Community Detection in Sparse Random Hypergraphs

Yizhe Zhu

Department of Mathematics
University of California Irvine

March 16, 2022

Random Tensors and Related Topics
CIRM

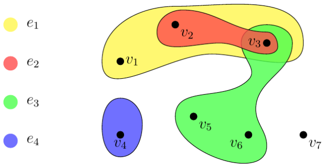
Joint work with Soumik Pal (Univeristy of Washington)
and Ludovic Stephan (EPFL)

Hypergraph

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

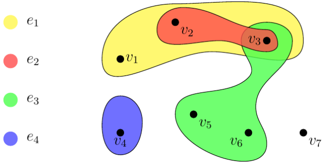
Hypergraph

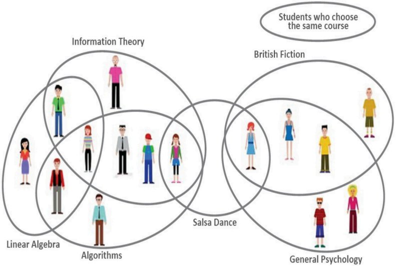
- $G = (V, H)$, V: vertex set, H: hyperedge set.



Hypergraph

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

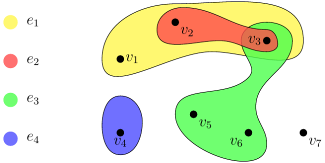


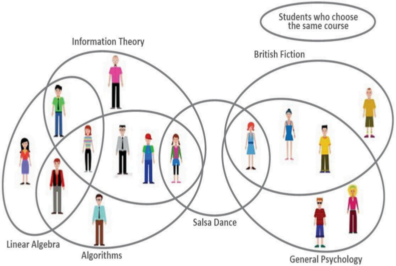


Ravindran '15

Hypergraph

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





Ravindran '15

- co-authorship network
- chat group in social network
- Protein interaction network

Higher-order network

siam news

HOME | HAPPENING NOW | GET INVOLVED | RESEARCH

SIAM NEWS **BLOG**

Research | January 21, 2021 Print

Higher-order Network Analysis Takes Off, Fueled by Old Ideas and New Data

By Austin R. Benson, David F. Gleich, and Desmond J. Higham

Quanta magazine Physics Mathematics Biology Computer Science Topics Archive

GRAPH THEORY

How Big Data Carried Graph Theory Into New Dimensions

4

Researchers are turning to the mathematics of higher-order interactions to better model the complex connections within their data.

Higher-order network

siam news

HOME | HAPPENING NOW | GET INVOLVED | RESEARCH

SIAM NEWS BLOG

Research | January 21, 2021 Print

Higher-order Network Analysis Takes Off, Fueled by Old Ideas and New Data

By Austin R. Benson, David F. Gleich, and Desmond J. Higham

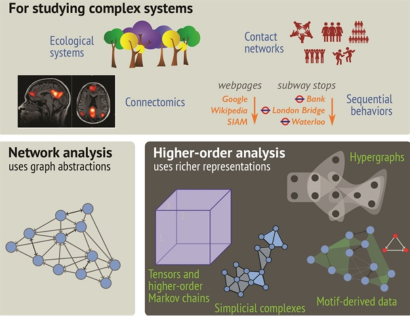
Quanta magazine | Physics | Mathematics | Biology | Computer Science | Topics | Archive

GRAPH THEORY

How Big Data Carried Graph Theory Into New Dimensions

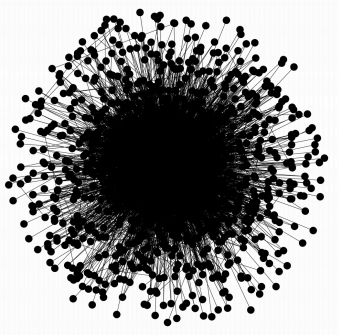
4

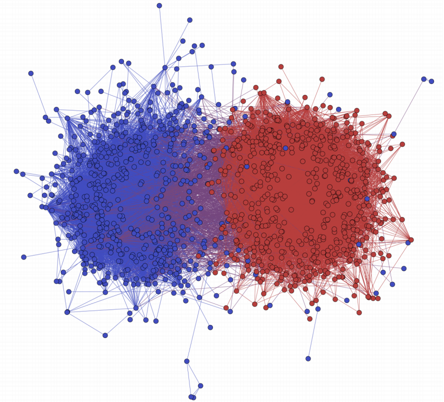
Researchers are turning to the mathematics of higher-order interactions to better model the complex connections within their data.



sinews.siam.org/Details-Page/higher-order-network-analysis-takes-off-fueled-by-old-ideas-and-new-data
www.quantamagazine.org/how-big-data-carried-graph-theory-into-new-dimensions-20210819/

Community detection





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

Community detection on random graphs

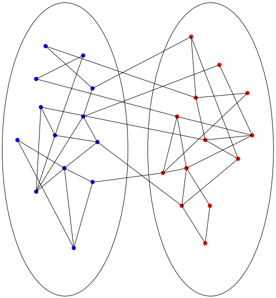
- Consider a (unknown) partition of n vertices into two *communities* of size $n/2$. Generate edges within each community with probability p. Generate edges across communities with probability $q < p$.

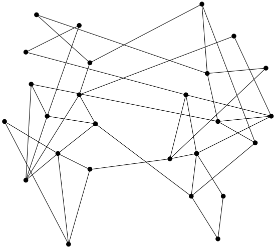
Community detection on random graphs

- 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).

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

Spectral method on the adjacency matrix

- Adjacency matrix A: symmetric, A_{ij} is independent Bernoulli for $i < j$.

Spectral method on the adjacency matrix

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

Spectral method on the adjacency matrix

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

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 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.

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.
- $\|A - \mathbb{E}A\| = O(\sqrt{n(p+q)})$ when $\frac{(p+q)n}{2} = \Omega(\log n)$. $o(n)$ vertices are mis-classified.

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.
- $\|A - \mathbb{E}A\| = O(\sqrt{n(p+q)})$ when $\frac{(p+q)n}{2} = \Omega(\log n)$. $o(n)$ vertices are mis-classified.

  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

Sparse SBMs

- Two communities of equal size. $\sigma : [n] \to \{-1, 1\}$.

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.

Sparse SBMs

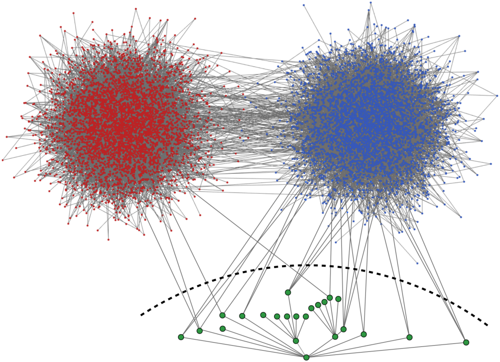
- Two communities of equal size. $\sigma : [n] \to \{-1, 1\}$.
- Bounded expected degrees: $p = \frac{a}{n}, q = \frac{b}{n}$. Impossible to recover σ exactly.
- Detection is possible (strictly better than random guessing) if and only if $(a-b)^2 > 2(a+b)$ (Kesten-Stigum threshold).

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 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 '12, '14, Massoulié '14, Bordenave-Lelarge-Massoulié '15.
Rich literature on SBMs in more general cases and different settings: survey by Abbe '18.

Bounded expected degrees



Abbe et al. '18, $a = 2.2$, $b = 0.06$, $n = 100000$, apply spectral method directly on A

When $p = \frac{a}{n}, q = \frac{b}{n}$, top eigenvectors are localized on high degree vertices.

Non-backtracking operator

Non-backtracking operator

The set of oriented edges:

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

$|\vec{E}| = 2|E|$.

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 $\vec{H}$.
For $u \to v, x \to y \in \vec{E}$,

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

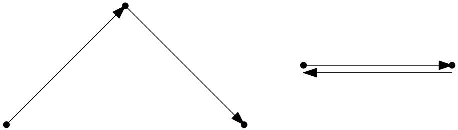
Non-backtracking operator

The set of oriented edges:

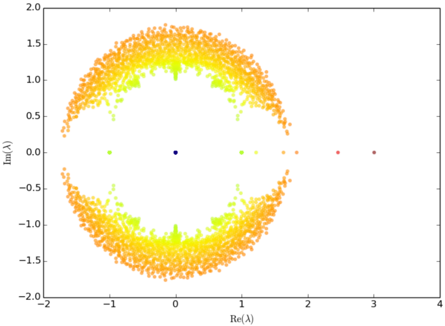
$$\vec{E} = \{u \to v : \{u, v\} \in E\}.$$

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

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



Spectrum of B

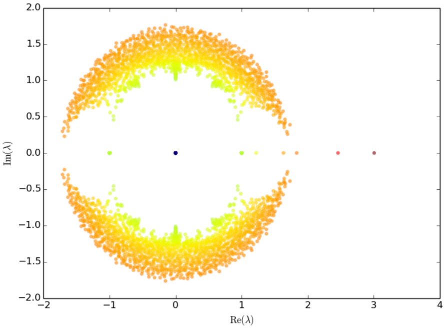


[Bordenave, Lelarge, Massoulié '15] Let $p = \frac{a}{n}, q = \frac{b}{n}$. Then if $(a-b)^2 > 2(a+b)$, with high probability,

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

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

Spectrum of B



[Bordenave, Lelarge, Massoulié '15] Let $p = \frac{a}{n}, q = \frac{b}{n}$. Then if $(a-b)^2 > 2(a+b)$, 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.

Hypergraph stochastic block model (HSBM)

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

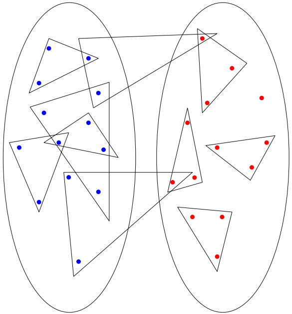
- Type assignment $\sigma : [n] \to \{-1, +1\}$.

Hypergraph stochastic block model (HSBM)

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

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

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

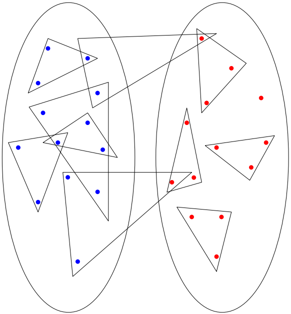


Hypergraph stochastic block model (HSBM)

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

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

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



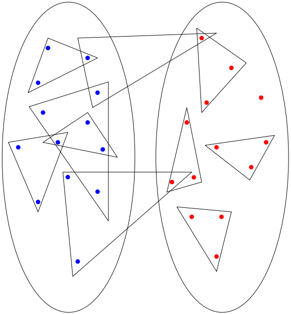
Task: observe G, construct a label estimator $\hat{\sigma} \in \{-1, +1\}^n$ correlated with the true σ.

Hypergraph stochastic block model (HSBM)

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

- Type 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.

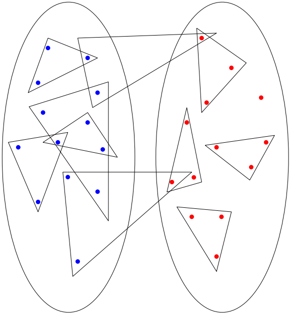
Ghoshdastidar-Dukkipati '14, '15, Chien-Lin-Wang '18, Kim-Bandeira-Goemans '18, Ahn-Lee-Suh '18, ...

Hypergraph stochastic block model (HSBM)

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

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

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



Task: observe G, construct a label estimator $\hat{\sigma} \in \{-1, +1\}^n$ correlated with the true σ.

Ghoshdastidar-Dukkipati '14, '15, Chien-Lin-Wang '18, Kim-Bandeira-Goemans '18, Ahn-Lee-Suh '18, ...
when expected degree (expected number of hyperedges containing a vertex) $d \to \infty$.

Sparse HSBM

- Detection: Angelini-Caltagirone-Krzakala-Zdeborová '15 conjectured a phase transition when $c_{\text{in}} = \frac{a}{\binom{n}{q-1}}, c_{\text{out}} = \frac{b}{\binom{n}{q-1}}$, based on the belief propagation algorithm.

Sparse HSBM

- Detection: Angelini-Caltagirone-Krzakala-Zdeborová '15 conjectured a phase transition when $c_{\text{in}} = \frac{a}{\binom{n}{q-1}}, c_{\text{out}} = \frac{b}{\binom{n}{q-1}}$, based on the belief propagation algorithm.
- $\alpha := (q-1)\frac{a+(2^{q-1}-1)b}{2^{q-1}}$, $\beta := (q-1)\frac{a-b}{2^{q-1}}$. $\beta^2 > \alpha$ is the detection threshold.

Sparse HSBM

- Detection: Angelini-Caltagirone-Krzakala-Zdeborová '15 conjectured a phase transition when $c_{\rm in} = \frac{a}{\binom{n}{q-1}}, c_{\rm out} = \frac{b}{\binom{n}{q-1}}$, based on the belief propagation algorithm.
- $\alpha := (q-1)\frac{a+(2^{q-1}-1)b}{2^{q-1}}$, $\beta := (q-1)\frac{a-b}{2^{q-1}}$. $\beta^2 > \alpha$ is the detection threshold.
- (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.
$T_{i_1,\dots,i_q} = 1$ if $\{i_1, \dots, i_q\}$ is a hyperedge.

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.

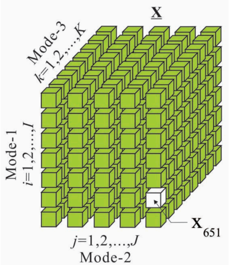


Figure: an order-3 tensor

Tensor

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

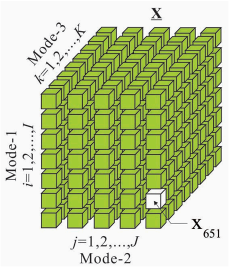


Figure: an order-3 tensor

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 $d = \omega(\log n)$.

Adjacency matrix

- Define the **adjacency matrix** of G as

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

Adjacency matrix

- Define the **adjacency matrix** of G as

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

- $\mathrm{tr} A^k$ counts the number of closed walks of length k in G: $(i_0, e_1, i_1, \ldots, i_{k-1}, e_k, i_0)$.

Adjacency matrix

- Define the **adjacency matrix** of G as

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

- $\mathrm{tr} A^k$ counts the number of closed walks of length k in G: $(i_0, e_1, i_1, \ldots, i_{k-1}, e_k, i_0)$.

The spectral method on A fails when average expected degree is $O(1)$.

Adjacency matrix

- Define the **adjacency matrix** of G as

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

- $\mathrm{tr} A^k$ counts the number of closed walks of length k in G: $(i_0, e_1, i_1, \dots, i_{k-1}, e_k, i_0)$.

The spectral method on A fails when average expected degree is $O(1)$.

[Pal-Z. '21]: spectral method on a matrix counting the self-avoiding walk of length $O(\log n)$ for HSBM achieves the conjectured threshold in Angelini et al '15, generalization of Massoulié '14.

Adjacency matrix

- Define the **adjacency matrix** of G as

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

- $\mathrm{tr} A^k$ counts the number of closed walks of length k in G: $(i_0, e_1, i_1, \dots, i_{k-1}, e_k, i_0)$.

The spectral method on A fails when average expected degree is $O(1)$.

[Pal-Z. '21]: spectral method on a matrix counting the self-avoiding walk of length $O(\log n)$ for HSBM achieves the conjectured threshold in Angelini et al '15, generalization of Massoulié '14.

What about the non-backtracking operator?

Adjacency matrix

- Define the **adjacency matrix** of G as

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

- $\mathrm{tr} A^k$ counts the number of closed walks of length k in G: $(i_0, e_1, i_1, \ldots, i_{k-1}, e_k, i_0)$.

The spectral method on A fails when average expected degree is $O(1)$.

[Pal-Z. '21]: spectral method on a matrix counting the self-avoiding walk of length $O(\log n)$ for HSBM achieves the conjectured threshold in Angelini et al '15, generalization of Massoulié '14.

What about the non-backtracking operator?
[Stephan, Z. '22]: Very efficient!

Non-backtracking operator for hypergraphs

Non-backtracking operator for hypergraphs

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

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

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

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

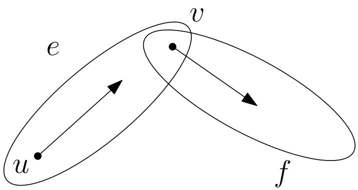
Non-backtracking operator for hypergraphs

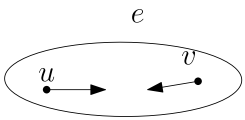
For a given hypergraph $G = (V, H)$, let $\vec{H}$ be the *oriented hyperedge* in G such that

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

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: Zeta function of hypergraphs.

Generate an HSBM from a probability tensor

- Consider an order-q *symmetric probability tensor* $\mathbf{P} \in \mathbb{R}^{r^q}$ and $\sigma : [n] \to [r]$.

Generate an HSBM from a probability tensor

- Consider 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

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

for any $e = \{v_1, \dots, v_q\}$, where

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

Generate an HSBM from a probability tensor

- Consider 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

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

for any $e = \{v_1, \dots, v_q\}$, where

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

- The proportion of each type is

$$\pi_i = \frac{\#\{v \in V \mid \sigma(v) = i\}}{n}.$$

Generate an HSBM from a probability tensor

- Consider 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

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

for any $e = \{v_1, \dots, v_q\}$, where

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

- The proportion of each type is

$$\pi_i = \frac{\#\{v \in V \mid \sigma(v) = i\}}{n}.$$

- Assume each vertex has the same expected degree d.

Generalized Kesten-Stigum threshold

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

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

Generalized Kesten-Stigum threshold

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

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

Generalized Kesten-Stigum threshold

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

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

Denote by r_0 the number of informative eigenvalues, or equivalently

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

The generalized Kesten-Stigum threshold conjectured in Angelini et al. '15.

Spectrum of B

Theorem (Stephan-Z., '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 $r_0 < i \leq qm$,*
$$|\lambda_i(B)| \leq \sqrt{(q-1)d} + o(1).$$

Spectrum of B

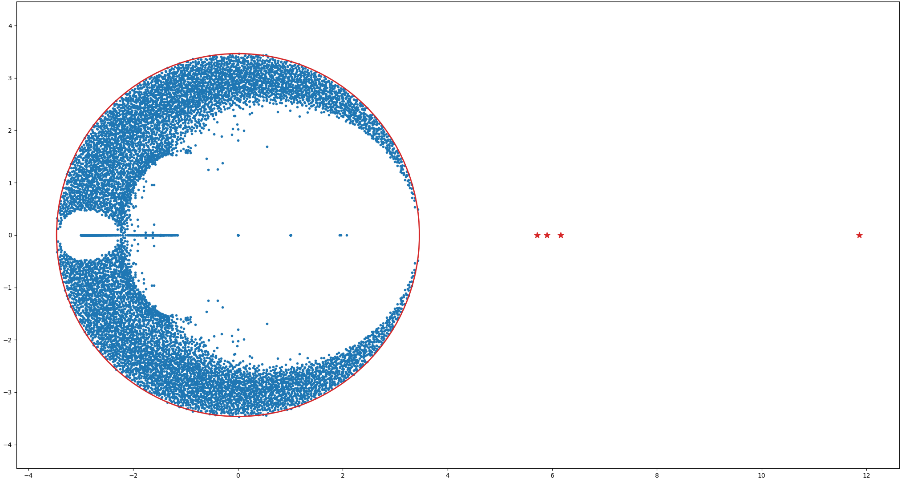
Theorem (Stephan-Z., '22)

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 $r_0 < i \leq qm$,*
$$|\lambda_i(B)| \leq \sqrt{(q-1)d} + o(1).$$

- Informative eigenvalues of $\mathbb{E}A$ above the Kesten-Stigum threshold can be seen in the spectrum of B outside the disk of radius $\sqrt{(q-1)d}$.
- Other eigenvalues of B are confined in the disk.

Spectrum of B



$n = 6000$, $q = r = 4$. The parameters c_{in} and c_{out} have been chosen so that $d = 4$ and $\mu_2 = 2$. The single eigenvalue is close to $(q-1)d = 12$ and the three eigenvalues are near $(q-1)\mu_2 = 6$.

Dimension reduction

B has size $q|H| \sim qdn$, could be very large!

Dimension reduction

B has size $q|H| \sim qdn$, could be very large! We also need a procedure to map eigenvectors of B into $\mathbb{R}^n$.

Dimension reduction

B has size $q|H| \sim qdn$, could be very large! We also need a procedure to map eigenvectors of B into $\mathbb{R}^n$.
Define the $2n \times 2n$ matrix $\tilde{B}$ as

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

and D is the diagonal *degree matrix* with $D_{ii} = \#\{e \in H : i \in e\}$.

Dimension reduction

B has size $q|H| \sim qdn$, could be very large! We also need a procedure to map eigenvectors of B into $\mathbb{R}^n$.
Define the $2n \times 2n$ matrix $\tilde{B}$ as

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

and D is the diagonal *degree matrix* with $D_{ii} = \#\{e \in H : i \in e\}$.

Lemma (Stephan-Z., '22)

The following Ihara-Bass formula holds:

$$\begin{aligned}\det(B - zI) =& (z-1)^{(q-1)|H|-n}(z+(q-1))^{|H|-n} \\ & \cdot \det\left(z^2 + (q-2)z - zA + (q-1)(D-I)\right).\end{aligned}$$

Dimension reduction

B has size $q|H| \sim qdn$, could be very large! We also need a procedure to map eigenvectors of B into $\mathbb{R}^n$.
Define the $2n \times 2n$ matrix $\tilde{B}$ as

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

and D is the diagonal *degree matrix* with $D_{ii} = \#\{e \in H : i \in e\}$.

Lemma (Stephan-Z., '22)

The following Ihara-Bass formula holds:

$$\begin{aligned}\det(B - zI) =& (z-1)^{(q-1)|H|-n}(z+(q-1))^{|H|-n} \\ & \cdot \det\left(z^2 + (q-2)z - zA + (q-1)(D-I)\right).\end{aligned}$$

The spectrum of $\tilde{B}$ is identical to the spectrum of B, except for possible trivial eigenvalues at -1 and $-(q-1)$.

Dimension reduction

B has size $q|H| \sim qdn$, could be very large! We also need a procedure to map eigenvectors of B into $\mathbb{R}^n$.
Define the $2n \times 2n$ matrix $\tilde{B}$ as

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

and D is the diagonal *degree matrix* with $D_{ii} = \#\{e \in H : i \in e\}$.

Lemma (Stephan-Z., '22)

The following Ihara-Bass formula holds:

$$\begin{aligned}\det(B - zI) =& (z-1)^{(q-1)|H|-n}(z+(q-1))^{|H|-n} \\ &\cdot \det\left(z^2 + (q-2)z - zA + (q-1)(D-I)\right).\end{aligned}$$

The spectrum of $\tilde{B}$ is identical to the spectrum of B, except for possible trivial eigenvalues at -1 and $-(q-1)$.

$q = 2$: Bass '92. Storm '06 for regular hypergraphs, stated in Angelini et al. '15.

Eigenvector overlaps

Theorem (Stephan-Z., '22)

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

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

Eigenvector overlaps

Theorem (Stephan-Z., '22)

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

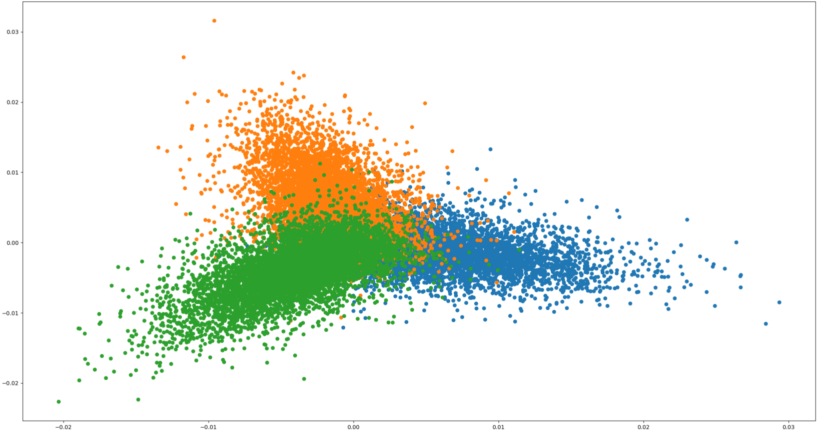
$$\langle \tilde{u}_i, \tilde{\phi}_i \rangle = \sqrt{\frac{1-\tau_i}{1+\frac{q-2}{(q-1)\mu_i}}} + o(1) \quad \textit{where } \tau_i = \frac{d}{(q-1)\mu_i^2}.$$

When $r = 2$, and

$$p_{i_1,\dots,i_q} = \begin{cases} c_{\mathrm{in}} & \text{if } \sigma(i_1) = \cdots = \sigma(i_q) \\ c_{\mathrm{out}} & \text{otherwise} \end{cases},$$

rounding the entries $\tilde{u}_2$ to ± 1 gives a correlated detection.

More than 2 blocks

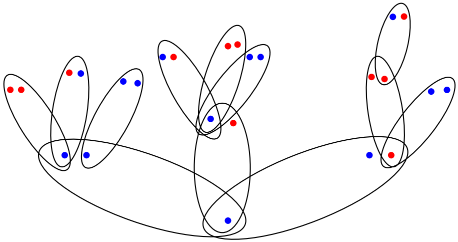


Scatter plot of the second and third eigenvector of $\tilde{B}$ under the symmetric HSBM with $q = 4$, $r = 3$ and $n = 20000$. The parameters c_{in} and c_{out} have been chosen so that $d = 4$ and $\mu_2 = 2$. The colors correspond to the actual label of each vertex.

vertices	1	2	$\cdots$	n
$\tilde{u}_2$	x_1	x_2	$\cdots$	x_n
$\tilde{u}_3$	y_1	y_2	$\cdots$	y_n

Local structure: Galton-Watson hypertree

Local structure: Galton-Watson hypertree

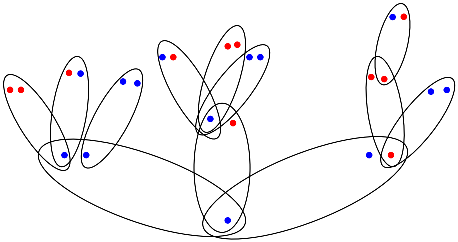


- Start from a *root* ρ with a given spin $\sigma(\rho)$;
- Generate $k = \text{Poi}(d)$ hyperedges intersecting only at ρ, yielding $k(q-1)$ *children*;
- For each hyperedge, fix an ordering of the $(q-1)$ associated children $v = (v_1, \ldots, v_{q-1})$. Assign a type to each $(q-1)$-tuple randomly such that

$$\mathbb{P}\left(\underline{\sigma}(v) = \underline{j}\right) = \frac{1}{d} \cdot p_{\sigma(\rho),\underline{j}} \cdot \prod_{\ell \in \underline{j}} \pi_\ell .$$

- Repeat the process for each child of ρ, treating as the root of an i.i.d Galton-Watson hypertree.

Local structure: Galton-Watson hypertree



- Start from a *root* ρ with a given spin $\sigma(\rho)$;
- Generate $k = \mathrm{Poi}(d)$ hyperedges intersecting only at ρ, yielding $k(q-1)$ *children*;
- For each hyperedge, fix an ordering of the $(q-1)$ associated children $v = (v_1, \dots, v_{q-1})$. Assign a type to each $(q-1)$-tuple randomly such that

$$\mathbb{P}\left(\underline{\sigma}(v) = \underline{j}\right) = \frac{1}{d} \cdot p_{\sigma(\rho),\underline{j}} \cdot \prod_{\ell \in \underline{j}} \pi_\ell.$$

- Repeat the process for each child of ρ, treating as the root of an i.i.d Galton-Watson hypertree.

[Pal-Z. '21]: considered 2-type Galton-Watson hypertrees.

Moment method and a bipartite representation

$\mathrm{tr} B^{\ell} = \#\{\text{closed non-backtracking walks of length } \ell\}$ with $\ell = \kappa \log n$.

Moment method and a bipartite representation

$\mathrm{tr} B^{\ell} = \#\{\text{closed non-backtracking walks of length } \ell\}$ with $\ell = \kappa \log n$.
Bounding bulk eigenvalues: high trace method on matrices modified from B.

Moment method and a bipartite representation

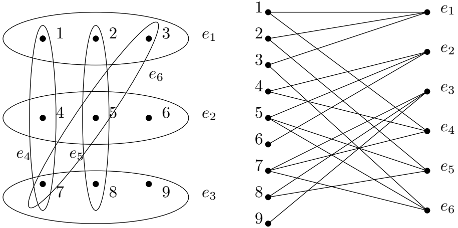
$\mathrm{tr} B^\ell = \#\{$closed non-backtracking walks of length $\ell\}$ with $\ell = \kappa \log n$.
Bounding bulk eigenvalues: high trace method on matrices modified from B.

- A non-backtracking walk on $\vec{H}$ of length ℓ: $(v_0 \to e_0, v_1 \to e_1, \dots, v_\ell \to e_\ell)$.

Moment method and a bipartite representation

$\mathrm{tr} B^{\ell} = \#\{\text{closed non-backtracking walks of length } \ell\}$ with $\ell = \kappa \log n$.
Bounding bulk eigenvalues: high trace method on matrices modified from B.

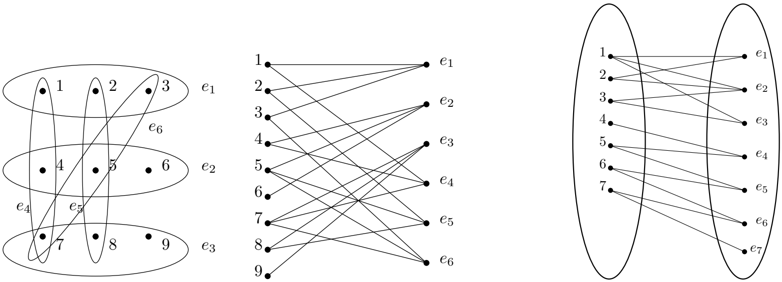
- A non-backtracking walk on $\vec{H}$ of length ℓ: $(v_0 \to e_0, v_1 \to e_1, \ldots, v_\ell \to e_\ell)$.
- A corresponding non-backtracking walk on the vertex space of a corresponding bipartite of length $2\ell + 1$: $(v_0, e_0, v_1, e_1, \ldots, v_\ell, e_\ell)$.



Moment method and a bipartite representation

$\mathrm{tr} B^\ell = \#\{$closed non-backtracking walks of length $\ell\}$ with $\ell = \kappa \log n$.
Bounding bulk eigenvalues: high trace method on matrices modified from B.

- A non-backtracking walk on $\vec{H}$ of length ℓ: $(v_0 \to e_0, v_1 \to e_1, \ldots, v_\ell \to e_\ell)$.
- A corresponding non-backtracking walk on the vertex space of a corresponding bipartite of length $2\ell + 1$: $(v_0, e_0, v_1, e_1, \ldots, v_\ell, e_\ell)$.



A closed non-backtracking walk: $(1, e_1, 2, e_2, 1, e_3, 3, e_2, 1)$.

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 and D, and it works down to the conjectured generalized KS threshold.

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 and D, and it works down to the conjectured generalized KS threshold.
- High trace methods can be applied to random hypergraphs with a proper definition of walks on hypergraphs.

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 and D, and it works down to the conjectured generalized KS threshold.
- High trace methods can be applied to random hypergraphs with a proper definition of walks on hypergraphs.
- Applicable to non-uniform hypergraphs. Possible application in sparse tensor completion.

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 and D, and it works down to the conjectured generalized KS threshold.
- High trace methods can be applied to random hypergraphs with a proper definition of walks on hypergraphs.
- Applicable to non-uniform hypergraphs. Possible application in sparse tensor completion.
- Open problem: impossibility for any algorithm below the generalized Kesten-Stigum threshold.

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 and D, and it works down to the conjectured generalized KS threshold.
- High trace methods can be applied to random hypergraphs with a proper definition of walks on hypergraphs.
- Applicable to non-uniform hypergraphs. Possible application in sparse tensor completion.
- Open problem: impossibility for any algorithm below the generalized Kesten-Stigum threshold.

Thank You!