Sparse random hypergraphs: Non-backtracking spectra and community detection

Ludovic Stephan

École Polytechnique Fédérale de Lausanne

Yizhe Zhu

University of California, Irvine

What is community detection ?

Figure taken from [Abbe '18]

What is community detection ?

$\simeq$

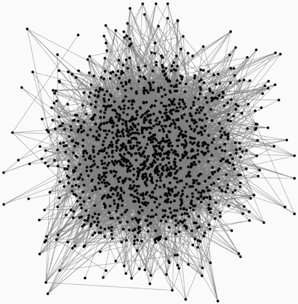


Figure: Scrambled and unscrambled graph

Figure taken from [Abbe '18]

A generative model : the Stochastic Block Model (SBM)

Generates a random graph with n vertices, k blocks.

Parameters: $P \in \mathbb{R}^{k \times k}$, $\pi \in \mathbb{R}^k$

- Type assignment σ i.i.d such that

$$\mathbb{P}(\sigma(x) = i) = \pi_i$$

- Graph generation: independent edges,

$$\mathbb{P}((x, y) \in E) = \frac{P_{\sigma(x)\sigma(y)}}{n}$$

SBM definition : Holland et al. '84

A generative model : the Stochastic Block Model (SBM)

Generates a random graph with n vertices, k blocks.

Parameters: $P \in \mathbb{R}^{k\times k}$, $\pi \in \mathbb{R}^k$

- Type assignment σ i.i.d such that

$$\mathbb{P}(\sigma(x) = i) = \pi_i$$

- Graph generation: independent edges,

$$\mathbb{P}((x, y) \in E) = \frac{P_{\sigma(x)\sigma(y)}}{n}$$

Goal: recover σ from $G \sim \mathrm{SBM}(n, \pi, P)$

SBM definition : Holland et al. ’84

Spectral methods

Simple case: $k = 2$, $\pi_i \equiv 1/2$, $P = \begin{pmatrix} p & q \\ q & p \end{pmatrix}$

$$\mathbb{E}[A] = \frac{1}{n} \left[\begin{array}{ccc|ccc} p & \dots & p & q & \dots & q \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ p & \dots & p & q & \dots & q \\ \hline q & \dots & q & p & \dots & p \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ q & \dots & q & p & \dots & p \end{array} \right]$$

Spectral methods

Simple case: $k = 2$, $\pi_i \equiv 1/2$, $P = \begin{pmatrix} p & q \\ q & p \end{pmatrix}$

$$\mathbb{E}[A] = \frac{1}{n}\left[\begin{array}{ccc|ccc} p & \dots & p & q & \dots & q \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ p & \dots & p & q & \dots & q \\ \hline q & \dots & q & p & \dots & p \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ q & \dots & q & p & \dots & p \end{array}\right] \quad v_1 = \left[\begin{array}{c} 1 \\ \vdots \\ 1 \\ \hline 1 \\ \vdots \\ 1 \end{array}\right]$$

$$\mu_1 = \frac{p+q}{2}$$

Spectral methods

Simple case: $k = 2$, $\pi_i \equiv 1/2$, $P = \begin{pmatrix} p & q \\ q & p \end{pmatrix}$

$$\mathbb{E}[A] = \frac{1}{n} \left[\begin{array}{ccc|ccc} p & \dots & p & q & \dots & q \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ p & \dots & p & q & \dots & q \\ \hline q & \dots & q & p & \dots & p \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ q & \dots & q & p & \dots & p \end{array}\right] \quad v_1 = \left[\begin{array}{c} 1 \\ \vdots \\ 1 \\ \hline 1 \\ \vdots \\ 1 \end{array}\right] \quad v_2 = \left[\begin{array}{c} 1 \\ \vdots \\ 1 \\ \hline -1 \\ \vdots \\ -1 \end{array}\right]$$

$$\mu_1 = \frac{p+q}{2} \quad \mu_2 = \frac{p-q}{2}$$

Spectral methods

In general :

- eigenvalues of $\mathbb{E}[A]$ the same as

$$M = P\,\mathrm{diag}(\pi)$$

- associated eigenvectors constant on the clusters

Spectral methods

In general :

- eigenvalues of $\mathbb{E}[A]$ the same as

$$M = P \,\mathrm{diag}(\pi)$$

- associated eigenvectors constant on the clusters

Constant average degree hypothesis:

$$\sum_j P_{ij}\pi_j = d \quad \forall i \in [k]$$

$\Rightarrow M\mathbf{1} = d\mathbf{1}$, $\mu_1 = d$, v_1 uninformative

Spectral methods: high degree

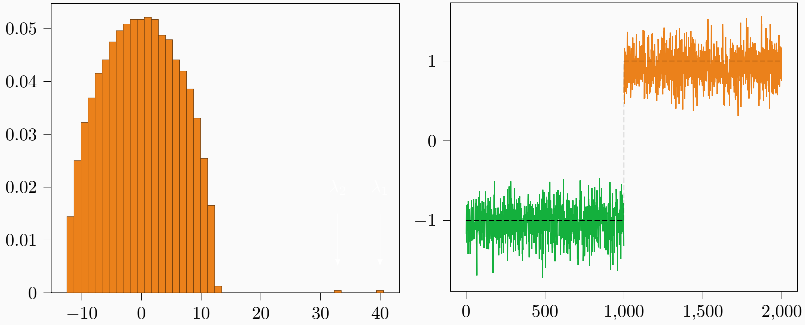


Figure: $n = 2000, d = 40$

When $p, q = \Omega(\log(n))$: $o(n)$ misclassifications by k-means

Feige-Ofek '05, Lei-Rinaldo '13

Spectral methods: low degree

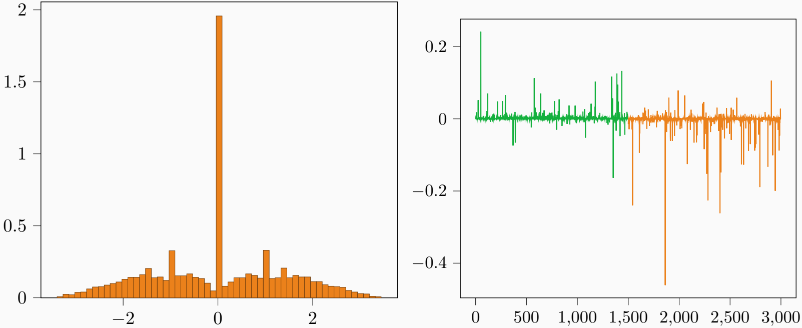


Figure: $n = 3000, d = 2$

High-degree vertices dominate the spectrum

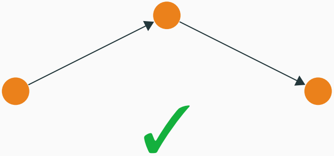
Krivelevich-Sudakov '01, Benaych-Georges et al. '19

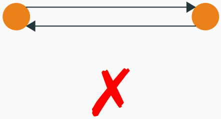
The non-backtracking matrix

Indexed by directed edges

$$\vec{E} = \{(x, y) \mid \{x, y\} \in E\}$$

$$B_{\vec{e},\vec{f}} = \begin{cases} 1 & \text{if } e_2 = f_1 \text{ and } e_1 \neq f_2 \\ 0 & \text{otherwise} \end{cases}$$





Spectral redemption theorem

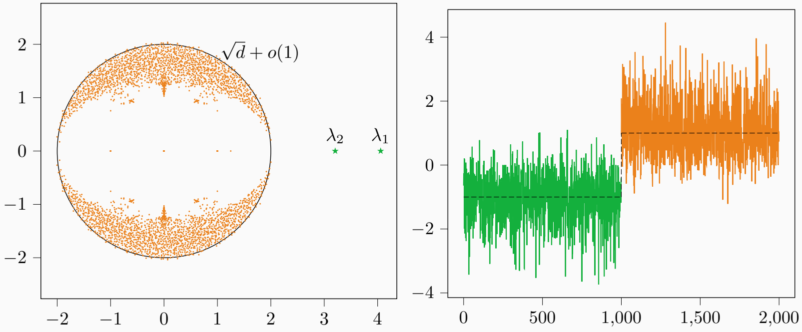


Figure: n=2000, d=4

$$\begin{cases} \lambda_i(B) = \mu_i + o(1) & \text{if } \mu_i^2 > d \\ |\lambda_i(B)| \leq \sqrt{d} & \text{otherwise} \end{cases}$$

Krzakala et al '13, Bordenave et al '15

Higher-dimensional relations

Published: 12 May 2004

Pizza and risk of acute myocardial infarction

S Gallus, A Tavani & C La Vecchia

European Journal o

10k Accesses | 6

Abstract

Objectives: Pizz

the data are limi

myocardial infar

Epidemiology | Free Access

Does pizza protect against cancer?

Silvano Gallus, Cristina Bosetti, Eva Negri, Renato Talamini, Maurizio Montella, Ettore Conti, Silvia Franceschi, Carlo La Vecchia

First published: 21 Ju

SECTIONS

Abstract

We analyzed th

network of cas

RESEARCH PAPERS: LIFESTYLE: NUTRITION

Pizza consumption and the risk of breast, ovarian and prostate cancer

Gallus, Silvano[a]; Talamini, Renato[b]; Bosetti, Cristina[a]; Negri, Eva[a]; Montella, Maurizio[c]; Franceschi, Silvia[d]; Giacosa, Attilio[e]; La Vecchia, Carlo[a f]

Author Information

European Journal of Cancer Prevention: February 2006 - Volume 15 - Issue 1 - p 74-76

doi: 10.1097/01.cej.0000186632.04625.f6

BUY

Metrics

Hypergraphs

$G = (V, H)$ with $H \subseteq 2^V$, q-uniform if $H \subseteq \binom{V}{q}$.

Figure from R. Mulas
Non-uniform hypergraphs: Chodrow et al. '22

HSBM

Immediate generalization of the SBM:

Affinity *tensor* P of dimension q

$$\mathbb{P}(\{x_1, \dots, x_q\} \in H) = \frac{P_{\underline{\sigma}(x)}}{\binom{n}{q-1}},$$

where $\underline{\sigma}(x) = (\sigma(x_1), \dots, \sigma(x_q))$.

Tensor representation

Intuitive way to represent adjacencies: adjacency tensor T

$$T_{x_1,\dots,x_q} = 1 \quad \Leftrightarrow \quad \{x_1, \dots, x_q\} \in H$$

Tensor representation

Intuitive way to represent adjacencies: adjacency tensor T

$$T_{x_1,\ldots,x_q} = 1 \quad \Leftrightarrow \quad \{x_1,\ldots,x_q\} \in H$$

Problem: everything with tensors is hard[1]!

[1] Hillar and Lim '09

Adjacency matrix

Defined as

$$A_{xy} = |\{e \in H \mid x, y \in e\}|$$

Same properties as before:

- $\mathbb{E}[A]$ is low-rank, eigenvalues

$$d \geq \mu_2 \geq \cdots \geq \mu_k$$

- Spectral method on A fails for $d = O(1)$

Non-backtracking matrix

Defined on pointed edges

$$\vec{H} = \{(x, e) \mid e \in H, x \in e\}$$

$$B_{(x,e),(y,f)} = \begin{cases} 1 & \text{if } y \in e,\ y \neq x,\ f \neq e \\ 0 & \text{otherwise} \end{cases}$$

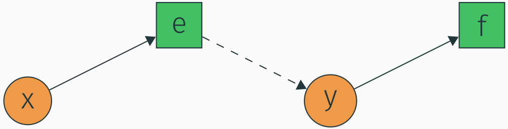
Storm '06

Non-backtracking matrix

Defined on pointed edges

$$\vec{H} = \{(x, e) \mid e \in H, x \in e\}$$

$$B_{(x,e),(y,f)} = \begin{cases} 1 & \text{if } y \in e,\ y \neq x,\ f \neq e \\ 0 & \text{otherwise} \end{cases}$$



Storm '06

Spectral redemption (again)

Same reasoning as the $q = 2$ case:

Conjecture (Angelini et al., '15)

If

$$(q-1)\mu_2^2 > d,$$

both BP and a spectral method based on B recover the communities.

Spectral redemption theorem (again)

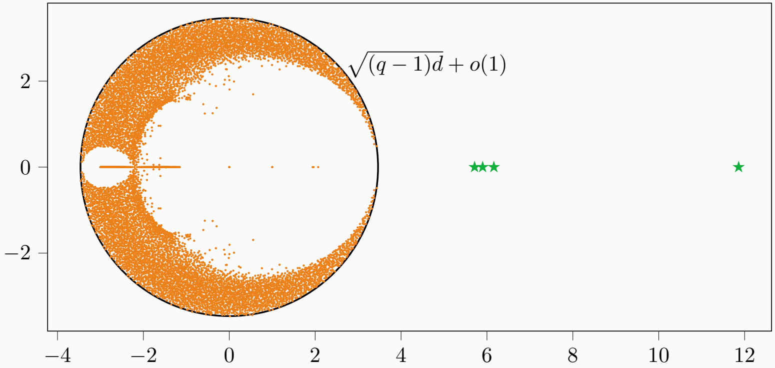


Figure: $n = 6000$, $k = q = 4$, $d = 4$, $\mu_2 = 2$

S.-Zhu, '22

Dimension reduction

Two problems:

- B has size $q|H| \sim qdn$, can be very large !
- need an embedding procedure into $\mathbb{R}^n$

We do both at once !

Angelini et al. '15

Dimension reduction

Two problems:

- B has size $q|H| \sim qdn$, can be very large !
- need an embedding procedure into $\mathbb{R}^n$

We do both at once !

Define

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

with D the diagonal degree matrix:

$$D_{xx} = |\{e \in H \,|\, x \in e\}| = [A\mathbf{1}]_x$$

Angelini et al. '15

Spectrum of $\tilde{B}$

Ihara-Bass extension:

Theorem (S.-Zhu '22)

For any $z \in \mathbb{C}$,

$$\det(B - zI) = (z-1)^{(q-1)m-n}(z+(q-1))^{m-n}\det(\tilde{B} - zI)$$

As a result,

$$Sp(B) = Sp(\tilde{B}) \cup \{1, -(q-1)\}$$

Bass '92 for q=2, Storm '06 for regular graphs

Eigenvectors of $\tilde{B}$

Theorem (S.-Zhu '22)

Assume that $(q-1)\mu_i^2 > d$*, and let* $\begin{pmatrix} u_i \\ v_i \end{pmatrix}$ *an eigenvector associated to* $\lambda_i(\tilde{B})$*. Then there exists an eigenvector* ϕ_i *of* $\mathbb{E}[A]$ *such that*

$$\langle v_i, \phi_i \rangle = \sqrt{\frac{1-\tau_i}{1+\frac{q-2}{(q-1)\mu_i}}} + o(1)$$

τ_i is the inverse SNR for μ_i:

$$\tau_i = \frac{d}{(q-1)\mu_i^2} < 1$$

Illustration

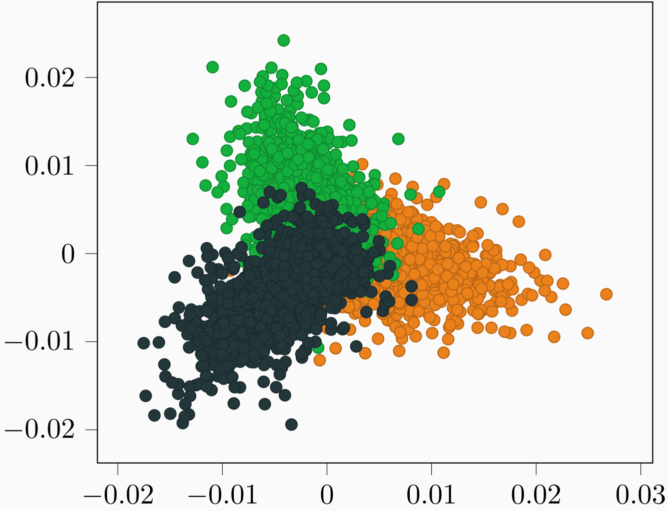


Figure: Plot of $(v_2(x), v_3(x))$ for $x \in V$. $n = 20000$, $q = d = 4$, $k = 3$.

Thank you !