Spectral algorithm without trimming or cleaning works for exact recovery in SBM

Yiqiao (Joe) Zhong

Princeton University

with Emmanuel Abbe, Jianqing Fan and Kaizheng Wang

January 11, 2018
Graphs are everywhere
Graphs are everywhere
Graphs are everywhere
Let $G = (V, E)$ be a graph with $n$ vertices, i.e., $|V| = n$. 

Goal: partition the vertices into several blocks (subsets) such that some criterion is satisfied.
Let $G = (V, E)$ be a graph with $n$ vertices, i.e., $|V| = n$.

**Goal**: partition the vertices into several blocks (subsets) such that some criterion is satisfied.
Let $G = (V, E)$ be a graph with $n$ vertices, i.e., $|V| = n$.

Goal: partition the vertices into several blocks (subsets) such that some criterion is satisfied.
Let $G = (V, E)$ be a graph with $n$ vertices, i.e., $|V| = n$.

**Goal**: partition the vertices into several blocks (subsets) such that some criterion is satisfied.

The deterministic approach vs. statistical model approach.
Stochastic Block Model (SBM): (two equal-sized blocks)

Goal: recover unknown index set $J \in [n]$ with $|J| = n/2$.

Observations:

$$A_{ij} \sim \begin{cases} 
    \text{Ber}(p_n), & \text{if } i, j \in J \text{ or } i, j \in J^c \\
    \text{Ber}(q_n), & \text{otherwise}
\end{cases}$$

for all $i \leq j$. Assume that $p_n \geq q_n$. Allow self-loops.
Stochastic Block Model (SBM): (two equal-sized blocks)

**Goal**: recover unknown index set $J \in [n]$ with $|J| = n/2$.

**Observations**:

$$A_{ij} \sim \begin{cases} 
\text{Ber}(p_n), & \text{if } i, j \in J \text{ or } i, j \in J^c \\
\text{Ber}(q_n), & \text{otherwise}
\end{cases}$$

for all $i \leq j$. Assume that $p_n \geq q_n$. Allow self-loops.
Equivalently, recover (estimate) block membership vector \( x \in \{\pm 1\}^n \).
Equivalentlly, recover (estimate) block membership vector $x \in \{\pm 1\}^n$.

**Exact recovery** find $\hat{x} = \hat{x}(A)$ such that $\hat{x} = x$ w.p. $1 - o(1)$. 

**Related phase transition** for weak recovery (detection): find $\hat{x}$ such that $\frac{1}{n} \# \{i \in [n]: \hat{x}_i = x_i\} > 0.5 + \varepsilon$ w.p. $1 - o(1)$. 

If $\frac{p}{n} = a \log n / n$, $\frac{q}{n} = b \log n / n$, then information limit for exact recovery: $\sqrt{a} - \sqrt{b} > \sqrt{2}$. 

No estimator achieves exact recovery if $\sqrt{a} - \sqrt{b} < \sqrt{2}$. 
Equivalently, recover (estimate) block membership vector $x \in \{\pm 1\}^n$.

**Exact recovery** find $\hat{x} = \hat{x}(A)$ such that $\hat{x} = x$ w.p. $1 - o(1)$.

If $p_n = a \log n / n$, $q_n = b \log n / n$, then **information limit** for exact recovery:

$$\sqrt{a} - \sqrt{b} > \sqrt{2}.$$ 

No estimator achieves exact recovery if $\sqrt{a} - \sqrt{b} < \sqrt{2}$. 
Equivalently, recover (estimate) block membership vector $x \in \{\pm 1\}^n$.

**Exact recovery** find $\hat{x} = \hat{x}(A)$ such that $\hat{x} = x$ w.p. $1 - o(1)$.

If $p_n = a \log n / n$, $q_n = b \log n / n$, then **information limit** for exact recovery:

$$\sqrt{a} - \sqrt{b} > \sqrt{2}.$$

No estimator achieves exact recovery if $\sqrt{a} - \sqrt{b} < \sqrt{2}$.

**Related phase transition:** **weak recovery (detection):** find $\hat{x}$ such that $\frac{1}{n} \# \{ i \in [n] : \hat{x}_i = x_i \} > 0.5 + \varepsilon$ w.p. $1 - o(1)$. 
What algorithms?

- Lots of works in the literature.

References: Abbe, Bandeira, and Hall [2014], Abbe and Sandon [2015], Yun and Proutiere [2016].
What algorithms?

- Lots of works in the literature.
- Efficient methods that works down to the threshold:

  Semidefinite relaxation;
  Spectral method with local refinement.

(Incomplete) references: Abbe et al. [2014], Abbe and Sandon [2015], Yun and Proutiere [2016].
What algorithms?

- Lots of works in the literature.
- Efficient methods that works down to the threshold:
  - Semidefinite relaxation;
  - Spectral method with local refinement.
What algorithms?

- Lots of works in the literature.
- Efficient methods that works down to the threshold:
  - Semidefinite relaxation;
  - Spectral method with local refinement.

(Incomplete) references: Abbe et al. [2014], Abbe and Sandon [2015], Yun and Proutiere [2016]
What algorithms?

- Lots of works in the literature.
- Efficient methods that works down to the threshold:
  - Semidefinite relaxation;
  - Spectral method with local refinement.
- (Incomplete) references: Abbe et al. [2014], Abbe and Sandon [2015], Yun and Proutiere [2016]
- One-shot spectral method works?
Does spectral algorithm work?

- Rank-2 structure (up to permutation):

\[
\mathbb{E}A = \left( \begin{array}{cc} p_n \mathbf{1}_{n/2} \times \mathbf{1}_{n/2} & q_n \mathbf{1}_{n/2} \times \mathbf{1}_{n/2} \\
q_n \mathbf{1}_{n/2} \times \mathbf{1}_{n/2} & p_n \mathbf{1}_{n/2} \times \mathbf{1}_{n/2} \end{array} \right) \cdot J J_c
\]

The first eigenvector \( u_1^* = \frac{1}{\sqrt{n}} \mathbf{1}_n \); the second

\[
u_2^* = \frac{1}{\sqrt{n}} \left( \begin{array}{c} \mathbf{1}_{n/2} \\ -\mathbf{1}_{n/2} \end{array} \right).
\]
Does spectral algorithm work?

- Rank-2 structure (up to permutation):

\[
\mathbb{E}A = \begin{pmatrix}
 p_n \mathbf{1}_{n/2 \times n/2} & q_n \mathbf{1}_{n/2 \times n/2} \\
 q_n \mathbf{1}_{n/2 \times n/2} & p_n \mathbf{1}_{n/2 \times n/2}
\end{pmatrix} \cdot J_J^c
\]

The first eigenvector \( u_1^* = \frac{1}{\sqrt{n}} \mathbf{1}_n \); the second

\[
u_2^* = \frac{1}{\sqrt{n}} \begin{pmatrix} \mathbf{1}_{n/2} ; -\mathbf{1}_{n/2} \end{pmatrix}.
\]

- Target: \( u_2 \), i.e., the second eigenvector of \( A \).
Does spectral algorithm work?

- **Good news** for exact recovery.

\[ p_n = a \log n / n, \quad q_n = b \log n / n, \]  
so  
\[ \lambda^*_1 = a + b 2 \log n, \quad \lambda^*_2 = a - b 2 \log n. \]

Eigenvectors preserve ordering:

\[ \begin{array}{c}
\text{Weyl's inequality: + Feige-Ofek's inequality:} \\
\parallel A - E_A \parallel_2 = O(\sqrt{\log n})
\end{array} \]

Contrast with sparser regime (weak recovery).

\[ ^1 \text{See Feige and Ofek [2005].} \]
Does spectral algorithm work?

- **Good news** for exact recovery.

- recall \( p_n = a \log n / n \), \( q_n = b \log n / n \), so \( \lambda_1^* = \frac{a+b}{2} \log n \), \( \lambda_2^* = \frac{a-b}{2} \log n \).

\(^1\)See Feige and Ofek [2005].
Does spectral algorithm work?

- **Good news** for exact recovery.

- recall $p_n = a \log n / n$, $q_n = b \log n / n$, so $\lambda_1^* = \frac{a+b}{2} \log n$, $\lambda_2^* = \frac{a-b}{2} \log n$.

- eigenvalues preserve ordering:

\[ \lambda_3 \ldots \lambda_2 \lambda_1 \]

\[ 0 \quad \lambda_2^* \quad \lambda_1^* \]

\[ O(\sqrt{\log n}) \]

\[ \Theta(\log n) \]

---

1See Feige and Ofek [2005].
Does spectral algorithm work?

- **Good news** for exact recovery.

recall $p_n = a \log n / n$, $q_n = b \log n / n$, so $\lambda_1^* = \frac{a+b}{2} \log n$, $\lambda_2^* = \frac{a-b}{2} \log n$.

- eigenvalues preserve ordering:

Weyl’s inequality + Feige-Ofek’s\(^1\): w.h.p

$$\|A - \mathbb{E}A\|_2 = O(\sqrt{\log n}).$$

---

\(^1\) See Feige and Ofek [2005].
Does spectral algorithm work?

- **Good news** for exact recovery.
- recall $p_n = a \log n / n$, $q_n = b \log n / n$, so $\lambda_1^* = \frac{a+b}{2} \log n$, $\lambda_2^* = \frac{a-b}{2} \log n$.
- eigenvalues preserve ordering:

- Weyl’s inequality + Feige-Ofek’s\(^1\): w.h.p
  \[ \| A - \mathbb{E} A \|_2 = \mathcal{O}(\sqrt{\log n}). \]

- Contrast with sparser regime (weak recovery).

\(^1\)See Feige and Ofek [2005].
Does spectral algorithm work?

- Implies consistency: $|\langle u_2^*, u_2 \rangle| \xrightarrow{p} 1$. 

Key insight:

$u^* = Au^* \lambda^2 \approx Au_2^* \lambda^2$. 

Under the $\ell_\infty$ norm.

That is, $u_2^* = Au_2^* \lambda_2$. 

Linearized (first order) term

Negligible (higher order) term
Does spectral algorithm work?

- Implies consistency: $|\langle u_2^*, u_2 \rangle| \xrightarrow{p} 1$.

- So, $1 - o(1)$ fraction of vertices have correct signs...but doesn’t solve exact recovery.
Does spectral algorithm work?

- Implies consistency: \( |\langle u_2^*, u_2 \rangle| \xrightarrow{p} 1. \)

- So, \( 1 - o(1) \) fraction of vertices have correct signs...but doesn’t solve exact recovery.

- Need **uniform** control. **Key insight:**
Does spectral algorithm work?

- Implies consistency: $|\langle u^*_2, u_2 \rangle| \xrightarrow{p} 1$.
- So, $1 - o(1)$ fraction of vertices have correct signs...but doesn’t solve exact recovery.
- Need **uniform** control. **Key insight:**

$$u_2 = \frac{Au^*_2}{\lambda^*_2} + \left( u_2 - \frac{Au^*_2}{\lambda^*_2} \right)$$

under the $\ell_\infty$ norm.
Does spectral algorithm work?

- Implies consistency: $|\langle u^*_2, u_2 \rangle| \xrightarrow{p} 1$.
- So, $1 - o(1)$ fraction of vertices have correct signs...but doesn’t solve exact recovery.
- Need **uniform** control. **Key insight:**

$$u_2 = \frac{Au^*_2}{\lambda^*_2} + \left( u_2 - \frac{Au^*_2}{\lambda^*_2} \right)$$

under the $\ell_\infty$ norm.
- That is, $u_2 = \frac{Au_2}{\lambda_2} \approx \frac{Au^*_2}{\lambda^*_2}$. 
Does spectral algorithm work?

**Left:** From a typical realization of $A$, distribution of 5000 coordinates.  

**Right:** From 100 realizations, three errors

1. $\sqrt{n}\|u_2 - u_2^*\|_\infty$  
2. $\sqrt{n}\|Au_2^*/\lambda_2^* - u_2^*\|_\infty$  
3. $\sqrt{n}\|u_2 - Au_2^*/\lambda_2^*\|_\infty$. 


Does spectral algorithm work? Yes!

Theorem

If $A \sim \text{SBM}(n, a\frac{\log n}{n}, b\frac{\log n}{n}, J)$, then with probability $1 - O(n^{-3})$ we have

$$\min_{s \in \{\pm 1\}} \| u_2 - sA u_2^*/\lambda_2^* \|_\infty \leq \frac{C}{\sqrt{n \log \log n}}.$$ 

where $C = C(a, b)$ is some constant only depending on $a$ and $b$. 
Does spectral algorithm work? Yes!

Let $\hat{x}_{eig}(A) = \text{sign}(u_2)$ be the simple eigenvector estimator.
Let $\hat{x}_{eig}(A) = \text{sign}(u_2)$ be the simple eigenvector estimator.

**Corollary**

Suppose $a > b > 0$ with $\sqrt{a} \neq \sqrt{b} + \sqrt{2}$. Then, whenever the MLE is successful, in the sense that $\hat{x}_{MLE} = x$ (up to sign) with probability $1 - o(1)$, we have

$$\hat{x}_{eig}(A) = \hat{x}_{MLE}(A) = x$$

with probability $1 - o(1)$, where $x$ is the sign indicator of the true communities.
Eigenvector analysis: a formal setup

Random matrix $A \in \mathbb{R}^{n \times n}$ symmetric, $(A_{ij})_{i \geq j}$ independent,

$E_A = A^*$.

Eigenpairs: $A \sim \{\lambda_j, u_j\}_{j=1}^n$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$;

$A^* \sim \{\lambda^*_j, u^*_j\}_{j=1}^n$, $\lambda^*_1 \geq \lambda^*_2 \geq \ldots \geq \lambda^*_n$.

Assume $A^*$ has rank $r = O(1)$, and $\lambda^*_1 \approx \lambda^*_r$.

How does $u_k$ look like?

Eigengap: $\Delta^* = \min\{\lambda^*_k - 1 - \lambda^*_k, \lambda^*_k - \lambda^*_k + 1\}$ for $k \in [r]$.

Spectral norm concentration: there exists $\gamma = o(1)$ such that $\|A - A^*\|_2 \leq \gamma \Delta^*$ w.h.p.

Delocalization (incoherence): $\|A^*\|_2 \to \infty \leq \gamma \Delta^*$, $\|u^*_k\|_\infty \leq \gamma$.

$\|X\|_2 \to \infty = \max_{m \in [n]} \|X_m\|_2$ is the maximum $\ell_2$ norm of rows.
**Eigenvector analysis: a formal setup**

**Random matrix:** $A \in \mathbb{R}^{n \times n}$ symmetric, $(A_{ij})_{i \geq j}$ independent, $\mathbb{E} A = A^*$.

**Eigenpairs:** $A \sim \{\lambda_j, u_j\}_{j=1}^n$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$;

$A^* \sim \{\lambda_j^*, u_j^*\}_{j=1}^n$, $\lambda_1^* \geq \lambda_2^* \geq \cdots \geq \lambda_n^*$.

Assume $A^*$ has rank $r$, $r = O(1)$, and $\lambda_i^* \asymp \lambda_i^*$. Fix $k \in [r]$. **How does $u_k$ look like?**
**Eigenvector analysis: a formal setup**

**Random matrix**: \( A \in \mathbb{R}^{n \times n} \) symmetric, \( (A_{ij})_{i \geq j} \) independent, \( \mathbb{E}A = A^* \).

**Eigenpairs**: \( A \sim \{ \lambda_j, u_j \}_{j=1}^n, \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n; \)
\( A^* \sim \{ \lambda_j^*, u_j^* \}_{j=1}^n, \lambda_1^* \geq \lambda_2^* \geq \cdots \geq \lambda_n^* \).

Assume \( A^* \) has rank \( r, r = O(1) \), and \( \lambda_1^* \approx \lambda_r^* \). Fix \( k \in [r] \). How does \( u_k \) look like?

**Eigengap**: \( \Delta^* = \min\{\lambda^*_{k-1} - \lambda^*_k, \lambda^*_k - \lambda^*_{k+1}\} \) for \( k \in [r] \).

**Spectral norm concentration**: there exists \( \gamma = o(1) \) such that \( \|A - A^*\|_2 \leq \gamma \Delta^* \) w.h.p.

**Delocalization (incoherence)**: \( \|A^*\|_{2 \rightarrow \infty} \leq \gamma \Delta^*, \|u^*_k\|_\infty \leq \gamma \).

\( \star \) \( \|X\|_{2 \rightarrow \infty} = \max_{m \in [n]} \|X_m\|_2 \) is the maximum \( \ell_2 \) norm of rows.
Row concentration assumption

\[ \phi \colon [0, +\infty) \to [0, +\infty) \text{ non-decreasing}, \]
\[ \phi \left( \frac{x}{x} \right) \text{ non-increasing on } (0, +\infty). \]

For any fixed \( w \in \mathbb{R}^n \) and \( m \in \mathbb{N} \),
\[ |(A - A^*)_m \cdot w| \leq \Delta^* \left\| w \right\|_\infty \phi \left( \sqrt{\frac{n}{2}} \left\| w \right\|_2 \right) \]
with probability \( 1 - o(n^{-1}) \).

\( \phi \) is allowed to change with \( n \).

Typical choices of \( \phi \) for Gaussian noise and Bernoulli noise.
Row concentration assumption

\[ \varphi : [0, +\infty) \to [0, +\infty) \text{ non-decreasing, } \varphi(x)/x \text{ non-increasing on } (0, +\infty). \]

For any fixed \( w \in \mathbb{R}^n \) and \( m \in [n] \),

\[ |(A - A^*)_m \cdot w| \leq \Delta^* \| w \|_\infty \varphi \left( \frac{\| w \|_2}{\sqrt{n} \| w \|_\infty} \right) \]

with probability \( 1 - o(n^{-1}) \). \( \varphi \) is allowed to change with \( n \).

\[ \varphi(x) \]

Bernoulli \hspace{2cm} Gaussian

Typical choices of \( \varphi \) for Gaussian noise and Bernoulli noise.
**Theorem**: Let \(s = \text{sgn}(u_k^T u_k^*)\). With probability \(1 - o(1)\),

\[
\|su_k - Au_k^*/\lambda_k^*\|_\infty \lesssim (\gamma + \varphi(\gamma))(1 + \varphi(1))\|u^*\|_\infty.
\]

Usually \(\varphi(1) = O(1)\). Then \(Au_k^*/\lambda_k^*\) approximates \(u_k\) well since

\[
\|su_k - Au_k^*/\lambda_k^*\|_\infty = o(\|u^*\|_\infty).
\]

Indeed, the first-order approximation (linearization) idea is correct.
One-slide proof idea

Proof idea = leave-one-out decoupling + Davis-Kahan's.

\[ u_k = A u_k / \lambda_k. \] Observe: \( A \) and \( u_k \) are weakly correlated.

For each \( m \in [n] \), introduce \( n \times n \) matrix

\[ A(m)_{ij} = A_{ij} \{ i \neq m, j \neq m \}. \]

Let \( u(m)_k \) be the eigenvector of \( A(m) \).

Decoupling: independence in \( m \)th coordinate of \( Au(m)_k \).

Davis-Kahan: \( \| u_k - u(m)_k \|_2 \) very small.
Proof idea = leave-one-out decoupling + Davis-Kahan’s.
Proof idea = leave-one-out decoupling + Davis-Kahan’s.

$u_k = Au_k / \lambda_k$. Observe: $A$ and $u_k$ are weakly correlated.
Proof idea = leave-one-out decoupling $+$ Davis-Kahan’s.

$u_k = Au_k / \lambda_k$. Observe: $A$ and $u_k$ are weakly correlated.

For each $m \in [n]$, introduce $n \times n$ matrix

$$ [A^{(m)}]_{ij} = A_{ij} 1_{\{i \neq m, j \neq m\}}. $$

Let $u^{(m)}_k$ be the eigenvector of $A^{(m)}$. 
One-slide proof idea

- Proof idea = leave-one-out decoupling + Davis-Kahan’s.
- $u_k = Au_k / \lambda_k$. Observe: $A$ and $u_k$ are weakly correlated.
- For each $m \in [n]$, introduce $n \times n$ matrix
  \[
  [A^{(m)}]_{ij} = A_{ij} \mathbf{1}_{\{i \neq m, j \neq m\}}.
  \]

  Let $u_k^{(m)}$ be the eigenvector of $A^{(m)}$.

- **Decoupling**: independence in $m$th coordinate of $Au_k^{(m)}$.
- **Dacis-Kahan**: $\|u_k - u_k^{(m)}\|_2$ very small.
Back to SBM, what about the linearized term?

Lemma (E. Abbe, A. Bandeira, G. Hall, 2014)

Suppose $a > b$, $\{W_i\}_{i=1}^{n/2}$ are i.i.d Ber($\frac{a \log n}{n}$), and $\{Z_i\}_{i=1}^{n/2}$ are i.i.d. Ber($\frac{b \log n}{n}$), independent of $\{W_i\}_{i=1}^{n/2}$. For any $\varepsilon \in \mathbb{R}$, we have the following tail bound:

$$\mathbb{P}\left(\sum_{i=1}^{n/2} W_i - \sum_{i=1}^{n/2} Z_i \leq \varepsilon \log n\right) \leq n^{-\left(\sqrt{a} - \sqrt{b}\right)^2/2 + \varepsilon \log (a/b)/2}.$$
Corollary

(i) If $\sqrt{a} - \sqrt{b} > \sqrt{2}$, then there exists $\eta = \eta(a, b) > 0$ and $s \in \{\pm 1\}$ such that with probability $1 - o(1)$,

$$\sqrt{n} \min_{i \in [n]} sz_i(u_2)_i \geq \eta.$$ 

As a consequence, our spectral method achieves exact recovery.

(ii) Let the misclassification rate be $r(\hat{z}, z)$. If $\sqrt{a} - \sqrt{b} \in (0, \sqrt{2}]$, then

$$\mathbb{E} r(\hat{z}, z) \leq n^{-(1+o(1))(\sqrt{a} - \sqrt{b})^2/2}.$$ 

This upper bound matches the minimax lower bound.
y-axis: $a$, x-axis: $b$, red curve: $\sqrt{a} - \sqrt{b} = \pm \sqrt{2}$. Fix $n = 300$. Heatmap from 100 realizations.
Simulations

Log plot of misclassification rate. Fix $b = 2$. x-axis: $a \in [2, 8]$, y-axis: $\log r(\hat{x}, x)/\log n$.

**Red**: theoretical, **black**: $n = 100$, **green**: $n = 500$, **blue**: $n = 5000$
Beyond SBM: 😑

- Extension to eigenspaces. ✓

Unsolved problems: 😞
Beyond SBM:

- Extension to eigenspaces. ✓
- Synchronization problems ($\mathbb{Z}_2$-synchronization). ✓

Unsolved problems: 😞

2References: Zhong and Boumal [2017], Chen et al. [2017], etc.
Generalizations and open problems

**Beyond SBM:**

- Extension to eigenspaces. ✓
- Synchronization problems ($\mathbb{Z}_2$-synchronization). ✓
- Matrix completion. ✓

**Unsolved problems:**

2 References: Zhong and Boumal [2017], Chen et al. [2017], etc.
Beyond SBM:

- Extension to eigenspaces. ✓
- Synchronization problems ($\mathbb{Z}_2$-synchronization). ✓
- Matrix completion. ✓
- Analyze iterative algorithms. ² ✓

Unsolved problems:

²References: Zhong and Boumal [2017], Chen et al. [2017], etc.
Generalizations and open problems

Beyond SBM:

- Extension to eigenspaces. ✓
- Synchronization problems ($\mathbb{Z}_2$-synchronization). ✓
- Matrix completion. ✓
- Analyze iterative algorithms. ✓

Unsolved problems:

- How to analyze normalized Laplacian?

---

References: Zhong and Boumal [2017], Chen et al. [2017], etc.
Generalizations and open problems

**Beyond SBM:**

- Extension to eigenspaces. ✓
- Synchronization problems ($\mathbb{Z}_2$-synchronization). ✓
- Matrix completion. ✓
- Analyze iterative algorithms. ✓

**Unsolved problems:**

- How to analyze normalized Laplacian?
- More than two blocks?

---

References: Zhong and Boumal [2017], Chen et al. [2017], etc.
Thank you!


