Computation of the Laplacian Spectral Barycentre Graph in a Soules Basis

François G. Meyer
Applied Mathematics
University of Colorado at Boulder, Boulder CO 80305
fmeyer@colorado.edu
https://orcid.org/0000-0002-1529-3796

January 30, 2025

Abstract

The main contribution of this work is a fast algorithm to compute the barycentre of a set of graphs based on a Laplacian spectral pseudo-distance. The core engine for the estimation of the barycentre is an algorithm that explores the large library of Soules bases, and returns a basis that leads to the reconstruction of a weighted graph whose spectrum is the sample mean spectrum, and whose geometry matches that of the sample mean adjacency matrix. We prove that when the graphs are random realizations of stochastic block models, then our algorithm reconstructs the population mean adjacency matrix. In addition to the theoretical analysis of the estimator of the barycentre graph, we perform Monte Carlo simulations to validate the theoretical properties of the estimator. This work is significant because it opens the door to the design of new spectral-based graph synthesis that have theoretical guarantees.

Keywords: Barycentre graph; Soules basis; Fréchet mean; Laplacian Spectral distance; statistical analysis of graph-valued data.

1 Introduction, problem statement, and related work

1.1 The barycentre graph

The design of machine learning algorithms that can analyze "graph-valued random variables" is of fundamental importance (e.g. [11, 18, 19, 20, 23, 28, 38, 41], and references therein). Such machine learning algorithms often require the computation of a "sample mean" graph that can summarize the topology and connectivity of a dataset of graphs, $\{G^{(1)}, \ldots, G^{(N)}\}$. Formally, we denote by $\mathcal{S}$ the set of $n \times n$ symmetric adjacency matrices with nonnegative weights, and we assume that the adjacency matrix $\mathbf{A}^{(k)}$ of the graph $G^{(k)}$ is sampled from a probability space $(\mathcal{S}, \mathbb{P})$. An example of a probability space is the stochastic block model (see Section 1.4). We equip the probability space $(\mathcal{S}, \mathbb{P})$ with a metric d to quantify proximity of graphs. Then, a notion of summary graph is provided by the concept of *barycentre* [33], or *Fréchet mean* [4], graph, $\widehat{\mu}_N[\mathbb{P}]$, which minimizes the sum of the squared distances to all the graphs in the ensemble,

$$\widehat{\mu}_N[\mathbb{P}] \stackrel{\text{def}}{=} \operatorname*{argmin}_{\mathbf{B} \in \mathcal{S}} \sum_{k=1}^{N} d^2(\mathbf{B}, \mathbf{A}^{(k)}). \tag{1}$$

In this work, we propose a fast algorithm to compute the barycentre of a set of graphs based on a Laplacian spectral pseudo-distance.

Before continuing, we introduce some notations. We denote by $[n] \stackrel{\text{def}}{=} \{1, \ldots, n\}$. We define $\mathbf{1} \stackrel{\text{def}}{=} [1 \cdots 1]^T$, and $\mathbf{J} = \mathbf{1}\mathbf{1}^T$. We use $\mathbf{A}$ to denote the adjacency matrix of a graph G, and $\mathbf{D}$ to denote the diagonal degree matrix. The symmetric normalized adjacency matrix, $\widehat{\mathbf{A}} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$, is defined by $\hat{a}_{ij} \stackrel{\text{def}}{=} a_{ij}/\sqrt{d_i d_j}$ if $d_i d_j \neq 0$, and is zero otherwise. The normalized Laplacian is defined by $\mathcal{L} \stackrel{\text{def}}{=} \mathrm{Id} - \widehat{\mathbf{A}}$. We denote by $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_n]$ the ascending sequence of eigenvalues of $\mathcal{L}$.

1.2 The Laplacian spectral pseudo-distance

The metric $\mathfrak{d}$ that is chosen in (1) to compute $\widehat{\mu}_N[\mu]$ influences the topological characteristics that $\widehat{\mu}_N[\mu]$ inherits from $\{G^{(1)}, \dots, G^{(N)}\}$ [25]. We advocate that the distance between graphs should be evaluated in the spectral domain, by comparing the eigenvalues of the normalized Laplacian, $\mathcal{L}^{(k)}$, of the respective graphs $G^{(k)}$. We define the Laplacian spectral pseudo-metric as

$$\mathfrak{d}(\mathcal{L}, \mathcal{L}') \stackrel{\text{def}}{=} \left\| \lambda(\mathcal{L}) - \lambda(\mathcal{L}') \right\|_2, \tag{2}$$

where $\lambda(\mathcal{L})$ and $\lambda(\mathcal{L}')$ are the vectors of eigenvalues of $\mathcal{L}$ and $\mathcal{L}'$ respectively. This pseudo-distance captures at multiple scales the structural and connectivity information in the graphs [10, 37]. Defining a pseudo-distance in the spectral domain alleviates the difficulty of solving the node correspondence problem, and in the case of the normalized Laplacian, it makes it possible to compare graphs of different sizes. When the graphs are realizations of a stochastic block model, the eigenvalues of $\mathcal{L}$ associated with each community are better separated from the bulk than the corresponding eigenvalues of $L \stackrel{\text{def}}{=} D - A$ [8].

1.3 From the spectrum to the Laplacian

In spite of the advantages of the pseudo-metric $\mathfrak{d}$ (2), the computation of (1) leads to two technical obstacles. The first challenge stems from the fact that $\mathfrak{d}$ is defined in the spectral domain, but the optimization (1) takes place in $\mathcal{S}$. This leads to the definition of a *realizable* sequence; we say that $\lambda = [\lambda_1, \dots, \lambda_n]$ is realizable if there exists $A \in \mathcal{S}$ whose Laplacian, $\mathcal{L}(A)$, satisfies $\lambda(\mathcal{L}(A)) = \lambda$. Further, we define $\mathcal{R}$ to be the *set of realizable sequences*. We can formally define the optimisation problem associated with the estimation of $\widehat{\mu}_N[\mathbb{P}]$ in (1),

$$\lambda\left(\widehat{\mu}_N[\mathbb{P}]\right) = \underset{\lambda \in \mathcal{R}}{\operatorname{argmin}} \sum_{k=1}^{N} \|\lambda - \lambda(\mathcal{L}^{(k)})\|_2^2. \tag{3}$$

If we relax this minimization problem ($\lambda \in \mathbb{R}^n$), then the solution to (3) is the sample mean $\widehat{\mathbb{E}}_N[\lambda] \stackrel{\text{def}}{=} N^{-1} \sum_{k=1}^{N} \lambda(\mathcal{L}^{(k)})$, which is in general not realizable.

Which brings us to the second difficulty in using a spectral pseudo-distance. The knowledge of the eigenvalues of the barycentre graph, $\lambda\left(\widehat{\mu}_N[\mathbb{P}]\right)$, is insufficient to reconstruct a graph; we need a set of eigenvectors that summarizes the distribution of eigenvectors associated with the respective Laplacian matrices $\mathcal{L}^{(k)}$. To address this problem, several authors have proposed to align the eigenvectors of the respective graph adjacency matrices [16] or Laplacian matrices [36]. Others [15] have proposed numerical methods to find the best SBM whose eigenvalues match the sample mean eigenvalues. More generally, we are interested in the question of solving a symmetric nonnegative inverse eigenvalue problem [35].

In this work, we prove that the sample mean vector of eigenvalues, $\widehat{\mathbb{E}}_N[\lambda]$ is in fact realizable, and we construct a Soules basis of eigenvectors that yields a sparse approximation of the sample mean adjacency matrix. We prove that when the graphs are random realizations of stochastic block models (SBM), then our method always reconstructs the population mean adjacency matrix.

1.4 The stochastic block model

To provide theoretical guarantees for the algorithms presented in this paper, we analyse the algorithms when the graphs are sampled from a stochastic block model (e.g. [1], and references therein). Stochastic block models provide universal approximants to graphs and can be used as building blocks to analyse more complex graphs [2, 15, 26, 40].

We define the general stochastic block model $\mathrm{SBM}(p, q, n)$. Let $\{B_k\}, 1 \leqslant k \leqslant M$ be a partition of the vertex set $[n]$ into M blocks (or communities). We define the vector $p = [p_1, \cdots, p_M]$ to be the edge probabilities within each block, and q to be the edge probability between blocks. The entries $a_{ij} = a_{ji}, i < j$ of the adjacency matrix A are independent (up to symmetry) and are distributed with Bernoulli distributions with parameter p_m if i and j are in the same block B_m, and parameter q if i and j are in distinct blocks.

We often represent SBM $(\mathbf{p}, \mathbf{q}, n)$ by the matrix of edge probabilities, or matrix of connection probabilities, $\mathbf{P} \overset{\text{def}}{=} \mathbb{E}[\mathbf{A}]$. We sometimes consider a balanced version of the model where all blocks have the same size, $|B_m| = n/M$, (in that case we assume without loss of generality that n is a multiple of M), and all the edge probabilities are equal, $p_1 = \cdots = p_M$.

1.5 Content of the paper: our main contributions

The main contribution of this work is a fast algorithm to compute the barycentre of a set of graphs based on a Laplacian spectral pseudo-distance. The core engine is an algorithm that explores the large library of Soules bases [13, 32] (which is organized as a binary tree [12]), and returns a Soules basis that can be used to construct the normalized Laplacian of a graph, whose eigenvalues are equal to the population mean spectrum. We prove that if the graphs are random realizations of SBM $(\mathbf{p}, q, n)$, the algorithm can reconstruct the population mean adjacency matrix associated with the SBM. This work is significant because it opens the door to the design of new spectral-based graph synthesis [31, 3] that have theoretical guarantees. We publicly share our code to facilitate future work [24].

1.6 Organization of the paper

In the next section, we review the construction of Soules bases [32, 12] and their properties. The reader who is already familiar with Soules bases can skip to Section 3 wherein we describe the construction of the best Soules basis for the approximation of $\mathbf{P}$ and derive the theoretical properties of the basis. In Section 4 we describe the reconstruction of the normalized Laplacian, and in Section 5, we define the estimator of the adjacency matrix of the barycentre graph. In section 6, we report the results of some experiments. In Section 7, we discuss the implications of our work. The proofs of some technical lemmata are left aside in Section 8.

2 Soules Bases

Soules bases [12, 32] were invented to provide a solution to the following symmetric nonnegative inverse eigenvalue problem: given and ordered sequence of eigenvalues, $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$, find an orthogonal matrix $\mathbf{\Psi}$ such that $\mathbf{\Psi}\,\mathrm{diag}\,(\lambda_1, \ldots, \lambda_n)\,\mathbf{\Psi}^T \geqslant 0$, where we write $\mathbf{A} \geqslant 0$ if the entries of the matrix $\mathbf{A}$ are nonnegative. The original construction [32] (see also [27] for a much earlier description of the same idea) provided a single basis. The authors in [12] realized that the transformation (4) could in fact be used to derive a library of so-called Soules-bases, each of which can be represented by a specific binary tree (see Fig. 2-Right). The Soules bases (or matrices) are orthogonal matrices that are constructed iteratively by applying a product of Givens rotations to a fixed vector $\boldsymbol{\psi}_1$ with nonnegative entries. Alternatively, a Soules basis is constructed iteratively using a simple transformation described in the next section.

2.1 Construction of Soules Bases

The construction of a Soules basis proceeds iteratively, starting at the coarsest level ($l = 1$) with a normalized vector $\boldsymbol{\psi}_1$ with nonnegative entries, whose support is the interval $I^{(1)} = [n]$ (see Fig. 2-left). This level, $l = 1$, is the coarsest level. The algorithm progresses from the top level $l = 1$ down to the finest level $l = n$. At any given level l, the set $[n]$ is partitioned into l ordered intervals $I_q^{(l)}, 1 \leqslant q \leqslant l$, with $\cup_{q=1}^{l} I_q^{(l)} = [n]$, and $I_q^{(l)} \cap I_p^{(l)} = \emptyset$ if $p \neq q$ (see Fig.2-right). When progressing from level l to $l+1$, one chooses an interval, $I_q^{(l)} = [i_0, i_1)$, and one chooses an index $k \in [i_0, i_1]$ and defines $I_q^{(l+1)} \overset{\text{def}}{=} [i_0, k]$, and $I_{q+1}^{(l+1)} \overset{\text{def}}{=} [k+1, i_1]$ (see Fig. 1-right). The split of $I_q^{(l)}$ into $I_q^{(l+1)}$ and $I_{q+1}^{(l+1)}$ triggers the construction of the Soules vector $\boldsymbol{\psi}_{l+1}$ (see Fig. 1-left), defined by

$$\boldsymbol{\psi}_{l+1}(i) \overset{\text{def}}{=} \frac{1}{\sqrt{\|\boldsymbol{\psi}(i_0 : i_1)\|}} \begin{cases} \dfrac{\|\boldsymbol{\psi}_1(k+1 : i_1)\|}{\|\boldsymbol{\psi}_1(i_0 : k)\|} \boldsymbol{\psi}_1(i) & \text{if } i_0 \leqslant i \leqslant k \\ -\dfrac{\|\boldsymbol{\psi}_1(i_0 : k)\|}{\|\boldsymbol{\psi}_1(k+1 : i_1)\|} \boldsymbol{\psi}_1(i) & \text{if } k+1 \leqslant i \leqslant i_1, \end{cases} \tag{4}$$

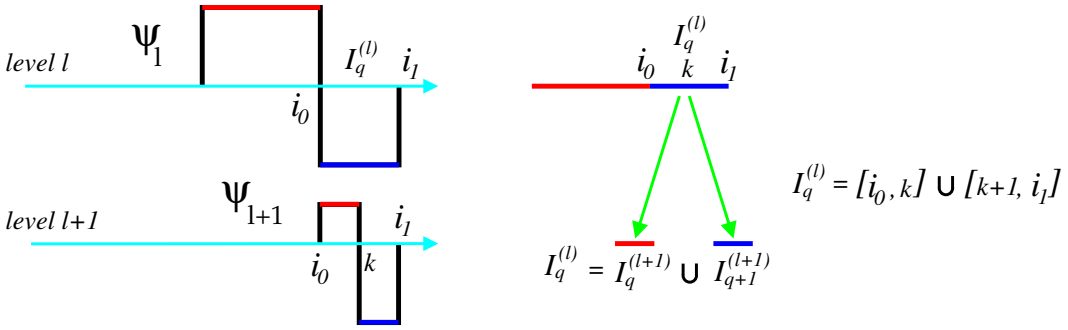

Figure 1: Left: the vector ψ_{l+1} is created by splitting a block of indices $I_q^{(l)} = [i_0, i_1]$ at level l into two sub-blocks, $[i_0, k] \cup [k+1, i_1]$ at level $l+1$. Right: a node is created in the Soules binary tree by splitting the interval $[i_0, i_1]$ into $[i_0, k] \cup [k+1, i_1]$

where the vectors $\psi_1(i_0 : i_1), \psi_1(i_0 : k)$, and $\psi_1(k+1 : i_1)$ are n-dimensional vectors where the nonzero entries are extracted from ψ_1 at the corresponding indices,

$$\begin{aligned}
\psi_1(i_0 : i_1) &= [0 \cdots 0 \;\; \psi_1(i_0) \cdots \psi_1(k) \; \psi_1(k+1) \cdots \psi_1(i_1) \; 0 \cdots 0]^T, \\
\psi_1(i_0 : k) &= [0 \cdots 0 \;\; \psi_1(i_0) \cdots \psi_1(k) \; 0 \cdots\cdots\cdots\cdots\cdots\cdots 0]^T, \\
\psi_1(k+1 : i_1) &= [0 \;\; \cdots\cdots\cdots\cdots\cdots\cdots 0 \; \psi_1(k+1) \cdots \psi_1(i_1) \; 0 \cdots 0]^T.
\end{aligned} \tag{5}$$

The new vector ψ_{l+1} oscillates over the block $I_q^{(l)}$: taking positive values in on $I_q^{(l+1)}$, and negative values on $I_{q+1}^{(l+1)}$ (see Fig. 1-left).

We observe that two Soules vectors ψ_l and ψ_m, $l \neq m$, are either nested, or they do not overlap. Indeed, either $\text{supp}(\psi_l) \cap \text{supp}(\psi_m) = \emptyset$, or (without loss of generality) $\text{supp}(\psi_l) \subset \text{supp}(\psi_m)$, $\text{supp}(\psi_l)$ coincides with one of the two intervals in $\text{supp}(\psi_m)$ where ψ_m is constant, and ψ_l is given by (4). In both cases, we have $\langle \psi_l, \psi_m \rangle = 0$.

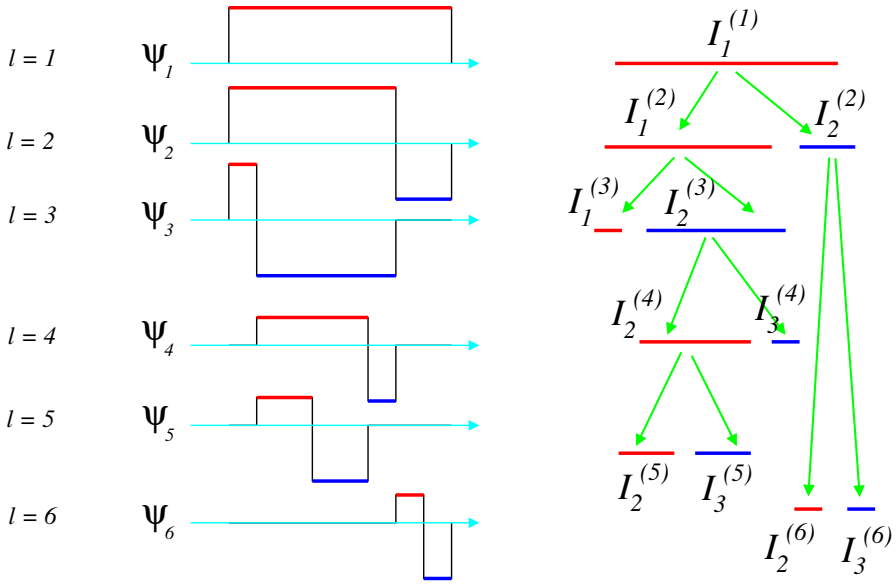

Figure 2: Left: starting from level $l = 1$, one Soules vector ψ_l is constructed at each level $l \geqslant 2$ by selecting and then splitting an interval $I_q^{(l)}$ over which an already existing vector ψ_m, $m \leqslant l$ keeps a constant value. Right: each Soules basis is associated with a binary tree. The leaves of the tree are intervals that are not split, $(I_1^{(3)}, I_2^{(5)}, I_3^{(5)}, I_3^{(4)}, I_2^{(6)}, I_3^{(6)})$. Each node with two children generates a vector ψ_l. The root of the tree $(I_1^{(1)})$ yields ψ_1, the original vector, and ψ_2, the vector created by the first split.

The iterative subdivision process can be described using a binary tree (see Fig.2-right) where a new vector is created at each node that has two children. In fact, each Soules basis is uniquely encoded by the geometry of the corresponding binary tree (see [34] for a similar construction of Walsh-Hadamard packets). Conversely, one can tailor the tree to guide the selection of the intervals $I_q^{(l)}$ that are split, and the shape of the vectors $\boldsymbol{\psi}_l$.

2.2 Properties of Soules Bases

The iterative procedure (4) leads to the construction of a set of n unit norm vectors that are mutually orthogonal. Therefore, after n steps $\begin{bmatrix}\boldsymbol{\psi}_1 & \cdots & \boldsymbol{\psi}_n\end{bmatrix}$ is an orthonormal matrix [12].
Using (4), we can derive the following lemma with a proof by induction.

Lemma 1 (See [12]). *Let $\begin{bmatrix}\boldsymbol{\psi}_1 & \cdots \boldsymbol{\psi}_n\end{bmatrix}$ be a Soules basis constructed according to (4). Then,*

$$\forall m = 1, \dots, n, \quad \mathbf{E}_m \overset{def}{=} \sum_{q=1}^{m} \boldsymbol{\psi}_q \boldsymbol{\psi}_q^T \geqslant 0, \quad \textit{and} \quad \mathbf{E}_n = \mathrm{Id}\,. \tag{6}$$

We now consider an ordered sequence of eigenvalues, $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$, and we define the associated diagonal matrix, $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$. We can write $\boldsymbol{\Lambda}$ as the sum of $n-1$ nonnegative diagonal matrices and a scaled version of the identity,

$$\boldsymbol{\Lambda} = \sum_{i=1}^{n-1} (\lambda_i - \lambda_{i+1}) \mathbf{D}_i + \lambda_n \,\mathrm{Id}, \tag{7}$$

where $\mathbf{D}_i$ is the diagonal matrix with the first i diagonal entries are equal to one, and zero afterwards, $\mathbf{D}_i = \operatorname{diag}(1, \cdots, 1, 0, \cdots, 0)$. Since $\boldsymbol{\Psi}\mathbf{D}_m\boldsymbol{\Psi}^T = \mathbf{E}_m = \sum_{q=1}^{m} \boldsymbol{\psi}_q \boldsymbol{\psi}_q^T$, we obtain the fundamental property of Soules bases.

Lemma 2 (See [12]). *Let $\boldsymbol{\Psi}$ be a Soules basis constructed according to (4). Let $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, where $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n$. Then, the off-diagonal entries of $\boldsymbol{\Psi}\boldsymbol{\Lambda}\boldsymbol{\Psi}^T$ are non-negative. In addition, if $\lambda_n \geqslant 0$, then $\boldsymbol{\Psi}\boldsymbol{\Lambda}\boldsymbol{\Psi}^T \geqslant 0$.*

We note that there has been some recent interest in Soules bases to solve various inverse eigenvalue problems [9, 29].

Remark 1. *The result in lemma 2 relies on the fact that the sequence of eigenvalues is decreasing (so that $\lambda_i - \lambda_{i+1} \geqslant 0$ in (7)). On the other hand, the eigenvalues of $\boldsymbol{\mathcal{L}}$ are by nature ranked in ascending order. The significance of this convention is that the index k of the eigenvalue λ_k of $\boldsymbol{\mathcal{L}}$ encodes the frequency of the corresponding eigenvector. Given an ascending sequence of eigenvalues of $\boldsymbol{\mathcal{L}}$, $0 = \lambda_1 < \lambda_2 \leqslant \cdots \leqslant \lambda_n$, we would like to apply lemma 2 to reconstruct a Laplacian matrix using a Soules basis. Since the off-diagonal entries of a normalized Laplacian $\boldsymbol{\mathcal{L}}$ are nonpositive, we need to work with $-\boldsymbol{\Lambda}$. Then, $0 = \lambda_1 > -\lambda_2 \geqslant \cdots \geqslant -\lambda_n$, and we can use lemma 2 to construct $\widehat{\boldsymbol{\mathcal{L}}}$ such that*

$$\widehat{\boldsymbol{\mathcal{L}}} = \boldsymbol{\Psi} \operatorname{diag}(\lambda_1, \dots, \lambda_n) \boldsymbol{\Psi}^T, \quad \textit{where } \widehat{\mathcal{L}}_{ij} \leqslant 0 \textit{ if } i \neq j. \tag{8}$$

Since we choose, $\boldsymbol{\psi}_1 = n^{-1/2}\mathbf{1}$, we have $\widehat{\boldsymbol{\mathcal{L}}}\mathbf{1} = \mathbf{0}$, and therefore $\widehat{\mathcal{L}}_{ii} \geqslant 0$. While the signs of the entries of $\widehat{\boldsymbol{\mathcal{L}}}$ match those of a normalized Laplacian, there is no guarantee that $\widehat{\boldsymbol{\mathcal{L}}}$ be a valid normalized Laplacian (but see a definite answer in the case of the combinatorial Laplacian, $\mathbf{L} = \mathbf{D} - \mathbf{A}$ in [9]).

This remark notwithstanding, we settle this question in section 4, in the case where all the graphs are sampled from $\mathrm{SBM}(p, q, n)$.

3 A Soules Basis for the sparse approximation of $\widehat{\mathbb{E}}_N[\mathbf{A}]$

3.1 Approximation of the graphs with a common SBM

The algorithm described in the following assumes that the N graphs $G^{(q)}$ in the sample can be approximated with a common stochastic block model, $\mathrm{SBM}(p, q, n)$. The question of approximating general graphs using SBMs, or

step graphons, has received a lot of attention recently (e.g. [2, 5, 6, 15, 17, 26, 30, 40], and references therein). Our experiments (not shown) demonstrate that many of these methods provide excellent approximations for a large class of (non SBM) graphs. This work focuses on the construction of the barycentre graph after an SBM approximation has been computed for each $G^{(k)}$. The study of the combined performance of the approximation with the computation of the barycentre is left for future work.

In the following, we therefore consider $\mathbf{A}^{(k)}$ to be the adjacency matrix of the SBM approximation of the graph $G^{(k)}$. We further assume that vertices of each $G^{(k)}$ have been rearranged in $\mathbf{A}^{(k)}$ according to the canonical ordering associated with the partition of $[n]$ into the blocks defined by $\mathrm{SBM}\left(\mathbf{p}, q, n\right)$.

3.2 Overview of the algorithm

Given N independent realizations of $\mathrm{SBM}\left(\mathbf{p}, q, n\right)$, represented by their adjacency matrices $\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(N)}$, we describe a greedy algorithm that constructs a Soules basis that provides at each level l a sparse approximation of the sample edge probability matrix $\widehat{\mathbb{E}}_N[\mathbf{A}] \overset{\text{def}}{=} N^{-1} \sum_{q=1}^{N} \mathbf{A}^{(q)}$. Given $\widehat{\mathbb{E}}_N[\mathbf{A}]$, algorithm 1 explores iteratively the binary tree of Soules vectors from the top level to the bottom level. At each level l, the algorithm selects a new Soules vector $\boldsymbol{\psi}^*_{l+1}$ that minimizes the residual approximation error between the sample mean adjacency matrix, $\widehat{\mathbb{E}}_N[\mathbf{A}]$, and its expansion in the first $l+1$ Soules vectors, $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_l, \boldsymbol{\psi}^*_{l+1}$,

$$\boldsymbol{\psi}^*_{l+1} = \underset{\boldsymbol{\psi}_{l+1} \text{ defined by } (4)}{\operatorname{argmin}} \left\| \widehat{\mathbb{E}}_N[\mathbf{A}] - \sum_{q=1}^{l} \langle \boldsymbol{\psi}_q, \widehat{\mathbb{E}}_N[\mathbf{A}] \rangle \boldsymbol{\psi}_q - \langle \boldsymbol{\psi}_{l+1}, \widehat{\mathbb{E}}_N[\mathbf{A}] \rangle \boldsymbol{\psi}_{l+1} \right\|_2^2. \tag{9}$$

3.2.1 Theoretical guarantees for the algorithm.

Our analysis of algorithm 1 is performed under the assumption that the input to the algorithm is not the sample mean adjacency matrix $\widehat{\mathbb{E}}_N[\mathbf{A}]$ but its population equivalent $\mathbb{E}[\mathbf{A}] = \mathbf{P}$. This assumption is realistic if the graph size n is large enough for some concentration phenomenon to be in effect. Our experiments (see Fig. 5-left) confirm the validity of this assumption. A finite sample analysis of the error bounds is left for future work.

The following lemma proves that the first M vectors of the Soules basis estimated by algorithm 1 recover the geometry of the SBM.

Lemma 3. *Let* $\mathbf{P}$ *be the population mean adjacency matrix of* $\mathrm{SBM}\left(\mathbf{p}, q, n\right)$ *defined by*

$$\mathbf{P} = \sum_{m=1}^{M} (p_m - q) \mathbf{1}_{B_m} \mathbf{1}_{B_m}^T + q\mathbf{J}, \tag{10}$$

where the M *blocks* B_m *form a partition of* $[n]$. *Let* $J_l, 1 \leqslant l \leqslant M$ *be the leaves in the binary Soules tree (these are intervals that are no longer split, see Fig. 2-right) after* M *steps of algorithm 1. Then, the blocks* $\{B_m\}$ *in (10) coincide with the intervals* $\{J_l\}$. *The entries of the matrix* $\mathbf{E}_M = \sum_{m=1}^{M} \boldsymbol{\psi}_m \boldsymbol{\psi}_m^T$ *satisfy*

$$e_M(i,j) = \begin{cases} \dfrac{1}{|J_m|} & \text{if } (i,j) \in J_m \times J_m, \\ 0 & \text{otherwise}, \end{cases} \tag{11}$$

where $|J_m|$ *is the length of the interval* J_m.

Proof. The proof of lemma 3 is provided in section 8.1. The first ingredient of the proof concerns the geometry of the matrix $\mathbf{E}_M$. A top-down exploration of the Soules binary tree, when the first Soules vector is $\boldsymbol{\psi}_1 = n^{-1/2}\mathbf{1}$, always results in a matrix $\mathbf{E}_M = \sum_{q=1}^{M} \boldsymbol{\psi}_q \boldsymbol{\psi}_q^T$ that is piecewise constant on square blocks aligned along the diagonal, and zero outside of the blocks. The second ingredient in the proof of lemma 3 specifically addresses the construction of each $\boldsymbol{\psi}_{l+1}$ in algorithm 1. We can prove that at each level l, the Soules vector $\boldsymbol{\psi}_{l+1}$ returned by algorithm 1 is aligned with the boundary of a block B_m of the edge probability matrix $\mathbf{P}$ defined by (10).

Algorithm 1 Top-down exploration of the Soules binary tree.

procedure BESTSOULESBASIS($\widehat{\mathbb{E}}_N[\boldsymbol{A}], \boldsymbol{\Psi}$)

▷ *Input: sample mean adjacency matrix $\widehat{\mathbb{E}}_N[\boldsymbol{A}]$; Output: $\boldsymbol{\Psi}$ the Soules matrix* ◁

for all levels $l \in \{1, \dots, n-1\}$ **do**

▷ *For each block $I_q^{(l)} = [i_0, i_1]$ which was not split at level l we split it using an index $k \in [i_0, i_1]$ and construct the eigenvector $\boldsymbol{\psi}_l$ associated with the split. We compute the coefficient $\langle \boldsymbol{\psi}_l \boldsymbol{\psi}_l^T, \widehat{\mathbb{E}}_N[\boldsymbol{A}] \rangle$* ◁

icoeff $\leftarrow 1$ ▷ *index of the tentative $\boldsymbol{\psi}_l$ at level l*

for all blocks $I_q^{(l)}$ at level l **do** ▷ *there are exactly l blocks at level l*

$i0 \leftarrow$ leftend$(I_q^{(l)})$ ▷ $I_q^{(l)} = [i_0, i_1]$

$i1 \leftarrow$ rightend$(I_q^{(l)})$

if $(i_0 < i_1)$ **then** ▷ *the block $I_q^{(l)}$ is not a leaf*

for all $k \in \{i_0, \dots, i_1\}$ **do**

$\boldsymbol{\psi}_l \leftarrow$ buildvector(B, k) ▷ *use (4) to construct $\boldsymbol{\psi}_l$*

coeff(icoeff) $\leftarrow \langle \boldsymbol{\psi}_l \boldsymbol{\psi}_l^T, \widehat{\mathbb{E}}_N[\boldsymbol{A}] \rangle$

icoeff $\leftarrow$ icoeff $+ 1$ ▷ *update the index of the next tentative $\boldsymbol{\psi}_l$*

end for ▷ *next index k so that $[i_0, i_1] = [i_0, k] \cup [k+1, i_1]$*

end if

end for ▷ *move to the next block at level l*

▷ *We have explored all the blocks at level l. We now find the block $[i_0, i_1]$ and the index k of the split that result in the largest $|\langle \boldsymbol{\psi}_l \boldsymbol{\psi}_l^T, \widehat{\mathbb{E}}_N[\boldsymbol{A}] \rangle|^2$. We save the corresponding $\boldsymbol{\psi}_l$ in $\boldsymbol{\Psi}$*

$(\text{bestSplit}, \text{bestBlock}) \leftarrow \underset{k \in B}{\text{argmax}} \underset{B}{\text{argmax}} \left(|\text{coeff}|^2 \right)$

$\boldsymbol{\psi}_l \leftarrow$ buildvector$(\text{bestBlock}, \text{bestSplit})$ ▷ *use (4) to construct $\boldsymbol{\psi}_l$*

$\boldsymbol{\Psi}(:, l) \leftarrow \boldsymbol{\psi}_l$ ▷ *save $\boldsymbol{\psi}_l$ in the Soules basis*

end for ▷ *go down to a finer level*

end procedure

4 The reconstruction of the normalized Laplacian

4.1 Description of the algorithm

In the following, we first present an algorithm to construct a matrix $\widehat{\mathcal{L}}$, whose spectrum is the sample mean spectrum $\widehat{\mathbb{E}}_N[\boldsymbol{\lambda}(\mathcal{L})]$. We prove that if the graphs are sampled from a balanced SBM with $p_1 = \cdots = p_M$ and $\widehat{\mathbb{E}}_N[\boldsymbol{\lambda}(\mathcal{L})] \approx \mathbb{E}[\boldsymbol{\lambda}(\mathcal{L})]$, then $\widehat{\mathcal{L}} = \mathcal{L}(\mathbf{P})$, the Laplacian of the population mean adjacency matrix. Our approach combines the structural information about the geometry of the block in the SBM, which is provided by lemma 3, along with the spectral information provided by the sample mean spectrum. Unfortunately, this solution, while theoretically satisfying, is numerically unstable. We propose a second estimator, which is numerically stable and has similar theoretical guarantees.

4.2 The complete reconstruction

We consider an SBM $(\mathbf{p}, q, n)$, where $\mathbf{p} = [p_1, \dots, p_M]$. We assume that we have access to N independent realizations of SBM $(\mathbf{p}, q, n)$, $G^{(1)}, \dots, G^{(N)}$. For each realization $G^{(q)}$, we compute the eigenvalues, $\boldsymbol{\lambda}(\mathcal{L}^{(q)})$, of the normalized

Laplacian $\mathcal{L}^{(q)}$. We estimate the sample mean spectrum of the normalized Laplacian,

$$\begin{bmatrix}\overline{\lambda}_1 & \dots & \overline{\lambda}_n\end{bmatrix}^T \stackrel{\text{def}}{=} \widehat{\mathbb{E}}_N\left[\lambda(\mathcal{L})\right] = \frac{1}{N}\sum_{q=1}^{N}\lambda(\mathcal{L}^{(q)}). \tag{12}$$

As explained in lemma 3, the Soules basis $\begin{bmatrix}\psi_1 & \cdots & \psi_n\end{bmatrix}$ generated by algorithm 1 recovers the geometry of the blocks of SBM (p, q, n). We propose to use this basis to estimate $\mathcal{L}$, and we define the following estimator of the normalized Laplacian,

$$\widehat{\mathcal{L}} \stackrel{\text{def}}{=} \sum_{q=1}^{n}\overline{\lambda}_q\psi_q\psi_q^T. \tag{13}$$

Because the ψ_q are Soules vectors, the spectrum of $\widehat{\mathcal{L}}$ always coincides with $\begin{bmatrix}\overline{\lambda}_1 & \dots & \overline{\lambda}_n\end{bmatrix}^T$. We can prove more, as explained in the next lemma. When the graphs are sampled from of a balanced SBM where $p_1 = \cdots = p_M$ and in the limit of large graph size (or when $\widehat{\mathbb{E}}_N\left[\lambda(\mathcal{L})\right] \approx \mathbb{E}[\lambda(\mathcal{L})]$), then $\widehat{\mathcal{L}}$ is equal to $\mathcal{L}(P)$, the normalized Laplacian associated with the edge probability matrix P.

4.2.1 Theoretical guarantees for the reconstruction.

Lemma 4. *Let P be the edge probability matrix of a balanced* SBM (p, q, n) *where $p_1 = \cdots = p_M = p$,*

$$P = \sum_{m=1}^{M}(p-q)1_{B_m}1_{B_m}^T + qJ. \tag{14}$$

Then, the estimator $\widehat{\mathcal{L}}$ defined in (13) is given by

$$\widehat{\mathcal{L}} = \mathrm{Id} - \left\{\frac{(p-q)}{p+(M-1)q}\Big(\sum_{m=1}^{M}\psi_m\psi_m^T\Big) + \frac{Mq}{p+(M-1)q}\psi_1\psi_1^T\right\}, \tag{15}$$

and therefore $\widehat{\mathcal{L}} = \mathcal{L}(P)$.

The proof of lemma 4, which is provided in section8.2, relies on estimates for the dominant eigenvalues of the symmetric normalized adjacency matrix $\widehat{A}$ when A is the adjacency matrix of a graph sampled from a balanced SBM (p, q, n) with M blocks [22].

4.3 A partial reconstruction

In practice, the estimator of the normalized Laplacian given by $\widehat{\mathcal{L}}$ in (13) is very poor. This numerical problem is perfectly natural: the geometry and edge density of the SBM is unfortunately encoded by the smallest eigenvalues of $\mathcal{L}$ (see lemma 3). The full expansion provided by (13) is plagued by the largest eigenvalues of $\mathcal{L}$, which come from the bulk created by the stochastic nature of the model. This issue is exacerbated by the fact that the high frequency eigenvectors (ψ_l with large l) have small support and therefore are localized around fine scale random structures present in the sample mean adjacency matrix, $\widehat{\mathbb{E}}_N[A]$.

The expression (15) suggests that $\mathcal{L}(P)$ depends only on the first M eigenvectors of the Soules basis. Therefore we propose the following estimator of the symmetric normalized adjacency matrix, $\widehat{A}(P)$,

$$\widetilde{A} \stackrel{\text{def}}{=} \frac{(p-q)}{p+(M-1)q}\Big(\sum_{m=1}^{M}\psi_m\psi_m^T\Big) + \frac{Mq}{p+(M-1)q}\psi_1\psi_1^T, \tag{16}$$

so that $\mathcal{L} = \mathrm{Id} - \widetilde{A}$. As is shown in lemma 4, $\widetilde{A} = \widehat{A}(P)$ when the graph is a balanced SBM (p, q, n) with $p_1 = \cdots = p_M = p$. For a general SBM, one therefore proposes to estimate $\mathcal{L}$ using a truncated reconstruction given

by

$$\widehat{\mathcal{L}_M} \stackrel{\text{def}}{=} \sum_{q=1}^{M} \overline{\lambda}_q \psi_q \psi_q^T. \tag{17}$$

In practice, one needs to estimate M, the number of eigenvalues outside the bulk. Fortunately, many estimators are available (e.g. [7, 14, 21, 39], and references therein).

4.3.1 Theoretical guarantees for the reconstruction.

In the case of a balanced SBM (p, q, n) where $p_1 = \cdots = p_M = p$, a simple calculation reveals that

$$\widetilde{A} = E_M - \widehat{\mathcal{L}_M}. \tag{18}$$

We recall (see lemma 3) that E_M is zero outside of the blocks $B_k \times B_k$ of the SBM. We can therefore interpret (18) as providing the reconstruction of $\widehat{A}(P)$ given by (16) inside the blocks of the SBM, once the offset given by E_M has been removed.

5 The reconstruction of the adjacency matrix

5.1 An estimator of the degree matrix

In order to recover the expected adjacency matrix, P, one needs an estimate, $\widehat{D}$, of the degree matrix to compute $\widehat{D}^{1/2}\widetilde{A}\widehat{D}^{1/2}$. Lemma 3 yields an estimate of the location of the blocks in the SBM. Our estimate of the degree matrix, $\widehat{D}$, is computed by averaging the degrees of all the nodes in each block of the sample mean adjacency matrix, $\widehat{\mathbb{E}}_N[A]$

$$\widehat{d}_i \stackrel{\text{def}}{=} \sum_{j \in J_m} \left[\widehat{\mathbb{E}}_N[A]\right]_{ij} \quad \text{if } i \in J_m,\ 1 \leqslant m \leqslant M, \tag{19}$$

The estimator $\widehat{D}$ proves to be extremely precise.

5.2 An estimator of the adjacency matrix

We combine $\widetilde{A}$ given by (16) with $\widehat{D}$ given by (19), to derive an estimator of the edge probability matrix,

$$\widehat{P} \stackrel{\text{def}}{=} \widehat{D}^{1/2}\widetilde{A}\widehat{D}^{1/2}. \tag{20}$$

6 Experiments

We compare our theoretical analysis to finite sample estimates, which were computed using numerical simulations. The software used to conduct the experiments is publicly available [24]. All graphs were generated using the SBM $(p, q, , n)$ model.

We first illustrate the selection of the Soules vectors (guaranteed by lemma 3). For this experiment, we use $M = 4$ communities of sizes $67, 133, 71, 241$ (see Fig. 3-left). The edge probabilities were given by $p_i = c_i(\log n)^2/n$, where the scaling factor c_i was chosen randomly in $[1, 4]$, and $q = 2\log n/n$. For this experiment, we consider a single realisation of the SBM ($N = 1$). This is clearly the least favorable scenario, where we expect that the estimation of the Soules basis is the most challenging. As shown in Fig. 3-right, the first three non trivial Soules vectors accurately detected the boundaries between the blocks, in spite of the very low contrast between the communities (see Fig. 3-left). This numerical evidence supports the theoretical analysis of lemma 3. We then evaluated the accuracy of (20).

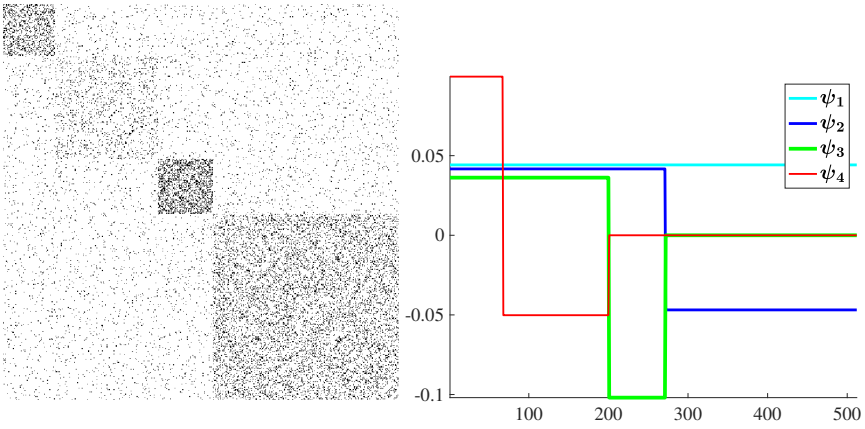

Figure 3: Left: a random realization of the SBM (p, q, n) model. We have $M = 4$ communities of sizes $67, 133, 71, 241$, the graph size is $n = 512$; the edge probability within community i was $p_i \propto (\log n)^2/n$, The edge probability across communities was $q = 2\log n/n$. Right: the first four trivial Soules vectors accurately detected the boundaries between the blocks, in spite of the very low contrast between the communities.

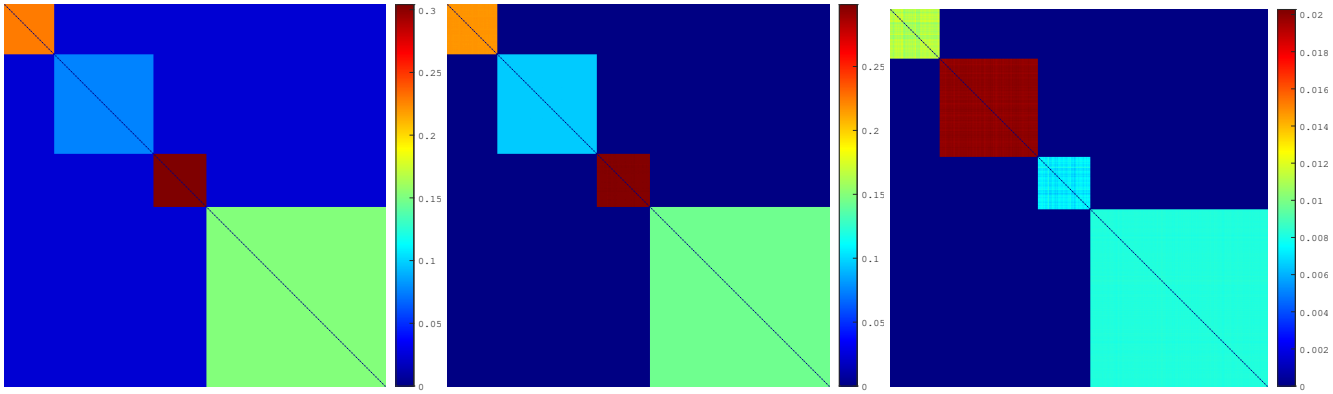

Figure 4: Left: original edge probability matrix P; center: the adjacency matrix of the barycentre graph $\widehat{P}$, given by (20); right:the residual error between P and $\widehat{P}$. The mean absolute error was $n^{-2}\left\|P - \widehat{P}\right\|_1 = 1.01e-05$.

Figure 4 displays the original edge probability matrix P (left), the reconstructed adjacency matrix, $\widehat{P}$ (center), using the top 4 Soules vectors, and the residual error $P - \widehat{P}$ (right). The mean absolute error, defined by

$$n^{-2}\left\|P - \widehat{P}\right\|_1 \stackrel{\text{def}}{=} \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|p_{ij} - \widehat{p}_{ij}\right|, \tag{21}$$

was $n^{-2}\left\|P - \widehat{P}\right\|_1 = 1.01e-05$.

Next we studied the effect of the graph size, n, on the mean absolute reconstruction error (see Fig. 5-left). We rescaled the $M = 4$ SBM model described in the previous paragraph, keeping the relative sizes of the communities the same, and increased the graph size from $n = 100$ to $n = 1,000$. For each n, we computed the mean absolute reconstruction error. As expected, the error decreases as a function of n. We found $n^{-2}\left\|P - \widehat{P}\right\|_1 \propto n^{-1.7}$. This experiment validates the theoretical derivations that were obtained in the limit of large graph sizes, when some concentration phenomenon is in effect, and we can replace $\widehat{\mathbb{E}}_N[A]$ with $\mathbb{E}[A] = P$ in our analysis of algorithm 1 (see section 4.2.1). The next experiment illustrates the effect of the number of blocks M in a balanced SBM (p, q, n) when the edge probabilities are equal, $p_1 = \cdots = p_M$. As is shown in (48), when M becomes large, then the first $M - 1$ non trivial eigenvalues λ_m of $\mathcal{L}$, converge to 1. Because these eigenvalues are no longer separated from the bulk, the truncated reconstruction (16) becomes numerically unstable, and the reconstruction error increases (see Fig. 5-right).

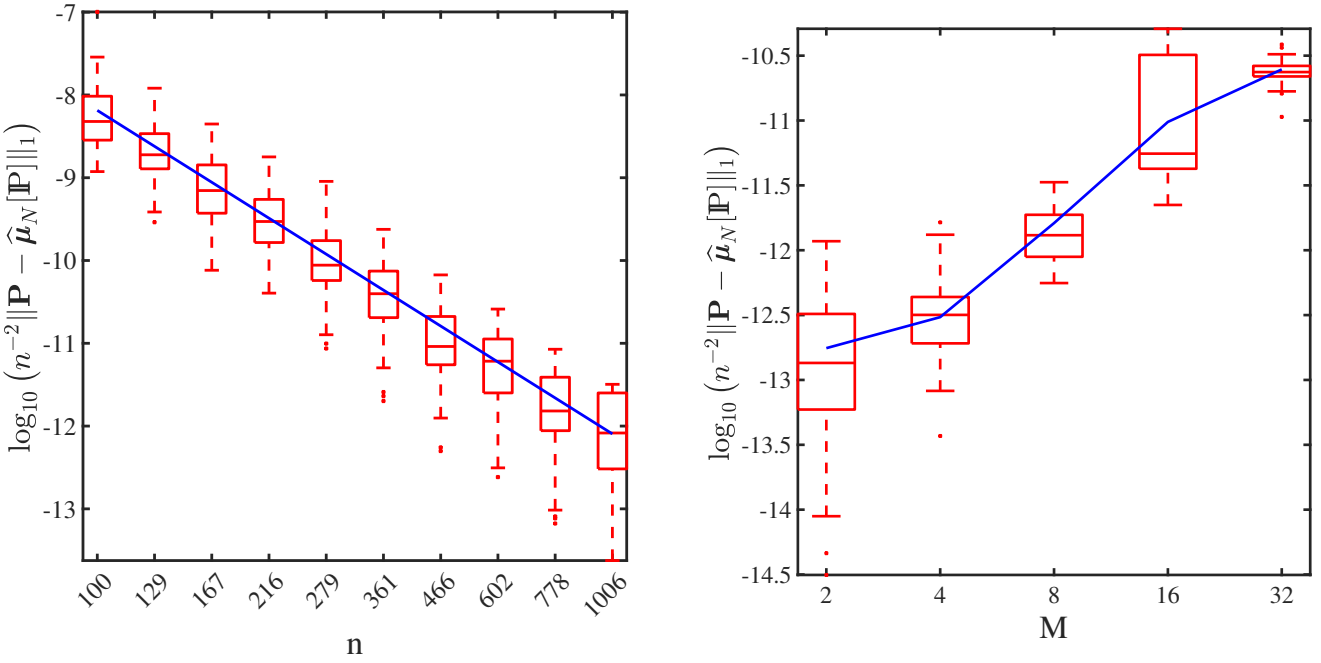

Figure 5: Left: mean absolute error $n^{-2}\left\|\mathbf{P}-\widehat{\mathbf{P}}\right\|_1$ as a function of the graph size, n. The graph is composed of $M=4$ communities, and is a scaled version of the graph shown in Fig. 4. Right: mean absolute error $n^{-2}\left\|\mathbf{P}-\widehat{\mathbf{P}}\right\|_1$ as a function of the number of blocks, M. Each graph is sampled from a balanced SBM (p,q,n) with M blocks of size n/M; $p_i=3(\log n)^2/n$, $q=2\log n/n$, and $n=512$.

7 Discussion

In this work, we proposed a fast algorithm to compute the barycentre of a set of graphs based on the Laplacian spectral pseudo-distance. An original contribution is an algorithm that explores the large library of Soules bases, and returns a basis that that can be used to construct the normalized Laplacian of a graph, whose eigenvalues are equal to the population mean spectrum. Our method combines the spectral information – provided by the sample mean of the first M nontrivial eigenvalues of the normalized Laplacian of the sample – with structural information given by the coarse scale Soules vectors, which are computed using the sample mean adjacency matrix.

We provided some theoretical guarantees in the context where the graphs are random realizations of stochastic block models. We proved that when the graphs are sampled from a balanced SBM with $p_1=\cdots=p_M$, then $\widehat{\mathcal{L}}=\mathcal{L}(\mathbf{P})$, and our approach reconstructs the edge probability matrix. In addition to the theoretical analysis of the estimator of the barycentre graph, we performed Monte Carlo simulations to validate the theoretical properties of the estimator.

Soules bases can always be used to construct non negative matrices with a prescribed set of eigenvalues. Our work is significant because we not only match the eigenvalues, but the coarse scale Soules vectors completely capture the community structure present in the graph. There has not been any work that takes advantage of the binary tree structure of the Soules bases; the present paper offers for the first time a principled method to quickly explore the large library of Soules bases, and construct a Soules basis that provides a sparse approximation of the sample mean adjacency matrix. We expect that this work will open the door to the design of new spectral-based graph synthesis that have theoretical guarantees.

8 Additional proofs

8.1 Proof of lemma 3

The proof of lemma 3 relies on two different results. We first show that a top-down exploration of the Soules binary tree, when the first Soules vector is $\psi_1=n^{-1/2}\mathbf{1}$, always result in a matrix $\mathbf{E}_M=\sum_{q=1}^{M}\psi_q\psi_q^T$ that is piecewise constant

on square blocks aligned along the diagonal, and zero outside of the blocks (see corollary 1). This property only relies on the fact that the sequence of $\boldsymbol{\psi}_m$ have nested supports.

The second result specifically addresses the construction of each $\boldsymbol{\psi}_m$ in algorithm 1. We prove in lemma 8 that at each level l, the Soules vector $\boldsymbol{\psi}_l$ returned by algorithm 1 is aligned with the boundary of a block B_m of the edge probability matrix $\mathbf{P}$. At level M, algorithm 1 has discovered all the M blocks. In the process, we prove several technical lemmata.

8.1.1 The tensor product $\boldsymbol{\psi}_l\boldsymbol{\psi}_l$

The first lemma is an elementary calculation that gives the expression of $\boldsymbol{\psi}_l\boldsymbol{\psi}_l^T$

Lemma 5. *We choose $\boldsymbol{\psi}_1 \stackrel{def}{=} n^{-1/2}\mathbf{1}$, and denote by $\boldsymbol{\psi}_l$ the Soules vector returned by algorithm 1 at level l. Let $\operatorname{supp}(\boldsymbol{\psi}_l) = [i_0, i_1]$, and let k be the location of the split in $[i_0, i_1]$ such that $\boldsymbol{\psi}_l\big|_{[i0,k]} > 0$ and $\boldsymbol{\psi}_l\big|_{[k+1,i_1]} < 0$ (see Fig. 1). Then,*

$$\boldsymbol{\psi}_l\boldsymbol{\psi}_l^T(i,j) = \frac{1}{i_1 - i_0 + 1}\begin{cases} \dfrac{i_1 - k}{k - i_0 + 1} & \text{if } i_0 \leqslant i,j \leqslant k \\ \dfrac{k - i_0 + 1}{i_1 - k} & \text{if } k+1 \leqslant i,j \leqslant i_1 \\ -1 & \text{if } \begin{cases} i_0 \leqslant i \leqslant k, \quad k+1 \leqslant j \leqslant i_1, \\ k+1 \leqslant i \leqslant i_1, \quad i_0 \leqslant j \leqslant k, \end{cases} \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

Proof. The proof is an elementary calculation based on the definition of $\boldsymbol{\psi}_l$ given by (4), and the observation that $\boldsymbol{\psi}_1(i) = n^{-1/2}$. Indeed, we know from (4) that $\boldsymbol{\psi}_l$ is piecewise constant, and given by

$$\boldsymbol{\psi}_l(i) = \frac{1}{\sqrt{i_1 1 - i_0 + 1}}\begin{cases} \dfrac{\sqrt{i_1 - k}}{\sqrt{k - i_0 + 1}} & \text{if } i_0 \leqslant i \leqslant k, \\ -\dfrac{\sqrt{k - i_0 + 1}}{\sqrt{i_1 - k}} & \text{if } k+1 \leqslant i \leqslant i_1, \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

The computation of the tensor product is immediate and yields the advertised result. □

8.1.2 The matrix E_M

In the following corollary, we describe the matrix $E_M = \sum_{m=1}^{M} \boldsymbol{\psi}_m\boldsymbol{\psi}_m^T$. When combined with lemma 8, we use this corollary to reconstruct the geometry of the blocks in the SBM.

Corollary 1. *Let J_m be the leaves in the binary Soules tree (these are intervals that are no longer split) after M steps of algorithm 1. Then $E_M \stackrel{def}{=} \sum_{m=1}^{M} \boldsymbol{\psi}_m\boldsymbol{\psi}_m^T$ is equal to*

$$e_M(i,j) = \begin{cases} \dfrac{1}{|J_l|} & \text{if } (i,j) \in J_l \times J_l, \\ 0 & \text{otherwise.} \end{cases} \tag{24}$$

Also,

$$\begin{cases} \sum_{m=2}^{M} \boldsymbol{\psi}_m\boldsymbol{\psi}_m^T(i,j) > 0 & \text{if } \exists q \in \{, 2 \ldots, M\}, \ (i,j) \in J_q \times J_q \\ \sum_{m=2}^{M} \boldsymbol{\psi}_m\boldsymbol{\psi}_m^T(i,j) < 0 & \text{otherwise} \end{cases} \tag{25}$$

Proof. We first observe that after M iterations of algorithm 1 there are M intervals J_m that are not split (the leaves in the binary tree shown in Fig. 2), where we count the construction of ψ_1 as the first iteration of the algorithm ($M = 1$). This can be proved by induction, after observing that an iteration of algorithm 1, described by (4), turns exactly one leaf in the tree into two leaves.

Next, we prove that E_M is nonnegative on each $J_l \times J_l$, $1 \leqslant l \leqslant M$. Since, each interval J_l is a leaf of the tree, the interval J_l is not further decomposed, and there exists a vector ψ_q such that $\psi_q\big|_{J_l} > 0$ or $\psi_q\big|_{J_l} < 0$ (see Fig. 2-right). We can therefore apply lemma 5, with $J_l = [i_0, k]$ or $J_l = [k, i_1]$, and $\psi_q \psi_q^T$ is constant on $J_l \times J_l$ (see (22)). All other vectors larger scale ψ_m such that $J_l \subset \operatorname{supp}(\psi_m)$, also keep a constant value on J_l, and therefore $\psi_m \psi_m^T$ is constant on $J_l \times J_l$. We conclude that E_M is constant on each $J_l \times J_l$, $1 \leqslant l \leqslant M$.

We can then prove by induction that

$$e_M(i,j) = \begin{cases} \dfrac{1}{|J_l|} & \text{if } (i,j) \in J_l \times J_l, \\ 0 & \text{otherwise.} \end{cases} \tag{26}$$

For $M = 1$ there is nothing to prove, since $\psi_1 = n^{-1/2}\mathbf{1}$, so $E_1 = n^{-1}J$. Now, assume that (26) holds for $M \geqslant 1$, then $E_{M+1} = E_M + \psi_{M+1} \times \psi_{M+1}$, and ψ_{M+1} is created by splitting an interval J_q, so there exists $q \in \{1, \ldots, M\}$, such that $\operatorname{supp}(\psi_{M+1}) = J_q$, and $\psi_{M+1}\big|_{J_m} = 0$ for all $m \neq q$. Since J_q is the only block that changes when going from M to $M+1$, E_{M+1} is equal to E_M on all the other blocks. Using the induction hypothesis, we then have for all $m \neq q$,

$$\forall (i,j) \in J_m \times J_m, \quad e_{M+1}(i,j) = e_M(i,j) = \frac{1}{|J_m|}. \tag{27}$$

We are left with the computation of E_{M+1} on J_q. Let us define i_0 and i_1 such that $J_q = [i_0, i_1]$, and let k be the index where J_q is split, $J_q = [i_0, k] \cup [k+1, i_1]$. Then using lemma 5 we have for all $(i,j) \in J_q \times J_q$,

$$\psi_{M+1} \times \psi_{M+1}(i,j) = \frac{1}{i_1 - i_0 + 1} \begin{cases} \frac{i_1 - k}{k - i_0 + 1} & \text{if } (i,j) \in [i_0, k] \times [i_0, k], \\ \frac{k - i_0 + 1}{i_1 - k} & \text{if } (i,j) \in [k+1, i_1] \times [k+1, i_1], \\ -1 & \text{otherwise.} \end{cases} \tag{28}$$

From the induction hypothesis, we have for all $(i,j) \in J_q \times J_q$, $e_M(i,j) = |i_1 - i_0 + 1|^{-1}$. Adding E_M and $\psi_{M+1} \times \psi_{M+1}$ yields for all $(i,j) \in J_q \times J_q$,

$$e_{M+1}(i,j) = \begin{cases} \frac{1}{k - i_0 + 1} & \text{if } (i,j) \in [i_0, k] \times [i_0, k], \\ \frac{1}{i_1 - k} & \text{if } (i,j) \in [k+1, i_1] \times [k+1, i_1], \\ 0 & \text{otherwise,} \end{cases} \tag{29}$$

which concludes the case for $M + 1$. By induction, (26) holds for all M.

We conclude the proof of corollary 1 by proving (25). Let $(i,j) \in \{1, \ldots, n\} \times \{1, \ldots, n\}$. If $\exists q \in \{1, \ldots, M\}$, such that (i,j) is in block $J_q \times J_q$ then $e_M(i,j) = |J_q|^{-1}$. Also, $\psi_\times \psi_1(i,j) = n^{-1}$, and thus

$$\sum_{m=2}^{M} \psi_m \psi_m^T(i,j) = \frac{1}{|J_q|} - \frac{1}{n} > 0, \tag{30}$$

since $|J_q| > 1$. Now, if (i,j) is not in any blocks $J_q \times J_q$, then $e_M(i,j) = 0$, and therefore $e_M(i,j) - \psi_1 \times \psi_1(i,j) = -n^{-1} < 0$. □

We now prove a series of lemmata that address the performance of algorithm 1 and its ability to detect the blocks of an SBM by aligning the successive $\boldsymbol{\psi}_m$ with the block boundaries. The proof hinges on the study of one iteration of algorithm 1, as explained in lemma 6. The proof of lemma 6 is a simple calculation that relies on the fact that both $\boldsymbol{\psi}_l\boldsymbol{\psi}_l^T$ and $\mathbf{P}$ are piecewise constant over $[i_0, i_1] \times [i_0, i_1]$. Then, $|\langle \boldsymbol{\psi}_l\boldsymbol{\psi}_l^T, \mathbf{P}\rangle|^2$ is maximum if the location of the zero-crossing of $\boldsymbol{\psi}_l$ is equal to the location of the jump in the SBM, $k = j$ (see Fig. 6). □

8.1.3 One iteration of algorithm 1

The next lemma studies a single iteration of algorithm 1, which leads to the construction of the Soules vector $\boldsymbol{\psi}_l$. We assume that $\operatorname{supp}(\boldsymbol{\psi}_l) = [i_0, i_1]$, and we consider the matrix $\mathbf{P}$ that is nonzero only on $[i_0, i_1] \times [i_0, i_1]$, and is piecewise constant on two blocks $J_0 \times J_0$ and $J_1 \times J_1$, where $J_0 = [i_0, j]$, and $J_1 = [j+1, i_1]$ (see Fig. 6),

$$\mathbf{P} = p_0(\mathbf{1}_{J_0}\mathbf{1}_{J_0}^T) + p_1(\mathbf{1}_{J_1}\mathbf{1}_{J_1}^T) + q(\mathbf{1}_{J_0}\mathbf{1}_{J_1}^T + \mathbf{1}_{J_1}\mathbf{1}_{J_0}^T). \tag{31}$$

We prove that in order to maximize $|\langle \boldsymbol{\psi}_l\boldsymbol{\psi}_l^T, \mathbf{P}\rangle|^2$, algorithm 1 must always align k (the zero-crossing of $\boldsymbol{\psi}_l$) with the jump in the SBM inside $\operatorname{supp}(\boldsymbol{\psi}_l\boldsymbol{\psi}_l^T)$ (see Fig. 6).

Lemma 6. *Let $\boldsymbol{\psi}_l$ be the Soules vector returned by algorithm 1 at level l with support $\operatorname{supp}(\boldsymbol{\psi}_l) \stackrel{def}{=} [i_0, i_1]$. We consider the matrix $\mathbf{P}$ that is nonzero only on $[i_0, i_1] \times [i_0, i_1]$, and is piecewise constant on two blocks $J_0 \times J_0$ and $J_1 \times J_1$, where $J_0 = [i_0, j]$, and $J_1 = [j+1, i_1]$ (see Fig. 6),*

$$\mathbf{P} = p_0(\mathbf{1}_{J_0}\mathbf{1}_{J_0}^T) + p_1(\mathbf{1}_{J_1}\mathbf{1}_{J_1}^T) + q(\mathbf{1}_{J_0}\mathbf{1}_{J_1}^T + \mathbf{1}_{J_1}\mathbf{1}_{J_0}^T). \tag{32}$$

Then, $|\langle \boldsymbol{\psi}_l\boldsymbol{\psi}_l^T, \mathbf{P}\rangle|^2$ is maximum if the location of the zero-crossing of $\boldsymbol{\psi}_l$ is equal to the location of the jump in the SBM, $k = j$ (see Fig. 6).

Proof. The proof relies on the computation of the inner-product between a Soules tensor product $\boldsymbol{\psi}_l\boldsymbol{\psi}_l^T$ and an SBM whose support coincide with the support of $\boldsymbol{\psi}_l\boldsymbol{\psi}_l^T$.

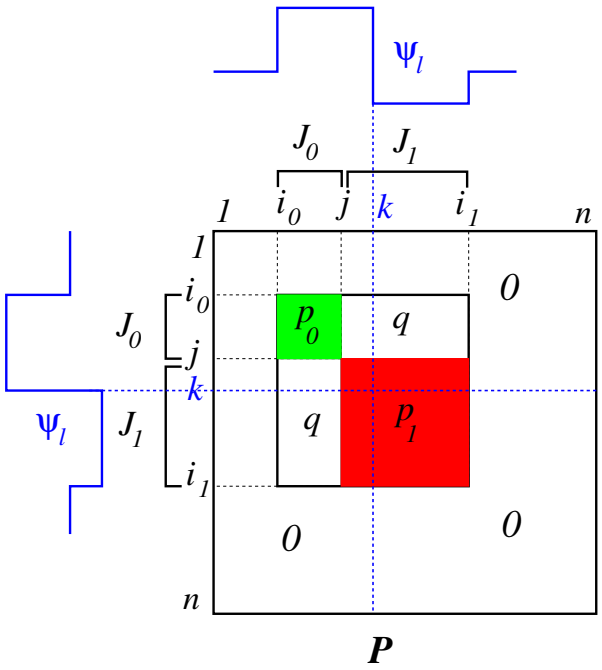

Figure 6: The vector $\boldsymbol{\psi}_l$ (in blue) is created by splitting a block of indices $I = [i_0, i_1]$ at level $l-1$ into two sub-blocks, $[i_0, k] \cup [k+1, i_1]$ at level l. We consider the matrix $\mathbf{P}$ that is nonzero only on $[i_0, i_1] \times [i_0, i_1]$, and is piecewise constant on two blocks $J_0 \times J_0$ (in green) and $J_1 \times J_1$ (in red), where $J_0 = [i_0, j]$, and $J_1 = [j+1, i_1]$

We use lemma 5, and we study two cases for the choice of $k \in [i_0, i_1]$. We have

$$\langle \psi_l \psi_l^T, P \rangle = p_0 \langle \psi_l \psi_l^T, \mathbf{1}_{J_0} \mathbf{1}_{J_0} \rangle^T + p_1 \langle \psi_l \psi_l^T, \mathbf{1}_{J_1} \mathbf{1}_{J_1}^T \rangle + c \langle \psi_l \psi_l^T, \mathbf{1}_{J_0} \mathbf{1}_{J_1}^T + \mathbf{1}_{J_1} \mathbf{1}_{J_0}^T \rangle. \tag{33}$$

Also, $\langle \psi_l \psi_l^T, \mathbf{1}_{J_q} \mathbf{1}_{J_r}^T \rangle = \langle \psi_l, \mathbf{1}_{J_q} \rangle \langle \psi_l, \mathbf{1}_{J_r} \rangle$, for $q, r \in \{0, 1\}$. We define $r_q \stackrel{\text{def}}{=} \langle \psi_l, \mathbf{1}_{J_q} \rangle$ for $q = 0, 1$. Then

$$\langle \psi_l = \psi_l^T, P \rangle = p_0 r_0^2 + 2 q r_0 r_1 + p_1 r_1^2. \tag{34}$$

The expression of the coefficients r_0 and r_1 can be derived by using (23). We give the details for the computation of r_0, the computation of r_1 is very similar. To compute r_0, we need to consider the two cases, $i_0 \leqslant k \leqslant j$ and $j \leqslant k \leqslant i_1$. We recall from (23) that we always have

$$\psi_l(i) = \frac{1}{\sqrt{i_1 - i_0 + 1}} \begin{cases} \dfrac{\sqrt{i_1 - k}}{\sqrt{k - i_0 + 1}} & \text{if } i_0 \leqslant i \leqslant k, \\ -\dfrac{\sqrt{k - i_0 + 1}}{\sqrt{i_1 - k}} & \text{if } k + 1 \leqslant i \leqslant i_1, \\ 0 & \text{otherwise.} \end{cases} \tag{35}$$

If $k \leqslant j$ then ψ_l changes sign over J_0 and we have

$$\begin{aligned} r_0 = \langle \psi_l, \mathbf{1}_{J_0} \rangle &= \frac{1}{\sqrt{i_1 - i_0 + 1}} \left\{ \sum_{i=i_0}^{k} \frac{\sqrt{i_1 - k}}{\sqrt{k - i_0 + 1}} - \sum_{i=k+1}^{j} \frac{\sqrt{k - i_0 + 1}}{\sqrt{i_1 - k}} \right\} \\ &= \sqrt{\frac{(k - i_0 + 1)(i_1 - k)}{i_1 - i_0 + 1}} \left(\frac{i_1 - j}{i_1 - k} \right). \end{aligned} \tag{36}$$

If $j \leqslant k$ then ψ_l is positive over J_0 (this is the case for Fig. 6) and we have

$$r_0 = \langle \psi_l, \mathbf{1}_{J_0} \rangle = \frac{1}{\sqrt{i_1 - i_0 + 1}} \sum_{i=i_0}^{j} \frac{\sqrt{i_1 - k}}{\sqrt{k - i_0 + 1}} = \sqrt{\frac{(k - i_0 + 1)(i_1 - k)}{i_1 - i_0 + 1}} \left(\frac{j - i_0 + 1}{k - i_0 + 1} \right). \tag{37}$$

A similar calculation yields r_1. If $k + 1 \leqslant j + 1$ then ψ_l is negative over J_1 and we have

$$r_1 = -\sqrt{\frac{(k - i_0 + 1)(i_1 - k)}{i_1 - i_0 + 1}} \left(\frac{i_1 - j}{i_1 - k} \right), \tag{38}$$

and if $j + 1 \leqslant k + 1$ (this is the case for Fig. 6) then ψ_l changes sign over J_1 and we have

$$r_1 = -\sqrt{\frac{(k - i_0 + 1)(i_1 - k)}{i_1 - i_0 + 1}} \left(\frac{j - i_0 + 1}{k - i_0 + 1} \right). \tag{39}$$

We are now ready to evaluate $\langle \psi_l \psi_l^T, P \rangle = p_0 r_0^2 + 2 q r_0 r_1 + p_1 r_1^2$. Again, we need to consider the following two cases. If $k \leqslant j$ then

$$\begin{aligned} \langle \psi_l \psi_l^T, P \rangle &= p_0 \frac{(k - i_0 + 1)(i_1 - k)}{i_1 - i_0 + 1} \left(\frac{i_1 - j}{i_1 - k} \right)^2 + p_1 \frac{(k - i_0 + 1)(i_1 - k)}{i_1 - i_0 + 1} \left(\frac{i_1 - j}{i_1 - k} \right)^2 \\ &\quad - 2q \frac{(k - i_0 + 1)(i_1 - k)}{i_1 - i_0 + 1} \left(\frac{i_1 - j}{i_1 - k} \right)^2 \\ &= \frac{(k - i_0 + 1)(i_1 - k)}{i_1 - i_0 + 1} \left(\frac{i_1 - j}{i_1 - k} \right)^2 \{ p_0 + p_1 - 2q \}, \end{aligned} \tag{40}$$

which is maximum when $k = j$. In the case where if $j \leqslant k$ we have

$$\langle \psi_l \psi_l^T, P \rangle = \frac{(k - i_0 + 1)(i_1 - k)}{i_1 - i_0 + 1} \left(\frac{j - i_0 + 1}{k - i_0 + 1} \right)^2 \{p_0 + p_1 - 2q\}, \tag{41}$$

which is also maximum when $k = j$. This concludes the proof that $\langle \psi_l \psi_l^T, P \rangle$ is maximal if $k = j$. □

Lemma 7 extends lemma 6 to the general edge probability matrix P of an SBM (see (10)); it is used to prove lemma 8 by induction. Lemma 7 can be proved using a proof by contradiction (using lemma 6).

Lemma 7. *Let P be the population mean adjacency matrix of* $\mathrm{SBM}(p, q, n)$ *defined by*

$$P = \sum_{m=1}^{M} (p_m - q) 1_{B_m} 1_{B_m}^T + qJ, \tag{42}$$

where the M blocks B_m form a partition of $[n]$. Then, the split that creates ψ_2 in algorithm 1, is always located at the boundary between two blocks B_m and B_{m+1}.

Proof. Let k be the index associated with the construction of ψ_2 and the subdivision of $[n]$. We need to prove that k coincides with the endpoint of a block B_m. By contradiction, if k does not correspond to the boundary between two blocks, then there exists $i_0 < i_1$ such that $B_m = [i_0, i_1]$ and $i_0 < k < i_1$. Since P is constant over the block $[i_0, i_1] \times [i_0, i_1]$ (see Fig. 6 with $p_0 = p_1 = q$), lemma 6 tells us that the value of P in $B_m \times B_m$ does not contribute to $|\langle \psi_2 \psi_2^T, P \rangle|^2$, and algorithm 1 should not have placed k in B_m. □

8.1.4 M iterations of algorithm 1

This last lemma guarantees that after M iterations of algorithm 1, the matrix E_M associated with the first M Soules vectors recovers the block geometry. Lemma 8 is proved by induction on M, using lemma 7.

Lemma 8. *Let P be the population mean adjacency matrix of* $\mathrm{SBM}(p, q, n)$ *defined by*

$$P = \sum_{m=1}^{M} (p_m - q) 1_{B_m} 1_{B_m}^T + qJ \tag{43}$$

Let $J_l, 1 \leqslant l \leqslant M$ be the leaves in the binary Soules tree (these are intervals that are no longer split) after M steps of algorithm 1. Then, the M blocks $\{B_m\}$ in (43) coincide with the M intervals $\{J_l\}$ discovered by algorithm 1.

Proof. We prove the result by induction on M. If $M = 1$, there is nothing to prove. If $M = 2$, then lemma 7 shows that ψ_2 recovers the block geometry. We assume that the result holds for all P with $m \leqslant M$ blocks given by (43). We consider the population mean adjacency matrix Q defined by

$$Q = \sum_{m=1}^{M+1} (p_m - q) 1_{C_m} 1_{C_m}^T + qJ, \tag{44}$$

where $\cup_{m=1}^{M+1} C_m = [n]$. Because of lemma 7, the first split of $[n]$, which leads to the construction of ψ_2 is aligned the boundary of a block $C_{m_0} = [i_0, k]$. Without loss of generality, we can assume that the cut is aligned with the endpoint of C_{m_0}. We can then partition $Q = Q_1 + Q_2$, where

$$Q_1 = \sum_{m=1}^{k} (p_m - q) 1_{C_m} 1_{C_m}^T + q 1_{[k]} 1_{[k]}^T \tag{45}$$

and

$$Q_2 = \sum_{m=k+1}^{M+1} (p_m - q) 1_{C_m} 1_{C_m}^T + q 1_{\{k+1,\ldots,n\}} 1_{\{k+1,\ldots,n\}}^T. \tag{46}$$

Again, because of lemma 7, the next splits happen (independently) in $\mathbf{Q}_1$, or $\mathbf{Q}_2$. We can use the induction hypothesis to argue that all further splits will be located along the blocks in $\mathbf{Q}_1$, or $\mathbf{Q}_2$. After M splits, the algorithm has detected all $M+1$ blocks. By induction, the result holds for all M. □

Lemma 3 is then a direct consequence of lemma 8 and corollary 1.

8.2 Proof of lemma 4

Proof. Let $\mathbf{P}$ be the edge probability matrix of a balanced SBM (p, q, n) where $p_1 = \cdots = p_M = p$. The authors in [22] provide the following estimates of the eigenvalues $\lambda(\widehat{\mathbf{A}})$ of the normalized adjacency matrix, $\widehat{\mathbf{A}} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$,

Lemma 9 (Proposition 4.3 of [22]). *The M largest eigenvalues of $\widehat{\mathbf{A}}$ are given by*

$$\lambda_m(\widehat{\mathbf{A}}) = \left\{ \begin{array}{c} 1 \text{ if } m = 1, \\ \dfrac{p-q}{p+(M-1)q} \text{ if } 2 \leqslant m \leqslant M, \\ 0 \text{ if } M+1 \leqslant m \leqslant n, \end{array} \right\} + \mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right), \tag{47}$$

in the limit of large n, almost surely [22]. Neglecting the $\mathcal{O}\left(\sqrt{\log n/n}\right)$ terms the eigenvalues of $\mathcal{L}$ are given by

$$\lambda_m(\mathcal{L}) = \begin{cases} 0 & \text{if } m = 1, \\ \dfrac{Mq}{p+(M-1)q} & \text{if } 2 \leqslant m \leqslant M, \\ 1 & \text{if } M+1 \leqslant m \leqslant n, \end{cases} \tag{48}$$

in the limit of large n. Substituting the expression of $\lambda_q(\mathcal{L})$ given by (48) for $\overline{\lambda}_q$ in (13), we get

$$\begin{aligned} \widehat{\mathcal{L}} &= \sum_{q=1}^{n} \overline{\lambda}_q \psi_q \psi_q^T = \sum_{q=2}^{M} \overline{\lambda}_q \psi_q \psi_q^T + \sum_{q=M+1}^{n} \overline{\lambda}_q \psi_q \psi_q^T \\ &= \left(1 - \frac{p-q}{p+(M-1)q}\right) \sum_{q=2}^{M} \psi_q \psi_q^T + \sum_{q=M+1}^{n} \psi_q \psi_q^T \\ &= \sum_{q=1}^{n} \psi_q \psi_q^T - \left\{ \frac{(p-q)}{p+(M-1)q} \Big(\sum_{m=1}^{M} \psi_m \psi_m^T \Big) + \frac{Mq}{p+(M-1)q} \psi_1 \psi_1^T \right\} \\ &= \mathrm{Id} - \left\{ \frac{(p-q)}{p+(M-1)q} \Big(\sum_{m=1}^{M} \psi_m \psi_m^T \Big) + \frac{Mq}{p+(M-1)q} \psi_1 \psi_1^T \right\} \\ &= \mathrm{Id} - \left\{ \frac{(p-q)}{p+(M-1)q} E_M + \frac{Mq}{p+(M-1)q} \psi_1 \psi_1^T \right\}, \end{aligned} \tag{49}$$

in the limit of large n. In the case of a balanced SBM, we have from (24),

$$e_M(i,j) = \begin{cases} \dfrac{M}{n} & \text{if } (i,j) \in J_l \times J_l, \\ 0 & \text{otherwise.} \end{cases} \tag{50}$$

We conclude that

$$\widehat{\mathcal{L}}_{ij} = \begin{cases} 1 - \dfrac{pM}{n(p+(M-1)q)} & \text{if } i = j \\ -\dfrac{pM}{n(p+(M-1)q)} & \text{if } (i,j) \in J_l \times J_l, i \neq j, \\ \dfrac{qM}{n(p+(M-1)q)} & \text{otherwise.} \end{cases} \tag{51}$$

We conclude the proof by computing the normalized Laplacian associated with $\mathbf{P}$, we have

$$\mathcal{L}(\mathbf{P}) = \mathrm{Id} - \frac{M}{n(p+(M-1)q)}\mathbf{P} \tag{52}$$

where we neglected p in the computation of the degree matrix. Whence,

$$\mathcal{L}(\mathbf{P})_{ij} = \begin{cases} 1 - \dfrac{pM}{n(p+(M-1)q)} & \text{if } i = j \\ -\dfrac{pM}{n(p+(M-1)q)} & \text{if } (i,j) \in J_l \times J_l, i \neq j, \\ \dfrac{qM}{n(p+(M-1)q)} & \text{otherwise.} \end{cases} \tag{53}$$

We conclude that $\widehat{\mathcal{L}}_{ij} = \mathcal{L}(\mathbf{P})_{ij}$. □

References

[1] Emmanuel Abbe, *Community detection and stochastic block models: recent developments*, Journal of Machine Learning Research **18** (2018), no. 177, 1–86.

[2] Edo M Airoldi, Thiago B Costa, and Stanley H Chan, *Stochastic blockmodel approximation of a graphon: Theory and consistent estimation*, Advances in Neural Information Processing Systems, 2013, pp. 692–700.

[3] Luca Baldesi, Athina Markopoulou, and Carter T Buttsc, *Spectral graph forge: A framework for generating synthetic graphs with a target modularity*, IEEE/ACM Transactions on Networking **27** (2019), no. 5, 2125–2136.

[4] Moïse Blanchard and Adam Quinn Jaffe, *Fréchet mean set estimation in the Hausdorff metric, via relaxation*, Bernoulli **31** (2025), no. 1, 432 – 456.

[5] Stanley Chan and Edoardo Airoldi, *A consistent histogram estimator for exchangeable graph models*, International Conference on Machine Learning, PMLR, 2014, pp. 208–216.

[6] Antoine Channarond, Jean-Jacques Daudin, and Stéphane Robin, *Classification and estimation in the Stochastic Blockmodel based on the empirical degrees*, Electronic Journal of Statistics **6** (2012), 2574 – 2601.

[7] Yi Yu D. Franco Saldaña and Yang Feng, *How many communities are there?*, Journal of Computational and Graphical Statistics **26** (2017), no. 1, 171–181.

[8] Shaofeng Deng, Shuyang Ling, and Thomas Strohmer, *Strong consistency, graph laplacians, and the stochastic block model*, Journal of Machine Learning Research **22** (2021), no. 117, 1–44.

[9] Karel Devriendt, Renaud Lambiotte, and Piet Van Mieghem, *Constructing Laplacian matrices with Soules vectors*, arXiv preprint arXiv:1909.11282 (2019).

[10] Claire Donnat and Susan Holmes, *Tracking network dynamics: A survey using graph distances*, The Annals of Applied Statistics **12** (2018), no. 2, 971–1012.

[11] Paromita Dubey and Hans-Georg Müller, *Fréchet change-point detection*, The Annals of Statistics **48** (2020), no. 6, 3312–3335.

[12] Ludwig Elsner, Reinhard Nabben, and Michael Neumann, *Orthogonal bases that lead to symmetric nonnegative matrices*, Linear Algebra and its Applications **271** (1998), no. 1-3, 323–343.

[13] SD Eubanks and Judith J McDonald, *On a generalization of soules bases*, SIAM journal on matrix analysis and applications **31** (2010), no. 3, 1227–1234.

[14] Xinjie Fan, Yuguang Yue, Purnamrita Sarkar, and Y. X. Rachel Wang, *On hyperparameter tuning in general clustering problems*, Proceedings of the 37th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 119, 2020, pp. 2996–3007.

[15] Daniel Ferguson and François G Meyer, *Theoretical analysis and computation of the sample Fréchet mean of sets of large graphs*, Information and Inference **12** (2023), no. 3, 1347–1404.

[16] Miquel Ferrer, Francesc Serratosa, and Alberto Sanfeliu, *Synthesis of median spectral graph*, Pattern Recognition and Image Analysis, 2005, pp. 139–146.

[17] Thorben Funke and Till Becker, *Stochastic block models: A comparison of variants and inference methods*, PLOS ONE **14** (2019), no. 4, 1–40.

[18] Debarghya Ghoshdastidar, Maurilio Gutzeit, Alexandra Carpentier, and Ulrike Von Luxburg, *Two-sample hypothesis testing for inhomogeneous random graphs*, The Annals of Statistics **48** (2020), no. 4, 2208–2229.

[19] Isabel Haasler and Pascal Frossard, *Bures-wasserstein means of graphs*, International Conference on Artificial Intelligence and Statistics, PMLR, 2024, pp. 1873–1881.

[20] Eric D Kolaczyk, Lizhen Lin, Steven Rosenberg, Jackson Walters, and Jie Xu, *Averages of unlabeled networks: Geometric characterization and asymptotic behavior*, The Annals of Statistics **48** (2020), no. 1, 514–538.

[21] Can M. Le and Elizaveta Levina, *Estimating the number of communities by spectral methods*, Electronic Journal of Statistics **16** (2022), no. 1, 3315 – 3342.

[22] Matthias Löwe and Sara Terveer, *Hitting times for random walks on the stochastic block model*, arXiv preprint arXiv:2401.07896 (2024), 1–26.

[23] Simón Lunagómez, Sofia C Olhede, and Patrick J Wolfe, *Modeling network populations via graph distances*, Journal of the American Statistical Association **116** (2021), no. 536, 2023–2040.

[24] François G. Meyer, *The spectral barycentre network*, https://github.com/francoismeyer/barycentre-network, 2025.

[25] François G. Meyer, *When does the mean network capture the topology of a sample of networks?*, Frontiers in Physics **12** (2024), 1–11.

[26] Sofia C Olhede and Patrick J Wolfe, *Network histograms and universality of blockmodel approximation*, PNAS **111** (2014), no. 41, 14722–14727.

[27] Hazel Perfect and Leon Mirsky, *Spectral properties of doubly-stochastic matrices*, Monatshefte für Mathematik **69** (1965), 35–57.

[28] Alexander Petersen and Hans-Georg Müller, *Fréchet regression for random objects with euclidean predictors*, The Annals of Statistics **47** (2019), no. 2, 691–719.

[29] Ievgen Redko, Marc Sebban, and Amaury Habrard, *Non-negative matrix factorization meets time-inhomogeneous markov chains*, OPT2020, 2020.

[30] Hannu Reittu, Ilkka Norros, Tomi Räty, Marianna Bolla, and Fülöp Bazsó, *Regular decomposition of large graphs: Foundation of a sampling approach to stochastic block model fitting*, Data Science and Engineering **4** (2019), 44–60.

[31] Alana Shine and David Kempe, *Generative graph models based on Laplacian spectra?*, The World Wide Web Conference, ACM, 2019, pp. 1691–1701.

[32] George W Soules, *Constructing symmetric nonnegative matrices*, Linear and Multilinear Algebra **13** (1983), no. 3, 241–251.

[33] Karl-Theodor Sturm, *Probability measures on metric spaces of nonpositive*, Heat kernels and analysis on manifolds, graphs, and metric spaces **338** (2003), 357.

[34] Christoph M Thiele and Lars F Villemoes, *A fast algorithm for adapted time–frequency tilings*, Applied and Computational Harmonic Analysis **3** (1996), no. 2, 91–99.

[35] Edwin R Van Dam and Willem H Haemers, *Developments on spectral characterizations of graphs*, Discrete Mathematics **309** (2009), no. 3, 576–586.

[36] David White and Richard C Wilson, *Spectral generative models for graphs*, ICIAP 2007, 2007, pp. 35–42.

[37] Peter Wills and François G Meyer, *Metrics for graph comparison: a practitioner's guide*, PLoS ONE **15(2)** (2020), 1–54.

[38] Hongtengl Xu, *Gromov-Wasserstein factorization models for graph clustering*, Proceedings of the AAAI conference on artificial intelligence, vol. 34(04), 2020, pp. 6478–6485.

[39] Bowei Yan, Purnamrita Sarkar, and Xiuyuan Cheng, *Provable estimation of the number of blocks in block models*, ICAIS, 2018, pp. 1185–1194.

[40] Jean-Gabriel Young, Guillaume St-Onge, Patrick Desrosiers, and Louis J Dubé, *Universality of the stochastic block model*, Physical Review E **98** (2018), no. 3, 032309.

[41] Daniele Zambon, Cesare Alippi, and Lorenzo Livi, *Change-point methods on a sequence of graphs*, IEEE Transactions on Signal Processing **67** (2019), no. 24, 6327–6341.