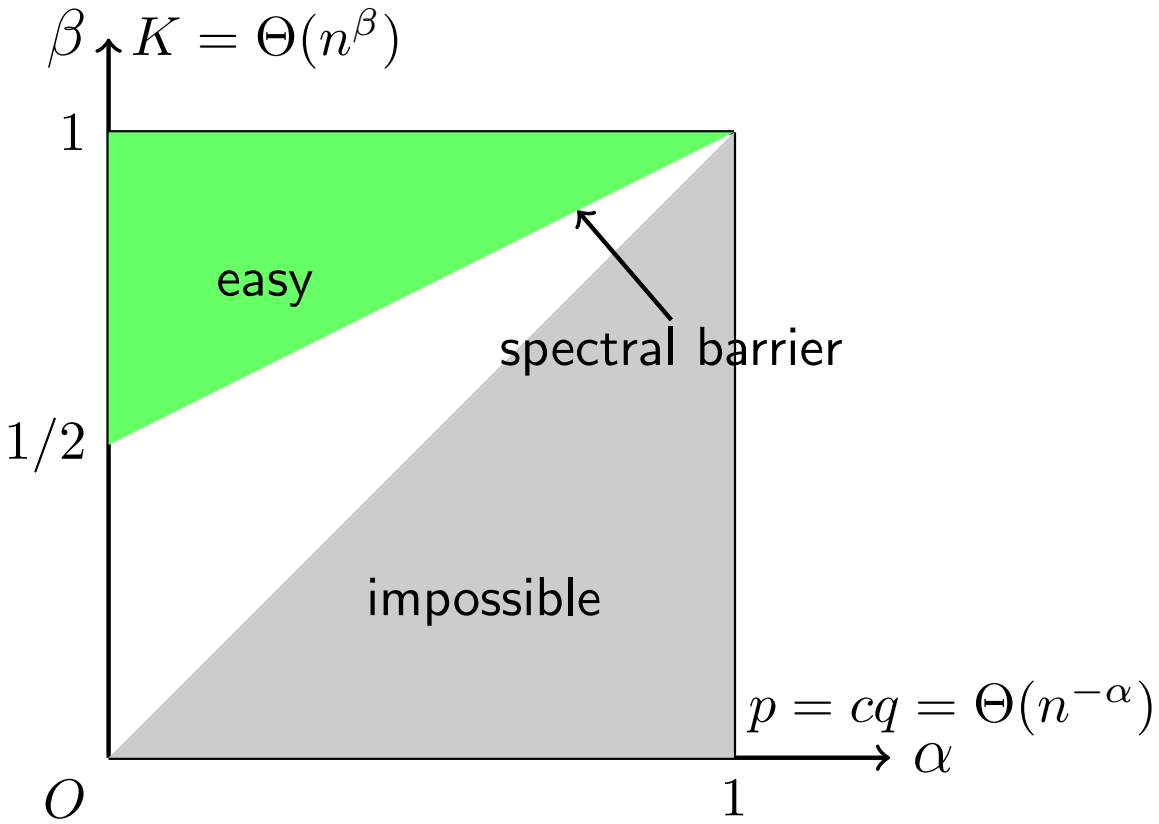
Conjecture on computational limit



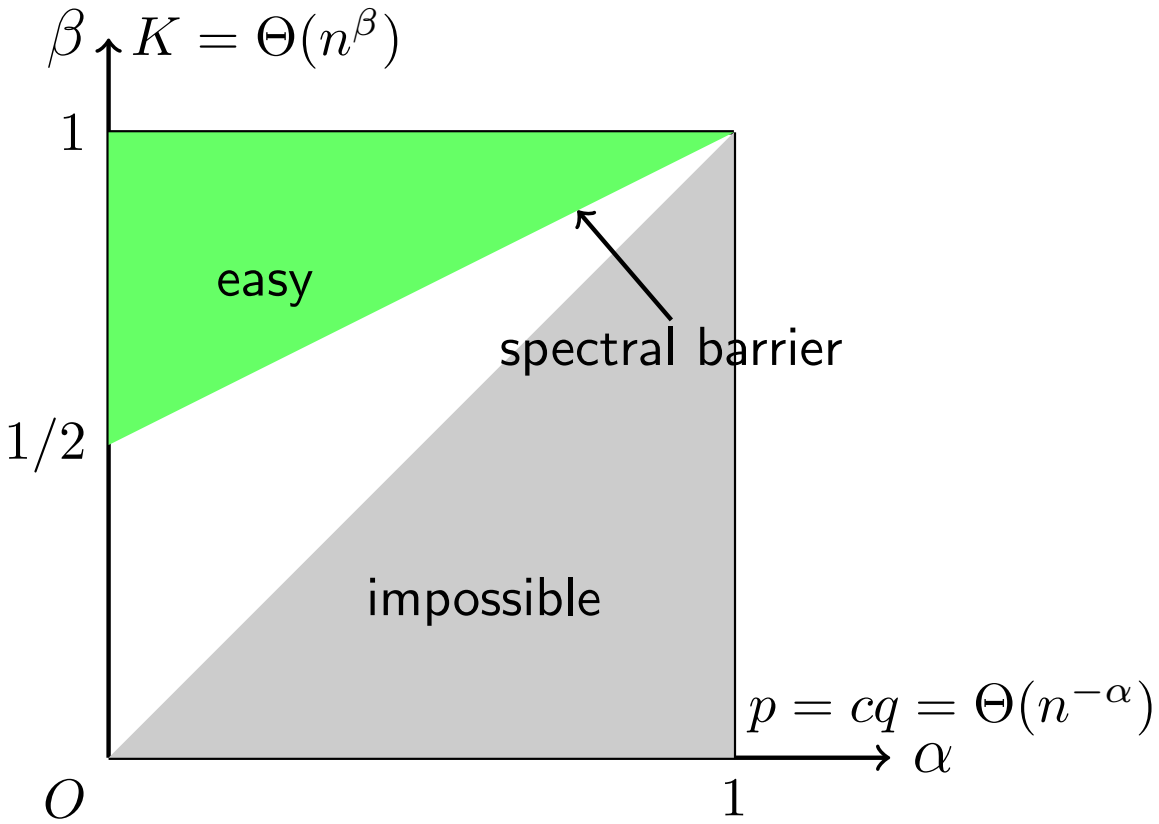
Conjecture on computational limit



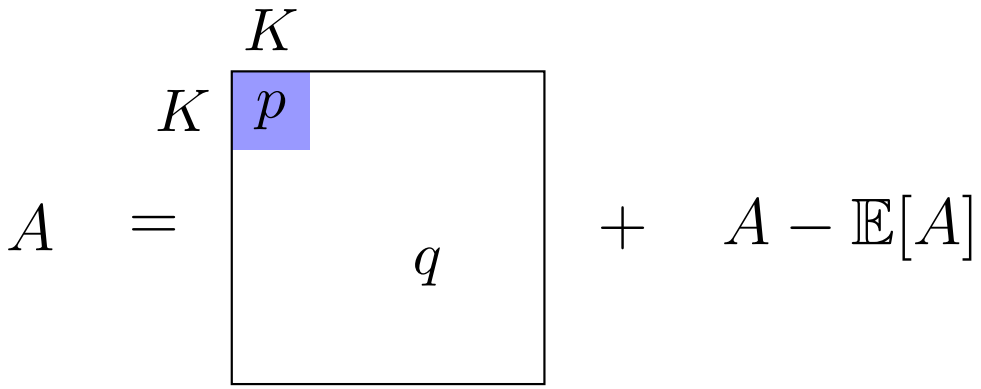
Conjecture on computational limit



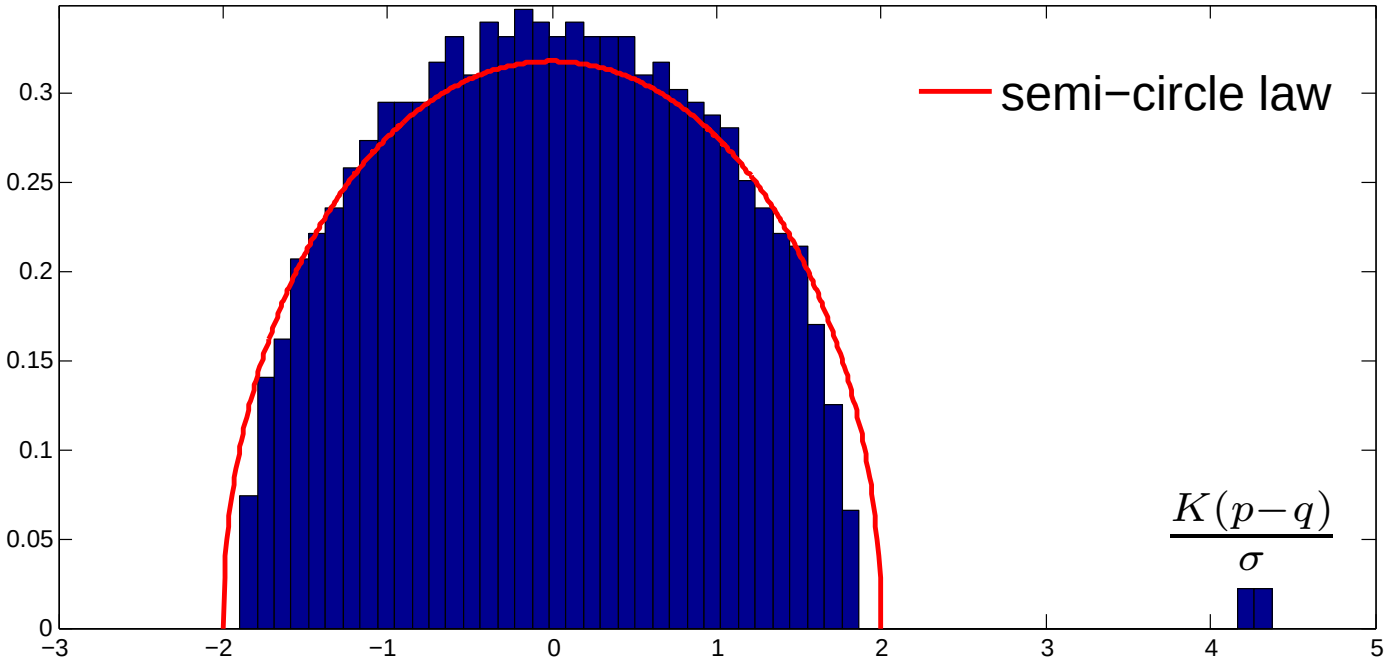
Conjecture on computational limit



Conjecture [Chen-Xu '14]: no polynomial-time algorithm succeeds beyond the spectral barrier [Nadakuditi-Newman '12]

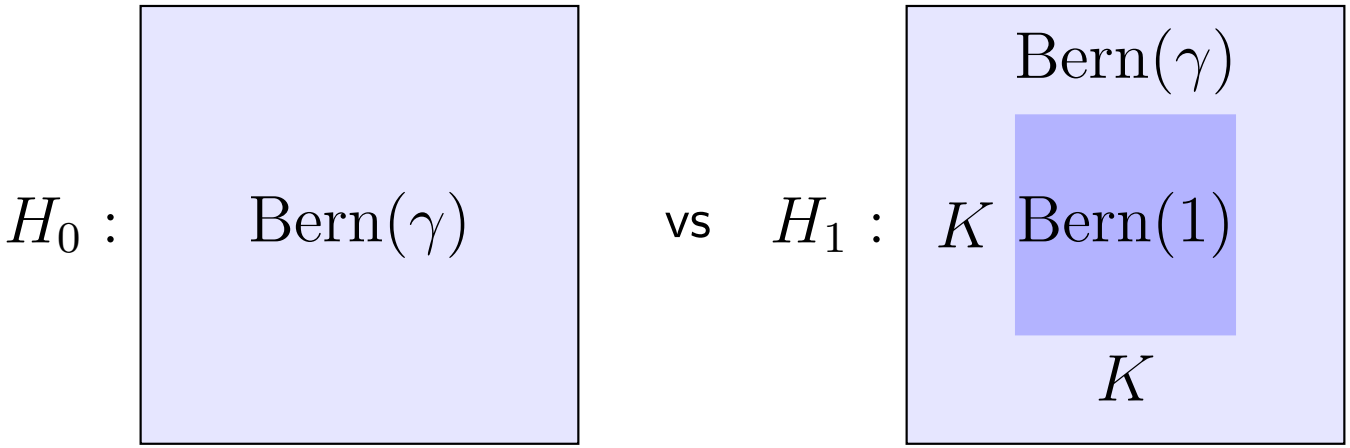

K
K
p
A
=
q
+
A − 𝔼[A]

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



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

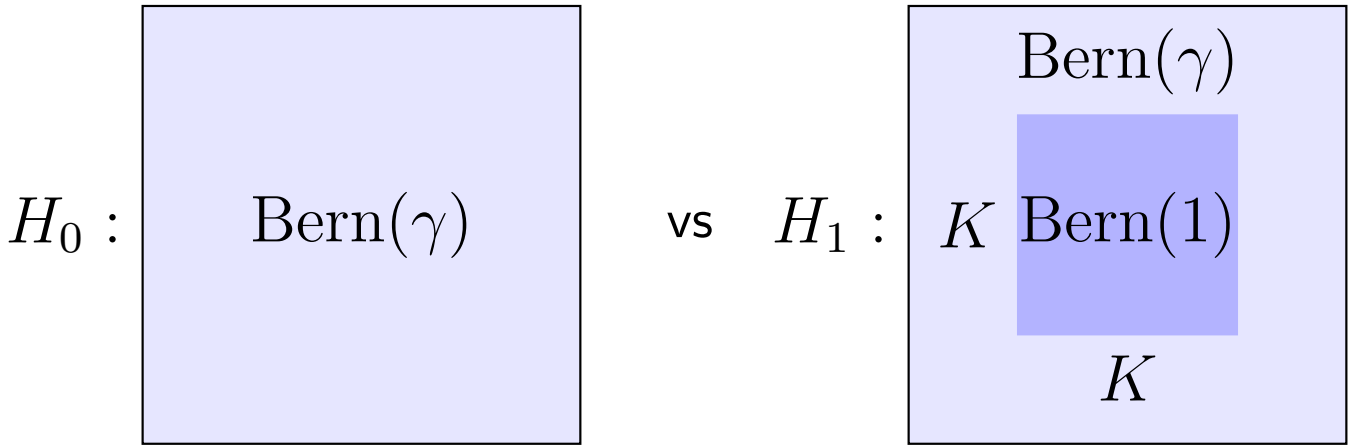
Planted clique hardness hypothesis



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] [Deshpande-Montanari '15] [Meka-Potechin-Wigderson '15]

Planted clique hardness hypothesis

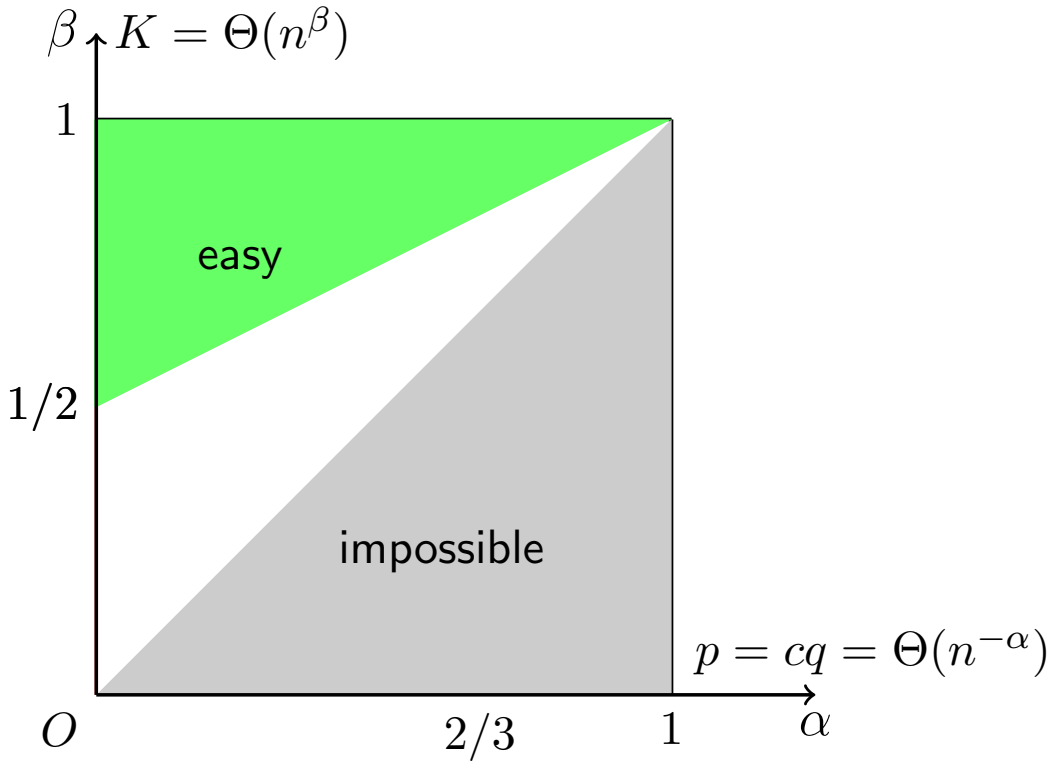


Intermediate regime: $\log n \ll K \ll \sqrt{n}$, $\gamma = \Theta(1)$

- detection is possible but believed to have high computational complexity: [Alon et al. '11] [Feldman et al. '13] [Deshpande-Montanari '15] [Meka-Potechin-Wigderson '15]
- various hardness results assuming Planted Clique hardness
  - detecting sparse principal component [Berthet-Rigollet '13]: $\gamma = \frac{1}{2}$
  - detecting sparse submatrix [Ma-W. '13, Cai-Liang-Rakhlin '15]: $\gamma = \frac{1}{2}$
  - cryptography [Applebaum-Barak-Wigderson '10]: $\gamma = 2^{-\log^{0.99} n}$

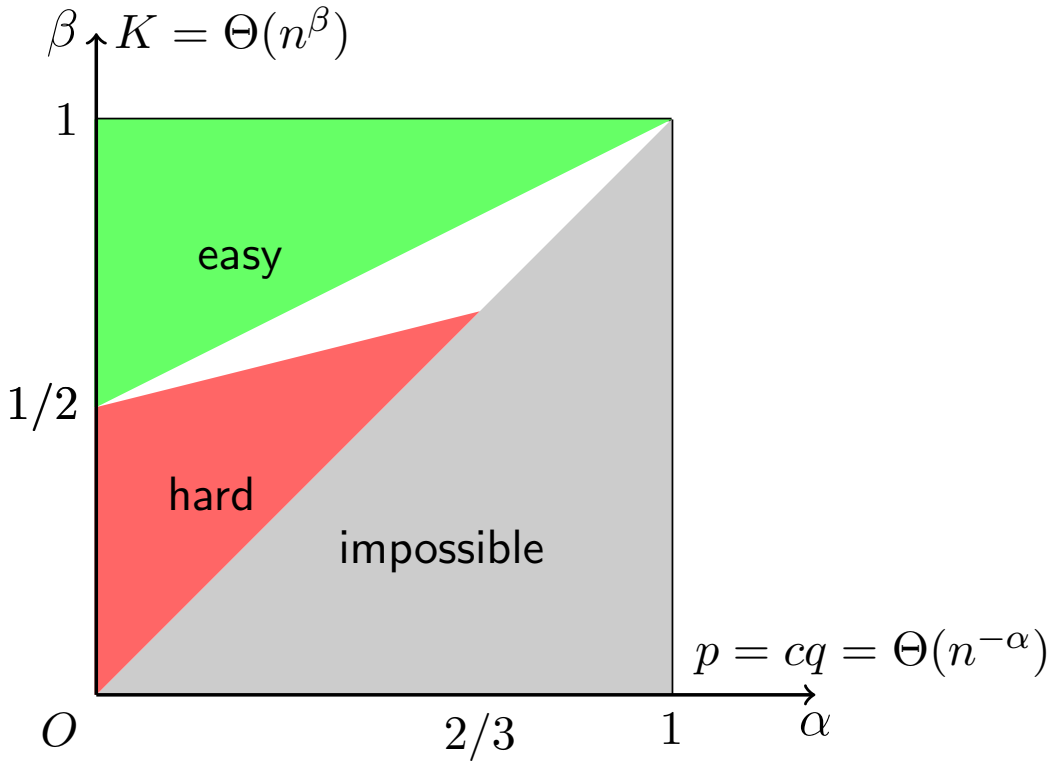
Hard regime for recovering a single cluster

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



Hard regime for recovering a single cluster

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



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

Proof step 1: Recovery is harder than *detection*

Recovery versus Detection [Arias-Castro-Verzelen '14] :

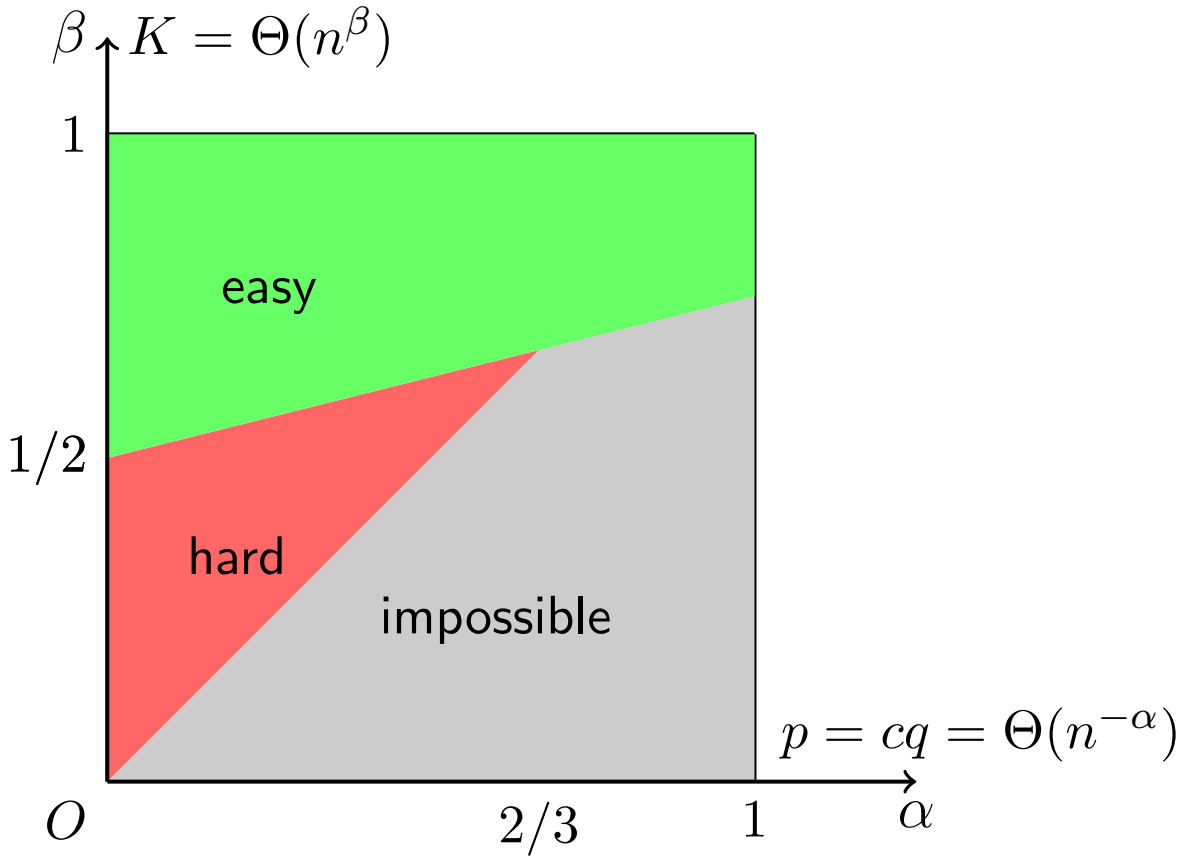
Proof step 1: Recovery is harder than *detection*

Recovery versus Detection [Arias-Castro-Verzelen '14] :

H_0 : $\mathrm{Bern}(q)$ vs H_1 : $\mathrm{Bern}(q)$, S $\mathrm{Bern}(p)$, S

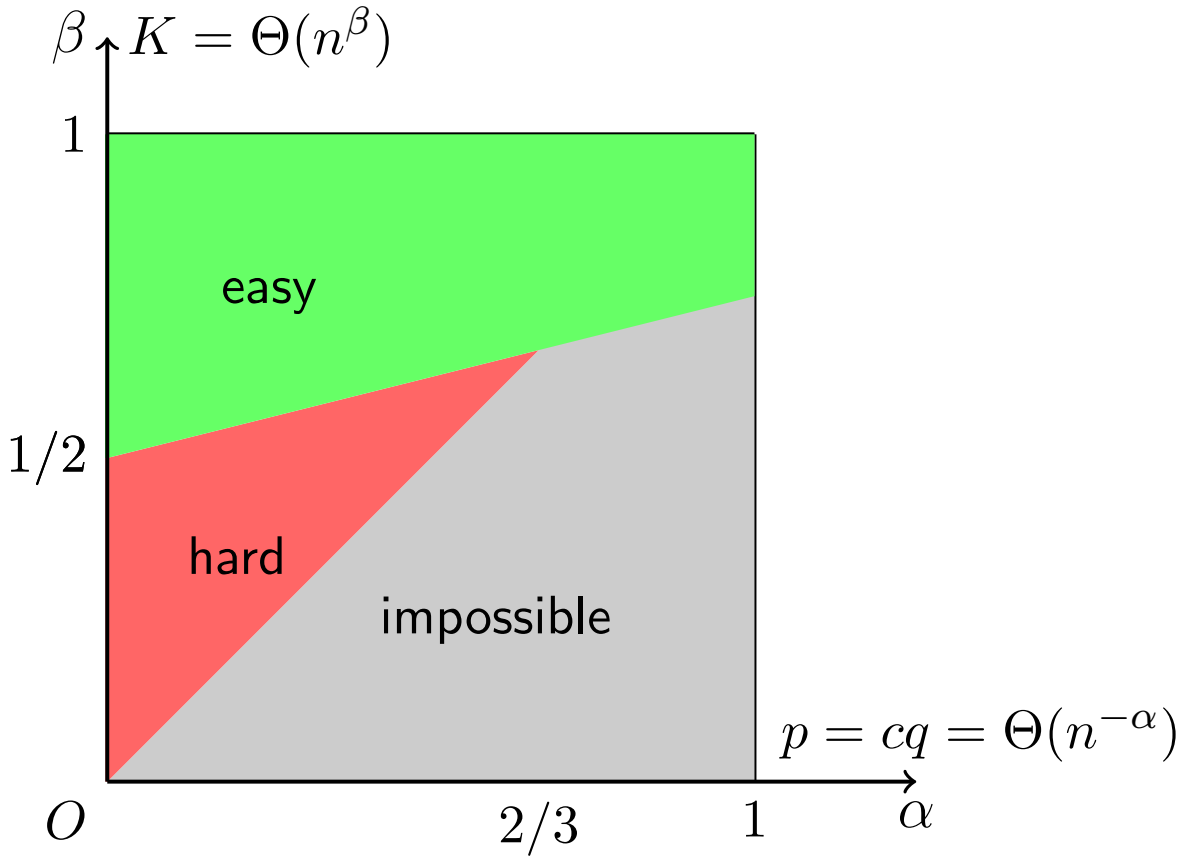
Each node is included in S with probability $\frac{K}{n}$

Proof step 2: Hardness for *detecting* a single cluster

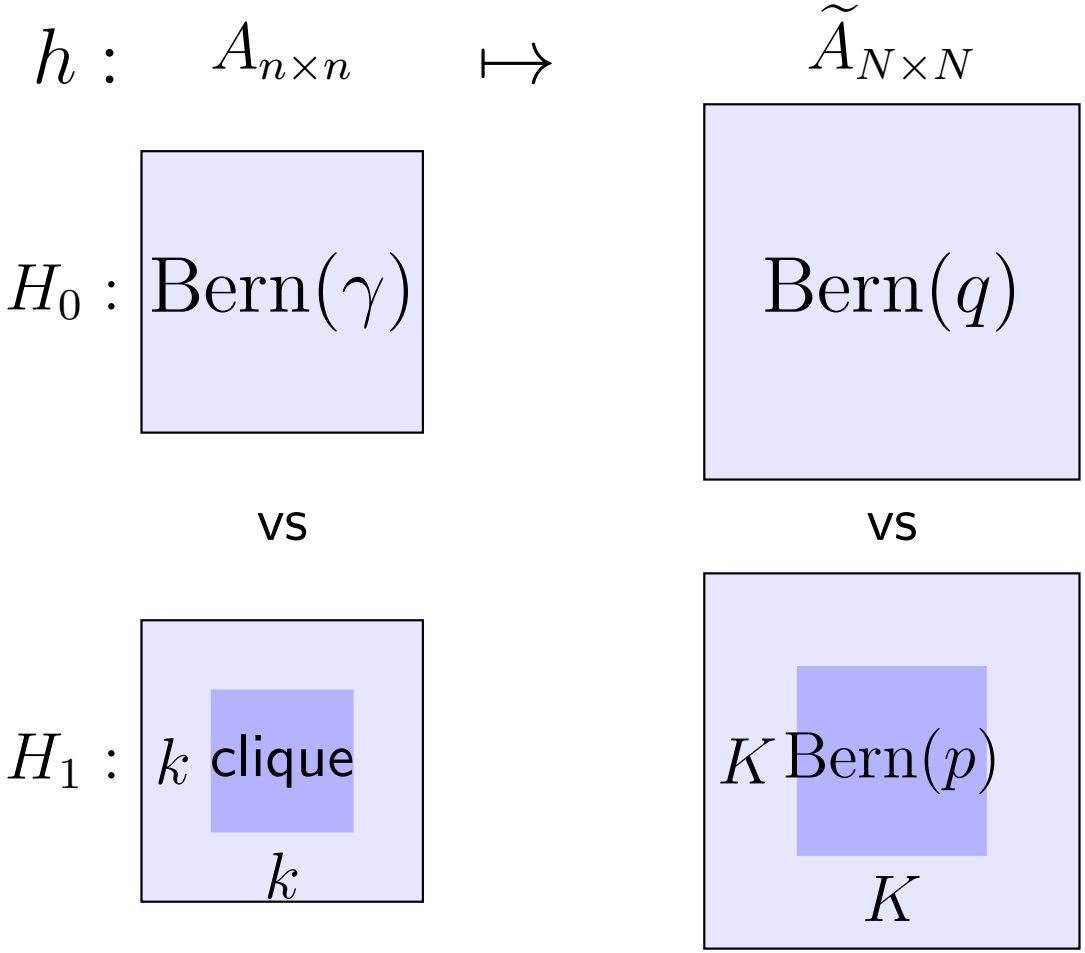


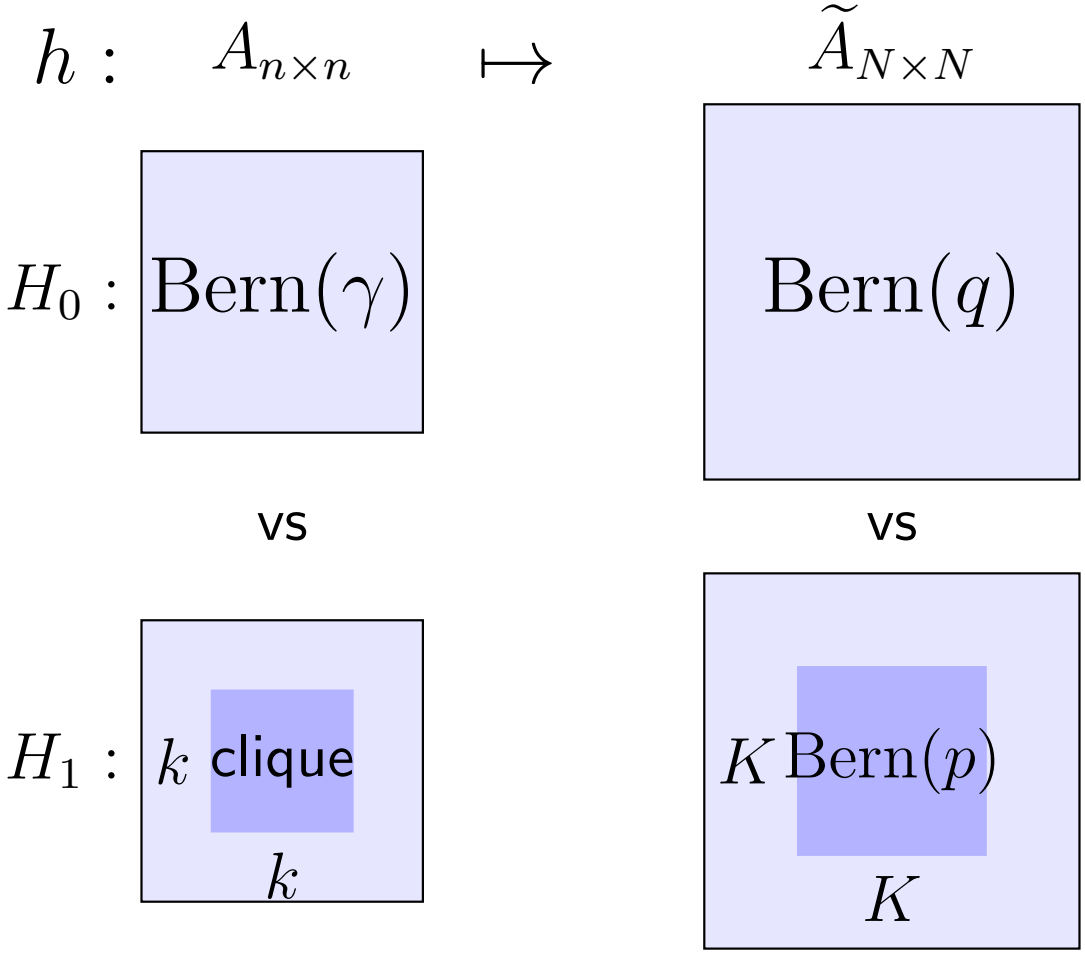
- Detecting a single cluster in the red regime is at least as hard as detecting a clique of size $K = o(\sqrt{n})$

Proof step 2: Hardness for *detecting* a single cluster



- Detecting a single cluster in the red regime is at least as hard as detecting a clique of size $K = o(\sqrt{n})$
- Reduced from Planted Clique detection in polynomial time

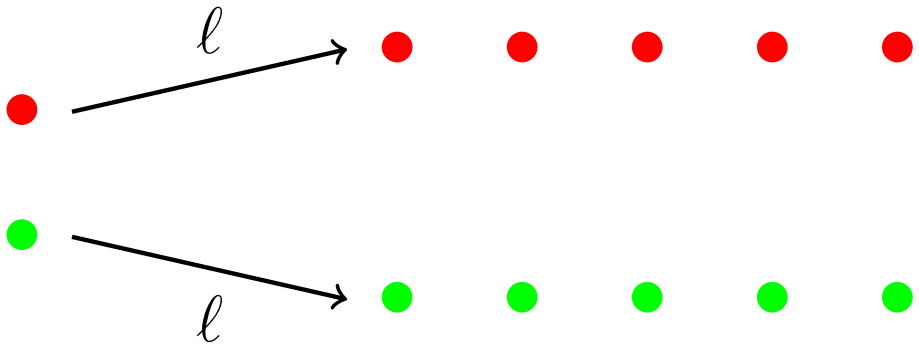

$h:$
$A_{n\times n}$
$\mapsto$
$\widetilde{A}_{N\times N}$
$H_0:$
$\mathrm{Bern}(\gamma)$
$\mathrm{Bern}(q)$
vs
vs
$H_1:$
k clique
k
K $\mathrm{Bern}(p)$
K



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

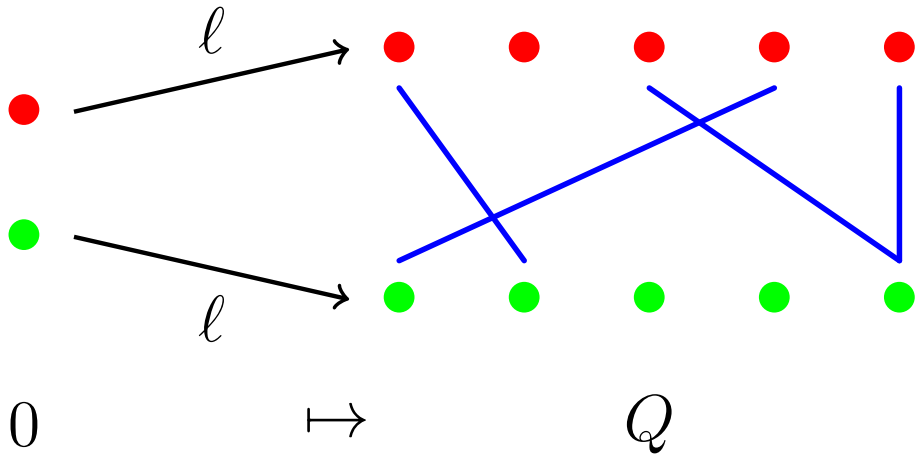
Given an integer ℓ, two probability distributions P, Q on $\{0, 1, \ldots, \ell^2\}$

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



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

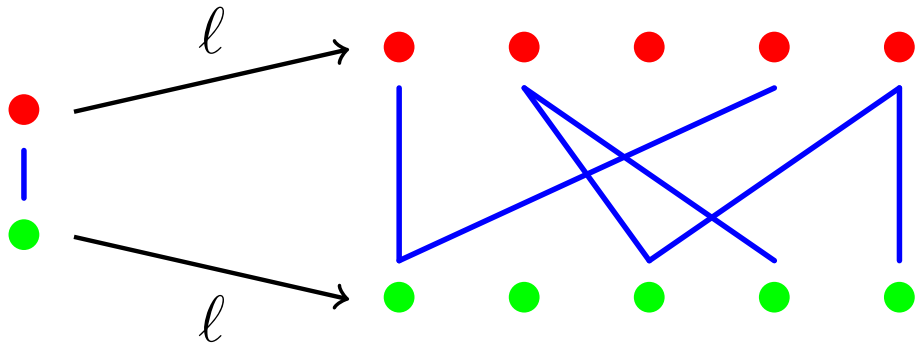
Split each node into ℓ new nodes $N = n\ell, K = k\ell$



Assign edges with distributions P, Q

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

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



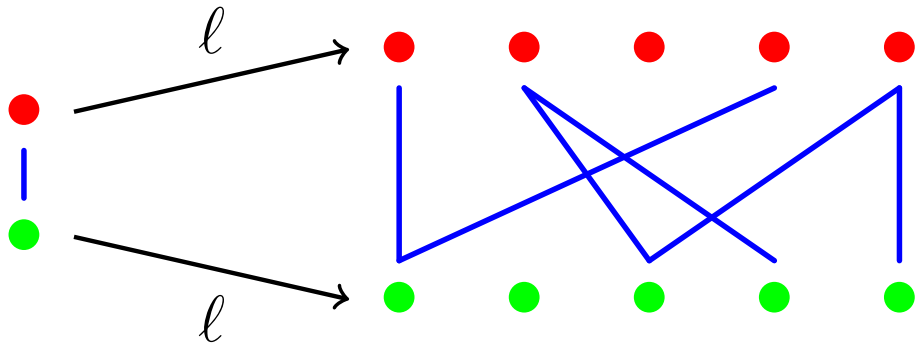
Assign edges with distributions P, Q

$$0 \mapsto Q$$

$$1 \mapsto P$$

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

$$0 \mapsto Q$$

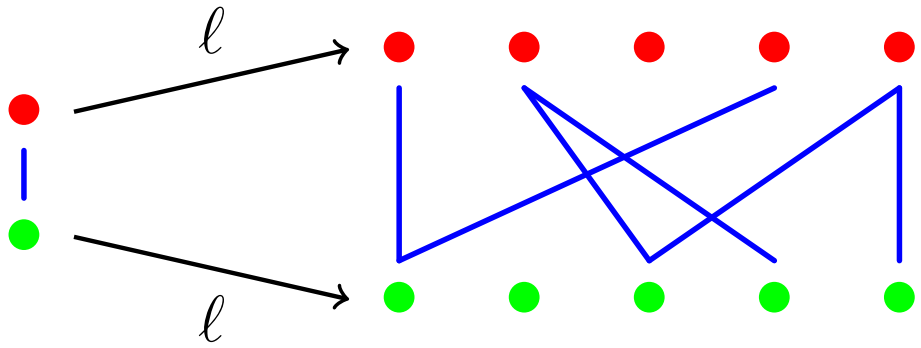
$$1 \mapsto P$$

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

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

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

$$0 \mapsto Q$$

$$1 \mapsto P$$

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

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

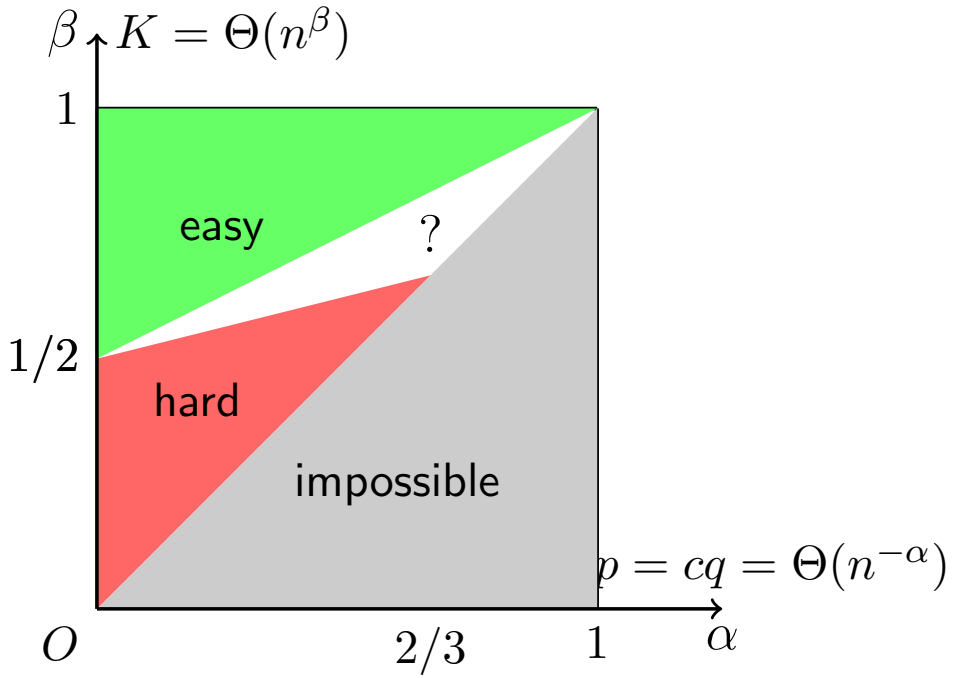
How to choose P, Q?

- 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
- Main effort: the law of the resulting graph is close to SBM in total variation

Concluding remarks

- Versatility of SDP as a simple, general purpose, computationally feasible methodology for community detection
- Construction of dual witness lacks a general recipe

Concluding remarks



References

- B. Hajek, Y. W. & J. Xu (2014). *Computational lower bounds for community detection on random graphs*. arXiv:1406.6625 (COLT '15)
- B. Hajek, Y. W. & J. Xu (2014). *Achieving exact cluster recovery threshold via semidefinite programming*. arXiv:1412.6156
- B. Hajek, Y. W. & J. Xu (2015). *Achieving exact cluster recovery threshold via semidefinite programming: Extensions*. arXiv:1502.07738

Formal statement of hardness of detecting a cluster

γ: edge probability in Planted Clique

Theorem

Assume Planted Clique Hypothesis holds for all $0 < \gamma \leq 1/2$. *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 ideas: Reduce from Planted Clique in polynomial-time

Map approximately:

- $\mathcal{G}(n, \gamma) \mapsto \mathcal{G}(N, q)$
- $\mathcal{G}(n, k, \gamma, 1) \mapsto \mathcal{G}(N, K, q, p)$

Bound the total variation distance

Lemma

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*

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

Bound the total variation distance

Lemma

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*

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

Proof ideas: $d_{\mathrm{TV}}(P, Q) \le \frac{1}{2}\sqrt{\chi^2(P, Q)}$ and use negative associations [Dubhashi-Ranjan '98] to get rid of dependency in calculating the χ^2 distance.

Bound the total variation distance

Lemma

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*

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

Proof ideas: $d_{\mathrm{TV}}(P, Q) \le \frac{1}{2}\sqrt{\chi^2(P, Q)}$ and use negative associations [Dubhashi-Ranjan '98] to get rid of dependency in calculating the χ^2 distance.

Apply the Lemma by choosing $q = \ell^{-2-\delta}$ so that $q\ell^2 \to 0$: $N = \ell^{\frac{2+\delta}{\alpha}}$, $K = \ell^{\frac{(2+\delta)\beta}{\alpha}}$, $n = \ell^{\frac{2+\delta}{\alpha}-1}$, $k = \ell^{\frac{(2+\delta)\beta}{\alpha}-1}$. Easy to check that

$$\alpha < \beta < \frac{1}{2} - \delta + \frac{\alpha(1+2\delta)}{4+2\delta} \Rightarrow \frac{\log k}{\log n} \le \frac{1}{2} - \delta$$

Spectral concentration

Theorem

Let A denote a symmetric and zero-diagonal random matrix, where the entries $\{A_{ij} : i < j\}$ are independent and $[0,1]$-valued. Assume that $\mathbb{E}[A_{ij}] \le p$, where $c_0 \log n/n \le p \le 1 - c_1$ for arbitrary constants $c_0 > 0$ and $c_1 > 0$. Then for any $c > 0$, there exists $c' > 0$ such that for any $n \ge 1$,

$$\mathbb{P}\left\{\|A - \mathbb{E}[A]\|_2 \le c'\sqrt{np}\right\} \ge 1 - n^{-c}.$$