Community Detection in Sparse Random Hypergraphs

Yizhe Zhu

Department of Mathematics
UCSD

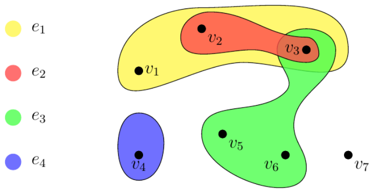
November 1, 2021

MSRI Seminar

Joint work with Soumik Pal (University of Washington)

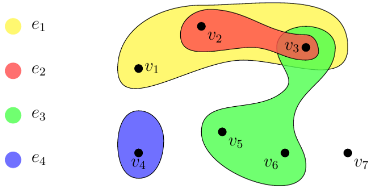
Hypergraph

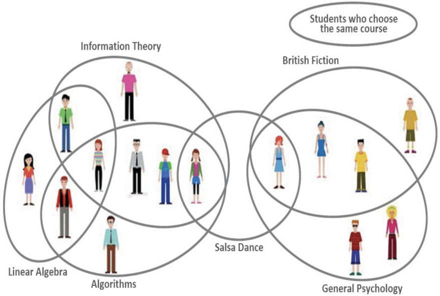
- $H = (V, E)$, V: vertex set, E: hyperedge set.



Hypergraph

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

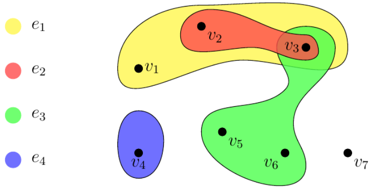


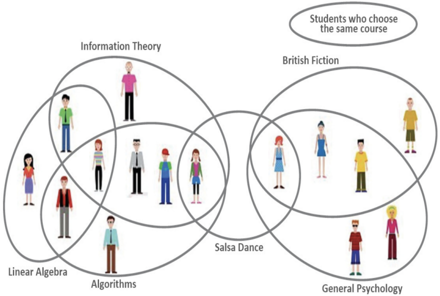


Ravindran (2015)

Hypergraph

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

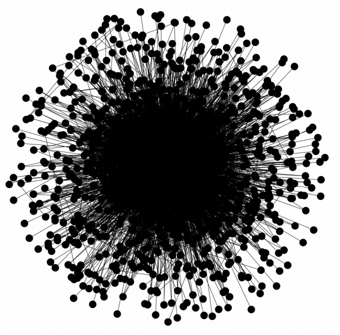


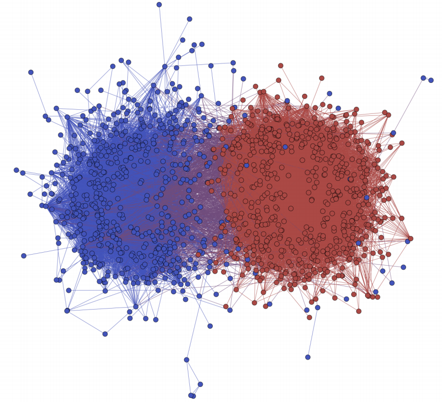


Ravindran (2015)

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

Community detection





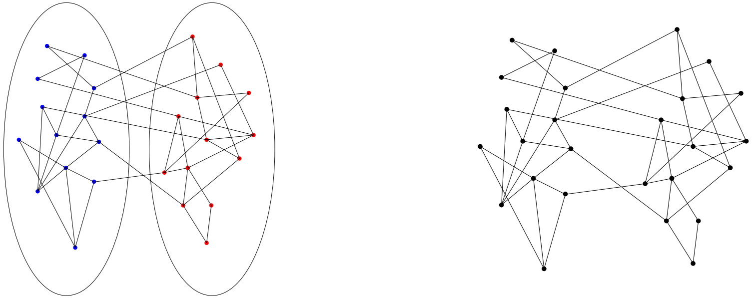
Political blogs data from Adamic-Glance (05). Figure from Abbe (18)

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 A

Spectral method on A

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

Spectral method on A

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

Spectral method on A

- 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 A

- 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 A

- 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 A

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

Spectral method on A

- 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. $v = (0.5, 1.1, -0.8, -0.4)$

Spectral method on A

- 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. $v = (0.5, 1.1, -0.8, -0.4)$
- $\|A - \mathbb{E}A\| = O(\sqrt{np})$ 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 phase transitions

Sparse SBMs: two phase transitions

Two communities of roughly equal size: assign labels $\sigma_i \in \{-1, +1\}$ uniformly and i.i.d. for $i \in [n]$.

Sparse SBMs: two phase transitions

Two communities of roughly equal size: assign labels $\sigma_i \in \{-1, +1\}$ uniformly and i.i.d. for $i \in [n]$.

- Logarithmic expected degrees: $p = \frac{a \log n}{n}, q = \frac{b \log n}{n}$.

Sparse SBMs: two phase transitions

Two communities of roughly equal size: assign labels $\sigma_i \in \{-1, +1\}$ uniformly and i.i.d. for $i \in [n]$.

- Logarithmic expected degrees: $p = \frac{a \log n}{n}, q = \frac{b \log n}{n}$.
- Exact recovery (recover vector σ up to a sign flip) is possible if and only if $(\sqrt{a} - \sqrt{b})^2 \geq 2$.

  Abbe-Bandeira-Hall (14), Mossel-Neeman-Sly (14).

Sparse SBMs: two phase transitions

Two communities of roughly equal size: assign labels $\sigma_i \in \{-1, +1\}$ uniformly and i.i.d. for $i \in [n]$.

- Logarithmic expected degrees: $p = \frac{a \log n}{n}, q = \frac{b \log n}{n}$.
- Exact recovery (recover vector σ up to a sign flip) is possible if and only if $(\sqrt{a} - \sqrt{b})^2 \geq 2$.

  Abbe-Bandeira-Hall (14), Mossel-Neeman-Sly (14).
- Bounded expected degrees: $p = \frac{a}{n}, q = \frac{b}{n}$.

Sparse SBMs: two phase transitions

Two communities of roughly equal size: assign labels $\sigma_i \in \{-1, +1\}$ uniformly and i.i.d. for $i \in [n]$.

- Logarithmic expected degrees: $p = \frac{a \log n}{n}, q = \frac{b \log n}{n}$.
- Exact recovery (recover vector σ up to a sign flip) is possible if and only if $(\sqrt{a} - \sqrt{b})^2 \geq 2$.

  Abbe-Bandeira-Hall (14), Mossel-Neeman-Sly (14).
- Bounded expected degrees: $p = \frac{a}{n}, q = \frac{b}{n}$.
- Detection is possible (strictly better than random guessing) if and only if $(a - b)^2 > 2(a + b)$.

Sparse SBMs: two phase transitions

Two communities of roughly equal size: assign labels $\sigma_i \in \{-1, +1\}$ uniformly and i.i.d. for $i \in [n]$.

- Logarithmic expected degrees: $p = \frac{a \log n}{n}, q = \frac{b \log n}{n}$.
- Exact recovery (recover vector σ up to a sign flip) is possible if and only if $(\sqrt{a} - \sqrt{b})^2 \geq 2$.

  Abbe-Bandeira-Hall (14), Mossel-Neeman-Sly (14).
- Bounded expected degrees: $p = \frac{a}{n}, q = \frac{b}{n}$.
- Detection is possible (strictly better than random guessing) if and only if $(a - b)^2 > 2(a + b)$.

  Decelle-Krzakala-Moore-Zdeborová (11), Mossel-Neeman-Sly (12, 14), Massoulié (14), Bordenave-Lelarge-Massoulié (15).

Sparse SBMs: two phase transitions

Two communities of roughly equal size: assign labels $\sigma_i \in \{-1, +1\}$ uniformly and i.i.d. for $i \in [n]$.

- Logarithmic expected degrees: $p = \frac{a \log n}{n}, q = \frac{b \log n}{n}$.
- Exact recovery (recover vector σ up to a sign flip) is possible if and only if $(\sqrt{a} - \sqrt{b})^2 \geq 2$.

  Abbe-Bandeira-Hall (14), Mossel-Neeman-Sly (14).
- Bounded expected degrees: $p = \frac{a}{n}, q = \frac{b}{n}$.
- Detection is possible (strictly better than random guessing) if and only if $(a - b)^2 > 2(a + b)$.

  Decelle-Krzakala-Moore-Zdeborová (11), Mossel-Neeman-Sly (12, 14), Massoulié (14), Bordenave-Lelarge-Massoulié (15).

A huge body of work for more general cases and different settings: survey by Abbe (18).

Bounded expected degrees

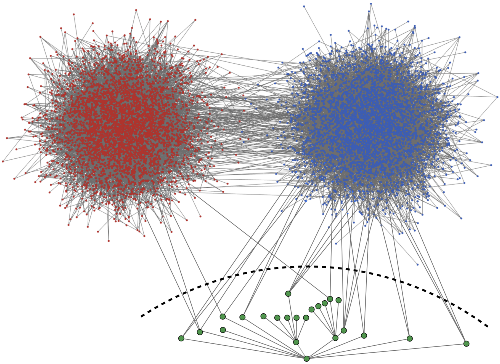


Figure: Abbe et al. (2018), $a = 2.2$, $b = 0.06$, $n = 100000$, apply spectral method directly on A

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

Detection by self-avoiding walks

Detection by self-avoiding walks

When $p = \frac{a}{n}, q = \frac{b}{n}$:

- Locally tree-like structure appears, few cycles.

Detection by self-avoiding walks

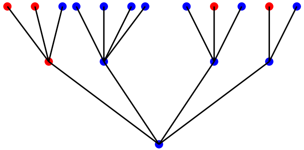
When $p = \frac{a}{n}, q = \frac{b}{n}$:

- Locally tree-like structure appears, few cycles.
- local neighborhood of $G(n, \frac{c}{n})$ is close to a Galton-Watson tree with offspring distribution Pois(c).

Detection by self-avoiding walks

When $p = \frac{a}{n}, q = \frac{b}{n}$:

- Locally tree-like structure appears, few cycles.
- local neighborhood of $G(n, \frac{c}{n})$ is close to a Galton-Watson tree with offspring distribution Pois(c).

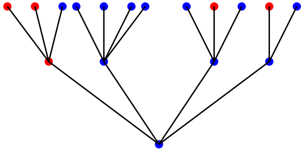


- Self-avoiding walks on trees are simple.

Detection by self-avoiding walks

When $p = \frac{a}{n}, q = \frac{b}{n}$:

- Locally tree-like structure appears, few cycles.
- local neighborhood of $G(n, \frac{c}{n})$ is close to a Galton-Watson tree with offspring distribution Pois(c).

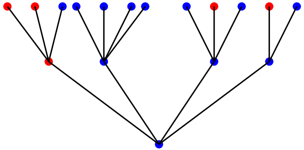


- Self-avoiding walks on trees are simple.
- Massoulié (14): **self-avoiding walk matrix** $B^{(\ell)}$.
- $B_{ij}^{(\ell)}$ = the number of self-avoiding walks of length $\ell = c \log n$ from i to j.

Detection by self-avoiding walks

When $p = \frac{a}{n}, q = \frac{b}{n}$:

- Locally tree-like structure appears, few cycles.
- local neighborhood of $G(n, \frac{c}{n})$ is close to a Galton-Watson tree with offspring distribution Pois(c).



- Self-avoiding walks on trees are simple.
- Massoulié (14): **self-avoiding walk matrix** $B^{(\ell)}$.
- $B_{ij}^{(\ell)}$ = the number of self-avoiding walks of length $\ell = c \log n$ from i to j.
- The second eigenvector of $B^{(\ell)}$ can be used to estimate $\sigma = (\sigma_1, \dots, \sigma_n)$, better than random guess.

Detection by self-avoiding walks

Let $\alpha := \frac{a+b}{2}, \beta := \frac{a-b}{2}$. Assume $\beta^2 > \alpha$.

Detection by self-avoiding walks

Let $\alpha := \frac{a+b}{2}, \beta := \frac{a-b}{2}$. Assume $\beta^2 > \alpha$.

- Show $v_2(B^{(\ell)}) \approx B^{(\ell)}\sigma$ (asymptotically aligned).

Detection by self-avoiding walks

Let $\alpha := \frac{a+b}{2}, \beta := \frac{a-b}{2}$. Assume $\beta^2 > \alpha$.

- Show $v_2(B^{(\ell)}) \approx B^{(\ell)}\sigma$ (asymptotically aligned).
- $(B^{(\ell)}\sigma)_i = \sum_j B_{ij}^{(\ell)}\sigma_j$.

Detection by self-avoiding walks

Let $\alpha := \frac{a+b}{2}, \beta := \frac{a-b}{2}$. Assume $\beta^2 > \alpha$.

- Show $v_2(B^{(\ell)}) \approx B^{(\ell)}\sigma$ (asymptotically aligned).
- $(B^{(\ell)}\sigma)_i = \sum_j B_{ij}^{(\ell)}\sigma_j$. Assume the ℓ-neighborhood of i is a tree, then $B_{ij}^{(\ell)} = 1$ if $d(i,j) = \ell$.

Detection by self-avoiding walks

Let $\alpha := \frac{a+b}{2}, \beta := \frac{a-b}{2}$. Assume $\beta^2 > \alpha$.

- Show $v_2(B^{(\ell)}) \approx B^{(\ell)}\sigma$ (asymptotically aligned).
- $(B^{(\ell)}\sigma)_i = \sum_j B_{ij}^{(\ell)}\sigma_j$. Assume the ℓ-neighborhood of i is a tree, then $B_{ij}^{(\ell)} = 1$ if $d(i,j) = \ell$.

$$(B^{(\ell)}\sigma)_i = \sum_{j:d(i,j)=\ell} \sigma_j.$$

Detection by self-avoiding walks

Let $\alpha := \frac{a+b}{2}, \beta := \frac{a-b}{2}$. Assume $\beta^2 > \alpha$.

- Show $v_2(B^{(\ell)}) \approx B^{(\ell)}\sigma$ (asymptotically aligned).
- $(B^{(\ell)}\sigma)_i = \sum_j B_{ij}^{(\ell)}\sigma_j$. Assume the ℓ-neighborhood of i is a tree, then $B_{ij}^{(\ell)} = 1$ if $d(i,j) = \ell$.

$$(B^{(\ell)}\sigma)_i = \sum_{j:d(i,j)=\ell} \sigma_j.$$

- The boundary of ℓ-neighborhood of i has size $\approx \alpha^\ell = n^{c \log \alpha}$, $|\sum_{j:d(i,j)=\ell} \sigma_j| \approx \beta^\ell = n^{c \log \beta}$.

Detection by self-avoiding walks

Let $\alpha := \frac{a+b}{2}, \beta := \frac{a-b}{2}$. Assume $\beta^2 > \alpha$.

- Show $v_2(B^{(\ell)}) \approx B^{(\ell)}\sigma$ (asymptotically aligned).
- $(B^{(\ell)}\sigma)_i = \sum_j B^{(\ell)}_{ij}\sigma_j$. Assume the ℓ-neighborhood of i is a tree, then $B^{(\ell)}_{ij} = 1$ if $d(i,j) = \ell$.
$$(B^{(\ell)}\sigma)_i = \sum_{j:d(i,j)=\ell} \sigma_j.$$
- The boundary of ℓ-neighborhood of i has size $\approx \alpha^\ell = n^{c\log\alpha}$, $|\sum_{j:d(i,j)=\ell}\sigma_j| \approx \beta^\ell = n^{c\log\beta}$. The sign of $(B^{(\ell)}\sigma)_i$ is correlated with σ_i.

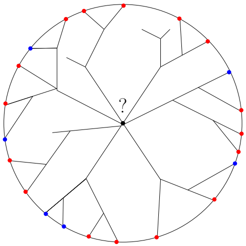
Detection by self-avoiding walks

Let $\alpha := \frac{a+b}{2}, \beta := \frac{a-b}{2}$. Assume $\beta^2 > \alpha$.

- Show $v_2(B^{(\ell)}) \approx B^{(\ell)}\sigma$ (asymptotically aligned).
- $(B^{(\ell)}\sigma)_i = \sum_j B^{(\ell)}_{ij}\sigma_j$. Assume the ℓ-neighborhood of i is a tree, then $B^{(\ell)}_{ij} = 1$ if $d(i,j) = \ell$.

$$(B^{(\ell)}\sigma)_i = \sum_{j:d(i,j)=\ell} \sigma_j.$$

- The boundary of ℓ-neighborhood of i has size $\approx \alpha^\ell = n^{c\log\alpha}$, $|\sum_{j:d(i,j)=\ell}\sigma_j| \approx \beta^\ell = n^{c\log\beta}$. The sign of $(B^{(\ell)}\sigma)_i$ is correlated with σ_i.



Hypergraph stochastic block model (HSBM)

H is *d-uniform* if each hyperedge has size d. Generate a random hypergraph H with label σ in two steps:

Hypergraph stochastic block model (HSBM)

H is *d-uniform* if each hyperedge has size d. Generate a random hypergraph H with label σ in two steps:

- labels $\{\sigma_i, i \in [n]\}$ are uniformly and i.i.d. drawn from $\{-1, +1\}$.
- Each hyperedge $e = \{v_1, \dots, v_d\}$ appears independently with probability

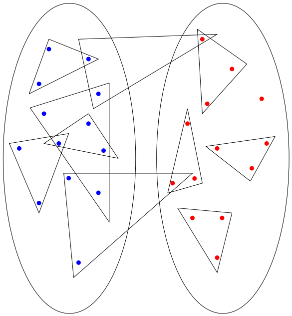
$$\mathbb{P}(e \in E) = \begin{cases} p & \text{if } \sigma_{v_1} = \cdots = \sigma_{v_d} \\ q & \text{otherwise.} \end{cases}$$

Hypergraph stochastic block model (HSBM)

H is *d-uniform* if each hyperedge has size d. Generate a random hypergraph H with label σ in two steps:

- labels $\{\sigma_i, i \in [n]\}$ are uniformly and i.i.d. drawn from $\{-1, +1\}$.
- Each hyperedge $e = \{v_1, \dots, v_d\}$ appears independently with probability

$$\mathbb{P}(e \in E) = \begin{cases} p & \text{if } \sigma_{v_1} = \cdots = \sigma_{v_d} \\ q & \text{otherwise.} \end{cases}$$

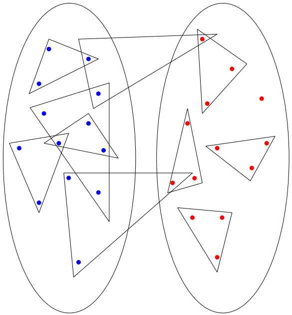


Hypergraph stochastic block model (HSBM)

H is *d-uniform* if each hyperedge has size d. Generate a random hypergraph H with label σ in two steps:

- labels $\{\sigma_i, i \in [n]\}$ are uniformly and i.i.d. drawn from $\{-1, +1\}$.
- Each hyperedge $e = \{v_1, \ldots, v_d\}$ appears independently with probability

$$\mathbb{P}(e \in E) = \begin{cases} p & \text{if } \sigma_{v_1} = \cdots = \sigma_{v_d} \\ q & \text{otherwise.} \end{cases}$$



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

Ghoshdastidar-Dukkipati (14, 15)

Community detection on HSBM

- Exact recovery: Chien-Lin-Wang (18), Kim-Bandeira-Goemans (18)
  $p = \frac{a \log n}{\binom{n}{d-1}}, q = \frac{b \log n}{\binom{n}{d-1}}$, exact recovery is possible if and only $(\sqrt{a} - \sqrt{b})^2 \geq 2^{d-1}$.

Community detection on HSBM

- Exact recovery: Chien-Lin-Wang (18), Kim-Bandeira-Goemans (18) $p = \frac{a \log n}{\binom{n}{d-1}}, q = \frac{b \log n}{\binom{n}{d-1}}$, exact recovery is possible if and only $(\sqrt{a} - \sqrt{b})^2 \geq 2^{d-1}$.
- Detection: Angelini-Caltagirone-Krzakala-Zdeborová (15) conjectured a phase transition when $p = \frac{a}{\binom{n}{d-1}}, q = \frac{b}{\binom{n}{d-1}}$.

Community detection on HSBM

- Exact recovery: Chien-Lin-Wang (18), Kim-Bandeira-Goemans (18)
  $p = \frac{a \log n}{\binom{n}{d-1}}, q = \frac{b \log n}{\binom{n}{d-1}}$, exact recovery is possible if and only $(\sqrt{a} - \sqrt{b})^2 \geq 2^{d-1}$.
- Detection: Angelini-Caltagirone-Krzakala-Zdeborová (15) conjectured a phase transition when $p = \frac{a}{\binom{n}{d-1}}, q = \frac{b}{\binom{n}{d-1}}$.
- Spectral method in the bounded expected degree regime?

Tensor

What we observe is an **adjacency tensor** T of order d with n^d many entries. $T_{i_1,\dots,i_d} = 1$ if $\{i_1, \dots, i_d\}$ is a hyperedge.

Tensor

What we observe is an **adjacency tensor** T of order d with n^d many entries. $T_{i_1,\ldots,i_d} = 1$ if $\{i_1, \ldots, i_d\}$ is a hyperedge.

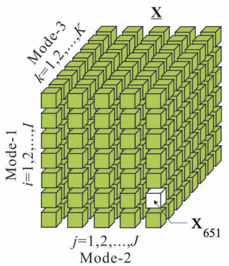


Figure: an order-3 tensor

Tensor

What we observe is an **adjacency tensor** T of order d with n^d many entries. $T_{i_1,\dots,i_d} = 1$ if $\{i_1, \dots, i_d\}$ is a hyperedge.

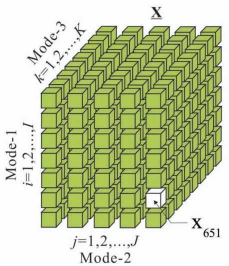


Figure: an order-3 tensor

- Eigenvalue, eigenvector, rank?

Tensor

What we observe is an **adjacency tensor** T of order d with n^d many entries.
$T_{i_1,\ldots,i_d} = 1$ if $\{i_1, \ldots, i_d\}$ is a hyperedge.

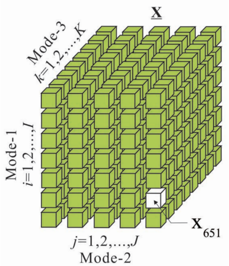


Figure: an order-3 tensor

- Eigenvalue, eigenvector, rank?
- Spectral norm
  $$\|T\|_2 = \sup_{x,y,z \in \mathbb{S}^{n-1}} \sum_{i,j,k} T_{ijk} x_i y_j z_k$$

Tensor

What we observe is an **adjacency tensor** T of order d with n^d many entries. $T_{i_1,\ldots,i_d} = 1$ if $\{i_1, \ldots, i_d\}$ is a hyperedge.

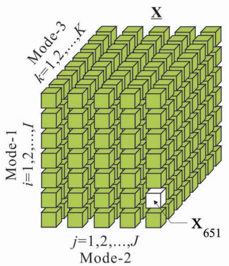


Figure: an order-3 tensor

- Eigenvalue, eigenvector, rank?
- Spectral norm
  $$\|T\|_2 = \sup_{x,y,z\in\mathbb{S}^{n-1}} \sum_{i,j,k} T_{ijk} x_i y_j z_k$$
- Most tensor problems are NP-hard (Hillar-Lim 13): rank, spectral norm, best low-rank approximation,...

Tensor

What we observe is an **adjacency tensor** T of order d with n^d many entries. $T_{i_1,\dots,i_d} = 1$ if $\{i_1,\dots,i_d\}$ is a hyperedge.

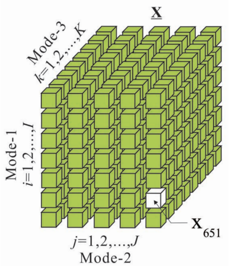


Figure: an order-3 tensor

- Eigenvalue, eigenvector, rank?
- Spectral norm
  $$\|T\|_2 = \sup_{x,y,z\in\mathbb{S}^{n-1}} \sum_{i,j,k} T_{ijk} x_i y_j z_k$$
- Most tensor problems are NP-hard (Hillar-Lim 13): rank, spectral norm, best low-rank approximation,...

Ke-Shi-Xia (20): Tensor unfolding and power iteration, $o(n)$ mis-classified vertices when the average degree $\gg \log^2(n)$.

Adjacency matrix of a hypergraph

- Define the **adjacency matrix** of H as $A_{ij} := \sum_{e \in E : \{i,j\} \subset e} T_e$, counting number of hyperedges containing i, j.

Adjacency matrix of a hypergraph

- Define the **adjacency matrix** of H as $A_{ij} := \displaystyle\sum_{e \in E : \{i,j\} \subset e} T_e$,

  counting number of hyperedges containing i, j.
- $A_{ij}^e := \mathbf{1}\{i, j \in e, e \in E\}$.

Adjacency matrix of a hypergraph

- Define the **adjacency matrix** of H as $A_{ij} := \sum_{e \in E : \{i,j\} \subset e} T_e$, counting number of hyperedges containing i, j.
- $A^e_{ij} := \mathbf{1}\{i, j \in e, e \in E\}$.
- $$\begin{aligned} \mathrm{tr} A^k &= \sum_{i_0, i_2, \ldots, i_{k-1}} A_{i_0 i_1} A_{i_2 i_3} \cdots A_{i_{k-1} i_0} \\ &= \sum_{\substack{i_0, i_1, \ldots, i_{k-1} \\ e_1, \ldots, e_k}} A^{e_1}_{i_0 i_1} \cdots A^{e_{k-1}}_{i_{k-2} i_{k-1}} A^{e_k}_{i_{k-1} i_0}, \end{aligned}$$

  which counts the number of closed walks of length k in H: $(i_0, e_1, i_1, \ldots, i_{k-1}, e_k, i_0)$.

Adjacency matrix of a hypergraph

- Define the **adjacency matrix** of H as $A_{ij} := \sum_{e \in E: \{i,j\} \subset e} T_e$,

  counting number of hyperedges containing i, j.
- $A_{ij}^e := \mathbf{1}\{i, j \in e, e \in E\}$.
- $$\begin{aligned} \mathrm{tr} A^k &= \sum_{i_0, i_2, \ldots, i_{k-1}} A_{i_0 i_1} A_{i_2 i_3} \cdots A_{i_{k-1} i_0} \\ &= \sum_{\substack{i_0, i_1, \ldots, i_{k-1} \\ e_1, \ldots, e_k}} A_{i_0 i_1}^{e_1} \cdots A_{i_{k-2} i_{k-1}}^{e_{k-1}} A_{i_{k-1} i_0}^{e_k}, \end{aligned}$$

  which counts the number of closed walks of length k in H: $(i_0, e_1, i_1, \ldots, i_{k-1}, e_k, i_0)$.

Detection by self-avoiding walks on hypergraphs

- A walk of length ℓ: $(v_0, e_1, v_1, \cdots e_\ell, v_\ell)$ such that $v_i \neq v_{i+1}$ and $\{v_{i-1}, v_i\} \subset e_i$.

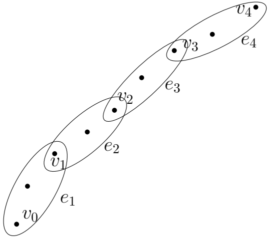
Detection by self-avoiding walks on hypergraphs

- A walk of length ℓ: $(v_0, e_1, v_1, \cdots e_\ell, v_\ell)$ such that $v_i \neq v_{i+1}$ and $\{v_{i-1}, v_i\} \subset e_i$.
- A **self-avoiding walk** of length ℓ in H is a walk $w = (v_0, e_1, v_1, \cdots, e_\ell, v_\ell)$ such that only consecutive hyperedges intersect at one vertex.

Detection by self-avoiding walks on hypergraphs

- A walk of length ℓ: $(v_0, e_1, v_1, \cdots e_\ell, v_\ell)$ such that $v_i \neq v_{i+1}$ and $\{v_{i-1}, v_i\} \subset e_i$.

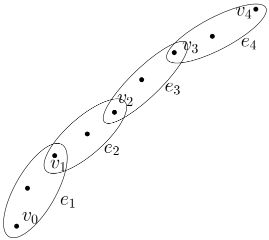
- A **self-avoiding walk** of length ℓ in H is a walk $w = (v_0, e_1, v_1, \cdots, e_\ell, v_\ell)$ such that only consecutive hyperedges intersect at one vertex.



Detection by self-avoiding walks on hypergraphs

- A walk of length ℓ: $(v_0, e_1, v_1, \cdots e_\ell, v_\ell)$ such that $v_i \neq v_{i+1}$ and $\{v_{i-1}, v_i\} \subset e_i$.

- A **self-avoiding walk** of length ℓ in H is a walk $w = (v_0, e_1, v_1, \cdots, e_\ell, v_\ell)$ such that only consecutive hyperedges intersect at one vertex.



- **self-avoiding walk matrix** $B^{(\ell)}$: $B_{ij}^{(\ell)}$ counts the number of self-avoiding walks of length ℓ from i to j.

Model parameters

- $\alpha := (d-1)\frac{a+(2^{d-1}-1)b}{2^{d-1}}$, expected degree of any vertex

Model parameters

- $\alpha := (d-1)\frac{a+(2^{d-1}-1)b}{2^{d-1}}$, expected degree of any vertex
- $\beta := (d-1)\frac{a-b}{2^{d-1}}$, discrepancy between numbers of $+,-$ labels of any vertex neighborhood

Model parameters

- $\alpha := (d-1)\frac{a+(2^{d-1}-1)b}{2^{d-1}}$, expected degree of any vertex
- $\beta := (d-1)\frac{a-b}{2^{d-1}}$, discrepancy between numbers of $+, -$ labels of any vertex neighborhood
- Angelini et al. (15): conjectured $\beta^2 = \alpha$ is the detection threshold for all $d \geq 2$.

Theorem (Pal-Z., 21)

Assume $\beta^2 > \alpha$. *Set* $\ell = c\log(n)$ *for a proper constant* c. *Let* x *be a unit second eigenvector of* $B^{(\ell)}$. *There exists a constant* t *such that, defining the label estimate* $\hat{\sigma}_i$ *as*

$$\hat{\sigma}_i = \begin{cases} +1 & \text{if } x_i\sqrt{n} \geq t, \\ -1 & \text{otherwise}, \end{cases}$$

then $\hat{\sigma}$ *is correlated with* σ *asymptotically almost surely.*

Theorem (Pal-Z., 21)

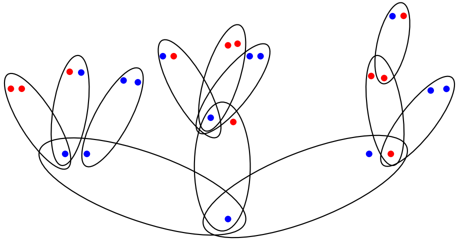
Assume $\beta^2 > \alpha$. Set $\ell = c\log(n)$ for a proper constant c. Let x be a unit second eigenvector of $B^{(\ell)}$. There exists a constant t such that, defining the label estimate $\hat{\sigma}_i$ as

$$\hat{\sigma}_i = \begin{cases} +1 & \text{if } x_i\sqrt{n} \geq t, \\ -1 & \text{otherwise}, \end{cases}$$

then $\hat{\sigma}$ is correlated with σ asymptotically almost surely.

- Dimension reduction: construct $B^{(\ell)}$ of n^2 entries from the adjacency tensor T of n^d entries.
- Spectral clustering: detect the community according to the second eigenvector.

Local structure: multi-type Poisson hypertrees



- Start with a root ρ with label $\tau(\rho)$, generate Pois $\left(\frac{\alpha}{d-1}\right)$ many hyperedges that pairwise intersects at ρ.
- Assign a *type* (the number of $+$ labels) to each hyperedge independently.
- Keep constructing subsequent generations by induction.

Proof sketch

- Show $v_2(B^{(\ell)}) \approx B^{(\ell)}\sigma$.

Proof sketch

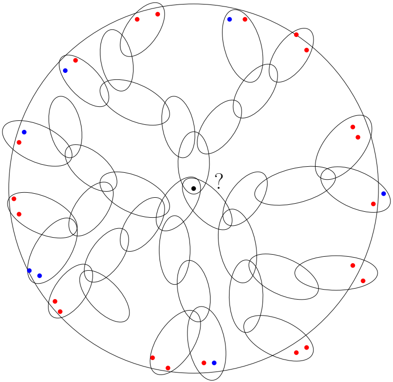
- Show $v_2(B^{(\ell)}) \approx B^{(\ell)}\sigma$.
- $(B^{(\ell)}\sigma)_i = \sum_j B^{(\ell)}_{ij}\sigma_j$. Assume the ℓ-neighborhood of i is a *hypertree*, then $B^{(\ell)}_{ij} = \mathbf{1}\{d(i,j)=\ell\} \implies (B^{(\ell)}\sigma)_i = \sum_{j:d(i,j)=\ell} \sigma_j$.

Proof sketch

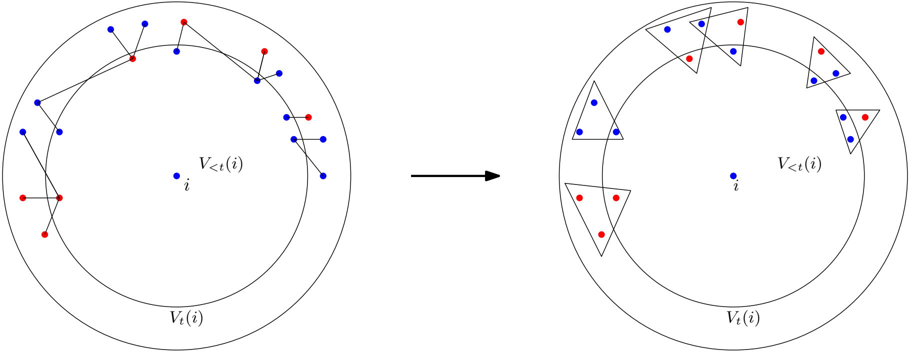
- Show $v_2(B^{(\ell)}) \approx B^{(\ell)}\sigma$.
- $(B^{(\ell)}\sigma)_i = \sum_j B_{ij}^{(\ell)}\sigma_j$. Assume the ℓ-neighborhood of i is a *hypertree*, then $B_{ij}^{(\ell)} = \mathbf{1}\{d(i,j)=\ell\} \implies (B^{(\ell)}\sigma)_i = \sum_{j:d(i,j)=\ell} \sigma_j$.
- The sign of $(B^{(\ell)}\sigma)_i$ is correlated with σ_i.

Proof sketch

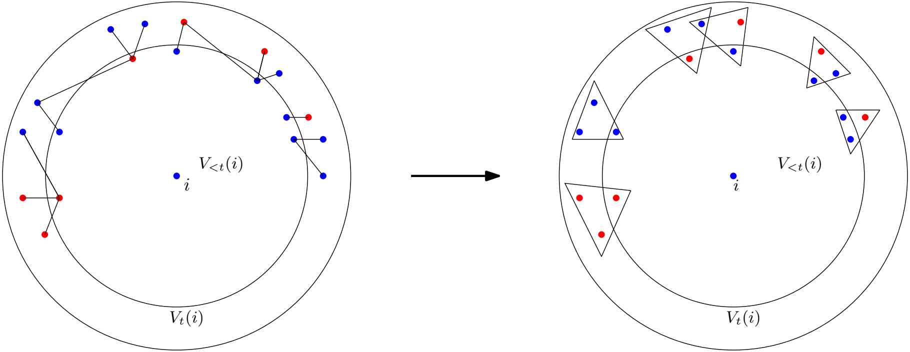
- Show $v_2(B^{(\ell)}) \approx B^{(\ell)}\sigma$.
- $(B^{(\ell)}\sigma)_i = \sum_j B^{(\ell)}_{ij}\sigma_j$. Assume the ℓ-neighborhood of i is a *hypertree*, then
$$B^{(\ell)}_{ij} = \mathbf{1}\{d(i,j)=\ell\} \implies (B^{(\ell)}\sigma)_i = \sum_{j:d(i,j)=\ell} \sigma_j.$$
- The sign of $(B^{(\ell)}\sigma)_i$ is correlated with σ_i.



Local Analysis



Local Analysis



Exploration process on hypergraphs. Control the boundary size and number of $\pm$ labels at distance t.

The moment method

The moment method

- Counting centered SAWs : $\Delta_{ij}^{(\ell)} := \sum_{w \in \mathrm{SAW}_{ij}} \prod_{t=1}^{\ell} (A_{i_{t-1} i_t}^{e_{i_t}} - \overline{A}_{i_{t-1} i_t}^{e_{i_t}})$.

The moment method

- Counting centered SAWs : $\Delta_{ij}^{(\ell)} := \sum_{w \in \mathrm{SAW}_{ij}} \prod_{t=1}^{\ell} (A_{i_{t-1} i_t}^{e_{i_t}} - \overline{A}_{i_{t-1} i_t}^{e_{i_t}})$.
- $B^{(\ell)} = \Delta^{(\ell)} + \sum_{m=1}^{\ell} (\Delta^{(\ell-m)} \overline{A} B^{(m-1)}) - \sum_{m=1}^{\ell} \Gamma^{(\ell,m)}$

The moment method

- Counting centered SAWs : $\Delta_{ij}^{(\ell)} := \sum_{w \in \mathrm{SAW}_{ij}} \prod_{t=1}^{\ell} (A_{i_{t-1}i_t}^{e_{i_t}} - \overline{A}_{i_{t-1}i_t}^{e_{i_t}})$.
- $B^{(\ell)} = \Delta^{(\ell)} + \sum_{m=1}^{\ell} (\Delta^{(\ell-m)} \overline{A} B^{(m-1)}) - \sum_{m=1}^{\ell} \Gamma^{(\ell,m)}$
- $\mathbb{E}\rho(\Delta^{(\ell)})^{2k} \leq \mathbb{E}\mathrm{tr}(\Delta^{(\ell)})^{2k}$, estimate by counting concatenations of $2k$ many self-avoiding walks of length ℓ.

The moment method

- Counting centered SAWs : $\Delta_{ij}^{(\ell)} := \sum_{w \in \mathrm{SAW}_{ij}} \prod_{t=1}^{\ell} (A_{i_{t-1}i_t}^{e_{i_t}} - \overline{A}_{i_{t-1}i_t}^{e_{i_t}})$.
- $B^{(\ell)} = \Delta^{(\ell)} + \sum_{m=1}^{\ell} (\Delta^{(\ell-m)} \overline{A} B^{(m-1)}) - \sum_{m=1}^{\ell} \Gamma^{(\ell,m)}$
- $\mathbb{E}\rho(\Delta^{(\ell)})^{2k} \leq \mathbb{E}\mathrm{tr}(\Delta^{(\ell)})^{2k}$, estimate by counting concatenations of $2k$ many self-avoiding walks of length ℓ.

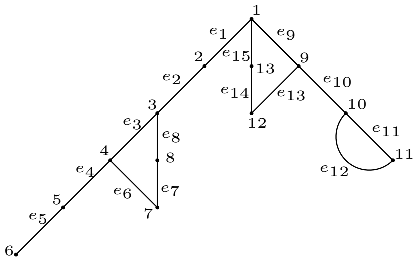


Figure: concatenations of 4 SAWs of length 5

Spectral gap for $B^{(\ell)}$

When $\beta^2 > \alpha$, $B^{(\ell)}$ has a spectral gap asymptotically almost surely:

- $\lambda_1(B^{(\ell)}) = \Theta(\alpha^\ell)$ up to a $\log n$ factor.
- $\lambda_2(B^{(\ell)}) = \Omega(\beta^\ell)$, and $\lambda_2(B^{(\ell)}) = O(n^{-\gamma}\alpha^\ell)$ for some $\gamma > 0$.
- $\lambda_3(B^{(\ell)}) = O(n^\epsilon \alpha^{\ell/2})$ for any $\epsilon > 0$.

Further Problems

- Non-backtracking operator for random hypergraphs with k blocks (work in progress with Ludovic Stephan)
- Non-uniform hypergraphs (with Ioana Dumitriu and Haixiao Wang)
- Impossibility for detection below the threshold
- Applications in tensor completion

Tensor Analog of Matrix Problems

Tensor Analog of Matrix Problems

Statistical and computational gap

- Tensor PCA: $X = \lambda v^{\otimes k} + Z$
  Montanari-Richard (14), Chen (18), Ben Arous-Mei-Montanari-Nica (17), Ben Arous-Gheissari-Jagannath (18), Wein-Alaoui-Moore (19), Huang-Huang-Yang-Cheng (20), Ding-Hopkins-Steurer (20), Ben Arous-Huang-Huang (21),...
- Tensor completion
  Jain-Oh (14), Ge-Huang-Jin-Yuan (15), Barak-Moitra (16), Xia-Yuan (17, 19), Yuan-Zhang (17), Ge-Ma (17), Potechin-Steurer (17), Montanari-Sun (18), Ghadermarzy-Plan-Yilmaz (18), ...

Tensor Analog of Matrix Problems

Statistical and computational gap

- Tensor PCA: $X = \lambda v^{\otimes k} + Z$
  Montanari-Richard (14), Chen (18), Ben Arous-Mei-Montanari-Nica (17), Ben Arous-Gheissari-Jagannath (18), Wein-Alaoui-Moore (19), Huang-Huang-Yang-Cheng (20), Ding-Hopkins-Steurer (20), Ben Arous-Huang-Huang (21),. . .
- Tensor completion
  Jain-Oh (14), Ge-Huang-Jin-Yuan (15), Barak-Moitra (16), Xia-Yuan (17, 19), Yuan-Zhang (17), Ge-Ma (17), Potechin-Steurer (17), Montanari-Sun (18), Ghadermarzy-Plan-Yilmaz (18), . . .

No such gap in many hypergraph community detection problems:

Exact recovery: Kim-Bandeira-Goemans (17, 18), Ahn-Lee-Suh (18), Chien-Lin-Wang (18), Zhang-Tan (21).

Conclusion

- Community detection on random hypergraphs can be analyzed by spectral methods on sparse random matrices.
- Moment methods can be applied to random hypergraphs.
- Sparse random tensors are not well understood.

Thank You!