Computation of the Laplacian Spectral Barycentre Network in a Soules Basis

François G. Meyer

Applied Mathematics
University of Colorado at Boulder

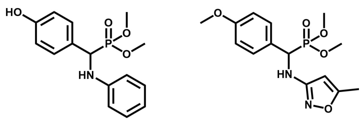
20th Workshop on Modelling and Mining Networks, WAW 2025

Supported by the National Science Foundation: CCF/CIF 1815971

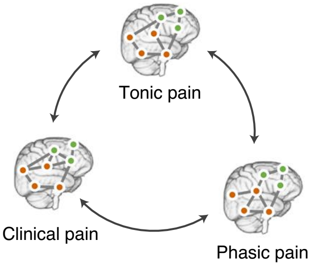
Introduction

Scientific questions:

- de novo design of molecules: aminophosphonate derivatives [1]



- biomarker for clinical pain [2]: comparison of pain networks (measured by fMRI);



Introduction: the barycentre of a set of networks (2)

- dataset of graphs, $\{G^{(1)}, \dots, G^{(T)}\}$; adjacency matrix of $G^{(t)}$: $\boldsymbol{A}^{(t)}$;

☞ Goal: find $\widehat{\mu}_T[\mathbb{P}]$ = graph that summarizes the topology and connectivity of $\{G^{(t)}\}$

Mathematical framework:

- $\mathcal{S}$ is the set of $n \times n$ symmetric adjacency matrices with nonnegative weights,
- the $n \times n$ adjacency matrix $\boldsymbol{A}^{(t)}$ is sampled from the probability space $(\mathcal{S}, \mathbb{P})$;
- we equip the probability space $(\mathcal{S}, \mathbb{P})$ with a metric d;
- *barycentre* [3], or *Fréchet mean* [4], graph, $\widehat{\mu}_T[\mathbb{P}]$,

$$\widehat{\mu}_T[\mathbb{P}] \stackrel{\text{def}}{=} \underset{\mathbf{B} \in \mathcal{S}}{\operatorname{argmin}} \sum_{t=1}^{T} d^2(\mathbf{B}, \boldsymbol{A}^{(t)}). \quad (1)$$

- choice of the distance d influences the topology/connectivity of $\widehat{\mu}_T[\mathbb{P}]$;

"!": the distance between graphs should be evaluated in the spectral domain.

Notations

- $[n] \stackrel{\text{def}}{=} \{1, \ldots, n\}$;
- $\mathbf{1} \stackrel{\text{def}}{=} [1 \cdots 1]^T$, and $\mathbf{J} = \mathbf{1}\mathbf{1}^T$;
- $O(n)$ is the orthogonal group;
- $G = (V, E)$ is an undirected unweighted graph;
- the adjacency matrix of G is denoted by $\mathbf{A}$; the degree matrix is denoted by $\mathbf{D}$;
- the symmetric normalized adjacency matrix, $\widehat{\mathbf{A}} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$, is defined by

$$\hat{a}_{ij} \stackrel{\text{def}}{=} a_{ij}/\sqrt{d_i d_j} \text{ if } d_i d_j \neq 0; \text{ and } \hat{a}_{ij} \stackrel{\text{def}}{=} 0 \text{ otherwise;} \tag{2}$$

- the normalized Laplacian is defined by $\mathcal{L} \stackrel{\text{def}}{=} \mathrm{Id} - \widehat{\mathbf{A}}$;
- the ascending sequence of eigenvalues $0 = \lambda_1 \leqslant \cdots \leqslant \lambda_n \leqslant 2$ of $\mathcal{L}$ is denoted by

$$\boldsymbol{\lambda}(\mathcal{L}) = \begin{bmatrix} \lambda_1 & \cdots & \lambda_n \end{bmatrix}. \tag{3}$$

The Laplacian spectral pseudo-metric

- we define the Laplacian spectral pseudo-metric as

$$\mathrm{d}(\mathcal{L}, \mathcal{L}') \stackrel{\text{def}}{=} \left\| \boldsymbol{\lambda}(\mathcal{L}) - \boldsymbol{\lambda}(\mathcal{L}') \right\|_2 \tag{4}$$

  where $\boldsymbol{\lambda}(\mathcal{L})$ and $\boldsymbol{\lambda}(\mathcal{L}')$ are the vectors of eigenvalues of $\mathcal{L}$ and $\mathcal{L}'$ respectively.

- $\mathrm{d}(\mathcal{L}, \mathcal{L}')$: differences – at multiple scales – in topology & connectivity [5, 6].
- no need to solve the node correspondence problem; possible to compare graphs of different sizes ($\mathcal{L}$ is the normalized Laplacian)

Goal: find

$$\widehat{\boldsymbol{\mu}}_T[\mathbb{P}] \stackrel{\text{def}}{=} \underset{\mathbf{B} \in \mathcal{S}}{\operatorname{argmin}} \sum_{t=1}^{T} \left\| \boldsymbol{\lambda}(\mathcal{L}(\mathbf{A}^{(t)})) - \boldsymbol{\lambda}(\mathcal{L}(\mathbf{B})) \right\|_2^2 \tag{5}$$

Technical difficulties :

1. $\left\| \boldsymbol{\lambda}\big(\mathcal{L}(\mathbf{A}^{(t)})\big) - \boldsymbol{\lambda}(\mathcal{L}(\mathbf{B})) \right\|_2$ is defined in the spectral domain ...
2. ... but the optimization (5) takes place in $\mathcal{S}$

From the spectrum to the Laplacian (1)

Solution to the technical difficulties

1. we say that $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_n]$ is *realizable* if

$$\exists \mathbf{A} \in \mathcal{S} \text{ whose normalized Laplacian, } \boldsymbol{\mathcal{L}}(\mathbf{A}), \text{ satisfies } \boldsymbol{\lambda}(\boldsymbol{\mathcal{L}}(\mathbf{A})) = \boldsymbol{\lambda}. \tag{6}$$

2. the *set of realizable sequences* is denoted by $\mathcal{R}$.

We seek $\widehat{\boldsymbol{\mu}}_T[\mathbb{P}]$ such that

$$\boldsymbol{\lambda}\big(\widehat{\boldsymbol{\mu}}_T[\mathbb{P}]\big) = \underset{\boldsymbol{\lambda} \in \mathcal{R}}{\operatorname{argmin}} \sum_{t=1}^{T} \|\boldsymbol{\lambda} - \boldsymbol{\lambda}(\boldsymbol{\mathcal{L}}^{(t)})\|_2^2. \tag{7}$$

If we relax this minimization problem ($\boldsymbol{\lambda} \in \mathbb{R}^n$ instead of $\boldsymbol{\lambda} \in \mathcal{R}$), then

1. the solution to (7) is the sample mean $\widehat{\mathbb{E}}_T[\boldsymbol{\lambda}] \stackrel{\text{def}}{=} T^{-1} \sum_{t=1}^{T} \boldsymbol{\lambda}(\boldsymbol{\mathcal{L}}(\mathbf{A}^{(t)}))$;
2. ... but $\widehat{\mathbb{E}}_T[\boldsymbol{\lambda}]$ has no guarantee to be realizable.

From the spectrum to the Laplacian (2)

... Yet more technical difficulties :

1. the knowledge of $\boldsymbol{\lambda}(\widehat{\mu}_T[\mathbb{P}])$ is insufficient to reconstruct a barycentre graph;
2. we need a basis of eigenvectors $\boldsymbol{\Psi} \in O(n)$ of a valid normalized Laplacian,

$$\exists \mathbf{A} \in \mathcal{S}, \quad \boldsymbol{\Psi} \operatorname{diag}\left(\widehat{\mathbb{E}}_T[\boldsymbol{\lambda}]\right) \boldsymbol{\Psi}^T = \mathrm{Id} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}, \tag{8}$$

   where $\mathbf{D}$ is the degree matrix associated to $\mathbf{A}$;
3. if $\boldsymbol{\Psi}$ satisfies (8), then we can define $\widehat{\mu}_T[\mathbb{P}]$ by

$$\widehat{\mu}_T[\mathbb{P}] \stackrel{\text{def}}{=} \mathbf{D}^{1/2}\left[\mathrm{Id} - \boldsymbol{\Psi} \operatorname{diag}\left(\widehat{\mathbb{E}}_T[\boldsymbol{\lambda}]\right) \boldsymbol{\Psi}^T\right]\mathbf{D}^{1/2}. \tag{9}$$

... Additional difficulties :

1. if $\mathbb{E}[\mathbb{P}]$ contains modular communities, rich clubs, hubs, trees, etc. then the graphs $\mathbf{A}^{(1)}, \dots, \mathbf{A}^{(T)}$ will share such topological structures;
2. we would like $\widehat{\mu}_T[\mathbb{P}]$ to also inherit such structures;

... Additional difficulties :

- given a random choice of $\boldsymbol{\Psi} \in \mathrm{O}(n)$ that satisfies (8), then $\widehat{\boldsymbol{\mu}}_T[\mathbb{P}]$ in (9), may have a very different topological structure than that of $\mathbb{E}[\mathbb{P}]$.

Informally, we need to impose that

$$\widehat{\boldsymbol{\mu}}_T[\mathbb{P}] \approx \mathbb{E}[\mathbb{P}]. \tag{10}$$

Remark: the trivial choice $\widehat{\boldsymbol{\mu}}_T[\mathbb{P}] = \widehat{\mathbb{E}}_T[\mathbb{P}]$ does not meet the constraint (7), since we have $\boldsymbol{\lambda}(\widehat{\mathbb{E}}_T[\mathbb{P}]) \neq \widehat{\mathbb{E}}_T[\boldsymbol{\lambda}]$ [7, 8].

Example of solution to (10): $\boldsymbol{\Psi}$ is an "average on $\mathrm{O}(n)$" of the bases of eigenvectors associated with the respective $\{\boldsymbol{\mathcal{L}}^{(1)}, \ldots, \boldsymbol{\mathcal{L}}^{(T)}\}$ of the graphs in the sample [9, 10]

From the spectrum to the Laplacian (4)

In summary, given $\widehat{\mathbb{E}}_T[\boldsymbol{\lambda}]$ (measured from the data) we seek $\widehat{\mu}_T[\mathbb{P}] \in \mathcal{S}$ such that,

$$\begin{cases} \mathcal{L}\left(\widehat{\mu}_T[\mathbb{P}]\right) = \boldsymbol{\Psi} \operatorname{diag}\left(\widehat{\mathbb{E}}_T[\boldsymbol{\lambda}]\right) \boldsymbol{\Psi}^T; \\ \boldsymbol{\Psi} \in O(n); \\ \widehat{\mu}_T[\mathbb{P}] \approx \mathbb{E}[\mathbb{P}]. \end{cases} \tag{11}$$

Original contributions

- we prove that it is possible to solve (11) using a "customized" *Soules* basis $\boldsymbol{\Psi}$;
- when $(\mathcal{S}, \mathbb{P})$ is the probability space associated with a balanced stochastic block model, we prove that $\widehat{\mu}_T[\mathbb{P}] = \mathbb{E}[\mathbb{P}]$.
- experiments on real-life graphs demonstrate that our approach works beyond the controlled environment of balanced stochastic block models;
- our theoretical analysis could probably be extended to a larger class of community networks.
- Theoretical details: arXiv:2502.00038 (2025), https://arxiv.org/abs/2502.00038

The setting: the stochastic block model

The stochastic block model

- we derive theoretical guarantees for our algorithms when the graphs are sampled from $(\mathcal{S}, \mathbb{P})$ = stochastic block model (e.g., [11]).
- quintessential exemplar of a network with community structure [12–14].
- universal approximants (under various norms or distances) [14–18]
  → building blocks to analyse more complex networks;
- a discrete version of step graphons [19–22], which are dense in the space of graphons for the topology induced by the cut-norm;
- amenable to a rigorous mathematical analysis;
- cutting edge of rigorous probabilistic analysis of random networks [23].

The stochastic block model $\mathrm{SBM}\left(\mathbf{p}, q, n\right)$

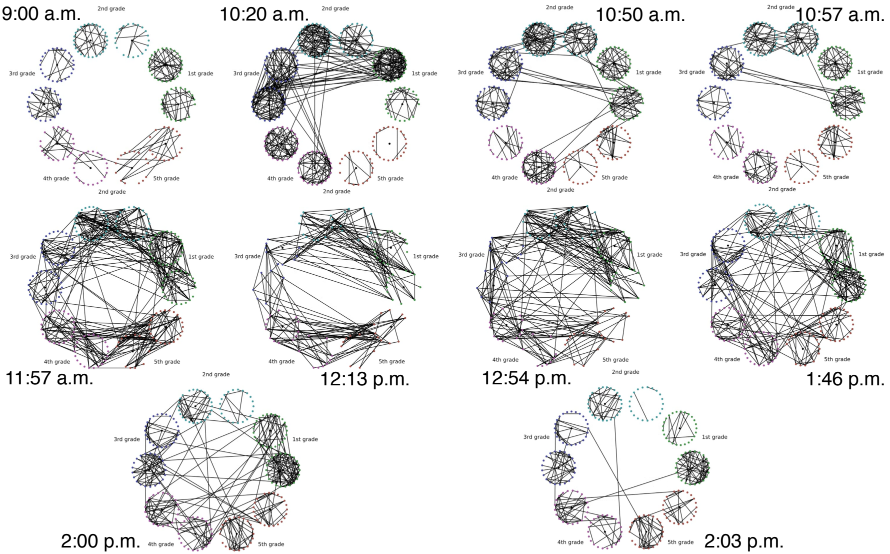
- Let $\{B_k\}, 1 \leqslant k \leqslant M$ be a partition of the vertex set $[n]$ into M contiguous blocks;
- $\mathbf{p} = [p_1, \cdots, p_M]$ is the vector of edge probabilities within each block;
- q is the edge probability between blocks;
- $\boldsymbol{A} \sim \mathrm{SBM}\left(\mathbf{p}, q, n\right)$ if
    1. $a_{ij} = a_{ji}, i < j$ are independent (up to symmetry);
    2. $a_{ij} \sim \mathrm{Bernoulli}(p_m)$ if $(i, j) \in B_m \times B_m$;
    3. $a_{ij} \sim \mathrm{Bernoulli}(q)$ if $(i, j) \in B_m \times B_{m'}$, $m \neq m'$;
    4. matrix of edge probabilities $\mathbf{P} \stackrel{\text{def}}{=} \mathbb{E}\left[\mathbb{P}\right]$;
    5. the $\mathrm{SBM}\left(p, q, n\right)$ is *balanced* if $|B_m| = n/M$, and $p_1 = \cdots = p_M$.

Intermezzo: a day in a French primary school

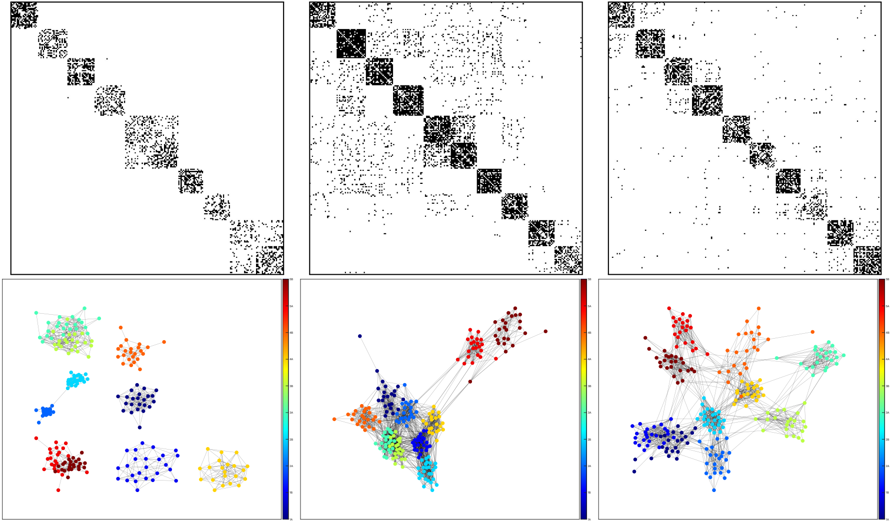
a day in a French primary school

- dynamic social-contact graphs collected in a French primary school [24–34]).
- students carried RFID tags that recorded (every 20 seconds) face-to-face contacts during two school days [24]
- primary school = five grades; each grade is divided into two classes (A & B);
- each student $(n = 232)$ is a node of the network.
- school day: 8:30 AM – 4:30 PM; changes in connectivity and topology:
  - 10:30 – 11:00 AM: morning recess;
  - 3:30 – 4:00 PM: afternoon recess;
  - two lunch periods: 12:00 PM– 1:00 PM, and 1:00 – 2:00 PM.
- divide the school day into morning and afternoon periods;
- morning period: $T = 35$ time intervals of $\approx$ 6 minutes;
- afternoon period: $T = 26$ time intervals of $\approx$ 6 minutes.
- for each time sample t, we construct an undirected unweighted graph $G^{(t)}$ by aggregating face-to-face contact events;

the contact network during significant connectivity and topological changes



the French primary school is a sequence of stochastic block models

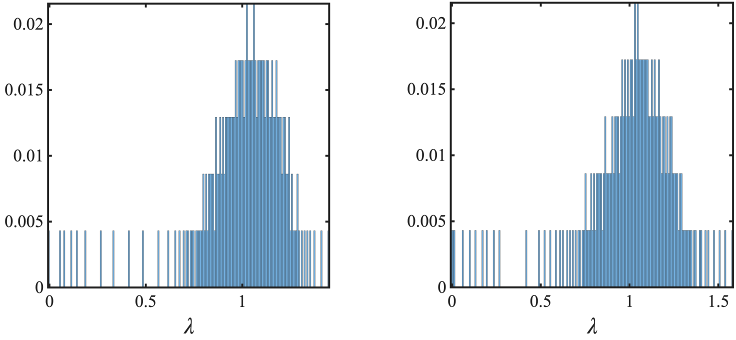


9:00 AM
beginning of the day

10:30 AM
morning recess

12:00 PM
end of morning period

Distribution of the eigenvalues of $\mathcal{L}(\mathbf{A}^{(t)})$



morning afternoon

- stochastic nature of the network $\longrightarrow$ bump-shaped *bulk* centered around 1;
- $0 = \lambda_1 \leqslant \cdots \leqslant \lambda_{10}$ are separated from the bulk;
- each of these 10 eigenvalues is associated with a specific community;

$\rightarrow$ signature of the stochastic block model [7, 8, 35–37].

The Soules bases

Soules' bases: formal definition

- Soules basis: orthogonal matrix that is constructed iteratively;
- at level 1, we choose $\psi_1 \stackrel{\text{def}}{=} n^{-1/2}\mathbf{1}$.
- at level l: apply a Givens rotations to ψ_l:

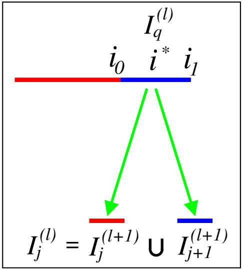
    ① the set $[n]$ is partitioned into l ordered intervals $I_j^{(l)}, 1 \leqslant q \leqslant l$.

    ② level $l \rightarrow$ level $l+1$: select an interval, $I_j^{(l)} = [i_0, i_1)$, and $i^* \in [i_0, i_1]$;

    ③ $I_j^{(l+1)} \stackrel{\text{def}}{=} [i_0, i^*]$, and $I_{j+1}^{(l+1)} \stackrel{\text{def}}{=} [i^*+1, i_1]$

    ④

$$\psi_{l+1}(i) \stackrel{\text{def}}{=} \frac{1}{\left\|\psi_l(i_0:i_1)\right\|} \begin{cases} \dfrac{\left\|\psi_l(i^*+1:i_1)\right\|}{\left\|\psi_l(i_0:i^*)\right\|}\psi_l(i) & \text{if} \quad i_0 \leqslant i \leqslant i^* \\ -\dfrac{\left\|\psi_l(i_0:i^*)\right\|}{\left\|\psi_l(i^*+1:i_1)\right\|}\psi_l(i) & \text{if} \quad i^*+1 \leqslant i \leqslant i_1, \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

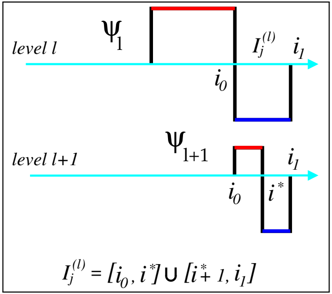


1. a node in the Soules binary tree is triggered by the splitting of $[i_0, i_1] = [i_0, i^*] \cup [i^* + 1, i_1]$.

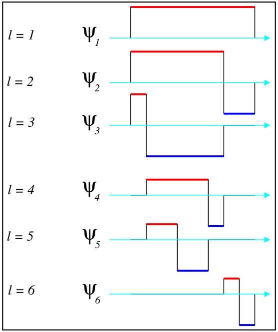
2. ψ_{l+1} is created by splitting $I_j^{(l)} = [i_0, i_1]$
$I_j^{(l)} = [i_0, i_1] = [i_0, i^*] \cup [i^* + 1, i_1]$;

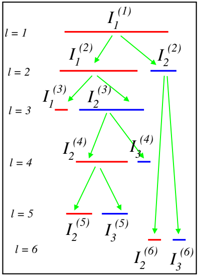
3. ψ_m and $\psi_{m'}$, $m \neq m'$, are either nested ... or they do not overlap;

☞ $\langle \psi_m, \psi_{m'} \rangle = 0$,



1. select $I_j^{(l)}$ such that $\forall 1 \leqslant k \leqslant l,\ \psi_k|_{I_j^{(l)}}$ is constant
2. split $I_j^{(l)}$ at node i^*: create ψ_{l+1}
3. $\psi_{l+1}|_{[i_0,i^*]} > 0,\ \psi_{l+1}|_{[i^*+1,i_1]} < 0$
4. a node in the binary tree is created by ψ_{l+1}





5. each Soules basis is associated with a binary tree;
6. the leaves are the intervals that are not split.

What is the Talk Really About?

Informal description of the results and line of attack

Idea 1.

For a balanced $\mathrm{SBM}\,(p,q,n)$ composed of M blocks,

$$\mathbb{E}[\mathcal{L}]_{ij} = \frac{M}{n(p+(M-1)q)} \begin{cases} -p & \text{if } \exists\, m \in [M],\ (i,j) \in B_m \times B_m, \\ 1 & \text{if } i = j, \\ -q & \text{otherwise.} \end{cases} \tag{13}$$

→ $\mathbb{E}[\mathcal{L}]$ for $\mathrm{SBM}\,(p,q,n)$ is constant over blocks $B_m \times B_m$;

☞ $\boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{\psi}_1 & \cdots & \boldsymbol{\psi}_n \end{bmatrix}$ solution to (11) should be designed such that

$$\mathcal{L}\big(\widehat{\boldsymbol{\mu}}_T[\mathbb{P}]\big) = \sum_{k=1}^{n} \widehat{\mathbb{E}}_T\,[\lambda_k]\,\boldsymbol{\psi}_k \boldsymbol{\psi}_k^T \tag{14}$$

is piecewise constant over the blocks $B_m \times B_m, 1 \leqslant m \leqslant M$.

Idea 2.

1. $\lambda_k(\mathcal{L})$ for a balanced $\mathrm{SBM}\,(p, q, n)$ composed of M blocks are given by [38],

$$\lambda_k(\mathcal{L}) = \begin{cases} 0 & \text{if } k = 1, \\ \dfrac{Mq}{p + (M-1)q} & \text{if } k = 2, \dots, M, \\ 1 & \text{if } k = M+1, \dots, n, \end{cases} \tag{15}$$

with probability converging to 1 as the graph size $n \to \infty$;

2. $\widehat{\mathbb{E}}_T[\lambda_j]$ converges for large n to the estimate above; we substitute $\widehat{\mathbb{E}}_T[\lambda_k]$ for the (large graph size n) estimates (15) in the eigendecomposition of $\mathcal{L}(\widehat{\mu}_T[\mathbb{P}])$ (14).

☛ Our goal: find $\widehat{\mu}_T[\mathbb{P}] \in \mathcal{S}$ such that

$$\begin{cases} \mathcal{L}(\widehat{\mu}_T[\mathbb{P}]) = \displaystyle\sum_{k=1}^{n} \psi_k \psi_k^T - \left\{ \frac{p-q}{p+(M-1)q} \Big(\sum_{j=1}^{M} \psi_j \psi_j^T \Big) + \frac{Mq}{p+(M-1)q} \psi_1 \psi_1^T \right\} \\ \begin{bmatrix} \psi_1 & \cdots & \psi_n \end{bmatrix} \in O(n); \\ \widehat{\mu}_T[\mathbb{P}] \approx \mathbb{E}[\mathbb{P}]. \end{cases} \tag{16}$$

Idea 2.

3. the comparison of the topology of $\widehat{\mu}_T[\mathbb{P}]$ with that of $\mathbb{E}[\mathbb{P}]$ for $\text{SBM}\,(\mathbf{p}, q, n)$,

$$\widehat{\mu}_T[\mathbb{P}] \approx \mathbb{E}[\mathbb{P}], \tag{17}$$

can be replaced by the equivalent condition,

$$\mathcal{L}\left(\widehat{\mu}_T[\mathbb{P}]\right) = \mathbb{E}[\mathcal{L}], \tag{18}$$

where $\mathbb{E}[\mathcal{L}]$ is given by (13). We combine (18) with (13), and (16) to get the program

☛ Our goal: find $\boldsymbol{\Psi} = \begin{bmatrix}\boldsymbol{\psi}_1 & \cdots & \boldsymbol{\psi}_n\end{bmatrix} \in O(n)$ such that

$$\begin{cases} \sum_{k=1}^{n} \boldsymbol{\psi}_k \boldsymbol{\psi}_k^T = \mathrm{Id}\,. \\ \boldsymbol{\psi}_1 = n^{-1/2}\mathbf{1}, \\ \sum_{k=1}^{M} \boldsymbol{\psi}_k \boldsymbol{\psi}_k^T(i,j) = \begin{cases} M/n & \text{if } \exists\, m \in [M],\ (i,j) \in B_m \times B_m, \\ 0 & \text{otherwise,} \end{cases} \end{cases} \tag{19}$$

Idea 3.

→ design an algorithm that explores the library of Soules bases [39], and returns $\boldsymbol{\Psi} = \begin{bmatrix} \psi_1 & \cdots & \psi_n \end{bmatrix}$, such that

$$\begin{cases} \psi_1 = n^{-1/2}\mathbf{1}, \\ \displaystyle\sum_{k=1}^{n} \psi_k \psi_k^T = \mathrm{Id}, \\ \displaystyle\sum_{k=1}^{M} \psi_k \psi_k^T(i,j) = \begin{cases} M/n & \text{if } \exists\, m \in [M],\ (i,j) \in B_m \times B_m, \\ 0 & \text{otherwise,} \end{cases} \end{cases} \tag{20}$$

Remarks

- $\psi_1 = n^{-1/2}\mathbf{1}$ is very standard for the construction of Soules bases;
  $\rightarrow$ each ψ_k is piecewise constant over $[n]$;
- the condition $\sum_{k=1}^{n} \psi_k \psi_k^T = \mathrm{Id}$ comes for free with Soules bases [39];
- the zero-crossing of $\psi_k \psi_k^T$ is aligned with the jumps between the blocks in $\mathbb{E}[\mathbb{P}]$;

Idea 3: some details

1. coarse scale: $\psi_1 = n^{-1/2}\mathbf{1}$ whose support is $[n]$;
2. finer scale: ψ_2 detect the largest gradient between any pair of blocks $B_m \times B_m$ and $B_{m'} \times B_{m'}$;

   → align the zero-crossing of $\psi_2\psi_2^T$ with the boundaries between two blocks of $\widehat{\mathbb{E}}_T[\mathbb{P}]$ with the largest jump in $\{p_1, \dots, p_M\}$;

   → ψ_2 maximizes $|\langle\psi_2\psi_2^T, \widehat{\mathbb{E}}_T[\mathbb{P}]\rangle|^2$;
3. next scales ($\psi_k, k \geqslant 3$): proceeds iteratively by detecting all the boundaries between the remaining blocks B_k;
4. theoretical analysis: we substitute the population mean $\mathbb{E}\left[\mathbb{P}\right]$ for the sample mean $\widehat{\mathbb{E}}_T[\mathbb{P}]$ (see details: `https://arxiv.org/abs/2502.00038`).

Algorithm 1: A greedy exploration of the Soules library

1. compute $\widehat{\mathbb{E}}_T[\mathbb{P}] \stackrel{\text{def}}{=} T^{-1}\sum_{t=1}^{T} \mathbf{A}^{(t)}$

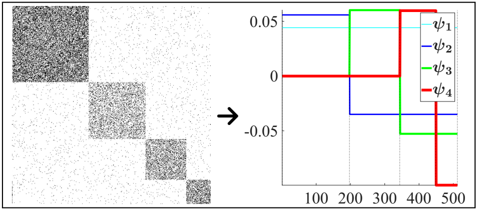
2. set $\boldsymbol{\psi}_1 = n^{-1/2}\mathbf{1}$; find $\boldsymbol{\psi}_2 = \underset{\boldsymbol{\psi}_2 \text{ defined by } (12)}{\operatorname{argmax}} |\langle \boldsymbol{\psi}_2 \boldsymbol{\psi}_2^T, \widehat{\mathbb{E}}_T[\mathbb{P}]\rangle|^2$.

3. $\boldsymbol{\psi}_3$ has its support inside either one of the two sets $\{\boldsymbol{\psi}_2 \geqslant 0\}$ or $\{\boldsymbol{\psi}_2 \leqslant 0\}$;

4. maximize the magnitude of the inner product between $\boldsymbol{\psi}_3\boldsymbol{\psi}_3^T$ and the reconstruction error, $\left[\widehat{\mathbb{E}}_T[\mathbb{P}] - \langle \widehat{\mathbb{E}}_T[\mathbb{P}], \boldsymbol{\psi}_2\boldsymbol{\psi}_2^T\rangle \boldsymbol{\psi}_2\boldsymbol{\psi}_2^T\right]$,

$$\boldsymbol{\psi}_3 = \underset{\boldsymbol{\psi}_3 \text{ defined by } (12)}{\operatorname{argmax}} \left|\langle \boldsymbol{\psi}_3 \boldsymbol{\psi}_3^T, \widehat{\mathbb{E}}_T[\mathbb{P}]\rangle\right|^2, \qquad (21)$$

5. repeat until we find $\boldsymbol{\psi}_n$.



Theoretical guarantees for Algorithm 1

- $\mathbb{E}[\mathbb{P}] = \mathbf{P}$ is the edge probability matrix of a balanced $\mathrm{SBM}(p, q, n)$;
- we observe that $\widehat{\mathbb{E}}_T[\mathbb{P}] \to \mathbb{E}[\mathbb{P}]$ when the graph size $n \to \infty$;
- we analyse the algorithm when its input is $\mathbb{E}[\mathbb{P}]$ (instead of $\widehat{\mathbb{E}}_T[\mathbb{P}]$);
- $\boldsymbol{\psi}_1 = n^{-1/2}\mathbf{1}$; $\begin{bmatrix}\boldsymbol{\psi}_1 & \cdots & \boldsymbol{\psi}_n\end{bmatrix}$ is the Soules basis returned by Algorithm 1.

***Lemma* 1.** We have

$$\sum_{k=1}^{M} \boldsymbol{\psi}_k \boldsymbol{\psi}_k^T(i,j) = \begin{cases} M/n & \text{if } \exists\, m \in [M],\ (i,j) \in B_m \times B_m, \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

***Corollary* 1.** $\begin{bmatrix}\boldsymbol{\psi}_1 & \cdots & \boldsymbol{\psi}_n\end{bmatrix}$ solves (20).

We reconstruct the normalized Laplacian of $\widehat{\boldsymbol{\mu}}_T[\mathbb{P}]$,

$$\mathcal{L}\left(\widehat{\boldsymbol{\mu}}_T[\mathbb{P}]\right) = \sum_{k=1}^{n} \widehat{\mathbb{E}}_T[\lambda_k]\, \boldsymbol{\psi}_k \boldsymbol{\psi}_k^T. \tag{23}$$

See https://arxiv.org/abs/2502.00038 for the proofs.

A partial reconstruction

1. in practice, the estimator $\mathcal{L}(\widehat{\mu}_T[\mathbb{P}])$ in (23) is very poor;
2. the full expansion (23) of $\mathcal{L}(\widehat{\mu}_T[\mathbb{P}])$ using $[\psi_1 \quad \cdots \quad \psi_n]$ is plagued by:
   a) $\lambda_n \geqslant \lambda_{n-1} \geqslant \cdots$ are noisy because they come from the bulk;
   b) $\psi_n, \psi_{n-1}, \ldots$ have small support and are therefore unstable;
3. ... but the reconstruction in (16) for a balanced SBM depends only on $[\psi_1 \cdots \psi_M]$

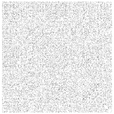
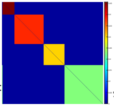
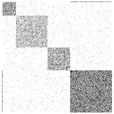
💡 replace the full reconstruction (23) with the following truncated estimator,

$$\widehat{\mathcal{L}}_M(\widehat{\mu}_T[\mathbb{P}]) \stackrel{\text{def}}{=} \mathrm{Id} - \sum_{k=1}^{M} \left(1 - \widehat{\mathbb{E}}_T[\lambda_k]\right) \psi_k \psi_k^T; \tag{24}$$

4. we propose the following estimator of the adjacency matrix of the barycentre graph,

$$\widehat{\mu}_T^M[\mathbb{P}] \stackrel{\text{def}}{=} \widehat{\mathbf{D}}^{1/2} \Big(\sum_{k=1}^{M} \left(1 - \widehat{\mathbb{E}}_T[\lambda_k]\right) \psi_k \psi_k^T \Big) \widehat{\mathbf{D}}^{1/2}. \tag{25}$$

A technical detail

- is $\boldsymbol{A} =$  a stochastic block model? ...
- yes ... it was generated by $\mathbf{P} =$ , and we applied a permutation on the adjacency matrix
- our algorithm necessitates that $\widehat{\mathbb{E}}_{\mathrm{T}}[\mathbb{P}]$ be "well-aligned", $\boldsymbol{A} =$ 

we aggregate the nodes into clusters wherein $\widehat{\mathbb{E}}_{\mathrm{T}}[\mathbb{P}]$ is approximately constant

- we use a spectral clustering algorithm based on the eigenvectors of $\mathcal{L}(\widehat{\mathbb{E}}_{\mathrm{T}}[\mathbb{P}])$
- equivalent to the approximation of each $\boldsymbol{A}^{(t)}$ using a step graphon;

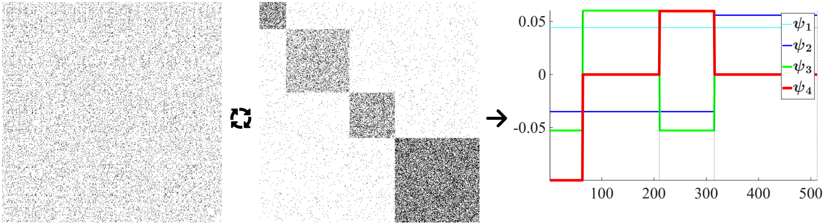
Spectral clustering of the nodes.

- the clustering of the nodes is not always accurate;
- but: the greedy algorithm relies on the M coarsest scale Soules basis, $\psi_1, \ldots, \psi_M$;
- ψ_k is determined by the computation of $\left|\langle \psi_k \psi_k^T, \widehat{\mathbb{E}}_T[\mathbb{P}] \rangle\right|^2$;

💡 the support of ψ_k is large for $k = 2, 3, \ldots$, and $\psi_k \psi_k^T$ is piecewise constant;

→ the noise in $\widehat{\mathbb{E}}_T[\mathbb{P}]$ is partly suppressed when computing $\left|\langle \psi_k \psi_k^T, \widehat{\mathbb{E}}_T[\mathbb{P}] \rangle\right|^2$;

- ψ_k are well aligned along the boundaries of large "noisy blocks" of $\widehat{\mathbb{E}}_T[\mathbb{P}]$.



Experiments

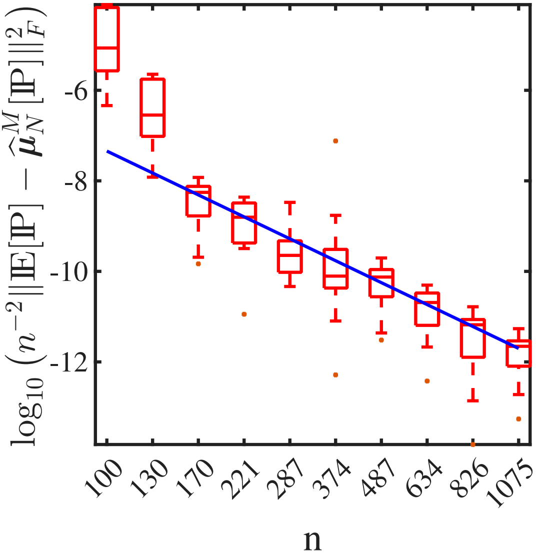
software: https://github.com/francoismeyer/barycentre-network

Rate of convergence of $\widehat{\mu}_T^M\left[\mathbb{P}\right]$ as a function of n

- $M = 4$ communities of sizes $63, 147, 105, 197$;
- edge probabilities $p_i = c_i \log n^2/n$, $c_i \sim \mathcal{U}[1,4]$, and $q = 2\log n/n$; the graphs are sparse and are connected almost surely.
- $\mathbf{A}^{(t)}$ is permuted with a different random permutation for each t;
- network size: $n = 100 \longrightarrow 1,075$; we compute the mean squared error,

$$n^{-2}\left\|\mathbb{E}[\mathbb{P}] - \widehat{\mu}_T^M\left[\mathbb{P}\right]\right\|_F^2 \stackrel{\text{def}}{=} \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|p_{ij} - \widehat{p}_{ij}\right|^2, \tag{26}$$

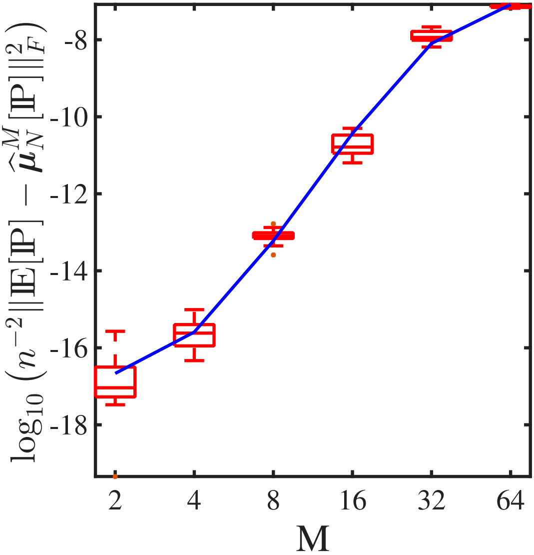
- we found $n^{-2}\left\|\mathbb{E}[\mathbb{P}] - \widehat{\mu}_T^M\left[\mathbb{P}\right]\right\|_F^2 \propto n^{-1.84}$;
- same order as $n^{-1}\log(M) + n^{-2}M^2$, the optimal (minimax) rate for the estimation of graphons under the mean squared error [18, 22, 40];
- concentration phenomenon is in effect: validates the theoretical derivations that were obtained in the limit $n \to \infty$;
- our approach works beyond the balanced stochastic block models.



Mean squared error $n^{-2}\left\|\mathbb{E}[\mathbb{P}] - \widehat{\boldsymbol{\mu}}_T^M\left[\mathbb{P}\right]\right\|_F^2$ as a function of the network size, n.

effect of the number of blocks M on the estimation of $\widehat{\boldsymbol{\mu}}_T^M\left[\mathbb{P}\right]$

- balanced $\mathrm{SBM}\left(\mathbf{p}, q, n\right)$, $n = 1,024$ nodes;
- M communities of sizes n/M;
- number of blocks: $M = 2 \longrightarrow 64$;
- when M becomes large, then $\lambda_2, \ldots, \lambda_M$ all converge to 1.
- $\lambda_2, \ldots, \lambda_M$ are no longer separated from the bulk;
- the truncated reconstruction (25) becomes numerically unstable,
- the mean squared error $n^{-2}\left\|\mathbb{E}[\mathbb{P}] - \widehat{\boldsymbol{\mu}}_T^M\left[\mathbb{P}\right]\right\|_F^2$ increases with M.



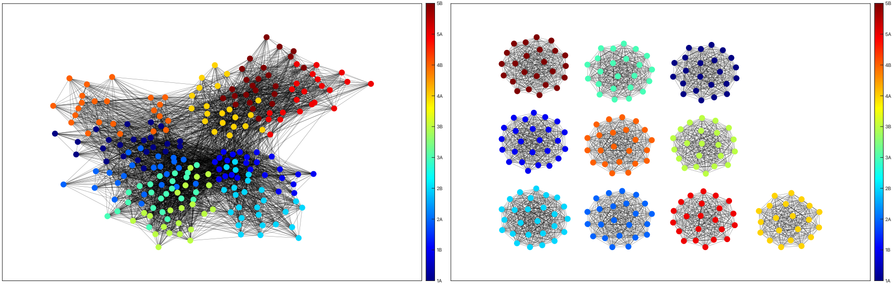
Mean squared error $n^{-2}\left\|\mathbb{E}[\mathbb{P}] - \widehat{\boldsymbol{\mu}}_T^M\left[\mathbb{P}\right]\right\|_F^2$ as a function of the number of blocks, M.

Real world networks

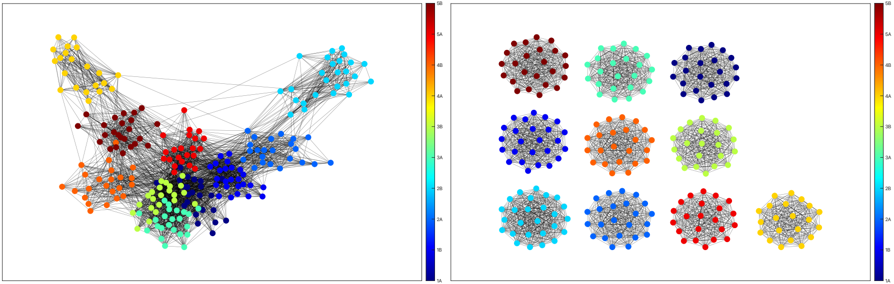
- French primary school dataset [24];
- exclude the lunch period because many students leave the school to take their lunch at home;
- morning period is divided into $T = 35$ time intervals of approximately 6 minutes; morning barycentre graph is computed using the T graphs;
- the afternoon is divided into $T = 26$ time intervals of approximately 6 minutes; afternoon barycentre graph is determined using the T graphs;
- For each $t = 1, \ldots, T$ we construct an undirected unweighted graph $G^{(t)}$, where the $n = 232$ nodes correspond to the students in the 10 classes;
- we consider the hypothesis that each class is a community of connected students;
- In fact, students in each class are only weakly connected (e.g., 9:00 AM, and 2:03 PM);

goal of the experiment: recover the communities determined by the classes.

Morning [top] and Afternoon [bottom] periods

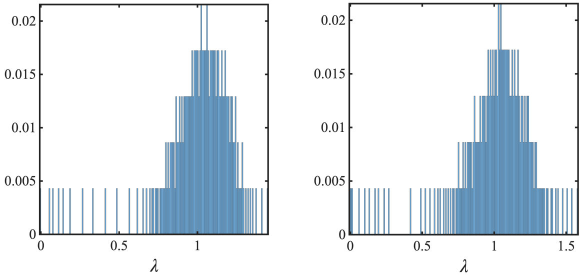


graph of the average network $\widehat{\mathbb{E}}_T[\mathbb{P}]$ barycentre graph $\widehat{\mu}_T^M[\mathbb{P}]$



☛ the Laplacian spectral barycentre graph $\widehat{\mu}_T[\mathbb{P}]$ recovers the classes

- recesses and lunchtime periods trigger significant changes in the number of links between the 10 classes;
- the community structure associated with the individual classes collapses in the sample mean adjacency matrix $\widehat{\mathbb{E}}_T[\mathbb{P}]$;
- in contrast $\boldsymbol{\lambda}\big(\mathcal{L}(\boldsymbol{A}^{(t)})\big), t = 1, \ldots, T$ are much more stable



morning afternoon

Thanks!

... Questions?

References

1. Bhagat, S. *et al.* α-Aminophosphonates as novel anti-leishmanial chemotypes: synthesis, biological evaluation, and CoMFA studies. *Med. Chem. Commun.* **5,** 665–670 (5 2014).

2. Lee, J.-J. *et al.* A neuroimaging biomarker for sustained experimental and clinical pain. *Nature medicine* **27,** 174–182 (2021).

3. Sturm, K.-T. Probability measures on metric spaces of nonpositive. *Heat kernels and analysis on manifolds, graphs, and metric spaces* **338,** 357 (2003).

4. Blanchard, M. & Jaffe, A. Q. Fréchet mean set estimation in the Hausdorff metric, via relaxation. *Bernoulli* **31,** 432 –456 (2025).

5. Donnat, C. & Holmes, S. Tracking network dynamics: A survey using graph distances. *The Annals of Applied Statistics* **12,** 971–1012 (2018).

6. Wills, P. & Meyer, F. G. Metrics for graph comparison: a practitioner's guide. *PLoS ONE* **15(2),** 1–54 (2020).

7. Athreya, A. *et al.* Eigenvalues of stochastic blockmodel graphs and random graphs with low-rank edge probability matrices. *Sankhya A,* 1–28 (2022).

8. Chakrabarty, A. *et al.* Eigenvalues Outside the Bulk of Inhomogeneous Erdős–Rényi Random Graphs. *Journal of Statistical Physics* **181,** 1746–1780 (2020).

9. Ferrer, M. *et al. Synthesis of median spectral graph* in *Pattern Recognition and Image Analysis* (2005), 139–146.

10. White, D. & Wilson, R. C. *Spectral generative models for graphs* in *ICIAP 2007* (2007), 35–42.

11. Abbe, E. Community detection and stochastic block models: recent developments. *Journal of Machine Learning Research* **18,** 1–86 (2018).

12. Faskowitz, J. *et al.* Weighted stochastic block models of the human connectome across the life span. *Scientific reports* **8,** 12997 (2018).

13. Thibeault, V. *et al.* The low-rank hypothesis of complex systems. *Nature Physics* **20 (2),** 294–302 (2024).

14. Young, J.-G. *et al.* Universality of the stochastic block model. *Phys. Rev. E* **98,** 032309 (3 2018).

15. Airoldi, E. M. *et al. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation* in *Advances in Neural Information Processing Systems* (2013), 692–700.

16. Ferguson, D. & Meyer, F. G. Theoretical analysis and computation of the sample Fréchet mean of sets of large graphs. *Information and Inference* **12,** 1347–1404 (Mar. 2023).

17. Gerlach, M. *et al.* A network approach to topic models. *Science advances* **4,** eaaq1360 (2018).

18. Olhede, S. C. & Wolfe, P. J. Network histograms and universality of blockmodel approximation. *PNAS* **111,** 14722–14727 (2014).

19. Borgs, C. *et al.* Private graphon estimation for sparse graphs. *Advances in Neural Information Processing Systems* **28** (2015).

20. Borgs, C. *et al.* in *Building Bridges II: Mathematics of László Lovász* 29–157 (Springer, 2020).

21. Doležal, M. *et al.* Relating the cut distance and the weak∗ topology for graphons. *Journal of Combinatorial Theory, Series B* **147,** 252–298 (2021).

22. Gao, C. *et al.* Rate-optimal graphon estimation. *The Annals of Statistics* **43,** 2624–2652 (2015).

23. Abbe, E. *et al.* Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory* **62,** 471–487 (2016).

24. Stehlé, J. *et al.* High-resolution measurements of face-to-face contact patterns in a primary school. *PloS one* **6,** e23176 (2011).

25. Bovet, A. *et al.* Flow stability for dynamic community detection. *Science advances* **8,** eabj3063 (2022).

26. Cencetti, G. & Barrat, A. Generating surrogate temporal networks from mesoscale building blocks. *Communications Physics* **8,** 159 (2025).

27. Contisciani, M. *et al.* Inference of hyperedges and overlapping communities in hypergraphs. *Nature communications* **13,** 7229 (2022).

28. Djurdjevac Conrad, N. *et al.* Detection of dynamic communities in temporal networks with sparse data. *Applied Network Science* **10,** 1 (2025).

29. Failla, A. *et al.* Describing group evolution in temporal data using multi-faceted events. *Machine Learning* **113,** 7591–7615 (2024).

30. Ferraz de Arruda, G. *et al.* Phase transitions and stability of dynamical processes on hypergraphs. *Communications Physics* **4,** 24 (2021).

31. Gauvin, L. *et al.* Detecting the community structure and activity patterns of temporal networks: a non-negative tensor factorization approach. *PloS one* **9,** e86028 (2014).

32. Gemmetto, V. *et al.* Mitigation of infectious disease at school: targeted class closure vs school closure. *BMC infectious diseases* **14,** 1–10 (2014).

33. Génois, M. & Barrat, A. Can co-location be used as a proxy for face-to-face contacts? *EPJ Data Science* **7,** 1–18 (2018).

34. Sattar, N. S. *et al.* Exploring temporal community evolution: algorithmic approaches and parallel optimization for dynamic community detection. *Applied Network Science* **8,** 64 (2023).

35. Avrachenkov, K. *et al. Spectral properties of random matrices for stochastic block model* in *International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks* (2015), 537–544.

36. Oliveira, R. I. Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. *arXiv preprint arXiv:0911.0600* (2009).

37. Zhang, X. *et al.* Spectra of random graphs with community structure and arbitrary degrees. *Physical review E* **89,** 042816 (2014).

38. Löwe, M. & Terveer, S. Hitting times for Random Walks on the stochastic block model. *arXiv preprint arXiv:2401.07896,* 1–26 (2024).

39. Elsner, L. *et al.* Orthogonal bases that lead to symmetric nonnegative matrices. *Linear Algebra and its Applications* **271,** 323–343 (1998).

40. Xu, J. *Rates of convergence of spectral methods for graphon estimation* in *International Conference on Machine Learning* (2018), 5433–5442.