gipsa-lab

Community detection in sparse time-evolving graphs with a dynamical Bethe-Hessian

Lorenzo Dall'Amico, Romain Couillet, Nicolas Tremblay

Abstract

- **Problem**: community detection in temporal graphs $\{\mathcal{G}_t\}_{t=1,\dots,T}$
- **Goal**: exploit time label correlation to improve clustering perfomance
- **Technique**: physics inspired approach for spectral clustering in sparse graphs

Dynamical Bethe-Hessian matrix for fast and improved community detection down to the detectability threshold. Extension to arbitrary number of classes.

Model and notation (I)

Dynamical degree-corrected stochastic block model

- $\{\mathcal{G}_t(\mathcal{V}_t, \mathcal{E}_t)\}_{t=1,\dots,T}$ sequence of graphs with n nodes and k classes.
- $\eta \in [0,1]$ class label persistence
- $\boldsymbol{\ell}_t \in \{1, \cdots, k\}^n$ label vector at time t. π_p fraction of nodes with label p

$$\forall\, i_{t+1} \in \mathcal{V}_{t+1}, \qquad \ell_{i_{t+1}} = \begin{cases} \ell_{i_t} & \text{w.p. } \eta \\ a & \text{w.p. } (1-\eta)\pi_a \end{cases}$$

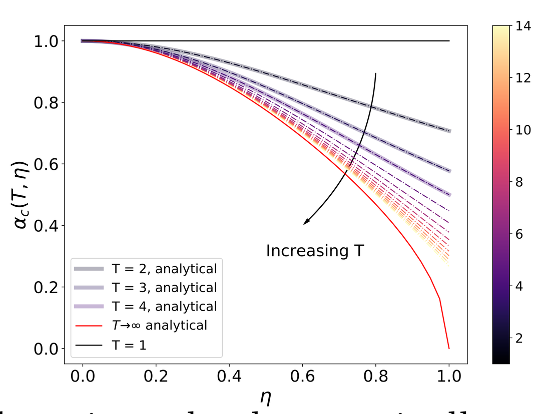
- C: class affinity matrix, $\Pi = \mathrm{diag}(\boldsymbol{\pi})$. $C, \Pi \in \mathcal{M}_{k\times k}$.
- $A^{(t)}$, $D^{(t)} = \mathrm{diag}(A^{(t)}\mathbf{1}) \in \mathcal{M}_{n\times n}$: adjacency and degree matrix of $\mathcal{G}_t$.
- $\boldsymbol{\theta}$: node connectivity. $\boldsymbol{\theta} \in \mathbb{R}^n$; $\frac{1}{n}\mathbf{1}^T\boldsymbol{\theta} = 1$; $\frac{1}{n}\mathbf{1}^T\boldsymbol{\theta}^2 = \Phi = O_n(1)$.
- $c = \frac{1}{n}\mathbf{1}^T A \mathbf{1} = O_n(1)$ average degree; $C\Pi\mathbf{1} = c\mathbf{1}$.

$$\mathbb{P}(A_{ij}^{(t)} = 1) = \theta_i\theta_j \frac{C_{\ell_{i_t}, \ell_{j_t}}}{n} \quad \text{i.i.d across time}$$

Dynamical detectability threshold (II)

For $k=2$ communities of equal size: $C_{\ell_i,\ell_j} = c_{\mathrm{in}}$ if $\ell_i = \ell_j$ and c_{out} otherwise, with $\lambda = \frac{c_{\mathrm{in}} - c_{\mathrm{out}}}{c_{\mathrm{in}} + c_{\mathrm{out}}}$.
Non trivial reconstruction if and only if [1]

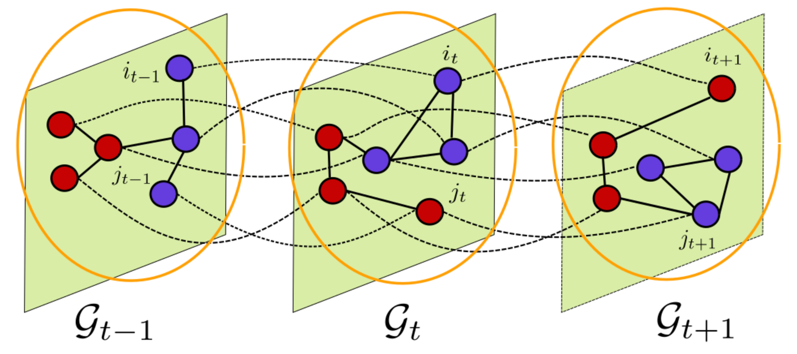
$$\alpha := \sqrt{c\Phi\lambda^2} > \alpha_c(T, \eta)$$



- Large T and η improve performance
- Analytical expression of $\alpha_c(T,\eta)$ up to $T=4$, otherwise solved numerically

[1] Ghasemian *et. al.* – *Detectability thresholds and optimal algorithms for community structure in dynamic networks* (2016)

Dynamical Bethe-Hessian matrix (III)



Dynamical Bethe-Hessian matrix

- Let $\xi, h \in (0,1)$
- Let $\phi_{t=1} = \phi_{t=T} = 1$ and $\phi_t = 2$, for all other t.

The dynamical Bethe-Hessian matrix $H_{\xi,h} \in \mathbb{R}^{nT\times nT}$ is defined as

$$(H_{\xi,h})_{i_t, j_{t'}} = \begin{cases} \left(\frac{\xi^2 D^{(t)} - \xi A^{(t)}}{1-\xi^2} + \frac{1 + h^2(\phi_t - 1)}{1-h^2} I_n\right)_{ij} & \text{if } t = t' \\ \left(-\frac{h}{1-h^2} I_n\right)_{ij} & \text{if } t = t' \pm 1, \end{cases}$$

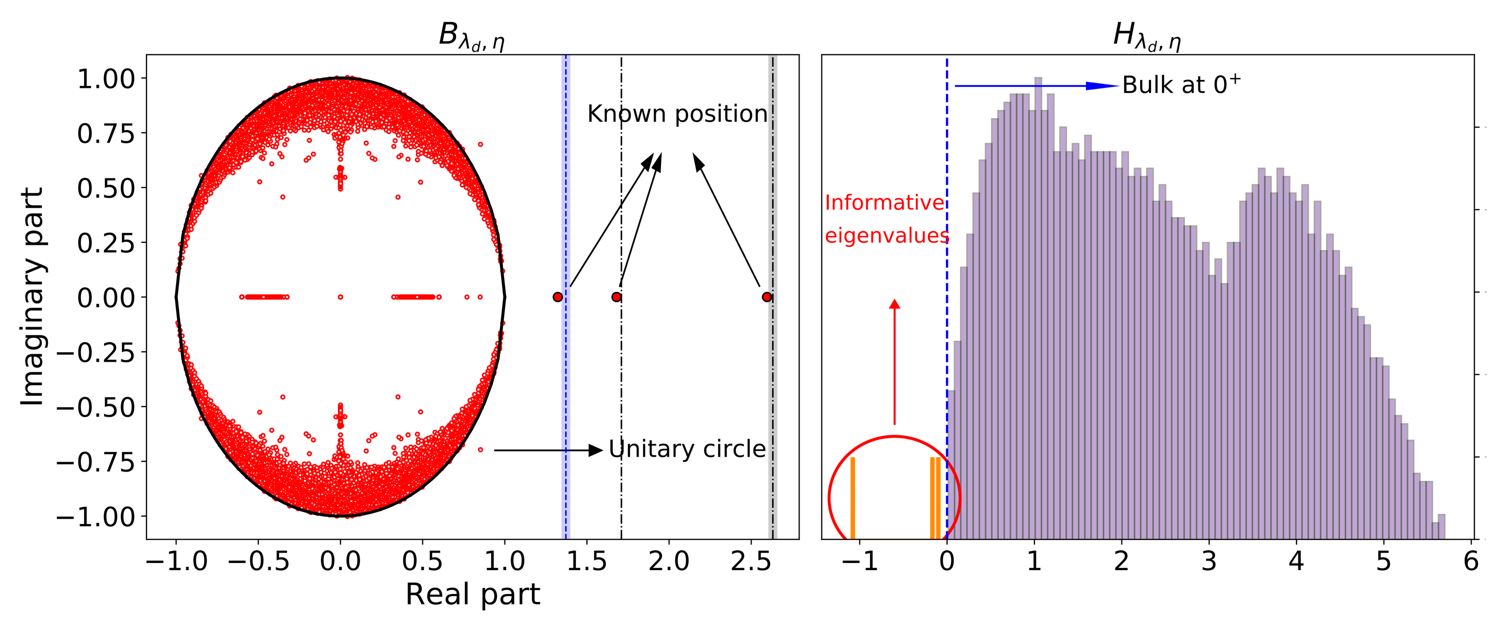
Dynamical non-backtracking matrix

$$(B_{\xi,h})_{(ij)(kl)} = \delta_{jk}(1 - \delta_{kl}) \times \begin{cases} \xi & \text{if } (kl) \text{ spatial edge} \\ h & \text{if } (kl) \text{ temporal edge} \end{cases}$$

Generalized Ihara Bass formula : if $\exists\, \boldsymbol{v}$ s.t. $B_{\xi,h}\boldsymbol{v} = \boldsymbol{v}$, then $\det[H_{\xi,h}] = 0$

Main result (IV)

Proposition: suppose the problem is **above the detectability threshold** $\boldsymbol{\alpha > \alpha_c(T, \eta)}$ and define $\boldsymbol{\xi = \lambda_d = \frac{\alpha_c(T,\eta)}{\sqrt{c\Phi}}}$. Then, as $n \to \infty$



Algorithm (V)

Input : adjacency matrices $\{A^{(t)}\}_{t=1,\dots,T}$ of the undirected graphs $\{\mathcal{G}_t\}_{t=1,\dots,T}$; label persistence, η; number of clusters k.

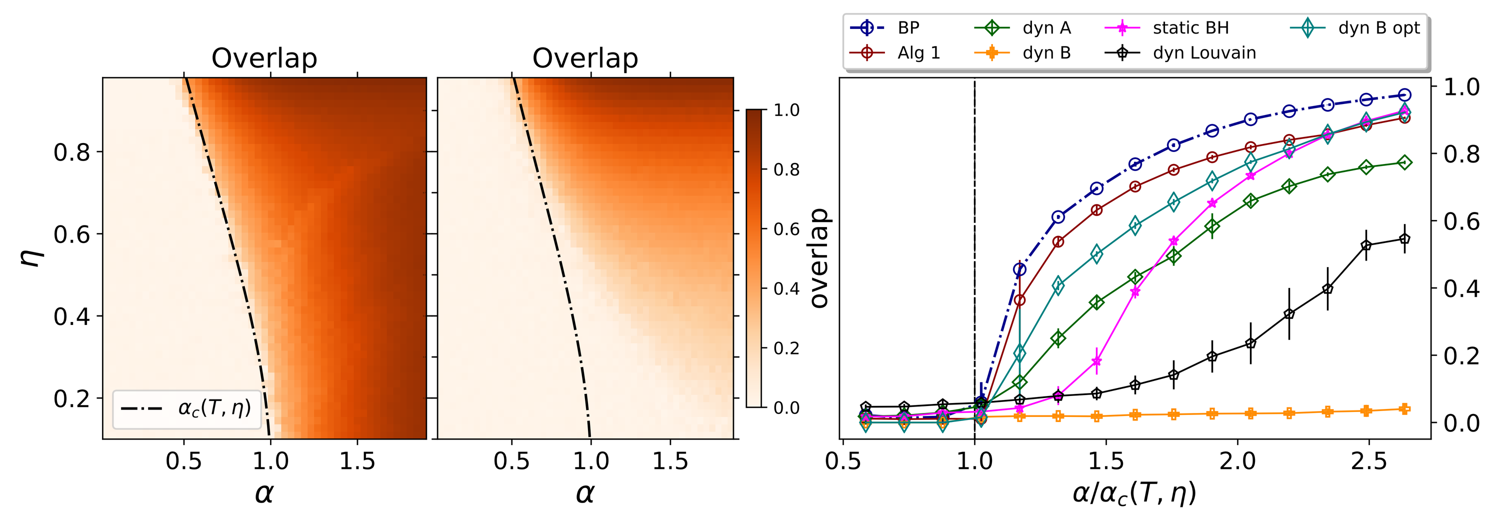
- Compute λ_d and build $H_{\lambda_d,\eta}$
- Stack the m eigenvectors of $H_{\lambda_d,\eta}$ with eigenvalue < 0 in of $X \in \mathbb{R}^{nT\times m}$
- Normalize the rows of $X_{i,:} \leftarrow X_{i,:}/\|X_{i,:}\|$
- For each t, estimate community labels using k-*means* on the rows $\{X_{i_t}\}_{i=1,\dots,n}$.

Return Estimated label vector $\hat{\boldsymbol{\ell}} \in \{1, \dots, k\}^{nT}$.

Efficient Julia implementation at lorenzodallamico.github.io/CoDeBetHe.jl

Performance (VI)

$$\mathrm{ov}(\boldsymbol{\ell}, \hat{\boldsymbol{\ell}}) = \max_{\bar{\boldsymbol{\ell}} \in \mathcal{P}(\hat{\boldsymbol{\ell}})} \frac{1}{1 - \frac{1}{k}}\left(\frac{1}{n}\sum_{i=1}^n \delta_{\ell_i, \bar{\ell}_i} - \frac{1}{k}\right),$$



[1] Ghasemian *et. al.* – *Detectability thresholds and optimal algorithms for community structure in dynamic networks* (2016)
[2] Mucha *at. al.* – *Community structure in time-dependent, multiscale, and multiplex networks* (2010)
[3] Keriven, Vaiter – *Sparse and smooth: improved guarantees for spectral clustering in the dynamic stochastic block model* (2020)
[4] Dall'Amico *et. al.* – *A unified framework for spectral clustering in sparse graphs* (2020)

Research openings

- How to estimate η for all $\alpha > \alpha_c(T,\eta)$?
- Generalization to the case in which $C, \Pi, \boldsymbol{\theta}$ also evolve through time
- Generalization to the case of $A^{(t)}$ with additional informative edge dependences.
- Preliminary investigation in the supplementary material.

Grenoble Images Parole Signal Automatique
UMR CNRS 5216 - Grenoble Campus
38400 Saint Martin d'Hères - FRANCE

cnrs · Grenoble INP · Université Grenoble Alpes