Information Limits for Recovering a Hidden Community

Bruce Hajek 1  Yihong Wu 2  Jiaming Xu 3

1University of Illinois at Urbana-Champaign
2University of Illinois at Urbana-Champaign → Yale
3Simons Institute, UC Berkeley → Purdue

July 12, 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 is recovery of \( C^* \) from \( A \) (almost exactly or exactly)
- Fruitful venue for studying computational aspects of statistical problems
Examples

### Planted dense subgraph

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

- \( A \) = 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

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

- \( A = \) adjancency 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

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

- \( A = \begin{pmatrix} \mu & \text{noise} \\ 0 & \end{pmatrix} \)
- [Shabalin et al '09, Butucea-Ingster '11, Kolar et al '11, Ma-W '13, Cai et al '15, ...]
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. $q$. 

Bruce Hajek, Yihong Wu, Jiaming Xu

Info Limits for Hidden Community
A community of $K$ vertices are chosen randomly.
A community of $K$ vertices are chosen randomly
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. $q$
Planted dense subgraph – graph view

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

$n = 200, K = 50, p = 0.3, q = 0.1$
Planted dense subgraph – adjacency matrix view

$n = 200$, $K = 50$, $p = 0.3$, $q = 0.1$
$n = 200, K = 50, p = 0.3, q = 0.1$
Assumption 1  $K \geq 2$, $K/n$ is bounded away from one as $n \to \infty$. 
**Assumptions**

**Assumption 1** \( K \geq 2, \ K/n \) is bounded away from one as \( n \to \infty \).

**Assumption 2** There exists a constant \( C \) such that for all \( n \),

\[
\psi''_Q(\lambda) \leq C \min\{D(P\|Q), D(Q\|P)\}, \quad \forall \lambda \in [-1, 1].
\]

where \( \psi_Q(\lambda) = \log \mathbb{E}_Q[\exp(\lambda L)] \) with \( L \triangleq \frac{dP}{dQ} \)
Assumptions

**Assumption 1** $K \geq 2$, $K/n$ is bounded away from one as $n \to \infty$.

**Assumption 2** There exists a constant $C$ such that for all $n$,

$$
\psi''_Q(\lambda) \leq C \min\{D(P\|Q), D(Q\|P)\}, \quad \forall \lambda \in [-1, 1].
$$

where $\psi_Q(\lambda) = \log \mathbb{E}_Q[\exp(\lambda L)]$ with $L \triangleq \frac{dP}{dQ}$

**Notes:**

- $\psi''_Q(\lambda) = \text{var}_{Q_\lambda}(L)$, where $dQ_\lambda = \exp(\lambda L - \psi_Q(\lambda))dQ$,
- Assumption 2 holds if $L \triangleq \frac{dP}{dQ}$ is bounded. e.g. in Bernoulli case, if $\frac{p}{q}$ and $\frac{1-p}{1-q}$ are bounded away from 0 and $\infty$.
- Assumption 2 holds if $P = \mathcal{N}(\mu, 1)$ and $Q = \mathcal{N}(0, 1)$ (more generally, exp. families under mild conditions): Then $L(x) = \mu(x - \frac{\mu}{2})$, $D(P\|Q) = D(Q\|P) = \mu^2/2$ and $\psi''_Q(\lambda) \equiv \mu^2$. 
Main result for weak (aka almost exact) recovery

**Theorem (Weak recovery)**

If

\[
K \cdot D(P\parallel Q) \to \infty \quad \text{and} \quad \liminf_{n \to \infty} \frac{(K - 1)D(P\parallel Q)}{\log \frac{n}{K}} > 2, \tag{1}
\]

then

\[
\mathbb{P}\{|\hat{C}_{\text{ML}} \triangle C^*| \leq 2K \epsilon\} \geq 1 - e^{-\Omega(K/\epsilon)},
\]

where \(\epsilon = \frac{1}{\sqrt{KD(P\parallel Q)}}\).

If there exists \(\hat{C}\) such that \(\mathbb{E}[|\hat{C} \triangle C^*|] = o(K)\), then

\[
K \cdot D(P\parallel Q) \to \infty \quad \text{and} \quad \liminf_{n \to \infty} \frac{(K - 1)D(P\parallel Q)}{\log \frac{n}{K}} \geq 2. \tag{2}
\]
Main result for weak (aka almost exact) recovery

**Theorem (Weak recovery)**

If

\[ K \cdot D(P\|Q) \to \infty \quad \text{and} \quad \liminf_{n \to \infty} \frac{(K - 1)D(P\|Q)}{\log \frac{n}{K}} > 2, \tag{1} \]

then

\[ \mathbb{P}\{|\hat{C}_{ML} \triangle C^*| \leq 2K\epsilon\} \geq 1 - e^{-\Omega(K/\epsilon)}, \]

where \( \epsilon = 1/\sqrt{KD(P\|Q)} \). [union bound]

If there exists \( \hat{C} \) such that \( \mathbb{E}[|\hat{C} \triangle C^*|] = o(K) \), then

\[ K \cdot D(P\|Q) \to \infty \quad \text{and} \quad \liminf_{n \to \infty} \frac{(K - 1)D(P\|Q)}{\log \frac{n}{K}} \geq 2. \tag{2} \]
Main result for weak (aka almost exact) recovery

**Theorem (Weak recovery)**

If

\[ K \cdot D(P \parallel Q) \to \infty \quad \text{and} \quad \liminf_{n \to \infty} \frac{(K - 1)D(P \parallel Q)}{\log \frac{n}{K}} > 2, \tag{1} \]

then

\[ \mathbb{P}\{|\hat{C}_{ML} \triangle C^*| \leq 2K\epsilon\} \geq 1 - e^{-\Omega(K/\epsilon)}, \]

where \( \epsilon = 1/\sqrt{KD(P \parallel Q)} \). [union bound]

If there exists \( \hat{C} \) such that \( \mathbb{E}[|\hat{C} \triangle C^*|] = o(K) \), then

\[ K \cdot D(P \parallel Q) \to \infty \quad \text{and} \quad \liminf_{n \to \infty} \frac{(K - 1)D(P \parallel Q)}{\log \frac{n}{K}} \geq 2. \tag{2} \]

[genie argument]
Main result for weak (aka almost exact) recovery

**Theorem (Weak recovery)**

If

\[ K \cdot D(P \| Q) \to \infty \quad \text{and} \quad \liminf_{n \to \infty} \frac{(K - 1)D(P \| Q)}{\log \frac{n}{K}} > 2, \]  

then

\[ \mathbb{P}\{\left| \hat{C}_{ML} \Delta C^* \right| \leq 2K\epsilon\} \geq 1 - e^{-\Omega(K/\epsilon)}, \]

where \( \epsilon = \frac{1}{\sqrt{KD(P \| Q)}} \). [union bound]

If there exists \( \hat{C} \) such that \( \mathbb{E}[\left| \hat{C} \triangle C^* \right|] = o(K) \), then

\[ K \cdot D(P \| Q) \to \infty \quad \text{and} \quad \liminf_{n \to \infty} \frac{(K - 1)D(P \| Q)}{\log \frac{n}{K}} \geq 2. \]  

[genie argument] [mutual information / rate distortion]
Main result for exact recovery

**Theorem (Exact recovery)**

If weak recovery suff. condition (1) and the following hold:

\[
\liminf_{n \to \infty} \frac{KE_Q \left( \frac{1}{K} \log \frac{n}{K} \right)}{\log n} > 1. \tag{3}
\]

then the maximum likelihood estimator satisfies \( P\{\hat{C}_{ML} = C^*\} \to 1. \)

If there exists an estimator \( \hat{C} \) such that \( P\{\hat{C} = C^*\} \to 1 \), then weak recovery nece. condition (2) and the following hold:

\[
\liminf_{n \to \infty} \frac{KE_Q \left( \frac{1}{K} \log \frac{n}{K} \right)}{\log n} \geq 1. \tag{4}
\]

Note: \( E_Q(\theta) = \psi^*_Q(\theta) \triangleq \sup_{\lambda \in \mathbb{R}} \lambda \theta - \psi_Q(\lambda) \) (Legendre transform of the log moment generating function)
VOTING WITH SUCCESSIVE WITHHOLDING ALGORITHM

1. Input: $n \in \mathbb{N}$, $K > 0$, distributions $P$, $Q$; observed matrix $A$;
   $\delta \in (0, 1)$ with $1/\delta$, $n\delta \in \mathbb{N}$.

2. (Partition): Partition $[n]$ into $1/\delta$ subsets $S_k$ of size $n\delta$.

3. (Approximate Recovery) For each $k = 1, \ldots, 1/\delta$, produce
   $\hat{C}_k \approx C^* \cap ([n]\setminus S_k)$.

4. (Cleanup) For each $k = 1, \ldots, 1/\delta$ compute
   $r_i = \sum_{j \in \hat{C}_k} L_{ij}$ for all
   $i \in S_k$ and return $\hat{\mathcal{C}}$, the set of $K$ indices in $[n]$ with the largest
   values of $r_i$. 
Lemma (Voting Converse for Exact Recovery)

Assume that $K \to \infty$ and $\lim \sup K/n < 1$. Let $L_i$ denote i.i.d. copies of $\log \frac{dP}{dQ}$. If there exists an estimator $\hat{C}$ such that $P\{\hat{C} = C^*\} \to 1$, then for any $K_o \to \infty$ such that $K_o = o(K)$, there exists a threshold $\theta_n$

\[
P \left[ \sum_{i=1}^{K-K_o} L_i \leq (K - 1)\theta_n - (K_o - 1)D(P\|Q) - 6\sigma \right] \leq \frac{2}{K_o}
\]

\[
Q \left[ \sum_{i=1}^{K-1} L_i \geq (K - 1)\theta_n \right] \leq \frac{1}{n - K},
\]

where $\sigma^2 = K_o \text{var}_P (L_1)$.
Let $\text{ALGX}=$ semi-definite programming relaxation or belief propagation. For both planted dense subgraph (Bernoulli) and submatrix localization (Gaussian):

- $K = \omega\left(\frac{n}{\log n}\right)$: $\text{ALGX}$ attains the info-theoretic limit with sharp constants.
- $K = \Theta\left(\frac{n}{\log n}\right)$: $\text{ALGX}$ is order-wise optimal, but strictly suboptimal by a constant factor.
- $K = o\left(\frac{n}{\log n}\right)$ and $K \to \infty$: $\text{ALGX}$ is order-wise suboptimal.
Open problem

Can the computational gap for exact recovery be bridged by any polynomial time algorithm? (SoS hardness result or reduction to planted clique would offer further evidence for “no” answer.)
Can the computational gap for exact recovery be bridged by any polynomial time algorithm? (SoS hardness result or reduction to planted clique would offer further evidence for “no” answer.)

Thank you!