Achieving Exact Cluster Recovery Threshold via Semidefinite Programming

Jiaming Xu

Department of Statistics, The Wharton School
University of Pennsylvania

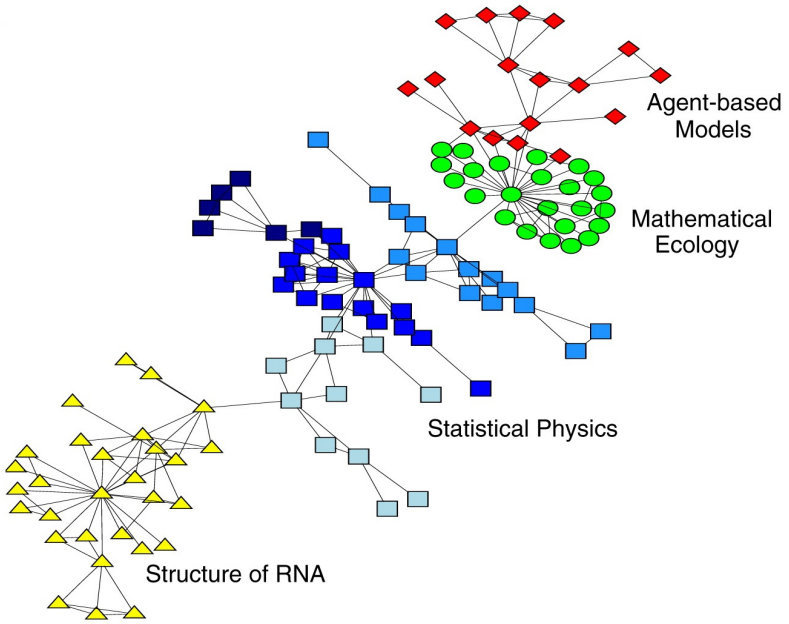
jiamingx@wharton.upenn.edu

Joint work with Bruce Hajek (Illinois) and Yihong Wu (Illinois)

June 17, 2015

Community detection in networks

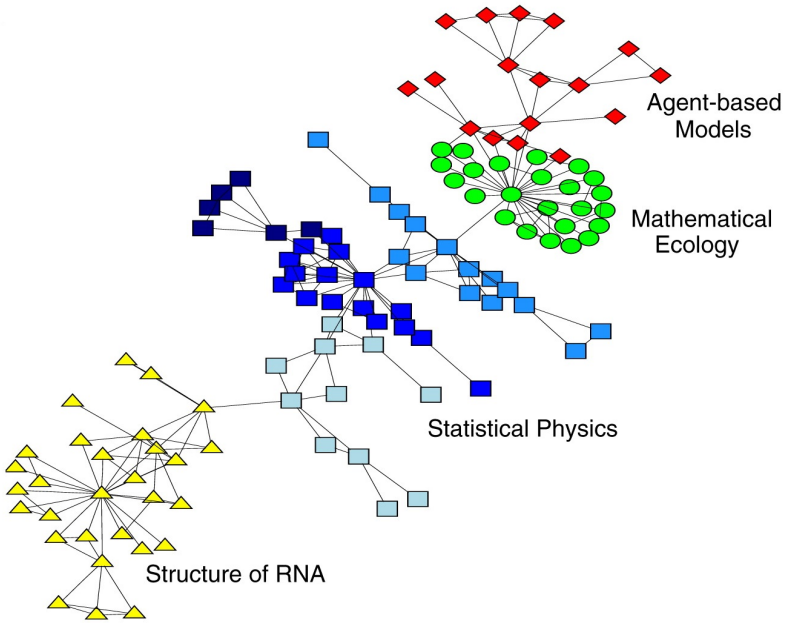
- Networks with community structures arise in many applications



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

Community detection in networks

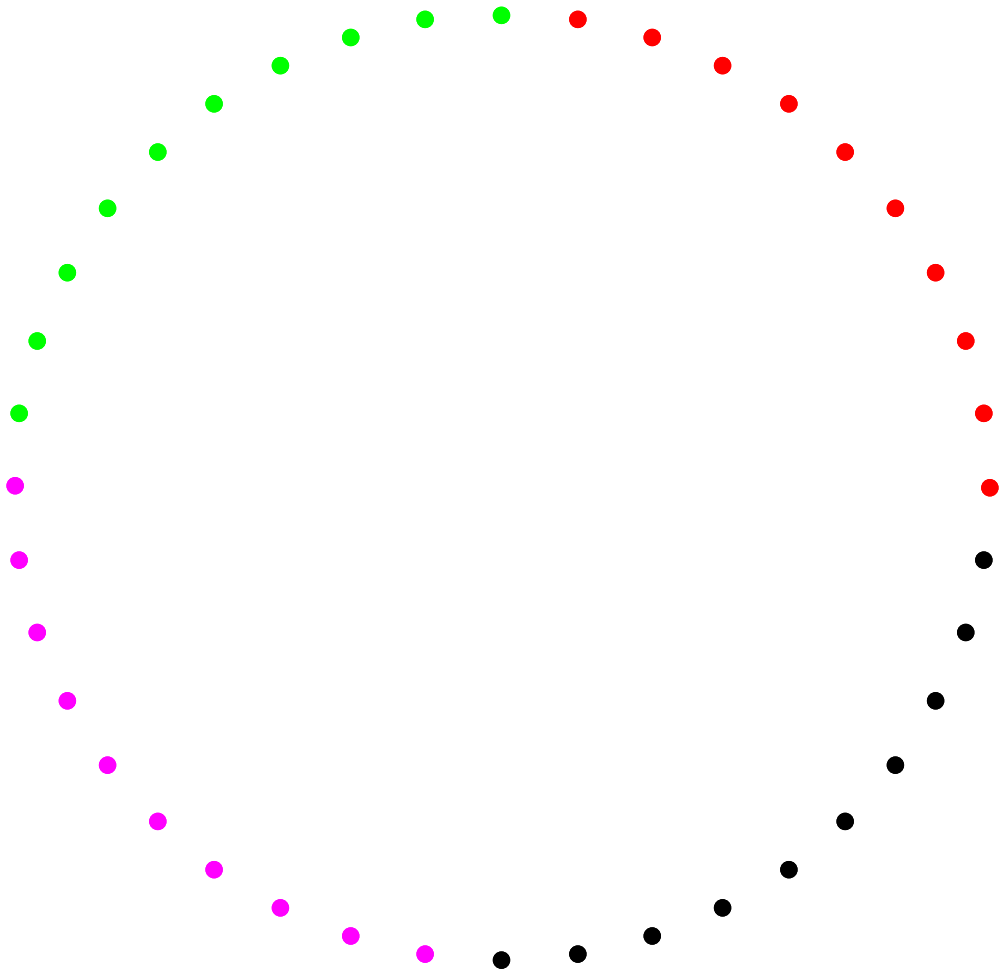
- Networks with community structures arise in many applications



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

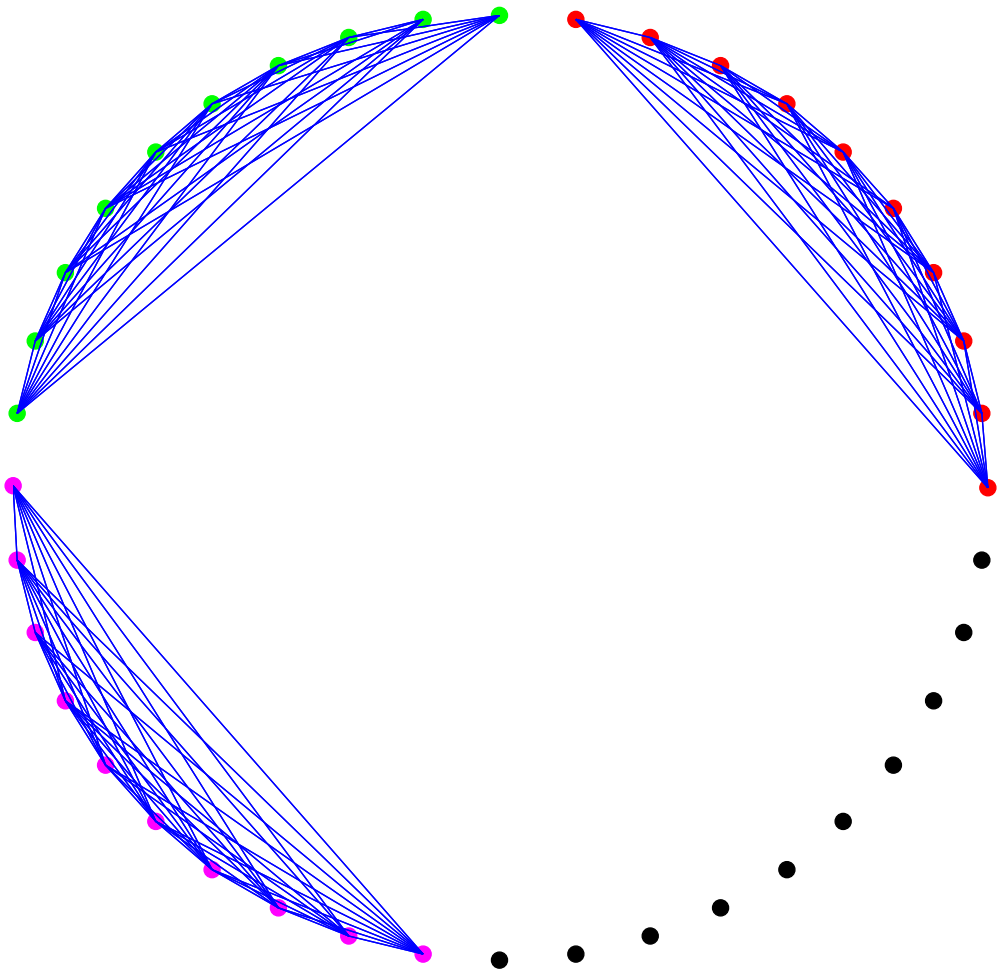
- Task: Discover underlying communities based on the network topology

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



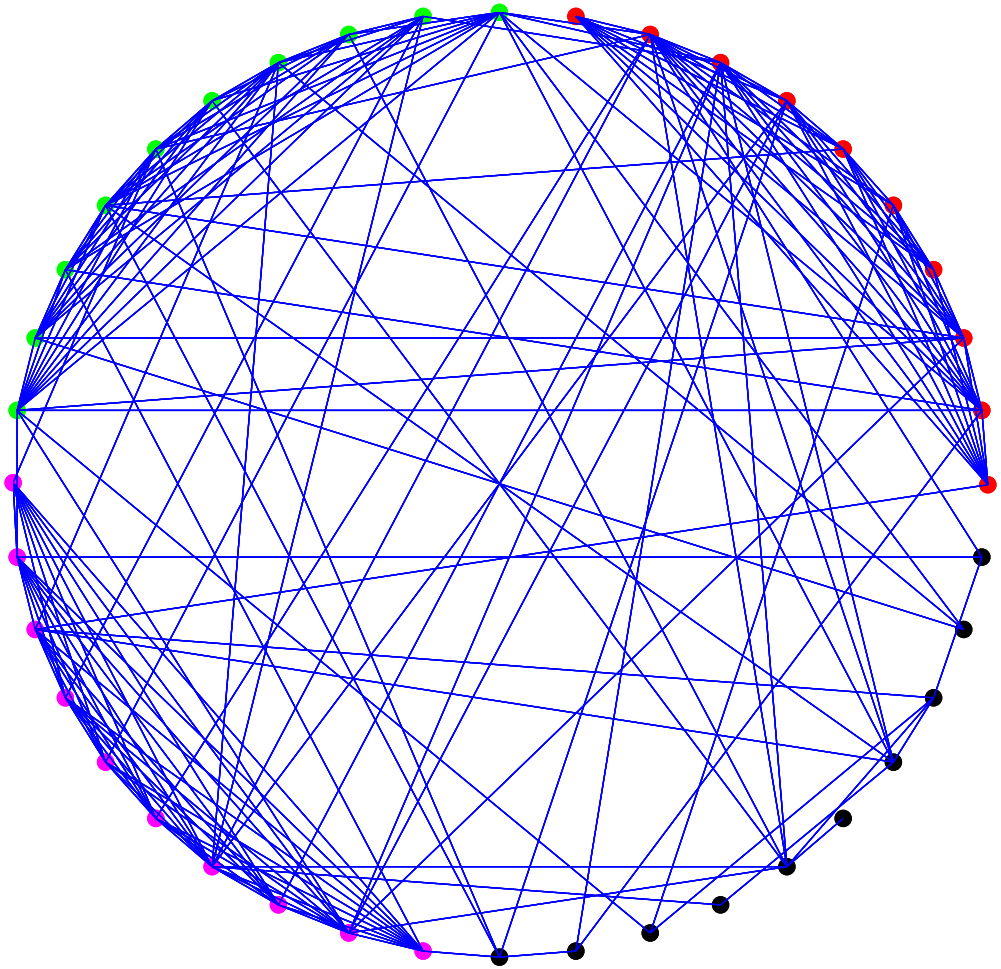
$n = 40,\ K = 10,\ r = 3$

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



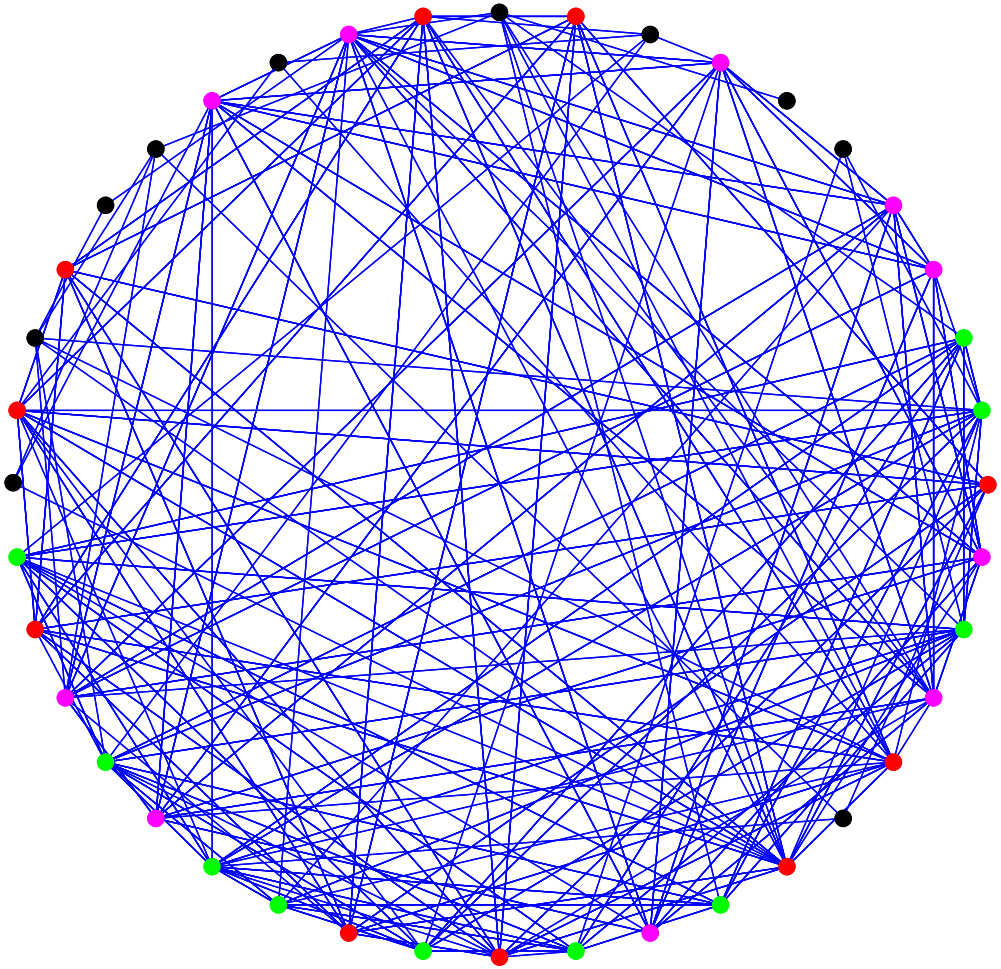
$p = 0.9$

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



$p = 0.9 \qquad q = 0.1$

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



$p = 0.9 \qquad q = 0.1$

Exact recovery

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

- Goal: exact recovery (strong consistency)

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

- Alternatives
  - almost exact recovery (weak consistency): [Mossel-Neeman-Sly '14, Abbe-Sandon '15, Montanari '15]...
  - correlated recovery: [Decelle-Krzakala-Moore-Zdeborova '11, Mossel-Neeman-Sly '12 '13, Massoulie '13]...

Objectives of this talk

- Information limit: When is exact recovery possible (impossible)?

Objectives of this talk

- Information limit: When is exact recovery possible (impossible)?
- Is the information limit achievable in polynomial time, e.g., via semidefinite programming?

remainder of the talk

1. Two equal-sized communities
2. A single community of linear size
3. Extensions and open problems

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. $\begin{cases} p = \frac{a \log n}{n} & \sigma_i^* = \sigma_j^* \\ q = \frac{b \log n}{n} & \sigma_i^* \neq \sigma_j^* \end{cases}$

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. $\begin{cases} p = \frac{a \log n}{n} & \sigma_i^* = \sigma_j^* \\ q = \frac{b \log n}{n} & \sigma_i^* \neq \sigma_j^* \end{cases}$

Remarks

- $a + b > 2$ is the connectivity threshold and necessary for exact recovery

Two equal-sized communities: MLE $\Rightarrow$ SDP relaxation

- Maximum likelihood estimator (MLE): Assume $p \geq q$

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

Two equal-sized communities: MLE $\Rightarrow$ SDP relaxation

- Maximum likelihood estimator (MLE): Assume $p \geq q$

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

$$\xrightarrow{\text{lift: } Y = \sigma\sigma^\top}$$

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

Two equal-sized communities: MLE $\Rightarrow$ SDP relaxation

- Maximum likelihood estimator (MLE): Assume $p \geq q$

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

$$\xrightarrow{\text{lift: } Y = \sigma\sigma^\top}$$

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

Two equal-sized communities: MLE $\Rightarrow$ SDP relaxation

- Maximum likelihood estimator (MLE): Assume $p \geq q$

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

$$\xrightarrow{\text{lift: } Y = \sigma\sigma^\top}$$

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

- Goal: $\mathbb{P}\left\{ \widehat{Y}_{\text{SDP}} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \right\} \rightarrow 1$

Two equal-sized communities: 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.*

Two equal-sized communities: 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-Wu-X. '14)

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

Two equal-sized communities: 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-Wu-X. '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]
- $\mathbb{P}\left\{\widehat{Y}_{\mathrm{SDP}} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\right\} = 1 - n^{-\Omega(1)}$

Two equal-sized communities: Dual certificate argument

Goal: $\mathbb{P}\left\{\widehat{Y}_{\mathrm{SDP}} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\right\} \to 1$

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

Two equal-sized communities: Dual certificate argument

Goal: $\mathbb{P}\left\{\widehat{Y}_{\mathrm{SDP}} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\right\} \to 1$

$$\widehat{Y}_{\mathrm{SDP}} = \arg\max_{Y} \langle A, Y \rangle$$

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

$$Y_{ii} = 1$$

$$\langle \mathbf{J}, Y \rangle = 0$$

dual variables

$$S \succeq 0$$

$$D = \mathrm{diag}\{d_i\}$$

$$\lambda \in \mathbb{R}$$

Two equal-sized communities: Dual certificate argument

Goal: $\mathbb{P}\left\{\widehat{Y}_{\mathrm{SDP}} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\right\} \to 1$

$$\widehat{Y}_{\mathrm{SDP}} = \arg\max_{Y} \langle A, Y \rangle$$

dual variables

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

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

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

- $d_i =$ (# of nbrs in own cluster) $-$ (# of nbrs in other cluster)
  $\sim \mathrm{Binom}(n/2-1, p) - \mathrm{Binom}(n/2, q)$

Two equal-sized communities: Dual certificate argument

Goal: $\mathbb{P}\left\{\widehat{Y}_{\mathrm{SDP}} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\right\} \to 1$

$$\widehat{Y}_{\mathrm{SDP}} = \arg\max_{Y} \langle A, Y \rangle \qquad \text{dual variables}$$

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

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

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

- $d_i =$ (# of nbrs in own cluster) $-$ (# of nbrs in other cluster)
  $\sim \mathrm{Binom}(n/2-1, p) - \mathrm{Binom}(n/2, q)$
- $S = D - A + \lambda \mathbf{J} \succeq 0$ if $\lambda \geq (p+q)/2$ and $\min d_i \geq \|A - \mathbb{E}[A]\|$

Two equal-sized communities: Dual certificate argument

Goal: $\mathbb{P}\left\{\widehat{Y}_{\mathrm{SDP}} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\right\} \to 1$

$$\widehat{Y}_{\mathrm{SDP}} = \arg\max_{Y} \langle A, Y \rangle \qquad \text{dual variables}$$

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

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

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

- $d_i =$ (# of nbrs in own cluster) $-$ (# of nbrs in other cluster)
  $\sim \mathrm{Binom}(n/2-1, p) - \mathrm{Binom}(n/2, q)$
- $S = D - A + \lambda \mathbf{J} \succeq 0$ if $\lambda \geq (p+q)/2$ and $\min d_i \geq \|A - \mathbb{E}[A]\|$
- $\min d_i = \Omega_P(\log n)$ if $\sqrt{a} - \sqrt{b} > \sqrt{2}$

Two equal-sized communities: Dual certificate argument

Goal: $\mathbb{P}\left\{\widehat{Y}_{\mathrm{SDP}} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\right\} \to 1$

$$\widehat{Y}_{\mathrm{SDP}} = \arg\max_{Y} \langle A, Y \rangle \qquad \text{dual variables}$$

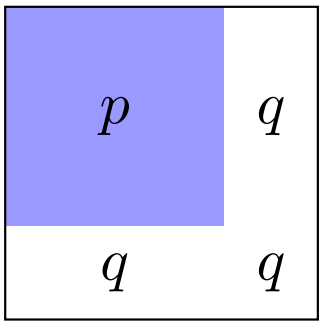
$$\text{s.t.} \quad Y \succeq 0 \qquad S \succeq 0$$

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

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

- $d_i =$ (# of nbrs in own cluster) $-$ (# of nbrs in other cluster)
  $\sim \mathrm{Binom}(n/2-1, p) - \mathrm{Binom}(n/2, q)$
- $S = D - A + \lambda \mathbf{J} \succeq 0$ if $\lambda \geq (p+q)/2$ and $\min d_i \geq \|A - \mathbb{E}[A]\|$
- $\min d_i = \Omega_P(\log n)$ if $\sqrt{a} - \sqrt{b} > \sqrt{2}$
- $\|A - \mathbb{E}[A]\| = O_P(\sqrt{\log n})$: 2nd-order stochastic dominance [Tomozei-Massoulié '14] + result for iid matrix [Seginer '00]

A single community: Planted dense subgraph model



- One cluster of size K plus $n - K$ outliers
- Connectivity p within cluster and q otherwise
- Linear community size: $K = \rho\, n$
- Relatively sparse graph: $p = \frac{a \log n}{n}$ and $q = \frac{b \log n}{n}$

A single community: Optimal recovery via SDP

Theorem (Hajek-Wu-X. '14)

- *If* $\rho f(a,b) > 1$, *recovery is achievable via SDP in polynomial-time.*
- *If* $\rho f(a,b) < 1$, *recovery is impossible.*

A single community: Optimal recovery via SDP

Theorem (Hajek-Wu-X. '14)

- *If* $\rho f(a,b) > 1$, *recovery is achievable via SDP in polynomial-time.*
- *If* $\rho f(a,b) < 1$, *recovery is impossible.*

Remarks

- $f(a,b) = a - \tau^* \log \frac{\mathrm{e}a}{\tau^*}$ with $\tau^* = \frac{a-b}{\log a - \log b}$
- Sufficiency: dual certificate argument
- Necessity: show MLE fails by swapping an in-cluster node with an out-cluster node

Extensions and open problems

SDP achieves sharp threshold:

- Two unequal-sized clusters $(\rho n, (1-\rho)n)$: $\eta(\rho, a, b) > 1$

Extensions and open problems

SDP achieves sharp threshold:

- Two unequal-sized clusters $(\rho n, (1-\rho)n)$: $\eta(\rho, a, b) > 1$
- Two clusters with unknown sizes: $\sqrt{a} - \sqrt{b} > \sqrt{2}$

Extensions and open problems

SDP achieves sharp threshold:

- Two unequal-sized clusters $(\rho n, (1-\rho)n)$: $\eta(\rho, a, b) > 1$
- Two clusters with unknown sizes: $\sqrt{a} - \sqrt{b} > \sqrt{2}$
- r equal-sized clusters: $\sqrt{a} - \sqrt{b} > \sqrt{r}$

Extensions and open problems

SDP achieves sharp threshold:

- Two unequal-sized clusters $(\rho n, (1-\rho)n)$: $\eta(\rho, a, b) > 1$
- Two clusters with unknown sizes: $\sqrt{a} - \sqrt{b} > \sqrt{2}$
- r equal-sized clusters: $\sqrt{a} - \sqrt{b} > \sqrt{r}$

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

Concluding remarks

- If community sizes are linear, information limit is attainable in polynomial-time via SDP
- If community sizes scale as n^β for $\beta < 1$, information limit might not be achievable in polynomial-time [Hajek-Wu-X. '14]

Concluding remarks

- If community sizes are linear, information limit is attainable in polynomial-time via SDP
- If community sizes scale as n^β for $\beta < 1$, information limit might not be achievable in polynomial-time [Hajek-Wu-X. '14]

References

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

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 $n \ge 1$,

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

A single community: MLE $\Leftrightarrow$ Densest K-subgraph

Assuming $p > q$ and $\xi =$ cluster indicator

- Maximum likelihood estimator (MLE)

$$\begin{aligned} \max_{\xi} \ & \sum_{i,j} A_{ij}\xi_i\xi_j \\ \text{s.t.} \ & \xi \in \{0,1\}^n \\ & \xi^\top \mathbf{1} = K \end{aligned}$$

A single community: MLE $\Leftrightarrow$ Densest K-subgraph

Assuming $p > q$ and $\xi =$ cluster indicator

- Maximum likelihood estimator (MLE)

$$\max_{\xi} \sum_{i,j} A_{ij}\xi_i\xi_j$$
$$\text{s.t.} \quad \xi \in \{0,1\}^n$$
$$\xi^\top \mathbf{1} = K$$

$\xleftrightarrow{\text{lift: } Z=\xi\xi^\top}$

$$\max_{Z} \langle A, Z\rangle$$
$$\text{s.t.} \quad \text{rank}(Z) = 1$$
$$Z_{ii} \le 1 \quad \forall i \in [n]$$
$$Z_{ij} \ge 0 \quad \forall i,j \in [n]$$
$$\langle \mathbf{I}, Z\rangle = K$$
$$\langle \mathbf{J}, Z\rangle = K^2$$

A single community: SDP relaxation

- Semidefinite programming (SDP) relaxation of MLE

$$\widehat{Z}_{\mathrm{SDP}} = \arg\max_{Z} \langle A, Z \rangle$$

$$\begin{aligned} \text{s.t.} \quad & Z \succeq 0 \\ & Z_{ii} \le 1, \quad \forall i \in [n] \\ & Z_{ij} \ge 0, \quad \forall i, j \in [n] \\ & \langle \mathbf{I}, Z \rangle = K \\ & \langle \mathbf{J}, Z \rangle = K^2 \end{aligned}$$

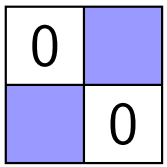
- goal: $\mathbb{P}\left\{ \widehat{Z}_{\mathrm{SDP}} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \right\} \to 1$

A single community: Dual certificate

$$
\begin{aligned}
\max_{Z} \ & \langle A, Z \rangle \\
\text{s.t.} \ & Z \succeq 0 \\
& Z_{ii} \leq 1 \\
& Z_{ij} \geq 0 \\
& \langle \mathbf{I}, Z \rangle = K \\
& \langle \mathbf{J}, Z \rangle = K^2
\end{aligned}
$$

A single community: Dual certificate

$$\begin{aligned} \max_{Z} \ & \langle A, Z \rangle && \text{dual variables} \\ \text{s.t.} \ & Z \succeq 0 && S \succeq 0 \\ & Z_{ii} \leq 1 && D = \text{diag}\{d_i\} \\ & Z_{ij} \geq 0 && B \geq 0 \\ & \langle \mathbf{I}, Z \rangle = K && \eta \in \mathbb{R} \\ & \langle \mathbf{J}, Z \rangle = K^2 && \lambda \in \mathbb{R} \end{aligned}$$

- $S\xi^* = 0 \Rightarrow d_i = e(i, C^*) - \lambda K - \eta$ if $i \in C^*$; $d_i = 0$ otherwise.
- $B =$ 

, where  $= b\mathbf{1}^\top$, with $b_i = \lambda - e(i, C^*)/K$ for $i \notin C^*$
- Set $\lambda = \tau^* \log n/n$ so that $\min_{i \notin C^*} b_i \geq 0$
- Set $\eta = ||A - \mathbb{E}[A]||$ such that $\lambda_2(S) > 0$ if $d_i \geq 0$
- $\min_{i \in C^*} e(i, C^*) - \lambda K = \Omega_P(\log n)$ and $\eta = O_P(\sqrt{\log n})$

A single community: Necessity

$$\rho f(a,b) < 1$$

$$\Rightarrow \quad \underbrace{\min_{i \in C^*} e(i, C^*)}_{K \text{ almost ind. } \mathrm{Binom}(K-1,\, p)} \quad < \quad \underbrace{\max_{j \notin C^*} e(i, C^*)}_{n-K \text{ ind. } \mathrm{Binom}(K,\, q)} \qquad \text{w.h.p.}$$

$\Rightarrow \exists i \in C^*, j \notin C^*$ such that swapping i, j increases the cluster density

$\Rightarrow$ MLE fails