Clustering the Nodes of Sparse Edge-Weighted Graphs via Non-Backtracking Spectra

Marianna Bolla, Reittu Hannu, Fatma Abdelkhalek
BME Math. Inst. Budapest, VTT Finland, Assiut Univ. Egypt
marib@math.bme.hu

FICC 2025 Berlin

April 28, 2025.

Outline

- Problem: spectral clustering the nodes of sparse graphs.
- Tool: non-backtracking matrix of simple graphs.
- Our advantages: k-means clustering with node representatives.
- Methods: belief propagation (number of clusters), EM algorithm (estimating the parameters).
- Sparse stochastic block model, inflation–deflation.
- Non-backtracking matrix of edge-weighted graphs.
- Edge percolation for simulated data.
- Application to quantum chemistry data.

Preliminaries

DENSE	SPARSE
Füredi-Komlós, Combinatorica (1981): $\mathbb{E}(a_{ij}) = \mu > 0 \quad (i \neq j)$	Percolated $G_n(p)$, $p = \frac{c}{n}$ $(n \to \infty)$ Erdős-Rényi
B, Wiley (2013): $\mathbb{E}(a_{ij}) = c_{ab}$ $(i \in V_a, j \in V_b)$, $1 \leq a, b \leq k$	Percolated SBM: $\mathbb{E}(a_{ij}) = \frac{c_{ab}}{n}$ $(i \in V_a, j \in V_b)$, $1 \leq a, b \leq k$
$\lambda_i(\mathbf{A})$ is aligned with $\lambda_i(\mathbb{E}\mathbf{A})$, $i = 1, \ldots, k \quad$ (F-K: $k = 1$)	$\lambda_i(\mathbf{B})$ is aligned with $\lambda_i(\mathbb{E}\mathbf{A})$, $i = 1, \ldots, k$ non-backtracking
Perturbation: Wigner-type matrix Subspace perturbation: Davis-Kahan	deformed Wigner matrix Bauer–Fike

Non-Backtracking (Hashimoto) matrix of simple graphs

$G = (V, E)$ simple graph, $|V| = n$, $|E| = m$;
The entries of the non-backtracking matrix $\mathbf{N} = (n_{ef})$ are indexed by the oriented edges (bidirected edges of E), $|E^{\rightarrow}| = 2m$:

$$n_{ef} = \delta_{e \to f} \delta_{f \neq e^{-1}}, \quad n_{i \to j, s \to l} = \delta_{js}(1 - \delta_{il}),$$

where $e = \{i \to j\}$ and $f = \{s \to l\}$ are oriented edges, and $e \to f$ with $e = (e_1, e_2)$ and $f = (f_1, f_2)$ means that $e_2 = f_1$;
$e^{-1} = \{j \to i\}$.

Alon, Benjamini, Lubetzky, Sodin, Non-backtracking random walks mix faster, Commun. Contemp. Math. (2007).

Relation to line-graphs

If $\mathbf{N} = \begin{pmatrix} \mathbf{N}_{11} & \mathbf{N}_{12} \\ \mathbf{N}_{21} & \mathbf{N}_{22} \end{pmatrix}$, *where the two (row/column) blocks correspond to the edges and their inverses (in the same order), then*

$$\mathbf{N}_{11}^* = \mathbf{N}_{22}, \quad \mathbf{N}_{22}^* = \mathbf{N}_{11}, \quad \mathbf{N}_{12}^* = \mathbf{N}_{12}, \quad \mathbf{N}_{21}^* = \mathbf{N}_{21}.$$

Further, $\mathbf{N}_{11} + \mathbf{N}_{12} + \mathbf{N}_{21} + \mathbf{N}_{22}$ *is equal to the* $m \times m$ *adjacency matrix of the line-graph of* G.

If the line-graphs of two simple graphs, provided they both have node-degrees at least 4, are isomorphic, then they are isomorphic too. However, two simple graphs are isomorphic if and only if, their non-backtracking graphs (with adjacency matrix $\mathbf{N}$) are isomorphic. The non-backtracking random walk is not Markovian (it has the memory that no going back in one step), but the random walk on the non-backtracking graph is Markovian (non-backtr. Laplacian).

Transpose (*), swapping, and parity-time invariance

$\mathbf{N}$ is not a normal matrix, even not always diagonalizable.
In particular, if there are nodes of degree 1 (for example, G is a tree), the algebraic multiplicity of the eigenvalue 0 is larger than the geometric one. However $\mathbf{N}$ exhibits some symmetry:
$n^*_{ef} = n_{e^{-1}\, f^{-1}}$.

With the notation $\breve{x}_e := x_{e^{-1}}$ for the coordinates of $\mathbf{x}, \breve{\mathbf{x}} \in \mathbb{R}^{2m}$: if $\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix}$, then $\breve{\mathbf{x}} = \begin{pmatrix} \mathbf{x}_2 \\ \mathbf{x}_1 \end{pmatrix}$ (swapping).

Consequently: if $\mathbf{x}$ is a right eigenvector of $\mathbf{N}$ corresponding to a real eigenvalue, then $\breve{\mathbf{x}}$ is a left eigenvector of $\mathbf{N}$ (and right eigenvector of $\mathbf{N}^*$) with the same real eigenvalue.

Eigenvalues of N (Ihara formula)

If G is connected, not a cycle and the minimum node-degree≥ 2, then **N** is **irreducible**. By **Frobenius thm**, it has a single real eigenvalue $\lambda_{max}(\mathbf{N}) > 0$ of largest modulus. Since the characteristic polynomial of **N** has real coefficients, its complex eigenvalues occur in conjugate pairs in the bulk, and there can be some other „structural" real eigenvalues out of the bulk.

Ihara formula: If G is not a tree, **N** has $m - n$ eigenvalues equal to 1 and $m - n$ eigenvalues equal to -1, whereas its further eigenvalues are those of the $2n \times 2n$ matrix

$$\mathbf{K} = \begin{pmatrix} \mathbf{O} & \mathbf{D_A} - \mathbf{I}_n \\ -\mathbf{I}_n & \mathbf{A} \end{pmatrix},$$

where **A** is the adjacency- and $\mathbf{D_A}$ is the degree-matrix of the graph (diagonal, contains the node-degrees=row-sums of **A**).

K always has at least one eigenvalue 1, the geometric multiplicity of which is equal to the number of the connected components of G and $\lambda_{max}(\mathbf{K}) = \lambda_{max}(\mathbf{N}) \leq \lambda_{max}(\mathbf{A})$.

Real eigenvalues and eigenvectors of N

Two auxiliary matrices are introduced: the $2m \times n$ matrix **End** has entries $end_{ei} = 1$ if i is the end-node of the (directed) edge e and 0, otherwise; the $2m \times n$ matrix **Start** has entries $start_{ei} = 1$ if i is the start-node of the (directed) edge e and 0, otherwise. Then for any vector $\mathbf{u} \in \mathbb{R}^n$ and for any directed edge $\{i \to j\}$:

$$(\mathbf{End}\,\mathbf{u})_{i\to j} := u_j \quad \text{and} \quad (\mathbf{Start}\,\mathbf{u})_{i\to j} := u_i.$$

Consequently, $\mathbf{End}\,\mathbf{u}$ is the $2m$-dimensional inflated version of the n-dimensional vector $\mathbf{u}$, where the coordinate u_j of $\mathbf{u}$ is repeated as many times, as many edges have end-node j; likewise, in the $2m$-dimensional inflated vector $\mathbf{Start}\,\mathbf{u}$, the coordinate u_i of $\mathbf{u}$ is repeated as many times, as many edges have start-node i.
As each edge is considered in both possible directions, these multiplicities are the node-degrees d_j and d_i, respectively.

Beyond the Ihara formula

$\mathbf{End}^* \mathbf{End} = \mathbf{Start}^* \mathbf{Start} = \mathrm{diag}(d_1, \ldots, d_n) = \mathbf{D_A}$

For any vector $\mathbf{x} \in \mathbb{R}^{2m}$, define

$$x_i^{out} := \sum_{j: j \sim i} x_{i \to j} \quad \text{and} \quad x_i^{in} := \sum_{j: j \sim i} x_{j \to i} \quad (i = 1, \ldots, n).$$

These become the coordinates of the n-dimensional (column) vectors $\mathbf{x}^{in}$ and $\mathbf{x}^{out}$: $\mathbf{x}^{out} = \mathbf{Start}^* \mathbf{x}$ and $\mathbf{x}^{in} = \mathbf{End}^* \mathbf{x}$.

If $\mathbf{x}$ is a (right) eigenvector of $\mathbf{N}^*$ with structural real eigenvalue μ, then the vector $\begin{pmatrix} \mathbf{x}^{out} \\ \mathbf{x}^{in} \end{pmatrix}$ is a right eigenvector of $\mathbf{K}$ with the same eigenvalue μ, where $\mathbf{x}^{out} = \frac{1}{\mu}(\mathbf{D_A} - \mathbf{I}_n)\mathbf{x}^{in}$.

Edge percolation

The edge percolation threshold for the giant component to appear in a sparse simple graph is $\beta > \frac{1}{\lambda_{max}(\mathbf{N})}$, where β is the edge retention probability, see Newman, M. E. J., Message passing methods on complex networks, Proc. R. Soc. London A (2023).
The proof uses the method of Belief Propagation (BP) (when the so-called message passing system of the approximating linear equations has a non-trivial and unstable trivial solution).
In the dense case, it happens at $\frac{1}{\lambda_{max}(\mathbf{A})} \leq \frac{1}{\lambda_{max}(\mathbf{N})}$, see Bollobás, B., Borgs, C., Chayes, J., and Riordan, O., Percolation on dense graph sequences, Ann. Probab. (2010).
More generally, we are looking for the number k, so that k strongly connected clusters (communities) can be detected (within the giant component) in a graph coming from the sparse stochastic block model. We are also looking for the clusters themselves.
The Erdős–Rényi graph $G_n(p)$ is a special case with $k = 1$, where the edges of the complete graph on n vertices are retained independently, with probability $\beta = p$.

The sparse stochastic block model SBM_k

The $k \times k$ *probability matrix* $\mathbf{P}$ of the random graph $G_n \in SBM_k$ has entries $p_{ab} = \frac{c_{ab}}{n}$, where the $k \times k$ symmetric affinity matrix $\mathbf{C} = (c_{ab})$ stays constant as $n \to \infty$. An edge between $i < j$ comes into existence, independently of the others, with probability p_{ab} if $i \in V_a$ and $j \in V_b$, where $(V_1, \ldots, V_k)$ is a partition of the node-set V into k disjoint clusters; $a_{ji} := a_{ij}$. It can be extended to the $i = j$ case when self-loops are allowed, or else, the diagonal entries of the adjacency matrix are zeros.

$\bar{\mathbf{A}}$: the $n \times n$ inflated matrix of the $k \times k$ $\mathbf{P}$: $\bar{a}_{ij} = p_{ab}$ if $i \in V_a$ and $b \in V_b$. When loops are allowed, then $\mathbb{E}(a_{ij}) = \bar{a}_{ij}$ for all $1 \leq i, j \leq n$. In the loopless case, the expected adjacency matrix $\mathbb{E}\mathbf{A}$ differs from $\bar{\mathbf{A}}$ with respect to the the main diagonal, but the diagonal entries are negligible.

Special cases

Sometimes $c_{ab} = c_{in}$ is the within-cluster ($a = b$) and $c_{ab} = c_{out}$ is the between-cluster ($a \neq b$) affinity. The network is called assortative if $c_{in} > c_{out}$, and disassortative if $c_{in} < c_{out}$. Of course, remarkable difference is needed between the two, to recognize the clusters.

The cluster sizes are $n_1, \ldots, n_k$ ($\sum_{i=1}^{k} n_i = n$), so the $k \times k$ diagonal matrix $\mathbf{R} := \mathrm{diag}(r_1, \ldots, r_k)$, where $r_a = \frac{n_a}{n}$ is the relative size of cluster a ($a = 1, \ldots, k$), is also a model parameter ($\sum_{a=1}^{k} r_a = 1$). It is nearly kept fixed as $n \to \infty$.

The model SBM_k is called symmetric if $r_1 = \cdots = r_k = \frac{1}{k}$ and all diagonal entries of the affinity matrix are equal to c_{in}, whereas the off-diagonal ones to c_{out}.

Average degrees

The average degree of a real world graph on m edges and n nodes is $\frac{2m}{n}$. The expected average degree of the random graph $G_n \in SBM_k$ is

$$c = \frac{1}{n}\sum_{a=1}^{k}\sum_{b=1}^{k} n_a n_b p_{ab} = \frac{1}{n^2}\sum_{a=1}^{k}\sum_{b=1}^{k} n_a n_b c_{ab} = \sum_{a=1}^{k} r_a c_a,$$

where $c_a = \sum_{b=1}^{k} r_b c_{ab}$ is the average degree of cluster a. It is valid only if self-loops are allowed. Otherwise, c_a and c should be decreased with a term of order $\frac{1}{n}$, but it will not make too much difference in the subsequent calculations.

Kesten–Stigum threshold

In Bordenave, C., Lelarge, M., Massoulié, L., Non-backtracking spectrum of random graphs: Community detection and non-regular Ramanujan graphs, Ann. Probab. (2018), the case when $c_a = c$ for all a is considered. (This is the hardest case, as otherwise the clusters could be distinguished by sorting the node-degrees.) In this case $\frac{1}{c}\bar{\mathbf{A}}$ is a stochastic matrix, and so, the spectral radius of $\bar{\mathbf{A}}$ is c.

In the symmetric case, $c = \frac{c_{in}+(k-1)c_{out}}{k}$ and the separation of the clusters only depends on the c_{in}, c_{out} relation. If c_{in} is „close" to c_{out}, then the groups cannot be distinguished. The detectability Kesten–Stigum threshold in the symmetric case is

$$|c_{in} - c_{out}| > k\sqrt{c} \iff \mu_2 = \cdots = \mu_k > \sqrt{c},$$

where $c = \mu_1 > \mu_2 = \cdots = \mu_k$ are the leading (real) eigenvalues of $\mathbf{N}$.

BP in the general sparse SBM_k model

Given the observed graph G on n nodes, if $c_1 = \cdots = c_k = c$, then eigenvalues of $\lambda(\mathbf{N}) > \sqrt{c}$ should be considered.

The eigenvalues of $\mathbf{N}$ and $\mathbf{RC}$ (those of $\bar{\mathbf{A}}$) are aligned, see Bordenave, C., Lelarge, M., Massoulié, L., Non-backtracking spectrum of random graphs: Community detection and non-regular Ramanujan graphs, Ann. Prob.

The matrix $\bar{\mathbf{A}}$ has rank k and its non-zero eigenvalues (ν's) with unit norm eigenvectors ($\mathbf{u}$'s), which are step-vectors over the clusters.

Let $\mathbf{x}$ be a unit-norm eigenvector of $\mathbf{N}$, corresponding to the eigenvalue μ that is close to the eigenvalue ν of the expected adjacency matrix, with corresponding eigenvector $\mathbf{u} \in \mathbb{R}^n$. If our graph is from the SBM_k model, then (without knowing its parameters) we know that $\mathbf{u}$ is a step-vector with at most k different coordinates. Then by the above citation,

$$\left\langle \mathbf{x}, \frac{\mathbf{End}\,\mathbf{u}}{\|\mathbf{End}\,\mathbf{u}\|} \right\rangle \geq \sqrt{1-\varepsilon} \geq 1 - \frac{1}{2}\varepsilon,$$

Assume that the expected adjacency matrix of the underlying random graph on n nodes and m edges has rank k with k single eigenvalues and corresponding unit-norm eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_k \in \mathbb{R}^n$. Assume that the non-backtracking matrix $\mathbf{N}$ of the random graph has k structural eigenvalues (aligned with those of the expected adjacency matrix) with eigenvectors $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{R}^{2m}$ such that

$$\left\langle \mathbf{x}_j, \frac{\mathbf{End}\,\mathbf{u}_j}{\|\mathbf{End}\,\mathbf{u}_j\|} \right\rangle \geq \sqrt{1-\varepsilon}, \quad j = 1, \ldots, k.$$

Then for the transformed vectors $\mathbf{D}_{\mathbf{A}}^{-1}\mathbf{x}_j^{in} \in \mathbb{R}^n$:

$$\textit{Obj. function of } k\textit{-means} \leq \sum_{j=1}^{k} \left\| \mathbf{D}_{\mathbf{A}}^{-1}\mathbf{x}_j^{in} - \frac{\mathbf{u}_j}{\|\mathbf{End}\,\mathbf{u}_j\|} \right\|^2 \leq k\varepsilon.$$

Non-Backtracking matrix of edge-weighted graphs

Let $G = (V, E)$ be the skeleton of an edge-weighted graph, $|V| = n$, $|E| = m$; the weight of edge $e = \{i, j\}$ is $W_e = w_{ij} = w_{ji} > 0$, where the remaining entries of the the $n \times n$ symmetric edge weight matrix **W** are zeros (including the diagonal).
Let the $2m \times 2m$ diagonal matrix **D** contain the positive edge-weights in its main diagonal (the first m diagonal entries are the same as the second m ones as $W_e = W_{e^{-1}}$). With them,

$$\mathbf{B} = \mathbf{ND} \quad \text{and} \quad \mathbf{B}^* = \mathbf{DN}^*.$$

The general entry of the $2m \times 2m$ non-backtracking matrix **B** is

$$b_{ef} = W_f \delta_{e \to f} \delta_{f \neq e^{-1}}.$$

Notation

We assume that there are constants C_1 and C_2 (independent of n):

$$C_1 \leq w_{ij} \leq C_2, \quad \text{for} \quad w_{ij} \neq 0.$$

Further, we assume that the skeleton's node degrees

$$d_i = |\{i : w_{ij} > 0, \quad j = 1, \ldots, n\}|, \quad i = 1, \ldots, n$$

are of constant order (it is the case in the k-cluster stochastic block model (SBM_k)).
Let $\mathbf{D}^{\mathbf{W}}$ denote the $n \times n$ diagonal matrix of diagonal entries

$$d_i^{\mathbf{W}} = \sum_{j=1}^{n} w_{ij}, \quad i = 1, \ldots, n,$$

that are the so-called generalized degrees. In the unweighted case (0-1 weights), $d_i^{\mathbf{W}} = d_i$ and $C_1 = C_2 = 1$; in general,

$$C_1 d_i \leq d_i^{\mathbf{W}} \leq C_2 d_i, \quad i = 1, \ldots, n.$$

Start- and End-matrices, in- and out-vectors

The **End** and **Start** matrices are defined as in the unweighted case:

$$\mathbf{End}^* \, \mathbf{D} \, \mathbf{End} = \mathbf{Start}^* \, \mathbf{D} \, \mathbf{Start} = \mathbf{D}^{\mathbf{W}} \quad \text{and} \quad \mathbf{Start}^* \, \mathbf{D} \, \mathbf{End} = \mathbf{W}.$$

For any vector $\mathbf{x} \in \mathbb{R}^{2m}$, the following n-dimensional vectors are introduced:

$$\mathbf{x}^{out} := \mathbf{Start}^* \mathbf{D} \mathbf{x} \quad \text{and} \quad \mathbf{x}^{in} := \mathbf{End}^* \mathbf{D} \mathbf{x}.$$

Coordinatewise, for $i = 1, \ldots, n$,

$$x_i^{out} = \sum_{j \colon j \sim i} w_{ij} x_{i \to j} = \sum_{e \colon e_1 = i} W_e x_e, \quad x_i^{in} = \sum_{j \colon j \sim i} w_{ij} x_{j \to i} = \sum_{e \colon e_2 = i} W_e x_e.$$

Tracing back the problem to lower order matrices

No counterpart of matrix $\mathbf{K}$ works here, but if we know a real eigenvalue μ of $\mathbf{B}$, we are able to find a linear system of equations for the *out*-transform of the corresponding eigenvector that is necessary for spectral clustering. With a Laplacian type equation, μ can also be concluded.

Proposition: Let $\mathbf{x}$ be a (right) eigenvector of $\mathbf{B}$ corresponding to a single positive real eigenvalue μ such that $\mu \neq w_{ij}$, $\forall i, j \in \{1, \ldots, n\}$. Then $\mathbf{y} = \mathbf{x}^{out}$ satisfies the homogeneous system of linear equations

$$[\mathbf{I}_n - \tilde{\mathbf{A}}(\mu) + \tilde{\mathbf{D}}(\mu)]\mathbf{y} = \mathbf{0}$$

with a Laplacian type coefficient matrix, where

$$\tilde{\mathbf{A}}(\mu)_{ij} = \frac{\mu w_{ij}}{\mu^2 - w_{ij}^2} \quad \text{and} \quad \tilde{\mathbf{D}}(\mu)_{ii} = \sum_{j=1}^{n} \frac{w_{ij}^2}{\mu^2 - w_{ij}^2},$$

with the understanding that $w_{ij} = 0$ whenever $i \not\sim j$.

Consequences

The above homogeneous system of linear equations for the coordinates of $\mathbf{y}$ must have a non-trivial solution, so

$$|\mathbf{I}_n - \tilde{\mathbf{A}}(\mu) + \tilde{\mathbf{D}}(\mu)| = 0.$$

This is not a polynomial (characteristic) equation, but it is a rational function of μ. By the assumptions of the Proposition, the denominators are not zeros, so we can multiply the determinant equations with them, and we obtain an high-degree (higher than n) polynomial of μ.
The leading positive real solutions $\mu_1 \geq \cdots \geq \mu_k$ are the same as the structural eigenvalues of $\mathbf{B}$. Their number will be denoted by k. The corresponding $\mathbf{y}_1, \ldots, \mathbf{y}_k$ can be obtained by solving the system of the above homogeneous linear equations (with only an $n \times n$ coefficient matrix).

Clustering

Proposition of Stephan, L., Massoulié, Non-backtracking spectra of inhomogeneous random graphs, Mathematical Statistics and Learning (2022) is applicable to the edge-weighted case too. At the instance, when the number of vertices is n: $\mathbf{P} = (p_{ij})$ is the $n \times n$ symmetric probability matrix of the edges and $\mathbb{W} = (W_{ij})$ is the $n \times n$ symmetric matrix of random weights of the edges. The proposition is applicable if $\bar{\mathbf{A}} := \mathbf{P} \circ \mathbb{EW}$ is a low rank matrix and the so obtained graph is sparse enough. A constant average degree can be guaranteed if, in the instance of n nodes, the p_{ij}'s are proportional to $\frac{1}{n}$. The authors only require for the average degree to be of order $o(\log n)$. In the classical literature, for the average degrees, the order $o(n)$ is considered as sparse. Nowadays the notion of intermediate density is introduced, e.g., for $\log n$ or $Poly(\log n)$ order average degrees, which is the case in the subsequent quantum chemistry examples.

The $SBM_k^{\mathbf{W}}$ model

The stochastic block models discussed before are special cases, where the weights are constantly 1.

In many practical situations, the entry W_{ij} of the $n \times n$ random weighted adjacency matrix $\mathbb{W}$ is $\tilde{w}_{ij}$ times a Bernoulli distributed random variable with parameter p_{ij}, for $1 \leq i < j \leq n$; these entries above the diagonal are independent of each other, while those below the diagonal are identical to them.

So the parameters of this distribution are contained in the symmetric matrices $\mathbf{P}$ and $\widetilde{\mathbf{W}} = (\tilde{w}_{ij})$ of real entries in (0,1]. In this way, the expected adjacency matrix is $\bar{\mathbf{A}} = \mathbf{P} \circ \widetilde{\mathbf{W}}$, with approximate matrix of variances $\mathbf{P} \circ \widetilde{\mathbf{W}} \circ \widetilde{\mathbf{W}}$ (if the entries of $\mathbf{P}$ are decreasing with n), and so, the above theory is applicable to it.

However, for given (large) n, we only observe a realization $\mathbf{W}$ from the the distribution $\mathbb{W}$, in the non-zero positions of which the entries are equal to those of $\widetilde{\mathbf{W}}$.

Even if the observed $\mathbf{W}$ is sparse, the expected weighted adjacency matrix $\bar{\mathbf{A}}$ is not sparse (it is full) $\Longrightarrow$ dense matrix techniques.

k-means clustering in the $SBM_k^{\mathbf{W}}$ model

Assume that the non-backtracking matrix $\mathbf{B}$ has k structural (real) eigenvalues with eigenvectors $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{R}^{2m}$. Then for the transformed vectors $\mathbf{W}^{-1}\mathbf{x}_j^{out} \in \mathbb{R}^n$:

$$\sum_{j=1}^{k} \left\| \mathbf{W}^{-1}\mathbf{x}_j^{out} - \frac{\mathbf{v}_j}{\|\mathbf{End}\, \mathbf{v}_j\|} \right\|^2 \leq k\varepsilon \frac{C_2^2}{C_1^2}.$$

If $\mathbf{v}_j$'s (eigenvectors of $\bar{\mathbf{A}}$) are step-vectors on k steps, then the left-hand side estimates from above the sum of the inner variances of the node representatives that are row vectors of

$$(\mathbf{W}^{-1}\mathbf{x}_1^{out}, \ldots, \mathbf{W}^{-1}\mathbf{x}_k^{out}).$$

To get the $\mathbf{x}_j^{out} \in \mathbb{R}^n$ vectors we do not need the $2m$-dimensional eigenvectors $\mathbf{x}_j$'s of $\mathbf{N}$, but the previous calculations can be used.

The SBM_k^β model: edges are retained with prob. β

- This is in accord with the fact, that in the $k = 1$ case (Erdős–Rényi model), $\lambda(\mathbf{N}) = c$ and $\beta = \frac{1}{c}$; special case of the percolation threshold $\frac{1}{\mu_1}$, see Newman, M. E. J., Message passing methods on complex networks, Proc. R. Soc. London A (2023), when the complete graph is edge-percolated.
- In the multiclass scenario, $\beta_i := \frac{c}{\mu_i^2}$ are further phase transitions, leading to i clusters, for $i = 1, \ldots, k_0$ until $\mu_{k_0} \geq \sqrt{c}$, but $\mu_{k_0+1} < \sqrt{c}$.
- This has relevance only if $\lambda_{max}(\mathbf{N}) > \sqrt{c}$, so eigenvalues of $\mathbf{N}$ greater than $\sqrt{c}$ give the phase transitions.
- Since $\mu_1 \geq \mu_2 \geq \ldots$, with larger β, larger number of clusters can be detected.
- $SBM_k^{\mathbf{W}}$ is a generalization of the SBM_k^β model, where the edges may have different edge-retention probability ($0 < w_{ij} \leq 1$ for the connected vertex-pair i, j); e.g., transmission of the infection depends on the randomly coupled individuals. Then the eigenvalues of $\mathbf{B} = \mathbf{ND}$ are used.

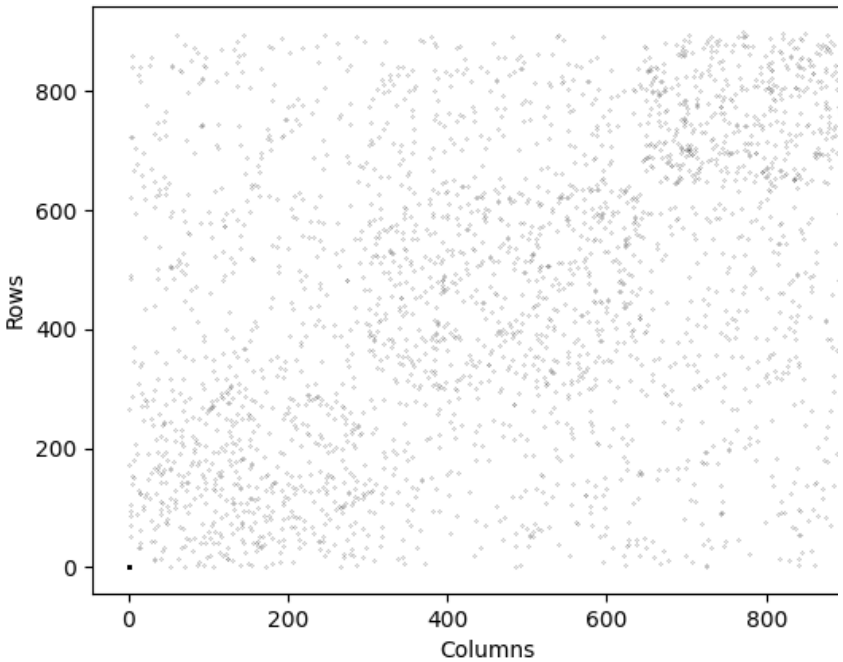
Simulation

A random graph was generated on $n = 900$ nodes, with parameter matrices $\mathbf{R} = \mathrm{diag}(\frac{35}{107}, \frac{42}{107}, \frac{30}{107})$ and

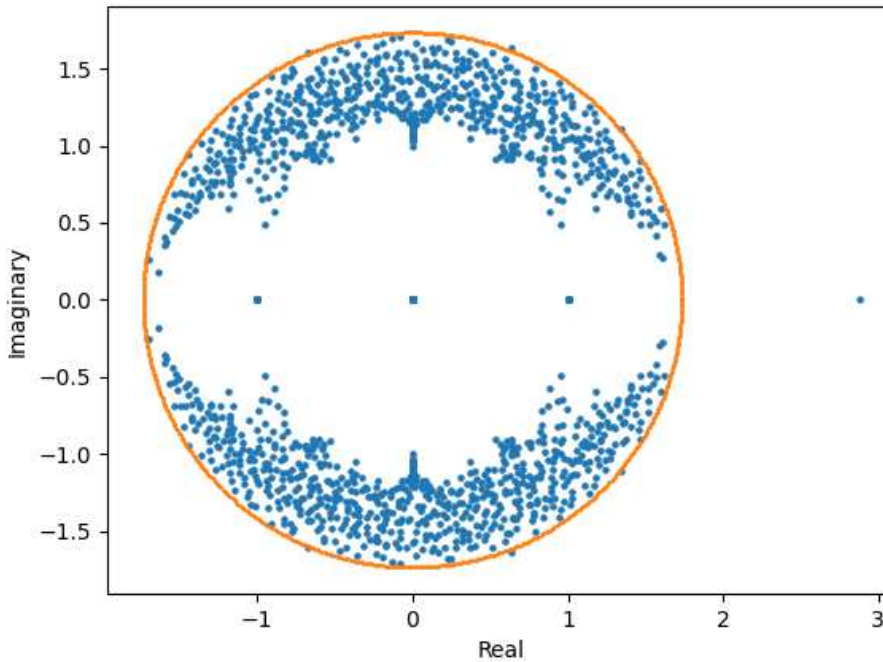
$$\mathbf{C} = \begin{pmatrix} 30 & 11.28 & 7.728 \\ 11.28 & 25 & 10.36 \\ 7.728 & 10.36 & 35 \end{pmatrix}$$

constructed so that the average degrees of the clusters be the same, i.e., $c_a = \sum_{b=1}^{3} r_b c_{ab}$ is the same for $a = 1, 2, 3$.
In the symmetric case, $\beta_2 = \cdots = \beta_k$, so all these phase transitions occur at the same time from the giant cluster to the k-cluster scenario.
Multiple transitions are spectacular if the eigenvalues of $\mathbf{B}$ greater than $\sqrt{c}$ are separated from each other and from $\sqrt{c}$.

$\beta_1 = c/\mu_1^2 < \beta = 0.183 < \beta_2 = c/\mu_2^2$

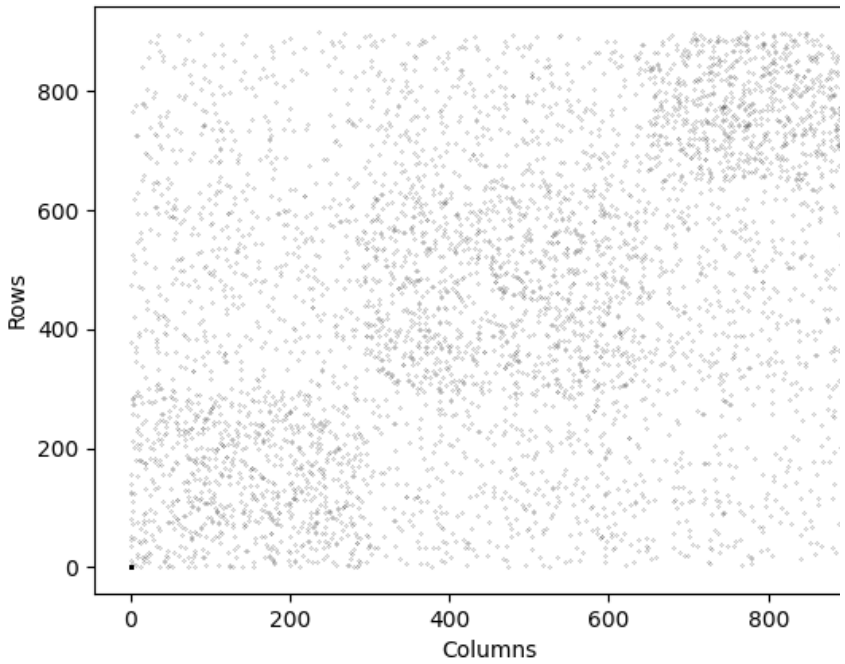


(a) Adjacency matrix.

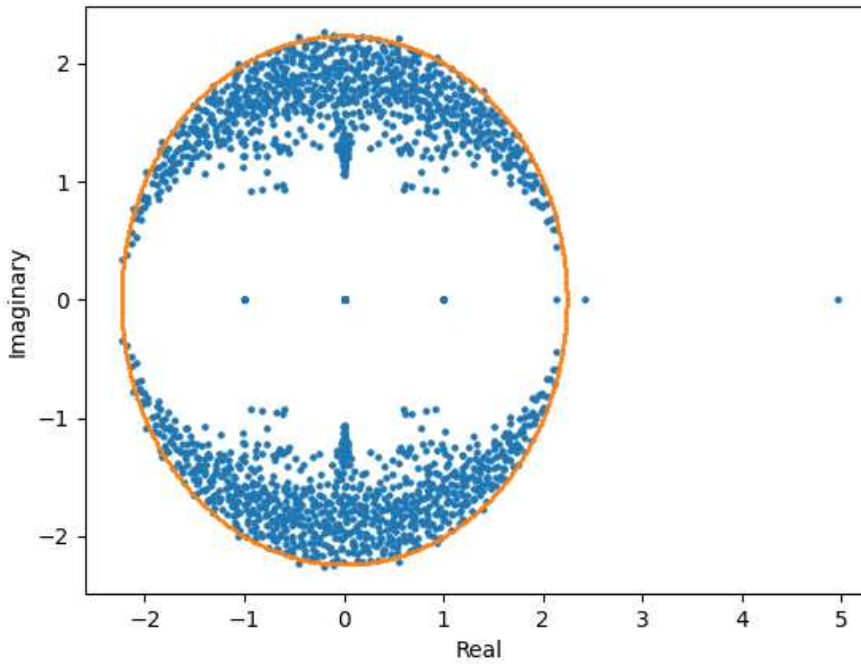


(b) Spectrum of matrix $\mathbf{K}$.

$\beta = \beta_2 = c/\mu_2^2 = 0.305$

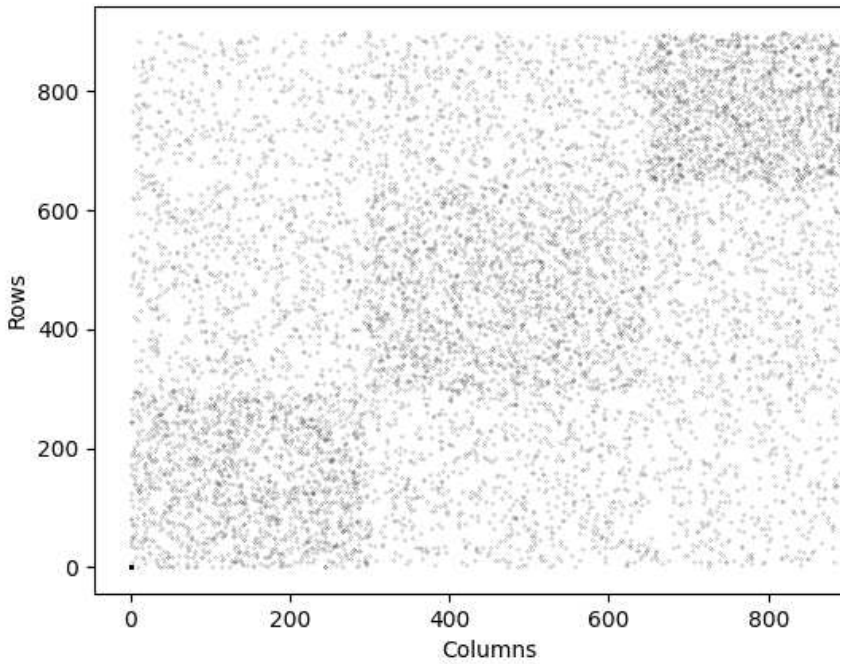


(c) Adjacency matrix.

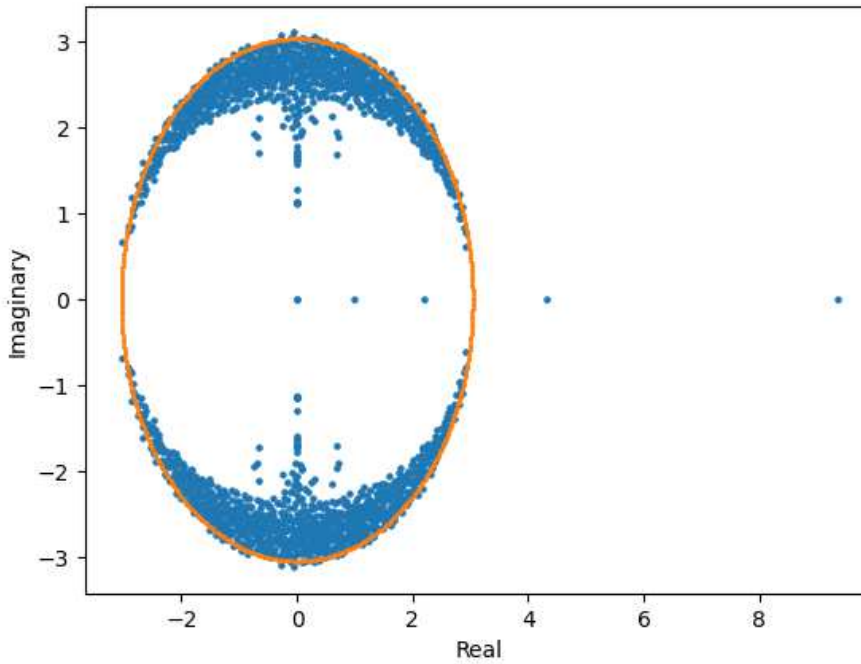


(d) Spectrum of matrix **K**.

$\beta_2 = c/\mu_2^2 < \beta = 0.563 < \beta_3 = c/\mu_3^2$

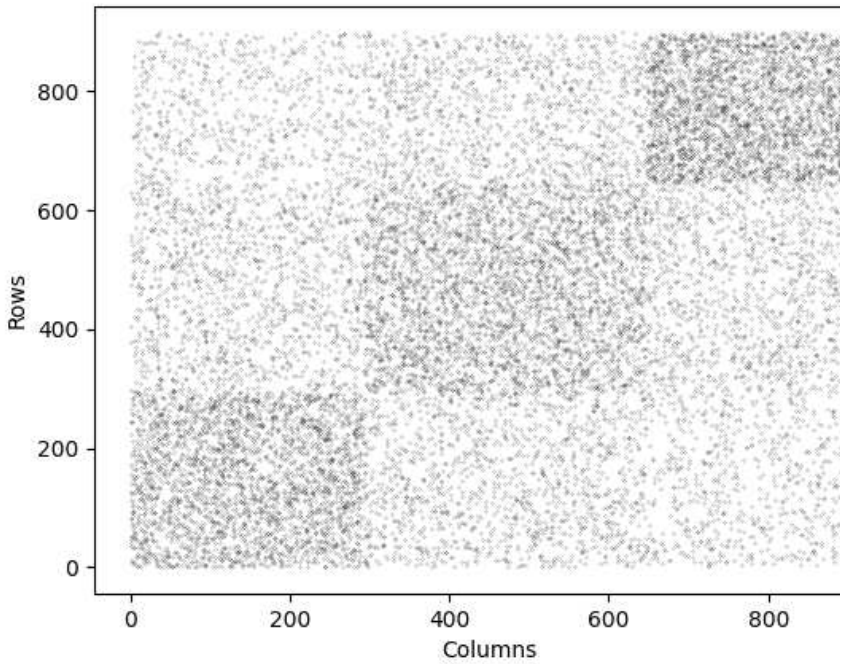


(e) Adjacency matrix.

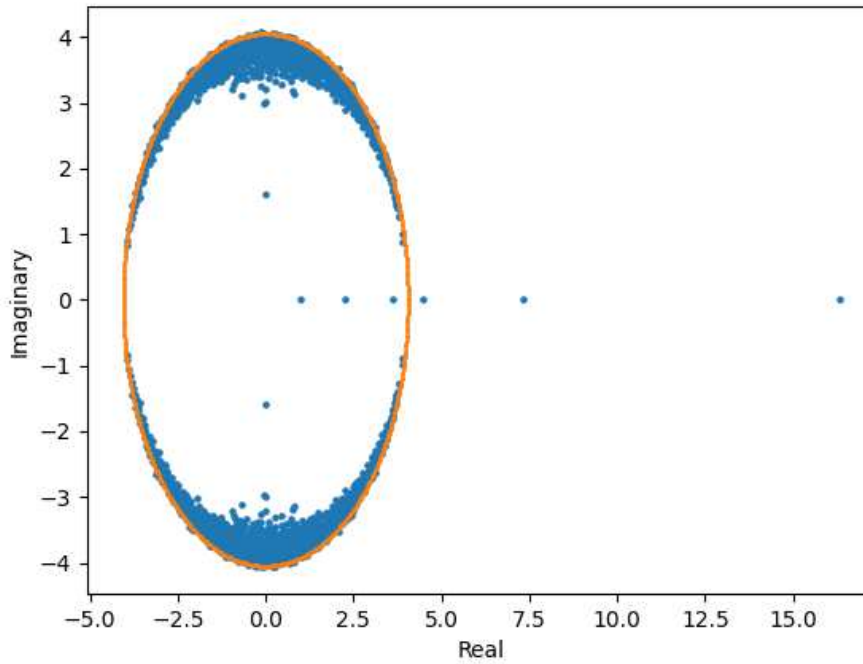


(f) Spectrum of matrix **K**.

$\beta = 1 \approx c/\mu_3^2$



(g) Adjacency matrix.

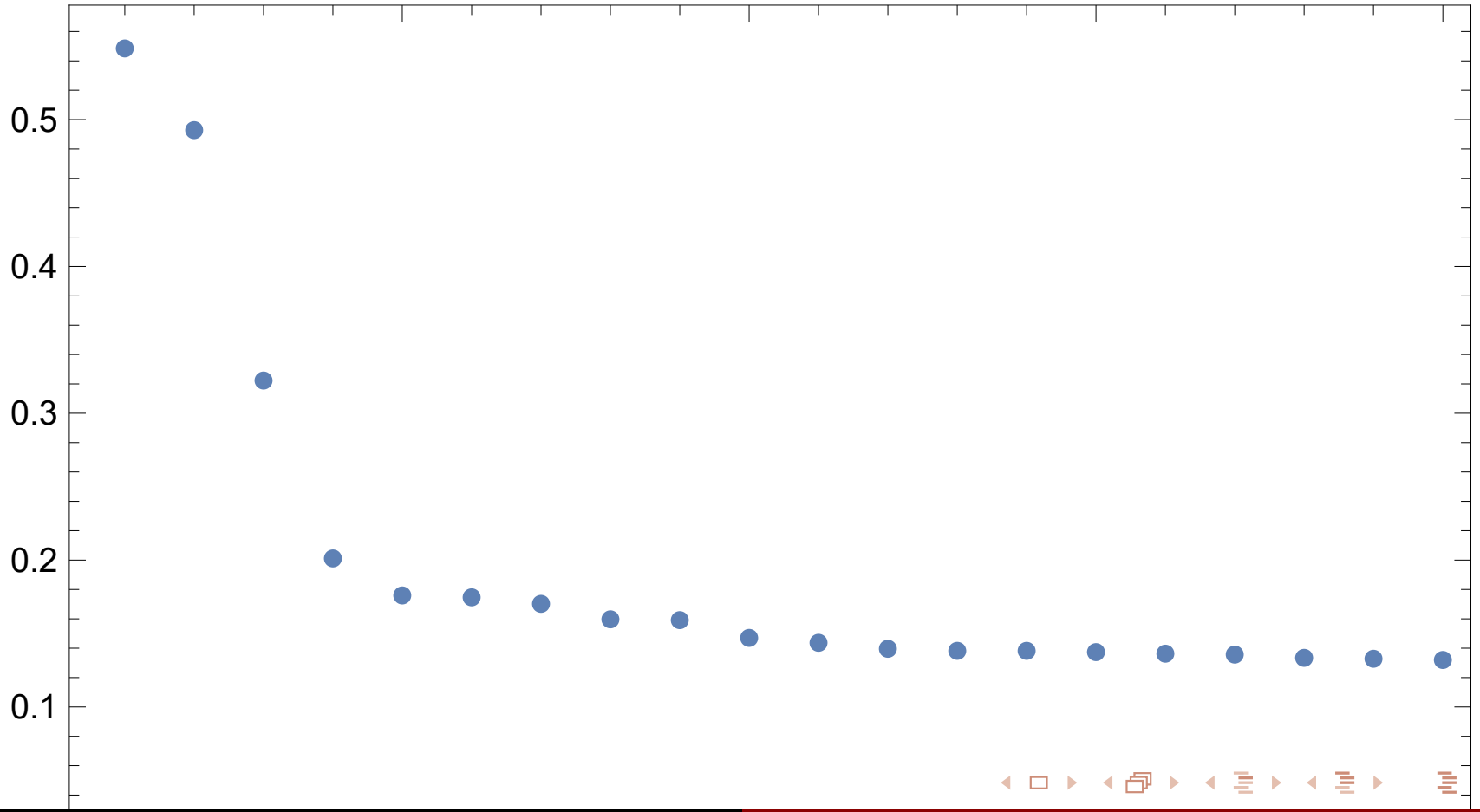


(h) Spectrum of matrix **K**.

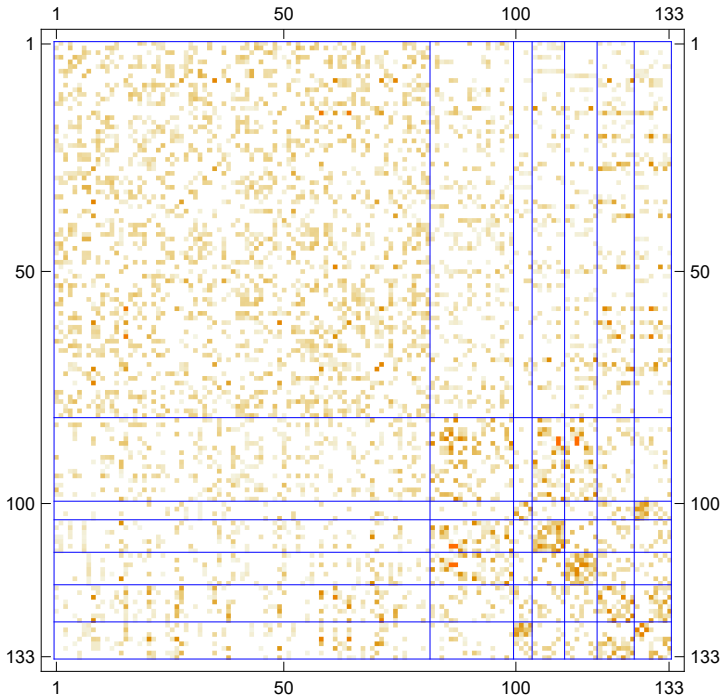
Quantum chemistry application: Real eigenvalues of the B matrix of the water molecule, $n = 133$, $m = 2210$

Nodes: wave functions (electronic networks).

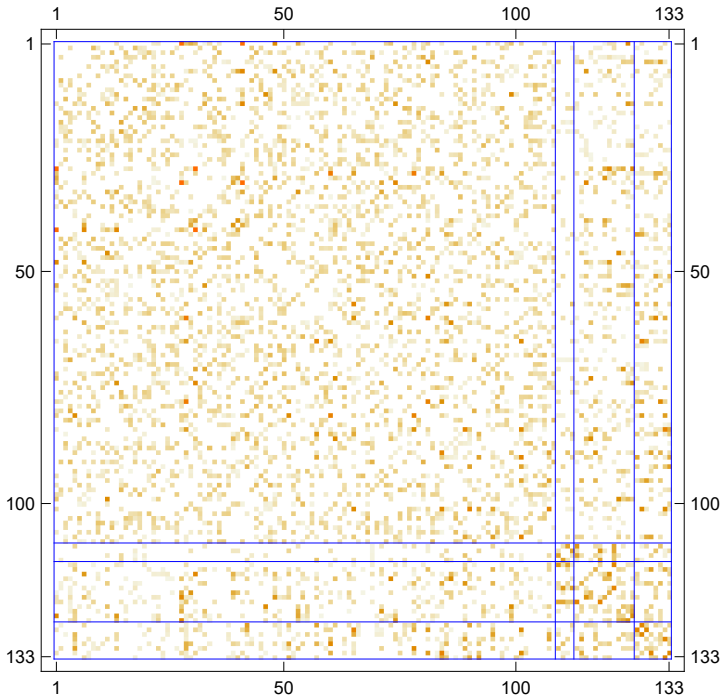
Edge-weights: $w_{ij} = h_{ij}^2$ $(i \neq j)$, where $\mathbf{H} = (h_{ij})$ Hamiltonian.



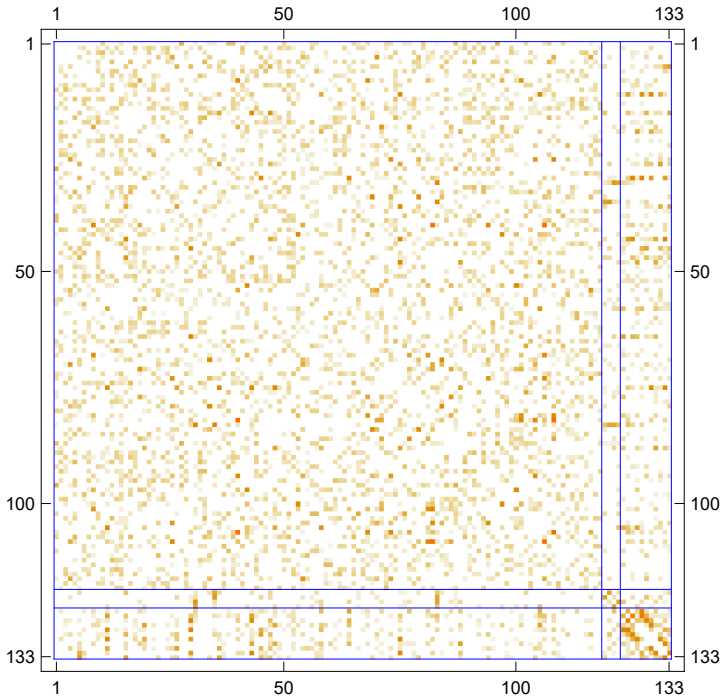
7 clusters obtained by the k-means algorithm applied to the out-vectors corresponding to the 7 leading B-eigenvectors



4 clusters obtained by the k-means algorithm applied to the out-vectors corresponding to the 7 leading B-eigenvectors



3 clusters obtained by the k-means algorithm applied to the out-vectors corresponding to the 7 leading B-eigenvectors



Acknowledgements

Simulation: Daniel Zhou, Budapest Semester in Mathematics (student)

Epidemic problem: László Lovász and Tamás Móri.

Thank you for your attention!