Community detection on large graphs using random matrix theory

Gabriel Ruault: gabriel.ruault@student-cs.fr

Contents

1 Introduction

The purpose of this report is to leverage random matrix theory to build and understand an algorithm that can detect communities in large graphs. In high dimension, the kmeans algorithm breaks down as high-dimensional vectors are equidistant. An approach leveraging spectral information rather than distance becomes necessary.

1.1 Modeling communities in large graphs

In order to get theoretical results, we will need a model for our graphs. We model the graph G using a probabilistic generative model inspired by the Degree-Corrected Stochastic Block Model (DCSBM), an extension of the classical Stochastic Block Model (SBM). This framework captures both community structure and heterogeneous node degrees, as observed in many real-world networks.

Consider a symmetric graph G with n nodes partitioned into K communities $C_1, \dots, C_K$, such that the proportion of nodes in community C_a satisfies $\frac{|C_a|}{n} \to c_a > 0$ as $n \to \infty$. Let $A \in \mathbb{R}^{n \times n}$ denote the adjacency matrix of G, where the entries are defined as:

$$A_{ij} \sim \text{Bernoulli}(q_i q_j C_{ab}), \quad \text{for } i \in C_a,\ j \in C_b.$$

Here, $q_i \in (0, 1)$ represents the intrinsic connection probability of node i, and C_{ab} is a weighting factor governing the connection probability between communities C_a and C_b. We assume:

$$C_{ab} = 1 + \frac{M_{ab}}{\sqrt{n}},$$

where $M_{ab} = \mathcal{O}(1)$ with respect to n. The intrinsic probabilities q_i are independent and identically distributed (i.i.d.) random variables, sampled from a distribution independent of the M_{ab}'s.

The influence of class on the connectivity between nodes reflects the presence of mesoscopic structure typical in networks such as social, biological, and communication systems. The way it scales as the graphs grows ensures that the signal distinguishing communities is weak and vanishes as $n \to \infty$, while still being detectable under spectral methods. It places the model in a sparse, weakly informative regime where phase transitions can be studied.

In practice we will study the modularity matrix B of the graph G which is then given by:

$$B = A - qq^*.$$

with $q \in \mathbb{R}^n$ the vector of q_i's, as it isolates community structure from degree effects.

1.2 The random matrix perspective

Main idea: There may be high dimensionality (due here to the size of the graph) but the real structure lives in a sub-manifold (here the classes). Speaking in the language of linear algebra, it means that the core information will live in low-rank deterministic matrices. Stochasticity will induce high-dimensional independent vectors and hence a high-rank matrix component which will be random.

Strategy: Random matrix theory helps us understand the random component. We can thus have access to the information in the deterministic component by looking at the eigenvalues of the total matrix.

2 Preliminary Observations

2.1 Rewriting the adjacency matrix

We want to see whether the modelling we made of a network can be reframed in the random matrix context described above. We consider q_i known for all i. A straightforward way to identify a deterministic component

is to take an expectation. Hence:

$$A = \mathbb{E}A + (A - \mathbb{E}A) = S + W \tag{1}$$

with

$$S_{ij} = \mathbb{E}[A_{ij}] = \mathbb{E}\big[\mathbb{E}(A_{ij} \mid C_{ab})\big] \underset{A_{ij} \sim \text{Binomial}}{=} \mathbb{E}[q_i q_j C_{ab}] \underset{\text{cond. on } q_i, q_j}{=} q_i q_j \mathbb{E}[C_{ab}] = q_i q_j (1 + \mathbb{E}[\frac{M_{ab}}{\sqrt{n}}]) \tag{2}$$

. In order to express the condition of belonging to a given class in a matrix form we need to use the indicators of the classes. Given this, we notice that S can be written as:

$$S = QJ(\mathbf{1}_K \mathbf{1}_K^T + \frac{\mathbb{E}[M]}{\sqrt{n}}) J^T Q \tag{3}$$

with $Q = \text{diag}(q_1, \ldots, q_n)$, $\mathbf{1}_K \mathbf{1}_K^T$ the matrix with 1 everywhere and $J = (j_1, j_2, \ldots, j_K) \in \mathbb{R}^{nK}$ the matrix of the indicators of the classes. Indeed, QJ multiplies the lines of J with the diagonal elements in Q and a line of J has a 1 in the class to which the node at that line belongs and 0 elsewhere. The same thing happens but with columns for $J^T Q$. Multiplying these with $\mathbb{E}[M]$ we get the correct elements which we need to scale with $\sqrt{n}$. The central matrix $\mathbf{1}_K \mathbf{1}_K^T + \frac{\mathbb{E}[M]}{\sqrt{n}} \in \mathcal{M}_K(\mathbb{R})$thus it is clear that $\text{rg}(S) \leq K$. By definition, $W = A - \mathbb{E}[A]$ has 0 mean. We thus obtain the following decomposition satisfying the requirements of the question:

$$\frac{A}{\sqrt{n}} = QJ(\frac{\mathbf{1}_K \mathbf{1}_K}{\sqrt{n}} + \frac{\mathbb{E}[M]}{n}) J^T Q + \frac{W}{\sqrt{n}} \tag{4}$$

2.2 Rewriting the modularity matrix

As before, but noticing that $q = QJ\mathbf{1}_K$ it is clear that $QJ\mathbf{1}_K \mathbf{1}_K^T J^T Q = qq^T$ and

$$\frac{B}{\sqrt{n}} = QJ \frac{\mathbb{E}[M]}{n} J^T Q + \frac{W}{\sqrt{n}} \tag{5}$$

2.3 Preliminary simulations

Using our model, we can simulate a graph by generating a modularity matrix and extract the spectrum of $\frac{B}{\sqrt{n}}$ for $K = 3$. In these preliminary simulations, we chose $n = 1000$.

Theoretical expectations:

- We observe that $\frac{W}{\sqrt{n}}$ ressembles a Wiener matrix given it is symmetric and with 0 mean. The variance is however non uniform across the whole matrix.
- We also observe the signal S contains both information about the classes (stored in $\mathbb{E}[M]$) and about the distribution of the node degrees (stored in Q).
- From the lecture on spiked-models, we know that we can expect each eigenvalue of the deterministic signal to give rise to an eigenvalue of the total matrix outside the bulk for a strong enough signal-to-noise ratio. Given its rank, S can yield at most $K = 3$ non-zero eigenvalues.

Simulation:

Initially, we generated M randomly but failed to observe spiking. In order to ensure that the eigenvalues of S were large enough and distinct, we chose to place ourselves in the basis where $\mathbb{E}[M]$ is diagonal so as to handpick the eigenvalues.

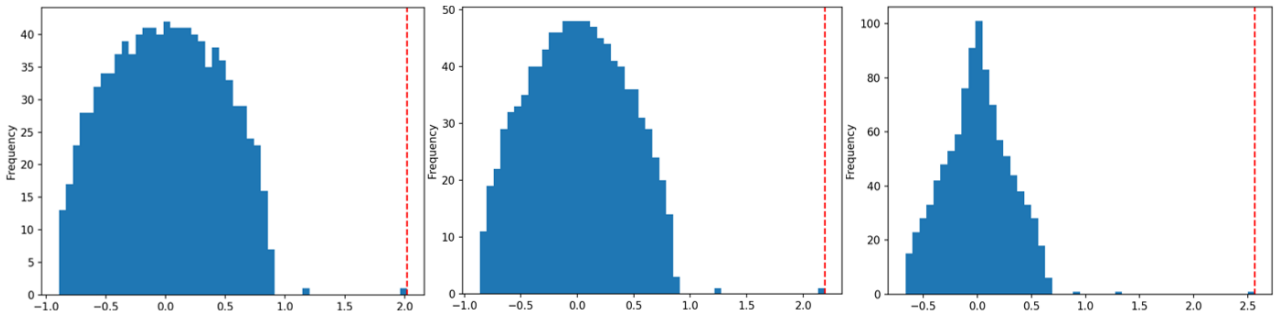

Figure 1: Histogram of eigenvalues of the modularity matrix $B/\sqrt{n}$ for different degree profiles q_i. All graphs are generated with fixed $K = 3$ communities and a fixed signal matrix $M = \sqrt{n}\text{diag}(0.3, 0.5, 0.4)$ Largest eigenvalue is dashed in red. **Left:** $q_i = 0.5$, **Middle:** $q_i \sim \mathcal{U}([0.2, 0.8])$, **Right:** $q_i \sim \mathcal{U}(\{0.1, 0.8\})$.

2.3.1 Case 1: $q_i = q_0$

In Fig. 2 we can make two observations:

- The lower q_0 is , the closer the bulk resembles a semi-circle. This is due to the fact that at low q_0, the impact of C_{ab} on the variance is reduced and variances are more uniform across the matrix.
- The higher q_0 is, the larger the out-of-bulk eigenvalues we observe. This is due to the fact that q_0 also impacts the signal. A large q_0 simply scales up the eigenvalues of S. This can be understood qualitatively: if the network is denser (high q_0), then the impact of a fixed noise is smaller.

In Fig. 1 we see case 1 provides the bulk which is closest to the semi-circular law. In this case we observe 2 out-of-bulk eigenvalues (although in other simulations with stronger signal, we obtained 3 out-of bulk eigenvalues). Intuitively, the near semi-circularity could be due to the fact that, q being uniform, and K being small, $\frac{W}{\sqrt{n}}$ is close to being a Wigner-matrix. It is a Wigner-matrix "by blocks" but with few blocks which justifies why we observe closer behaviour to the pure Wigner situation. We also remark that in this case, the eigenvalues directly only contain information about the classes given there is no information about connectivity as it is uniform.

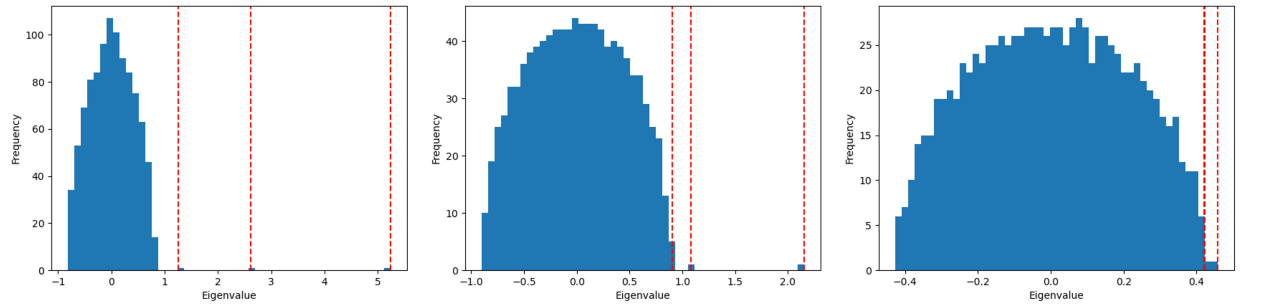

Figure 2: Histogram of eigenvalues of the modularity matrix $B/\sqrt{n}$ for different values of homogeneous $q = q_0$. All graphs are generated with $K = 3$ communities and a fixed signal matrix $M = \sqrt{n}\text{diag}(0.3, 0.5, 0.4)$ Largest K eigenvalues are dashed in red. **Left:** $q_0 = 0.8$, **Middle:** $q_0 = 0.5$, **Right:** $q_0 = 0.2$.

2.3.2 Case 2: $q_i \sim \mathcal{U}([q_0 - \delta, q_0 + \delta])$

In this case, we can have intuition that the signal still mainly stores information about the classes and not the connectivity given that uniformly sampled connectivity does not provide a rich structure on the graph. In practice, we observe in Fig. 1 that indeed the out-of-bulk eigenvalues are very close to those seen in case-1 for homogeneous q. We will see later however that the connectivity impacts the isolated eigenvectors.

As seen in Fig. 3, with a small spread δq, we recover a situation similar to that of case 1. With a larger spread, we see the distribution becomes sharper and more unlike the semi-circular law as the variances in W are now more inhomogeneous. The change is not drastic because $Var(q_i q_j) = O(\delta q^2 q_0^2) + O(\delta q^4)$ meaning it is quite small and we are not so far away from case 1.

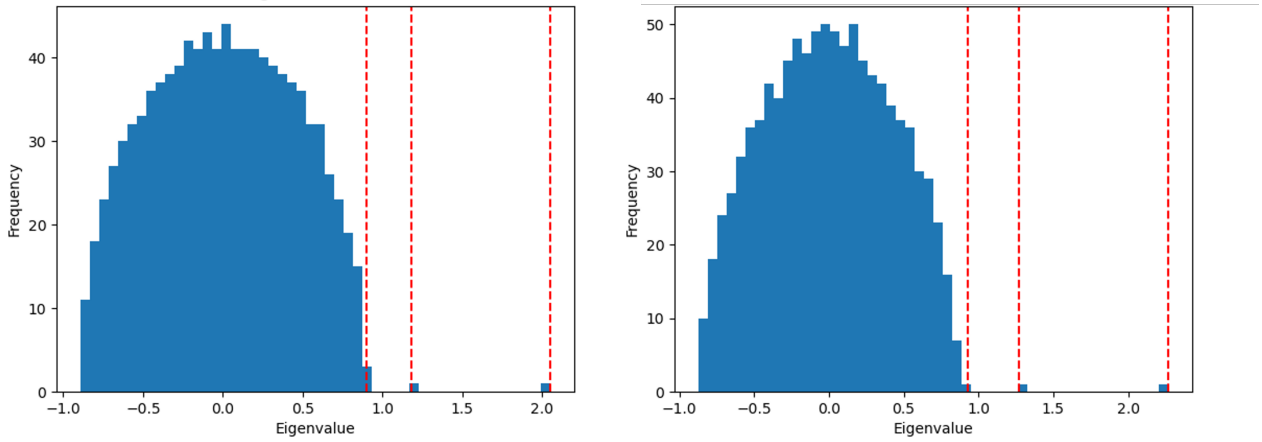

Figure 3: Histogram of eigenvalues of the modularity matrix $B/\sqrt{n}$ for different values of the spread δq for $q \sim \mathcal{U}([q_0 - \delta q, q_0 + \delta q])$ and $q_0 = 0.5$. All graphs are generated with $K = 3$ communities and a fixed signal matrix $M = \sqrt{n}\text{diag}(0.3, 0.5, 0.4)$ Largest K eigenvalues are dashed in red. **Left:** $\delta q = 0.1$, **Right:** $\delta q = 0.3$.

2.3.3 Case 3: $q \sim \mathcal{U}(\{q^{(1)}, q^{(2)}\})$

This corresponds to picking q distributed evenly between $q^{(1)}$ and $q^{(2)}$. As shown in Fig. 4, if the spread between $q^{(1)}$ and $q^{(2)}$ is large then the distribution of q being clearly bimodal, the noise becomes far from Wigner and the bulk is significantly distorted. We also observe in Fig. 1 that while the out-of-bulk eigenvalues are very similar in the first two cases, in the third case, the number of eigenvalues (3 instead of 2) and their values is changed. This is due to the large value of $q^{(2)}$ which scales the signal eigenvalues out of the noise.

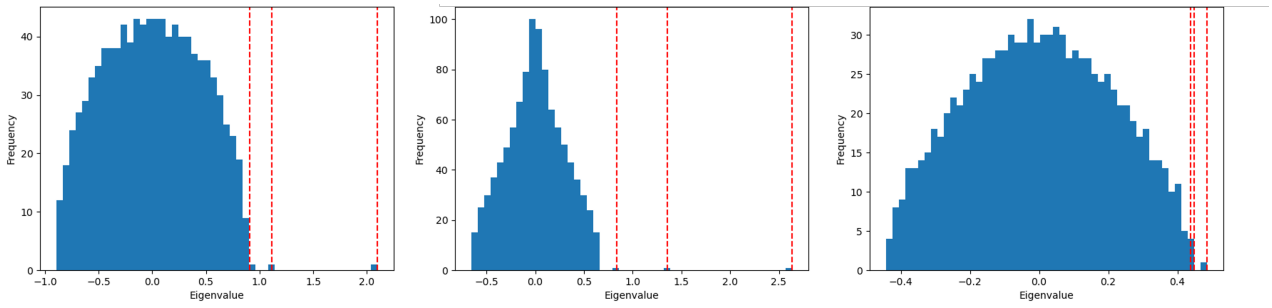

Figure 4: Histogram of eigenvalues of the modularity matrix $B/\sqrt{n}$ for different values of $q^{(1)}$ and $q^{(2)}$ and $q \sim \mathcal{U}(\{q^{(1)}, q^{(2)}\})$. All graphs are generated with $K = 3$ communities and a fixed signal matrix $M = \sqrt{n}\text{diag}(0.3, 0.5, 0.4)$ Largest K eigenvalues are dashed in red. **Left:** $q^{(1)} = 0.45$, $q^{(2)} = 0.55$, **Middle:** $q^{(1)} = 0.1$, $q^{(2)} = 0.8$ **Right:** $q^{(1)} = 0.15$, $q^{(2)} = 0.25$

Remark: we note that in order to test values of q_i close to 0.1 or 0.9 in cases 2 and 3, we had to adapt the range of the eigenvalues of M and make them small enough so that the binomial coefficient $q_i q_j C_{ab} \in [0, 1]$

2.4 Deducing a community detection algorithm

2.4.1 Eigenvectors

We use the same parameter values as those used in Fig. 1 and plot the eigenvectors associated with eigenvalues. We observe that with larger eigenvalues for $\mathbb{E}[M]$ and $K = 3$ out-of-bulk eigenvalues, we have K meaningful extreme eigenvectors. When we only have 2 out-of-bulk eigenvalues as in cases 1 and 2 for our given values of $\mathbb{E}[M]$, we only get two meaningful eigenvectors. Indeed, when the third eigenvalue from the signal gets drowned in the noise, the third largest eigenvalue is not related to the signal but only to the noise. We plot the four largest eigenvalues to see which are meaningful and to get visually interpretable results, we order the classes in the graph.

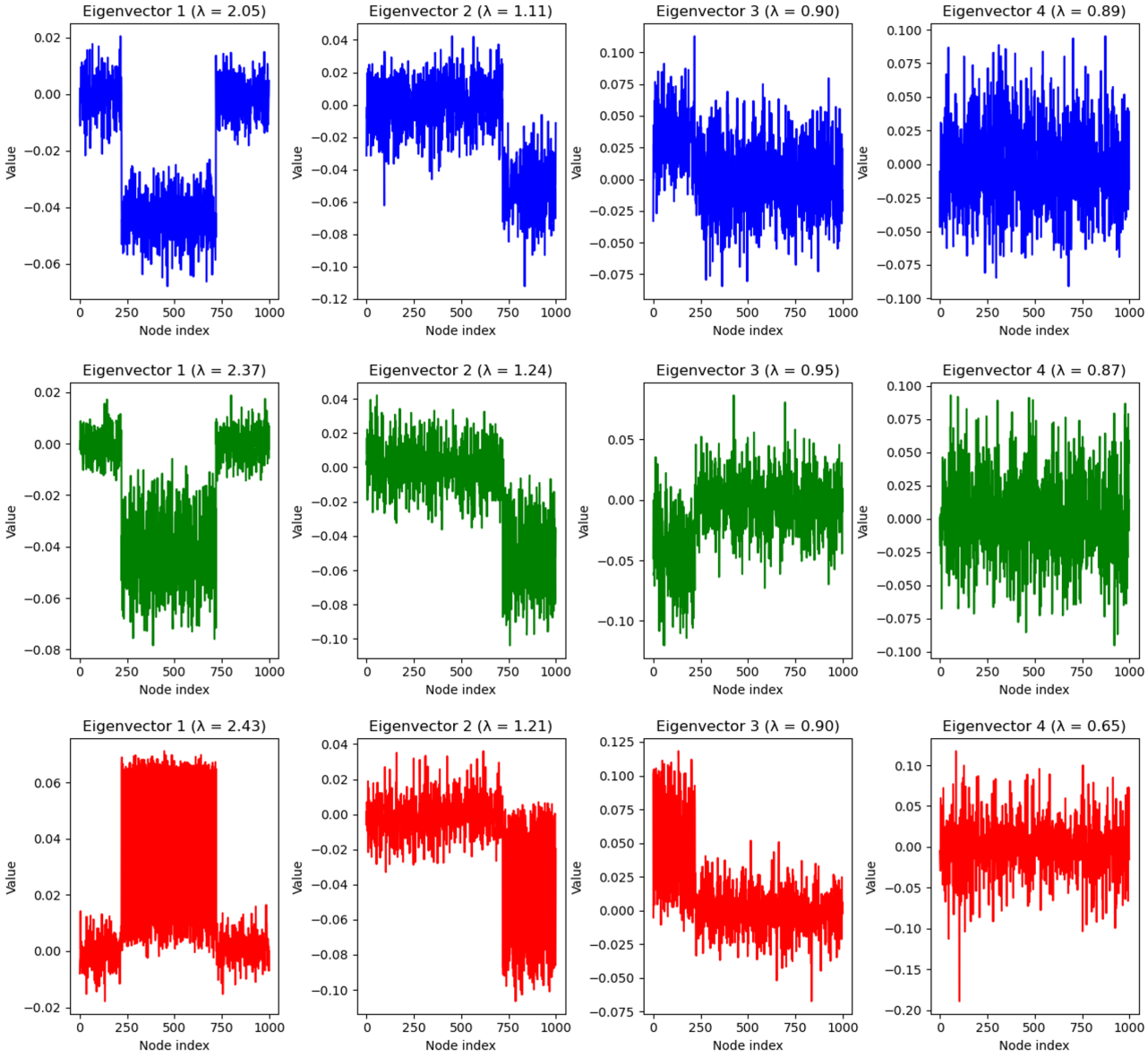

Figure 5: Plot of the eigenvectors associated with the 4 largest eigenvalues of the modularity matrix $B/\sqrt{n}$ for different distributions of q. All graphs are generated with fixed $K = 3$ communities and a fixed signal matrix $M = \sqrt{n}\text{diag}(0.3, 0.5, 0.4)$. **Top** $q_i = 0.5$, **Middle:** $q_i \sim \mathcal{U}([0.2, 0.8])$, **Bottom:** $q_i \sim \mathcal{U}(\{0.1, 0.8\})$.

In Fig. 5, we see that the first two eigenvectors look like scaled and noisy versions of the class indicators. As we saw previously, the third largest eigenvector is outside of the bulk in case 3 of Fig. 1. This explains why here the third largest eigenvector still looks similar to a class indicator. In the other two cases, we see the third eigenvector is very noisy but still contains information about class structure, as the third largest eigenvalue is ever so slightly outside the bulk (hard to visualise). As expected, the smaller eigenvectors are only related to the noise and do not yield information about the signal.

Unlike what we had anticipated, we do not clearly observe yet the fact that in cases 2 and 3, information about the classes is mixed with information about the connectivity structure.

2.4.2 Embeddings:

For the community detection algorithm, we embed the nodes into a 3D space defined by the $K = 3$ eigenvectors as seen in Fig. 6. Noting the k-th eigenvector $v_k = (v_k^{(1)}, \cdots, v_k^{(n)})$, node a is encoded as $\tilde{v}_a = (v_1^{(a)}, \cdots, v_K^{(a)})$.

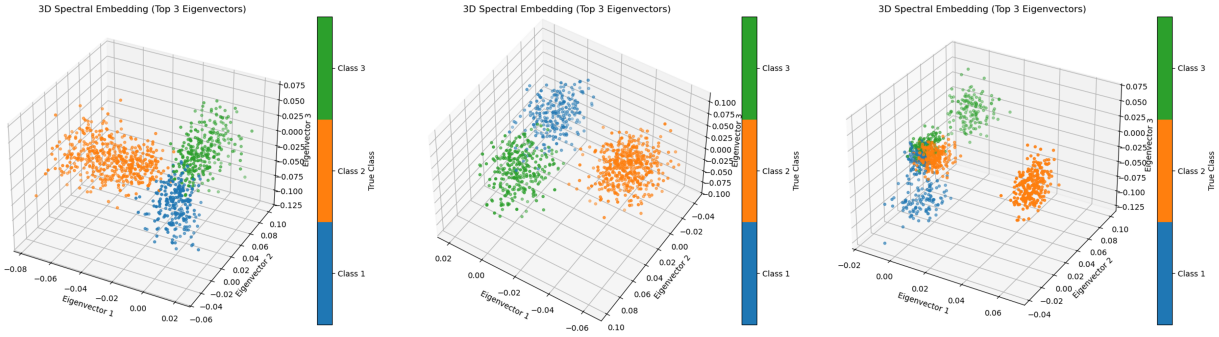

Figure 6: 3D embedding of the graph nodes in a space spanned by the $K = 3$ largest eigenvectors of the modularity matrix. All graphs are generated with fixed $K = 3$ communities and a fixed signal matrix $M = \sqrt{n}\text{diag}(0.3, 0.5, 0.4)$. **Left** $q_i = 0.5$, **Middle:** $q_i \sim \mathcal{U}([0.2, 0.8])$, **Right:** $q_i \sim \mathcal{U}(\{0.1, 0.8\})$.

Fig. 6 confirms our intuition that with inhomogeneous connectivity, the class information is mixed with connectivity.

- Case 3: In case 3 with the bimodal distribution of node connectivities, we see that the classes are de-doubled with three distant clusters (probably corresponding to a high-connectivity regime), and three close-up clusters (probably corresponding to the low connectivity regime). The clusters corresponding to the low connectivity regime are so close it would be hard for an unsupervised algorithm to distinguish between them.
- Case 2: Similarly, we see that unlike the isotropic clusters in case 1, in case 2 the class clusters diffuse from the center to the edges, hinting at a continuum of connectivity regimes from low connectivity to high.

Before we formalise our community detection algorithm, we remark that we build B to take out the degree structure and exhibit the class structure. qq^T will yield a perturbation that scales with n and is not normalized. Connectivity information might hence dominate class information in the largest eigenvalue of A.

2.4.3 Community detection algorithm

As a community detection algorithm we suggest:

- Obtain the adjacency matrix A
- Estimate its node connectivities and build B (this seems to be far from trivial)

- Extract the K eigenvectors corresponding to the K largest eigenvalues depending on the number of classes you believe there to be
- Embed the nodes in a K-dimensional space using the eigenvectors
- Perform k-means on these embeddings

As mentioned above, when the q_i are in homogeneous, clustering will be more difficult due to the appearance of sub clusters and to the short distance between clusters corresponding to small q_i.

3 Theoretical study of the homogeneous case

We now try to support the intuition we gained through our preliminary simulations with theoretical results. To keep things tractable, we consider the homogeneous case and consider in this section that $q_i = q_0 \in (0,1)$ for all i. We chose M diagonal.

3.1 A generalized Wigner theorem

We accept the following results from the lecture notes. For a matrix $X_n \in \mathbb{R}^{n\times n}$ with independent zero-mean entries whose variance is equal to (or converges uniformly to) $1/n$ and whose fourth moment is finite and does not depend on n, its spectral measure

$$\frac{1}{n}\sum_{i=1}^{n}\delta_{\lambda_i(X_n)}$$

converges to the semicircle law P_{sc} with density

$$P_{\text{sc}}(dx) = \frac{1}{2\pi}\sqrt{(4-x^2)_+}\,dx.$$

Furthermore, for $z \in \mathbb{C}^+$, the Stieltjes transform $g_{\text{sc}}(z)$ is the unique solution in $\mathbb{C}^+$ of the equation

$$g = -\frac{1}{z+g} \tag{6}$$

which can be extended on $\mathbb{R}\setminus[-2,2]$. We also know from the isotropic Wigner theorem that, for all deterministic vectors $u, v \in \mathbb{R}^n$, with probability one

$$u^*(X_n - zI_n)^{-1}v - g_{\text{sc}}(z)u^*v \to 0. \tag{7}$$

Moreover,

$$g'_{\text{sc}}(z) = \frac{g^2_{\text{sc}}(z)}{1-g^2_{\text{sc}}(z)}. \tag{8}$$

3.2 Verifying conditions for Wigner theorem

Under the homogeneity hypothesis, we can write

$$\frac{B}{\sqrt{n}} = \frac{q_0^2 J\text{diag}(m_1,\cdots,m_K)J^T}{n} + \frac{W}{\sqrt{n}} \tag{9}$$

We also notice that

$$\mathbb{V}[W_{ij}] = \mathbb{E}[\mathbb{E}[A_{ij}^2|C_{ab}]] - \mathbb{E}[\mathbb{E}[A_{ij}|C_{ab}]]^2 = \mathbb{E}[q_0^2 C_{ab}(1-q_0^2C_{ab})]$$

with $C_{ab} = 1 + \frac{M_{ab}}{n} \xrightarrow[n\to\infty]{\text{a.s.}} 1$, independently on i, j. Thus, since almost sure convergence implies convergence in law:

$$\mathbb{V}[W_{ij}] \xrightarrow[n\to\infty]{\text{uniformly wrt } ij} \mathbb{E}[q_0^2(1-q_0^2)] = q_0^2(1-q_0^2) \underset{\text{def}}{=} \sigma^2$$

Thus,

$$\mathbb{V}\left[\frac{W_{ij}}{\sigma\sqrt{n}}\right]\cdot n \xrightarrow[n\to\infty]{\text{uniformly wrt } ij} 1$$

Given by properties of Bernoulli random variables we have $\mathbb{E}[A^k_{ij}] = q_0^2\mathbb{E}[C_{ab}]$, we can show quite easily the fourth moment is a decreasing function of n, hence finite and with a bound that does not depend on n. I however could not prove it was explicitly independent on n. We nonetheless assume the Wigner theorem mentioned for $\frac{W}{\sigma\sqrt{n}}$.

3.3 Condition for asymptotic existence of isolated eigenvalues in the spectrum of $\frac{B}{\sqrt{n}}$

We assume there exists an isolated eigenvalue λ_k such that λ_k is an eigenvalue of $\frac{B}{\sqrt{n}}$ but not of $\frac{W}{\sqrt{n}}$. We assume there might be up to K such eigenvalues hence the notation λ_k. Since the eigenvalues of $\frac{W}{\sqrt{n\sigma}}$ are in $[-2,2]$, necessarily $\frac{\lambda_k}{\sigma} \in \mathbb{R}\setminus[-2,2]$.

$$\lambda_k \in \mathrm{sp}\left(\frac{B}{\sqrt{n}}\right) \Leftrightarrow 0 = \det\left(\frac{B}{\sqrt{n}} - \lambda I_n\right) = \det\left(\frac{W}{\sqrt{n}} - \lambda I_n + \frac{q_0^2 J\mathrm{diag}(m_1,\cdots,m_K)J^T}{n}\right)$$

Since $\lambda_k \notin \mathrm{sp}\left(\frac{W}{\sqrt{n}}\right)$, we can factorise:

$$0 = \det\left(I_n + \frac{1}{\sigma}\left(\frac{W}{\sqrt{n\sigma}} - \frac{\lambda}{\sigma}I_n\right)^{-1}\frac{q_0^2 J\mathrm{diag}(m_1,\cdots,m_K)J^T}{n}\right)$$

Writing $Q\left(\frac{\lambda}{\sigma}\right) = \left(\frac{W}{\sqrt{n\sigma}} - \frac{\lambda}{\sigma}I_n\right)^{-1}$, and using the Sylvester identity:

$$\det\left(I_n + \frac{1}{\sigma}Q\left(\frac{\lambda}{\sigma}\right)\frac{q_0^2 J\mathrm{diag}(m_1,\cdots,m_K)J^T}{n}\right) = \det\left(I_K + \frac{1}{\sigma}J^T Q\left(\frac{\lambda}{\sigma}\right)J\cdot\frac{q_0^2\mathrm{diag}(m_1,\cdots,m_K)}{n}\right)$$

and we rewrite

$$J^T Q\left(\frac{\lambda}{\sigma}\right)J = \left[j_k^* Q\left(\tfrac{\lambda}{\sigma}\right) j_l\right]_{kl}$$

Using the isotropic Wigner theorem, and by continuity of the determinant, we observe that asymptotically: [

$$J^T Q\left(\frac{\lambda}{\sigma}\right)J = g_{sc}\left(\frac{\lambda}{\sigma}\right)\mathrm{diag}(n_1,\cdots,n_K)$$

since the indicatrices are orthogonal, with n_k the number of nodes in class k. We thus have asymptotically:

$$\lambda_k \in \mathrm{sp}\left(\frac{B}{\sqrt{n}}\right) \Leftrightarrow \det\left(I_K + \mathrm{diag}\left(\frac{q_0^2 g_{sc}\left(\frac{\lambda}{\sigma}\right)n_1 m_1}{\sigma n},\cdots,\frac{q_0^2 g_{sc}\left(\frac{\lambda}{\sigma}\right)n_K m_K}{\sigma n}\right)\right) = 0$$

$$\Leftrightarrow \exists k \in \{1,\cdots,K\}, \quad 1 + \frac{q_0^2 g_{sc}\left(\frac{\lambda}{\sigma}\right)n_k m_k}{\sigma n} = 0$$

This gives K possible values of λ_k associated with K different conditions, each of which can be written as:

$$g_{sc}\left(\frac{\lambda}{\sigma}\right) = -\frac{\sigma}{q_0^2 c_k m_k} \tag{10}$$

where we have taken the limit $\frac{n_k}{n} \to c_k$. When does this equation have a solution in $\mathbb{R}\setminus[-2,2]$? We know from Eq. 6 that for $x \in \mathbb{R}\setminus[-2,2]$

$$g_{sc}(x)^2 + \lambda g_{sc}(x) + 1 = 0 \quad \Leftrightarrow \quad g_{sc}(x) = \frac{-x \pm \sqrt{x^2-4}}{2} \tag{11}$$

Since g_{sc} is holomorphic, it is continuous where defined, thus we must select one of the two branches. The only branch which tends to 0 as $x \to +\infty$ is:

$$g_{sc}(x) = \frac{-x + \sqrt{x^2 - 4}}{2} \tag{12}$$

This asymptotic behaviour is required by the definition of the Stieltjes transform. For $x < 0$, we similarly pick:

$$g_{sc}(x) = \frac{-x + \text{sign}(x)\sqrt{x^2 - 4}}{2} \tag{13}$$

From this, we see g_{sc} is continuously increasing on $]-\infty, -2[\cup]2, +\infty[$, with values in $]-1, 0[\cup]0, 1[$. Hence, by the intermediate value theorem, a condition for the existence of an isolated λ_k is:

$$\frac{\sigma}{q_0^2 c_k |m_k|} < 1 \tag{14}$$

3.4 Asymptotic position of isolated eigenvalues

Since $g_{sc}(]-\infty, -2[) =]0, 1[$ on, $g_{sc}(]2, -\infty[) =]-1, 0[$, and $]0, 1[\cap]-1, 0[= \emptyset$ and g_{sc} is increasing on its domain, there is unicity of the solution when it exists. We solve for it explicitly: If $y < 0$, we solve $g_{sc}(x) = y$ for $x > 0$, and find

$$x = \frac{-(y^2 + 1)}{y} \tag{15}$$

We get the same answer for $y > 0$. Taking $x = \frac{\lambda_k}{\sigma}$ and $y = -\frac{\sigma}{q_0^2 m_k c_k}$, we obtain

$$\lambda_k = \frac{\sigma^2 + (q_0^2 m_k c_k)^2}{q_0^2 m_k c_k} \tag{16}$$

Thus, whenever the condition established in Eq. 14 is verified for a given combination of m_k, c_k, q_0, if n is large enough, there will be an isolated eigenvalue given by Eq. 16. As expected, there can indeed be up to K such isolated eigenvalues. We verify this experimentally with the following parameter values, for $K = 3$ classes.

Parameter	**Value**
n	5000
q_0	0.7
c_k	$\{0.5, 0.2, 0.3\}$
$\min \lvert m_k \rvert$	$\{2.0, 5.1, 3.4\}$
Simulation m_k	$\{-3, 7, 4\}$
Theoretical λ_k	$\{-1.075, 1.050, -1.013\}$

Table 1: Summary of parameters for the simulation in the homogeneous case: $\min |m_k|$ refers to the theoretical condition in Eq. 14 while Simulation m_k is the value we selected for the simulation.

We obtain the following eigenvalue distribution:

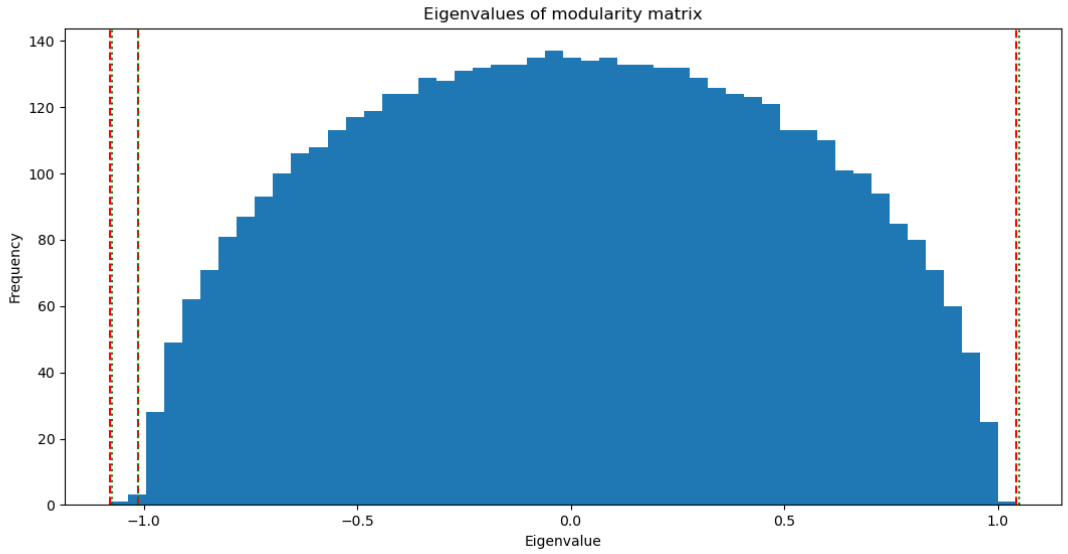

Figure 7: Histogram of eigenvalues of the modularity matrix $B/\sqrt{n}$ with parameters described in Tab. 1 The empirical K largest eigenvalues are dashed in red and the theoretical isolated eigenvalues are dashed in green.

We can see in Fig. 7 that the red and green lines are almost superposed. The empirical values of the largest K eigenvalues are $\{\,1.045, -1.053, -1.018\,\}$. Compared to the theoretical values, this corresponds to an average relative error of 1%. We selected simulation values of m_k close to the minimum required values for the existence of isolated eigenvalues and obtained empirical isolated eigenvalues close to the theoretical ones. This confirms our theoretical results.

3.5 Asymptotic alignment between isolated eigenvectors and class indicators

We will study a single isolated eigenvalue λ_{iso} and its associated eigenvector, assuming single multiplicity. It is easy to see that with the given hypotheses, B is symmetric. We can thus apply the spectral theorem:

$$\begin{aligned}\frac{B}{\sqrt{n}} &= \mathcal{O}_p \begin{pmatrix} \lambda_{\text{iso}} & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \mathcal{O}_n^* \\ &= \begin{bmatrix} v_{\text{iso}} & \mathcal{O}_{n-1} \end{bmatrix} \begin{pmatrix} \lambda_{\text{iso}} & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \begin{bmatrix} v_{\text{iso}}^* \\ \mathcal{O}_{n-1}^* \end{bmatrix}\end{aligned}$$

By the properties of orthogonal matrices:

$$\tilde{Q}(z) = \left(\frac{B}{\sqrt{n}} - zI_n\right)^{-1} = \begin{bmatrix} v_{\text{iso}} & \mathcal{O}_{n-1} \end{bmatrix} \begin{pmatrix} \frac{1}{\lambda_{\text{iso}}-z} & & \\ & \ddots & \\ & & \frac{1}{\lambda_n - z} \end{pmatrix} \begin{bmatrix} v_{\text{iso}}^* \\ \mathcal{O}_{n-1}^* \end{bmatrix}$$

Since the eigenvalue λ_{iso} is isolated, we can integrate over a circle Γ around it:

$$\frac{1}{2\pi i} \oint_\Gamma {j_a}^* \widetilde{Q}(z) j_a \, dz$$

$$= {j_a}^* \begin{bmatrix} v_{\text{iso}} & \mathcal{O}_{n-1} \end{bmatrix} \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 0 \end{pmatrix} \begin{bmatrix} v_{\text{iso}}^* \\ \mathcal{O}_{n-1}^* \end{bmatrix} j_a = {j_a}^* v_{\text{iso}} v_{\text{iso}}^* j_a$$

by the residue theorem. We now rewrite:

$${j_a}^* \widetilde{Q}(z) j_a = \frac{1}{\sigma} {j_a}^* \left(\frac{q_0^2 J \text{diag}(m_1, \cdots, m_K) J^T}{n\sigma} + \frac{W}{\sqrt{n}\sigma} - \frac{z}{\sigma} I_n \right)^{-1} j_a$$

Using the Woodbury identity with

$$A = \frac{W}{\sqrt{n}\sigma} - \frac{z}{\sigma} I_n, \quad U = \frac{q_0^2}{n\sigma} J \text{diag}(m_1, \cdots, m_K), \quad V^* = J^T$$

we get:

$${j_a}^* \left(A + UV^*\right)^{-1} j_a = {j_a}^* \left(Q\left(\frac{z}{\sigma}\right) - Q\left(\frac{z}{\sigma}\right) \frac{q_0^2}{n\sigma} J D_M \left(I_K + \frac{q_0^2}{n\sigma} J^T Q\left(\frac{z}{\sigma}\right) J D_M \right)^{-1} J^T Q\left(\frac{z}{\sigma}\right) \right) j_a$$

with $Q\left(\frac{z}{\sigma}\right) = \left(\frac{W}{\sqrt{n}\sigma} - \frac{z}{\sigma} I_n\right)^{-1}$ and $D_M = \text{diag}(m_1, \cdots, m_K)$. We now notice:

$$j_a^* Q\left(\frac{z}{\sigma}\right) J = \left[j_a^* Q\left(\tfrac{z}{\sigma}\right) j_1, \cdots, j_a^* Q\left(\tfrac{z}{\sigma}\right) j_K \right]$$

$$J^T Q\left(\frac{z}{\sigma}\right) j_a^* = \begin{bmatrix} j_1^* Q\left(\frac{z}{\sigma}\right) j_a \\ \vdots \\ j_K^* Q\left(\frac{z}{\sigma}\right) j_a \end{bmatrix}$$

$$J^T Q\left(\frac{z}{\sigma}\right) J = \left[j_k^* Q\left(\tfrac{z}{\sigma}\right) j_l \right]_{k,l}$$

We now use the isotropic Wigner theorem to obtain the following asymptotic result as $n \to \infty$:

$$\sigma {j_a}^* \widetilde{Q}(z) j_a = {j_a}^* Q\left(\frac{z}{\sigma}\right) j_a - \frac{q_0^2}{n\sigma} \left[0 \cdots n_a g_{sc}\left(\tfrac{z}{\sigma}\right) \cdots 0 \right] D_M \left(I_K + \frac{q_0^2 g_{sc}\left(\frac{z}{\sigma}\right)}{n\sigma} \text{diag}(n_1, \cdots, n_K) D_M \right)^{-1} \begin{bmatrix} 0 \\ \vdots \\ n_a g_{sc}\left(\frac{z}{\sigma}\right) \\ \vdots \\ 0 \end{bmatrix}$$

where we again used the fact that ${j_k}^* j_l = \delta_{kl} n_k$. All matrices are diagonal. Taking inverses and scalar products, we get:

$$\sigma {j_a}^* \widetilde{Q}(z) j_a = {j_a}^* Q\left(\frac{z}{\sigma}\right) j_a - \frac{q_0^2}{n\sigma} n_a g_{sc}\left(\frac{z}{\sigma}\right) m_a \left(\frac{1}{1 + \frac{q_0^2 g_{sc}\left(\frac{z}{\sigma}\right)}{n\sigma} n_a m_a} \right) n_a g_{sc}\left(\frac{z}{\sigma}\right)$$

We observe an extra n_a factor compared to n, which motivates the normalization of j_a. Thus, we write:

$$\frac{1}{n_a} \cdot \frac{1}{2\pi i} \oint_\Gamma {j_a}^* \widetilde{Q}(z) j_a \, dz = \left| \left\langle \frac{j_a}{\sqrt{n_a}} \middle| v_{\text{iso}} \right\rangle \right|^2 = \frac{1}{\sigma} \cdot \frac{1}{2\pi i} \oint_\Gamma {j_a}^* Q\left(\frac{z}{\sigma}\right) j_a - \left(\frac{\frac{q_0^2}{\sigma} \cdot \frac{n_a}{n} g_{sc}^2\left(\frac{z}{\sigma}\right) m_a}{1 + \frac{q_0^2}{\sigma} \cdot \frac{n_a}{n} g_{sc}\left(\frac{z}{\sigma}\right) m_a} \right) dz$$

By the residue theorem, the first term integrates to zero since, by construction, the bulk spectrum has no eigenvalue within Γ. Asymptotically, we have:

$$\frac{n_a}{n} \to c_a$$

We now consider $\lambda_{\text{iso}} = \lambda_k$, with $k \in \{1, 2, \ldots, K\}$.

3.5.1 Case 1: $k = a$

From Eq. 10, we know:

$$g_{sc}\left(\frac{\lambda_k}{\sigma}\right) = -\frac{\sigma}{q_0^2 m_k c_k}$$

Thus, the function

$$f(z) = \frac{\frac{q_0^2}{\sigma} \cdot \frac{n_a}{n} g_{sc}^2\left(\frac{z}{\sigma}\right) m_a}{1 + \frac{q_0^2}{\sigma} \cdot \frac{n_a}{n} g_{sc}\left(\frac{z}{\sigma}\right) m_a}$$

has a pole at λ_k. Since g_{sc} is holomorphic, so is the denominator, and its zeros are isolated. We can therefore choose Γ small enough so that the integrand has no other poles inside Γ. Assuming the pole is simple, we apply the residue theorem:

$$-\frac{1}{\sigma} \cdot \frac{1}{2\pi i} \oint_\Gamma f(z)\, dz = -\frac{1}{\sigma} \lim_{z \to \lambda_k} (z - \lambda_k) \cdot f(z)$$

$$= -\lim_{z \to \lambda_k} \frac{(z - \lambda_k)}{\sigma} \cdot \frac{g_{sc}^2\left(\frac{z}{\sigma}\right)}{-g_{sc}\left(\frac{\lambda_k}{\sigma}\right) + g_{sc}\left(\frac{z}{\sigma}\right)} = \frac{1}{\sigma} \cdot \frac{g_{sc}^2\left(\frac{\lambda_k}{\sigma}\right)}{g'_{sc}\left(\frac{\lambda_k}{\sigma}\right)}$$

Using the known identity $g'_{sc}(x) = 1 - g_{sc}^2(x)$, we obtain:

$$|\langle \frac{j_k}{\sqrt{n_k}} \,|\, v_k \rangle|^2 = 1 - g_{sc}^2\left(\frac{\lambda_k}{\sigma}\right) = 1 - \left(\frac{\sigma}{q_0^2 m_k c_k}\right)^2 \tag{17}$$

3.5.2 Case 2: $a \neq k$

In this case, $f(z)$ is holomorphic in a neighborhood of λ_k, and the integral of a holomorphic function over a closed contour yields zero:

$$|\langle \frac{j_a}{\sqrt{n_a}} \,|\, v_k \rangle|^2 = 0 \tag{18}$$

3.6 Empirical validation of theoretical results

We perform the following simulations:

Quantity	Value
q_0	0.7
c_k	$\{0.5, 0.2, 0.3\}$
Simulation m_k	$\{-10, 15, -8\}$
Theoretical $\lvert\langle \frac{j_k}{\sqrt{n_k}} \,\rvert\, v_k\rangle\rvert^2$	$\{0.9584, 0.8843, 0.8193\}$
Empirical $\lvert\langle \frac{j_k}{\sqrt{n_k}} \,\rvert\, v_k\rangle\rvert^2$, n = 1000	$\{0.9615, 0.8749, 0.8284\}$
Empirical $\lvert\langle \frac{j_k}{\sqrt{n_k}} \,\rvert\, v_k\rangle\rvert^2$, n = 3000	$\{0.9589, 0.9013, 0.8132\}$
Empirical $\lvert\langle \frac{j_k}{\sqrt{n_k}} \,\rvert\, v_k\rangle\rvert^2$, n = 5000	$\{0.9583, 0.8820, 0.8198\}$

Table 2: parameters, empirical and theoretical results regarding alignement between class indicators and the isolated K eigenvectors for the simulation in the homogeneous case:

The average relative errors between theoretical and empirical values in the cases $n = 1000, n = 3000, n = 5000$ are $0.8\%, 0.9\%, 0.1\%$. While it is surprising to see an increase in error for $n = 3000$ compared to $n = 1000$ (probably due to noise and only estimation for class 2 is worse), with $n = 5000$, we clearly notice an improved agreement between theory and practice as n grows.

Fig. 8 confirms that the isolated eigenvectors are asymptotically essentially orthogonal to all but one class indicator.

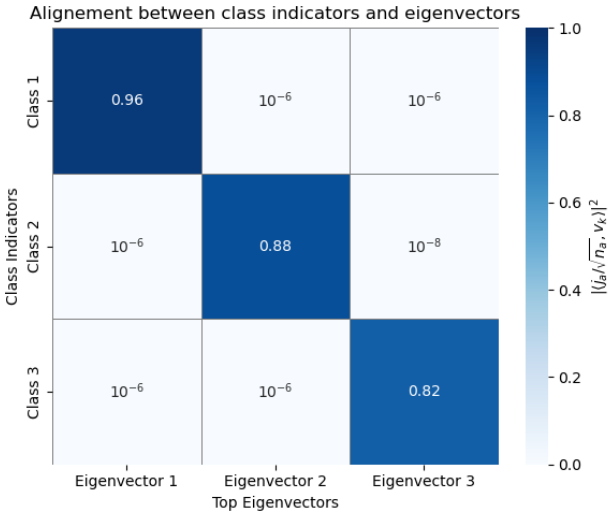

Figure 8: Heatmap of the values of the alignment between the $K = 3$ isolated eigenvectors and the class indicators. The parameters used are defined in Tab. 2 with $n = 5000$

3.7 Evaluating the performance of our algorithm in the homogeneous case

3.7.1 Estimating the modularity matrix

In general one has access only to A and needs to build B from A to use our suggested algorithm. While it is difficult in the general case to estimate q_i from the adjacency matrix, it is tractable to estimate q_0 in the homogeneous case. Indeed, we notice:

$$\sum_{ij} A_{ij} = \sum_{a,b} \sum_{i_a j_b} A_{i_a j_b} \underset{LLN}{=} \sum_{a,b} n_a n_b \mathbb{E}[A_{i_a i_b}] = \sum_{a,b} c_a c_b n^2 q_0^2 (1 + \frac{M_{ab}}{\sqrt{n}})$$

$$\frac{1}{n^2} \sum_{ij} A_{ij} \to \sum_{ab} c_a c_b q_0^2 = \sum_a c_a \sum_b c_b q_0^2 = q_0^2$$

We now can build B, find the isolated eigenvectors, perform an embedding and use k-means on this embedding.

3.7.2 Modeling the noise

To be able to tell how close to optimality our algorithm is, it would be useful to have an estimate of the theoretical error.We notice in Fig. 5 that the extracted eigenvectors look like noised versions of the class indicators. We can model this noise as Gaussian and try to estimate it. We suggest the following model:

$$v_k = \gamma_k (j_k + Z_k) \text{ with } Z_k \sim \mathcal{N}(0, \sigma_k^2 I_n)$$

and a scaling factor γ_k to account for the fact we until now only have results about alignment, not the magnitude of the eigenvectors. We chose to fix the scaling factor by imposing $||v_k|| = 1$ which gives the equation:

$$\gamma_k^2 \langle j_k + Z_k \,|\, j_k + Z_k \rangle = 1 \tag{19}$$

We write

$$\alpha_k = 1 - \left(\frac{\sigma}{q_0^2 m_k c_k} \right)^2$$

and Eq. 17 becomes:

$$|\langle \frac{j_k}{\sqrt{n_k}} \,|\, v_k \rangle|^2 = |\langle \frac{j_k}{\sqrt{n_k}} \,|\, \gamma_k (j_k + Z_k) \rangle|^2 = \alpha_k \tag{20}$$

Using the law of large numbers, we approximate the following terms:

$$\langle j_k \,|\, Z_k \rangle \approxeq 0$$
$$\langle Z_k \,|\, Z_k \rangle \approxeq n\sigma_k^2$$
$$(\langle j_k \,|\, Z_k \rangle)^2 \approxeq n_k \sigma_k^2$$

and Eq. 19,20 become

$$\gamma_k^2(n_k + n\sigma_k^2) = 1$$
$$\gamma_k^2(n_k + \sigma_k^2) = \alpha_k$$

from which we deduce as $n \to \infty$

$$\sigma_k^2 \to \frac{1-\alpha_k}{\alpha_k} c_k \tag{21}$$

$$n\gamma_k^2 \to \frac{1}{c_k + \sigma_k^2} = \frac{\alpha_k}{c_k} \tag{22}$$

We confirm this result experimentally:

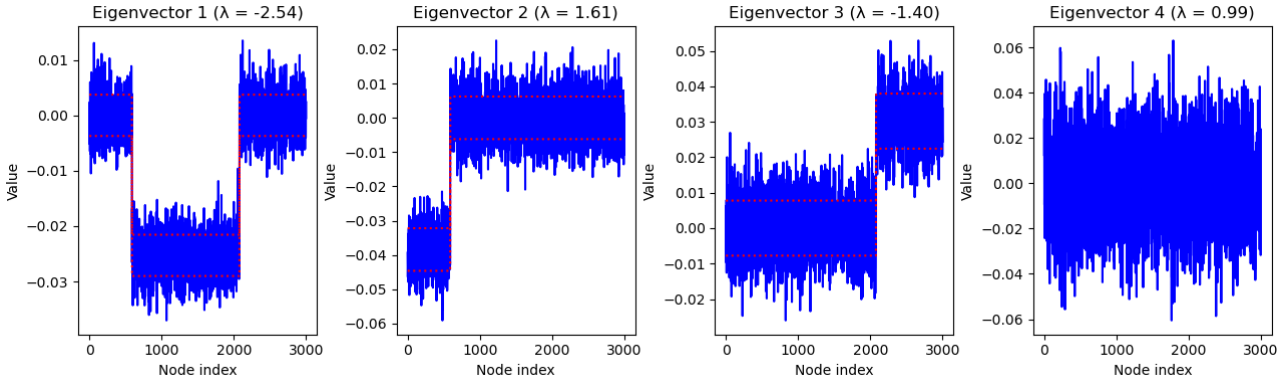

Figure 9: Plot of the eigenvectors associated with the 4 largest eigenvalues of the modularity matrix. The dashes represent $j_k \pm \sigma_k$ with σ_k the theoretical class variance defined in Eq. 21 The parameters used are defined in Tab. 2 with $n = 3000$

Although the Gaussianity of the noise was suggested by the central-limit theorem, we can now confirm this hypothesis by plotting the distribution of $\gamma_1 j_1 - v_1$ and comparing it to $\mathcal{N}(0, \gamma_1^2\sigma_1^2)$ as done in Fig. 10.

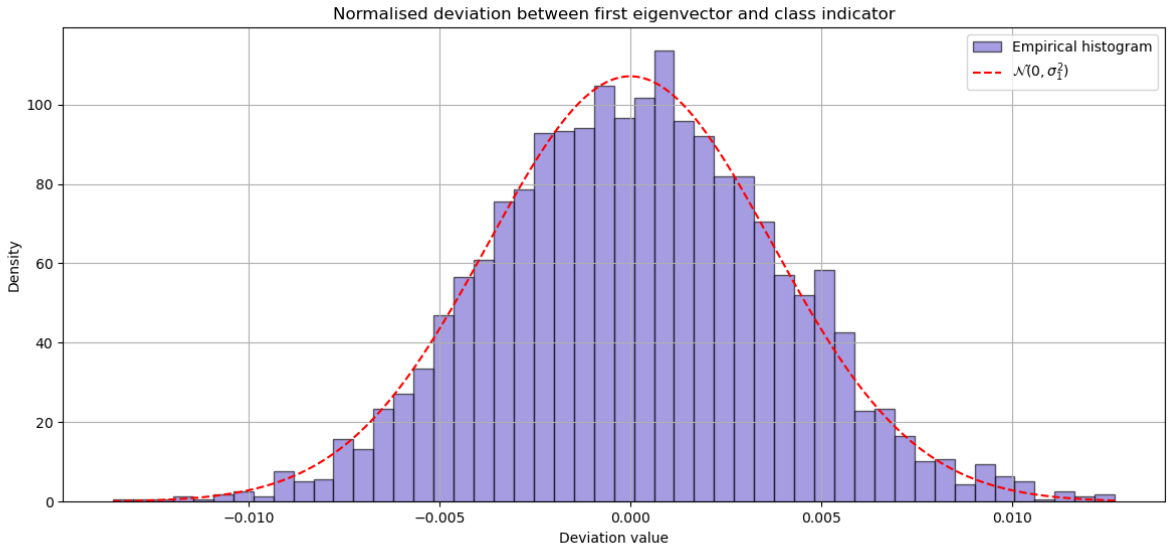

Figure 10: Histogram of the normalised deviations between the isolated eigenvector associated with the largest eigenvalue and the corresponding class indicator appropriately scaled. The theoretical noise $\mathcal{N}(0, \gamma_1^2\sigma_1^2)$ defined in Eq. 21 is plotted for comparison. The parameters used are defined in Tab. 2 with $n = 3000$

3.7.3 Estimating the error

Having modeled the noise, we now model the probability of misclassifying a node. We assume node a belongs to class i. We embed it using the components$(v_1^{(a)}, \cdots, v_K^{(a)}) = \tilde{v}_a$. Noting $\tilde{j}_k = (\delta_{ki})_{i=1:K}$ we can write

$$\tilde{v}_a = \gamma_i \tilde{j}_i + \tilde{Z}_a, \tilde{Z}_a \sim \mathcal{N}(0, \text{diag}(\gamma_1^2\sigma_1^2, \cdots, \gamma_K^2\sigma_K^2))$$

The probability of misclassifying a is the probability of $\tilde{v}_a$ being closer to at least another $\tilde{j}_k$ than $\tilde{j}_a$ which we rewrite as:

$$\mathbb{P}(\text{error}|\text{true class} = \text{i}) = \mathbb{P}\left(||\tilde{v}_a - \gamma_i\tilde{j}_i||^2 \geq \min_{k\neq i} ||\tilde{v}_a - \gamma_k\tilde{j}_k||^2\right)$$

Note that this means $\tilde{v}_a$ could be closer to more than one $\tilde{j}_k$ than to $\tilde{j}_a$. This is the core difficulty in computing this error as it forces us to compute the measure of Voronoi cells instead of simple half-spaces. We define the Voronoi cell V_i as

$$V_i = \left\{x \in \mathbb{R}^K \;\middle|\; \|x - \gamma_j\tilde{j}_i\| \leq \|x - \gamma_k j_k\| \quad \forall k \neq i\right\}$$

and

$$\mathbb{P}(\text{error}|\text{true class} = \text{i}) = \int_{V_i} \mathcal{N}(0, \text{diag}(\gamma_1^2\sigma_1^2, \cdots, \gamma_K^2\sigma_K^2))dx^K$$

$$\mathbb{P}(\text{error}) = \sum_{i=1}^{K} \mathbb{P}(\text{error}|\text{true class} = \text{i})c_i$$

To have intuition for this error we can still try to compute analytically the probability of misclassifying node a, whose true class is i, as being from class k

$$\mathbb{P}(\text{predicted class} = \text{k}|\text{true class} = \text{i}) = \mathbb{P}(||\tilde{v}_a - \gamma_i\tilde{j}_i|| \geq ||\tilde{v}_a - \gamma_k\tilde{j}_k||)$$

$$||\tilde{v}_a - \gamma_i \tilde{j}_i|| \geq ||\tilde{v}_a - \gamma_k \tilde{j}_k|| \Leftrightarrow \gamma_k^2 \mathcal{N}(0, \sigma_k^2) - \gamma_i^2 \mathcal{N}(0, \sigma_i^2) \geq \frac{\gamma_i^2 + \gamma_k^2}{2}$$

This expression makes geometric sense as it means the noise represents more than half the distance between the two classes. Using Eq. 21 and Eq. 22 we rewrite this as

$$\mathcal{N}(0,1) \geq \frac{1}{2} \left(\frac{\frac{\alpha_i}{c_i} + \frac{\alpha_k}{c_k}}{\frac{\alpha_i \sigma_i}{c_i} + \frac{\alpha_k \sigma_k}{c_k}} \right) \tag{23}$$

Hence the probability of error is:

$$\mathbb{P}(\text{predicted class} = \text{k} — \text{true class} = \text{i}) = 1 - \Phi \left(\frac{1}{2} \frac{\frac{\alpha_i}{c_i} + \frac{\alpha_k}{c_k}}{\frac{\alpha_i \sigma_i}{c_i} + \frac{\alpha_k \sigma_k}{c_k}} \right) \tag{24}$$

with Φ the repartition function of the normal law.

4 Dealing with the heterogeneous case

In practice, node connectivity is not homogeneous. Can we adapt our algorithm to perform in this case as well?

4.1 Where our algorithm fails in the heterogeneous case

As seen in the preliminary remarks, we remark that for very heterogeneous distributions of q our algorithm no longer works as the embeddings splits in multiple regimes corresponding to different connectivities. Whenever q_i is allowed to be small for some nodes, their embeddings gather near the origin of $\mathbb{R}^K$. This makes a compact cluster where it is difficult to distinguish between classes. This behaviour is due to the signal containing information about connectivity q. This information is then found in the eigenvectors.

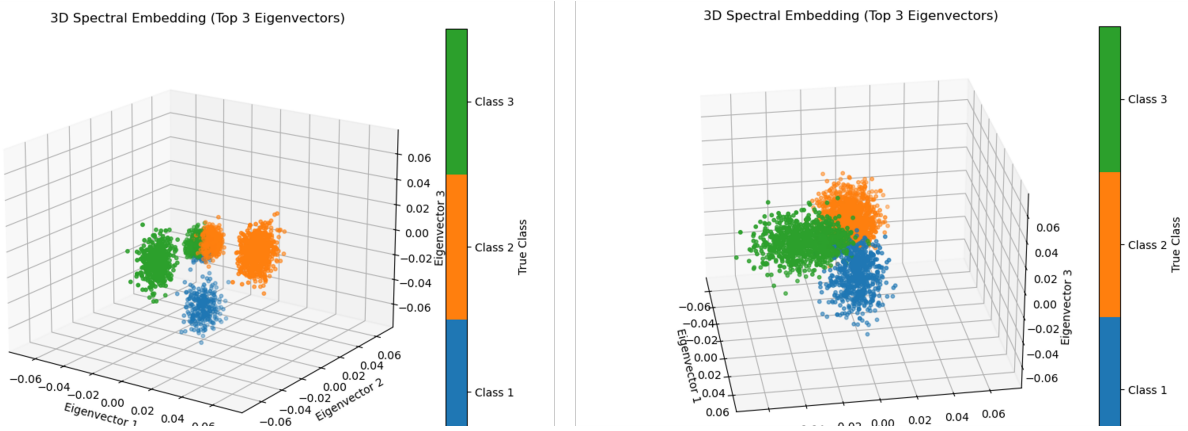

Figure 11: Node embeddings for heterogeneous q_i.Parameter values: $n = 1000, K = 3, M = \text{diag}(-14, 15, 13), c_1 = 0.2, c_2 = 0.5, c_3 = 0.3$ **Left** $q_i \sim \mathcal{U}(\{0.1, 0.8\})$, **Right:** $q_i \sim \mathcal{U}([0.1, 0.8])$.

The issue essentially arises in the bimodal case as a significant portion of the graph cannot be classified correctly (the center cluster). This is clear in Fig. 12 where, in the bimodal case, k-means has an accuracy of 76%. One can see k-means incorrectly classifies the entire center cluster as belonging to class 2. This is caused by γ_2 being small enough (since c_2 is large) that the distant orange cluster is closest to every center node. Since 50% of the nodes are in the center cluster and 50% of those are from class two, this justifies the 76% accuracy. In contrast, in case-2, since only few nodes are associated with a small q_i, a small minority is misclassified.

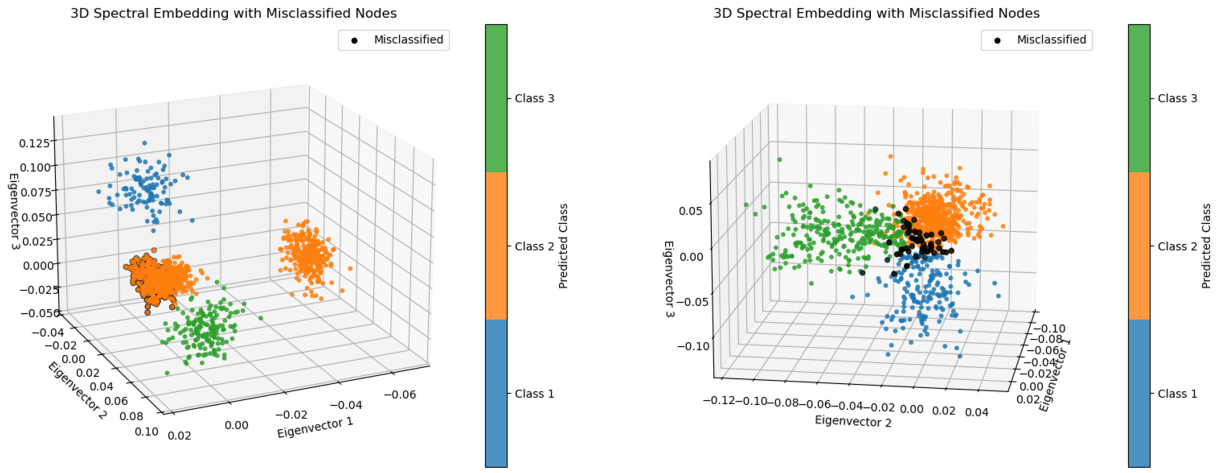

Figure 12: Predicted classes by k-means after Hungarian matching. Misclassified nodes are in black.Parameter values: $n = 1000, K = 3, M = \text{diag}(-14, 15, 13), c_1 = 0.2, c_2 = 0.5, c_3 = 0.3$ **Left** $q_i \sim \mathcal{U}(\{0.1, 0.8\})$, Accuracy = 76% **Right:** $q_i \sim \mathcal{U}([0.1, 0.8])$, Accuracy = 94% .

4.2 Approaches to tackling this issue

4.2.1 Matrix normalisation:

A first idea which is very common when dealing with graphs is normalisation: One can limit the influence between different values of q in the signal by multiplying B by the inverse of the degree matrix $D = \text{diag}(\sum_{i=1}^{n} A_{ij})_{i=1:n}$ One can then compute $D^{-1}B$ which will rescale the importance of the columns of B so as to not give excessive importance to columns corresponding to highly connected nodes. We will favour performing a symetric form of normalisation so as to keep B symmetric:

$$\tilde{B} = D^{-\frac{1}{2}} B D^{-\frac{1}{2}}$$

In the case $q_i \sim \mathcal{U}(\{\,0.1, 0.8\,\})$, taking the same parameter values as in Fig. 11, we now use symmetric matrix normalisation on B and obtain:

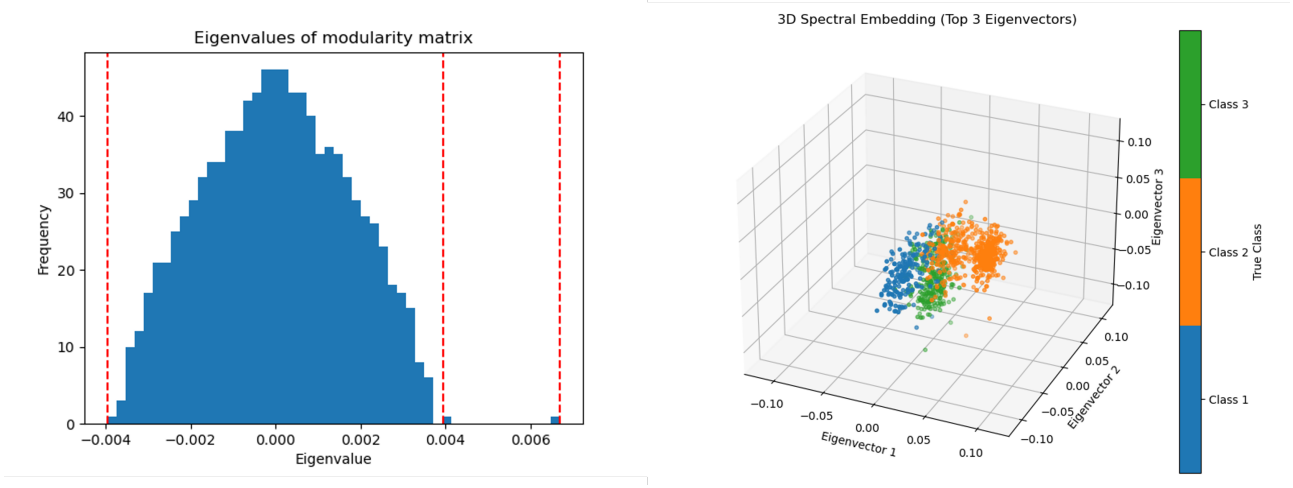

Figure 13: Eigenvalue distribution (Left) and node embeddings (Right) for a symmetric normalisation of B: $\tilde{B} = D^{-\frac{1}{2}}BD^{-\frac{1}{2}}$ and heterogeneous$q_i$distribution $q_i \sim \mathcal{U}(\{0.1, 0.8\})$. The largest 3 eigenvalues are dotted in red. Classification accuracy: 86% Parameter values: $n = 1000, K = 3, M = \text{diag}(-14, 15, 13), c_1 = 0.2, c_2 = 0.5, c_3 = 0.3$

As shown in Fig. 13, after matrix-normalisation, the embeddings are no longer divided in sub-clusters corresponding to different connectivity regimes. With an accuracy of 86% we see matrix-normalisation had a positive impact.

4.2.2 Eigenvector normalisation:

We can notice in Fig. 11 that when the clusters split into connectivity-related clusters, the sub-clusters related to a given class are not completely unrelated. They indeed seem to be aligned along the same direction. In order to leverage this angular information, we project the embedding on a sphere by normalising all the embedding vectors so that they have unit norm in the embedding space. We apply this embedding normalisation in the same example as previously, without matrix normalisation:

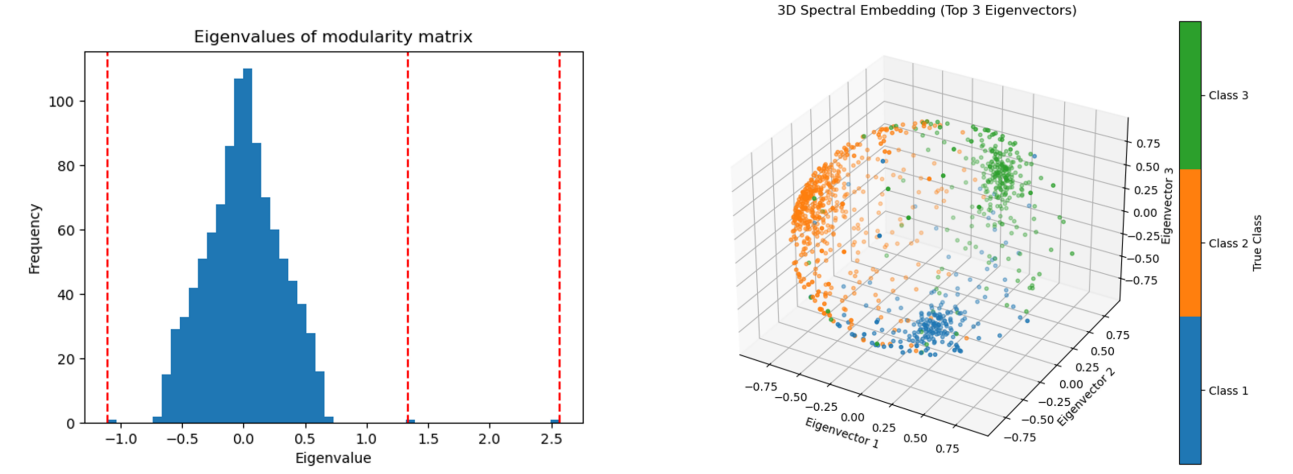

Figure 14: Eigenvalue distribution (Left) and node embeddings (Right) for normalised embeddings and heterogeneous q_i distribution $q_i \sim \mathcal{U}(\{0.1, 0.8\})$.The largest 3 eigenvalues are dotted in red. Classification accuracy: 86% Parameter values: $n = 1000, K = 3, M = \text{diag}(-14, 15, 13), c_1 = 0.2, c_2 = 0.5, c_3 = 0.3$

In Fig. 14, we notice projecting onto the sphere effectively undoes the multiplicity caused by the bimodal q_i distribution. We obtain the exact same accuracy with vector normalisation as we did for matrix normalisation. However, we can see that for the same parameter values, it is easier to extract isolated eigenvalues for an unormalised B since in Fig 13, the third largest eigenvalue is on the edge of the bulk. This suggests that matrix normalization will more often fail to extract three meaningful eigenvectors and work accordingly. In addition, normalising K-dimensional vectors is computationally easier than performing $n \times n$ matrix multiplication, so we will favour the vector normalisation approach.

As a final remark, we should notice that these methods work only if we have access to B and hence to q. This is far from obvious when q is heterogeneous. One approach we tried with some success was to approximate qq^T with a normalised dd^t with $d = (\sum_j A_{ij})_{i=1:n}$

5 Conclusions and future work

We have put forward an unsupervised community detection algorithm grounded in random matrix theory. It consists in using the largest eigenvectors of the modularity matrix to construct a low dimensional embedding in which kmeans is applied. We obtained theoretical guarantees in the case of homogeneous connectivity and suggested using eigenvector normalisation in the general case. Performance was tested against simulated graphs according to the Degree-Corrected Stochastic Block Model. A next step would be to test this algorithm on real graph datasets. We emphasize our method only succeeds if one has access to the modularity matrix B and hence the connectivities q. Deducing q from the adjacency matrix A is far from obvious when q is heterogeneous. One approach which could be tried is to approximate qq^T with a normalised dd^t with $d = (\sum_j A_{ij})_{i=1:n}$