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

Yiqiao (Joe) Zhong

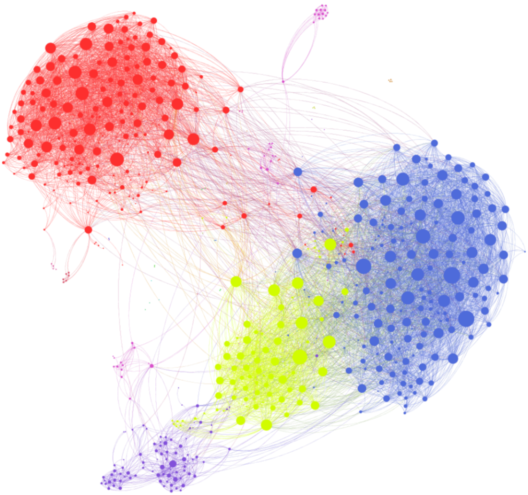
Princeton University

with **Emmanuel Abbe, Jianqing Fan and Kaizheng Wang**

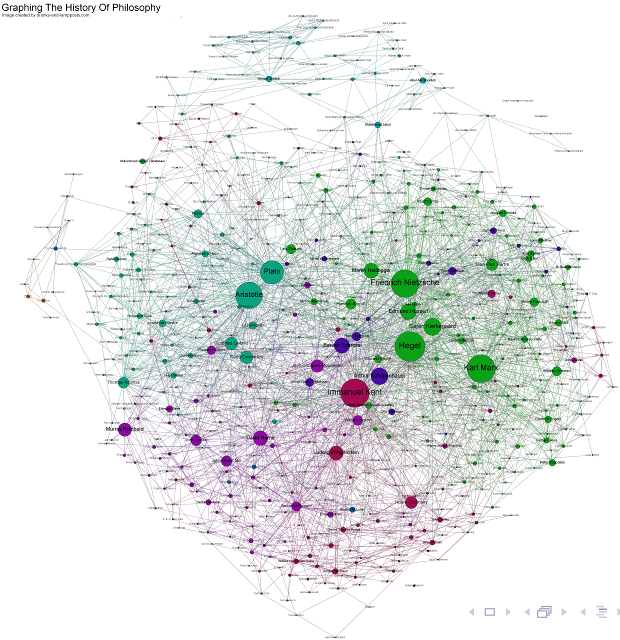
January 11, 2018

Graphs are everywhere

Graphs are everywhere



Graphs are everywhere



Clustering of graphs

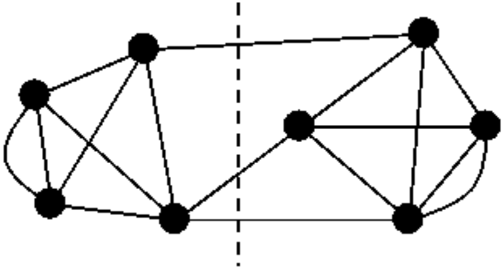
- Let $G = (V, E)$ be a graph with n vertices, i.e., $|V| = n$.

- 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., $|V| = n$.

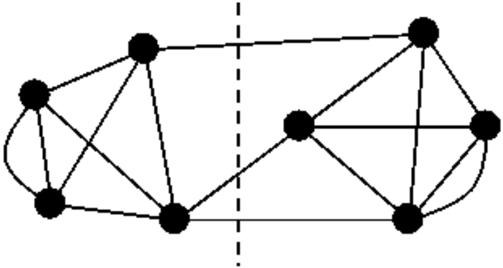
- **Goal** : partition the vertices into several blocks (subsets) such that some criterion is satisfied.



Clustering of graphs

- 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:

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

for all $i \leq j$. Assume that $p_n \geq q_n$. Allow self-loops.

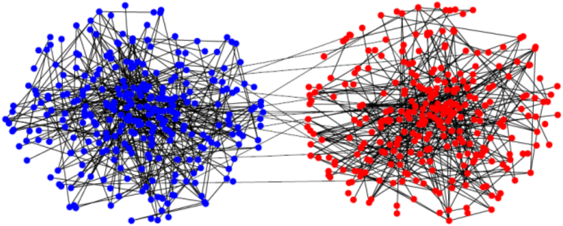
- **Stochastic Block Model (SBM):** (two equal-sized blocks)

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

  Observations:

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

for all $i \leq j$. Assume that $p_n \geq q_n$. Allow self-loops.



- Equivalently, recover (estimate) block membership vector $x \in \{\pm 1\}^n$.

- Equivalently, recover (estimate) block membership vector $x \in \{\pm 1\}^n$.
- **Exact recovery** find $\widehat{x} = \widehat{x}(A)$ such that $\widehat{x} = x$ w.p. $1 - o(1)$.

- Equivalently, recover (estimate) block membership vector $x \in \{\pm 1\}^n$.
- **Exact recovery** find $\widehat{x} = \widehat{x}(A)$ such that $\widehat{x} = x$ 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 $\widehat{x} = \widehat{x}(A)$ such that $\widehat{x} = x$ 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 $\widehat{x}$ such that $\frac{1}{n} \# \{i \in [n] : \widehat{x}_i = x_i\} > 0.5 + \varepsilon$ w.p. $1 - o(1)$.

What algorithms?

- Lots of works in the literature.

What algorithms?

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

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?

- Rank-2 structure (up to permutation):

$$\begin{array}{cc} \mathbb{E}A = \begin{pmatrix} p_n \mathbf{1}_{\frac{n}{2}\times\frac{n}{2}} & q_n \mathbf{1}_{\frac{n}{2}\times\frac{n}{2}} \\ q_n \mathbf{1}_{\frac{n}{2}\times\frac{n}{2}} & p_n \mathbf{1}_{\frac{n}{2}\times\frac{n}{2}} \end{pmatrix}. & \begin{array}{c} J \\ J^c \end{array} \end{array}$$

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

$$u_2^* = \frac{1}{\sqrt{n}} \big(\underset{J}{\mathbf{1}_{n/2}} \;\; ; \;\; \underset{J^c}{-\mathbf{1}_{n/2}} \big).$$

Does spectral algorithm work?

- Rank-2 structure (up to permutation):

$$\mathbb{E}A = \begin{array}{cc} \begin{pmatrix} p_n \mathbf{1}_{\frac{n}{2}\times\frac{n}{2}} & q_n \mathbf{1}_{\frac{n}{2}\times\frac{n}{2}} \\ q_n \mathbf{1}_{\frac{n}{2}\times\frac{n}{2}} & p_n \mathbf{1}_{\frac{n}{2}\times\frac{n}{2}} \end{pmatrix}. & \begin{matrix} J \\ J^c \end{matrix} \end{array}$$

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

$$u_2^* = \frac{1}{\sqrt{n}} \begin{array}{ccc} (\quad \mathbf{1}_{n/2} & ; & -\mathbf{1}_{n/2} \quad). \\ J & & J^c \end{array}$$

- Target: u_2, i.e., the second eigenvector of A.

Does spectral algorithm work?

- **Good news** for exact recovery.

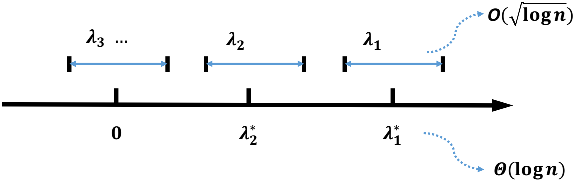
[1] See Feige and Ofek [2005].

Does spectral algorithm work?

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

[1] See Feige and Ofek [2005].

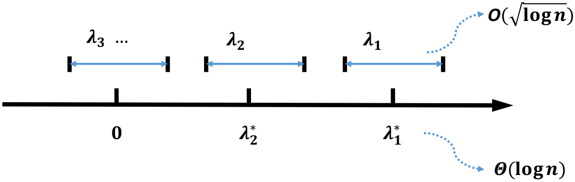
- **Good news** for exact recovery.
- recall $p_n = a \log n/n$, $q_n = b \log n/n$, so $\lambda_1^* = \frac{a+b}{2} \log n$, $\lambda_2^* = \frac{a-b}{2} \log n$.
- eigenvalues preserve ordering:



[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:



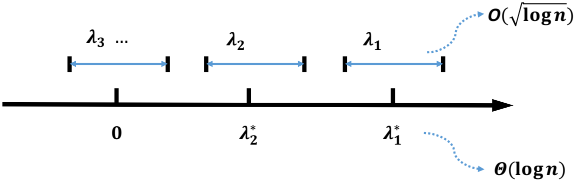
- Weyl's inequality + Feige-Ofek's[1]: w.h.p

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

[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

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

- Contrast with sparser regime (weak recovery).

[1] See Feige and Ofek [2005].

Does spectral algorithm work?

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

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**:

$$\boldsymbol{u}_2 = \underbrace{\frac{\boldsymbol{A}\boldsymbol{u}_2^*}{\boldsymbol{\lambda}_2^*}}_{\text{Linearized (first-order) term}} + \overbrace{\left(\boldsymbol{u}_2 - \frac{\boldsymbol{A}\boldsymbol{u}_2^*}{\boldsymbol{\lambda}_2^*}\right)}^{\text{Negligible (higher-order) term}}$$

under the ℓ_∞ 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**:

Negligible (higher-order) term

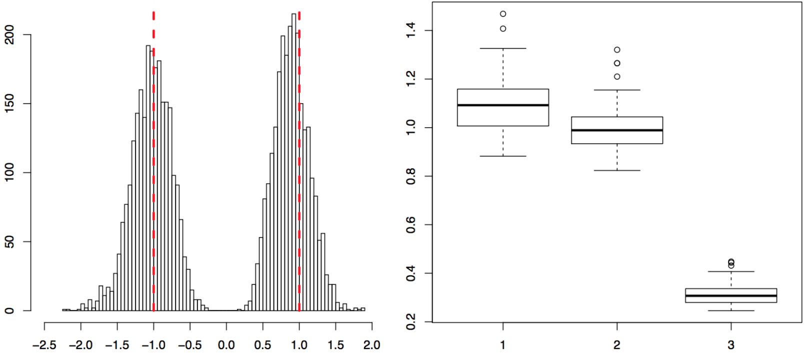
$$\boldsymbol{u}_2 = \frac{\boldsymbol{A}\boldsymbol{u}_2^*}{\boldsymbol{\lambda}_2^*} + \left(\boldsymbol{u}_2 - \frac{\boldsymbol{A}\boldsymbol{u}_2^*}{\boldsymbol{\lambda}_2^*}\right)$$

Linearized (first-order) term

under the ℓ_∞ norm.

- That is, $\mathbf{u_2} = \frac{\mathbf{A}\mathbf{u_2}}{\lambda_2} \approx \frac{\mathbf{A}\mathbf{u_2^*}}{\lambda_2^*}$.

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

$$\min_{s \in \{\pm 1\}} \|u_2 - sAu_2^*/\lambda_2^*\|_\infty \leq \frac{C}{\sqrt{n} \log\log n}.$$

where $C = C(a, b)$ *is some constant only depending on a and b.*

Does spectral algorithm work? Yes!

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

Does spectral algorithm work? Yes!

Let $\widehat{x}_{eig}(A) = \mathrm{sign}(u_2)$ 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 $\widehat{x}_{MLE} = x$ (up to sign) with probability $1 - o(1)$, we have

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

with probability $1 - o(1)$, where x is the sign indicator of the true communities.

Eigenvector analysis: a formal setup

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 $\lambda_1^* \asymp \lambda_r^*$. 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 $\lambda_1^* \asymp \lambda_r^*$. 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): $\|A^*\|_{2\to\infty} \leq \gamma\Delta^*$, $\|u_k^*\|_\infty \leq \gamma$.
★ $\|X\|_{2\to\infty} = \max_{m\in[n]} \|X_{m\cdot}\|_2$ is the maximum ℓ_2 norm of rows.

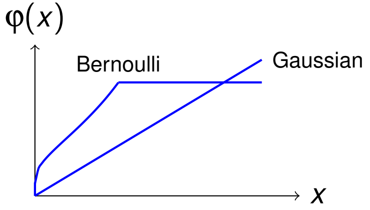
Row concentration assumption

Row concentration assumption

$\varphi : [0,+\infty) \to [0,+\infty)$ non-decreasing, $\varphi(x)/x$ non-increasing on $(0,+\infty)$. For any fixed $w \in \mathbb{R}^n$ and $m \in [n]$,

$$|(A - A^*)_{m\cdot} w| \le \Delta^* \|w\|_\infty \varphi\left(\frac{\|w\|_2}{\sqrt{n}\|w\|_\infty}\right)$$

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



Typical choices of φ for Gaussian noise and Bernoulli noise.

ℓ_∞ approximation of u_k

Theorem: Let $s = \mathrm{sgn}(u_k^T u_k^*)$. 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^*/λ_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

One-slide proof idea

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

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.

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 $A^{(m)}$.

- 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 mth 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:

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

Exact recovery for SBM

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 $1-o(1)$,

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

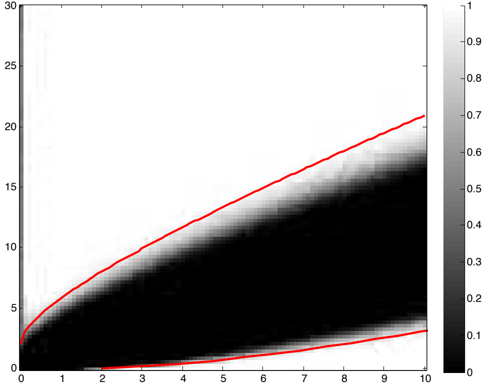
As a consequence, our spectral method **achieves exact recovery**.

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

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

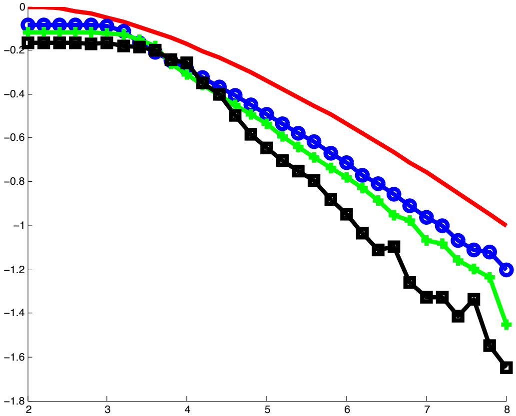
This upper bound **matches the minimax lower bound**.

Simulations



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: $\log r(\widehat{x}, x)/\log n$.
Red: theoretical, **black**: $n = 100$, **green**: $n = 500$, **blue**: $n = 5000$

Beyond SBM: 🙂

- Extension to eigenspaces. ✓

Unsolved problems: 🙁

[2] References: Zhong and Boumal [2017], Chen, Fan, Ma, and Wang [2017], etc.

Beyond SBM: 🙂

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

Unsolved problems: 😕

[2] 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. ✓

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: 😕

[2] 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.[2] ✓

Unsolved problems: 😕

- How to analyze normalized Laplacian?

[2] 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.[2] ✓

Unsolved problems:

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

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

Thank you!

E. Abbe and C. Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pages 670–688, 2015. doi: 10.1109/FOCS.2015.47. URL http://dx.doi.org/10.1109/FOCS.2015.47.

Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *arXiv preprint arXiv:1405.3267*, 2014.

Yuxin Chen, Jianqing Fan, Cong Ma, and Kaizheng Wang. Spectral method and regularized MLE are both optimal for top-K ranking. *arXiv preprint arXiv:1707.09971*, 2017.

Uriel Feige and Eran Ofek. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27(2):251–275, 2005.

Se-Young Yun and Alexandre Proutiere. Optimal cluster recovery in the labeled stochastic block model. In *Advances in Neural Information Processing Systems*, pages 965–973, 2016.

Yiqiao Zhong and Nicolas Boumal. Near-optimal bounds for phase synchronization. *SIAM Journal on Numerical Analysis (to appear)*, 2017.