Semidefinite Programs for Exact Recovery of a Hidden Community (and Many Communities)

Bruce Hajek $^1$  Yihong Wu $^1$  Jiaming Xu $^2$

$^1$University of Illinois at Urbana-Champaign

$^2$Simons Institute, UC Berkeley

June 23, 2016
Hidden community model [Deshpande-Montanari ’13]

- Data: $n \times n$ symmetric matrix $A$ with empty diagonal
- Community $C^* \subset [n]$ of size $K$ uniform at random, such that
  \[ A_{ij} \sim \begin{cases} 
  P & \text{both } i \text{ and } j \in C \\
  Q & \text{otherwise}
  \end{cases} \]
- $(K, P, Q)$ varies with $n$
- Goal: exact recovery of $C$ from $A$
  \[ \mathbb{P}\{\hat{C} = C^*\} \xrightarrow{n \to \infty} 1 \]
- Fruitful venue for studying computational aspects of statistical problems
Examples

Planted dense subgraph

\[ P = \text{Bern}(p), \quad Q = \text{Bern}(q), \quad p > q \]

- \( A = \) adjacency matrix of \( G(n, q) \) planted with \( G(K, p) \)
- [Alon et al '98, McSherry '01, Arias-Castro-Verzelen '14, Chen-Xu 14, Montanari '15, ...]
Examples

Planted dense subgraph

\[ P = \text{Bern}(p), \ Q = \text{Bern}(q), \quad p > q \]

- \( A = \) adjacency matrix of \( G(n, q) \) planted with \( G(K, p) \)
- [Alon et al ’98, McSherry ’01, Arias-Castro-Verzelen ’14, Chen-Xu 14, Montanari ’15, ...]

Submatrix localization

\[ P = \mathcal{N}(0, \mu), \ Q = \mathcal{N}(0, 1), \quad \mu > 0 \]

- \( A = \begin{pmatrix} \mu & \text{noise} \\ 0 & \end{pmatrix} \)
- [Shabalin et al ’09, Butucea-Ingster ’11, Kolar et al ’11, Ma-W ’13, Cai et al ’15, ...]
Running example:
Plated Dense Subgraph
A community of $K$ vertices are chosen randomly. For every pair of nodes in the community, add an edge w.p. $p$. For other pairs of nodes, add an edge w.p. $q$. 
A community of $K$ vertices are chosen randomly
1. A community of $K$ vertices are chosen randomly.
2. For every pair of nodes in the community, add an edge w.p. $p$. 
1. A community of $K$ vertices are chosen randomly
2. For every pair of nodes in the community, add an edge w.p. $p$
3. For other pairs of nodes, add an edge w.p. $q$
1. A community of $K$ vertices are chosen randomly
2. For every pair of nodes in the community, add an edge w.p. $p$
3. For other pairs of nodes, add an edge w.p. $q$
Planted dense subgraph – adjacency matrix view

$n = 200$, $K = 50$, $p = 0.3$, $q = 0.1$
$n = 200, K = 50, p = 0.3, q = 0.1$
$n = 200, K = 50, p = 0.3, q = 0.1$
Computational gap in planted Clique

- $p = 1$
- $q = \Omega(1)$

- $K = \Omega(\log n)$: exact recovery is possible via maximum likelihood
- $K = \Omega(\sqrt{n})$: exact recovery is attainable in poly-time [Alon et al. ’98]
- $K = o(\sqrt{n})$: exact recovery is believed to be hard [Deshpande-Montanari ’15] [Meka-Potechin-Wigderson ’15], ...
Computational gap in planted Clique

- $p = 1$
- $q = \Omega(1)$

- $K = \Omega(\log n)$: exact recovery is possible via maximum likelihood
- $K = \Omega(\sqrt{n})$: exact recovery is attainable in poly-time [Alon et al. ’98]
- $K = o(\sqrt{n})$: exact recovery is believed to be hard [Deshpande-Montanari ’15] [Meka-Potechin-Wigderson ’15], ...

What about dense subgraphs clique?
Linear community size

- $K = \rho n$
- $p = \frac{a \log n}{n}$ and $q = \frac{b \log n}{n}$

Theorem (Hajek-W-Xu Trans. IT 16)

- If $\rho > \rho^*$, exact recovery is possible in polynomial-time.
- If $\rho < \rho^*$, exact recovery is impossible.

Remarks

- $\rho^* = 1/(a - \tau^* \log \frac{ea}{\tau^*})$ with $\tau^* = \frac{a-b}{\log a - \log b}$
- Convex (SDP) relaxation works
Sublinear community size

[Hajek-W-Xu, COLT '15]

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

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

\[ K = \Omega(n) : \text{SDP works} \]
\[ K = n^{1-\epsilon} : \text{no known poly-time algorithm} \]
\[ \text{Where is the SDP barrier?} \]
Sublinear community size

[Hajek-W-Xu, COLT ’15]

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

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

- \( K = \Omega(n) \): SDP works
- \( K = n^{1-\epsilon} \): no known poly-time algorithm
- Where is the SDP barrier? \( K = \Theta\left(\frac{n}{\log n}\right) \)
Main results: For both planted dense subgraph (Bernoulli) and submatrix localization (Gaussian)

- $K = \omega\left(\frac{n}{\log n}\right)$: SDP attains the info-theoretic limit with sharp constants
- $K = \Theta\left(\frac{n}{\log n}\right)$: SDP is order-wise optimal, but strictly suboptimal by a constant factor
- $K = o\left(\frac{n}{\log n}\right)$ and $K \to \infty$: SDP is order-wise suboptimal
Log-likelihood ratio matrix $L$

\[ L_{ij} = \log \frac{dP}{dQ}(A_{ij}), \quad i \neq j, \quad L_{ii} = 0 \]

Let $\xi = \text{indicator of } C$.

Maximum likelihood estimator = find densest $K$-subgraph

\[ \hat{\xi}_{\text{MLE}} = \arg \max_{\xi} \sum_{i,j} L_{ij} \xi_i \xi_j \]

s.t. $\xi \in \{0, 1\}^n$

\[ \langle \xi, 1 \rangle = K. \]
\( \hat{Z}_{\text{MLE}} = \arg \max_{Z} \langle L, Z \rangle \)

s.t. \( \text{rank}(Z) = 1 \)

\( Z_{ii} \leq 1 \quad \forall i \in [n] \)

\( Z_{ij} \geq 0, \quad \forall i, j \in [n] \)

\( \langle I, Z \rangle = K \)

\( \langle J, Z \rangle = K^2 \)
Natural SDP relaxation:

\[ \hat{Z}_{\text{SDP}} = \arg \max_Z \langle L, Z \rangle \]

s.t.
\[ Z \succeq 0 \]
\[ Z_{ii} \leq 1 \quad \forall i \in [n] \]
\[ Z \geq 0 \]
\[ \langle I, Z \rangle = K \]
\[ \langle J, Z \rangle = K^2 \]
Semidefinite programming

Natural SDP relaxation:

\[
\hat{Z}_{\text{SDP}} = \arg \max_Z \langle L, Z \rangle \\
\text{s.t. } Z \succeq 0 \\
Z_{ii} \leq 1 \quad \forall i \in [n] \\
Z \succeq 0 \\
\langle I, Z \rangle = K \\
\langle J, Z \rangle = K^2
\]

Goal:

\[
P\left\{ \hat{Z}_{\text{SDP}} = \hat{Z}_{\text{MLE}} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\} \to 1
\]
Define

\[ e(i, C^*) = \sum_{j \in C^*} L_{ij}, \quad i \in [n], \quad \beta = -D(Q\|P). \]
Analysis of SDP

Theorem

- **Sufficient condition:** \( \hat{Z}_{\text{SDP}} = Z^* \), if

\[
\min_{i \in C^*} e(i, C^*) - \max_{j \notin C^*} \left\{ \max_{j \in C^*} e(j, C^*), K \beta \right\} > \| L - \mathbb{E} [L] \| - \beta
\]
Theorem

• **Sufficient condition:** $\hat{Z}_{SDP} = Z^*$, if

$$\min_{i \in C^*} e(i, C^*) - \max \left\{ \max_{j \notin C^*} e(j, C^*), K\beta \right\} > \|L - \mathbb{E}[L]\| - \beta$$

• **Necessary condition:** If $Z^* \in \hat{Z}_{SDP}$, then

$$\min_{i \in C^*} e(i, C^*) - \max_{j \notin C^*} e(j, C^*) \geq \sup_{1 \leq a \leq K} \left\{ V(a) - \frac{a}{K} \max_{j \notin C^*} e(j, C^*) \right\},$$

where

- $V(a) = \max\{\langle L_{\overline{C^*} \times \overline{C^*}}, Z \rangle : Z \succeq 0, Z \succeq 0, \text{Tr}(Z) = 1, \langle J, Z \rangle = a\}$ is the value of an (simpler) auxiliary SDP.
Remarks

• To apply this result, \( \min, \max, \| L - \mathbb{E} [L] \| \), etc concentrate

• Sufficient condition proof: construction of dual witnesses (standard)
Proof of necessary condition

- Primal proof: random perturbation of the ground truth to establish integrality gap

\[ \begin{align*}
Z^* &= 1 \\
Z^* &= 0 \\
Z^* &= 0
\end{align*} \]
Proof of necessary condition

- Primal proof: random perturbation of the ground truth to establish integrality gap

\[
\begin{align*}
Z^* &= 1 \\
Z^* &= 0
\end{align*}
\]
Proof of necessary condition

- Primal proof: random perturbation of the ground truth to establish integrality gap

\[ Z^* = 1 \quad + \]
\[ - \quad + \]

- Dual proof: non-existence of dual witness
Multiple communities
$k$ communities: $\text{MLE} \Rightarrow \text{SDP relaxation}$

**SBM with $k$ communities and parameter $(p, q)$**

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

s.t. $\theta_\ell \in \{0, 1\}^n$

$$\langle \theta_\ell, 1 \rangle = n/k$$

$$\langle \theta_\ell, \theta_\ell' \rangle = 0, \ell \neq \ell'$$
SBM with $k$ communities and parameter $(p, q)$

\[
\begin{align*}
\text{max} & \quad \sum_{\ell=1}^{k} \langle A, \theta_\ell \theta_\ell^\top \rangle \\
\text{s.t.} & \quad \theta_\ell \in \{0, 1\}^n, \quad \langle \theta_\ell, 1 \rangle = n/k, \\
\end{align*}
\]

lift: $Z = \sum_{\ell=1}^{k} \theta_\ell \theta_\ell^\top$

\[
\begin{align*}
\text{max} & \quad \langle A, Z \rangle \\
\text{s.t.} & \quad \text{rank}(Z) = k, \quad Z_{ii} = 1 \quad \forall i \in [n], \\
& \quad Z_{ij} \geq 0, \quad \sum_j Z_{ij} = n/k
\end{align*}
\]
**k communities: MLE ⇒ SDP relaxation**

---

SBM with $k$ communities and parameter $(p, q)$

\[
\begin{align*}
\text{max} \quad & \sum_{\ell=1}^{k} \langle A, \theta_\ell \theta_\ell^\top \rangle \\
\text{s.t.} \quad & \theta_\ell \in \{0, 1\}^n \quad \text{lift: } Z = \sum_{\ell=1}^{k} \theta_\ell \theta_\ell^\top \\
\langle \theta_\ell, 1 \rangle &= n/k \\
\langle \theta_\ell, \theta_\ell' \rangle &= 0, \ell \neq \ell' \\
\langle \theta_\ell, \theta_\ell' \rangle &= 0, \ell \neq \ell' \\
\langle \theta_\ell, 1 \rangle &= n/k
\end{align*}
\]

\[
\begin{align*}
\text{max} \quad & \langle A, Z \rangle \\
\text{s.t.} \quad & Z \succeq 0 \\
& Z_{ii} = 1 \quad \forall i \in [n] \\
& Z_{ij} \geq 0, \quad \sum_j Z_{ij} = n/k
\end{align*}
\]
SBM with $k$ communities and parameter $(p, q)$

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

s.t. $\theta_\ell \in \{0, 1\}^n$

$$\langle \theta_\ell, \mathbf{1} \rangle = n/k$$

$$\langle \theta_\ell, \theta_{\ell'} \rangle = 0, \ell \neq \ell'$$

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: $\mathbb{P}\left\{ \hat{Z}_{\text{SDP}} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{pmatrix} \right\} \to 1$
Theorem (Hajek-W-Xu ’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.
Theorem (Hajek-W-Xu ’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]
Theorem (Hajek-W-Xu ’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]
When does SDP cease to be optimal?

Theorem (Hajek-W.-Xu ’16)

- \( k \ll \log n \): SDP achieves the optimal exact recovery threshold.
- \( k \geq c \log n \): SDP is suboptimal by a constant factor.
- \( k \gg \log n \): SDP is order-suboptimal.

Remarks

- A “hard but informationally possible“ regime is conjectured to exist for exact recovery when \( k \gg \log n \) [Chen-Xu ’14]
Some remaining problems

- Can the computational gap for exact recovery be bridged by any polynomial time algorithm? (SoS hardness result or reduction to PC would offer further evidence for “no” answer.)
- Approximate recovery? (Current proof only rules out exact recovery.)
Some remaining problems

- Can the computational gap for exact recovery be bridged by any polynomial time algorithm? (SoS hardness result or reduction to PC would offer further evidence for “no” answer.)
- Approximate recovery? (Current proof only rules out exact recovery.)

Thank you!
Let \( M = L_{(C^*)^c \times (C^*)^c} \) denote the submatrix of \( L \) outside the community. For \( a \in \mathbb{R} \), consider the (random) value of the following SDP:

\[
V(a) \triangleq \max_Z \langle M, Z \rangle \\
\text{s.t.} \quad Z \succeq 0 \\
\quad Z \succeq 0 \\
\quad \text{Tr}(Z) = 1 \\
\quad \langle J, Z \rangle = a.
\]
Theorem (Necessary condition for SDP)

If $Z^* \in \hat{Z}_{SDP}$, then

$$\min_{i \in C^*} e(i, C^*) - \max_{j \notin C^*} e(j, C^*) \geq \sup_{1 \leq a \leq K} \left\{ V(a) - \frac{a}{K} \max_{j \notin C^*} e(j, C^*) \right\}. \quad (2)$$

Weaker necessary condition (set $a = K$):

$$\min_{i \in C^*} e(i, C^*) \geq V(K)$$
Phase diagram for the Gaussian model with $K = \frac{\rho n}{\log n}$ and $\mu = \frac{\mu_0 \log n}{\sqrt{n}}$. 

Bruce Hajek, Yihong Wu and Jiaming Xu