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
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]
Statistical and computational challenges

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

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

$p = 0.9$

$q = 0.1$
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 $Y^*$

- **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

- **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 $A$

$Y^* \rightarrow A \rightarrow \hat{Y}$
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 becomes more difficult with smaller \( K \), smaller \( p \), or \( p - q \).
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]: $r = 1$, $p = 1$, $0 < q < 1$

![Diagram of a clique with parameters $q$ and $clique$]
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)?
\( p = cq = \Theta(n^{-\alpha}) \)

\( K = \Theta(n^{\beta}) \)

large clusters

small clusters

dense graph \( p = cq = \Theta(n^{-\alpha}) \) \rightarrow \) sparse graph
small clusters

large clusters

$K = \Theta(n^\beta)$

$p = cq = \Theta(n^{-\alpha})$

dense graph \rightarrow sparse graph

hard

easy
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

\[ p = cq = \Theta(n^{-\alpha}) \]

\[ K = \Theta(n^\beta) \]
Necessary conditions for cluster recovery

\[ p = cq = \Theta(n^{-\alpha}) \]

\[ K = \Theta(n^\beta) \]

Proof: \( Y^* \xrightarrow{} A \xrightarrow{} \hat{Y} \). Show \( I(Y^*; A) \lesssim H(Y^*) \) and use Fano’s inequality.
Necessary conditions for cluster recovery

\[ p = cq = \Theta(n^{-\alpha}) \]

Proof: \( Y^* \rightarrow A \rightarrow \hat{Y} \). Show \( I(Y^*; A) \lesssim H(Y^*) \) and use Fano’s inequality
Maximum likelihood estimator: $\hat{Y} = \arg \max_Y P(A|Y)$

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

\[ Y^* \rightarrow A \rightarrow \hat{Y} \]

If \( p > q \), maximum likelihood estimation reduces to

\[
\max_Y \sum_{i,j} A_{ij} Y_{ij} \leftarrow \text{# of in-cluster edges}
\]

s.t. \( Y \) is a cluster matrix with \( r \) clusters of size \( K \)
Maximum likelihood estimator: \( \hat{Y} = \arg \max_Y \mathbb{P}(A|Y) \)

\[
Y^* \rightarrow A \rightarrow \hat{Y}
\]

If \( p > q \), maximum likelihood estimation reduces to

\[
\max_Y \sum_{i,j} A_{ij} Y_{ij} \leftarrow \# \text{ of in-cluster edges}
\]

s.t. \( Y \) is a cluster matrix with \( r \) clusters of size \( K \)

Q: When \( Y^* \) is the optimal solution to MLE?
Sufficient conditions for maximum likelihood estimation

\[ p = cq = \Theta(n^{-\alpha}) \]

\[ K = \Theta(n^\beta) \]
Sufficient conditions for maximum likelihood estimation

\[ p = cq = \Theta(n^{-\alpha}) \]

\[ K = \Theta(n^\beta) \]

Proof: Concentration inequality + union bound (needs non-trivial counting)
\[ \max_Y \sum_{i,j} A_{ij} Y_{ij} := f(Y) \]

s.t. \( Y \) is a cluster matrix with \( r \) clusters of size \( K \).
\[
\max_Y \sum_{i,j} A_{ij} Y_{ij} := f(Y) \\
\text{s.t. } Y \text{ is a cluster matrix with } r \text{ clusters of size } K
\]

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

Space of all cluster matrices
\[
\max_Y \sum_{i,j} A_{ij} Y_{ij} := f(Y)
\]

s.t. \( Y \) is a cluster matrix with \( r \) clusters of size \( K \)

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

Space of all cluster matrices

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

\[
\log |V_t| \lesssim t \log n/K
\]

\[
\log \mathbb{P}\{f(Y) \geq f(Y^*)\} \lesssim -tD(p\|q)
\]

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

Exact cluster recovery is possible if and only if

\[ K \cdot D(q \| p) \gtrsim \log(rK) \quad \text{and} \quad K \cdot D(p \| q) \gtrsim \log n, \]  

(1)
Theorem (Informal)

Exact cluster recovery is possible if and only if

\[ K \cdot D(q \parallel p) \gtrsim \log(rK) \quad \text{and} \quad K \cdot D(p \parallel q) \gtrsim \log n, \quad (1) \]

\[ q \approx p: \text{(2) simplifies to } K(p - q)^2 \gtrsim q(1 - q) \log n \]
Key idea in information limit

\[ S \sim \text{Bin}(K - 1, \rho) \quad T_1 \sim \text{Bin}(K, q) \quad T_2 \]
Key idea in information limit

\[ S \sim \text{Bin}(K - 1, p) \quad T_1 \sim \text{Bin}(K, q) \quad T_2 \]

\[ \mathbb{P}\{S < T_1\} \lesssim ? \]
Key idea in information limit

\[ S \sim \text{Bin}(K - 1, p) \quad T_1 \sim \text{Bin}(K, q) \quad T_2 \]

\[ \mathbb{P}\{S < T_1\} \lesssim e^{-K \min\{D(q\|p), D(p\|q)\}} \]
Key idea in information limit

\[ S \sim \text{Bin}(K - 1, p) \quad T_1 \sim \text{Bin}(K, q) \quad T_2 \]

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

\[ \mathbb{P}\{ S < \max\{ T_1, \ldots, T_{r-1} \} \} \lesssim r \cdot e^{-K \min\{D(q\|p), D(p\|q)\}} \]
Key idea in information limit

\[ S \sim \text{Bin}(K - 1, p) \quad T_1 \sim \text{Bin}(K, q) \quad T_2 \]

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

\[ S \sim \text{Bin}(K - 1, p) \quad T_1 \sim \text{Bin}(K, q) \quad T_2 \]

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

Exact cluster recovery is possible if and only if

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

- \( q \approx p \): (2) simplifies to \( K(p - q)^2 \gtrsim q(1 - q) \log n \)
**Theorem (Informal)**

Exact cluster recovery is possible if and only if

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

- \( q \asymp p \): (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 \( K(\sqrt{p} - \sqrt{q})^2 > \log n \)
Theorem (Informal)

Exact cluster recovery is possible if and only if

\[ K \cdot D(q||p) \gtrsim \log(rK) \quad \text{and} \quad K \cdot D(p||q) \gtrsim \log n, \]  

(2)

- \( q \approx p \): (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

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

- \( q \asymp p \): (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

\[
\max_Y \sum_{ij} A_{ij} Y_{ij} \\
\text{s.t. } \| Y \|_* \leq rK \\
\sum_{ij} Y_{ij} = rK^2, \ Y_{ij} \in [0, 1].
\]
Polynomial-time recovery: Convex relaxation of MLE

\[ \alpha \]
\[ \beta \]
\[ K = \Theta(n^\beta) \]
\[ p = cq = \Theta(n^{-\alpha}) \]

impossible

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

Spectral barrier prevents spectrum of

[Nadakuditi-Newman '12]
Polynomial-time recovery: Convex relaxation of MLE

$K = \Theta(n^\beta)$

impossible

$p = cq = \Theta(n^{-\alpha})$

$\alpha$

$\beta$

$1$

$1/2$

$O$

$1$
Polynomial-time recovery: Convex relaxation of MLE

\[ K = \Theta(n^\beta) \]

\[ p = cq = \Theta(n^{-\alpha}) \]

Spectral barrier prevents spectrum of \( A \) revealing clusters

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

\[ \frac{1}{2} \leq \alpha \leq 1 \]
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

\[ K = \Theta(n^\beta) \]

\[ p = cq = \Theta(n^{-\alpha}) \]

- **Conjecture** on computational limit: No polynomial-time algorithm succeeds beyond the green regime
- **Spectral barrier** prevents spectrum of \( A \) revealing clusters

[Nadakuditi-Newman ’12]
\[ A = \begin{bmatrix} K & p & q \\ K & p & p \\ q & p \end{bmatrix} + A - \mathbb{E}[A] \]
$A = \begin{bmatrix} K \\ p \\ q \\ p \\ p \\ q \\ p \end{bmatrix} + A - \mathbb{E}[A]$

Eigenvalue distribution of $\frac{A-q11^\top}{\sigma}$ for $\sigma = \sqrt{Kp + (n-K)q}$
Complexity theoretic lower bounds conditional on Planted Clique hardness
Planted Clique hardness

\[ H_0 : \text{Ber}(\gamma) \quad \text{vs} \quad H_1 : \begin{bmatrix} K \\ \text{Ber}(1) \end{bmatrix} \]

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 $\gamma > 0$

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 $K = \Theta(\sqrt{n})$
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

$H_0 : \text{Ber}(q)$  vs  $H_1 : S \text{Ber}(p) \\ S$

Each node is included in $S$ with probability $\frac{K}{n}$
Detection of a single cluster

\[ H_0 : \text{Ber}(q) \quad \text{vs} \quad H_1 : \begin{array}{c} S \\ \text{Ber}(p) \end{array} \]

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

Complexity theoretic lower bounds

Reduced from Planted Clique in polynomial time
\[ h : A_{n \times n} \rightarrow \tilde{A}_{N \times N} \]

\[ H_0 : \text{Ber}(\gamma) \quad \text{vs} \quad \text{Ber}(q) \]

\[ H_1 : k \text{ clique} \quad \text{vs} \quad K \text{ Ber}(p) \]

\[ \tilde{A}_{N \times N} \text{ is agnostic to the clique and can be computed in P-time} \]
$h : \ A_{n \times n} \rightarrow \tilde{A}_{N \times N}$

$H_0 : \text{Ber}(\gamma)$

$H_1 : \begin{array}{c} k \text{ clique} \\ k \end{array}$

$\text{vs}$

$\text{vs}$

$\text{vs}$

$K \text{ Ber}(p)$

$h : A \mapsto \tilde{A}$ is agnostic to the clique and can be computed in P-time
Given an integer $\ell$, two probability distributions $P, Q$ on $\{0, 1, \ldots, \ell^2\}$

Split each node into $\ell$ new nodes

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

Matching $H_0$: $(1 - \gamma)Q + \gamma P = \text{Bin}(\ell^2, q)$

Matching $H_1$ approximately: $P \approx \text{Bin}(\ell^2, p)$ in total variation distance
Given an integer $\ell$, two probability distributions $P, Q$ on $\{0, 1, \ldots, \ell^2\}$

Split each node into $\ell$ new nodes

$\begin{align*}
N &= n\ell, \\
K &= k\ell
\end{align*}$

Assign edges with distributions $P, Q$

Matching $H_0$:

$$(1 - \gamma)Q + \gamma P = \text{Bin}(\ell^2, q)$$

Matching $H_1$ approximately:

$$P \approx \text{Bin}(\ell^2, p)$$

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

Split each node into $\ell$ new nodes

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

Assign edges with distributions $P, Q$

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

Split each node into \( \ell \) new nodes
\[
N = n\ell, \quad K = k\ell
\]

Assign edges with distributions \( P, Q \)

\[
\begin{align*}
H_0 & : \quad \text{Ber}(\gamma) & (1 - \gamma)Q + \gamma P \\
H_1 & : \quad \text{Ber}(1) \quad \text{(in-clique)} & P \quad \text{(in-cluster)}
\end{align*}
\]
Given an integer $\ell$, two probability distributions $P, Q$ on $\{0, 1, \ldots, \ell^2\}$

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

Assign edges with distributions $P, Q$

$H_0 : \quad \text{Ber}(\gamma)$
$H_1 : \quad \text{Ber}(1) \ (\text{in-clique})$

$P, Q$

How to choose $P, Q$?

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

1. Cluster recovery under planted cluster model
2. Information limit: Necessary and sufficient conditions for cluster recovery
3. Computational limit
4. Empirical study
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
  \[
  \max_Y \sum_{i<j} (A_{ij} - \lambda) Y_{ij}
  \]
  \[
  \text{s.t. } Y \succeq 0, Y_{ii} = 1, \forall i
  \]
  \[
  Y_{ij} \in [0, 1], \forall i \neq j
  \]
- Solve for $\hat{Y}$ and use k-means with $k = 2$ on $\hat{Y}$

Theory suggests $q < \lambda < p$ [Chen et al. '13, Cai and Li '14]

Choose $\lambda = \text{median degree}$ and fraction of mis-classified nodes: $\epsilon = 195/1222 \approx 0.16$.
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

\[
\max_Y \sum_{i<j} (A_{ij} - \lambda) Y_{ij} \\
\text{s.t. } Y \succeq 0, Y_{ii} = 1, \forall i \\
Y_{ij} \in [0, 1], \forall i \neq j
\]

- Solve for \( \hat{Y} \) and use k-means with \( k = 2 \) on \( \hat{Y} \)
- Theory suggests \( q < \lambda < p \) [Chen et al. ’13, Cai and Li ’14]
Pre-processing: Ignore directions and select the largest connected component with 1222 nodes, 16,714 edges

Convex relaxation of ML estimation

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

Solve for \(\hat{Y}\) and use k-means with \(k = 2\) on \(\hat{Y}\)

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 \( \{ d_i \} \)

\[
\mathbb{P}[A_{ij} = 1] \approx \frac{d_i d_j}{\sum_k d_k}
\]
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 \( B_{ij} = A_{ij} - \lambda_{ij}, \forall i \neq j \):

\[
\max_Y \sum_{i<j} B_{ij} Y_{ij}
\]

s.t. \( Y \succeq 0, Y_{ii} = 1, \forall i \)

\( Y_{ij} \in [0, 1], \forall i \neq j \)

\( B \) 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 \( B_{ij} = A_{ij} - \lambda_{ij}, \forall i \neq j \)

\[
\max_{Y} \sum_{i<j} B_{ij} Y_{ij}
\]

s.t. \( Y \succeq 0, Y_{ii} = 1, \forall i \)

\( Y_{ij} \in [0, 1], \forall i \neq j \)

\( B \) is known as modularity matrix [Newman ’06]

Fraction of mis-classified nodes: \( \epsilon = \frac{62}{1222} \approx 0.05 \)
$p = cq = \Theta(n^{-\alpha})$

$K = \Theta(n^\beta)$

\( \beta \)

1

1/2

1/2

\( \alpha \)

2/3

1

impossible

hard

open

easy

\( r = 1 \)

References


Summary

\[ p = cq = \Theta(n^{-\alpha}) \]
\[ K = \Theta(n^\beta) \]

References