SPECTRAL CLUSTERING FOR MULTICLASS ERDÖS-RÉNYI GRAPHS

Mohamed-Ali Belabbas
Harvard University
School of Engineering and Applied Sciences
Cambridge, MA 02138

ABSTRACT
In this article, we study the properties of the spectral analysis of multiclass Erdös-Rényi graphs. With a view towards using the embedding afforded by the decomposition of the graph Laplacian for subsequent processing, we analyze two basic geometric properties, namely interclass intersection and interclass distance. We will first study the dyadic two-class case in details and observe the existence of a phase transition for the interclass intersection. We then focus on the general multiclass case, where we introduce an appropriate notion of diagonal concentration and derive a statistical model that allows sampling graphs whose expected diagonal concentration is fixed. The simulations provided yield useful guidelines for practitioners to choose appropriately parameters in the context of spectral clustering.

Index Terms— Community detection, Non-Euclidean datasets, Random graph models, Spectral graph theory

1. INTRODUCTION
From biology to anthropology to computer science, the detection of groups in graphical datasets is a ubiquitous problem in science and engineering [1, 2, 3]. In this paper, we will adopt a statistical approach and focus our investigation on the classical ensemble of Erdös-Rényi graphs through their Laplacians.

Roughly speaking, our objective in this paper is to answer the question of how much can one hope to extract from the spectral embedding of mixtures of Erdös-Rényi (ER) graphs. In practice, this embedding is often followed by either a machine learning algorithm such as an SVM, a linear classifier or the k-means [4] algorithm, or by an estimation of the hidden parameters of the model. Our results will impose limits on how well one can hope these methods will perform. Specifically, the contributions of this paper are in the definition and evaluation of geometric measure that describe the suitability of the spectral embedding for subsequent processing and the ensuing rules for choosing parameters. As a consequence of this analysis, we will also uncover a new, as far as the author is aware, phase transition related to the spectral embedding.

From a historical perspective, using the spectrum of the Laplacian to understand the geometric properties of a space is an old idea that finds its root in differential geometry [5] and that has found applications in applied mathematics [6] and graph theory [7] early on. It is only in recent years that it has seen a widespread use in signal processing, pattern recognition and decision theory, where it provides new tools to understand non-Euclidean datasets [8].

The paper is organized as follows: after a brief overview of the relevant notions from spectral graph theory, we introduce two simple geometric measures of the embedding of a graph. We then analyze in details the case of 2-class graphs before discussing the general case of \( n \)-classes. For the latter part, we introduce the notion of diagonal concentration of the parameters as well as a stochastic model for them.

2. ELEMENTS OF SPECTRAL GRAPH THEORY
Here we very briefly summarize basic proprieties of spectral graph theory, and refer the reader to [7, 9] for more information.

Given a graph \( G = (V, E) \) where \( V \) is the set of vertices, or nodes, and \( E \) is the set of edges, we define the degree of a vertex as the number of edges originating from it. We assume throughout the paper that \( |V| = m \).

We define the adjacency matrix by

\[
A_{ij} = \begin{cases} 
1 & \text{if } (i, j) \in E, \\
0 & \text{otherwise},
\end{cases}
\]  

and the degree matrix \( D \) as a diagonal matrix whose entries are the degrees of the nodes. The graph Laplacian is then given by

\[
L(G) = D - A.
\]

This definition is derived from the usual Laplacian \( \Delta = \sum_i \frac{\partial^2}{\partial x_i^2} \) in \( \mathbb{R}^n \), as the two operators exhibit similar properties on their respective domains of definition.

The importance of the graph Laplacian for community detection and more broadly for network analysis comes from a result of Fiedler [10]. He showed that for the eigenvectors associated with the smallest nonzero eigenvalue, nowadays called Fiedler vectors, the subgraph induced by nonpositive vertices (i.e., vertices of \( G \) for which the corresponding entry of the eigenvector is nonpositive) and the subgraph induced by nonnegative vertices are both connected.

This suggests using the eigenvectors \( (u_1, \ldots, u_n) \) of the Laplacian, ordered such that the corresponding eigenvalues are non-decreasing, to embed the data in \( \mathbb{R}^k \), for \( k \leq m - 1 \) according to

\[
v_k = (u_{2k}, u_{2k+1}, \ldots, u_{k+1}).
\]

This method is known as spectral graph embedding [7]. We will denote throughout the paper by \( v_i \) the spectral embedding of node \( i \) as given by the above equation.

2.1. Erdös-Rényi Graphs
We now consider a definition of multiclass Erdös-Rényi graphs, also known as Erdös-Rényi mixtures or simple stochastic blockmodels. For a graph \( G = (V, E) \), assume we are given a function \( c(\cdot) \) from the set of vertices to \( \{1, \ldots, n\} \). This function characterizes the class to which node \( i \) belongs. We define the Bernoulli parameter

\[
\theta_{ij} = P(\text{node } i \text{ connects to node } j) = \begin{cases} 
\theta & \text{if } c(i) = c(j), \\
1 - \theta & \text{otherwise},
\end{cases}
\]  

for \( 0 < \theta < 1 \) and \( i, j \neq k \). The graph Laplacian is then given by

\[
L(G) = \sum_{i=1}^{n} \sum_{j=1}^{n} \delta_{ij} \frac{\partial^2}{\partial x_i^2} \frac{\partial^2}{\partial x_j^2},
\]

where \( \delta_{ij} \) is the Kronecker delta function. The eigenvalues and eigenvectors of this Laplacian are used to define the spectral embedding of node \( i \), which is then used to infer the class to which node \( i \) belongs.
matrix $P$ to be an $n \times n$ symmetric matrix with entries $p_{kl} \in [0,1]$, where each entry corresponds to a given class $c(i)$.

The probability of having an edge between vertices $i$ and $j$ is then given by a Bernoulli trial with parameter $p_{c(i),c(j)}$, independently of the presence or absence of other edges. In terms of the adjacency matrix, we have

$$\Pr(A_{ij} = 1) = p_{c(i),c(j)}.$$  

We say that the graph is balanced if each class contains the same number of vertices. We will also randomize the class function $c$ by introducing the vector $\mu$ which belongs to the unit simplex in $\mathbb{R}^n$ and setting $\Pr(c(i) = k) = \mu(k)$. Clearly, there are $n^m$ possible class assignments; we call $C$ the set of all such possible $c$.

In this randomized case, we obtain from Bayes’ rule

$$\Pr(A_{ij} = 1) = \sum_{c \in C} \Pr(A_{ij} | c) \Pr(c)$$

$$= \sum_{c \in C} p_{c(i),c(j)} \Pr(c)$$

where $\Pr(c) = \prod_{i=1}^n \mu(c(i))$. It is clear that evaluating these probabilities exactly is computationally expensive and we will thus use Monte-Carlo simulations in this paper.

3. MEASURING THE QUALITY OF AN EMBEDDING

As mentioned above, spectral approaches are often the first step in a more elaborate algorithm, and can be followed by, e.g., $k$-means or hierarchical clustering over the embedded graph. Here we focus exclusively on this first stage, and consider two fundamental geometric properties:

1. Interclass distance: The minimal distance, in the embedding, of points belonging to different classes;

2. Interclass Intersection: The amount by which the convex hulls of vertices belonging to different classes intersect.

In general, deciding whether or not two polyhedra intersect is an NP-complete problem [11]. Therefore we need to seek an approximation that can be evaluated quickly enough to yield meaningful computational experiments.

Precisely, we define the Interclass distance as

$$ICD = \min_{i,j \in V : c(i) \neq c(j)} \|v_i - v_j\|.$$  

Furthermore, we define $ICI(i,j)$, the Interclass Intersection for classes $i$ and $j$, as the number of vertices of $i$ that belong to the convex hull of the vertices of $j$. Observe that $ICI(i,j)$ is not symmetric, i.e. $ICI(i,j)$ is not necessarily equal to $ICI(j,i)$. The interclass intersection is then given by

$$ICI = \sum_{i,j} ICI(i,j)$$

Note that this is an approximation since the convex hulls of two set of points may have a non-empty intersection, while no points of one set belongs to the hull of the other.

4. THE DYADIC CASE

We let

$$P = \begin{bmatrix} p_1 & p_{12} \\ p_{12} & p_2 \end{bmatrix}$$

be the parameter matrix of interest for the dyadic case. Throughout this section, we fix $p_1 = p_2$ and we consider a spectral embedding in $\mathbb{R}^2$.

We will establish lower bounds on the performance of the spectral approach by conducting three experiments that address different characteristics of the output. First, for a fixed value of $P$ we show the density of the embedded vertices in $\mathbb{R}^2$ in two cases: balanced graphs and unbalanced graphs, drawn at random from the model of Section 2.1. The plots in Figure 1 show that balanced classes appear to be on average better separated than unbalanced ones. We will come back to this later in this section.

Fig. 1. Panel (a): The density of vertices of balanced dyadic ER graphs embedded in $\mathbb{R}^2$ with $p_1 = p_2 = 0.6$ and $p_{12} = 0.2$. The density is estimated using $10^5$ draws on graphs with $|V| = 30$. Panel (b): The density of unbalanced ER graphs with all other parameters set as in Panel (a). Without loss of generality, we embedded the largest group consistently on the right, which explains the asymmetry in the density.

While the density provides a good indication as to what the performance of the spectral approach is, there is a loss of information during the averaging over nodes of a same class. In contrast, the following experiments—evaluating the ICD and ICI as functions of $P$—do not perform such an averaging.

Fig. 2. Panel (a): ICD for 2-class Erdős-Rényi graphs as a function of $P$. The ICD is averaged over $10^5$ draws for each value of $P$; the graphs have 40 vertices and 20 in each class Panel (b): ICI for 2-class Erdős-Rényi graphs as a function of $P$, computed using the same set of samples as in Panel (a).

Though it is intuitively clear that separating the classes will be difficult for values of $P$ corresponding to weak in-class edge probability coupled to strong out-of-class edge probability, the experiments Figure 2 allows us to quantify this statement. Additionally,
there are two remarkable aspects to these results. In Figure 2(a), we see that the ICD is essentially zero below a fairly well-marked boundary, above which it then increases close to linearly. More surprising perhaps, the ICI seem to undergo a phase transition along a similarly shaped boundary. Phase transitions in Erdős-Rényi graphs have been observed before [12, 13] but, as far as the author is aware, never in this context. This suggests the existence of a simple criterion for when the spectral embedding will fail to separate different classes, which will form the subject of future work.

We conclude this section with an investigation of how the above results depend on the relative size of the classes. In Figure 3, we show in Panel (a) (resp. Panel (b)) the ICD (resp. the ICI) as a function of the number of elements of class 1 for a fixed value of $P$.

![Figure 3](image1)

**Fig. 3.** Panel (a): ICD as a function of the number of elements in class 1 of for $p_1 = p_2 = 0.6$ and $p_{12} = 0.2$. The ICD is averaged over $10^5$ randomized draws. Panel (b): ICI on the same set of samples as in Panel (a).

We observe that the ICD is roughly constant over the whole range of cardinalities of the subclasses, but increases strongly as one class becomes much smaller than the other. This stems from the fact that, as the ratio of class sizes becomes very skewed, the embedding of the points corresponding to the class of larger cardinality tend to concentrate around their center of mass, as shown in the example of Figure 4(b), whereas the embedding of the less populated class has a much larger variance, which explains that on average the interclass distance will be higher.

The ICI is affected in the opposite manner by this phenomenon: whereas balanced classes tend to be relatively concentrated around their centers of mass, which are rather close as per Figure 3(a), the higher variance of the embedding of the less populated class increases the chance that its convex hull will surround points of the more populated class. This is illustrated in Figure 4.

![Figure 4](image2)

**Fig. 4.** Panel (a): Example of embedding of in 2D of a graph with equal class sizes. Panel (b): Example of embedding of in 2D of a graph with extremely unbalanced class sizes, with the main grouping consisting of 34 points clustered closely together.

5. THE $N$-ADIC CASE

We now turn our attention to the general $n$-class problem. The simulations of the previous section hint at the fact that the ratio of the diagonal elements of $P$ to its off-diagonal elements play an important role in determining the properties of the resultant spectral embedding. Taking inspiration from this observation, we define below a new probability distribution over the space of matrices $P$ which allows us to perform our computational experiments in the $n$-adic case.

5.1. A stochastic model

We introduce the following definition of the diagonal concentration of a binomial parameter matrix $P$,

$$r(P) = 2 \frac{\text{trace}(P)}{e^T P e - \text{trace}(P)},$$

where $e$ is the $n$-vector of all 1.

Because its parameters allow a variety of densities over the interval $[0, 1]$, the beta distribution is very useful for modelling the entries of $P$. We thus rely on it to specify a distribution for binomial parameter matrices for which $E(r(P)) = r$ is fixed by the user. For ease of derivation of the model, we will assume that the diagonal entries are i.i.d. beta($\alpha_{on}, \beta_{on}$) and that the off-diagonal entries are also i.i.d. but distributed according to beta($\alpha_{off}, \beta_{off}$).

Under these assumptions, the expected diagonal concentration is given by

$$E(r(P)) = nE(\text{beta}(|\alpha_{on}, \beta_{on}|)) = \frac{1}{\sum_{i=1}^{n(n-1)/2} \text{beta}(\alpha_{off}, \beta_{off})}$$

(2)

Determining the parameters $\alpha_{on}, \beta_{on}, \alpha_{off}, \beta_{off}$ such that the expected diagonal concentration is given by $r$ is an ill-posed problem, hence we will fix some of these parameters now. We first recall that the expected of a beta random variable is $E(\text{beta}(\alpha, \beta)) = \frac{\alpha}{\alpha + \beta}$. Intuitively, we want the trace of $P$ to be larger than the sum of its off-diagonal elements, so that the samples graphs have on average more connections within a given class than between classes. We thus take $\alpha_{on} = n - 2$ and $\beta_{on} = 2$ and thus the expectation of the diagonal elements will be $\frac{n-2}{n}$. The first term in Equation (2) is then

$$\frac{n\alpha}{\alpha + \beta} = (n - 2).$$

The second term is harder to deal with, but following the same intuition as above, we take $\alpha_{off} = 2$ and seek to find $\beta_{off}$ so that the diagonal concentration is $r$. One then finds after some calculations that

$$E \left[ \sum_{i=1}^{n(n-1)/2} \text{beta}(2, \beta_{off}) \right] = \frac{1}{2n - 1} \beta_{off} + \frac{1}{n}.$$  

(3)

A simple substitution in Equation (2) then yields

$$\beta_{off} = \left( r - \frac{2n-4}{n(n-1)} \right) \left( n(n-1) - 1 \right).$$

(4)

5.2. Simulations

Having established a model that allows us to sample random graphs while controlling the associated diagonal concentration, we can
study how the ICI and ICD depend on the number of classes and dimensionality of the embedding.

From Figure 5, we observe that having a large number of classes \( n \) appears to decrease the quality of the embedding, as the ICD decreases while the ICI increases with \( n \), faster than a larger embedding dimension or a higher diagonal concentration help improve it. Indeed, we plot in Figure 5(a) the ICD for graphs with expected diagonal concentration \( r = 10 \) as a function of the number of classes and the embedding dimension and observe a sharp drop as \( n \) increases. Similarly, we plot in Figure 5(b) the ICI as a function of the number of classes and the diagonal concentration and observe that as \( n \) increases and for fixed \( r \), the ICI increases almost linearly.

![Figure 5](image-url) **Fig. 5. Panel (a):** ICD as a function of the embedding dimension and the number of classes for random ER graphs with \( r = 10 \). The graphs have 40 nodes and each point in the graphic is averaged over 100 Bernoulli parameter matrices \( P \) and 100 realizations of graphs per drawn \( P \). **Panel (b):** ICI as a function of the diagonal concentration and the number of classes for fixed \( k = 2 \). The graphs have 50 nodes and all other parameters are similar to the one of Panel (a).

We conclude by showing the ICI and ICD as a function of the diagonal concentration in Figure 6. As the dimension increases, we see that the ICI is relatively low and constant with regard to the diagonal concentration, see Figure 6(b). On the contrary, we observe in Figure 6(a) a steady increase in the ICD, which is more pronounced as the dimension increases.

![Figure 6](image-url) **Fig. 6. Panel (a):** ICD as a function of the diagonal concentration for embedding dimensions 2, 3 and 4. The curves are averaged over 100 random \( P \) and for each \( P \) we sampled 100 random graphs of size 40 with 5 classes. **Panel (b):** The ICI in a setting similar to the ones of Panel (a).

## 6. SUMMARY AND OUTLOOK

Whether one seeks to estimate the underlying parameters of multiclass Erdős-Rényi random graphs or to perform some processing on a realization of such a graph, the embedding obtained from a spectral analysis of the Graph Laplacian plays a distinguished role. In order to better understand the fundamental limits of the spectral point of view, we have investigated two basic measures of the quality of this embedding, termed ICI and ICD.

In the dyadic case, we observed a phase transition for the ICI, suggesting the existence of a threshold below which spectral methods fail to yield good results. While we were able to evaluate the ICI and ICD for all possible values of \( P \) in the dyadic case, we introduced the notion of diagonal concentration of the Bernoulli parameter matrix in order to understand the \( n \)-adic case. We derived a stochastic model for matrices of constant expected diagonal concentration and used it as a basis of our computational experiments, which in turn suggested guidelines for choosing the dimension of the corresponding embedding. In both cases, unbalanced classes tend to be harder to isolate that balanced ones, though we have seen that competing effects are at play.

## 7. REFERENCES


