Semidefinite Programming Relaxations for Recovering Hidden Communities

Jiaming Xu

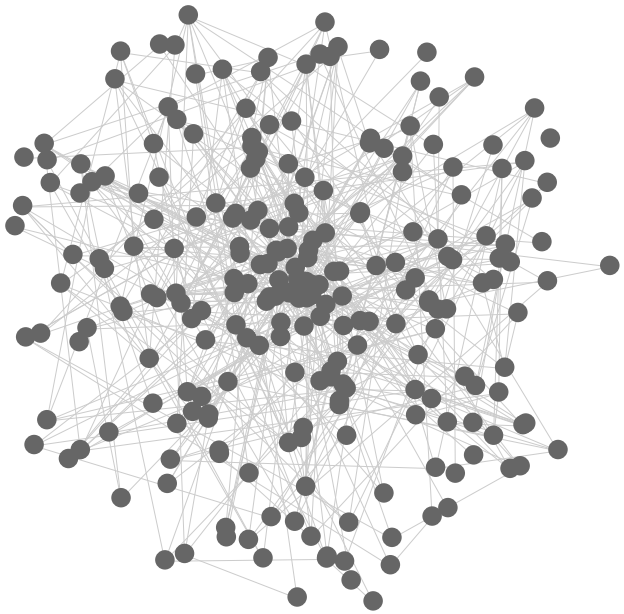
Krannert School of Management
Purdue University

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

December 17, 2016

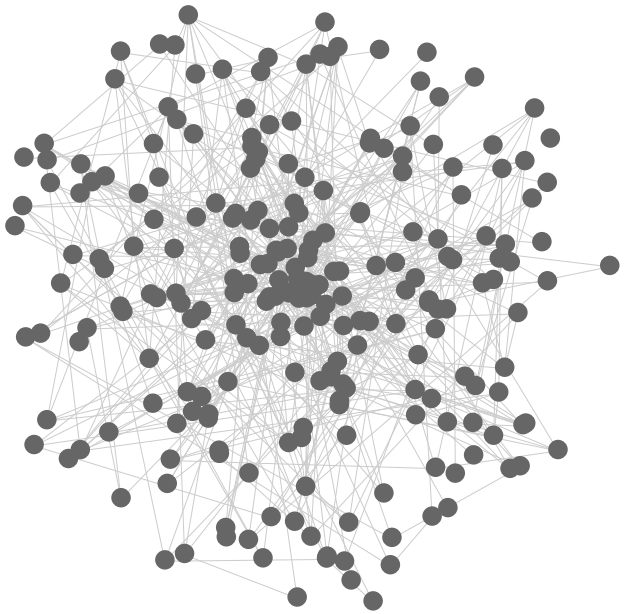
Community detection in networks

- Observe local pairwise interactions between objects, e.g., social networks, biological networks ...



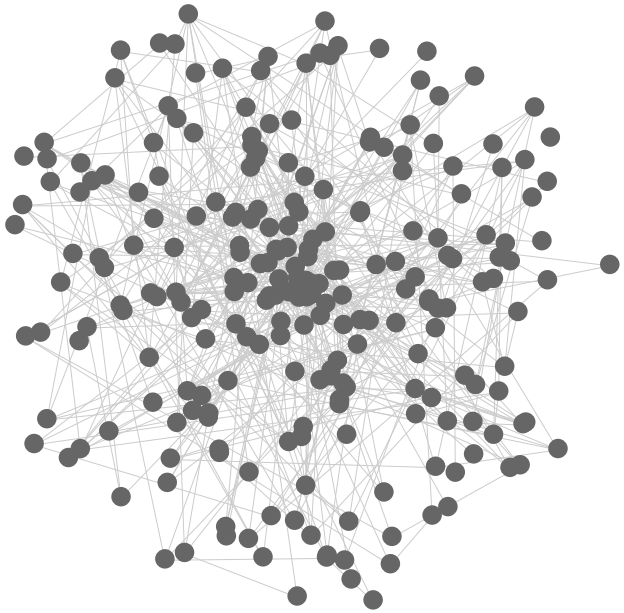
Community detection in networks

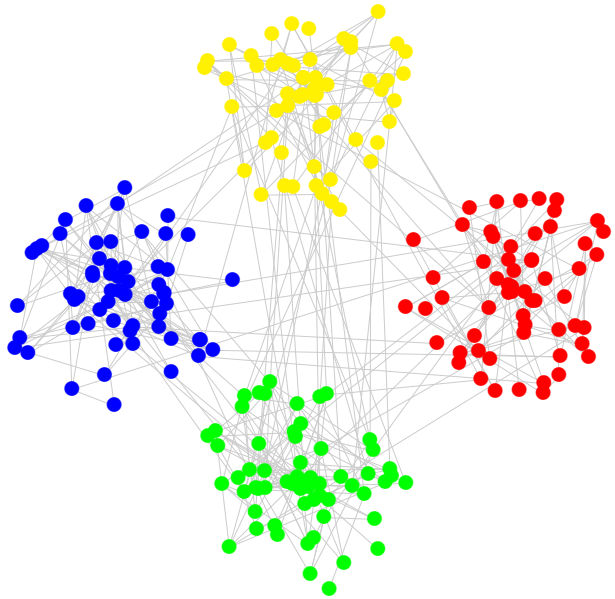
- Observe local pairwise interactions between objects, e.g., social networks, biological networks ...
- Interested in global properties of objects, e.g., similarity



Community detection in networks

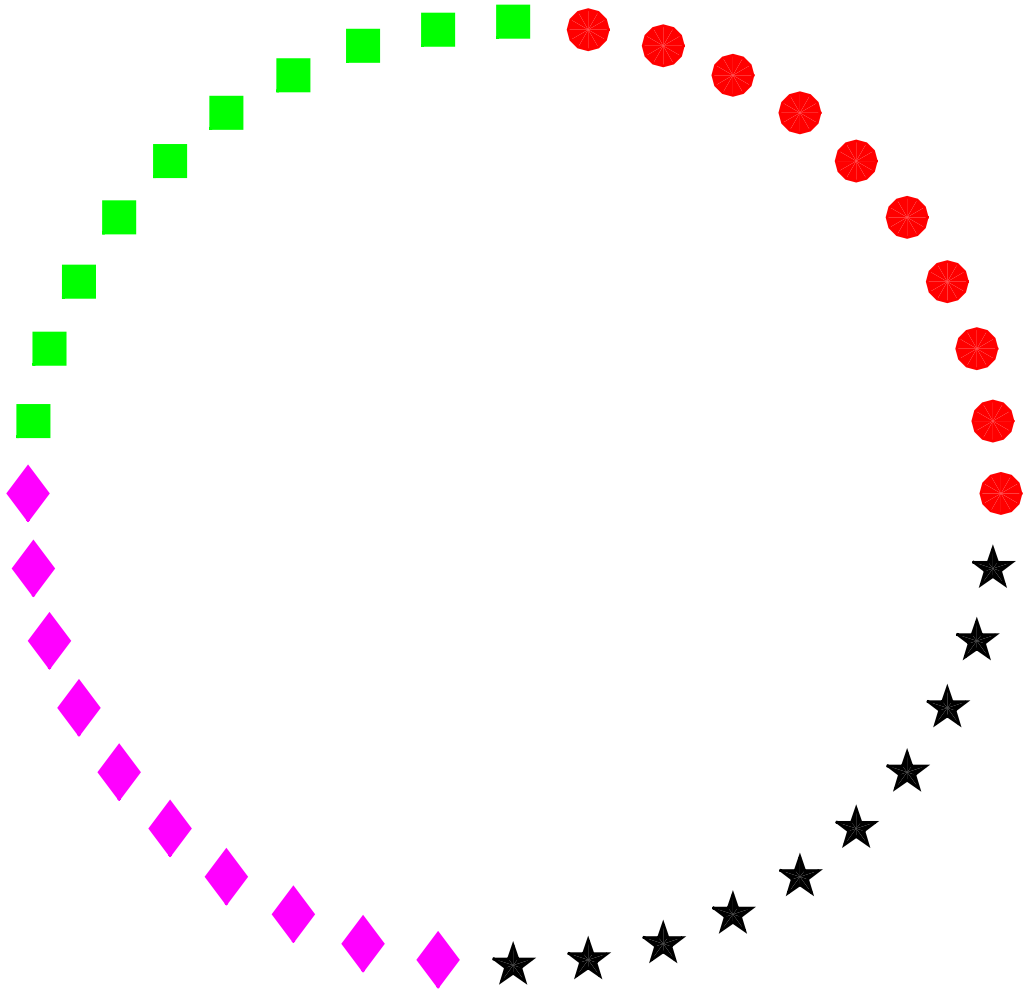
- Observe local pairwise interactions between objects, e.g., social networks, biological networks ...
- Interested in global properties of objects, e.g., similarity



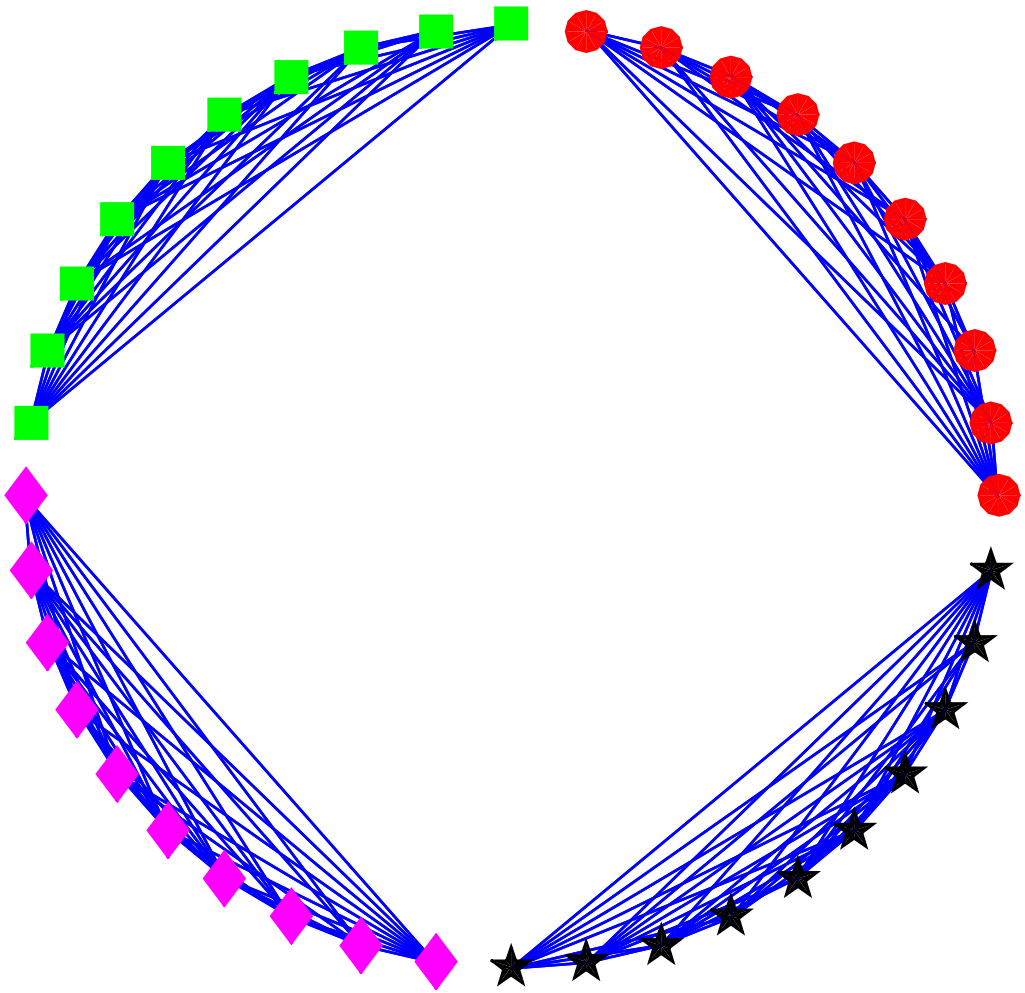


Goal: identify communities of similar objects, related to clustering and graph partitioning.

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

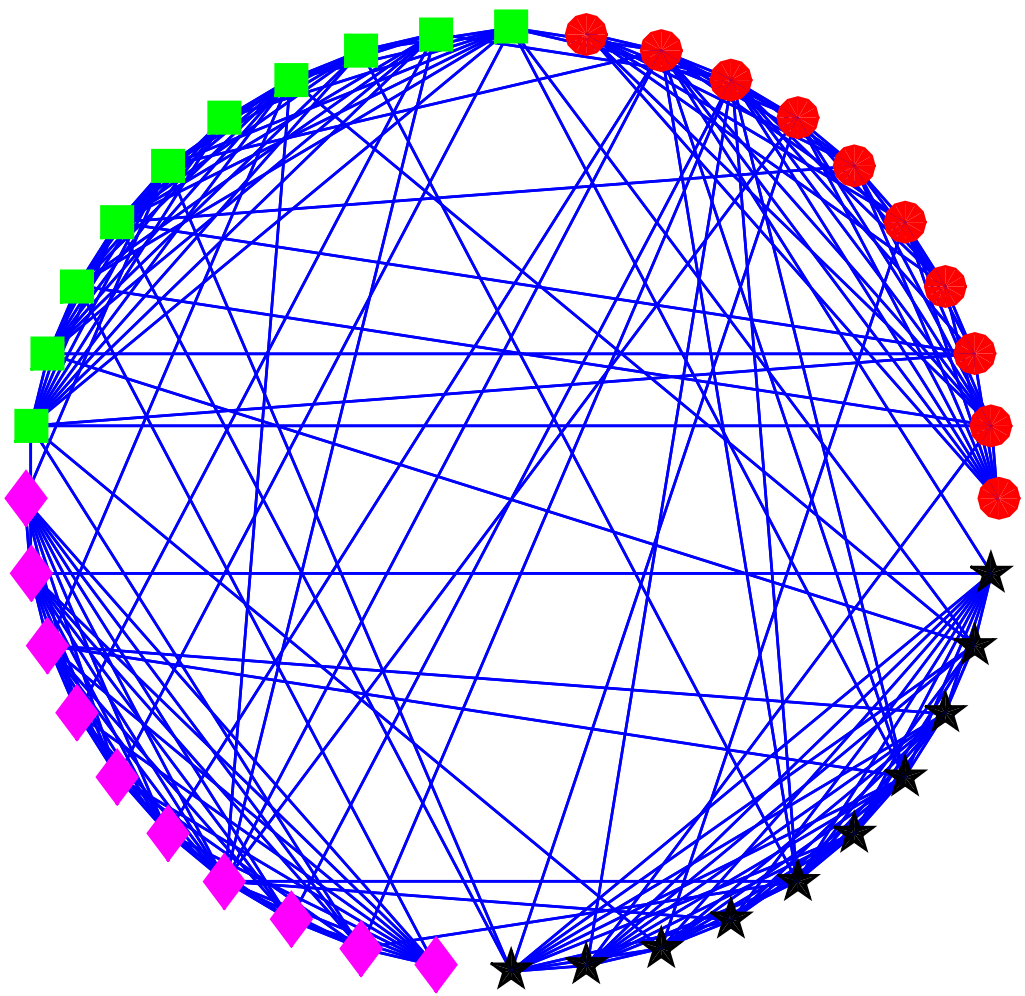


Stochastic block model [Holland-Laskey-Leinhardt '83]
Planted partition model [Condon-Karp 01']



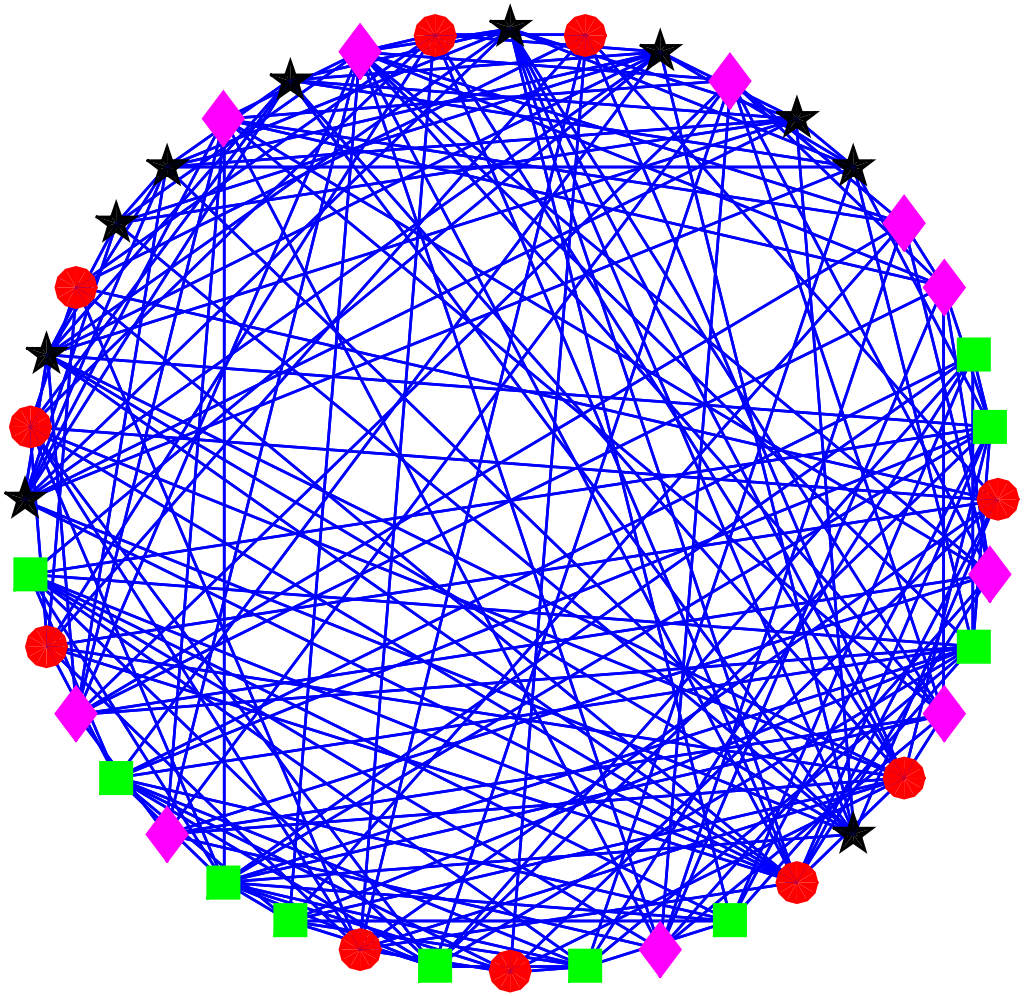
$p = 0.8$

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



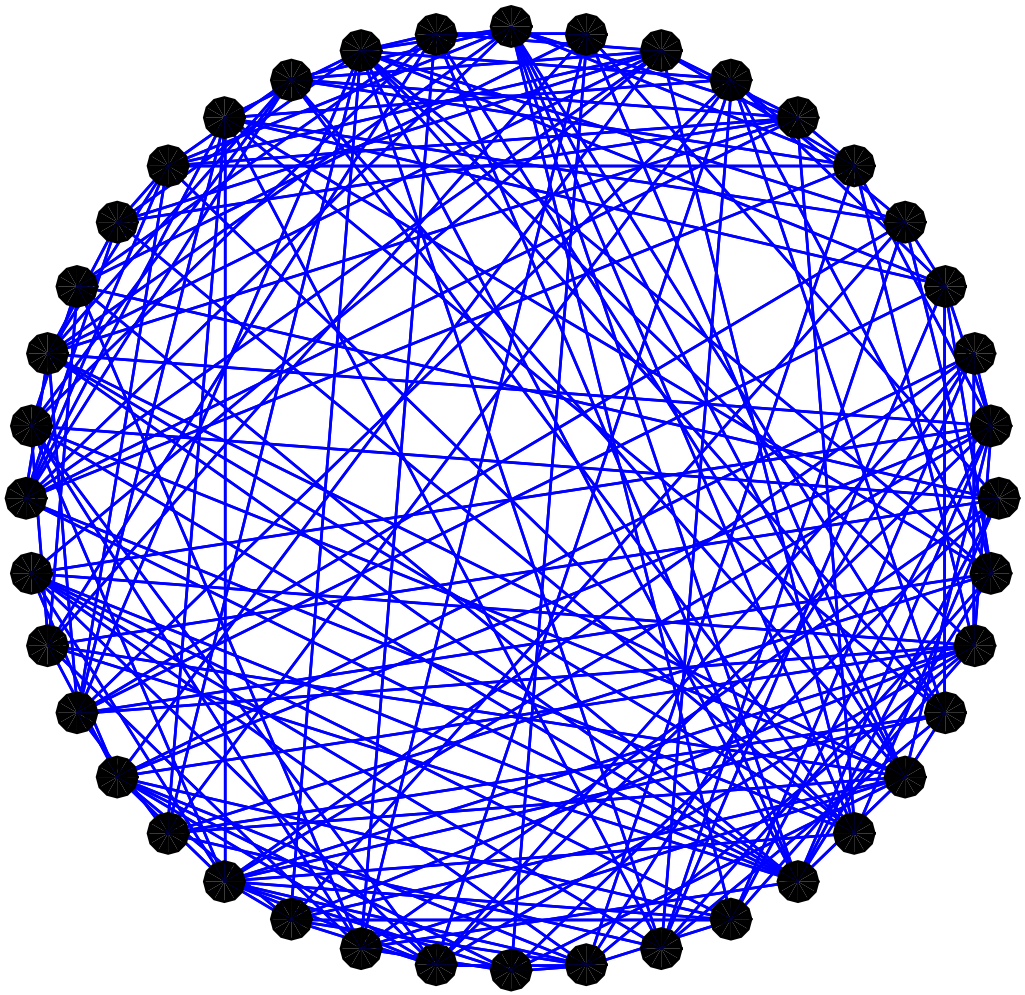
$p = 0.8 \quad q = 0.09$

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



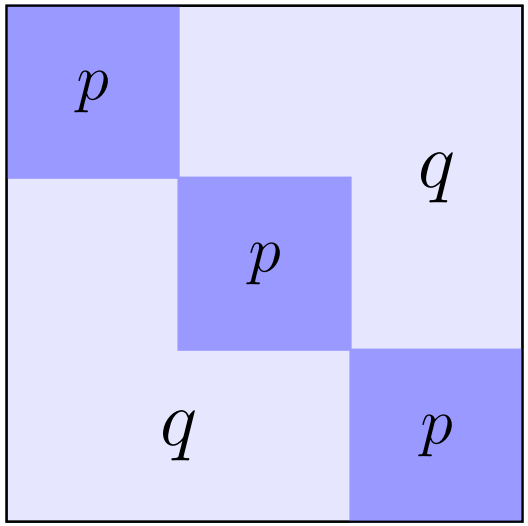
$p = 0.8 \quad q = 0.09$

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



$p = 0.8 \qquad q = 0.09$

SBM - adjacency matrix view



- n: total number of nodes
- k: number of communities
- p: **within-community** edge prob. q: **across-community** edge prob.

Exact recovery

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

- Goal: exact recovery

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

- Alternatives
  - almost exact recovery:
    [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

Exact recovery:

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

- 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. Multiple equal-sized communities
3. Conclusions

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)

For two equal-sized communities with $p = a \log n/n$ and $q = b \log n/n$:

- *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)

For two equal-sized communities with $p = a \log n/n$ and $q = b \log n/n$:

- *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)

For two equal-sized communities with $p = a\log n/n$ and $q = b\log n/n$:

- *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

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

Two equal-sized communities: Dual certificate argument

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

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

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

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

Two equal-sized communities: Dual certificate argument

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

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

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

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

- $d_i =$ (# 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

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

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

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

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

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

Two equal-sized communities: Dual certificate argument

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

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

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

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

- $d_i =$ (# of nbrs in own cluster) $-$ (# of nbrs in other cluster)
  $\sim \mathrm{Binom}(n/2-1, p) - \mathrm{Binom}(n/2, q)$
- $S = D - A + \lambda \mathbf{J} \succeq 0$ 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

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

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

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

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

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

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

$$\max \sum_{\ell=1}^{k} \langle A, \boldsymbol{\theta}_{\boldsymbol{\ell}} \boldsymbol{\theta}_{\boldsymbol{\ell}}^{\top} \rangle$$

$$\text{s.t.} \quad \boldsymbol{\theta}_{\boldsymbol{\ell}} \in \{0,1\}^n$$

$$\langle \boldsymbol{\theta}_{\boldsymbol{\ell}}, \mathbf{1} \rangle = n/k$$

$$\langle \boldsymbol{\theta}_{\boldsymbol{\ell}}, \boldsymbol{\theta}_{\boldsymbol{\ell}'} \rangle = 0, \ell \neq \ell'$$

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

$$\max \sum_{\ell=1}^{k} \langle A, \boldsymbol{\theta}_{\boldsymbol{\ell}} \boldsymbol{\theta}_{\boldsymbol{\ell}}^{\top} \rangle$$

$$\text{s.t.} \quad \boldsymbol{\theta}_{\boldsymbol{\ell}} \in \{0,1\}^n$$

$$\langle \boldsymbol{\theta}_{\boldsymbol{\ell}}, \mathbf{1} \rangle = n/k$$

$$\langle \boldsymbol{\theta}_{\boldsymbol{\ell}}, \boldsymbol{\theta}_{\boldsymbol{\ell}'} \rangle = 0, \ell \neq \ell'$$

$$\xleftrightarrow{\text{lift: } Z = \sum_{\ell=1}^{k} \boldsymbol{\theta}_{\boldsymbol{\ell}} \boldsymbol{\theta}_{\boldsymbol{\ell}}^{\top}}$$

$$\max \langle A, Z \rangle$$

$$\text{s.t.} \quad \text{rank}(Z) = k$$

$$Z_{ii} = 1 \quad \forall i \in [n]$$

$$Z_{ij} \geq 0, \quad \sum_{j} Z_{ij} = n/k$$

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

$$\max \sum_{\ell=1}^{k} \langle A, \boldsymbol{\theta}_{\boldsymbol{\ell}} \boldsymbol{\theta}_{\boldsymbol{\ell}}^{\top} \rangle$$

$$\text{s.t.} \quad \boldsymbol{\theta}_{\boldsymbol{\ell}} \in \{0,1\}^n$$

$$\langle \boldsymbol{\theta}_{\boldsymbol{\ell}}, \mathbf{1} \rangle = n/k$$

$$\langle \boldsymbol{\theta}_{\boldsymbol{\ell}}, \boldsymbol{\theta}_{\boldsymbol{\ell}'} \rangle = 0, \ell \neq \ell'$$

$$\xleftrightarrow{\text{lift: } Z = \sum_{\ell=1}^{k} \boldsymbol{\theta}_{\boldsymbol{\ell}} \boldsymbol{\theta}_{\boldsymbol{\ell}}^{\top}}$$

$$\max \langle A, Z \rangle$$

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

$$Z_{ii} = 1 \quad \forall i \in [n]$$

$$Z_{ij} \geq 0, \quad \sum_{j} Z_{ij} = n/k$$

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

$$\max \sum_{\ell=1}^{k} \langle A, \boldsymbol{\theta}_{\boldsymbol{\ell}} \boldsymbol{\theta}_{\boldsymbol{\ell}}^{\top} \rangle$$

s.t. $\boldsymbol{\theta}_{\boldsymbol{\ell}} \in \{0,1\}^n$

$\langle \boldsymbol{\theta}_{\boldsymbol{\ell}}, \mathbf{1} \rangle = n/k$

$\langle \boldsymbol{\theta}_{\boldsymbol{\ell}}, \boldsymbol{\theta}_{\boldsymbol{\ell}'} \rangle = 0, \ell \neq \ell'$

$\xleftrightarrow{\text{lift: } Z = \sum_{\ell=1}^{k} \boldsymbol{\theta}_{\boldsymbol{\ell}} \boldsymbol{\theta}_{\boldsymbol{\ell}}^{\top}}$

$$\max \langle A, Z \rangle$$

s.t. $Z \succeq 0$

$Z_{ii} = 1 \quad \forall i \in [n]$

$Z_{ij} \geq 0, \quad \sum_j Z_{ij} = n/k$

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

k equal-sized communities: optimal recovery via SDP

Theorem (Hajek-Wu-X. '15)

For a fixed k communities with $p = a \log n/n$ and $q = b \log n/n$.

- *If $\sqrt{a} - \sqrt{b} > \sqrt{k}$, exact recovery is attained via SDP in poly-time.*
- *If $\sqrt{a} - \sqrt{b} < \sqrt{k}$, exact recovery is impossible.*

k equal-sized communities: optimal recovery via SDP

Theorem (Hajek-Wu-X. '15)

For a fixed k communities with $p = a \log n/n$ and $q = b \log n/n$.

- *If $\sqrt{a} - \sqrt{b} > \sqrt{k}$, exact recovery is attained via SDP in poly-time.*
- *If $\sqrt{a} - \sqrt{b} < \sqrt{k}$, exact recovery is impossible.*

Remarks

- Extended to $k = o(\log n)$ in [Agarwal-Bandeira-Koiliaris-Kolla '15]
- Extended to the case with multiple unequal-sized clusters [Perry-Wein '15]
- Heterogeneous setting: [Yun-Proutiere '14] and [Abbe-Sandon '15]

When does SDP cease to be optimal?

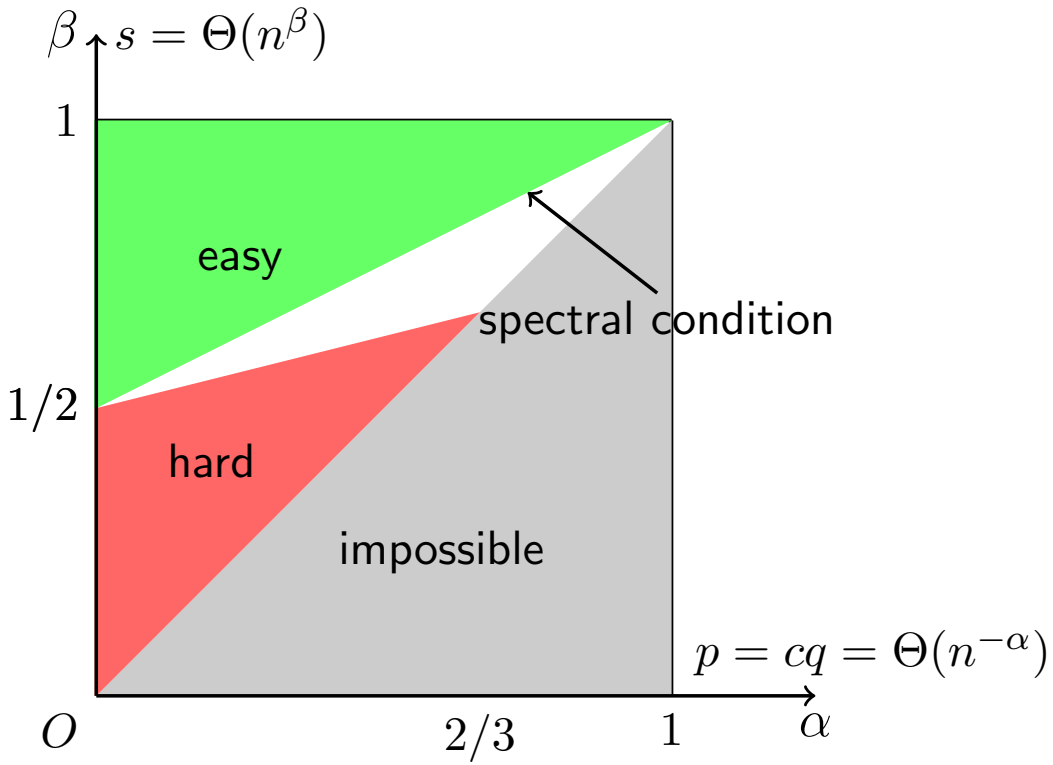
Theorem (Hajek-Wu-X. 'COLT16)

- *If $k \ll \log n$, SDP achieves the optimal exact recovery threshold.*
- *If $k \geq c \log n$, SDP is suboptimal by a constant factor.*
- *If $k \gg \log n$, SDP is order-suboptimal.*

Remarks

- A "hard but informationally possible" regime is conjectured to exist for exact recovery when $k \gg \log n$ [Chen-X. '14]

Concluding remarks



References

- B. Hajek, Y. Wu & J. X. *Achieving exact cluster recovery threshold via semidefinite programming*. (Transactions on IT '16)
- B. Hajek, Y. Wu & J. X. *Achieving exact cluster recovery threshold via semidefinite programming: Extensions*. (Transactions on IT '16)
- B. Hajek, Y. Wu & J. X. *Semidefinite programs for exact recovery of a hidden community*. (COLT'16)

SDP in real networks

- Y. Chen, X. Li, and J. X. (2015), Convexified modularity maximization for degree-corrected stochastic block models. arXiv:1512.08425.
- Code available at http://people.orie.cornell.edu/yudong.chen/cmm