Spectral clustering in sparse heterogeneous networks

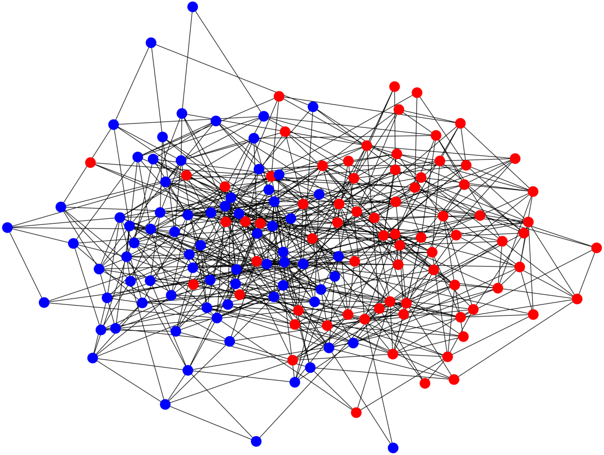
Lorenzo Dall'Amico

Romain Couillet, Nicolas Tremblay

Laboratoire Gipsa-lab, UMR 5216, CNRS, UGA
11 rue des mathématiques 38420 Grenoble, France
lorenzo.dall-amico@gipsa-lab.fr

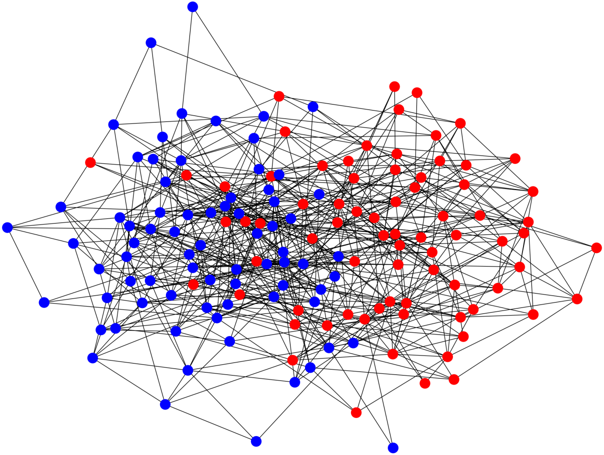
May 10, 2019

What and why



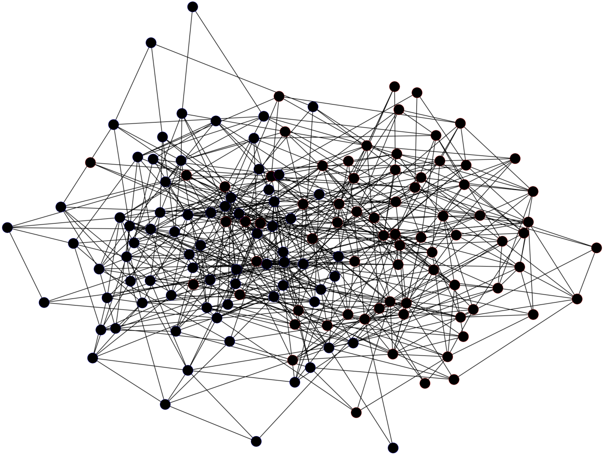
What and why

The solution...



What and why

The problem



The spectral techniques

Information inside the eigenvectors

The spectral techniques

Information inside the eigenvectors

FAST

The spectral techniques

Information inside the eigenvectors

FAST

UNSUPERVISED

The spectral techniques

Information inside the eigenvectors

FAST — TRACTABLE — UNSUPERVISED

The spectral techniques

Information inside the eigenvectors

FAST

TRACTABLE

UNSUPERVISED

Dense regime

$d \sim n$

The spectral techniques

Information inside the eigenvectors

FAST

TRACTABLE

UNSUPERVISED

Dense regime

$d \sim n$

Average degree

The spectral techniques

Information inside the eigenvectors

FAST

TRACTABLE

UNSUPERVISED

Dense regime

$d \sim n$

Average degree

Size of the network

Example

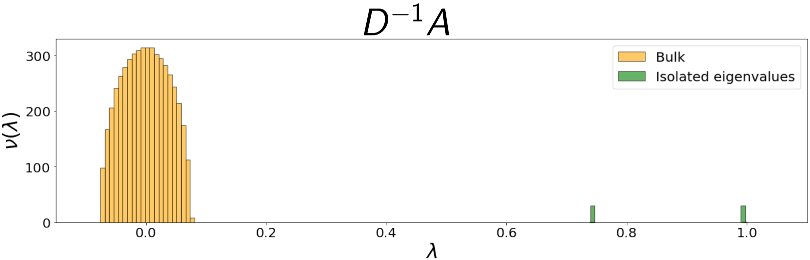
- A : adjacency matrix, $A_{ij} = 1$ if (ij) are connected, zero else

Example

- A : adjacency matrix, $A_{ij} = 1$ if (ij) are connected, zero else
- D : degree matrix $D = diag(A\mathbb{1})$

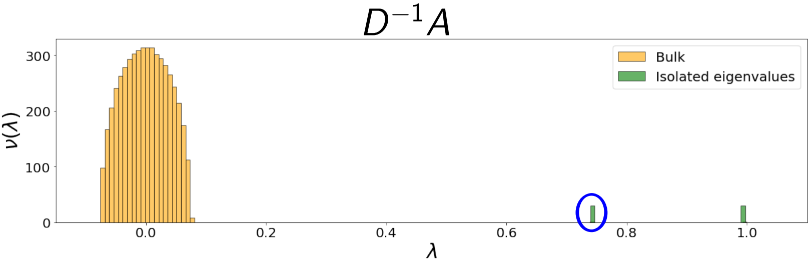
Example

- A : adjacency matrix, $A_{ij} = 1$ if (ij) are connected, zero else
- D : degree matrix $D = diag(A\mathbb{1})$

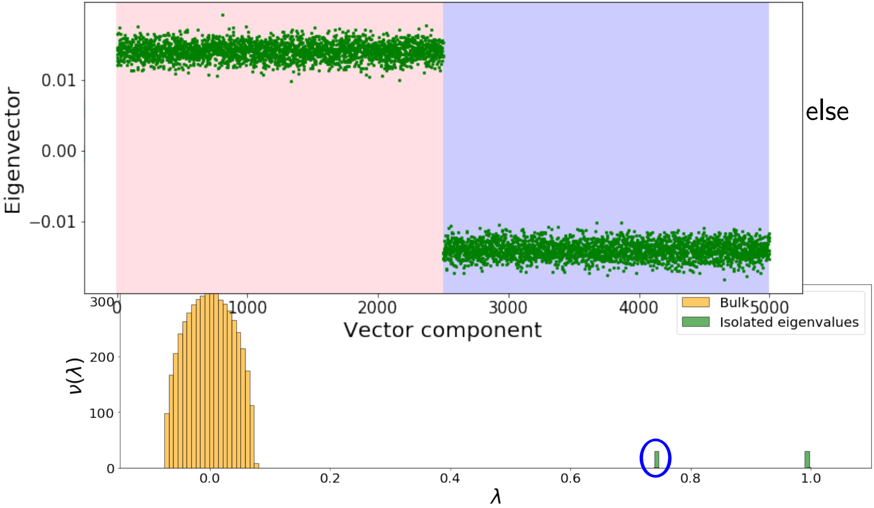


Example

- A : adjacency matrix, $A_{ij} = 1$ if (ij) are connected, zero else
- D : degree matrix $D = diag(A\mathbb{1})$



Example


Eigenvector
0.01
0.00
−0.01
0
1000
2000
3000
4000
5000
Vector component
else
300
200
100
0
$\nu(\lambda)$
0.0
0.2
0.4
0.6
0.8
1.0
λ
Bulk
Isolated eigenvalues

Sparsity

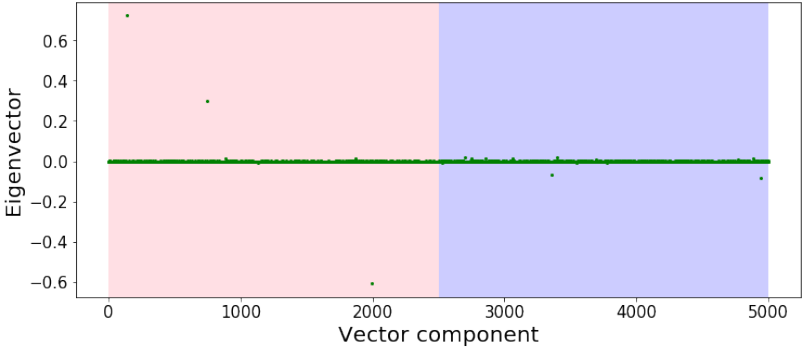
Real networks are sparse

$$d \approx const$$

Sparsity

Real networks are sparse

$$d \approx const$$



Two solutions

- Non-backtracking matrix[1] $B \in \{0,1\}^{2|\mathcal{E}|\times 2|\mathcal{E}|}$
  $B_{(ij),(kl)} = \delta_{jk}(1-\delta_{il})$

[1] Krzakala *et al*, Spectral redemption in clustering sparse networks, PNAS 2013

Two solutions

- Non-backtracking matrix[1] $B \in \{0,1\}^{2|\mathcal{E}| \times 2|\mathcal{E}|}$
  $B_{(ij),(kl)} = \delta_{jk}(1 - \delta_{il})$
- Bethe-Hessian matrix[2]
  $H_r = (r^2 - 1)I_n + D - rA, \; r \in \mathbb{R}$

[1] Krzakala *et al.*, Spectral redemption in clustering sparse networks, PNAS 2013
[2] Saade *et al.*, Spectral clustering of graphs with the Bethe-Hessian, NIPS 2014

Two solutions

- Non-backtracking matrix[1] $B \in \{0,1\}^{2|\mathcal{E}|\times 2|\mathcal{E}|}$
  $B_{(ij),(kl)} = \delta_{jk}(1-\delta_{il})$
- Bethe-Hessian matrix[2]
  $H_r = (r^2 - 1)I_n + D - rA, \; r \in \mathbb{R}$

[1] Krzakala *et al.*, Spectral redemption in clustering sparse networks, PNAS 2013
[2] Saade *et al.*, Spectral clustering of graphs with the Bethe-Hessian, NIPS 2014

Two solutions

- Non-backtracking matrix[1] $B \in \{0,1\}^{2|\mathcal{E}|\times 2|\mathcal{E}|}$
  $B_{(ij),(kl)} = \delta_{jk}(1-\delta_{il})$
- Bethe-Hessian matrix[2]
  $H_r = (r^2-1)I_n + D - rA,\ r \in \mathbb{R}$

Important theoretical results:

[1] Krzakala *et al.*, Spectral redemption in clustering sparse networks, PNAS 2013

[2] Saade *et al.*, Spectral clustering of graphs with the Bethe-Hessian, NIPS 2014

Two solutions

- Non-backtracking matrix[1] $B \in \{0,1\}^{2|\mathcal{E}|\times 2|\mathcal{E}|}$
  $B_{(ij),(kl)} = \delta_{jk}(1-\delta_{il})$
- Bethe-Hessian matrix[2]
  $H_r = (r^2-1)I_n + D - rA, \; r \in \mathbb{R}$

Important theoretical results:

1. Work in the sparse regime

[1] Krzakala *et al.*, Spectral redemption in clustering sparse networks, PNAS 2013
[2] Saade *et al.*, Spectral clustering of graphs with the Bethe-Hessian, NIPS 2014

Two solutions

- Non-backtracking matrix[1] $B \in \{0,1\}^{2|\mathcal{E}|\times 2|\mathcal{E}|}$
  $B_{(ij),(kl)} = \delta_{jk}(1-\delta_{il})$
- Bethe-Hessian matrix[2]
  $H_r = (r^2-1)I_n + D - rA, \; r \in \mathbb{R}$

Important theoretical results:

1. Work in the sparse regime
2. Work asymptotically down to the detectability threshold

[1] Krzakala *et al.*, Spectral redemption in clustering sparse networks, PNAS 2013
[2] Saade *et al.*, Spectral clustering of graphs with the Bethe-Hessian, NIPS 2014

Two solutions

- Non-backtracking matrix[1] $B \in \{0,1\}^{2|\mathcal{E}|\times 2|\mathcal{E}|}$
  $B_{(ij),(kl)} = \delta_{jk}(1-\delta_{il})$
- Bethe-Hessian matrix[2]
  $H_r = (r^2 - 1)I_n + D - rA, \; r \in \mathbb{R}$

Important theoretical results:

1. Work in the sparse regime
2. Work asymptotically down to the detectability threshold

Only for homogeneous degree distribution

[1] Krzakala *et al.*, Spectral redemption in clustering sparse networks, PNAS 2013

[2] Saade *et al.*, Spectral clustering of graphs with the Bethe-Hessian, NIPS 2014

The DC-SBM

$$\mathbb{P}(A_{ij} = 1 | \sigma_i, \sigma_j, q_i, q_j) = q_i q_j \frac{C(\sigma_i, \sigma_j)}{n}$$

The DC-SBM

$$\mathbb{P}(A_{ij} = 1 | \sigma_i, \sigma_j, q_i, q_j) = q_i q_j \frac{C(\sigma_i, \sigma_j)}{n}$$

$$\mathbb{E}[q] = 1, \quad \mathbb{E}[q^2] = \Phi, \quad C = \begin{pmatrix} c_{\rm in} & c_{\rm out} \\ c_{\rm out} & c_{\rm in} \end{pmatrix}, \quad c = \frac{c_{\rm in} + c_{\rm out}}{2}$$

The DC-SBM

$$\mathbb{P}(A_{ij} = 1 | \sigma_i, \sigma_j, q_i, q_j) = q_i q_j \frac{C(\sigma_i, \sigma_j)}{n}$$

$$\mathbb{E}[q] = 1, \quad \mathbb{E}[q^2] = \Phi, \quad C = \begin{pmatrix} c_{\rm in} & c_{\rm out} \\ c_{\rm out} & c_{\rm in} \end{pmatrix}, \quad c = \frac{c_{\rm in} + c_{\rm out}}{2}$$

Detectability threshold[3] :

$$\alpha := \frac{c_{\rm in} - c_{\rm out}}{\sqrt{c}} \geq \frac{2}{\sqrt{\Phi}}$$

[3] Gulikers *et al.* An impossibility result for reconstruction in the degree-corrected stochastic block model, The Annals of Applied Probability 2018

Goal

$$H_r = (r^2 - 1)I_n + D - rA, \quad r \in \mathbb{R}$$

Second smallest eigenvector of H_r for:

Goal

$$H_r = (r^2 - 1)I_n + D - rA, \quad r \in \mathbb{R}$$

Second smallest eigenvector of H_r for:

- Community detection in sparse and heterogeneous networks

Goal

$$H_r = (r^2 - 1)I_n + D - rA, \quad r \in \mathbb{R}$$

Second smallest eigenvector of H_r for:

- Community detection in sparse and heterogeneous networks
- Reach the detectability threshold

Goal

$$H_r = (r^2 - 1)I_n + D - rA, \quad r \in \mathbb{R}$$

Second smallest eigenvector of H_r for:

- Community detection in sparse and heterogeneous networks
- Reach the detectability threshold

Pick the good value of r to:

Goal

$$H_r = (r^2 - 1)I_n + D - rA, \quad r \in \mathbb{R}$$

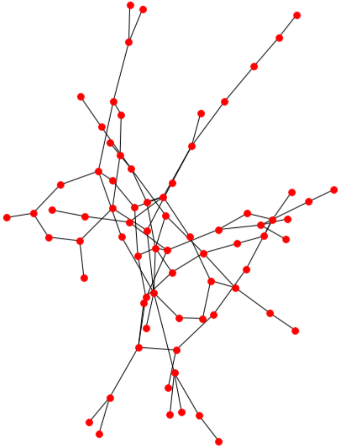
Second smallest eigenvector of H_r for:

- Community detection in sparse and heterogeneous networks
- Reach the detectability threshold

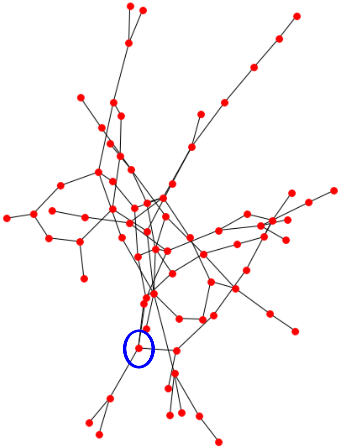
Pick the good value of r to:

- Retrieve $\boldsymbol{\sigma}$ regardless of $\boldsymbol{q}$

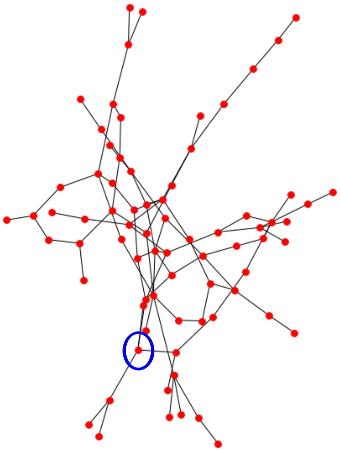
Tree like approximation

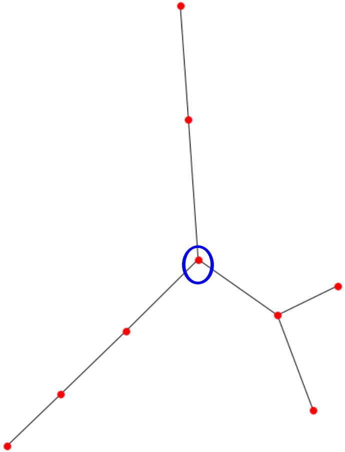


Tree like approximation

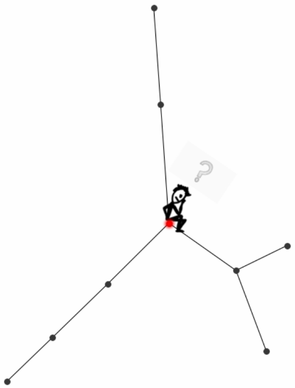


Tree like approximation



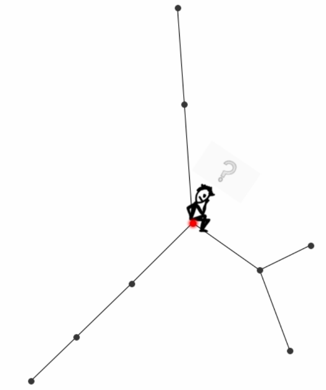


Tree like approximation



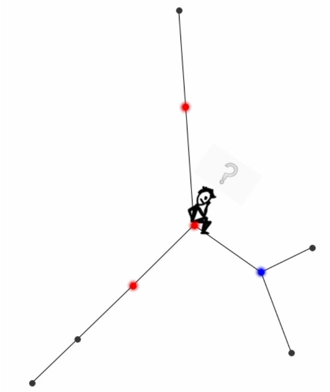
Tree like approximation

$$\mathbb{P}(\sigma_i = \sigma_j | A_{ij} = 1) = \frac{c_{\rm in}}{c_{\rm in} + c_{\rm out}} \qquad \mathbb{P}(\sigma_i \neq \sigma_j | A_{ij} = 1) = \frac{c_{\rm out}}{c_{\rm in} + c_{\rm out}}$$



Tree like approximation

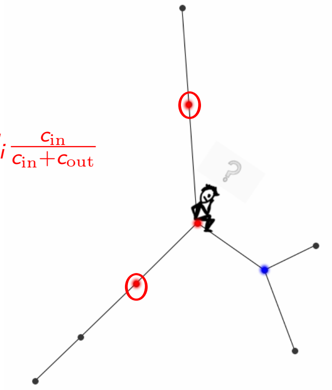
$$\mathbb{P}(\sigma_i = \sigma_j | A_{ij} = 1) = \frac{c_{\rm in}}{c_{\rm in}+c_{\rm out}} \qquad \mathbb{P}(\sigma_i \neq \sigma_j | A_{ij} = 1) = \frac{c_{\rm out}}{c_{\rm in}+c_{\rm out}}$$



Tree like approximation

$$\mathbb{P}(\sigma_i = \sigma_j | A_{ij} = 1) = \frac{c_{\rm in}}{c_{\rm in}+c_{\rm out}} \qquad \mathbb{P}(\sigma_i \neq \sigma_j | A_{ij} = 1) = \frac{c_{\rm out}}{c_{\rm in}+c_{\rm out}}$$

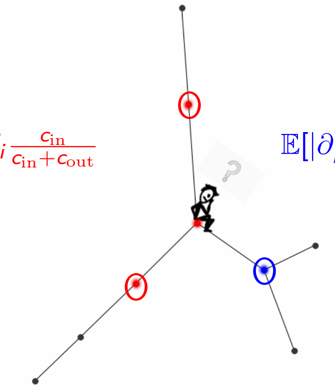
$$\mathbb{E}[|\partial_i^{(s)}|] = d_i \frac{c_{\rm in}}{c_{\rm in}+c_{\rm out}}$$



Tree like approximation

$$\mathbb{P}(\sigma_i = \sigma_j | A_{ij} = 1) = \frac{c_{\rm in}}{c_{\rm in}+c_{\rm out}} \qquad \mathbb{P}(\sigma_i \neq \sigma_j | A_{ij} = 1) = \frac{c_{\rm out}}{c_{\rm in}+c_{\rm out}}$$

$$\mathbb{E}[|\partial_i^{(s)}|] = d_i \frac{c_{\rm in}}{c_{\rm in}+c_{\rm out}} \qquad \mathbb{E}[|\partial_i^{(o)}|] = d_i \frac{c_{\rm out}}{c_{\rm in}+c_{\rm out}}$$



The optimal choice or r

What is the informative eigenvector of H_r?

The optimal choice or r

What is the informative eigenvector of H_r?

What is r such that : $H_r\boldsymbol{\sigma} \approx \lambda\boldsymbol{\sigma}$?

The optimal choice or r

What is the informative eigenvector of H_r?

What is r such that : $H_r\boldsymbol{\sigma} \approx \lambda\boldsymbol{\sigma}$?

$$\mathbb{E}[[((r^2-1)I_n+D-rA)\boldsymbol{\sigma}]_i]$$

The optimal choice or r

What is the informative eigenvector of H_r?

What is r such that : $H_r\boldsymbol{\sigma} \approx \lambda\boldsymbol{\sigma}$?

$$\mathbb{E}[[((r^2-1)I_n+D-rA)\boldsymbol{\sigma}]_i] = \sigma_i\left[(r^2-1) + d_i\left(1 - r\frac{c_{\rm in} - c_{\rm out}}{c_{\rm in} + c_{\rm out}}\right)\right]$$

The optimal choice or r

What is the informative eigenvector of H_r?

What is r such that : $H_r\boldsymbol{\sigma} \approx \lambda\boldsymbol{\sigma}$?

$$\mathbb{E}[[((r^2-1)I_n+D-rA)\boldsymbol{\sigma}]_i] = \sigma_i\left[(r^2-1)+d_i\left(1-r\frac{c_{\rm in}-c_{\rm out}}{c_{\rm in}+c_{\rm out}}\right)\right]$$

$$r_{\rm opt} = \zeta \equiv \frac{c_{\rm in}+c_{\rm out}}{c_{\rm in}-c_{\rm out}} \equiv \zeta_\alpha = \frac{2\sqrt{c}}{\alpha}$$

From heuristics arguments...

$$Ov = 2\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{\sigma_i,\tilde{\sigma}_i} - \frac{1}{2}\right)$$

Explicit expression of the overlap

$$\mathbb{E}[\mathrm{Ov}] \simeq \frac{1}{n}\sum_{i=1}^{n}\mathrm{erf}\left[\sqrt{\frac{\alpha^2 d_i}{8c-2\alpha^2}\left(\frac{c\Phi-\zeta_\alpha^2}{c\Phi-1}\right)}\right]$$

From heuristics arguments...

$$Ov = 2\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{\sigma_i,\tilde{\sigma}_i} - \frac{1}{2}\right)$$

Explicit expression of the overlap

$$\mathbb{E}[\mathrm{Ov}] \simeq \frac{1}{n}\sum_{i=1}^{n}\mathrm{erf}\left[\sqrt{\frac{\alpha^2 d_i}{8c-2\alpha^2}\left(\frac{c\Phi-\zeta_\alpha^2}{c\Phi-1}\right)}\right]$$

From heuristics arguments...

$$Ov = 2\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{\sigma_i,\tilde{\sigma}_i} - \frac{1}{2}\right)$$

Explicit expression of the overlap

$$\mathbb{E}[\mathrm{Ov}] \simeq \frac{1}{n}\sum_{i=1}^{n}\mathrm{erf}\left[\sqrt{\frac{\alpha^2 d_i}{8c - 2\alpha^2}\left(\frac{c\Phi - \zeta_\alpha^2}{c\Phi - 1}\right)}\right]$$

$$\lim_{c_{\mathrm{out}}\to 0} 8c - 2\alpha^2 = 0$$

From heuristics arguments...

$$Ov = 2\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{\sigma_i,\tilde{\sigma}_i} - \frac{1}{2}\right)$$

Explicit expression of the overlap

$$\mathbb{E}[\mathrm{Ov}] \simeq \frac{1}{n}\sum_{i=1}^{n}\mathrm{erf}\left[\sqrt{\frac{\alpha^2 d_i}{8c-2\alpha^2}\left(\frac{c\Phi-\zeta_\alpha^2}{c\Phi-1}\right)}\right]$$

$$\lim_{\alpha\to\alpha_c} c\Phi - \zeta_\alpha^2 = 0$$

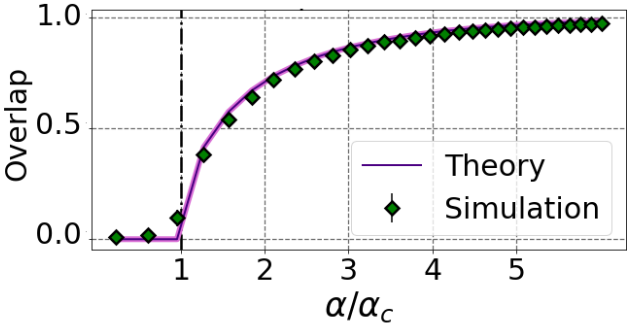


Figure: Simulated versus theoretical overlap for q_i distributed according to a power law. Average over 10 realizations. Both figures : The following parameters were used: $n = 5000$, $c_{\mathrm{out}} = 8$, $c_{\mathrm{in}} = 9 \to 61$.

How to estimate $r_{\rm opt} = \zeta_\alpha$?

From linearisation of BP

$$B\boldsymbol{w} = \zeta_\alpha \boldsymbol{w}$$

How to estimate $r_{\rm opt} = \zeta_\alpha$?

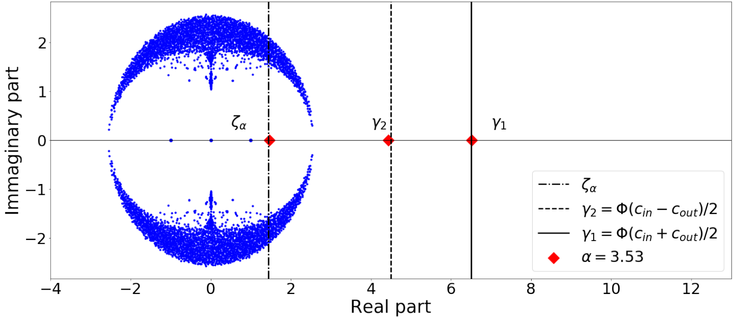
From linearisation of BP

$$\color{red}{B}\boldsymbol{w} = \zeta_\alpha \boldsymbol{w}$$

How to estimate $r_{\text{opt}} = \zeta_\alpha$?

From linearisation of BP

$$B\boldsymbol{w} = \zeta_\alpha \boldsymbol{w}$$

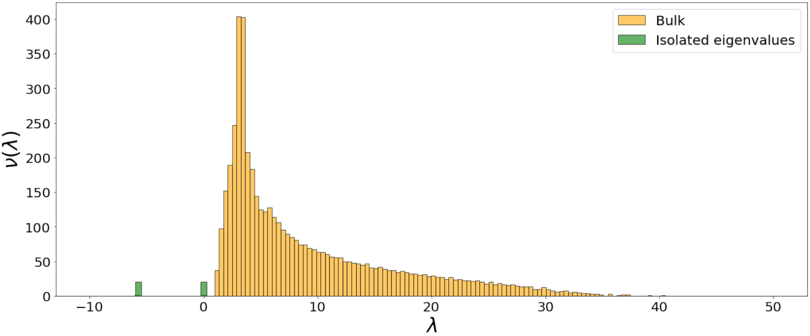


Property

$$Bw = \zeta_\alpha w \rightarrow det[H_{\zeta_\alpha}] = 0$$

Property

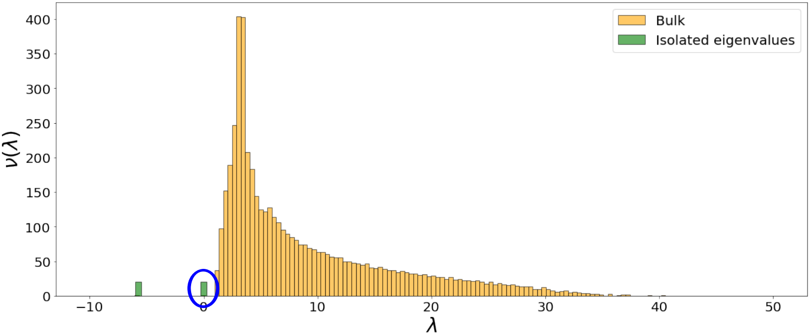
$$Bw = \zeta_\alpha w \rightarrow det[H_{\zeta_\alpha}] = 0$$



Spectrum of H_{ζ_α}

Property

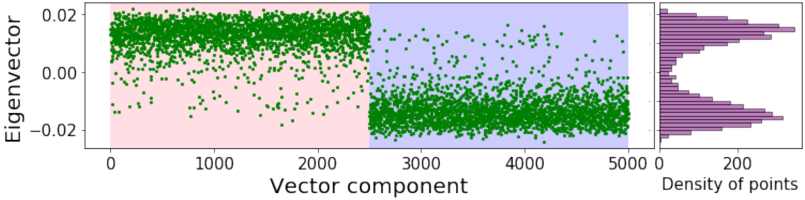
$$\boldsymbol{B}\mathbf{w} = \zeta_\alpha \mathbf{w} \rightarrow det[H_{\zeta_\alpha}] = 0$$



Spectrum of H_{ζ_α}

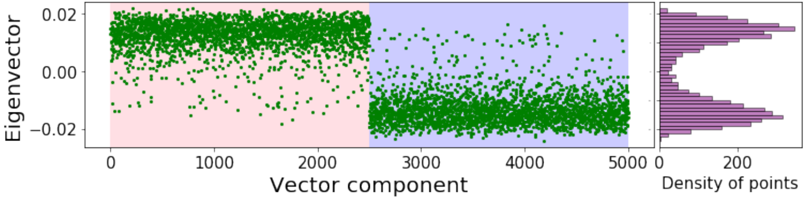
The eigenvector

Optimal value $r = \zeta_\alpha = \frac{c_{\rm in}+c_{\rm out}}{c_{\rm in}-c_{\rm out}}$

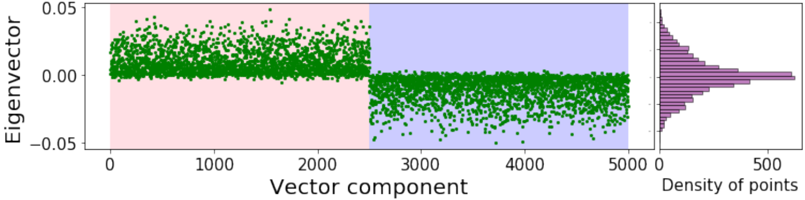


The eigenvector

Optimal value $r = \zeta_\alpha = \frac{c_{\rm in}+c_{\rm out}}{c_{\rm in}-c_{\rm out}}$



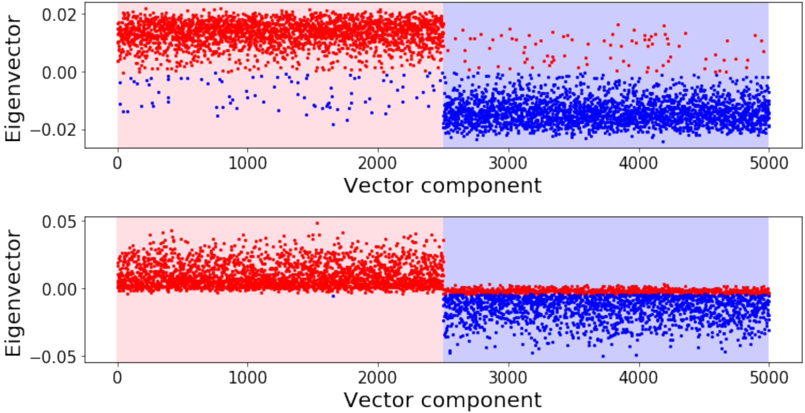
Initially proposed[4] value $r = \frac{\sum_i d_i^2}{\sum_i d_i}$



[4] Saade *et al.*, Spectral clustering of graphs with the Bethe-Hessian, NIPS 2014.

k - means

Optimal value $r = \zeta_\alpha = \frac{c_{\rm in}+c_{\rm out}}{c_{\rm in}-c_{\rm out}}$



The overlap

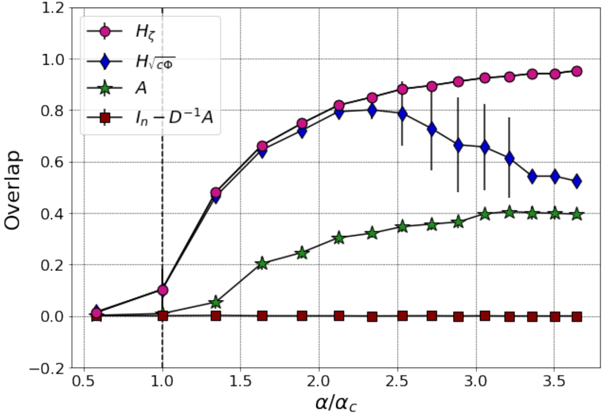


Figure: Overlap as a function of α/α_c for a power law degree distribution

More than two classes

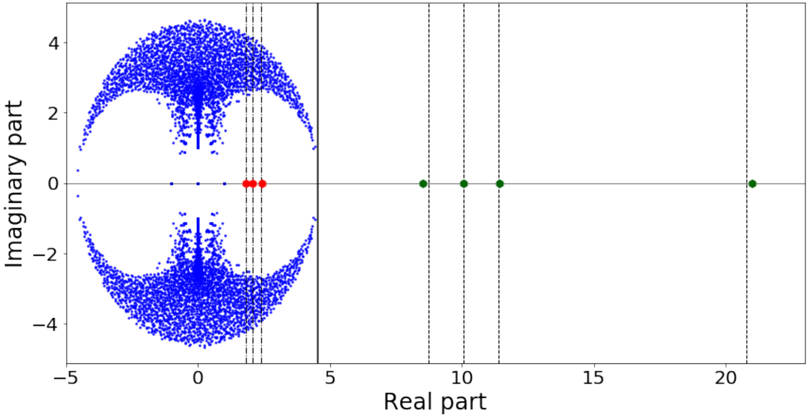


Figure: Spectrum of B for 4 classes

Future steps

- Theoretical support for our findings

Future steps

- Theoretical support for our findings
- Algorithmic optimization (eigencounts techniques)

Thank you!