Non-backtracking Spectra of Random Hypergraphs and Community Detection

Yizhe Zhu

Department of Mathematics
University of California Irvine

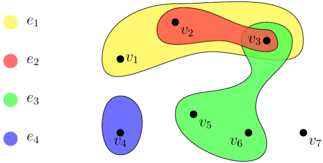
January 4, 2023
Tensor Representation, Completion, Modeling and Analytics of Complex Data

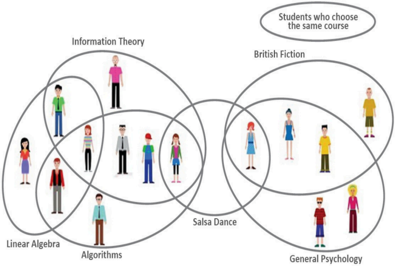
JMM 2023, Boston

Joint work with Ludovic Stephan (EPFL)

Hypergraph

- $G = (V, H)$, V: vertex set, H: hyperedge set.

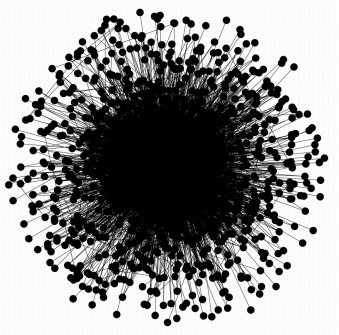


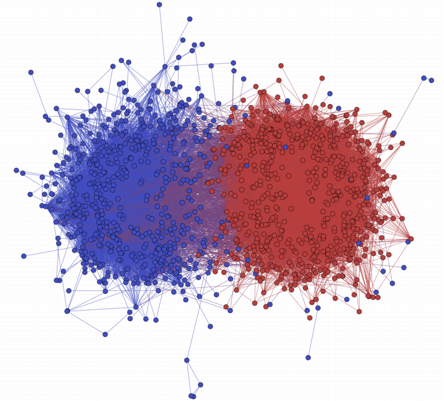


Ravindran '15

- Higher-order networks: co-authorship, chat group, protein interaction

Community detection





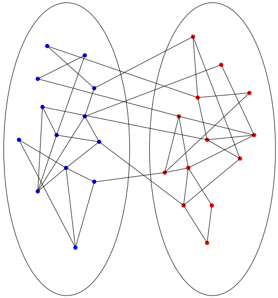
Political blogs data from Adamic-Glance '05. Figure from Abbe '18

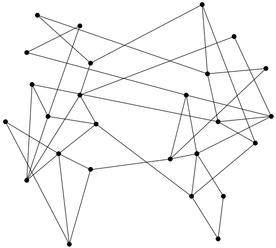
Community detection on random graphs

- Consider a (unknown) partition of n vertices into two *communities* of size $n/2$. Generate edges within each community with probability p. Generate edges across communities with probability $q < p$.
- **Stochastic block model** $\mathcal{G}(n, p, q)$. Holland et al. '83.

Community detection on random graphs

- Consider a (unknown) partition of n vertices into two *communities* of size $n/2$. Generate edges within each community with probability p. Generate edges across communities with probability $q < p$.
- **Stochastic block model** $\mathcal{G}(n,p,q)$. Holland et al. '83.
- Task: observe a graph $G \sim \mathcal{G}(n,p,q)$, find the unknown partition with high probability (efficiently and accurately).





Spectral method on the adjacency matrix

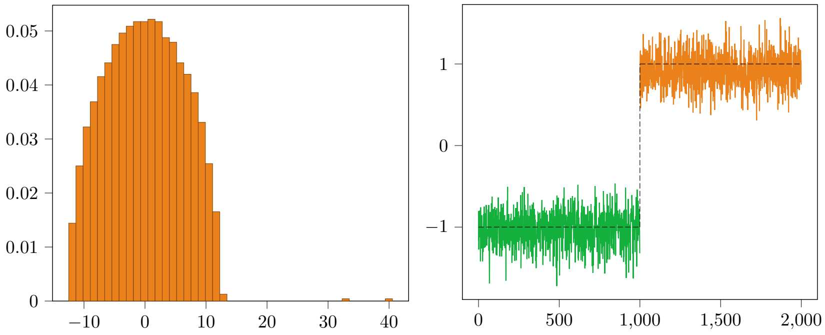
- Adjacency matrix A: symmetric, A_{ij} is independent Bernoulli for $i < j$.
- $\mathbb{E}A = \left[\begin{array}{cc|cc} p & p & q & q \\ p & p & q & q \\ \hline q & q & p & p \\ q & q & p & p \end{array}\right], \quad \lambda_1(\mathbb{E}A) = \frac{(p+q)n}{2}, \quad \lambda_2(\mathbb{E}A) = \frac{(p-q)n}{2}.$
- $v_1(\mathbb{E}A) = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, v_2(\mathbb{E}A) = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}.$
- $A = \mathbb{E}A + (A - \mathbb{E}A)$, low rank + noise.

Spectral method on the adjacency matrix

- Adjacency matrix A: symmetric, A_{ij} is independent Bernoulli for $i < j$.
- $\mathbb{E}A = \left[\begin{array}{cc|cc} p & p & q & q \\ p & p & q & q \\ \hline q & q & p & p \\ q & q & p & p \end{array}\right], \quad \lambda_1(\mathbb{E}A) = \frac{(p+q)n}{2}, \quad \lambda_2(\mathbb{E}A) = \frac{(p-q)n}{2}.$
- $v_1(\mathbb{E}A) = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, v_2(\mathbb{E}A) = \begin{bmatrix} 1 \\ 1 \\ -1 \\ -1 \end{bmatrix}.$
- $A = \mathbb{E}A + (A - \mathbb{E}A)$, low rank + noise.
- If A is concentrated around $\mathbb{E}A$, then $v_2(A) \approx v_2(\mathbb{E}(A))$.
- Spectral method: observe A, compute $v_2(A)$, use the signs of the entries in $v_2(A)$ to recover the community.

Feige–Ofek '05, Lei–Rinaldo '13, Le–Levina–Vershynin '16, Benaych Georges–Bordenave–Knowles '17, Latala–van Handel–Youssef '17, Alt–Ducatez–Knowles '19, Tikhomirov–Youssef '19

Spectral method on A: dense regime



Exact recovery when $p = \frac{a \log n}{n}, q = \frac{b \log n}{n}$ and $(\sqrt{a} - \sqrt{b})^2 \geq 2$.

[Abbe-Bandeira-Hall '15, Mossel-Neeman-Sly '16, Abbe-Fan-Wang-Zhong '20].

Sparse SBMs

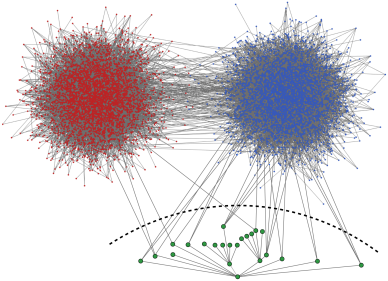
- Two communities of equal size. $\sigma : [n] \to \{-1, 1\}$.
- Bounded expected degrees: $p = \frac{a}{n}, q = \frac{b}{n}$. Impossible to recover σ exactly.
- Detection of σ is possible (strictly better than random guessing) if and only if $(a-b)^2 > 2(a+b)$ (Kesten-Stigum threshold).

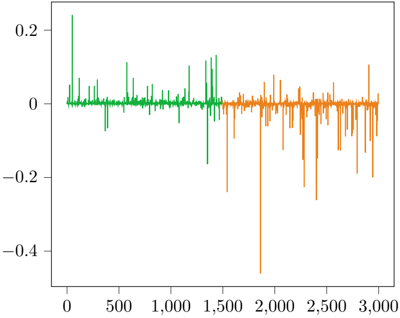
Decelle-Krzakala-Moore-Zdeborová '11, Mossel-Neeman-Sly '15, '18, Massoulié '14, Bordenave-Lelarge-Massoulié '18.

Sparse SBMs

- Two communities of equal size. $\sigma : [n] \to \{-1, 1\}$.
- Bounded expected degrees: $p = \frac{a}{n}, q = \frac{b}{n}$. Impossible to recover σ exactly.
- Detection of σ is possible (strictly better than random guessing) if and only if $(a-b)^2 > 2(a+b)$ (Kesten-Stigum threshold).

Decelle-Krzakala-Moore-Zdeborová '11, Mossel-Neeman-Sly '15, '18, Massoulié '14, Bordenave-Lelarge-Massoulié '18.





Top eigenvectors of A are localized on high degree vertices.

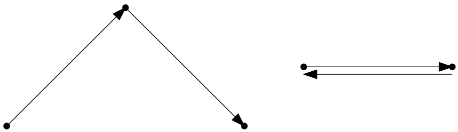
Non-backtracking operator

The set of oriented edges:

$$\vec{E} = \{u \to v : \{u, v\} \in E\}.$$

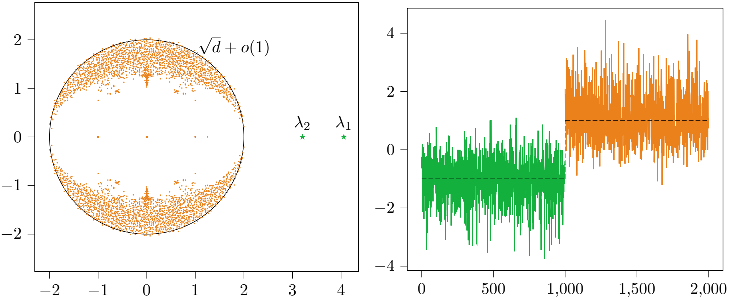
$|\vec{E}| = 2|E|$. The non-backtracking operator B is defined on $\vec{E}$.
For $u \to v, x \to y \in \vec{E}$,

$$B_{u\to v, x\to y} = \mathbf{1}_{v=x}\mathbf{1}_{u\neq y}.$$



Bordenave-Collins '19, Bordenave '20, Brito-Dumitriu-Harris '22, Benaych-Georges, Bordenave, Knowles '21, Stephan-Massoulié '20, Bordenave-Coste-Nadakuditi '20,. . .

Spectral method on B



[Bordenave, Lelarge, Massoulié '18] Let $p = \frac{a}{n}, q = \frac{b}{n}$. If $(a-b)^2 > 2(a+b)$, then with high probability,

$$\lambda_1(B) = \frac{a+b}{2} + o(1), \quad \lambda_2(B) = \frac{a-b}{2} + o(1), \quad |\lambda_3(B)| \leq \sqrt{\frac{a+b}{2}} + o(1).$$

The second eigenvector of B can be used to detect σ.

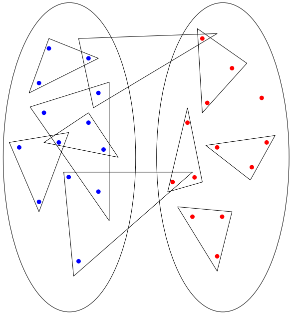
A fails but B works (optimally)!

Hypergraph stochastic block model (HSBM)

G is *q-uniform* if each hyperedge has size q.

- Community assignment $\sigma : [n] \to \{-1, +1\}$.
- Each hyperedge $e = \{v_1, \ldots, v_q\}$ appears independently with probability

$$\mathbb{P}(e \in H) = \begin{cases} c_{\mathrm{in}} & \text{if } \sigma_{v_1} = \cdots = \sigma_{v_q} \\ c_{\mathrm{out}} & \text{otherwise.} \end{cases}$$



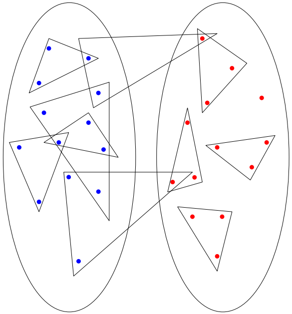
Task: observe G, construct a label estimator $\hat{\sigma} \in \{-1, +1\}^n$ correlated with the true σ. Ghoshdastidar-Dukkipati '14, Chien-Lin-Wang '18, Kim-Bandeira-Goemans '18, Ahn-Lee-Suh '18, ... when expected degree (expected number of hyperedges containing a vertex) $d \to \infty$.

Hypergraph stochastic block model (HSBM)

G is *q-uniform* if each hyperedge has size q.

- Community assignment $\sigma : [n] \to \{-1, +1\}$.
- Each hyperedge $e = \{v_1, \dots, v_q\}$ appears independently with probability

$$\mathbb{P}(e \in H) = \begin{cases} c_{\rm in} & \text{if } \sigma_{v_1} = \cdots = \sigma_{v_q} \\ c_{\rm out} & \text{otherwise.} \end{cases}$$

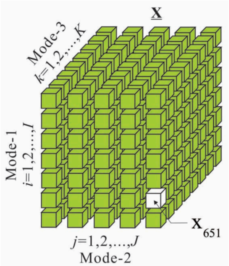


Task: observe G, construct a label estimator $\hat{\sigma} \in \{-1, +1\}^n$ correlated with the true σ. Ghoshdastidar-Dukkipati '14, Chien-Lin-Wang '18, Kim-Bandeira-Goemans '18, Ahn-Lee-Suh '18, ... when expected degree (expected number of hyperedges containing a vertex) $d \to \infty$.

- Detection: [Angelini-Caltagirone-Krzakala-Zdeborová '15] conjectured a phase transition when $c_{\rm in} = \frac{a}{\binom{n}{q-1}}, c_{\rm out} = \frac{b}{\binom{n}{q-1}}$.
- Provable spectral method in the bounded expected degree regime?

Tensor

The **adjacency tensor** T: sparse random tensor of order q with n^q many entries. $T_{i_1,\ldots,i_q} = 1$ if $\{i_1, \ldots, i_q\}$ is a hyperedge.



Most tensor problems are NP-hard (Hillar-Lim '13): rank, spectral norm, best low-rank approximation,. . .

Tucker decomposition: Ghoshdastidar-Dukkipat '17, Ke-Shi-Xia '20 for $d = \omega(\log^2 n)$.

Adjacency matrix

Define the **adjacency matrix** of G as

$$A_{ij} := \{\text{number of hyperedges containing } i, j\}.$$

Spectral method on A fails when $d = O(1)$.

Does the non-backtracking method work for random hypergraphs?
[Stephan-Z. '22]: Yes, and efficient for a more general HSBM model.

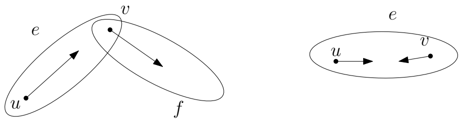
Non-backtracking operator for hypergraphs

For a given hypergraph $G = (V, H)$, let $\vec{H}$ be the *oriented hyperedge* in G such that

$$\vec{H} = \{(v, e) : v \in e \cap V, e \in H\}, \quad |\vec{H}| = q|H|.$$

B: a matrix indexed by $\vec{H}$ such that

$$B_{(u \to e),(v \to f)} = \begin{cases} 1 & \text{if } v \in e \setminus \{u\}, f \neq e, \\ 0 & \text{otherwise.} \end{cases}$$



Storm '06, Angelini-Caltagirone-Krzakala-Zdeborová '15, Dumitriu-Z. '21.

Generate an HSBM from a probability tensor

- An order-q *symmetric probability tensor* $\mathbf{P} \in \mathbb{R}^{r^q}$ and $\sigma : [n] \to [r]$.
- Each hyperedge of size q is included in H with probability

$$\mathbb{P}(e \in H) = \frac{p_{\underline{\sigma}(e)}}{\binom{n}{q-1}},$$

  where $\underline{\sigma}(e) = \underline{\sigma}(\{v_1, \dots, v_q\}) := (\sigma(v_1), \dots, \sigma(v_q))$.
- The proportion of each community is $\pi_i, i \in [r]$. Assume each vertex has the same expected degree d.

Generate an HSBM from a probability tensor

- An order-q *symmetric probability tensor* $\mathbf{P} \in \mathbb{R}^{r^q}$ and $\sigma : [n] \to [r]$.
- Each hyperedge of size q is included in H with probability

$$\mathbb{P}(e \in H) = \frac{p_{\underline{\sigma}(e)}}{\binom{n}{q-1}},$$

where $\underline{\sigma}(e) = \underline{\sigma}(\{v_1, \dots, v_q\}) := (\sigma(v_1), \dots, \sigma(v_q))$.

- The proportion of each community is $\pi_i, i \in [r]$. Assume each vertex has the same expected degree d.

The nonzero eigenvalues of $\mathbb{E}A$ are given by

$$|\mu_r| \le \cdots \le |\mu_2| \le \mu_1 = d.$$

Denote by r_0 the number of informative eigenvalues, or equivalently

$$(q-1)\mu_{r_0+1}^2 \le d < (q-1)\mu_{r_0}^2.$$

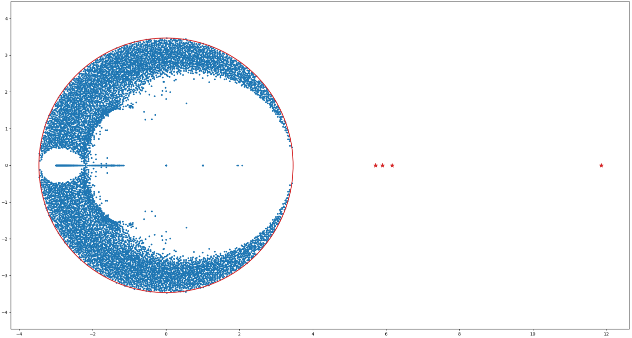
Same as the generalized KS threshold conjectured in [Angelini et al. '15].

Spectrum of B

Theorem (Stephan-Z., FOCS 22)

Let G be a hypergraph generated according to the HSBM with m hyperedges, and B be its non-backtracking matrix and $|\lambda_1(B)| \geq |\lambda_2(B)| \geq \cdots \geq |\lambda_{qm}(B)|$. Then with high probability:

1. *For any $i \in [r_0]$, $\lambda_i(B) = (q-1)\mu_i + o(1)$.*
2. *For all $r_0 < i \leq qm$, $|\lambda_i(B)| \leq \sqrt{(q-1)d} + o(1)$.*



$n = 6000$, $r = 4$. The parameters c_{in} and c_{out} have been chosen so that $d = 4$ and $\mu_2 = 2$.

Dimension reduction

B has size $q|H| \sim qdn$, but qd could be a large constant. We also need to embed eigenvectors into $\mathbb{R}^n$. Define the $2n \times 2n$ matrix $\tilde{B}$ as

$$\tilde{B} = \begin{pmatrix} 0 & (D-I) \\ -(q-1)I & A-(q-2)I \end{pmatrix},$$

where D is the diagonal *degree matrix* with $D_{ii} = (q-1)^{-1} \sum_j A_{ij}$.

Lemma (Stephan-Z., FOCS 22)

The following Ihara-Bass formula holds:

$$\det(B - zI) = (z-1)^{(q-1)|H|-n} (z+(q-1))^{|H|-n} \det\left(\tilde{B} - zI\right).$$

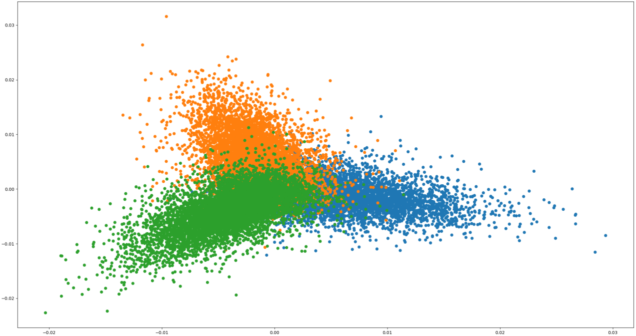
B and $\tilde{B}$ share the same non-trivial eigenvalues.

Eigenvector overlaps

Theorem (Stephan-Z., FOCS 22)

For $i \in [r_0]$, let $\tilde{u}_i$ be the last n entries of the i-th eigenvector of $\tilde{B}$ with $\|\tilde{u}_i\| = 1$. Then with high probability, there exists a unit eigenvector $\tilde{\phi}_i$ of $\mathbb{E}A$ associated to μ_i such that

$$\langle \tilde{u}_i, \tilde{\phi}_i \rangle = \sqrt{\frac{1-\tau_i}{1+\frac{q-2}{(q-1)\mu_i}}} + o(1) \quad \textit{where } \tau_i = \frac{d}{(q-1)\mu_i^2}.$$



Scatter plot of the second and third eigenvector of $\tilde{B}$ under the symmetric HSBM with $q = 4$, $r = 3$.

Conclusions

- Community detection for sparse random hypergraphs can be reduced to an eigenvector problem of a $2n \times 2n$ non-normal matrix constructed from A (**without** the knowledge of T), and it works down to the conjectured generalized KS threshold.
- Ihara-Bass formula and non-backtracking method for non-uniform hypergraphs [Chodrow-Eikmeier-Haddock '22].