Fundamental Limits for Community Detection

Jiaming Xu [1]

Joint work with Yudong Chen[2], Bruce Hajek[1], Yihong Wu[1]

[1]ECE, University of Illinois at Urbana-Champaign

[2]EECS, University of California, Berkeley

October 15, 2014

Community detection in networks

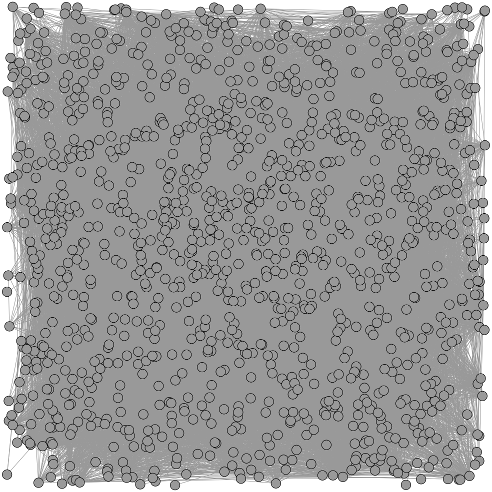
- Given a network
  - e.g. friendship networks on facebook
  - e.g. protein-protein interaction networks

Community detection in networks

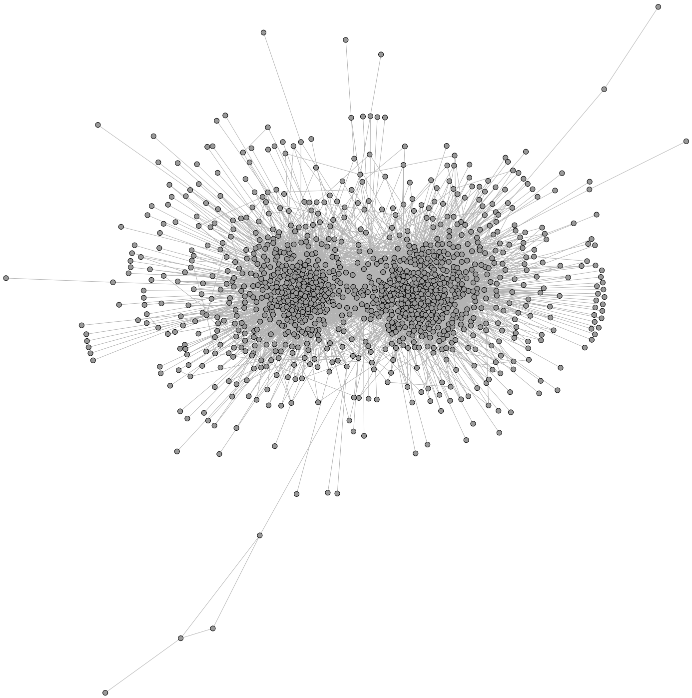
- Given a network
  - e.g. friendship networks on facebook
  - e.g. protein-protein interaction networks
- Task: Identify groups of similar nodes (communities)
  - Existence of edge or not indicates similarity
  - Communities: Densely-connected internally

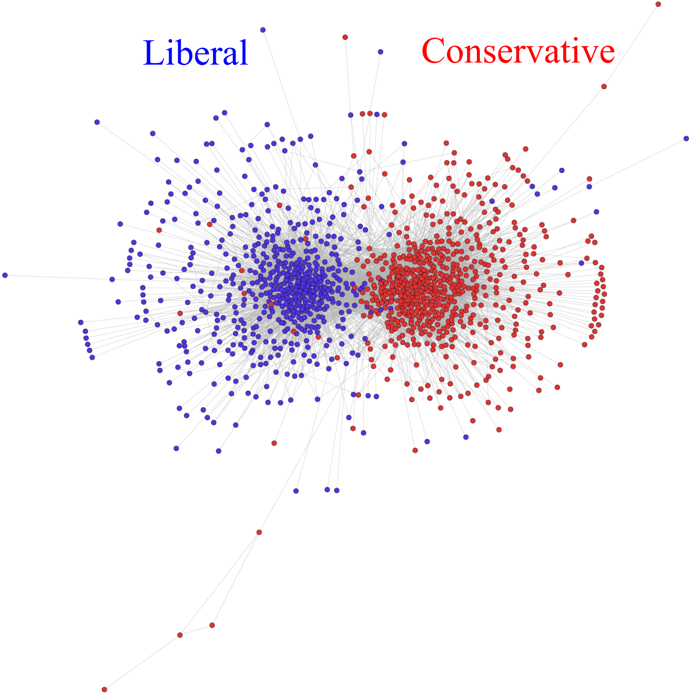
Community detection in networks

- Given a network
  - e.g. friendship networks on facebook
  - e.g. protein-protein interaction networks
- Task: Identify groups of similar nodes (communities)
  - Existence of edge or not indicates similarity
  - Communities: Densely-connected internally
- Graph clustering: Identify densely-connected groups of nodes



Political blog Network [Adamic and Glance '05]




Liberal
Conservative

Statistical and computational challenges

- From a statistical perspective
  - A large number of (small) communities
  - The observed network is sparse

Statistical and computational challenges

- From a statistical perspective
  - A large number of (small) communities
  - The observed network is sparse
- From a computational perspective
  - Large solution space

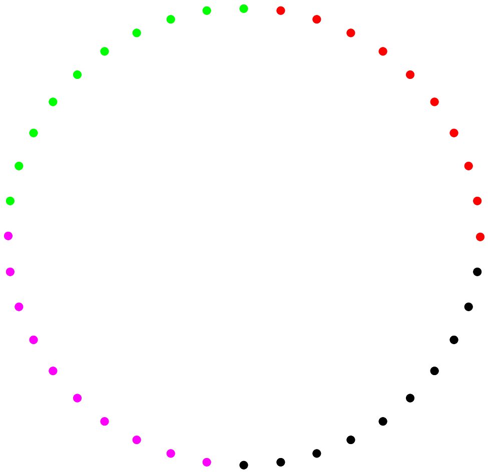
Statistical and computational challenges

- From a statistical perspective
  - A large number of (small) communities
  - The observed network is sparse
- From a computational perspective
  - Large solution space

Question

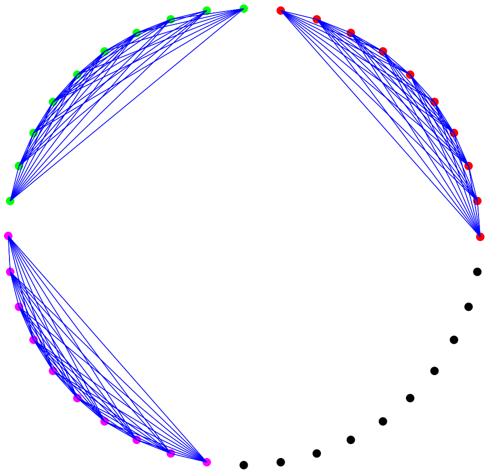
- Is there a computationally efficient and statistically optimal community detection algorithm?

Planted Cluster Model



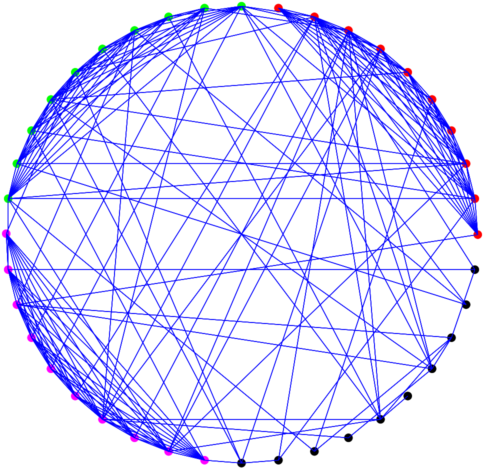
$n = 40, K = 10, r = 3$

Planted Cluster Model



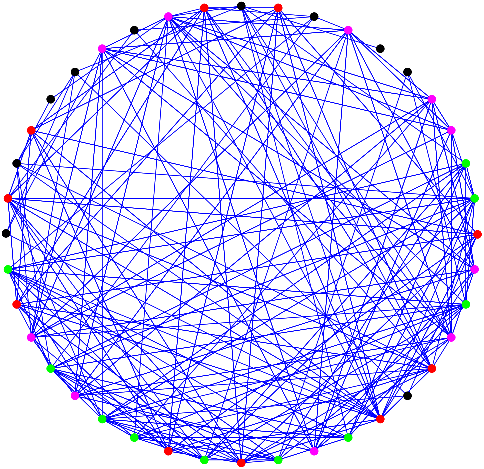
$p = 0.9$

Planted Cluster Model



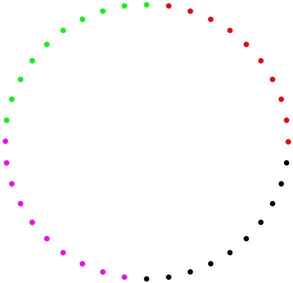
$p = 0.9 \quad q = 0.1$

Planted Cluster Model

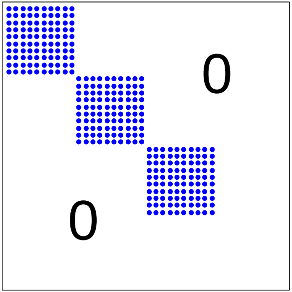


$p = 0.9 \quad q = 0.1$

Cluster recovery as structured matrix recovery

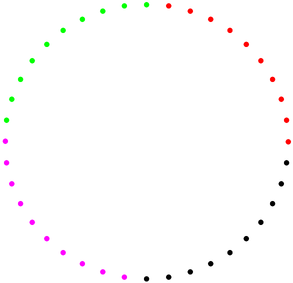


True clusters

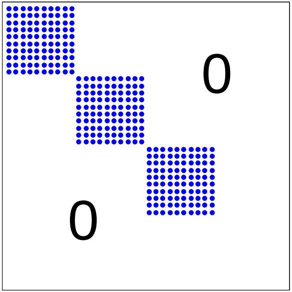


True cluster matrix Y^*

Cluster recovery as structured matrix recovery



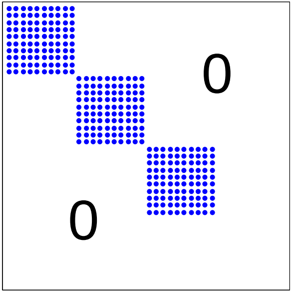
True clusters



True cluster matrix Y^*

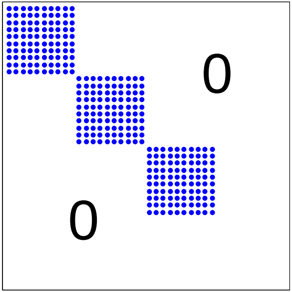
- Binary: $Y^* \in \{0, 1\}^{n \times n}$
- Low rank: $\text{rank}(Y^*) = r \ll n$
- Sparse: $\#$ of ones in Y^* is $rK^2 \ll n^2$
- Positive semi-definite: $Y^* \succeq 0$

Cluster recovery as structured matrix recovery

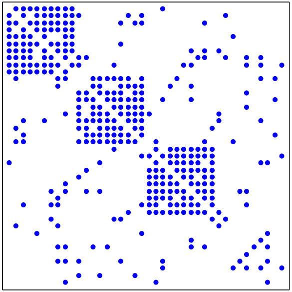


True cluster matrix Y^*

Cluster recovery as structured matrix recovery

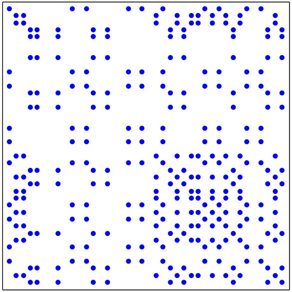


True cluster matrix Y^*

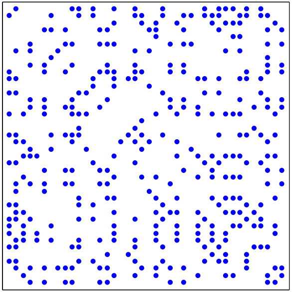


Adjacency matrix A

Cluster recovery as structured matrix recovery

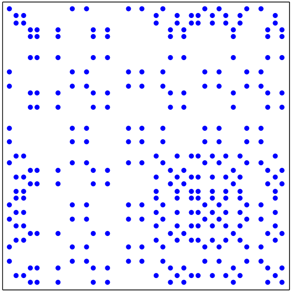


True cluster matrix Y^*

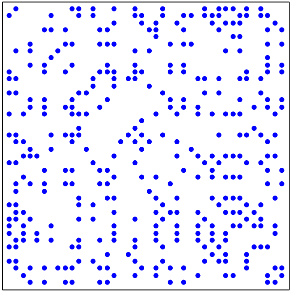


Adjacency matrix A

Cluster recovery as structured matrix recovery



True cluster matrix Y^*



Adjacency matrix A

$$Y^* \longrightarrow A \longrightarrow \widehat{Y}$$

Cluster recovery under planted cluster model

- Model parameters n, K, r, p, q
  - $n = \#$ of nodes, $K =$ size of clusters, $r = \#$ of clusters
  - $p =$ in-cluster edge probability
  - $q =$cross-cluster edge probability

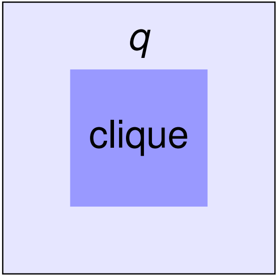
Cluster recovery under planted cluster model

- Model parameters n, K, r, p, q
  - $n = \#$ of nodes, $K =$ size of clusters, $r = \#$ of clusters
  - $p =$ in-cluster edge probability
  - $q =$cross-cluster edge probability
- Cluster recovery becomes more difficult with
  - Smaller K
  - Smaller p or $p - q$

Related work on cluster recovery

Planted cluster model covers several classical planted models

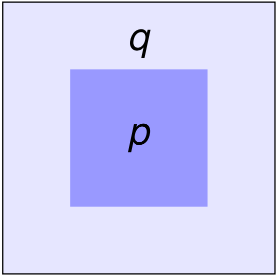
- Planted clique [McSherry '01] : $r = 1$, $p = 1$, $0 < q < 1$



Related work on cluster recovery

Planted cluster model covers several classical planted models

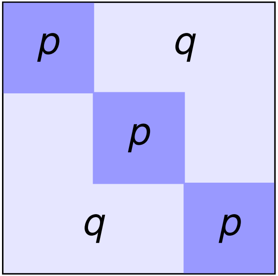
- Planted clique [McSherry '01] : $r = 1$, $p = 1$, $0 < q < 1$
- Planted dense subgraph [Arias-Castro-Verzelen '13] : $r = 1$, $0 < q < p < 1$



Related work on cluster recovery

Planted cluster model covers several classical planted models

- Planted clique [McSherry '01] : $r = 1$, $p = 1$, $0 < q < 1$
- Planted dense subgraph [Arias-Castro-Verzelen '13] : $r = 1$, $0 < q < p < 1$
- Planted partition [Condon-Karp '01] /Stochastic blockmodel [Holland et al. '83] : $n = rK$



Related work on cluster recovery

- Special case: Two clusters of size $n/2$
  - [Abbe et al. '14, Mossel et al. '14] Assume $p = \frac{a \log n}{n}, q = \frac{b \log n}{n}$. Exact recovery is possible if and only if

$$K(\sqrt{p} - \sqrt{q})^2 > \log n$$

  - [Decelle et al. '11, Mossel et al. '12 '13, Massoulié '13] Assume $p = \frac{a}{n}, q = \frac{b}{n}$. Correlated recovery is possible if and only if

$$K(p - q)^2 > p + q$$

Related work on cluster recovery

- Special case: Two clusters of size $n/2$
  - [Abbe et al. '14, Mossel et al. '14] Assume $p = \frac{a \log n}{n}, q = \frac{b \log n}{n}$. Exact recovery is possible if and only if

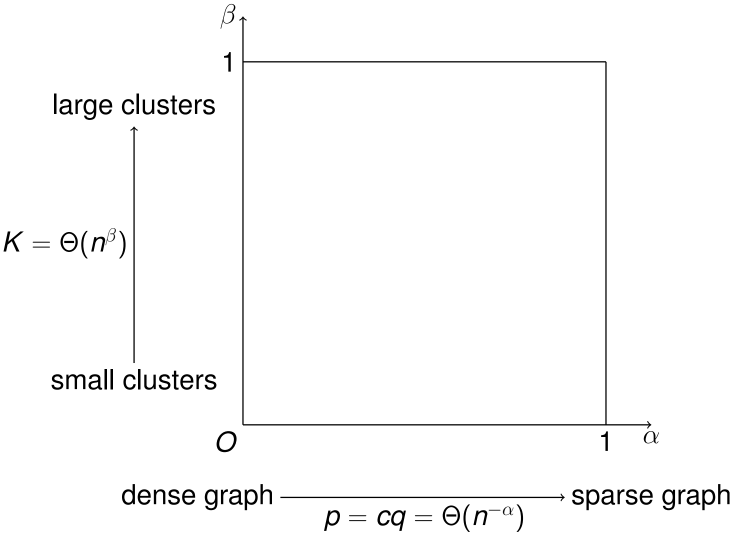
$$K(\sqrt{p} - \sqrt{q})^2 > \log n$$

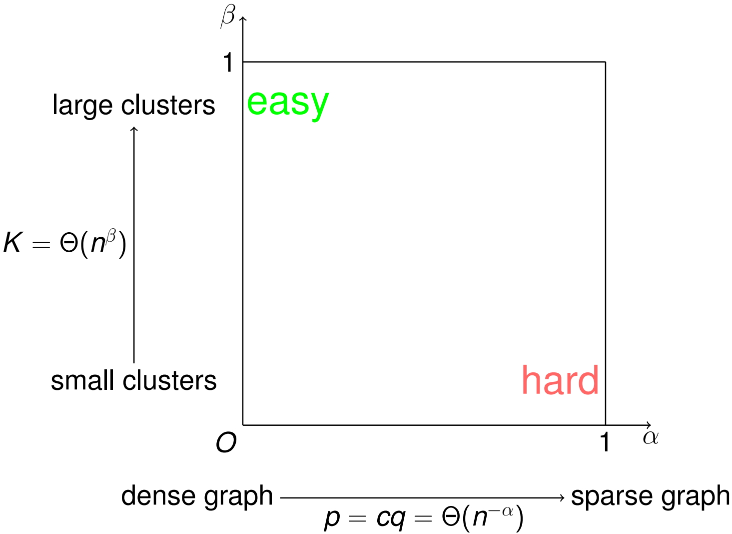
  - [Decelle et al. '11, Mossel et al. '12 '13, Massoulié '13] Assume $p = \frac{a}{n}, q = \frac{b}{n}$. Correlated recovery is possible if and only if

$$K(p-q)^2 > p + q$$

Two fundamental limits unclear in general

- Information limit: In which regime is exact recovery possible (impossible)?
- Computational limit: In which regime is exact recovery computationally easy (hard)?


β
1
large clusters
$K = \Theta(n^{\beta})$
small clusters
O
1
α
dense graph
sparse graph
$p = cq = \Theta(n^{-\alpha})$

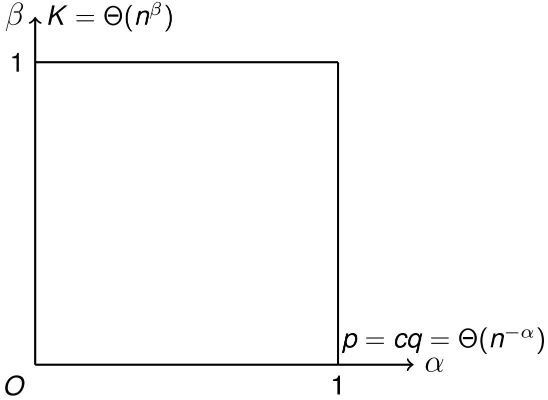


β
1
easy
large clusters
$K = \Theta(n^{\beta})$
small clusters
hard
O
1
α
dense graph
sparse graph
$p = cq = \Theta(n^{-\alpha})$

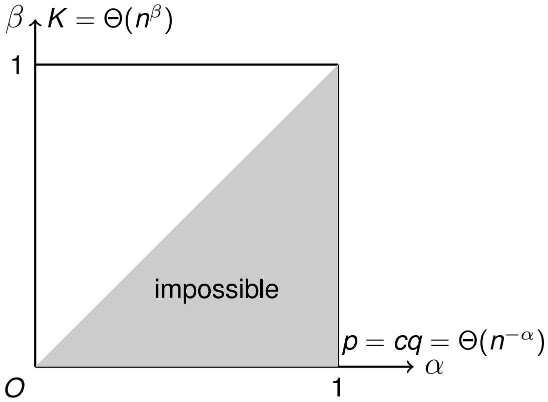
Outline

1. Cluster recovery under planted cluster model
2. Information limit: Necessary and sufficient conditions for cluster recovery
3. Computational limit
4. Empirical study

Necessary conditions for cluster recovery

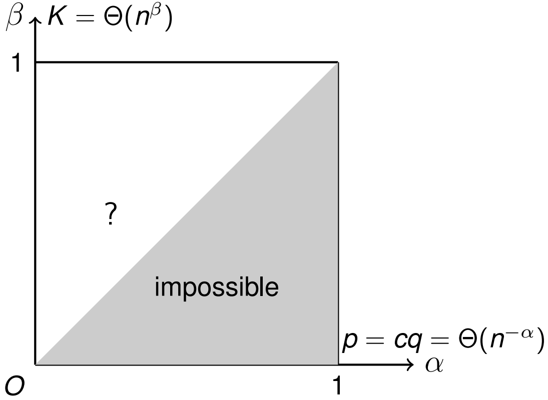


Necessary conditions for cluster recovery



Proof: $Y^* \longrightarrow A \longrightarrow \hat{Y}$. Show $I(Y^*; A) \lesssim H(Y^*)$ and use Fano's inequality

Necessary conditions for cluster recovery



Proof: $Y^* \longrightarrow A \longrightarrow \hat{Y}$. Show $I(Y^*; A) \lesssim H(Y^*)$ and use Fano's inequality

Sufficient conditions for maximum likelihood estimation

Maximum likelihood estimator: $\hat{Y} = \arg\max_Y \mathbb{P}(A|Y)$

$$Y^* \longrightarrow A \longrightarrow \hat{Y}$$

Sufficient conditions for maximum likelihood estimation

Maximum likelihood estimator: $\hat{Y} = \arg\max_Y \mathbb{P}(A|Y)$

$$Y^* \longrightarrow A \longrightarrow \hat{Y}$$

If $p > q$, maximum likelihood estimation reduces to

$$\begin{aligned} \max_Y \quad & \sum_{i,j} A_{ij} Y_{ij} \quad \leftarrow \# \text{ of in-cluster edges} \\ \text{s.t.} \quad & Y \text{ is a cluster matrix with } r \text{ clusters of size } K \end{aligned}$$

Sufficient conditions for maximum likelihood estimation

Maximum likelihood estimator: $\hat{Y} = \arg\max_Y \mathbb{P}(A|Y)$

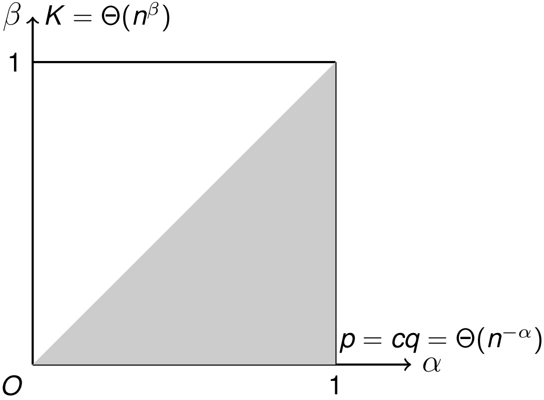
$$Y^* \longrightarrow A \longrightarrow \hat{Y}$$

If $p > q$, maximum likelihood estimation reduces to

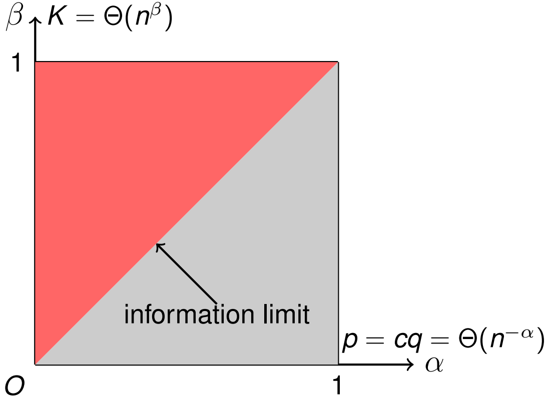
$$\begin{aligned} \max_Y \quad & \sum_{i,j} A_{ij} Y_{ij} \quad \leftarrow \# \text{ of in-cluster edges} \\ \text{s.t.} \quad & Y \text{ is a cluster matrix with } r \text{ clusters of size } K \end{aligned}$$

Q: When Y^* is the optimal solution to MLE?

Sufficient conditions for maximum likelihood estimation



Sufficient conditions for maximum likelihood estimation



Proof: Concentration inequality + union bound (needs non-trivial counting)

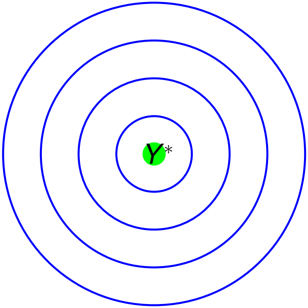
$$\max_{Y} \quad \sum_{i,j} A_{ij} Y_{ij} := f(Y)$$

s.t. $\quad Y$ is a cluster matrix with r clusters of size K

$$\max_{Y} \quad \sum_{i,j} A_{ij} Y_{ij} := f(Y)$$

s.t. $\quad Y$ is a cluster matrix with r clusters of size K

Define Hamming distance $d_H(Y, Y^*)$

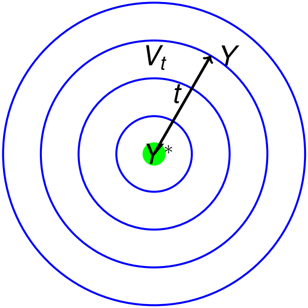


Space of all cluster matrices

$$\max_{Y} \quad \sum_{i,j} A_{ij} Y_{ij} := f(Y)$$

s.t. $\quad Y$ is a cluster matrix with r clusters of size K

Define Hamming distance $d_H(Y, Y^*)$



Space of all cluster matrices

Given $d_H(Y, Y^*) = t$

- $\log |V_t| \lesssim t \log n/K$
- $\log \mathbb{P}\{f(Y) \geq f(Y^*)\} \lesssim -t D(p\|q)$

So need $K \cdot D(p\|q) \gtrsim \log n$

Information limit for cluster recovery

Theorem (Informal)

Exact cluster recovery is possible if and only if

$$K \cdot D(q\|p) \gtrsim \log(rK) \quad \text{and} \quad K \cdot D(p\|q) \gtrsim \log n, \tag{1}$$

Information limit for cluster recovery

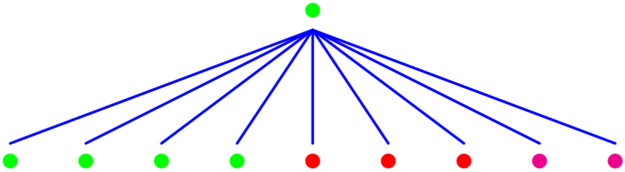
Theorem (Informal)

Exact cluster recovery is possible if and only if

$$K \cdot D(q\|p) \gtrsim \log(rK) \quad \text{and} \quad K \cdot D(p\|q) \gtrsim \log n, \tag{1}$$

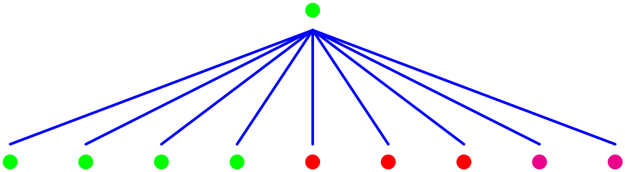
- $q \asymp p$: (2) simplifies to $K(p-q)^2 \gtrsim q(1-q)\log n$

Key idea in information limit



$S \sim \mathrm{Bin}(K-1, p)$ $T_1 \sim \mathrm{Bin}(K, q)$ T_2

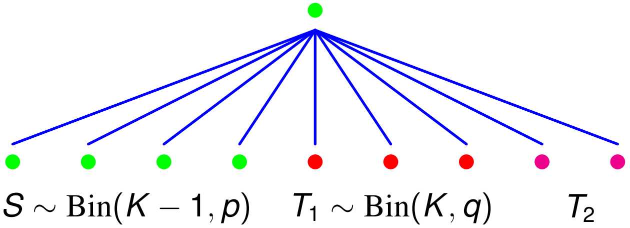
Key idea in information limit



$S \sim \mathrm{Bin}(K-1, p) \quad T_1 \sim \mathrm{Bin}(K, q) \quad T_2$

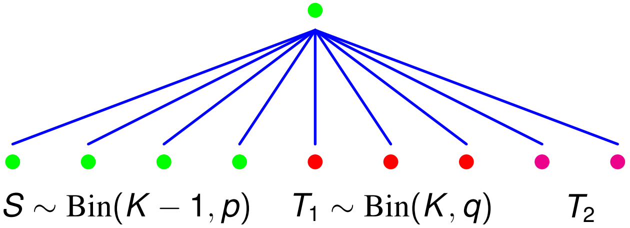
- $\mathbb{P}\{S < T_1\} \lesssim$?

Key idea in information limit



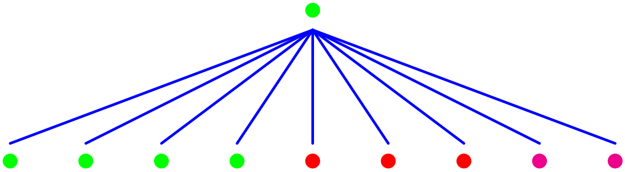
- $\mathbb{P}\{S < T_1\} \lesssim e^{-K \min\{D(q\|p), D(p\|q)\}}$

Key idea in information limit



- $\mathbb{P}\{S < T_1\} \lesssim e^{-K \min\{D(q\|p), D(p\|q)\}}$
- $\mathbb{P}\{S < \max\{T_1, \ldots, T_{r-1}\}\} \lesssim r \cdot e^{-K \min\{D(q\|p), D(p\|q)\}}$

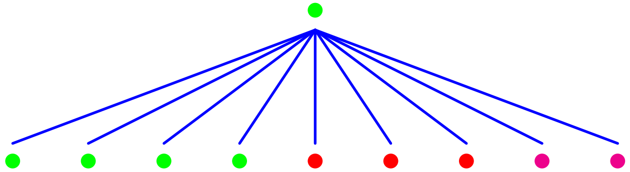
Key idea in information limit



$S \sim \mathrm{Bin}(K-1, p)$ $\quad T_1 \sim \mathrm{Bin}(K, q)$ $\quad T_2$

- $\mathbb{P}\{S < T_1\} \lesssim e^{-K \min\{D(q\|p), D(p\|q)\}}$
- $\mathbb{P}\{S < \max\{T_1, \ldots, T_{r-1}\}\} \lesssim r \cdot e^{-K \min\{D(q\|p), D(p\|q)\}}$
- $\mathbb{P}\{S < \max\{T_1, \ldots, T_{r-1}\} \text{ for all nodes}\} \lesssim nr \cdot e^{-K \min\{D(q\|p), D(p\|q)\}}$

Key idea in information limit



$S \sim \mathrm{Bin}(K-1, p)$ $\quad T_1 \sim \mathrm{Bin}(K, q)$ $\quad T_2$

- $\mathbb{P}\{S < T_1\} \lesssim e^{-K \min\{D(q\|p), D(p\|q)\}}$
- $\mathbb{P}\{S < \max\{T_1, \ldots, T_{r-1}\}\} \lesssim r \cdot e^{-K \min\{D(q\|p), D(p\|q)\}}$
- $\mathbb{P}\{S < \max\{T_1, \ldots, T_{r-1}\} \text{ for all nodes}\} \lesssim nr \cdot e^{-K \min\{D(q\|p), D(p\|q)\}}$
- If $K \min\{D(q\|p), D(p\|q)\} \gtrsim \log n$, then for every node, its color is the same as the most representative color among its neighbors

Information limit for cluster recovery

Theorem (Informal)

Exact cluster recovery is possible if and only if

$$K \cdot D(q \| p) \gtrsim \log(rK) \quad \text{and} \quad K \cdot D(p \| q) \gtrsim \log n, \tag{2}$$

- $q \asymp p$: (2) simplifies to $K(p-q)^2 \gtrsim q(1-q)\log n$

Information limit for cluster recovery

Theorem (Informal)

Exact cluster recovery is possible if and only if

$$K \cdot D(q\|p) \gtrsim \log(rK) \quad \text{and} \quad K \cdot D(p\|q) \gtrsim \log n, \tag{2}$$

- $q \asymp p$: (2) simplifies to $K(p-q)^2 \gtrsim q(1-q)\log n$
- [Abbe et al. '14, Mossel et al. '14] $p = a\log n/n, q = b\log n/n$: Exact recovery is possible if and only if $K(\sqrt{p}-\sqrt{q})^2 > \log n$

Information limit for cluster recovery

Theorem (Informal)

Exact cluster recovery is possible if and only if

$$K \cdot D(q\|p) \gtrsim \log(rK) \quad \text{and} \quad K \cdot D(p\|q) \gtrsim \log n, \tag{2}$$

- $q \asymp p$: (2) simplifies to $K(p-q)^2 \gtrsim q(1-q)\log n$
- [Abbe et al. '14, Mossel et al. '14] $p = a \log n/n, q = b \log n/n$: Exact recovery is possible if and only if $K(\sqrt{p} - \sqrt{q})^2 > \log n$
- [Decelle et al. '11, Mossel et al. '12 '13, Massoulié '13] $p = a/n, q = b/n$: Correlated recovery is possible if and only if $K(p-q)^2 > p + q$

Information limit for cluster recovery

Theorem (Informal)

Exact cluster recovery is possible if and only if

$$K \cdot D(q\|p) \gtrsim \log(rK) \quad \text{and} \quad K \cdot D(p\|q) \gtrsim \log n, \tag{2}$$

- $q \asymp p$: (2) simplifies to $K(p-q)^2 \gtrsim q(1-q)\log n$
- [Abbe et al. '14, Mossel et al. '14] $p = a\log n/n, q = b\log n/n$: Exact recovery is possible if and only if $K(\sqrt{p}-\sqrt{q})^2 > \log n$
- [Decelle et al. '11, Mossel et al. '12 '13, Massoulié '13] $p = a/n, q = b/n$: Correlated recovery is possible if and only if $K(p-q)^2 > p+q$

Question

Q: Is the information limit efficiently achievable in general?

Outline

1. Cluster recovery under planted cluster model
2. Information limit: Necessary and sufficient conditions for cluster recovery
3. Computational limit
    - A polynomial-time cluster recovery algorithm
    - Complexity theoretic lower bounds
4. Empirical study

Polynomial-time recovery: Convex relaxation of MLE

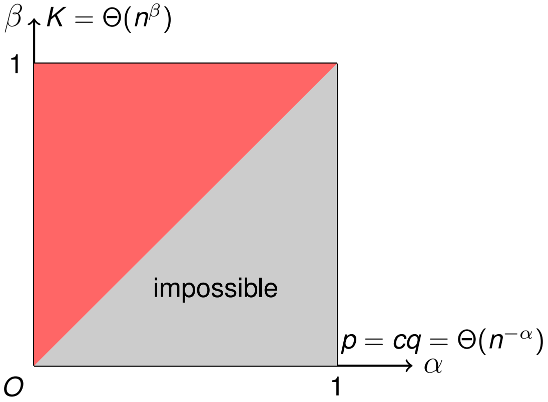
- $\text{rank}(Y^*) = r \ll n$
- Nuclear norm $\|\cdot\|_*$ (sum of singular values) is a convex surrogate for rank function: $\|Y^*\|_* = rK$

Polynomial-time recovery: Convex relaxation of MLE

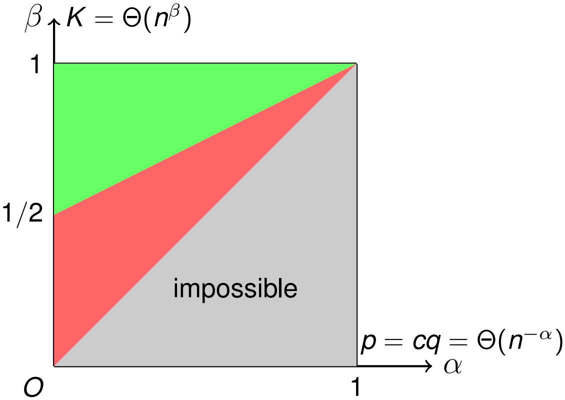
- $\text{rank}(Y^*) = r \ll n$
- Nuclear norm $\|\cdot\|_*$ (sum of singular values) is a convex surrogate for rank function: $\|Y^*\|_* = rK$
- A convex relaxation of MLE

$$\begin{aligned}
\max_{Y} \ & \sum_{ij} A_{ij} Y_{ij} \\
\text{s.t.} \ & \|Y\|_* \le rK \\
& \sum_{ij} Y_{ij} = rK^2, \ \ Y_{ij} \in [0,1].
\end{aligned}$$

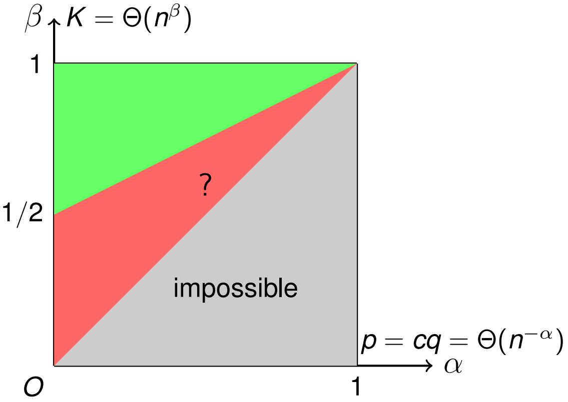
Polynomial-time recovery: Convex relaxation of MLE



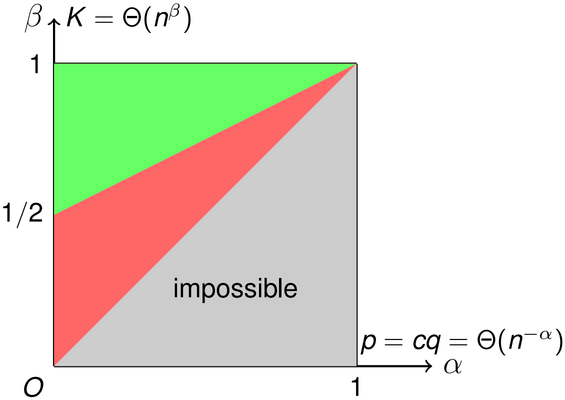
Polynomial-time recovery: Convex relaxation of MLE



Polynomial-time recovery: Convex relaxation of MLE

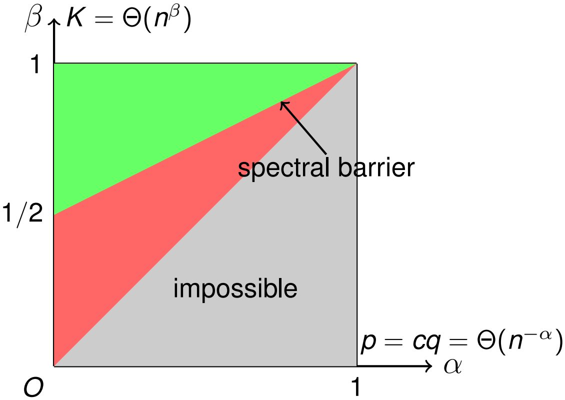


Polynomial-time recovery: Convex relaxation of MLE



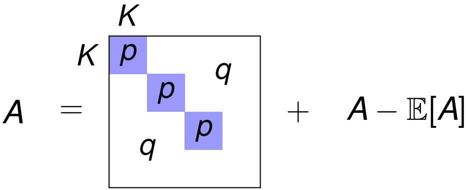
- Conjecture on computational limit: No polynomial-time algorithm succeeds beyond the green regime

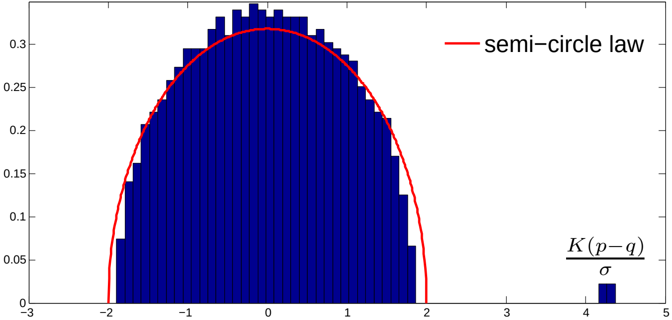
Polynomial-time recovery: Convex relaxation of MLE



- Conjecture on computational limit: No polynomial-time algorithm succeeds beyond the green regime
- Spectral barrier prevents spectrum of A revealing clusters [Nadakuditi-Newman '12]

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

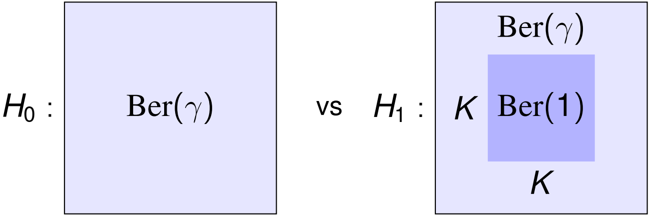




Eigenvalue distribution of $\frac{A-q\mathbf{1}\mathbf{1}^\top}{\sigma}$ for $\sigma = \sqrt{Kp + (n-K)q}$

Complexity theoretic lower bounds conditional on Planted Clique hardness

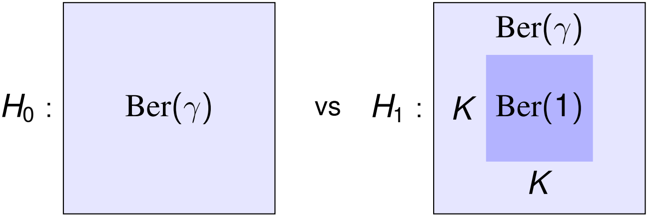
Planted Clique hardness



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

- detection is possible but believed to have high computational complexity

Planted Clique hardness

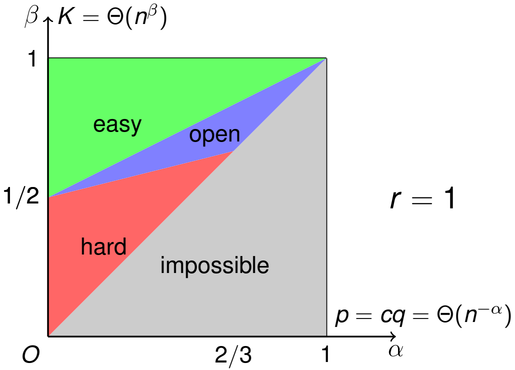


Intermediate regime: $\log n \ll K \ll \sqrt{n}$, $\gamma = \Theta(1)$

- detection is possible but believed to have high computational complexity
- 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]: $\gamma = 2^{-\log^{0.99} n}$

Conditional hardness for recovering a single cluster

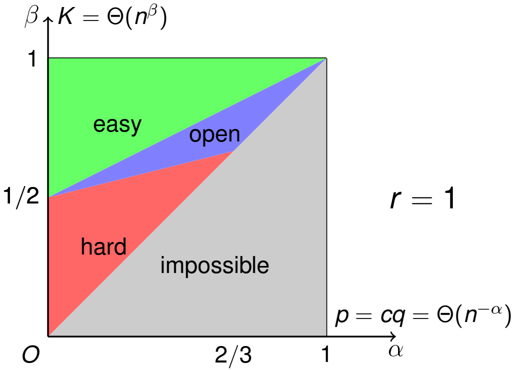
Assuming Planted Clique hardness for any constant $\gamma > 0$



$r = 1$

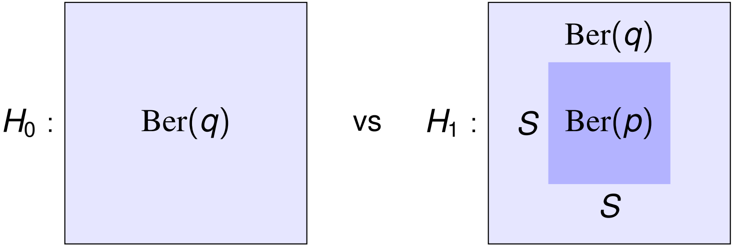
Conditional hardness for recovering a single cluster

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



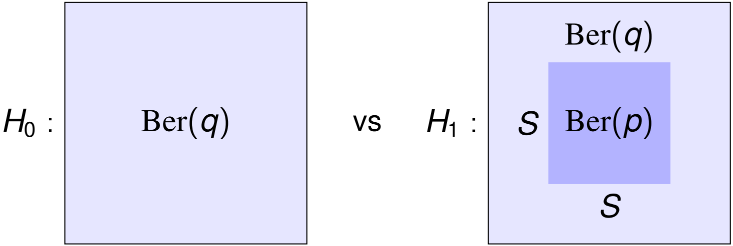
- Proof step 1: Recovery is “harder” than detection
- Proof step 2: Detecting a single cluster in the red regime is at least as hard as detecting a clique of size $K = o(\sqrt{n})$

Detection of a single cluster



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

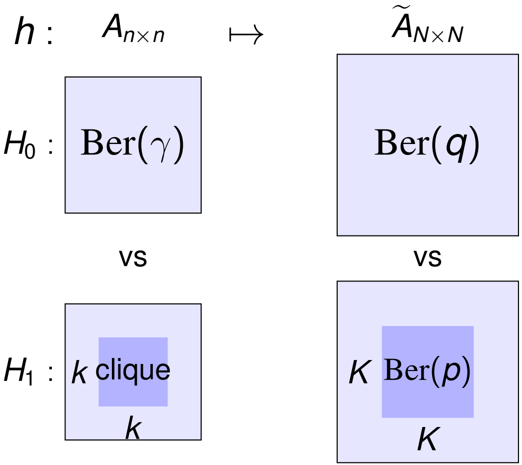
Detection of a single cluster



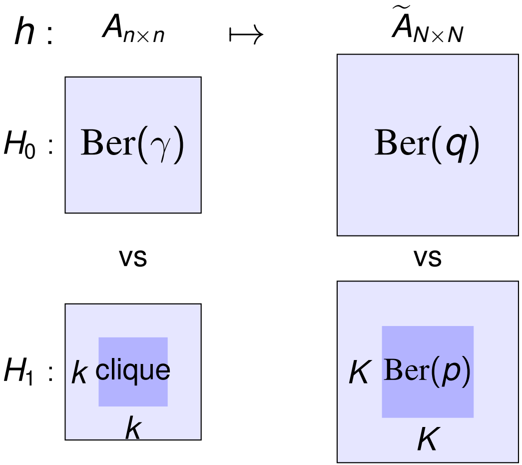
Each node is included in S with probability $\frac{K}{n}$

Complexity theoretic lower bounds

Reduced from Planted Clique in polynomial time



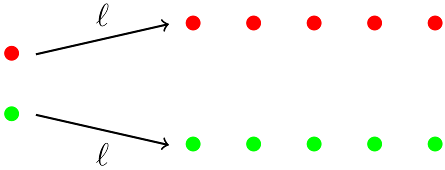
$h : A_{n \times n} \mapsto \widetilde{A}_{N \times N}$
H_0 :
Ber(γ)
Ber(q)
vs
vs
H_1 :
k clique
k
K Ber(p)
K



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

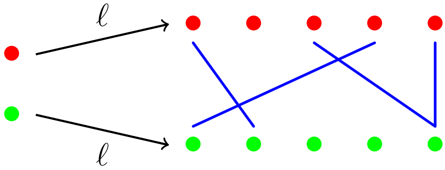
Given an integer ℓ, two probability distributions P, Q on $\{0, 1, \ldots, \ell^2\}$

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

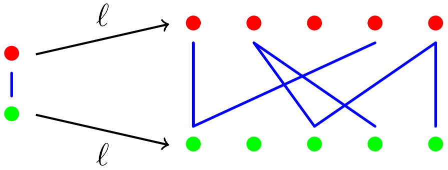


$0 \quad \mapsto \quad Q$

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



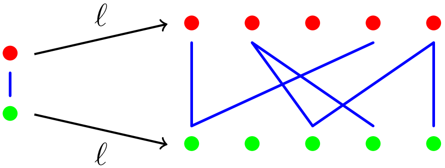
Assign edges with distributions P, Q

$0 \mapsto Q$

$1 \mapsto P$

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

$$0 \mapsto Q$$

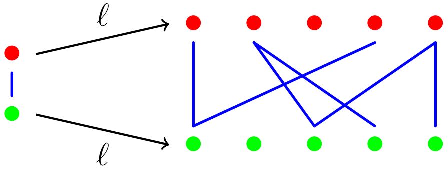
$$1 \mapsto P$$

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

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

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

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

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

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

How to choose P, Q?

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

Outline

Empirical study on political blog network

- Pre-processing: Ignore directions and select the largest connected component with 1222 nodes, $16,714$ edges

Empirical study on political blog network

- Pre-processing: Ignore directions and select the largest connected component with 1222 nodes, $16,714$ edges
- Convex relaxation of ML estimation

$$\begin{aligned} \max_{Y} \quad & \sum_{i<j} (A_{ij} - \lambda) Y_{ij} \\ \text{s.t.} \quad & Y \succeq 0, Y_{ii} = 1, \forall i \\ & Y_{ij} \in [0,1], \ \forall i \neq j \end{aligned}$$

- Solve for $\widehat{Y}$ and use k-means with $k = 2$ on $\widehat{Y}$

Empirical study on political blog network

- Pre-processing: Ignore directions and select the largest connected component with 1222 nodes, $16,714$ edges
- Convex relaxation of ML estimation

$$\begin{aligned} \max_{Y} \quad & \sum_{i<j} (A_{ij} - \lambda) Y_{ij} \\ \text{s.t.} \quad & Y \succeq 0,\ Y_{ii} = 1, \forall i \\ & Y_{ij} \in [0,1],\ \forall i \neq j \end{aligned}$$

- Solve for $\widehat{Y}$ and use k-means with $k = 2$ on $\widehat{Y}$
- Theory suggests $q < \lambda < p$ [Chen et al. '13, Cai and Li '14]

Empirical study on political blog network

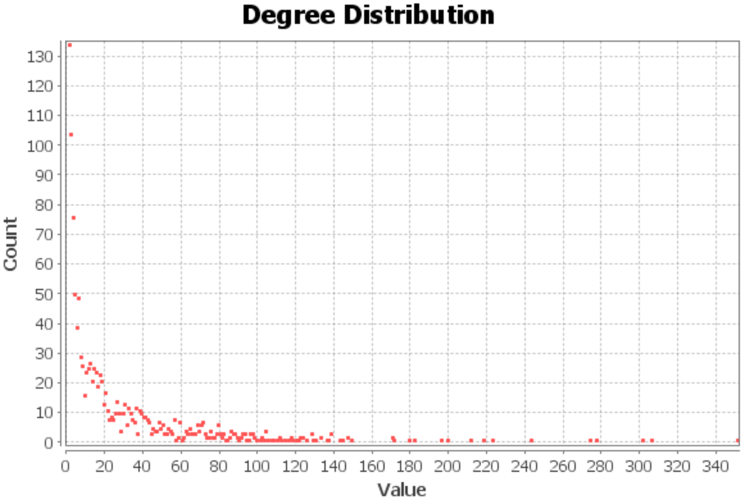
- Pre-processing: Ignore directions and select the largest connected component with 1222 nodes, $16,714$ edges
- Convex relaxation of ML estimation

$$\begin{aligned} \max_{Y} \quad & \sum_{i<j} (A_{ij} - \lambda) Y_{ij} \\ \text{s.t.} \quad & Y \succeq 0, Y_{ii} = 1, \forall i \\ & Y_{ij} \in [0,1], \ \forall i \neq j \end{aligned}$$

- Solve for $\widehat{Y}$ and use k-means with $k = 2$ on $\widehat{Y}$
- Theory suggests $q < \lambda < p$ [Chen et al. '13, Cai and Li '14]
- Choose $\lambda = \frac{\text{median degree}}{n}$ and fraction of mis-classified nodes: $\epsilon = 195/1222 \approx 0.16$

Degree distribution of political blog network

High degree variation: Max degree 351, mean degree 27, median degree 13



Convex relaxation of MLE with degree correction

- Given a random graph uniformly chosen with a fixed degree sequence $\{d_i\}$

$$\mathbb{P}[A_{ij} = 1] \approx \frac{d_i d_j}{\sum_k d_k}$$

Convex relaxation of MLE with degree correction

- Given a random graph uniformly chosen with a fixed degree sequence $\{d_i\}$

$$\mathbb{P}[A_{ij} = 1] \approx \frac{d_i d_j}{\sum_k d_k}$$

- Choose $\lambda_{ij} = \frac{d_i d_j}{\sum_k d_k}$ and let $B_{ij} = A_{ij} - \lambda_{ij}, \forall i \neq j$

$$\begin{aligned} \max_{Y} \quad & \sum_{i<j} B_{ij} Y_{ij} \\ \text{s.t.} \quad & Y \succeq 0,\ Y_{ii} = 1, \forall i \\ & Y_{ij} \in [0,1],\ \forall i \neq j \end{aligned}$$

- B is known as modularity matrix [Newman '06]

Convex relaxation of MLE with degree correction

- Given a random graph uniformly chosen with a fixed degree sequence $\{d_i\}$

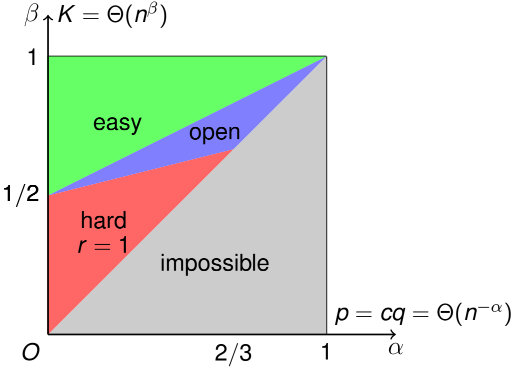
$$\mathbb{P}[A_{ij} = 1] \approx \frac{d_i d_j}{\sum_k d_k}$$

- Choose $\lambda_{ij} = \frac{d_i d_j}{\sum_k d_k}$ and let $B_{ij} = A_{ij} - \lambda_{ij}, \forall i \neq j$

$$\begin{aligned} \max_Y \quad & \sum_{i<j} B_{ij} Y_{ij} \\ \text{s.t.} \quad & Y \succeq 0,\ Y_{ii} = 1, \forall i \\ & Y_{ij} \in [0, 1],\ \forall i \neq j \end{aligned}$$

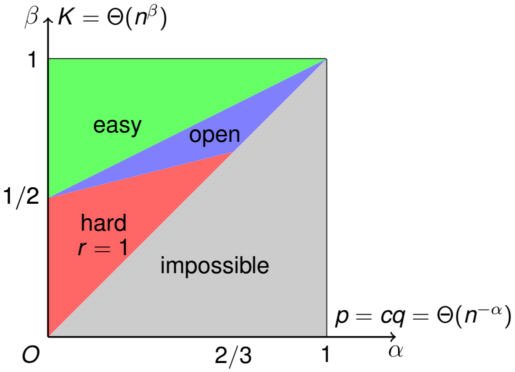
- B is known as modularity matrix [Newman '06]
- Fraction of mis-classified nodes: $\epsilon = 62/1222 \approx 0.05$

Summary



β
$K = \Theta(n^\beta)$
1
1/2
easy
open
hard
$r = 1$
impossible
$p = cq = \Theta(n^{-\alpha})$
O
2/3
1
α

Summary



References

- Y. Chen & X. (2014) *Statistical-Computational Tradeoffs in Planted Problems and Submatrix Localization with a Growing Number of Clusters and Submatrices*. arXiv:1402.1267.
- B. Hajek, Y. Wu & X. (2014) *Computational Lower Bounds for Community Detection on Random Graphs*. arXiv:1406.6625.