Exact Recovery Threshold in the Binary Censored Block Model

Jiaming Xu

Department of Statistics, The Wharton School
University of Pennsylvania

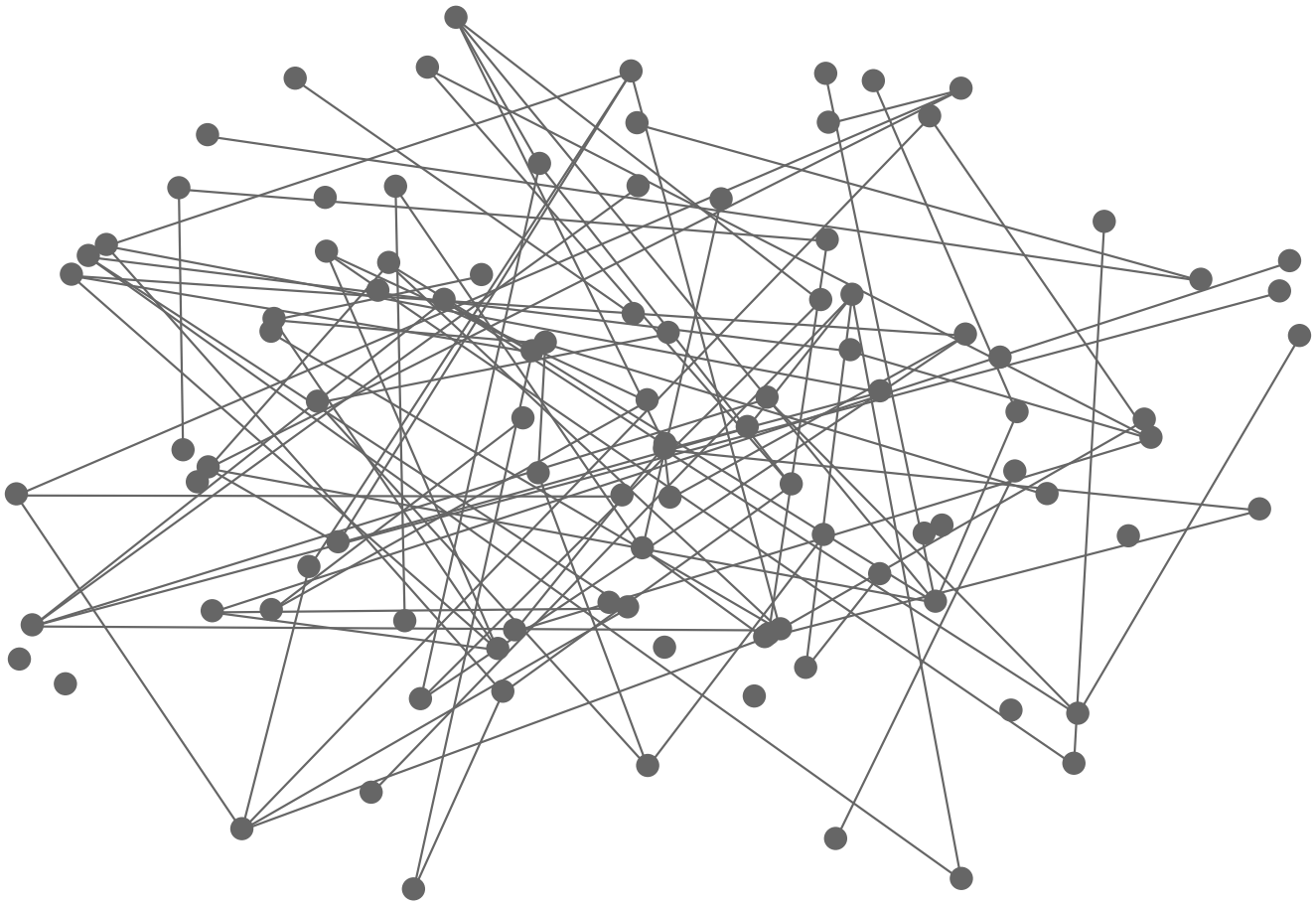
jiamingx@wharton.upenn.edu

Joint work with Bruce Hajek and Yihong Wu

October 13, 2015

Binary censored block model

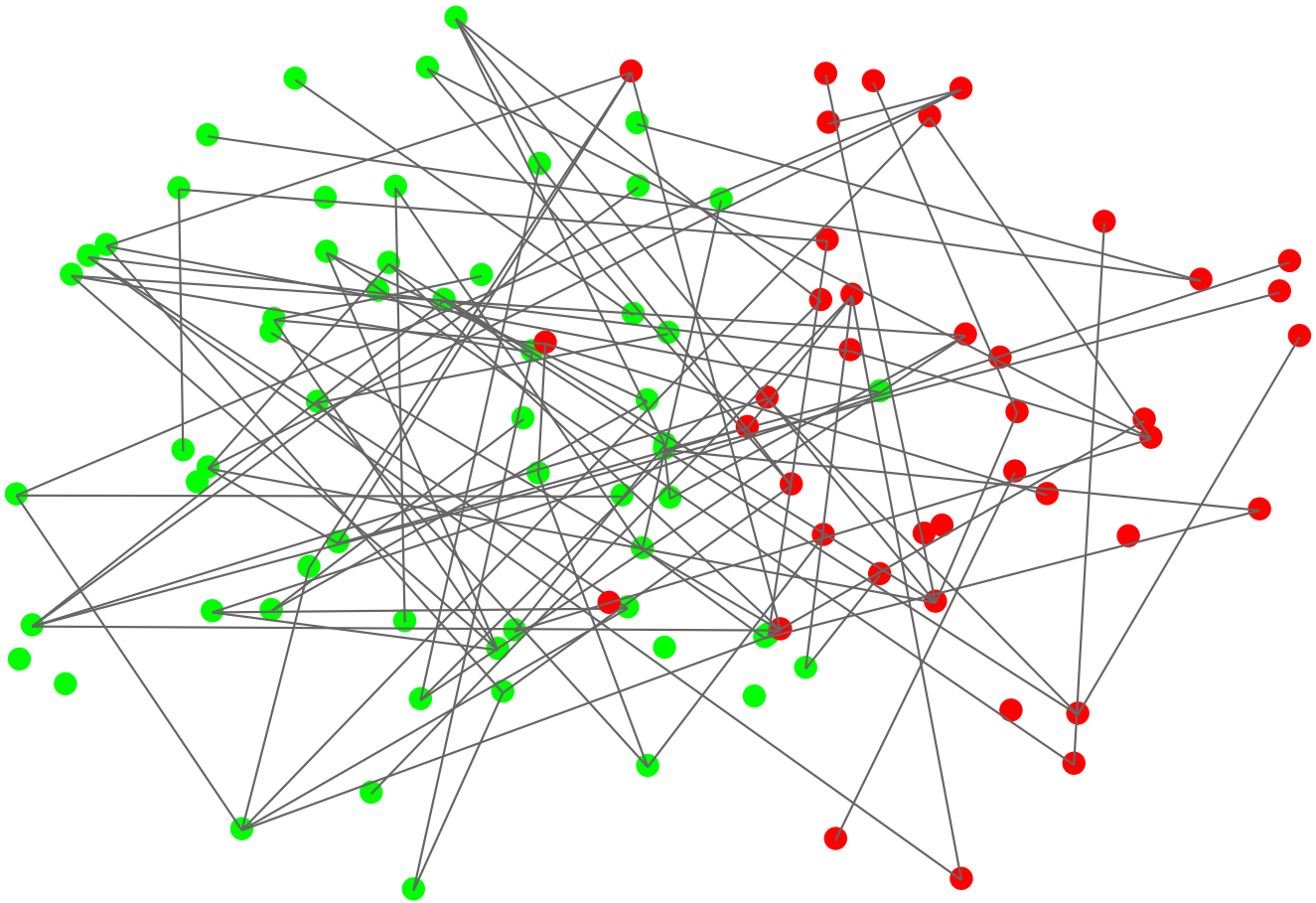
$G = ([n], E)$ and $\epsilon \in [0, 1/2]$



Binary censored block model

$$G = ([n], E) \text{ and } \epsilon \in [0, 1/2]$$

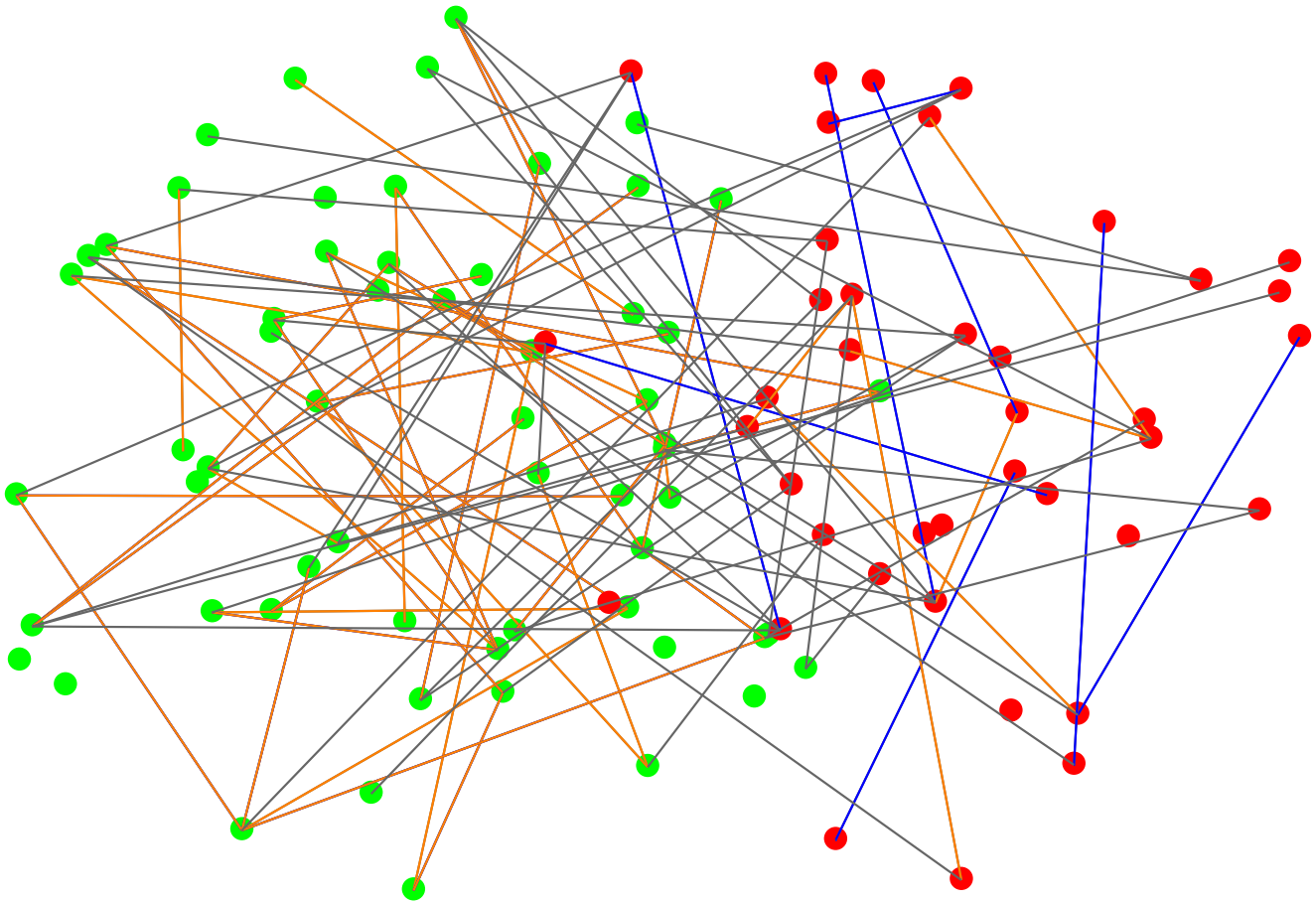
1. Color the vertices in green or red arbitrarily



Binary censored block model

$$G = ([n], E) \text{ and } \epsilon \in [0, 1/2]$$

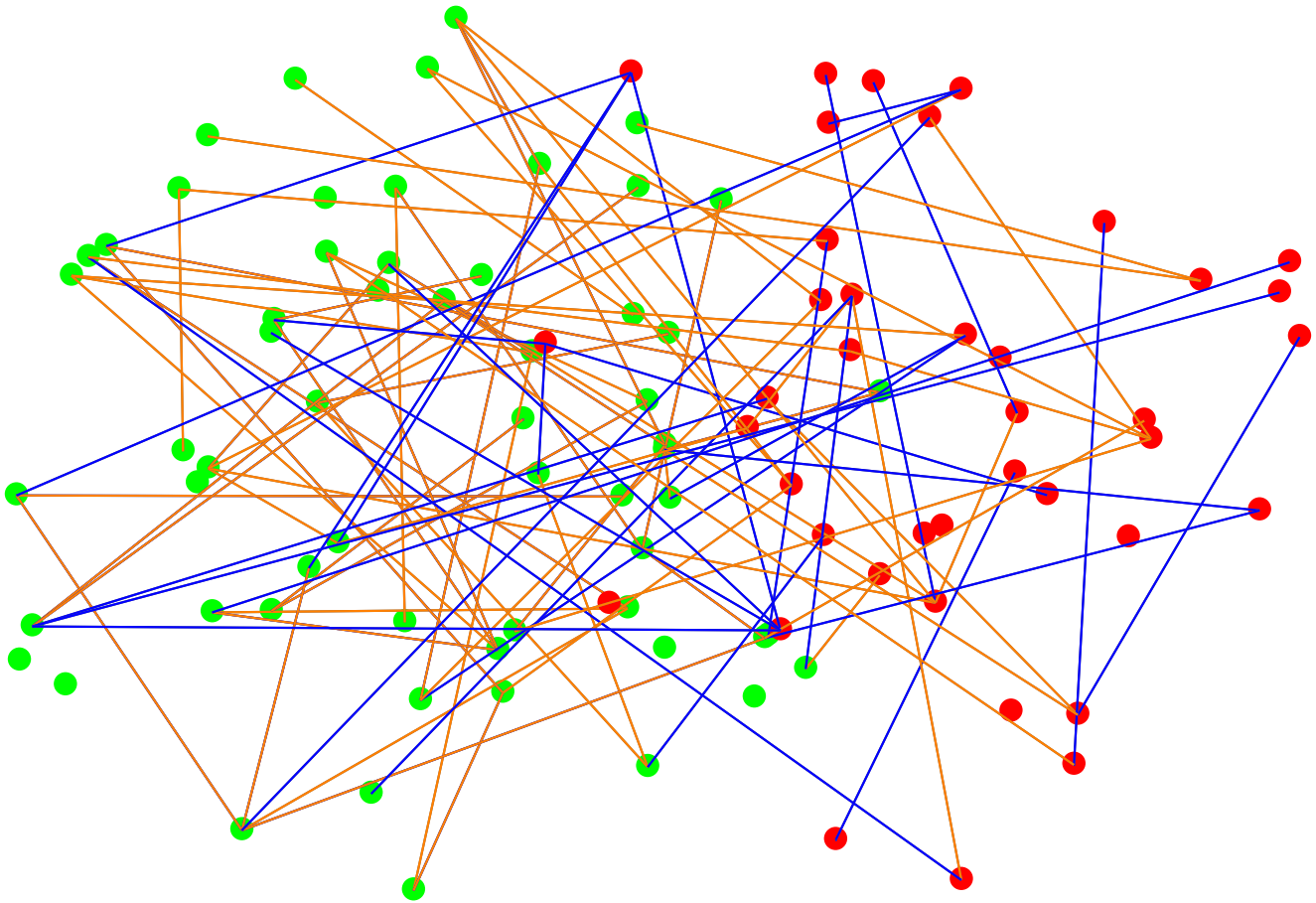
1. Color the vertices in green or red arbitrarily
2. If endpoints in same color, color edge in blue (orange) w.p. $1 - \epsilon$ (ϵ)



Binary censored block model

$$G = ([n], E) \text{ and } \epsilon \in [0, 1/2]$$

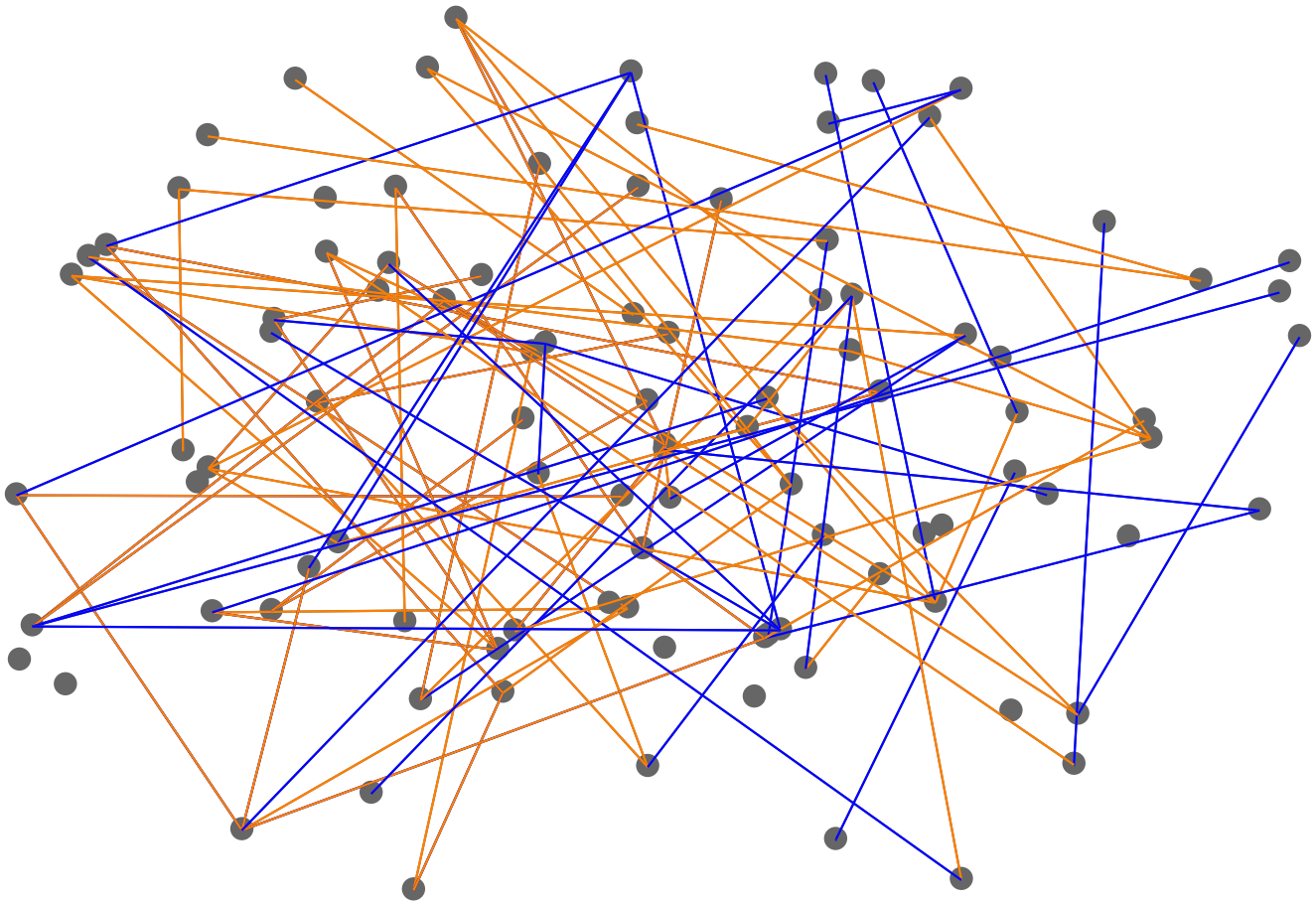
1. Color the vertices in green or red arbitrarily
2. If endpoints in same color, color edge in blue (orange) w.p. $1 - \epsilon$ (ϵ)
3. If endpoints in diff. colors, color edge in blue (orange) w.p. ϵ ($1 - \epsilon$)



Binary censored block model

$$G = ([n], E) \text{ and } \epsilon \in [0, 1/2]$$

1. Color the vertices in green or red arbitrarily
2. If endpoints in same color, color edge in blue (orange) w.p. $1 - \epsilon$ (ϵ)
3. If endpoints in diff. colors, color edge in blue (orange) w.p. ϵ ($1 - \epsilon$)



Exact recovery

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

- Goal: exact recovery

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

- Alternatives: correlated recovery [Heimlicher-Lelarge-Massoulie '11], [Lelarge-Massoulie-Xu '13], [Saade-Krzakala-Lelarge-Zdeborova '15] ...

Exact recovery

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

- Goal: exact recovery

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

- Alternatives: correlated recovery [Heimlicher-Lelarge-Massoulie '11], [Lelarge-Massoulie-Xu '13], [Saade-Krzakala-Lelarge-Zdeborova '15] ...

Focus of this talk

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

Remainder of the talk

1. From MLE to semidefinite programming
2. Background graph is Erdős-Rényi
3. Background graph is regular
4. Concluding remarks

Model and definitions

- $G = ([n], E)$
- n nodes partitioned into two communities of arbitrary size ($\sigma_i^* = \pm 1$)
- For every edge $(i, j) \in E$, observe a label $L_{ij} \in \{\pm 1\}$

$$\sigma_i^* \sigma_j^* \to BSC(\epsilon) \to L_{ij}$$

Model and definitions

- $G = ([n], E)$
- n nodes partitioned into two communities of arbitrary size ($\sigma_i^* = \pm 1$)
- For every edge $(i, j) \in E$, observe a label $L_{ij} \in \{\pm 1\}$

$$\sigma_i^* \sigma_j^* \to BSC(\epsilon) \to L_{ij}$$

- For every (i, j),

$$A_{ij} = \begin{cases} 0 & \text{if } (i,j) \notin E \\ L_{ij} & \text{otherwise} \end{cases}$$

MLE $\Rightarrow$ SDP relaxation

- Maximum likelihood estimator (MLE):
  maximize (# of in-cluster + edges) + (# of cross-cluster - edges)

$$\max_{\sigma} \langle A, \sigma\sigma^\top \rangle$$
$$\text{s.t.} \quad \sigma_i \in \{\pm 1\} \quad i \in [n]$$

MLE $\Rightarrow$ SDP relaxation

- Maximum likelihood estimator (MLE):
  maximize (# of in-cluster + edges) + (# of cross-cluster - edges)

$$\max_{\sigma} \langle A, \sigma\sigma^\top \rangle$$
$$\text{s.t.} \quad \sigma_i \in \{\pm 1\} \quad i \in [n]$$

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

$$\max_{Y} \langle A, Y \rangle$$
$$\text{s.t.} \quad \text{rank}(Y) = 1$$
$$Y_{ii} = 1 \quad i \in [n]$$

MLE $\Rightarrow$ SDP relaxation

- Maximum likelihood estimator (MLE):
  maximize (# of in-cluster + edges) + (# of cross-cluster - edges)

$$\max_{\sigma} \langle A, \sigma\sigma^\top \rangle$$
$$\text{s.t.} \quad \sigma_i \in \{\pm 1\} \quad i \in [n]$$

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

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

MLE $\Rightarrow$ SDP relaxation

- Maximum likelihood estimator (MLE):
  maximize (# of in-cluster + edges) + (# of cross-cluster - edges)

$$\max_{\sigma} \langle A, \sigma\sigma^\top \rangle$$
$$\text{s.t.} \quad \sigma_i \in \{\pm 1\} \quad i \in [n]$$

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

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

- Goal: $\mathbb{P}\left\{ \widehat{Y}_{\text{SDP}} = (\sigma^*)(\sigma^*)^\top = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \right\} \to 1$

Background graph is Erdős-Rényi

- $G = ([n], E) \sim \mathcal{G}(n, p)$ with $p = \frac{a \log n}{n}$

Background graph is Erdős-Rényi

- $G = ([n], E) \sim \mathcal{G}(n, p)$ with $p = \frac{a \log n}{n}$
- $a > 1$ is the connectivity threshold and necessary for exact recovery

Background graph is Erdős-Rényi

- $G = ([n], E) \sim \mathcal{G}(n, p)$ with $p = \frac{a \log n}{n}$
- $a > 1$ is the connectivity threshold and necessary for exact recovery
- What is the minimal possible a for exact recovery?

Erdős-Rényi case: optimal recovery via SDP

Theorem (Hajek-Wu-X. '15)

- *If* $a(\sqrt{1-\epsilon}-\sqrt{\epsilon})^2 > 1$, *recovery is achieved via SDP in poly-time.*
- *If* $a(\sqrt{1-\epsilon}-\sqrt{\epsilon})^2 < 1$, *recovery is impossible.*

Erdős-Rényi case: optimal recovery via SDP

Theorem (Hajek-Wu-X. '15)

- *If* $a(\sqrt{1-\epsilon}-\sqrt{\epsilon})^2 > 1$, *recovery is achieved via SDP in poly-time.*
- *If* $a(\sqrt{1-\epsilon}-\sqrt{\epsilon})^2 < 1$, *recovery is impossible.*

Remarks

- The minimum possible a:

$$a^*(\epsilon) = \frac{1}{(\sqrt{1-\epsilon}-\sqrt{\epsilon})^2}$$

Erdős-Rényi case: optimal recovery via SDP

Theorem (Hajek-Wu-X. '15)

- *If* $a(\sqrt{1-\epsilon}-\sqrt{\epsilon})^2 > 1$, *recovery is achieved via SDP in poly-time.*
- *If* $a(\sqrt{1-\epsilon}-\sqrt{\epsilon})^2 < 1$, *recovery is impossible.*

Remarks

- The minimum possible a:

$$a^*(\epsilon) = \frac{1}{(\sqrt{1-\epsilon}-\sqrt{\epsilon})^2}$$

- Previous work [Abbe-Bandeira-Bracher-Singer '14]: As $\epsilon \to 1/2$,
  - $a^*(\epsilon) = \frac{2+o(1)}{(1-2\epsilon)^2}$
  - SDP succeeds if $a > \frac{4+o(1)}{(1-2\epsilon)^2}$

Background graph is regular

- $G = ([n], E)$ is regular with degree $d = \lfloor a \log n \rfloor$

Background graph is regular

- $G = ([n], E)$ is regular with degree $d = \lfloor a \log n \rfloor$
- $a > 0$ is the connectivity threshold and necessary for exact recovery

Background graph is regular

- $G = ([n], E)$ is regular with degree $d = \lfloor a \log n \rfloor$
- $a > 0$ is the connectivity threshold and necessary for exact recovery
- What is the minimal possible a for exact recovery?

Background graph is regular

- $G = ([n], E)$ is regular with degree $d = \lfloor a \log n \rfloor$
- $a > 0$ is the connectivity threshold and necessary for exact recovery
- What is the minimal possible a for exact recovery?
- Second largest eigenvalue of $\frac{1}{d} A_G$:

$$\lambda_2 \triangleq \max_{\langle x, \mathbf{1} \rangle = 0, \|x\|_2 = 1} x^\top \left(\frac{1}{d} A_G \right) x.$$

Background graph is regular

- $G = ([n], E)$ is regular with degree $d = \lfloor a \log n \rfloor$
- $a > 0$ is the connectivity threshold and necessary for exact recovery
- What is the minimal possible a for exact recovery?
- Second largest eigenvalue of $\frac{1}{d} A_G$:

$$\lambda_2 \triangleq \max_{\langle x, \mathbf{1} \rangle = 0, \|x\|_2 = 1} x^\top \left(\frac{1}{d} A_G \right) x.$$

- If $\lambda_2 = 1$, G is disconnected

Regular case: recovery via SDP

Theorem (Hajek-Wu-X. '15)

- *If* $a\ D(\epsilon\lambda_2 + \frac{1}{2}(1-\lambda_2)\|\epsilon) > 1$, *recovery is achieved via SDP.*

Regular case: recovery via SDP

Theorem (Hajek-Wu-X. '15)

- *If* $a\ D(\epsilon\lambda_2 + \frac{1}{2}(1-\lambda_2)\|\epsilon) > 1$, *recovery is achieved via SDP.*

Theorem (Abbe-Bandeira-Bracher-Singer '14)

- *If* $a\ D(1/2\|\epsilon) < 1$, *recovery is impossible.*

Regular case: recovery via SDP

Theorem (Hajek-Wu-X. '15)

- *If* $a\ D(\epsilon\lambda_2 + \frac{1}{2}(1-\lambda_2)\|\epsilon) > 1$, *recovery is achieved via SDP.*

Theorem (Abbe-Bandeira-Bracher-Singer '14)

- *If* $a\ D(1/2\|\epsilon) < 1$, *recovery is impossible.*

Remarks

- If $\lambda_2 = o_P(1)$, e.g., G is random regular, then the minimum possible a:

$$a^*(\epsilon) = \frac{1}{D(1/2\|\epsilon)}$$

Regular case: recovery via SDP

Theorem (Hajek-Wu-X. '15)

- *If* $a\ D(\epsilon\lambda_2 + \frac{1}{2}(1-\lambda_2)\|\epsilon) > 1$, *recovery is achieved via SDP.*

Theorem (Abbe-Bandeira-Bracher-Singer '14)

- *If* $a\ D(1/2\|\epsilon) < 1$, *recovery is impossible.*

Remarks

- If $\lambda_2 = o_P(1)$, e.g., G is random regular, then the minimum possible a:

$$a^*(\epsilon) = \frac{1}{D(1/2\|\epsilon)}$$

- If $\lambda_2 = \Omega_P(1)$, the minimum possible a is unknown

Conclusion remarks

- $G \sim \mathcal{G}(n, a \log n/n)$:

$$a^*(\epsilon) = \frac{1}{(\sqrt{1-\epsilon} - \sqrt{\epsilon})^2} = \frac{1}{H^2(\epsilon, 1-\epsilon)}$$

- G is random d-regular with $d = \lfloor a \log n \rfloor$:

$$a^*(\epsilon) = \frac{1}{D(1/2 \| \epsilon)} \leq \frac{1}{H^2(\epsilon, 1-\epsilon)}$$

Conclusion remarks

- $G \sim \mathcal{G}(n, a \log n/n)$:

$$a^*(\epsilon) = \frac{1}{(\sqrt{1-\epsilon} - \sqrt{\epsilon})^2} = \frac{1}{H^2(\epsilon, 1-\epsilon)}$$

- G is random d-regular with $d = \lfloor a \log n \rfloor$:

$$a^*(\epsilon) = \frac{1}{D(1/2\|\epsilon)} \leq \frac{1}{H^2(\epsilon, 1-\epsilon)}$$

- G is any deterministic graph: sharp recovery threshold is open [Abbe-Bandeira-Bracher-Singer '14] [Chen-Suh-Goldsmith '15]

Conclusion remarks

- $G \sim \mathcal{G}(n, a \log n/n)$:

$$a^*(\epsilon) = \frac{1}{(\sqrt{1-\epsilon} - \sqrt{\epsilon})^2} = \frac{1}{H^2(\epsilon, 1-\epsilon)}$$

- G is random d-regular with $d = \lfloor a \log n \rfloor$:

$$a^*(\epsilon) = \frac{1}{D(1/2 \| \epsilon)} \leq \frac{1}{H^2(\epsilon, 1-\epsilon)}$$

- G is any deterministic graph: sharp recovery threshold is open [Abbe-Bandeira-Bracher-Singer '14] [Chen-Suh-Goldsmith '15]

Reference:

B. Hajek, Y. Wu, and J. Xu, "Achieving exact cluster recovery threshold via semidefinite programming: Extensions." arXiv 1502.07738, Feb. 2015.

Proof ideas in Erdős-Rényi case: dual certificate argument

Goal: $\mathbb{P}\left\{\widehat{Y}_{\mathrm{SDP}} = (\sigma^*)(\sigma^*)^\top = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\right\} \to 1$

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

Proof ideas in Erdős-Rényi case: dual certificate argument

Goal: $\mathbb{P}\left\{\widehat{Y}_{\mathrm{SDP}} = (\sigma^*)(\sigma^*)^\top = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\right\} \to 1$

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

dual variables

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

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

Proof ideas in Erdős-Rényi case: dual certificate argument

Goal: $\mathbb{P}\left\{\widehat{Y}_{\mathrm{SDP}} = (\sigma^*)(\sigma^*)^\top = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\right\} \to 1$

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

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

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

- $d_i = \sum_{j=1}^n A_{ij}\sigma_i^*\sigma_j^*$ is i.i.d. sum of $(1-p)\delta_0 + p(1-\epsilon)\delta_{+1} + p\epsilon\delta_{-1}$

Proof ideas in Erdős-Rényi case: dual certificate argument

Goal: $\mathbb{P}\left\{\widehat{Y}_{\mathrm{SDP}} = (\sigma^*)(\sigma^*)^\top = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\right\} \to 1$

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

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

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

- $d_i = \sum_{j=1}^n A_{ij}\sigma_i^*\sigma_j^*$ is i.i.d. sum of $(1-p)\delta_0 + p(1-\epsilon)\delta_{+1} + p\epsilon\delta_{-1}$
- $S = D - A \succeq 0$ if $\min d_i \geq \|A - \mathbb{E}[A]\|$

Proof ideas in Erdős-Rényi case: dual certificate argument

Goal: $\mathbb{P}\left\{\widehat{Y}_{\mathrm{SDP}} = (\sigma^*)(\sigma^*)^\top = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\right\} \to 1$

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

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

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

- $d_i = \sum_{j=1}^n A_{ij}\sigma_i^*\sigma_j^*$ is i.i.d. sum of $(1-p)\delta_0 + p(1-\epsilon)\delta_{+1} + p\epsilon\delta_{-1}$
- $S = D - A \succeq 0$ if $\min d_i \geq \|A - \mathbb{E}[A]\|$
- $\min d_i = \Omega_P(\log n)$ if $a(\sqrt{1-\epsilon} - \sqrt{\epsilon})^2 > 1$

Proof ideas in Erdős-Rényi case: dual certificate argument

Goal: $\mathbb{P}\left\{\widehat{Y}_{\mathrm{SDP}} = (\sigma^*)(\sigma^*)^\top = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\right\} \to 1$

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

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

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

- $d_i = \sum_{j=1}^n A_{ij}\sigma_i^*\sigma_j^*$ is i.i.d. sum of $(1-p)\delta_0 + p(1-\epsilon)\delta_{+1} + p\epsilon\delta_{-1}$
- $S = D - A \succeq 0$ if $\min d_i \geq \|A - \mathbb{E}[A]\|$
- $\min d_i = \Omega_P(\log n)$ if $a(\sqrt{1-\epsilon} - \sqrt{\epsilon})^2 > 1$
- $\|A - \mathbb{E}[A]\| = O_P(\sqrt{\log n})$: symmetrization method + result for iid matrix [Seginer '00]

Erdős-Rényi case: necessity

$$a(\sqrt{1-\epsilon}-\sqrt{\epsilon})^2 < 1$$

$\Rightarrow$ $\min d_i < 0$ w.h.p. $\quad$ d_i i.i.d. sum of $(1-p)\delta_0 + p(1-\epsilon)\delta_{+1} + p\epsilon\delta_{-1}$

$\Rightarrow \exists i$: # of incident + edges in own cluster plus - edges in other cluster

$<$ # of incident + edges in other cluster plus - edges in own cluster

$\Rightarrow$ MLE fails

Erdős-Rényi case: necessity

$$a(\sqrt{1-\epsilon}-\sqrt{\epsilon})^2 < 1$$

$\Rightarrow$ $\min d_i < 0$ w.h.p. $\quad$ d_i i.i.d. sum of $(1-p)\delta_0 + p(1-\epsilon)\delta_{+1} + p\epsilon\delta_{-1}$

$\Rightarrow \exists i$: # of incident $+$ edges in own cluster plus - edges in other cluster

$<$ # of incident $+$ edges in other cluster plus - edges in own cluster

$\Rightarrow$ MLE fails

Sharp threshold

- $a(\sqrt{1-\epsilon}-\sqrt{\epsilon})^2 > 1 \Rightarrow \min d_i = \Omega(\log n) \Rightarrow$ SDP succeeds
- $a(\sqrt{1-\epsilon}-\sqrt{\epsilon})^2 < 1 \Rightarrow \min d_i = -\Omega(\log n) \Rightarrow$ MLE fails