Achieving Exact Cluster Recovery Threshold via Semidefinite Programming

Jiaming Xu

Department of Statistics, The Wharton School
University of Pennsylvania

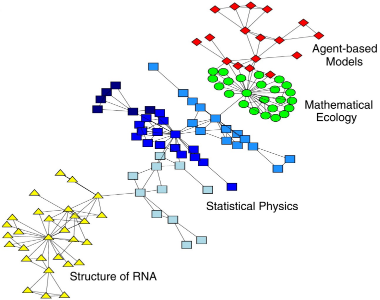
jiamingx@wharton.upenn.edu

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

November 10, 2015

Community detection in networks

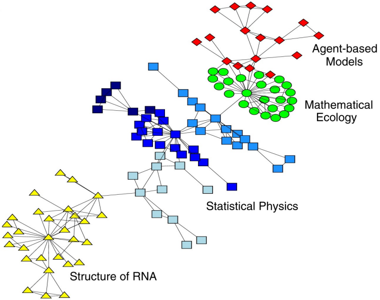
- Networks with community structures arise in many applications



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

Community detection in networks

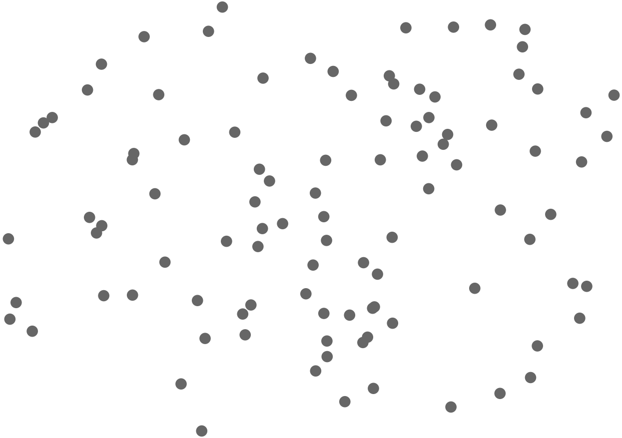
- Networks with community structures arise in many applications



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

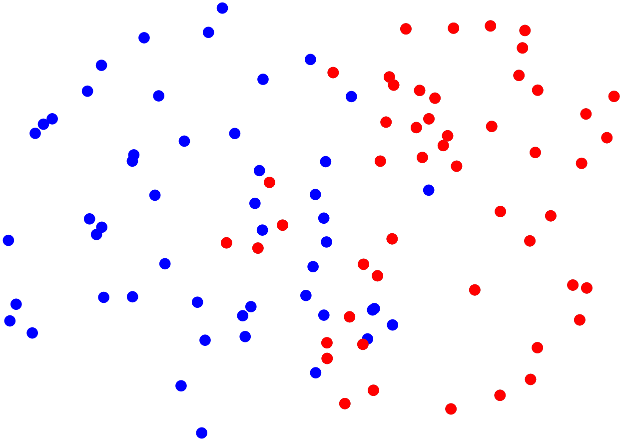
- Task: Discover underlying communities based on the network topology

Stochastic block model [Holland-Laskey-Leinhardt '83] Planted partition model [Condon-Karp '01]



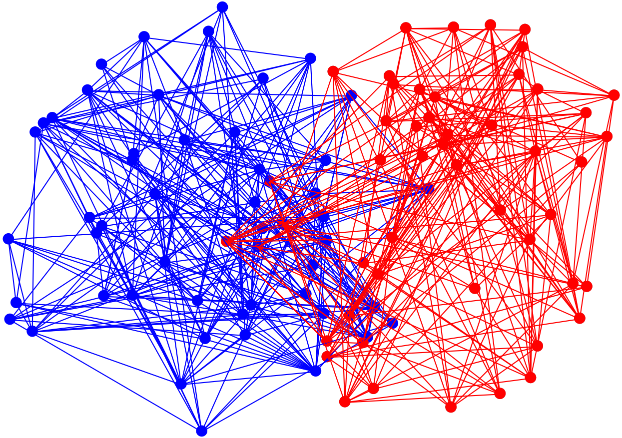
Stochastic block model [Holland-Laskey-Leinhardt '83]
Planted partition model [Condon-Karp '01]

1. n nodes are randomly partitioned into r equal-sized communities



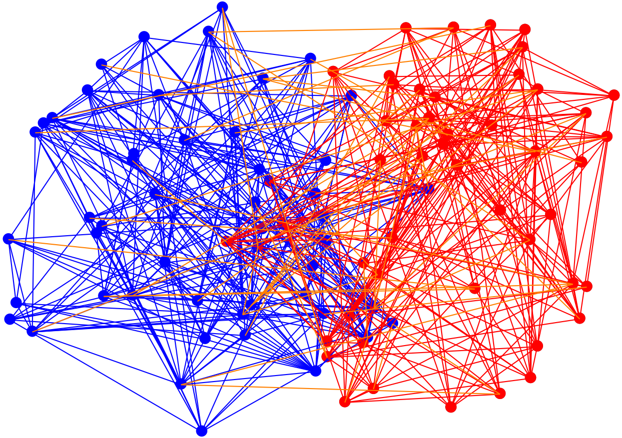
Stochastic block model [Holland-Laskey-Leinhardt '83] Planted partition model [Condon-Karp '01]

1. n nodes are randomly partitioned into r equal-sized communities
2. For every pair of nodes in same community, add an edge w.p. p



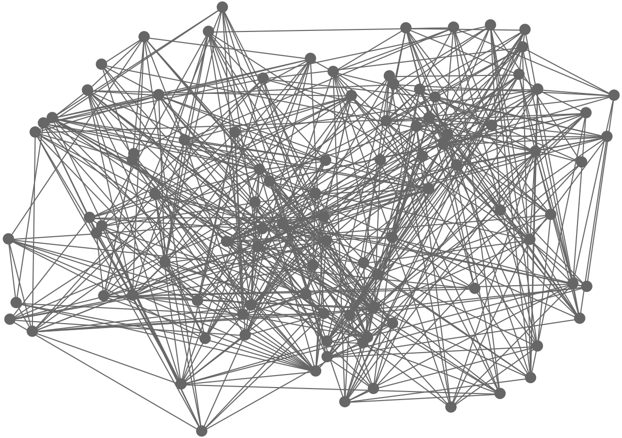
Stochastic block model [Holland-Laskey-Leinhardt '83] Planted partition model [Condon-Karp '01]

1. n nodes are randomly partitioned into r equal-sized communities
2. For every pair of nodes in same community, add an edge w.p. p
3. For every pair of nodes in diff. community, add an edge w.p. q



Stochastic block model [Holland-Laskey-Leinhardt '83] Planted partition model [Condon-Karp '01]

1. n nodes are randomly partitioned into r equal-sized communities
2. For every pair of nodes in same community, add an edge w.p. p
3. For every pair of nodes in diff. community, add an edge w.p. q



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, Zhang-Zhou'15, Yun-Proutiere '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)?
- Is the information limit achievable in polynomial time, e.g., via semidefinite programming?

Remainder of the talk

1. r equal-sized communities
2. Two unequal-sized communities
3. Extensions and open problems

r equal-sized communities

Model:

- n nodes partitioned into r clusters of size K $(n = rK)$.
- $i \sim j$ independently w.p. $\begin{cases} p = \frac{a \log n}{n} & \text{if } i \text{ and } j \text{ in same cluster} \\ q = \frac{b \log n}{n} & \text{otherwise.} \end{cases}$

r equal-sized communities

Model:

- n nodes partitioned into r clusters of size K $(n = rK)$.
- $i \sim j$ independently w.p. $\begin{cases} p = \frac{a \log n}{n} & \text{if } i \text{ and } j \text{ in same cluster} \\ q = \frac{b \log n}{n} & \text{otherwise.} \end{cases}$

Remarks

- $\frac{a+(r-1)b}{r} > 1$ is the connectivity threshold and necessary for exact recovery

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

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

$$\max_{\xi} \langle A, \sum_{k=1}^{r} \xi_k \xi_k^\top \rangle \quad \to \text{\# of in-cluster edges}$$

$$\text{s.t.} \quad \xi_k \in \{0,1\}^n, \quad \xi_k^\top \mathbf{1} = K, \quad \xi_k^\top \xi_{k'} = 0, k \neq k'$$

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

$$\max_{\xi} \langle A, \sum_{k=1}^{r} \xi_k \xi_k^\top \rangle \quad \to \text{\# of in-cluster edges}$$

$$\text{s.t.} \quad \xi_k \in \{0,1\}^n, \quad \xi_k^\top \mathbf{1} = K, \quad \xi_k^\top \xi_{k'} = 0, k \neq k'$$

$$\xRightarrow{Y = \sum_{k=1}^{r} \xi_k \xi_k^\top} \qquad \begin{aligned} \max_{Y} \ & \langle A, Y \rangle \\ \text{s.t.} \ & \text{rank}(Y) = r \\ & Y_{ii} = 1 \\ & Y_{ij} \geq 0, \quad \sum_j Y_{ij} = K \end{aligned}$$

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

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

$$\max_{\xi} \langle A, \sum_{k=1}^{r} \xi_k \xi_k^\top \rangle \quad \to \text{\# of in-cluster edges}$$

$$\text{s.t.} \quad \xi_k \in \{0,1\}^n, \quad \xi_k^\top \mathbf{1} = K, \quad \xi_k^\top \xi_{k'} = 0, k \neq k'$$

$$\xrightarrow{Y = \sum_{k=1}^{r} \xi_k \xi_k^\top} \quad \max_{Y} \langle A, Y \rangle$$

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

$$Y_{ii} = 1$$

$$Y_{ij} \geq 0, \quad \sum_j Y_{ij} = K$$

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

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

$$\max_{\xi} \langle A, \sum_{k=1}^{r} \xi_k \xi_k^\top \rangle \quad \rightarrow \text{\# of in-cluster edges}$$

$$\text{s.t.} \quad \xi_k \in \{0,1\}^n, \quad \xi_k^\top \mathbf{1} = K, \quad \xi_k^\top \xi_{k'} = 0, k \neq k'$$

$$\xrightarrow{Y = \sum_{k=1}^{r} \xi_k \xi_k^\top} \quad \max_{Y} \langle A, Y \rangle$$

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

$$Y_{ii} = 1$$

$$Y_{ij} \geq 0, \quad \sum_j Y_{ij} = K$$

- Goal: $\mathbb{P}\left\{ \widehat{Y}_{\text{SDP}} = \begin{bmatrix} 1 & & & 0 \\ & 1 & & \\ & & 1 & \\ 0 & & & 1 \end{bmatrix} \right\} \rightarrow 1$

r equal-sized communities: optimal recovery via SDP

Theorem ([Hajek-Wu-X. '15])

SDP achieves the optimal recovery threshold $\sqrt{a} - \sqrt{b} > \sqrt{r}$.

r equal-sized communities: optimal recovery via SDP

Theorem ([Hajek-Wu-X. '15])

SDP achieves the optimal recovery threshold $\sqrt{a} - \sqrt{b} > \sqrt{r}$*.*

Remarks

- The case of $r = 2$: conjectured in [Abbe-Bandeira-Hall '14]; resolved by [Hajek-Wu-X. '14] and independently by [Bandeira '15]

r equal-sized communities: optimal recovery via SDP

Theorem ([Hajek-Wu-X. '15])

SDP achieves the optimal recovery threshold $\sqrt{a} - \sqrt{b} > \sqrt{r}$.

Remarks

- The case of $r = 2$: conjectured in [Abbe-Bandeira-Hall '14]; resolved by [Hajek-Wu-X. '14] and independently by [Bandeira '15]
- Extended to $r = o(\log n)$ in [Agarwal-Bandeira-Koiliaris-Kolla '15]

r equal-sized communities: optimal recovery via SDP

Theorem ([Hajek-Wu-X. '15])

SDP achieves the optimal recovery threshold $\sqrt{a} - \sqrt{b} > \sqrt{r}$.

Remarks

- The case of $r = 2$: conjectured in [Abbe-Bandeira-Hall '14]; resolved by [Hajek-Wu-X. '14] and independently by [Bandeira '15]
- Extended to $r = o(\log n)$ in [Agarwal-Bandeira-Koiliaris-Kolla '15]

r equal-sized communities: optimal recovery via SDP

Theorem ([Hajek-Wu-X. '15])

SDP achieves the optimal recovery threshold $\sqrt{a} - \sqrt{b} > \sqrt{r}$.

Remarks

- The case of $r = 2$: conjectured in [Abbe-Bandeira-Hall '14]; resolved by [Hajek-Wu-X. '14] and independently by [Bandeira '15]
- Extended to $r = o(\log n)$ in [Agarwal-Bandeira-Koiliaris-Kolla '15]
- The optimal recovery threshold is also obtained independently in [Yun-Proutiere '14] and [Abbe-Sandon '15] using two-step procedures

r equal-sized communities: dual certificate argument

Goal: $\mathbb{P}\left\{ \widehat{Y}_{\rm SDP} = \begin{bmatrix} 1 & & & 0 \\ & 1 & & \\ & & 1 & \\ 0 & & & 1 \end{bmatrix} \right\} \to 1$

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

r equal-sized communities: dual certificate argument

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

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

dual variables

$S \succeq 0$

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

$B \geq 0$

$\lambda_i \in \mathbb{R}$

r equal-sized communities: dual certificate argument

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

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

dual variables

s.t.	$Y \succeq 0$	$S \succeq 0$
	$Y_{ii} = 1$	$D = \text{diag}\{d_i\}$
	$Y_{ij} \geq 0$	$B \geq 0$
	$\sum_j Y_{ij} = K$	$\lambda_i \in \mathbb{R}$

- $S\,\xi_k = 0 \Rightarrow d_i = e(i, C_k) - \lambda_i K - \sum_{j \in C_k} \lambda_j$ for $i \in C_k$

r equal-sized communities: dual certificate argument

Goal: $\mathbb{P}\left\{ \widehat{Y}_{\rm SDP} = \begin{bmatrix} 1 & & & 0 \\ & 1 & & \\ & & 1 & \\ 0 & & & 1 \end{bmatrix} \right\} \to 1$

$\widehat{Y}_{\rm SDP} = \arg\max_{Y} \langle A, Y \rangle$	dual variables
s.t. $Y \succeq 0$	$S \succeq 0$
$Y_{ii} = 1$	$D =$ diag $\{d_i\}$
$Y_{ij} \geq 0$	$B \geq 0$
$\sum_j Y_{ij} = K$	$\lambda_i \in \mathbb{R}$

- $S\, \xi_k = 0 \Rightarrow d_i = e(i, C_k) - \lambda_i K - \sum_{j \in C_k} \lambda_j$ for $i \in C_k$
- $\lambda_i = \max_{k' \neq k} e(i, C_{k'})/K - q/2 + \sqrt{\log n}/(2K)$ for $i \in C_k$

r equal-sized communities: dual certificate argument

Goal: $\mathbb{P}\left\{ \widehat{Y}_{\rm SDP} = \begin{bmatrix} 1 & & & 0 \\ & 1 & & \\ & & 1 & \\ 0 & & & 1 \end{bmatrix} \right\} \to 1$

$\widehat{Y}_{\rm SDP} = \arg\max_{Y} \langle A, Y \rangle$	dual variables
s.t. $Y \succeq 0$	$S \succeq 0$
$Y_{ii} = 1$	$D = \text{diag}\{d_i\}$
$Y_{ij} \geq 0$	$B \geq 0$
$\sum_j Y_{ij} = K$	$\lambda_i \in \mathbb{R}$

- $S\,\xi_k = 0 \Rightarrow d_i = e(i, C_k) - \lambda_i K - \sum_{j \in C_k} \lambda_j$ for $i \in C_k$
- $\lambda_i = \max_{k' \neq k} e(i, C_{k'})/K - q/2 + \sqrt{\log n}/(2K)$ for $i \in C_k$
- $B_{C_k \times C_k} = \mathbf{0}$ and $B_{C_k \times C_{k'}} = y_{kk'} \mathbf{1}^\top + \mathbf{1} z_{kk'}^\top$ rank 2 !

r equal-sized communities: dual certificate argument

Goal: $\mathbb{P}\left\{ \widehat{Y}_{\rm SDP} = \begin{bmatrix} 1 & & 0 & \\ & 1 & & \\ 0 & & 1 & \\ & & & 1 \end{bmatrix} \right\} \to 1$

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

dual variables

s.t.	$Y \succeq 0$	$S \succeq 0$
	$Y_{ii} = 1$	$D = \text{diag}\{d_i\}$
	$Y_{ij} \geq 0$	$B \geq 0$
	$\sum_j Y_{ij} = K$	$\lambda_i \in \mathbb{R}$

- $S\,\xi_k = 0 \Rightarrow d_i = e(i, C_k) - \lambda_i K - \sum_{j \in C_k} \lambda_j$ for $i \in C_k$
- $\lambda_i = \max_{k' \neq k} e(i, C_{k'})/K - q/2 + \sqrt{\log n}/(2K)$ for $i \in C_k$
- $B_{C_k \times C_k} = \mathbf{0}$ and $B_{C_k \times C_{k'}} = y_{kk'} \mathbf{1}^\top + \mathbf{1} z_{kk'}^\top$ rank 2 !
- $S = D - A - B + \lambda \mathbf{1}^\top + \mathbf{1} \lambda^\top \succeq 0$ if $\min d_i \geq \|A - \mathbb{E}[A]\|$

r equal-sized communities: dual certificate argument

Goal: $\mathbb{P}\left\{ \widehat{Y}_{\rm SDP} = \begin{bmatrix} 1 & & & 0 \\ & 1 & & \\ & & 1 & \\ 0 & & & 1 \end{bmatrix} \right\} \to 1$

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

	dual variables
s.t. $Y \succeq 0$	$S \succeq 0$
$Y_{ii} = 1$	$D = \text{diag}\{d_i\}$
$Y_{ij} \geq 0$	$B \geq 0$
$\sum_j Y_{ij} = K$	$\lambda_i \in \mathbb{R}$

- $S\,\xi_k = 0 \Rightarrow d_i = e(i, C_k) - \lambda_i K - \sum_{j \in C_k} \lambda_j$ for $i \in C_k$
- $\lambda_i = \max_{k' \neq k} e(i, C_{k'})/K - q/2 + \sqrt{\log n}/(2K)$ for $i \in C_k$
- $B_{C_k \times C_k} = \mathbf{0}$ and $B_{C_k \times C_{k'}} = y_{kk'} \mathbf{1}^\top + \mathbf{1} z_{kk'}^\top$ rank 2 !
- $S = D - A - B + \lambda \mathbf{1}^\top + \mathbf{1} \lambda^\top \succeq 0$ if $\min d_i \geq \| A - \mathbb{E}[A] \|$
- $\min d_i = \Omega_P(\log n)$ if $\sqrt{a} - \sqrt{b} > \sqrt{r}$; $\| A - \mathbb{E}[A] \| = O_P(\sqrt{\log n})$

r equal-sized communities: necessity

Genie argument: reveal the membership of all clusters except cluster 1 and 2

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

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

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

$\Rightarrow$ MLE fails

r equal-sized communities: necessity

Genie argument: reveal the membership of all clusters except cluster 1 and 2

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

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

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

$\Rightarrow$ MLE fails

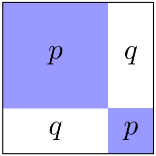
Sharp threshold

- $\sqrt{a} - \sqrt{b} > \sqrt{r} \Rightarrow \min d_i = \Omega(\log n) \Rightarrow$ 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 $n-K$ ($K = \rho n$):



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

Two unequal-sized clusters: known size

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

p	q
q	p

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

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

$$Y_{ii} = 1, \quad i \in [n]$$

$$\langle \mathbf{J}, Y\rangle = (2K-n)^2$$

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

Two unequal-sized clusters: known size

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

$$\begin{pmatrix} p & q \\ q & p \end{pmatrix}$$

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

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

$$Y_{ii} = 1, \quad i \in [n]$$

$$\langle \mathbf{J}, Y\rangle = (2K-n)^2$$

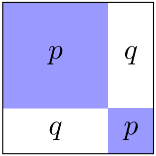
Goal: $\mathbb{P}\left\{ \widehat{Y}_{\mathrm{SDP}} = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \right\} \to 1$

Theorem ([Hajek-Wu-X. '15])

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

Two unequal-sized clusters: known size

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



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

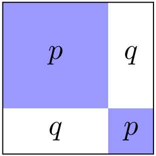
Goal: $\mathbb{P}\left\{ \widehat{Y}_{\rm SDP} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \right\} \to 1$

Theorem ([Hajek-Wu-X. '15])

SDP achieves the optimal recover 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

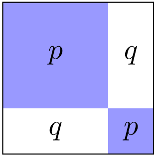


Two clusters of size K and $n-K$ ($K = 0, 1, \ldots, n$):

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

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 $n-K$ ($K = 0, 1, \ldots, n$):

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

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.

Extensions and open problems

Extensions: r unequal-sized clusters $(\rho_1 n, \ldots, \rho_r n)$

- SDP sufficient condition and information necessary bound is within a factor of 4 [Hajek-Wu-X. '15]

Extensions and open problems

Extensions: r unequal-sized clusters $(\rho_1 n, \ldots, \rho_r n)$

- SDP sufficient condition and information necessary bound is within a factor of 4 [Hajek-Wu-X. '15]
- Optimality of SDP relaxation is proved in [Perry-Wein '15]

Extensions and open problems

Extensions: r unequal-sized clusters $(\rho_1 n, \ldots, \rho_r n)$

- SDP sufficient condition and information necessary bound is within a factor of 4 [Hajek-Wu-X. '15]
- Optimality of SDP relaxation is proved in [Perry-Wein '15]

Open problem:

- Optimality of SDP relaxation in the general SBM remains open
- Sharp threshold for the general SBM 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]
- What happens if community sizes scale as $\frac{n}{\log n}$?

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]
- What happens if community sizes scale as $\frac{n}{\log n}$?

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)
- B. Hajek, Y. Wu & J. X. (2015). *Semidefinite programs for exact recovery of a hidden community*. arXiv:1511.xxxx