Revisiting the Bethe-Hessian: Improved Community Detection in Sparse Heterogeneous Graphs

Lorenzo Dall'Amico, *lorenzo.dall-amico@gipsa-lab.fr*
Romain Couillet, *romain.couillet@gipsa-lab.fr*
Nicolas Tremblay, *nicolas.tremblay@gipsa-lab.fr*

Abstract

Problem position

- **Problem**: detect communities of unweighted and undirected graph $\mathcal{G}$
- **Technique**: statistical physics-inspired spectral clustering
- **Model constraints**: sparse and heterogenous networks

Contribution

- Spectral algorithm based on $H_r = (r^2-1)I_n + D - rA$, provably performing down to the detectability threshold. Caracterization of optimal r.
- Connection with main spectral algorithms based on B, $D_\tau^{-1}A$, $D-A$, $D^{-1}A$.

Model and notation (I)

Degree-corrected stochastic block model

- n nodes, k classes
- $\boldsymbol{\ell} \in \{1, \cdots, k\}^n$ label vector. $\boldsymbol{\pi}_p$ fraction of nodes with label p
- C: class affinity matrix, $\Pi = \mathrm{diag}(\boldsymbol{\pi})$. $C, \Pi \in \mathcal{M}_{k\times k}$.
- A: adjacency matrix; $D = \mathrm{diag}(A\mathbf{1})$: degree matrix. $A, D \in \mathcal{M}_{n\times n}$.
- $\boldsymbol{\theta}$: node connectivity predisposition. $\boldsymbol{\theta} \in \mathbb{R}^n$; $\frac{1}{n}\mathbf{1}^T\boldsymbol{\theta} = 1$; $\frac{1}{n}\mathbf{1}^T\boldsymbol{\theta}^2 = \Phi = O_n(1)$.
- $c = \frac{1}{n}\mathbf{1}^T A \mathbf{1} = O_n(1)$ average degree; $C\Pi\mathbf{1} = c\mathbf{1}$.

$$\mathbb{P}(A_{ij} = 1) = \theta_i\theta_j \frac{C_{\ell_i,\ell_j}}{n}$$

Detectability threshold

For $k=2$ communities of equal size: $C_{\ell_i,\ell_j} = c_{\text{in}}$ if $\ell_i = \ell_j$ and c_{out} otherwise, non trivial reconstruction iff

$$\alpha := \frac{c_{\text{in}} - c_{\text{out}}}{\sqrt{c}} > \frac{2}{\sqrt{\Phi}} := \alpha_c$$

- Gulikers *et al.* – An impossibility result for reconstruction in a degree-corrected planted-partition model – 2015

Resilience to degree heterogeneity (II)

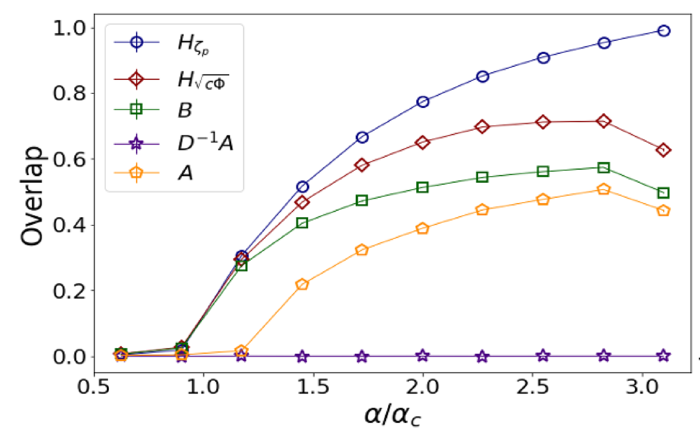
for large n, $\mathcal{G} \to$ GWT locally

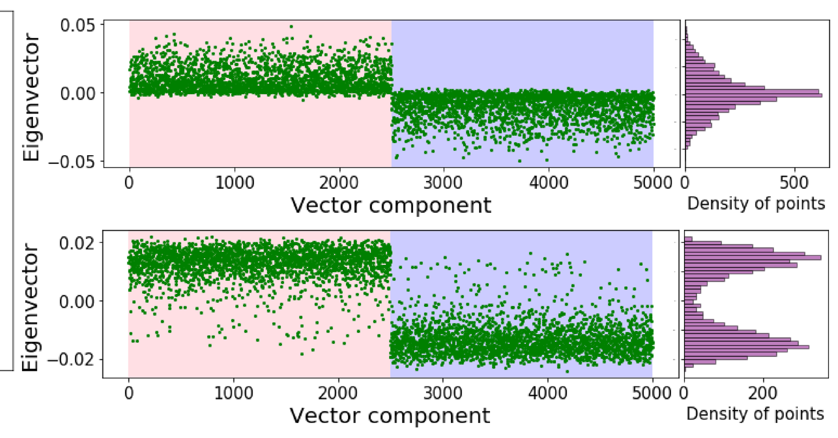
$$\mathbb{P}(\ell_j|\ell_i, A_{ij} = 1) = \frac{C_{\ell_i,\ell_j}}{c}$$
$$C\Pi\boldsymbol{v}_p = s_p(C\Pi)\boldsymbol{v}_p$$
$$\boldsymbol{u}_i = \boldsymbol{v}_{p,\ell_i}$$

$$\mathbb{E}[H_r\boldsymbol{u}_p|A] = \Big[(r^2-1)I_n + D\underbrace{\left(1 - r\frac{s_p(C\Pi)}{c}\right)}_{=0}\Big]\boldsymbol{u}_p$$

For $r = \zeta_p \equiv \frac{c}{s_p(C\Pi)}$ then $H_{\zeta_p}\boldsymbol{u}_p \approx \lambda_p\boldsymbol{u}_p$.

H_{ζ_p} has an informative eigenvector not spoiled by D.





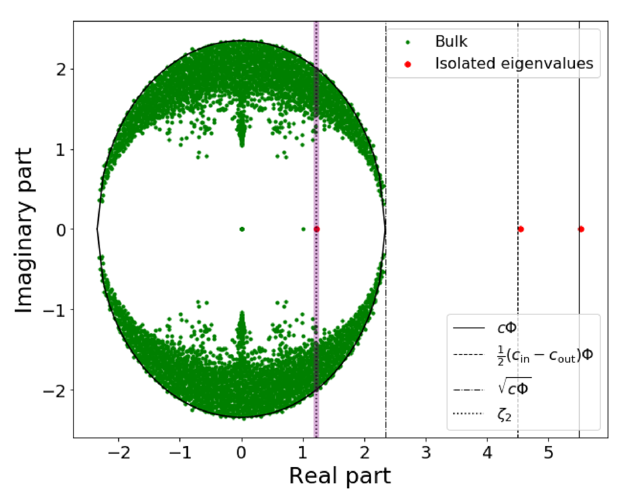
$$Ov = \max_{\hat{\ell}\in\mathcal{P}} \frac{1}{1-\frac{1}{k}}\left(\sum_{i=1}^n \delta(\hat{\ell}_i, \ell_i) - \frac{1}{k}\right)$$

we have theoretical prediction of Ov for $k=2$

For easy problems, $\zeta_p \to 1$ and $H_{\zeta_p} \to D - A$.

Linearization of belief propagation (III)

From the linearization of BP, $\boldsymbol{\delta_p}$ are informative



- Coste, Yizhe – Eigenvalues of the non-backtracking operator detached from the bulk – 2019

$$B_{(ij)(kl)} = \delta_{jk}(1-\delta_{il}), \forall\ (ij),(kl) \in \mathcal{E}^d$$
$$B\boldsymbol{\delta}_p = \zeta_p\boldsymbol{\delta}_p$$

$$s_p(B) = s_p(C\Pi)\Phi \quad \text{for } 1 \le p \le k$$
$$\zeta_p = \frac{s_1(B)}{s_p(B)}$$

Est 1

Ihara-Bass formula:

$$\boldsymbol{x}_p = \sum_{j\in\partial i}\delta_{p,ij}\ , \qquad H_{\zeta_p}\boldsymbol{x}_p = 0$$

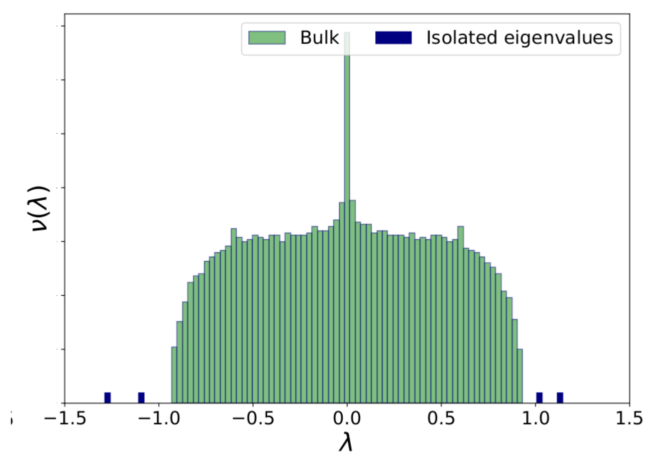
Est 2

Two ways to estimate ζ_p

Regularized Laplacian matrix (IV)

Proposition: For $\tau = \zeta_p^2 - 1$ and $n \gg 1$ then, with high probability, the p eigenvalues of $D_\tau^{-1}A$ with largest modulus are isolated.

Proposition: If the p largest eigenvalues of $D_\tau^{-1}A$ are isolated, then the p largest eigenvalues of $D_{\tau'}^{-1}A$ are also isolated for $\tau' > \tau > 0$.



$$H_{\zeta_p}\boldsymbol{x}_p = 0, \quad \text{then} \quad D_{\zeta_p^2-1}^{-1}A\boldsymbol{x}_p = \frac{1}{\zeta_p}\boldsymbol{x}_p$$
$$D_\tau = D + \tau I_n$$

For easy problems $D_{\zeta_p^2-1}^{-1}A \to D^{-1}A$

Algorithm (V)

Input: adjacency matrix of undirected graph $\mathcal{G}$

- Detect the number of classes: $\hat{k} \leftarrow \left|\left\{i : s_i\left((\rho(B)-1)I_n + D - \sqrt{\rho(B)}A\right) < 0\right\}\right|$
- For $1 \le p \le \hat{k}$:
  $\zeta_p \leftarrow s_{n-p+1}[(\zeta_p^2 - 1)I_n + D - \zeta_p A] = 0$
  $X_{:k} \leftarrow \boldsymbol{x}_p : [(\zeta_p^2-1)I_n + D - \zeta_p A]\boldsymbol{x}_p = 0$

Output: community labels obtained from k-means on the rows of X.

Research openings

- Embed node available information: graph semi-supervised learning
- Extension to dynamical models

Some real networks (VI)

Dataset	n	c	Φ	k	Est 1	Est 2	A	$H_{\sqrt{c\Phi}}$	B	L^{rw}	L_τ^{sym}
Polblogs	1222	27,4	3	2	**0.43**	**0.43**	0.23	0.27	0.24	0	**0.43**
Tv	3892	8.9	3	41	0.57	**0.8**	0.51	0.58	0.55	0.55	**0.8**
Facebook	4039	43.7	2.4	55	0.53	**0.78**	0.43	0.49	0.49	**0.78**	0.57
Power grid	4941	2.7	1.5	25	0.38	**0.93**	0.18	0.37	0.31	**0.93**	0.85
GrQc	5242	5.5	3.1	29	0.53	0.53	0.45	0.49	0.49	0.42	**0.79**
Politicians	5908	14.1	3	62	0.65	**0.85**	0.48	0.54	0.5	0.83	0.74
GNutella P2P	6301	6.6	2.7	5	0.22	**0.34**	0.15	0.15	0.20	0	**0.34**
Wikipedia	7115	28.3	5.1	22	0.22	0.23	0.15	0.17	0.17	0.23	**0.27**
Vip	11565	11.6	4.4	53	0.35	**0.62**	0.27	0.33	0.3	0.55	0.54
HepPh	12008	19.7	6.6	60	0.41	0.37	0.42	0.42	0.42	0.11	**0.52**