Università degli Studi di Padova

Department of Mathematics "Tullio Levi-Civita"

Master Degree in Mathematics

Stochastic Block Model with k communities: a spectral algorithm with optimal recovery

Master thesis

Supervisor

Prof. Marco Formentin

Candidate

Veronica Todesco

1206679

Academic Year 2020-2021

23RD April, 2021

Contents

Introduction **3**

1 Community detection in networks **8**
1.1 Graphs and random graphs 9
1.2 The Community Detection Problem 11
1.3 Different methods to solve the community detection problem . . 14
1.4 The Stochastic Block Model 18

2 Spectral method for a dense Stochastic Block Model with two communities **20**
2.1 Dense and sparse graphs 20
2.2 The spectral method 22
2.2.1 Step 1: bounding the error 24
2.2.2 Step 2: application of Davis - Kahan Theorem 25
2.2.3 Résumé of the method as Spectral Algorithm 26

3 Spectral method for a sparse Stochastic Block Model with two communities **28**
3.1 Spectral Partition: a first algorithm producing a correct vertex partition 30
3.2 Refinement of the method with the addition of a correction algorithm 48

4 Spectral method for a Stochastic Block Model with k communities **55**

Conclusions **75**

A Appendix **77**
A.1 Concentration inequalities 77
A.1.1 Markov inequality 77
A.1.2 Chernoff inequality 78
A.1.3 Bernstein inequality 80
A.2 Binomial coefficient 80
A.3 Properties on matrices 80
A.3.1 Eigenvalues and singular values 80
A.3.2 Matrix norm 81
A.4 ε-nets 82
A.5 Orders of approximation 83

Introduction

This thesis discusses a spectral method to solve the community detection problem in a Stochastic Block Model with k communities.

Network science is a modern discipline spanning the natural, social and computer science, as well as engineering, biology, economics and ecology. In many real-world networks it is possible to highlight a community structure, namely we can identify clusters of vertices that are strongly connected to each other. We should consider a community as a subset of vertices sharing the same idea, the same belief or that are linked by a particular relation. This is why the field of community detection has arised, becoming an important topic in modern network science.

Starting from random graph theory, in the last decades mathematicians have proposed different techniques to solve the community detection problem. In this thesis we discuss a spectral method: a remarkable observation shows that the expected value of the adjacency matrix of the input random graph contains all the information about the community structure of the graph. More precisely, it is possible to detect the clusters gathering the vertices according to their values in the eigenvectors of the expected value of the adjacency matrix. Unfortunately, we cannot deduce such average adjacency matrix just from a realization of the random graph: the spectral technique aims to extract the information about the community structure directly from the adjacency matrix.

In order to develop the method, we consider random graphs generated by the Stochastic Block Model (SBM). This is the simplest and most studied model of random graphs with community structure and it is built with the minimal assumptions: the edge distribution is described by an internal probability within the communities which is greater than the edge probability between vertices belonging to different communities. In this way, the internal density is certainly higher than external densities among communities.

This thesis focuses on the method proposed by P. Chin, A. Rao and V. Vu in [5], and the general layout follows the definitions and results from [16].
We have previously said that the starting point is the SBM: being more precisely, we consider a highly *sparse* SBM, where the mean number of edges per vertex approaches a constant as the number of vertices of the graph grows. In such a situation we cannot exploit the properties of denser graphs, as in the standard method explained in [16]. The main result of [5] is the *optimal* relation between the error rate produced by the method and the quantities describing the edge probabilities within the graph.
Let us take a random graph such that

- every pair of vertices is connected with probability $\frac{a}{n}$ if they belong to the same community;
- every pair of vertices is connected with probability $\frac{b}{n}$ if they belong to different communities,

where $n > a > b > 0$ and the number of vertices of the random graph is $\Theta(n)$.

First consider the easier case of $k = 2$ communities. The problem can be seen as a variant of the hidden bipartition problem, which has been studied by many researchers in theoretical computer science, starting with the work of Bui et al., [3]. Earlier papers, like [12], deal with a and b large, and it is known that we can efficiently obtain a complete recovery if $a, b \geq C \log n$ and $\frac{(a-b)^2}{a+b} \geq C\frac{\log n}{n}$ for a sufficiently large constant C (see for example [17]). Moreover, taking an even denser graph in which the internal/external probabilities are $p > q > 0$ fixed (so that they do not decrease as n^{-1}) like in [16], with high probability we can recover the communities with a spectral clustering algorithm correctly up to a small number of mislabeled vertices.
This work, on the contrary, deals with sparser random graphs. In a previous paper [6], Coja-Oghlan proved that it is possible to find a vertex partition, up to a small error rate and with high probability, with a polynomial time algorithm if $a, b > C_1$ and $\frac{(a-b)^2}{a+b} > C_2 \log(a + b)$, for suitable constants $C_1, C_2 > 0$. Coja-Oghlan proved this result as part of a more general problem, and his algorithm was rather involved. Furthermore, the result is not yet sharp and it has been conjectured that the log term is removable. A different approach is given, for instance, by the work of Zhang and Zhou, [18]: they proved a minimax rate result that suggested that there is a constant $c > 0$ such that, if $\frac{(a-b)^2}{a+b} \leq c \log \frac{1}{\gamma}$, then with high probability it is not possible to find the partition (in expectation), regardless the algorithm. The term γ refers to an error rate: in this manuscript we will say that a candidate vertex partition is γ-*correct* if the number of mislabeled vertices in each community is less than a fraction γ.
In the paper we focus on, [5], the authors prove, as we will exhibit in this thesis, an optimal relation for a and b involving the error rate γ produced by the method. The *optimality* refers to the fact that we look for a lower bound which is the best possible in relation to the result of Zhang and Zhou. Specifically, we will show that with high probability it is possible to find a γ-*correct* vertex partition with error rate $\gamma > 0$ using a simple spectral algorithm if $a > b > C_1$ and

$$\frac{(a-b)^2}{a+b} \geq C_2 \log \frac{1}{\gamma}$$

for suitable constant $C_1, C_2 > 0$.

In the general case $k > 2$, the problem becomes more complicated, both for the development of a method and for a good choice on the assumptions. In the last part of this thesis, we will prove that, with high probability, we can recover the vertex partition using a simple spectral algorithm and with a small error rate $\gamma > 0$ when $a > b > C_3$ and

$$(a-b)^2 \geq C_4\, k^2\, a\, \log \frac{1}{\gamma}$$

with $C_3, C_4 > 0$ suitable constants.

The thesis is organized as follows:

- in Chapter 1, we first recall some elementary definitions of graph theory. Then, we thoroughly describe the community detection problem and some applications to real networks (biological and social networks, the World Wide Web and other examples). Afterwards, we briefly present largely studied techniques to solve the problem that have different approaches with respect to the spectral technique: methods based on optimisation, methods using random walks and the spin dynamics. Finally we define the Stochastic Block Model, that will be the starting point for the rest of the work;

- in Chapter 2, we exhibit a spectral algorithm to recover the vertex partition in a dense SBM with $k = 2$ communities, following the results of [16]. After defining the difference between dense and sparse random graphs, we formulate a first SBM and we explain the fundamental ideas that lay the foundations of the spectral technique. Since it is possible to deduce the community structure from the expected value of the adjacency matrix, the idea is to rewrite the adjacency matrix as a difference of matrices involving its expected value. Exploiting the properties of dense graphs and applying Davis-Kahan theorem, we prove that the eigenspace of the adjacency matrix is *close* to the eigenspace of its expected value: using a spectral algorithm directly on the adjaceny matrix, with high probability we can find a good approximation of the vertex partition;

- in Chapter 3, we introduce a sparse SBM with $k = 2$ communities, as described in [5]. In this case, we cannot use the properties of dense graphs, so we need to proceed with a different approach. The general idea is the same as in the dense case, but the method requires additional steps, and to bound the involved matrices we use the following trick: we zero-out some rows and columns related to vertices with high degree. In this way, we lose some information on the random graph, but we can apply the spectral algorithm called Partition (Algorithm 3) and find a vertex partition with a small error rate;

- in Chapter 4, we generalize the sparse SBM model introduced in Chapter 3 to the case $k > 2$. As pointed out in [5], it is not obvious how to make approximations when there are $k > 2$ communities. The method requires several additional steps and each step works on different sets of edges and vertices, chosen randomly. We call the resulting algorithm k-Partition (Algorithm 8). In this thesis we prove the *correctness* of such an algorithm and the *optimality* of the relation among the probabilities of the model and the error rate.

To conclude this introductive part, we briefly collect here the main results that we are going to prove in the development of the thesis:

- **Spectral method for a dense Stochastic Block Model with two communities** (Chapter 2): consider a random graph with $2n$ vertices and such that every pair of vertices is connected with probability p or $q < p$ if they belong to the same community or to different ones, respectively.
  Let A be the adjacency matrix of the random graph and let $\bar{A} := \mathbb{E}[A]$. We rewrite A as
  $$A = \bar{A} + R,$$

where the $2n \times 2n$ matrix R is a sort of "noise". We will see that, with high probability,

$$\|R\| \leq C\sqrt{n}$$

for a suitable constant $C > 0$. A key point of the spectral method is that the eigenvector of $\bar{A}$, related to the second higher eigenvalue, contains all the informations about the two communities of the graph. Applying Davis-Kahan Theorem (see Theorem 2) to A and $\bar{A}$, we deduce that the distance between the eigenvectors of A and $\bar{A}$ corresponding to their second higher eigenvalue is small. This result permits to prove that, sorting the vertices according to their values in the second eigenvector of A, with high probability we find a vertex partition which is correct up to C/μ^2 vertices, where $\mu = \min\{p - q, 2q\}$ and $C > 0$ is a suitable constant.

- **Spectral method for a sparse Stochastic Block Model with two communities** (Chapter 3): consider a random graph with $2n$ vertices and such that every pair of vertices is connected with probability $\frac{a}{n}$ or $\frac{b}{n}$, with $n > a > b > 0$, if they belong to the same communitiy or to different ones, respectively.
  As before, let A_0 be the adjacency matrix of the random graph and let $\bar{A_0} := \mathbb{E}[A_0]$. The idea is again to rewrite A_0 involving $\bar{A_0}$, but now it is not easy to find a bound for the norm of their difference. For this reason, we apply the deletion, namely we zero-out the rows and columns of A_0 and $\bar{A_0}$ related to vertices with high degree. The resulting matrices are A and $\bar{A}$ and we put

  $$A := \bar{A} + E := \bar{A_0} + \Delta + E.$$

  We prove that, with high probability,

  $$\|\Delta\| \leq 1 \qquad \text{and} \qquad \|E\| \leq C\sqrt{a+b}$$

  for a constant $C > 0$, and applying a revised version of Davis-Kahan theorem we prove that the eigenspaces of A and $\bar{A_0}$ are close. The resulting algorithm, called Partition (Algorithm 3), consists of two parts: previously a spectral algorithm produces a first candidate vertex partition, then there is a correction process that detects the mislabeled vertices. The final result is that, given as input a SBM with connection probabilities $\frac{a}{n}$ and $\frac{b}{n}$, with high probability the algorithm Partition produces a vertex partition with a small error rate γ when $a > b > C_1$ and

  $$\frac{(a-b)^2}{a+b} \geq C_2 \log \frac{1}{\gamma}.$$

  In particular we prove that when the first part of the algorithm outputs a partition with error rate 0.1, we get the following optimal relation for the final error rate:

  $$\frac{(a-b)^2}{a+b} = \frac{1}{0.072} \log \frac{2}{\gamma} \approx 13.89 \log \frac{2}{\gamma}.$$

- **Spectral method for a sparse Stochastic Block Model with k communities** (Chapter 4): consider a random graph with n vertices and

such that every pair of vertices is connected with probability $\frac{a}{n}$ or $\frac{b}{n}$, with $n > a > b > 0$, if they belong to the same communitiy or to different ones, respectively.
When there are $k > 2$ communities, it is not obvious how to approximate the eigenspaces of the studied matrices. The algorithm developed for this general case is inspired by the algorithm Partition of the 2-communities case, but it requires many additional steps. Firstly, we randomly divide all the vertices of the graph into two subsets Y and Z, and all the edges into *Red* and *Blue* edges, (as in Figure 4.1). The final algorithm, k-Partition (Algorithm 8), is made of three steps: the first sub-routine outputs a partition of Z as vertices of *Red* edges; then there is a correction process to better the partition on Z; finally we label the vertices in Y according to the number of *Blue* connections with the communities in Z. Each step works on different sets of edges and vertices, and the final candidate vertex partition is built from the two partitions of the random sets Y and Z. Thus, given a SBM with k communities with connection probabilities $\frac{a}{n}$ and $\frac{b}{n}$, we prove that the algorithm k-Partition outputs with high probability a vertex partition with a small error rate $\gamma > 0$ when $a > b > C_1$ and

$$(a-b)^2 \geq C_2\, k^2\, a \log \frac{1}{\gamma}$$

for $C_1, C_2 > 0$ suitable constants. In particular, the requirement

$$\frac{(a-b)^2}{a} = \Omega(k^2)$$

is optimal. We prove that when the first sub-routine outputs a partition of Z with error rate 0.1, the final error rate satisfy the following optimal relation:

$$\frac{(a-b)^2}{k(a+b)} = \frac{1}{0.0324} \log \frac{2k}{\gamma} \leq 31 \log \frac{2k}{\gamma}.$$

Chapter 1

Community detection in networks

In the 18th century, a Swiss mathematician, Leonhard Euler, introduced for the first time the notion of graph. Euler wanted to answer a popular question of his time and that's how the following anecdote led to the creation of a new branch of mathematics: *"if we are in the centre of the city of Königsberg* (at the time a Prussian city, now Kaliningrad, Russia) *is it possible to cross each of the seven bridges of the city only once?"*
The crucial step made by the mathematician was to encapsulate all the relevant information in a simplified map of the city in which real distances did not matter any more: different parts of the city (large or small) were described by points called *vertices* and the links between them (through bridges) were lines called *edges*. The map of the city became a *graph*.
Starting from this kind of problems, graph theory became more and more elaborated and nowadays it is applied to many different fields.

In this thesis, we focus on the class of *random graphs*, namely graphs in which the presence of an edge between two vertices depends on a probability distribution. Many real-world graphs (or *networks*) tend to have a community structure: it means that there are clusters of tightly connected vertices. For this reason, one of the topics of modern network science is the *Community Detection Problem*, which consists in finding a method to recover the communities of vertices that naturally take shape within a random graph.
This is an important issue, indeed there are several applications in really different contexts including biology, the Internet, food webs, transport networks or the brain structure. Besides, the community structure is an important determinant of the behaviour of some processes on networks, such as information diffusion or virus spreading: the community structure can both enforce as well as inhibit diffusion processes. Another particular case is that of a structure of overlapping communities: in some circumstances, a vertex can belong to different clusters at the same time. For instance, in social networks the vertices represent people, and each person can be assigned to different communities representing its family, friends, work colleagues and any other kind of relation.
Due to the many applications, the mathematicians have been looking for different methods to solve the problem. Among them, it is possible to encounter

methods for which it is necessary to know the model of graph, and others that only require a particular property (like, for example, the connectedness). Nowadays, there are techniques based on different mathematical tools like the random walk, Hamiltonian functions or minmax theorems. In this thesis, we are going to study a spectral method that relies on the eigenstructure of the adjacency matrix of the input random graph. Among all the models of networks with a community structure, the most used and simple is the *Stochastic Block Model*. It can be considered as a "zero model" since it is built over the simplest assumptions.

In this first chapter, we are going to define and describe thoroughly the Community Detection problem, some applications and some resolution methods. Moreover, we introduce the Stochastic Block Model, that will be the starting point for the following chapters.

1.1 Graphs and random graphs

First, let us recall some notions about graphs:

Definition 1. *A **graph** is an ordered pair $G = (V, E)$, where V is a set of **vertices** (or **nodes**) and $E \subseteq \{\{i, j\} \,|\, i, j \in V\}$ is a set of **edges**.*

The graph **order** is the number of its vertices and the graph **size** is the number of its edges.
Inside the class of graphs, we can distinguish:

- **directed** graphs, in which the edges have an orientation ($\{i, j\} \neq \{j, i\}$ for every pair of vertices i and j);
- **undirected** graphs, in which all edges are bidirectional ($\{i, j\} = \{j, i\}$ for every pair of vertices i and j).

For the purposes of this thesis, we will only consider undirected graphs.

For any vertex $i \in V$ of a graph G, the **degree** of i, $\delta(i)$, is the number of edges that are incident to it. Moreover, given $G = (V, E)$ with $|V| = n$,

- the number of edges of G is
$$|E| = \frac{1}{2} \sum_{i \in V} \delta(i); \tag{1.1}$$
- the maximal possible number of edges in a graph with no loops (i.e. self edges) is
$$\nu_{max} := \frac{n(n-1)}{2}; \tag{1.2}$$
- the density of G is
$$\Delta_G := \frac{|E|}{\nu_{max}}. \tag{1.3}$$

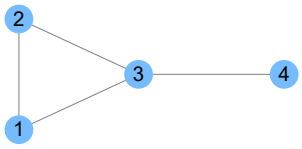
The structure of a graph can be represented by means of a matrix:

Definition 2. *The **adjacency matrix** A of an undirected graph $G = (V, E)$, with $|V| = n$, is a symmetric $n \times n$ matrix whose entries are defined as*

$$A_{i,j} = \begin{cases} 1 & \text{if vertices } i \text{ and } j \text{ are connected by an edge} \\ 0 & \text{otherwise} \end{cases} \tag{1.4}$$

for any $i, j \in V$.

In Figure 1.1 we see the translation from a graph to its adjacency matrix or viceversa.



$$\begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Figure 1.1: Given a simple graph G on 4 vertices, we see how it can be represented by a 4×4 matrix.

Let us now define

Definition 3. *A **random graph** is a graph G in which, given a set of vertices, the edges between vertices are added according to a probability distribution.*

Since the use of graphs is related to the representation of particular problems, during the decades many graph models have been introduced to describe different phenomena. The most thoroughly studied and simple model of random graph is due to the two mathematicians Paul Erdös and Alfred Rényi:

Definition 4. *The **Erdös-Rényi** model is a random graph $G := G(n, p)$ constructed on a set of n vertices by connecting every pair of distinct vertices independently with probability p.*

The degree of a vertex in a random graph is the number of edges incident to that vertex, as for a graph. The **expected degree** of every vertex in $G(n, p)$ is

$$d := (n - 1)p. \tag{1.5}$$

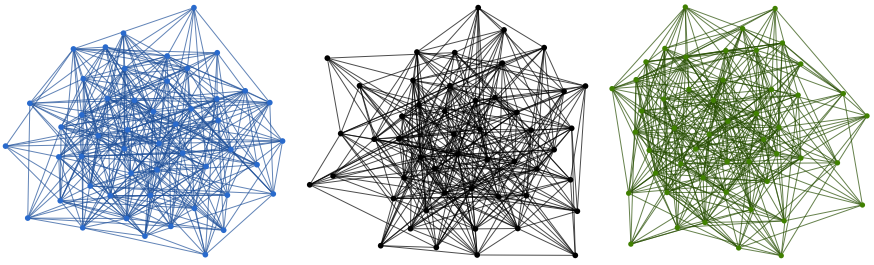


Figure 1.2: Three random graphs generated by the Erdös-Rényi model $G(50, 0.3)$. According to the probability $p = 0.3$, these three graphs have, starting from the left, 364, 371 and 374 edges.

Another different and useful model in network science is the **Barabási-Albert model**. Starting from a set of m_0 vertices, the graph is built through successive time-steps. At each step, a new vertex is added to the graph and is linked to the old ones through $m \leq m_0$ edges (see Figure 1.3). This kind of random graph is an important tool to represent the evolution of the Internet, the World Wide Web or the social networks.

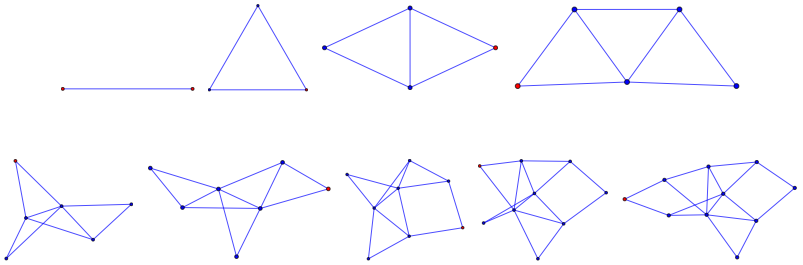


Figure 1.3: This sequence shows nine subsequent steps of the Barabási-Albert model. For every new vertex (the red ones) we add $m = 2$ new edged to the graph.

There exist many other random graph models, that we do not report in this work. For more details, see [8] or [4].

1.2 The Community Detection Problem

Most networks of interest display community structure, i.e. their vertices are organized into groups called communities, clusters or modules. Likewise, communities could represent proteins with similar function in protein-protein interaction networks, groups of friends in social networks, websites on similar topics on the Web graph and so on. Identifying communities may offer insight on how the network is organized.

The **Community Detection Problem** on a network consists in identifying the community structure of a graph accurately and efficiently.
This is an ill-defined problem. There are no universal protocols on the fundamental ingredients, like the definition of community itself, nor on other crucial issues like the validation of algorithms and the comparison of their performances. Thus, the field of community detection has been expanding greatly since the 80's with a remarkable diversity of models and algorithms developed in different fields. Nowadays, it is one of the most popular topics of modern network science.

The classical view of a community structure is given by a partition of vertices in which the different communities are well separated from each other (as in Figure 1.4). In this case, we should say that the density of edges inside each cluster is much higher than the density of edges between the clusters.

However, there are some situations in which communities may overlap, sharing some of the vertices. For instance, in social networks individuals can belong

to different circles at the same time, like family, friends and work colleagues. A subdivision of a network into overlapping communities is called **cover** and one speaks of soft clustering, as opposed to hard clustering, which deals with division into non-overlapping groups called **partitions**.
In this thesis, we are going to examine graphs with classical community structures.

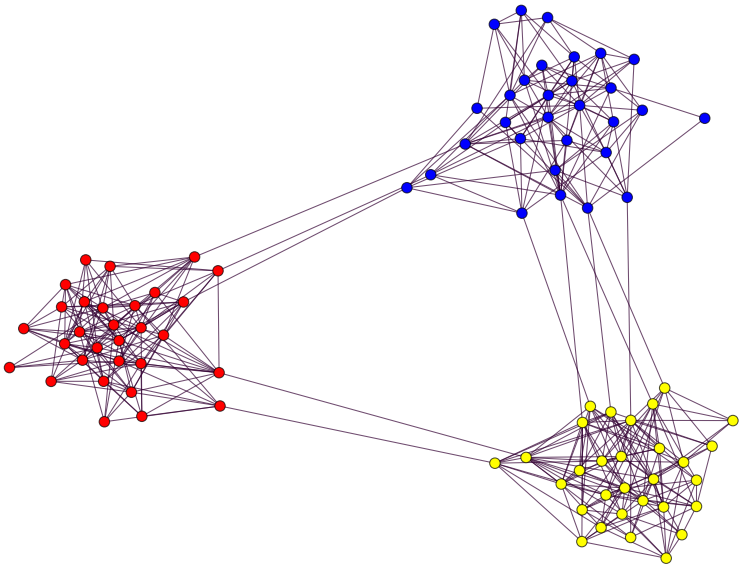


Figure 1.4: A classical and explicative representation of a community structure given by three clusters of vertices.

What is a community?

A basic issue in community detection is the lack of a universal definition of community.

One of the most used concept is that of **clique**. A clique is a complete graph, that is, a subgraph such that each of its vertices is connected to all the others. It is also a maximal subgraph, meaning that it is not included in a larger complete subgraph. The notion of cliques, even if useful, cannot be considered a good candidate for a community definition: while a clique has the largest possible internal edge density, communities are not complete graphs, in general.

Other definitions are based on the idea that a vertex must be adjacent to some minimum number of other vertices in the same subgraph. A **k-plex** is a maximal subgraph in which each vertex is adjacent to all other vertices of the subgraph except at most k of them.

For a proper community definition, one should take into account both the internal cohesion of the candidate subgraph and its separation from the rest of the network. A simple idea that has received a great popularity is that a community is a subgraph such that the number of internal edges is larger than the number of external edges.

Furthermore, what one should be really focusing on is the probability that vertices share edges with a subgraph. The existence of communities implies that vertices interact more strongly with the other members of their community than they do with vertices of the other communities. This is the reason why edge densities end up being higher within communities than between them. We can formulate that by saying that vertices of the same community have a higher probability to form edges with their partners than with the other vertices.

But is a definition of community really necessary? Actually not, most techniques to detect communities in networks do not require a precise definition of community. But defining clusters beforehand is a useful starting point, that allows one to check the reliability of the final results.

Accuracy of the detection

A further important question is related to the accuracy of a certain method: once a clustering technique has produced as output a candidate community structure, the efficiency of the method is determined by the error rate between the candidate and the real partition. As for the definition of community, there is not a universal notation that classifies an *optimal*, *exact* or *weak* recovery. For instance, in [1], they define some recovery requirements that differ from those given in [8].

In this thesis, at the right time we are going to define two notions of accuracy: the γ-*correctness* (Definition 6) and *optimality* (that we will introduce through an important result, namely Lemma 10). These definitions refer directly to [5], which is the fundamental base of this work.

Applications of community detection in real models

The community detection is a useful topic that applies in many different fields and situations. We mention here some of the most common applications:

- **Biological networks**: in recent times, the amount of information available on interactions involving proteins, genes, metabolic processes, etc., has allowed a great developement in genomics. In order to study cellular systems, the graph representation is regularly used: protein-protein interaction networks (PIN), gene regulatory networks (GRN) and metabolic networks (MN) are now standard objects of investigation in biology.

  Biological networks are characterized by a remarkable modular organization, reflecting functional associations between their components. For instance, proteins tend to be associated in two types of cellular modules, and identifying them is fundamental to uncover the organization and dynamics of cell functions.

  A further studied example is that of a network of gene co-occurrence to find groups of related genes. Communities of genes related to colon cancer can be helpful to identify the function of the genes.

- **Social networks**: social networks are networks in which the vertices are people, or sometimes groups of people, and the edges represent some form of social interaction between them.

Networks describing social interactions between people have been studied for decades. At present, we cannot avoid to think about the applications in epidemiology. Some studies dealing with this topic focus on strongly connected communities, in which a virus diffuses rapidly; others are related to predictions and factors into the social structure determining the spread. The recent work [13], for instance, is about containment, social factors, social structure and the underlying social contact networks. The aim is to guide communities on striking trade-offs between the spread of epidemic and societal inconvenience/economic distress.

Another class is that of online social networks. Recently new interaction modes between individuals have been born, like mobile phone communications and online interactions enabled by the Internet. In particular, the social networks as Facebook, Twitter etc. Such new social exchanges can be accurately monitored for very large systems, including millions of individuals. The same holds for the spread of news on the net. The applications to this field are similar to those for epidemics: an echo chamber is a sort of strongly connected community in which fake news or personal beliefs are amplified, and different ideas are not allowed. Detecting echo chambers implies finding communities in which all the members agree to the same idea and it helps to stop the spread of fake news.

- **The World Wide Web**: the first information network, the World Wide Web, is probably the best known example: the vertices are web pages consisting of text, pictures or other information, and the edges are the hyperlinks that allow us to navigate from page to page. Since hyperlinks run in one direction only, the Web is a directed graph. Within the Web graph, we can highlight clusters of websites concerning similar topics and thus recover a community structure. Another famous example is **the Internet**, the physical network of computers, routers and modems which are linked via cables or wireless signals. An ever more modern field is that of IoT (Internet of Things), whose networks are witnessing a drastic increase over the years. The need for identifying communities within such networks can serve as a strong complexity reduction mean for many discovery and identification services. The idea of communities in IoT networks is also motivated by the emerging concept of socializing IoT devices.

- Most papers refer to one or more other previous papers, and one can construct a network in which the vertices are papers and the edges are the citations among them. Since a citation implies a correlation with the indicated previuos work, **citation networks** are networks of relatedness of subject matter. Similarly, we can build a **collaboration network** in which the vertices represent a set of scientists and the edges indicate the collaborations among them.

1.3 Different methods to solve the community detection problem

Within the field of community detection, plenty of different algorithms have been introduced as possible resolution methods. They can be grouped in categories,

based on different criteria.

In this thesis we are going to develop a **spectral method**, which approaches the problem studying the eigenvectors of the adjacency matrix related to the input graph. Here we briefly present other popular methods:

- **Methods based on optimisation**: optimisation techniques have received the greatest attention in the literature. The goal is finding an extremum, usually the maximum, of a function indicating the quality of a clustering, over the space of all possible clusterings. Quality functions can express the goodness of a partition or of single clusters. The most popular quality function is the *modularity*. It estimates the quality of a partition of the network in communities. Taken a graph G with $n = |V|$ vertices, the general expression of modularity is

$$\mathcal{Q} = \frac{1}{2m} \sum_{i,j \in V} (A_{i,j} - P_{i,j})\, \delta(C_i, C_j), \tag{1.6}$$

  where m is the number of edges of the network, $A_{i,j}$ is an element of the $n \times n$ adjacency matrix, $P_{i,j}$ is the null model term and in the Kronecker delta C_i and C_j indicate the communities of vertices i and j. The term $P_{i,j}$ indicates the average adjacency matrix of an ensemble of networks, derived by randomising the original graph, such to preserve some of its features. Therefore, modularity measures how different the original graph is from such randomisations. The concept was inspired by the idea that, randomising the network structure, communities are destroyed, so the comparison between the actual structure and its randomisation reveals how non-random the group structure is. A standard choice is $P_{i,j} = k_i k_j / 2m$, k_i and k_j being the degrees of i and j, and corresponds to the expected number of edges joining vertices i and j if the edges of the network were rewired such to preserve the degree of all vertices, on average. This yields the classic form of modularity

$$\mathcal{Q} = \frac{1}{2m} \sum_{i,j \in V} \left(A_{i,j} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j). \tag{1.7}$$

  Other choices of the null model term allow us to incorporate specic features of network structure, like bipartiteness, correlations, signed edges, space embeddedness, etc..

  Optimisation techniques work looking for the maxima of such function and gathers the vertices according to the configurations that satisfy the extremum condition.

- **Methods using random walks**: communities can be identified by running dynamical processes on the network. Random walk dynamics is by far the most exploited: if communities have high internal edge density and are well-separated from each other, random walkers would be trapped in each cluster for a long time before finding a way out and migrating to another cluster.
  A first class of techniques consists in using random walk dynamics to estimate the similarity between pairs of vertices. For instance, in the popular

method *Walktrap* the similarity between vertices i and j is given by the probability that a random walker moves from i to j in a fixed number of steps t. The parameter t has to be large enough, to allow the exploration of a significant portion of the graph, but not too big since one would approach the stationary limit. If there is a pronounced community structure, pairs of vertices in the same cluster are much more easily reachable by a random walk than pairs of vertices in different clusters, so the vertex similarity is expected to be considerably higher within groups than between them. This class of methods have a computational complexity wich is higher than quadratic in the number n of vertices (on sparse graphs), so they cannot be used on large networks.
A different approach is the so called *Infomap*. It is born as answer to the problem of finding a parsimonious way to describe an infinitely long random walk taking place on the graph. The simplest description is obtained by listing sequentially all vertices reached by the random walker, each vertex being described by a unique codeword. However, if the network has a community structure, there may be a more compact description, which follows the principle of geographic maps, where there are multiple cities and streets with the same name across regions. Vertex codewords could be recycled among different communities, which play the role of regions/states, and vertices with identical name are distinguished by specifying the community they belong to. If clusters are well separated from each other, transitions between clusters are infrequent, so it is advantageous to use the map, with the communities as regions, because in the description of the random walk the codewords of the clusters will not be repeated many times, while there is a considerable saving in the description due to the limited length of the codewords used to denote the vertices (see Figure 1.5).

In both methods, running a random walker does not require many informations on the graph. Thus, these technique are really useful and can be implemented in different contexts.

- **Spin dynamics**: this is another effective technique in network clustering. It consists in defining a spin model on the network, namely we assign a set of spin variable s_i to the vertices and define an Hamiltonian function $\mathcal{H}(s)$ expressing the energy of the spin configuration s. For community detection, vertex spins s_i are given by ± 1 or 0 and 1, and the global configuration s is an integer value. Contributions to the energy are usually given by spin-spin interaction. The coupling of a spin-spin interaction can be *ferromagnetic* (negative) or *antiferromagnetic* (positive), if the energy is lower when the spins are equal or not, respectively. The goal is to find the spin configurations that minimise the Hamiltonian $\mathcal{H}(s)$. A popular model consists in rewarding edges between vertices in the same cluster and non-edges between vertices in different clusters, and at the same time penalising edges between vertices of different clusters and non-edges between vertices in the same cluster. This way, if the edge density within communities is larger than the edge density between communities, as it often happens, having equal spin values for vertices in the same cluster would considerably lower the energy of the configuration.

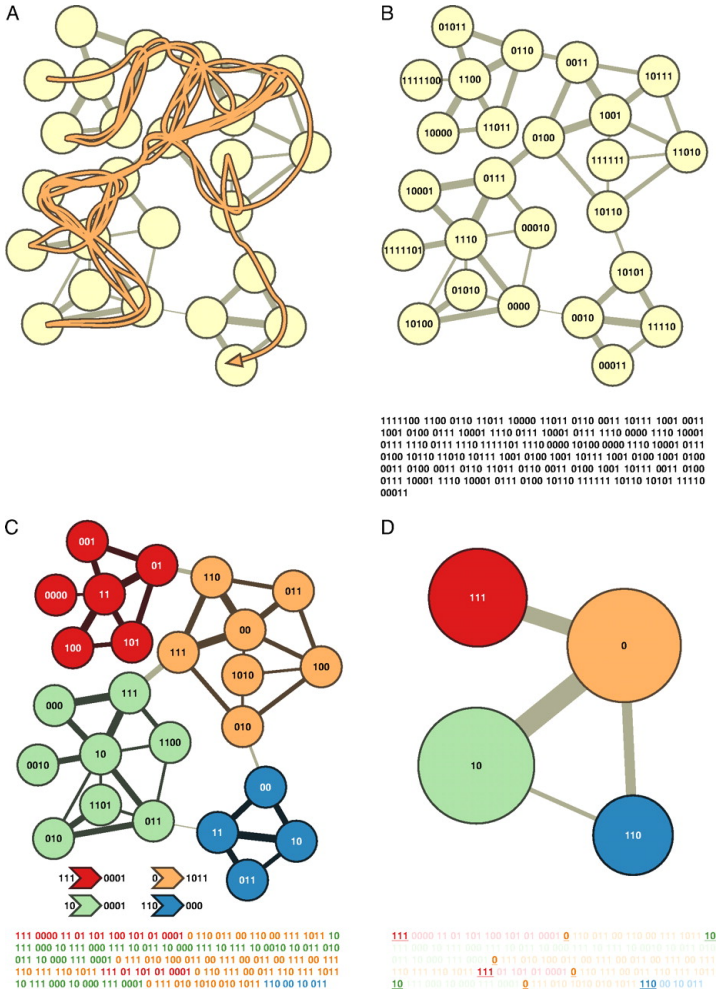


Figure 1.5: This figure is taken from [15]. It represents the Infomap. The random walk in (A) can be described as a sequence of the vertices, each labeled with unique codewords (B), or by dividing the graph in regions and using unique codewords only for the vertices of the same region (C). In this way, the same codeword can be used for multiple vertices, at the cost of indicating when the random walker leaves a region to enter a new one, as in that case one has to specify the codeword of the new region, to unambiguously locate the walker. This network has four communities (indicated by the colours in (C)), and in this case the map-like description of (C) is more parsimonious than the one in (B). This is shown by looking at the actual code needed in either case (bottom of the figures), which is clearly shorter for (C). In (D) the transitions between the clusters are highlighted.

A general expression for the Hamiltonian in the case of a graph with $n = |V|$ vertices is

$$\mathcal{H}(s) := -\sum_{i,j \in V} \left[a_{i,j} A_{i,j} - b_{i,j}(1 - A_{i,j}) \right] \delta(s_i, s_j), \tag{1.8}$$

where $A_{i,j}$ is an element of the adjacency matrix related to the graph, $a_{i,j}, b_{i,j} \geq 0$ are arbitrary coefficients and the Kronecker delta selects only

the pairs of vertices with the same spin value.
A popular model is obtained setting $a_{i,j} = 1 - b_{i,j}$ and $b_{i,j} = \gamma P_{i,j}$ where γ is a parameter and $P_{i,j}$ a null model term, expressing the expected number of edges running between vertices i and j under a suitable randomisation of the graph structure. The resulting Hamiltonian, as defined in [14], is

$$\mathcal{H}_{RB}(s) = -\sum_{i,j \in V} (A_{i,j} - \gamma P_{i,j})\, \delta(s_i, s_j). \tag{1.9}$$

If $\gamma = 1$ and $P_{i,j} = k_i k_j / 2m$, k_l being the degree of vertex l and m the total number of graph edges, the Hamiltonian (1.9) coincides with the modularity in (1.7), up to the sign. Consequently, modularity can be interpreted as the Hamiltonian of a spin dynamics.

This class of methods works looking for the minima of the Hamiltonian function $\mathcal{H}(s)$ by varying the possible configurations of s. The found configurations permit to separate the vertices of the graph according to the same spin, and thus obtain a partition.

1.4 The Stochastic Block Model

In this thesis, we focus on a particular model of networks with community structure: the Stochastic Block Model (SBM). It is the most simple and studied model, since it just takes the basic assumptions that permit the presence of a vertex partition within a graph. In the rest of the thesis, we will study a resolution method to solve the community detection in random graphs generated by this model.

Definition 5. *Let n and $k < n$ be positive integers. The (general) **Stochastic Block Model** is a random graph with n vertices divided in k communities $(V_1, \dots, V_k)$. The probability that two vertices of the graph are connected depends exclusively on their group membership: for $i, j = 1, \dots, k$, let*

$$p_{i,j} := \mathbb{P}(\textit{a vertex in } V_i \textit{ is connected to a vertex in } V_j)$$
$$p_{i,i} := \mathbb{P}(\textit{two vertices in } V_i \textit{ are connected}).$$

Then, all the $p_{i,j}$ form a $k \times k$ symmetric matrix W: the diagonal elements of W are the probabilities that vertices of the same community are neighbours, whereas the off-diagonal elements give the edge probabilities between different communities. We denote this model by $G(n, W)$.

Note that it is a natural extension of the Erdös-Rényi random graph (as defined in Def. 4) to the case in which we differentiate the probabilities inside and outside clusters of vertices. Indeed, if all the entries of W are the same, let us say $p_{i,j} = p$ for any $i, j = 1, \dots, k$, then the SBM collapses to the Erdös-Rényi random graph and no meaningful reconstruction of communities is possible.

In Definition 5 we do not refer to the dimensions of the communities V_i: indeed, in general the communities may have different sizes. However, in this thesis we will only consider the case

$$|V_i| = \frac{|V|}{k} \quad \text{for any } i = 1, \dots, k. \tag{1.10}$$

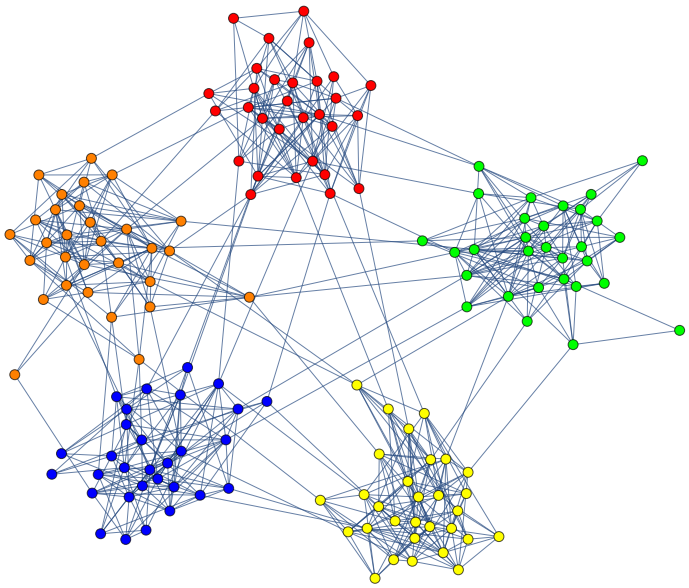


Figure 1.6: A SBM with five communities generated by $G(150, 0.3, 0.005)$, namely with $p_{i,i} = 0.3$ for every $i = 1, \dots, 5$ and $p_{i,j} = 0.005$ for every $i \neq j$.

The main result of this thesis consists in solving the community detection problem for a particular case of SBM using a **spectral algorithm**. This method produces a vertex partition exploiting the properties of the eigenvectors of the adjacency matrix related to the input graph. In order to exhibit the method, we will proceed this way:

- in Chapter 2 we consider the SBM $G(2n, p, q)$ with two communities (V_1, V_2) such that $|V_i| = n$. It is a random graph over $2n$ vertices where p is the probability inside a community and q is the edge probability between the two communities. In particular, to get a higher density inside the clusters, we will take $p > q$;
- in Chapter 3 we consider another SBM with two communities in which the probabilities inside/outside the communities are of the order $\sim \frac{1}{n}$. Namely, we will study $G\left(2n, \frac{a}{n}, \frac{b}{n}\right)$ with $n > a > b > 0$;
- in Chapter 4 we finally analyze the general model with $k > 2$ communities $G\left(n, \frac{a}{n}, \frac{b}{n}\right)$, where the inside and outside edge probabilities are the same for all the k blocks.

Chapter 2

Spectral method for a dense Stochastic Block Model with two communities

In this chapter we want to illustrate a spectral method to solve the community detection problem in a *dense* random graph generated by the Stochastic Block Model. The main result in this work deals with *sparse* random graphs, but it is useful and on completion to first understand how to manage a simpler case.

First, we want to focus on the definitions of dense and sparse graphs, highlighting some properties and differences. Afterwards, we introduce a methodology that will be the starting point for the central part of this thesis. Namely, we will analyze step by step a spectral method which will allow us to recover the community structure of a random graph.

For theoretical and technical results, we refer to the appendix.

2.1 Dense and sparse graphs

Let $G(n,p)$ be the Erdös-Rényi random graph: as introduced in Chapter 1 (see Definition 4), it is a graph with n vertices where every pair of distinct vertices is connected with probability p. It is the simplest and most studied model of random graphs: in the following exposition, we will define density/sparsity properties only for graphs $G \sim G(n,p)$.

The notion of density/sparsity is related to the amount of edges that are present in a graph. The expected degree for each vertex of $G(n,p)$ is

$$d := (n-1)p.$$

In this thesis, we consider the two following cases:

- we take a random graph in which the probability $p < 1$ is fixed and consequently the expected degree of each vertex is $d = (n-1)p \sim n$ (**dense**);
- we take a random graph in which the probability p goes to zero as $\frac{1}{n}$. The expected degree is $d \sim k$ for a constant k, thus it is an infinitesimal fraction of all the possible edges (**sparse**).

The main reason that leads to first consider dense graphs is their regularity: for n very big, the degrees of all vertices approximately equal the expected degree d. This property is no more true for sparse graphs. More precisely, we prove the following statement, in which we suppose $d \geq C \log n$ for a certain absolute constant $C > 0$, and thus we can extend to our dense graph definition:

Proposition 1. *Let $G \sim G(n,p)$ be a random graph with expected degree satisfying $d \geq C \log n$, for an absolute constant C. Then, with high probability, all vertices of G have degree between $(1-\delta)d$ and $(1+\delta)d$, for every small $\delta > 0$.*

Proof. Let i be a fixed vertex of G and let d_i be its degree. In particular, $d_i = \sum_{j=1}^{n-1} X_j$, where $X_j \sim Ber(p)$ are independent indicators of an edge between vertices i and j. So, observing that $\mathbb{E}[d_i] = d$, we apply Chernoff inequality 2 and get, for $\delta \in (0,1]$,

$$\mathbb{P}\left(|d_i - d| \geq \delta d\right) \;\leq\; 2\exp\left(-cd\delta^2\right). \tag{2.1}$$

Taking the union bound over all the possible choices for the vertex $i \in \{1,\dots,n\}$, we can use (2.1) so that

$$\mathbb{P}\left(\exists i \leq n : |d_i - d| \geq \delta d\right) \;\leq\; \sum_{i=1}^{n} \mathbb{P}\left(|d_i - d| \geq \delta d\right) \;\leq\; 2n\exp(-cd\delta^2). \tag{2.2}$$

The hypothesis stated that $d \geq C \log n$ for an absolute constant C.
Thus, for C sufficiently large, the complementary of (2.2) becomes

$$\mathbb{P}\left(\forall i \leq n : |d_i - d| < \delta d\right) \;\geq\; 1 - 2n\exp(-c\delta^2 C \log n) \;\geq\; 1 - \epsilon$$

for every $\delta > 0$ and $\epsilon > 0$.

□

From now on, we represent a dense and sparse random graph with the notations $G(n,p)$ and $G(n,\frac{k}{n})$ (with k a constant), where n denotes the number of vertices and the probabilities reflect the same behaviour as in our definition of dense/sparse. In Figure 2.1 we see examples of graphs built this way. As the number of vertices increases, we recognize the regularity in the dense graph and not in the sparse one, as it was expected to be.

The Stochastic Block Model is a random graph model with community structure, in which the edge probability between two vertices depends exclusively on their community membership (see Definition 5).
Later on, we are going to exhibit a method to solve the community detection problem for the SBM with two communities $G(2n,p,q)$ with $1 > p > q > 0$: this will be our dense example, since we can recognize the trend $d \sim n$.
Whereas, in the next chapter we will focus on a sparse version of the SBM with probabilities going to zero as $\frac{1}{n}$. Namely, we will study $G(2n,\frac{a}{n},\frac{b}{n})$ with $n > a > b > 0$. The same construction holds for the problem with k communities, that we are going to study in Chapter 4.

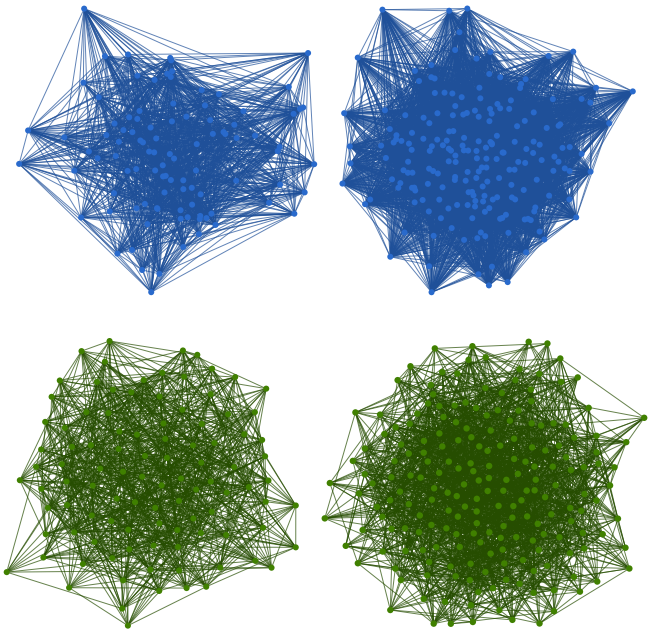


Figure 2.1: The blue graphs are dense graphs generated by $G(100, 0.3)$ and $G(200, 0.3)$ respectively. For the first one, the expected degree is $d_1 = 29.7$. We computed that the vertex degrees belong to the interval $[21, 40]$: it means that we get a fraction $\delta_1 = 0.347$ such that every vertex has degree in $[(1 - \delta_1)d_1, (1 + \delta_1)d_1]$. Similarly, for the left graph we expect $d_2 = 59.7$ and we found a degree interval $[45, 75]$: in this case, $\delta_2 = 0.256 < \delta_1$, so we can see that, as the number of vertices n grows, the vertex degrees approach the expected degree.
The green graphs are sparse graphs generated by $G(100, 20/100)$ and $G(200, 20/200)$. The expected degrees are $d_3 = 19.8$ and $d_4 = 19.9$ respectively, so, as n grows, we approach to $k = 20$. Moreover, looking at the values of the vertex degrees, we can compute a distance from d_i such that $\delta_3 = 0.465$ and $\delta_4 = 0.507$, thus we do not see an approach to the expected degree as the the number of vertices increases.

2.2 The spectral method

We now try to solve the community detection problem for the Stochastic Block Model $G(2n, p, q)$. It is a random graph of $2n = |V|$ vertices with a community structure given by two subsets of vertices $V_1, V_2 \subset V$, each of size n. We generate such a graph with the following distribution:

- an edge between vertices belonging to the same community appears with probability p;
- an edge between vertices belonging to different communities appears with probability q,

where $1 > p > q > 0$ are fixed.
Given the random graph G, our aim is to recover the partition (V_1, V_2).

A fundamental tool to represent a graph is the adjacency matrix. Let A be the adjacency matrix related to G: recall that an entry (i, j) is either 1 or

0 depending on the presence or absence of an edge between vertices i and j. Namely, in our particular model, each $A_{i,j}$ can be seen as a Bernoulli variable with parameter p if i and j belong to the same community or with parameter q otherwise.

Consider $\bar{A} = \mathbb{E}[A]$: due to what just observed, the entries of this matrix are given by the expected values of $Ber(p)$ and $Ber(q)$. So, under a permutation of rows and columns, we can collect the vertices in communities in such a way that the $2n \times 2n$ matrix $\bar{A}$ looks like

$$\bar{A} = \mathbb{E}[A] = \left(\begin{array}{ccc|ccc} p & \dots & p & q & \dots & q \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ p & \dots & p & q & \dots & q \\ \hline q & \dots & q & p & \dots & p \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ q & \dots & q & p & \dots & p \end{array}\right).$$

Now we compute the eigenvalues and eigenvectors of $\bar{A}$ in the case $n = 2$:

$$\det(\bar{A} - x\mathbb{I}) = x^2(x^2 - 4px + 4(p^2 - q^2))$$

thus, the eigenvalues are 0 with multiplicity 2, $\frac{4p+4q}{2}$ and $\frac{4p-4q}{2}$.
For a general n, since $\operatorname{rank}(\bar{A}) = 2$, the only non zero eigenvalues are

$$\lambda_1 = \left(\frac{p+q}{2}\right) 2n, \qquad \lambda_2 = \left(\frac{p-q}{2}\right) 2n. \tag{2.3}$$

The crucial point is the structure of the corresponding eigenvectors, so let us compute them: for the eigenspace related to λ_1 in the simple case $n = 2$, we need to solve

$$\begin{cases} -(p+2q)x_1 + px_2 + qx_3 + qx_4 = 0 \\ px_1 - (p+2q)x_2 + qx_3 + qx_4 = 0 \\ qx_1 + qx_2 - (p+2q)x_3 + px_4 = 0 \\ qx_1 + qx_2 + px_3 - (p+2q)x_4 = 0 \end{cases}$$

for any vector $\bar{x} = (x_1, x_2, x_3, x_4)$.
The only solution is given for $x_1 = x_2 = x_3 = x_4$.
Similarly, the eigenspace corresponding to λ_2 is generated by a vector such that $x_1 = x_2, x_3 = x_4, x_1 = -x_3$. Hence, the eigenvectors of the 4×4 case $\bar{A}$ are

$$u_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \text{ and } u_2 = \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}. \tag{2.4}$$

Above all, the point is that the eigenvector related to the second eigenvalue shows explicitly the community structure: indeed, it is sufficient to separate the coordinates ± 1 to detect the vertex partition. The structure of the second eigenvector keeps the same also for the general dimension n, so our method will

focus on the signs of the coordinates of such a vector: according to a positive or negative value, the vertices will be assigned to two subsets, V_1 and V_2, generating a partition.

Nevertheless, from the input random graph G we can just deduce the adjacency matrix A, and not its expected value $\bar{A}$. The aim of the spectral method is to study the eigenvectors of the adjacency matrix A and to extract community information from them. Namely, we will see that we can use the second eigenvector of A to accurately estimate the eigenvector of $\bar{A}$ exhibiting the partition. Producing a partition looking at the signs of the coordinates of such a vector, the number of mislabeled vertices will be small. Thus, we will produce a vertex partition with a small error.

2.2.1 Step 1: bounding the error

Let us rewrite the adjacency matrix A as

$$A = \bar{A} + R$$

where we keep the previous definition $\bar{A} = \mathbb{E}[A]$ and we consider matrix R as a sort of "noise". First, observe that

$$\|\bar{A}\| = \lambda_1 \sim n.$$

This estimate is an operator norm property, as recalled in the appendix A.3.2. On the other hand, a bound for $\|R\|$ needs deeper results.
Recall that a random variable X is *sub-gaussian* if there exists K such that, for any $t > 0$,

$$\mathbb{P}\left(|X| \geq t\right) \leq 2\exp\left(-\frac{t^2}{K^2}\right)$$

i.e. X has a sub-gaussian tail. Within this class, we can find, for example, Bernoulli and bounded random variables (for further details, see [16]). It holds

Theorem 1. *Let A be an $m \times n$ random matrix whose entries $A_{i,j}$ are independent mean-zero sub-gaussian random variables. Then, for any $t > 0$, we have*

$$\|A\| \leq CK(\sqrt{m} + \sqrt{n} + t)$$

with probability at least $1 - 2\exp(-t^2)$ for a certain constant C.
Here $K = \max_{i,j} \|A_{i,j}\|_{\psi_2}$ and $\|A_{i,j}\|_{\psi_2} := \inf\left\{s > 0 : \mathbb{E}\left[\exp\left({A_{i,j}}^2/s^2\right)\right] \leq 2\right\}$.

Namely, for our purpose we need the following

Corollary 1. *Let A be an $n \times n$ symmetric random matrix whose entries $A_{i,j}$ on and above the diagonal are independent mean-zero sub-gaussian random variables. Then, for any $t > 0$, we have*

$$\|A\| \leq CK(\sqrt{n} + t)$$

with probability at least $1-4\exp(-t^2)$ for a constant C. Here $K = \max_{i,j} \|A_{i,j}\|_{\psi_2}$.

Proof. Let us decompose A into its upper-triangular part A^+ and its lower-triangular part A^-, so that

$$A = A^+ + A^-,$$

and without loss of generality we put the diagonal in A^+. Then we can apply the previous Theorem to both A^+ and A^- and get simultaneously

$$\|A^+\| \leq CK(\sqrt{n} + t) \qquad \|A^-\| \leq CK(\sqrt{n} + t).$$

Finally, since $\|A\| \leq \|A^+\| + \|A^-\|$, with probability at least $1 - 4\exp(-t^2)$ the statement is true.

□

Since R is a symmetric matrix whose entries are mean-zero and bounded, we can apply Corollary 1 with $t := \sqrt{2n}$ and get that, for a certain constant C,

$$\|R\| \leq C\sqrt{n} \text{ with probability at least } 1 - 4e^{-2n}. \tag{2.5}$$

2.2.2 Step 2: application of Davis - Kahan Theorem

The next step consists in showing that there is a small number of vertices of G having representatives in the second eigenvector of A and in the second eigenvector of $\bar{A}$ with opposite sign. This will imply that a great part of information that we found from $\bar{A}$, can be extracted directly from A.
Keeping in mind this idea, we need the following

Theorem 2 (Davis - Kahan)**.** *Let S and T be symmetric matrices with the same dimensions. Fix i and assume that the ith eigenvalue of S is well separated from the rest of the spectrum:*

$$\min_{j:j\neq i} |\lambda_i(S) - \lambda_j(S)| = \delta > 0.$$

Then, the angle between the eigenvectors of S and T corresponding to the ith largest eigenvalues (as a number between 0 and $\pi/2$) satisfies

$$\sin \angle(v_i(S), v_i(T)) \leq \frac{2\|S - T\|}{\delta}.$$

In particular, the unit eigenvectors $v_i(S)$ and $v_i(T)$ are close to each other up to a sign:

$$\exists \theta \in \{-1, 1\} : \ \|v_i(S) - \theta v_i(T)\| \leq \frac{2^{3/2}\|S - T\|}{\delta}. \tag{2.6}$$

Since we would like to apply Davis - Kahan theorem to our matrices, we need to check that the hypothesis are satisfied. More precisely, taking $S = \bar{A} = \mathbb{E}[A]$ and $T = A = \bar{A} + R$, we need to see if the second non zero eigenvalue of $\bar{A}$ is well separated from the rest of the spectrum of $\bar{A}$, where recall from (2.3) that

$$\mathrm{Spec}(\bar{A}) = \left\{ \lambda_0 = 0, \ \lambda_1 = \left(\frac{p+q}{2}\right) 2n, \ \lambda_2 = \left(\frac{p-q}{2}\right) 2n \right\}.$$

Then

$$\delta = \min_{j:j\neq 2} |\lambda_2(\bar{A}) - \lambda_j(\bar{A})| = \min\{\lambda_2, \lambda_1 - \lambda_2\} = \min\{(p-q)n, 2qn\} := \mu n.$$

This means that the hypothesis of Theorem 2 is satisfied. Then, recalling the estimate (2.5) for $\|R\|$, we apply (2.6) to the *unit* eigenvectors of A and $\bar{A}$ ($v_2(A)$ and $v_2(\bar{A})$) related to their second eigenvalue: with probability $1-4e^{-2n}$, there exists $\theta \in \{-1, 1\}$ such that

$$\|v_2(\bar{A}) - \theta v_2(A)\| \leq \frac{2^{3/2}\|\bar{A} - A\|}{\delta} \leq \frac{C'\sqrt{n}}{\mu n} = \frac{C'}{\mu\sqrt{n}}. \tag{2.7}$$

Multiplying both sides of (2.7) by $\sqrt{2n}$, we reconduce to $u_2(\bar{A})$, which we already computed explicitly in (2.4) and we found was a key tool to detect the partition. Therefore, we get

$$\|u_2(\bar{A}) - \theta u_2(A)\| \leq \frac{C''}{\mu} \tag{2.8}$$

which we can rewrite taking the square of the norm and expliciting the sum as

$$\sum_{j=1}^{2n} \left|u_2(\bar{A})_j - \theta u_2(A)_j\right|^2 \leq \frac{C''^2}{\mu^2} := \frac{C}{\mu^2}. \tag{2.9}$$

For every $j = 1, \ldots, 2n$, we know from (2.4) that $u_2(\bar{A})_j = \pm 1$: this means that, when the signs of $v_2(\bar{A})_j$ and $v_2(A)_j$ disagree, the contribution in the sum is at least 1.

In other words, defined $\mu = \min\{p-q, 2q\}$, we have deduced that we can bound the number of vertices with opposite signs in the second eigenvector of A and $\bar{A}$ with

$$\left|\left\{j : \ \text{sign}(v_2(\bar{A})_j) \neq \text{sign}(v_2(A)_j)\right\}\right| \leq \frac{C}{\mu^2} \tag{2.10}$$

for a constant $C > 0$.

2.2.3 Résumé of the method as Spectral Algorithm

In conclusion, given a random graph $G \sim G(2n, p, q)$ characterized by the inside presence of two clusters, we can detect the community structure studying its adjacency matrix A. Indeed, we can use the eigenvector $v_2(A)$ of A related to the second largest eigenvalue as estimate of the second eigenvector of $\mathbb{E}[A]$, whose signs identify the partition: sorting the vertices according to their values in $v_2(A)$, we recover the communities. The mislabeled vertices amount to a small quantity bounded by $\frac{C}{\mu^2}$, where $\mu = \min\{p-q, 2q\}$.

We can summarize the developed method as

Algorithm 1 Spectral Algorithm

1: Input: take the graph G.

2: Compute the adjacency matrix A of the graph.

3: Compute the eigenvector $v_2(A)$ corresponding to the second largest eigenvalue.

4: Sort the vertices according to their values in $v_2(A)$: if $v_2(A)_j > 0$ put vertex j in V_1, otherwise put vertex j in V_2.

5: Output the partition (V_1, V_2).

Following the steps collected in **Spectral Algorithm**, with the previous computations we have then proved the following

Theorem 3. *Let $G \sim G(2n, p, q)$ with $p > q$ and $\mu = \min(p-q, 2q)$. Then, with probability al least $1 - 4e^{-2n}$, the **Spectral Algorithm** identifies the two communities V_1, V_2 of G correctly up to C/μ^2 misclassified vertices, for a constant $C > 0$.*

It is important to observe that the quantity of misclassified vertices is bounded by a constant that does not depend on the total number of vertices $2n$, thus for n large the set of mislabeled vertices becomes negligible.
Moreover, to apply the spectral method, we only require that the graph is dense enough ($q \geq$ constant) and that the probability p is great enough with respect to q ($p - q \geq$ constant).

In the next chapter, we will analyze a spectral method for the community detection in a *sparse* graph (i.e. with expected degree $d \sim k$, for k constant). The starting point will be the same, and the idea will already be to exploit the informations in the expected value of the adjacency matrix. However, dealing with probabilities of the order $\sim \frac{1}{n}$, we will proceed in a different way and we will introduce new steps in the method in order to recover the communities.

Chapter 3

Spectral method for a sparse Stochastic Block Model with two communities

In this chapter we study a spectral algorithm for a *sparse* Stochastic Block Model with two communities. Recall that we defined as a *sparse* graph a random graph in which the mean number of edges per vertex approaches a constant as the number of vertices of the graph grows.

Starting from the method that we are going to develop in the case of two communities, in the next chapter we will extend the procedure to the case of k communities and prove the main result of this thesis.

Consider a set V of $2n$ vertices. Denote with V_1, V_2 the two subsets of vertices, each of size n, representing the community structure. We generate a random graph with the following distribution:

- an edge between vertices belonging to the same community appears with probability $\frac{a}{n}$;
- an edge between vertices belonging to different communities appears with probability $\frac{b}{n}$,

with $a > b > 0$, and set $d := a + b$.

We will solve the community detection problem for this random graph $G\left(2n, \frac{a}{n}, \frac{b}{n}\right)$ for a large n, recovering a great portion of the two blocks.

We define the *correctness* of the recovery in the following way:

Definition 6. *Let $G \sim G(2n, \frac{a}{n}, \frac{b}{n})$ be a random graph with a community structure given by the partition (V_1, V_2) of V. Let $(V_1^{'}, V_2^{'})$ be a candidate partition obtained as output of any method. Then, we define $(V_1^{'}, V_2^{'})$ a γ- **correct partition** of V if*

$$|V_i \cap V_i^{'}| \geq (1 - \gamma)n \tag{3.1}$$

or equivalently

$$|V_i \setminus V_i^{'}| \leq \gamma n \tag{3.2}$$

for $i = 1, 2$.

Since γ plays the role of an error rate, we would like to compute a candidate γ-correct partition with γ as small as possible.

Let us now introduce the real protagonists of our computations.
Let A_0 be the adjacency matrix of the random graph $G(2n, \frac{a}{n}, \frac{b}{n})$ generated as in our model. A_0 is a symmetric $2n \times 2n$ matrix with $(A_0)_{i,j} = 1$ if there is an edge between verteces i and j, and 0 otherwise.
Define

$$\bar{A_0} := \mathbb{E}[A_0] \quad \text{and} \quad E_0 := A_0 - \bar{A_0}.$$

As already seen in the case of dense graphs, the entries of A_0 can be viewed as Bernoulli variables with probabilities $\frac{a}{n}$ or $\frac{b}{n}$. In particular, the matrix $\bar{A_0}$ is given by the expected values of $Ber\,(a/n)$ and $Ber\,(b/n)$.
Replacing $p = \frac{a}{n}$ and $q = \frac{b}{n}$ in the computation of the eigenvalues of $\bar{A_0} = \mathbb{E}[A_0]$ as in (2.3), we get

$$\lambda_1 = a + b, \qquad \lambda_2 = a - b. \tag{3.3}$$

What is crucial, however, is the structure of the eigenvectors: the entries of the eigenvector related to λ_2 are ± 1 and they allow us to recognize the two communities (for the computations, see (2.4)).

Since we do not have $\bar{A_0}$ as a starting information, the idea of the spectral method is to use the second eigenvector of A_0 to approximately identify the partition. So, we rewrite $A_0 = \bar{A_0} + E_0$ with the hope that A results close to $\bar{A_0}$. Applying the methodology introduced in 2.2, we can bound $\bar{A_0}$ as in 2.2.1 and it is possible to bound E_0 with Bernstein inequality (A.1.3) (for further details, see [16]). Then, in order to apply Davis-Kahan (Theorem 2), we get a distance δ between eigevalues such that

$$\delta = \min(\lambda_2, |\lambda_1 - \lambda_2|) = \min(a - b, 2b) := \mu n$$

where μ becomes of the order of $\frac{1}{n}$. This means that the estimate as in (2.9) for the number of mislabeled vertices becomes of the kind of $\frac{C}{\mu^2} \sim n^2$ for a constant $C > 0$. Therefore, we would obtain a partition with a very large error.

In order to avoid this problem, we proceed in a different way: we modify E_0 and the matrices A_0, $\bar{A_0}$ deleting the rows and columns corresponding to vertices with high degree.
We fix the bound for the **deletion** at a degree $\delta = 20d = 20(a + b)$: it means that if the vertex j has degree greater than $20d$, the deletion acts zeroing out the j-th row and j-th column of the given matrix (as showed in (3.4)). Repeat this way for every vertex with such a property.

$$\begin{pmatrix} x_{1,1} & \dots & x_{1,j} & \dots & x_{1,2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{j,1} & \dots & x_{j,j} & \dots & x_{j,2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{2n,1} & \dots & x_{2n,j} & \dots & x_{2n,2n} \end{pmatrix} \Rightarrow \begin{pmatrix} x_{1,1} & \dots & 0 & \dots & x_{1,2n} \\ \vdots & \vdots & 0 & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & 0 & \vdots & \vdots \\ x_{2n,1} & \dots & 0 & \dots & x_{2n,2n} \end{pmatrix} \tag{3.4}$$

The deletion leads to the loss of part of information, but bounding the vertex degrees will be an important tool to limit the size of the involved matrices.

Then, let $A, \bar{A}, E$ be the revised matrices obtained from $A_0, \bar{A}_0, E_0$ after the deletion. Moreover, if $\Delta := \bar{A} - \bar{A}_0$, we put

$$A = \bar{A} + E = \bar{A}_0 + \Delta + E.$$

Starting from these new matrices and assuming the absence of loops, in the next sections we will work this way:

- in section 3.1 we define a spectral algorithm, **Spectral Partition**, for which we show the correctness of the output (with correctness we refer to Definition 6). This algorithm works on the eigenvectors of the deleted version of the adjacency matrix (see (3.4)) and produces a candidate vertex partition for the input random graph $G\left(2n, \frac{a}{n}, \frac{b}{n}\right)$.

  It will rely on the following

  Theorem 4. *There are constant $C_0, C_1 > 0$ such that, for any $a > b > C_0$ and $\gamma > 0$ satisfying*

  $$\frac{(a-b)^2}{a+b} \geq C_1 \tag{3.5}$$

  we can find a γ-correct partition with probability $1 - o(1)$ using Spectral Partition.

- in section 3.2 we use Spectral Partition to build a more robust algorithm **Partition**, in which we preserve the intrinsic randomness of the community detection problem and we add a sub-routine algorithm, **Correction**, that guarantees a small error rate γ. A key point will be the relation between γ and the quantities a and b describing the probabilities between vertices. So, we will use Theorem 4 replacing hypothesis (3.5) with

  $$\frac{(a-b)^2}{a+b} \geq C_1 \log \frac{1}{\gamma}. \tag{3.6}$$

  In particular, we are going to show that if Spectral Partition, run over a subset of edges chosen randomly, gives a 0.1-correct partition, then the sub-routine Correction, run over the other part of edges, outputs a γ-correct partition with $\gamma = 2\exp\left(-0.072\frac{(a-b)^2}{a+b}\right)$.

3.1 Spectral Partition: a first algorithm producing a correct vertex partition

In this section we are going to define a spectral algorithm that exploits the structure of the eigenvectors of the adjacency matrix related to the input random graph $G\left(2n, \frac{a}{n}, \frac{b}{n}\right)$. After some computations, it will be possible to separate the vertices of the graph into two subsets $V_1^{'}, V_2^{'}$ and get a candidate solution for the community detection problem. We will show that the candidate partition $(V_1^{'}, V_2^{'})$ is γ-correct with respect to the real partition (V_1, V_2), i.e.

$$|V_i \cap V_i^{'}| \geq (1-\gamma)n$$

for a small error rate $\gamma > 0$.

The main result for this section is the following

Theorem 5. *There are constant $C_0, C_1 > 0$ such that, for any $a > b > C_0$ and $\gamma > 0$ satisfying*

$$\frac{(a-b)^2}{a+b} \geq C_1 \tag{3.7}$$

we can find a γ-correct partition with probability $1 - o(1)$ using ***Spectral Partition***

where **Spectral Partition** is defined as:

Algorithm 2 Spectral Partition

1: Input: take the adjacency matrix A_0, $d := a + b$.

2: Deletion: zero out all the rows and columns of A_0 corresponding to vertices whose degree is bigger than $20d$, and obtain the matrix A.

3: Find the eigenspace W corresponding to the top two eigenvalues of A.

4: Compute w_1, the projection of all-ones vector in to W.

5: Define w_2 as the unit vector in W perpendicular to w_1.

6: Sort the vertices according to their values in w_2, and let $V_1^{'} \subset V$ be the top n vertices, and $V_2^{'} \subset V$ be the remaining n vertices.

7: Output $(V_1^{'}, V_2^{'})$.

Recall that

$$A_0 \text{ is the adjacency matrix related to } G\left(2n, \frac{a}{n}, \frac{b}{n}\right),$$
$$\bar{A_0} = \mathbb{E}[A_0],$$
$$E_0 = A_0 - \bar{A_0}.$$

After the deletion, namely the zeroing of rows and columns related to vertices with degree greater than $20d = 20(a+b)$ (see (3.4)), we defined $A, \bar{A}$ and E. Moreover, we rewrote the revised adjacency matrix after the deletion as

$$A = \bar{A_0} + \Delta + E$$

with

$$\Delta \;=\; \bar{A} - \bar{A_0} \;=\; \mathbb{E}[A] - \mathbb{E}[A_0].$$

In order to prove Theorem 5, later on we are going to see, step by step, that:

- Step 1: with high probability, $\|\Delta\| = \mathcal{O}(1)$;
- Step 2: with high probability, $\|E\| = \mathcal{O}(\sqrt{d})$;
- Step 3: using the bounds from Step 1 and Step 2, we apply Davis - Kahan theorem and obtain that the angle between the eigenspaces of $\bar{A_0}$ and $A = \bar{A_0} + \Delta + E$ is small;

- Step 4: once the angle condition is satisfied, we can find a vector (w_2 of Spectral Partition) in the eigenspace of A that is close to the second eigenvector of $\bar{A}_0$. Sorting the vertices according to their values in w_2, we get a candidate partition in which the number of mislabeled vertices is small. In this way, Spectral Partition outputs a γ-correct partition (as defined in Def. 6).

Step 1: bounding the distance between the expected value of the adjacency matrix and the expected value of a revised adjacency matrix

In order to bound the norm of the matrix Δ, we first prove a result on the maximal number of vertices with degree greater than $20d$:

Lemma 1. *There exists a constant d_0 such that if $d := a + b \geq d_0$, then with probability $1 - exp(-\Omega(a^{-2}n))$, not more than $a^{-3}n$ vertices have degree $\geq 20d$.*

Proof. Consider $X \subset V$ of size $|X| = cn$, where $c < 1$ is a constant and V is the set of $2n$ vertices of our model.
We will first bound the probability that all the vertices in this set have degree greater than $20d$; then, we will consider the case $c = a^{-3}$ and conclude thanks to the union bound on all the possible choices for X.
Let us define

$$E(X) = \{\text{edges with both end points in } X\}$$
$$E(X, X^c) = \{\text{edges with exactly one end point in } X\}.$$

We claim that if every vertex $x \in X$ has degree $\delta(x) \geq 20d$, then either $|E(X)| \geq 2cnd$ or $|E(X, X^c)| \geq 8cnd$.
Indeed, suppose that $|E(X, X^c)| < 8cnd$: the relation for the total amount of edges from vertices of X is given by

$$2|E(X)| + |E(X, X^c)| \geq 20cnd \Rightarrow |E(X)| \geq 6cnd$$

which is certainly greater than $2cnd$. Analogously, if $|E(X)| < 2cnd$, we get that $|E(X, X^c)| \geq 16cnd$ which is certainly greater than $8cnd$.

Now, in order to apply Chernoff inequality (see Chernoff 1), we want to find an upper and lower bound for the expected number of edges $\mu_{E(X)} := \mathbb{E}[E(X)]$. Recall the fact that $V = V_1 \cup V_2$ with $|V_i| = n$ for $i = 1, 2$; in particular, this means that $X \subset V_1 \cup V_2$:

- to find an upper bound, we consider the case in which we maximize the number of edges with probability $\frac{a}{n} > \frac{b}{n}$, i.e. $X \subset V_1$ and $X \cap V_2 = \emptyset$, or viceversa. We get
$$\mu_{E(X)} \leq \frac{1}{2}(cn)^2 \frac{a}{n};$$
- to find a lower bound, we consider the case in which we maximize the number of edges with probability $\frac{b}{n} < \frac{a}{n}$, it means when there are $\frac{1}{2}cn$ vertices of X in V_1 and $\frac{1}{2}cn$ vertices in V_2: then,
$$\mu_{E(X)} \geq 2\frac{1}{2}(\frac{1}{2}cn)^2 \frac{a}{n}.$$

Therefore, we obtain

$$\frac{1}{4}(cn)^2\frac{a}{n} \leq \mu_{E(X)} \leq \frac{1}{2}(cn)^2\frac{a}{n}. \tag{3.8}$$

Let

$$\delta_1 =: \frac{2}{c}.$$

We observe that

$$\delta_1 \leq \frac{2cnd}{\mu_{E(X)}} \tag{3.9}$$

indeed

$$\frac{2cnd}{\mu_{E(X)}} \geq \frac{2cnd}{\frac{1}{2}(cn)^2\frac{a}{n}} = \frac{2d}{\frac{1}{2}ca} \geq \frac{2}{\frac{1}{2}c} \geq \frac{2}{c}.$$

Then, Chernoff 1 gives

$$\begin{aligned}
\mathbb{P}(|E(X)| \geq 2cnd) &\leq e^{-\mu_{E(X)}}e^{2cnd}\Big(\frac{\mu_{E(X)}}{2cnd}\Big)^{2cnd} \\
&= \Big(e^{1-\frac{\mu_{E(X)}}{2cnd}}\frac{\mu_{E(X)}}{2cnd}\Big)^{2cnd} \\
&\underset{\text{for (3.9)}}{\leq} \Big(e^{1-\frac{c}{2}}\frac{c}{2}\Big)^{2cnd} \\
&\leq \Big(e^{(\frac{2}{c}-1)\frac{c}{2}}\frac{1}{2/c}\Big)^{\frac{2}{c}\mu_{E(X)}} \\
&\leq \Big(\frac{\exp(\delta_1-1)}{{\delta_1}^{\delta_1}}\Big)^{\mu_{E(X)}} \\
&\underset{\text{for (3.8)}}{\leq} \exp\Big(\Big(\frac{2}{c}-1-\frac{2}{c}\log\Big(\frac{2}{c}\Big)\Big)\frac{1}{4}(cn)^2\frac{a}{n}\Big) \\
&\leq \exp\Big(-\frac{1}{c}\log\Big(\frac{1}{c}\Big)\frac{1}{4}(cn)^2\frac{a}{n}\Big) \\
&= \exp\Big(-\frac{1}{4}\log\Big(\frac{1}{c}\Big)acn\Big)
\end{aligned} \tag{3.10}$$

where we used the fact that the quotient in brackets is smaller than 1 and that c is small.

Similarly for $\mu_{E(X,X^c)} := \mathbb{E}[E(X, X^c)]$:

- to find an upper bound, we consider the case in which we maximize the number of edges with probability $\frac{b}{n} < \frac{a}{n}$, it means when there are $\frac{1}{2}cn$ vertices in V_1 and $\frac{1}{2}cn$ vertices in V_2, so

$$\mu_{E(X,X^c)} \leq 2(\frac{1}{2}cn)(n-\frac{1}{2}cn)\frac{a}{n} + 2(\frac{1}{2}cn)(n-\frac{1}{2}cn)\frac{b}{n} < 2cn(n-\frac{1}{2}cn)\frac{a}{n};$$

- to find a lower bound, we maximize the number of edges with probability $\frac{a}{n}$, i.e. $X \subset V_1$ and $X \cap V_2 = \emptyset$, or viceversa:

$$\mu_{E(X,X^c)} \geq (cn)(n-cn)\frac{a}{n} = c(1-c)n^2\frac{a}{n}.$$

Therefore,

$$c(1-c)n^2\frac{a}{n} \leq \mu_{E(X,X^c)} \leq c(2-c)n^2\frac{a}{n}. \tag{3.11}$$

Let

$$\delta_2 := 4.$$

As before,

$$\delta_2 \leq \frac{8cnd}{\mu_{E(X,X^c)}} \tag{3.12}$$

because

$$\frac{8cnd}{\mu_{E(X,X^c)}} \geq \frac{8cnd}{c(2-c)n^2\frac{a}{n}} = \frac{8d}{(2-c)a} \geq \frac{8}{2-c} \geq 4.$$

Then we apply Chernoff inequality 1 and get

$$\begin{aligned}
\mathbb{P}(|E(X,X^c)| \geq 8cnd) &\leq \left(\frac{\exp(\delta_2 - 1)}{{\delta_2}^{\delta_2}}\right)^{\mu_{E(X,X^c)}} \\
&\leq \exp\left((3 - 4\log(4))\, c(1-c)n^2\frac{a}{n}\right) \\
&\leq \exp(-c(1-c)an). \qquad (3.13)
\end{aligned}$$

Substituting $c = a^{-3}$ in (3.10) and (3.13) we get the bound for the probability that a subset $X \subset V$, with $|X| = a^{-3}n$, has all the vertices with degree greater than $20d$:

$$\mathbb{P}\left(|E(X)| \geq 2cnd\right) \leq \exp\left(-\frac{3}{4}\log(a)a^{-2}n\right) \tag{3.14}$$

$$\mathbb{P}\left(|E(X,X^c)| \geq 8cnd\right) \leq \exp\left(-a^{-2}n\right). \tag{3.15}$$

Using the binomial coefficient property A.2, we see that the subsets with cardinality cn in V are at most

$$\begin{aligned}
\binom{2n}{cn} &\leq \left(\frac{2ne}{cn}\right)^{cn} \\
&= \left(\frac{2}{c}\right)^{cn} e^{cn} \\
&= \exp\left[cn + cn\log\left(\frac{2}{c}\right)\right]
\end{aligned}$$

$$= \exp\Big[-c(\log\Big(\frac{c}{2}\Big) - 1\Big)n\Big] \tag{3.16}$$

so substituting $c = a^{-3}$ in (3.16) we get

$$\binom{2n}{a^{-3}n} \leq \exp\left[-a^{-3}\left(\log\left(\frac{a^{-3}}{2}\right) - 1\right)n\right]$$

$$= \exp\left[-a^{-3}n(-3\log(a) - \log(2) - 1\right]$$

$$= \exp\left[3a^{-3}\log(a)n + a^{-3}n\left(\log(2) + 1\right)\right]$$

$$\leq \exp[4a^{-3}\log(a)n] \tag{3.17}$$

where the last inequality holds because we can take a big enough so that $\log(2) + 1 \leq \log(a)$.
Finally, we join (3.14), (3.15) and (3.17) to conclude with the union bound:

$$\mathbb{P}\{\text{every set of order } a^{-3}n \text{ has every vertex with degree} \geq 20d\}$$

$$\leq \exp(4a^{-3}\log(a)n)\ \exp(a^{-2}n)$$

$$= \exp\Big(-a^{-2}n\Big(1 - 4\frac{\log a}{a}\Big)\Big)$$

$$\leq \exp(-\Omega(a^{-2}n))$$

since we can choose a big.
Then we can state that with probability $1 - \exp(-\Omega(a^{-2}n))$, there are at most $a^{-3}n$ vertices with degree $\geq 20d$.

□

Now, let us use Lemma 1 to bound $\|\Delta\|$:
recalling the definition of $\Delta = \bar{A} - \bar{A_0}$, we observe that the zero entries of the matrix $\bar{A}$, obtained after the deletion of $\bar{A_0}$, correspond to the non-zero entries of Δ. So, since we should represent the matrix Δ, under a permutation of rows and columns, as

$$\Delta \simeq \left(\begin{array}{c|c} \mathbb{O} & \mathbb{O} \\ \hline \mathbb{O} & \bar{a_{i,j}} \end{array}\right) - \left(\begin{array}{c|c} \bar{a_{h,k}} & \bar{a_{h,j}} \\ \hline \bar{a_{i,k}} & \bar{a_{i,j}} \end{array}\right) = \left(\begin{array}{c|c} -\bar{a}_{h,k} & -\bar{a}_{h,j} \\ \hline -\bar{a}_{i,k} & \mathbb{O} \end{array}\right),$$

if there are at most $a^{-3}n$ vertices with degree greater than $20d$, then the non-zero entries of Δ are certainly at most

$$|\{\Delta_{i,j} \neq 0\}| \leq 4a^{-3}n^2.$$

Moreover, every entry $\Delta_{i,j} \leq \frac{a}{n}$, so we use the Hilbert-Schmidt norm (see A.3.2) and get

$$\|\Delta\|_{HS} = \left(\sum_{i=1}^{2n} \sum_{j=1}^{2n} (\Delta_{i,j})^2 \right)^{1/2} \leq \left(4a^{-3}n^2 \left(\frac{a}{n} \right)^2 \right)^{1/2} = \frac{2}{\sqrt{a}}$$

and, since a can be large, we can state that $\|\Delta\|_{HS} \leq 1$.
In particular, recalling form A.3.2 that $\|M\| \leq \|M\|_{HS}$ for every matrix M, we finally obtain that

Corollary 2. *Let $d = a + b \geq d_0$ and $\Delta = \bar{A} - \bar{A_0} = \mathbb{E}[A] - \mathbb{E}[A_0]$. Then, for d_0 sufficiently large, $\|\Delta\| \leq 1$ with probability $1 - \exp(-\Omega(a^{-2}n))$.*

Step 2: bounding the distance between a revised adjacency matrix and its expected value

Now we focus on the matrix $E = A - \bar{A} = A - \mathbb{E}[A]$.
We will proceed this way: at first, we introduce three statements (Lemma 2, 3 and 4) for $n \times n$ symmetric matrices with bounded entries distribution. Then we use those statements to prove a key result (Lemma 5) that we are going to apply to the $2n \times 2n$ matrix E.

Let us introduce the results we need:

Lemma 2. *Let M be a random symmetric matrix of size n with zero diagonal whose entries above the diagonal are independent with the following distribution:*

$$M_{i,j} = \begin{cases} 1 - p_{i,j} & \text{with probability } p_{i,j} \\ -p_{i,j} & \text{with probability } 1 - p_{i,j} \end{cases}.$$

Let $\sigma^2 \geq C_1 \frac{\log n}{n}$ be a quantity such that $p_{i,j} \leq \sigma^2$ for all i, j, where C_1 is a constant. Then with probability $1 - o(1)$, $\|M\| \leq C_2 \sigma \sqrt{n}$ for some constant $C_2 > 0$.

Lemma 3. *Let $\widetilde{G} = (\widetilde{V}, \widetilde{E})$ be any graph whose adjacency matrix is denoted by $\widetilde{A}$, and x, y be any two unit vectors. Let $\widetilde{d}$ be such that the maximum degree is $\leq c_1 \widetilde{d}$. Further, let $\widetilde{d}$ satisfy the property that for any two subsets of vertices $S, T \subset \widetilde{V}$ one of the following holds for some constants c_2 and c_3:*

$$\frac{e(S,T)}{|S||T|\frac{\widetilde{d}}{n}} \leq c_2 \tag{3.18}$$

$$e(S,T) \log \left(\frac{e(S,T)}{|S||T|\frac{\widetilde{d}}{n}} \right) \leq c_3 |T| \log \frac{n}{|T|} \tag{3.19}$$

where $e(S,T)$ is the number of edges between S and T.
Then $\sum_H x_i \widetilde{A}_{i,j} y_j \leq \max(16,\, 8c_1,\, 32c_2,\, 32c_3) \sqrt{\widetilde{d}}$.
Here $H := \{(i,j) \mid |x_i y_j| \geq \sqrt{\widetilde{d}}/n\}$.

Lemma 4. *Let $\widetilde{d} := \sigma^2 n$. Then with probability $1 - o(1)$, the maximum degree in the graph G_A is $\leq 20\widetilde{d}$ and for any $S, T \subset V$ one of the conditions of Lemma 3 holds.*

The proofs of Lemma 2 and Lemma 4 will be a starting point for later results. Thus, now we are going to prove them, whereas we consider Lemma 3 as a fact.

Proof (Lemma 2) . Consider a $\frac{1}{2}$-net $\mathcal{N}$ of the unit sphere $\mathcal{S}^n$.
For a property of nets, we have $|\mathcal{N}| \leq 5^n$ (see A.4 for definitions and properties). Referring to an equivalent definition for the norm of a matrix (see A.3.2), to show our statement it suffices to prove that for all $x, y \in \mathcal{N}$ there exists a constant $C_2' > 0$ such that with probability $1 - o(1)$,

$$|x^T M y| \leq C_2' \sigma \sqrt{n}.$$

Let $x, y \in \mathcal{N}$ and define a pair

$$(i,j) : \begin{cases} \text{light} & \text{if } |x_i y_j| \leq \frac{\sigma}{\sqrt{n}} \\ \text{heavy} & \text{otherwise} \end{cases} .$$

Moreover, let L and H be the classes of light and heavy pairs. Then we can rewrite

$$x^T M y = \sum_{i,j=1}^{n} x_i M_{i,j} y_j = \sum_{(i,j)\in L} x_i M_{i,j} y_j + \sum_{(i,j)\in H} x_i M_{i,j} y_j. \tag{3.20}$$

The goal becomes to bound the two summands.

First, let us focus on the light couples.
Define

$$X := \sum_{(i,j)\in L} x_i M_{i,j} y_j = \sum_{(i,j)\in L, i>j} M_{i,j} a_{i,j}$$

where

$$a_{i,j} = \begin{cases} x_i y_j + x_j y_i & \text{if } (i,j),(j,i) \in L \\ x_i y_j & \text{if } (i,j) \in L \\ x_j y_i & \text{if } (j,i) \in L \end{cases} .$$

By definition of light pairs, $|a_{i,j}| \leq 2\frac{\sigma}{\sqrt{n}}$. Moreover, since x, y are unit vectors,

$$\begin{aligned} \sum_{(i,j)\in L, i>j} {a_{i,j}}^2 &\leq \sum_{(i,j)\in L, i>j} \left(x_i y_j + x_j y_i\right)^2 \\ &\leq \sum_{(i,j)\in L, i>j} 2\left(x_i^2 y_j^2 + x_j^2 y_i^2\right) \\ &= 2\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{j=1}^{n} y_j^2\right) + 2\left(\sum_{j=1}^{n} x_j^2\right)\left(\sum_{i=1}^{n} y_i^2\right) \\ &\leq 4. \end{aligned} \tag{3.21}$$

Then, we would like to apply Bernstein inequality (see A.1.3) for X, indeed the hypothesis are satisfied:

$$X = \sum_{(i,j)\in L} x_i M_{i,j} y_j = \sum_{(i,j)\in L, i>j} M_{i,j} a_{i,j}$$

where

$$|M_{i,j} a_{i,j}| \leq 2\frac{\sigma}{\sqrt{n}}$$

for every $(i,j) \in L$ and

$$\begin{aligned}
\sum_{(i,j)\in L} \mathbb{E}[M_{i,j}^2 a_{i,j}^2] &= \sum_{(i,j)\in L} a_{i,j}^2 \left[(1-p_{i,j})^2 p_{i,j} + p_{i,j}^2 (1-p_{i,j}) \right] \\
&= \sum_{(i,j)\in L} a_{i,j}^2 p_{i,j} (1-p_{i,j}) \\
&\leq \sigma^2 \sum_{(i,j)\in L, i>j} a_{i,j}^2 \\
&\underset{\text{for (3.21)}}{\leq} 4\sigma^2.
\end{aligned}$$

So we apply Bernstein:

$$\mathbb{P}(X > t) \leq \exp\left(\frac{-\frac{1}{2}t^2}{4\sigma^2 + \frac{1}{3} 2 \frac{\sigma}{\sqrt{n}} t} \right). \tag{3.22}$$

Setting $t = 10\sigma\sqrt{n}$ in (3.22), we get

$$\mathbb{P}\left(X > 10\sigma\sqrt{n}\right) \leq \exp\left(\frac{-50\sigma^2 n}{4\sigma^2 + \frac{20}{3}\sigma^2} \right) = \exp\left(-3n \frac{50}{32} \right) \leq \exp(-3n) \tag{3.23}$$

and we can conclude that

$$\left| \sum_{(i,j)\in L} x_i M_{i,j} y_j \right| \leq 10\sigma\sqrt{n} \tag{3.24}$$

with probability at least $1 - \exp(-3n)$.

Now let us focus on the heavy pairs. Since x, y are unit vectors, and thanks to the definition of heavy, we observe that

$$1 \geq \sum_{(i,j)\in H} x_i^2 y_j^2 = \sum_{(i,j)\in H} |x_i y_j| |x_i y_j| \geq \frac{\sigma}{\sqrt{n}} \sum_{(i,j)\in H} |x_i y_j|$$

which implies

$$\sum_{(i,j)\in H} |x_i y_j| \leq \frac{\sqrt{n}}{\sigma}. \tag{3.25}$$

Define

$$A_{i,j} := M_{i,j} + p_{i,j}$$

and rewrite

$$\sum_{(i,j)\in H} x_i M_{i,j} y_j = \sum_{(i,j)\in H} x_i A_{i,j} y_j - \sum_{(i,j)\in H} p_{i,j} x_i y_j. \tag{3.26}$$

Since $p_{i,j} \leq \sigma^2$ by hypothesis, we easily bound the second term of (3.26) as

$$\sum_{(i,j)\in H} p_{i,j} |x_i y_j| \leq \sigma^2 \frac{\sqrt{n}}{\sigma} = \sigma\sqrt{n}. \tag{3.27}$$

Finally, dealing with the first term of (3.26), we observe that A can be seen as the adjacency matrix of a certain graph G_A, in fact

$$A_{i,j} = M_{i,j} + p_{i,j} = \begin{cases} (1 - p_{i,j}) + p_{i,j} = 1 & \text{with probability } p_{i,j} \\ -p_{i,j} + p_{i,j} = 0 & \text{with probability } 1 - p_{i,j} \end{cases}.$$

Lemma 4 guarantees that A satisfies the hypothesis of Lemma 3 with $\widetilde{d} = \sigma^2 n$ and $H = \{\ (i,j) \mid |x_i y_j| \geq \sigma/\sqrt{n}\ \}$. Therefore with probability $1 - o(1)$ holds

$$\Big| \sum_{(i,j)\in H} x_i A_{i,j} y_j \Big| \leq C' \sigma \sqrt{n} \tag{3.28}$$

for some constant $C' > 0$.

Finally, we collect relations (3.24), (3.27) and (3.28) to conclude that with high probability

$$|x^T M y| \leq C_2' \sigma \sqrt{n}$$

and then

$$\|M\| \leq C_2' \sigma \sqrt{n}.$$

□

Proof (Lemma 4). Let k be a vertex in G_A and let

$$X := \sum_{i=1}^{n} A_{i,k}$$

be the random variable denoting the number of edges incident on it. Having the same hypothesis as in Lemma 2, we deduce that

$$\mu := \mathbb{E}[X] = \sum_{i=1}^{n} p_{i,k} \leq \sigma^2 n. \tag{3.29}$$

Thanks to Chernoff inequality 1, we find the desired bound for the maximum degree of the graph G_A : let $l \geq 4$ and recall the hypothesis $\sigma^2 \geq C_1 \frac{\log n}{n}$,

$$\mathbb{P}(X > l\sigma^2 n) \leq e^{-\mu} \Big(\frac{e\mu}{l\sigma^2 n} \Big)^{l\sigma^2 n}$$

$$= \left(e^{l - \frac{\mu}{\sigma^2 n}} \left(\frac{\mu}{l\sigma^2 n} \right)^l \right)^{\sigma^2 n}$$

$$\leq \left(e^{4 - \frac{\sigma^2 n}{\sigma^2 n}} \left(\frac{\sigma^2 n}{l\sigma^2 n} \right)^l \right)^{\sigma^2 n}$$

$$= \left(e^3 l^{-l} \right)^{\sigma^2 n}$$

$$= \exp\left(\sigma^2 n (3 - l \log l) \right)$$

$$= \exp\Big(3\sigma^2 n \Big(1 - \frac{l \log l}{3} \Big) \Big)$$

$$\leq \exp\left(-\frac{\sigma^2 n l \log l}{3} \right)$$

$$\leq \exp\left(-C_1 \frac{\frac{\log n}{n} n l \log l}{3} \right)$$

$$\leq \exp\left(-\frac{l \log n}{3} \right). \tag{3.30}$$

Applying (3.30) with $l = 20$ and taking the union bound over all the n vertices, we can state that, with probability $1 - o(1)$, the maximum degree is $\leq 20\sigma^2 n$.

Now, let $S, T \subset V$ be two subsets of vertices with $|S| \leq |T|$, and let $X := e(S,T)$ be the number of edges between S and T. If $e_{i,j}$ is the edge between $i \in S$ and $j \in T$, we have that

$$\mathbb{E}[X] = \sum_{i \in S, j \in T} p_{i,j} e_{i,j} \leq \sigma^2 |S||T|. \tag{3.31}$$

If $|T| \geq \frac{n}{e}$, since we have shown that the maximum degree in the graph G_A is $\leq 20\sigma^2 n$, we get that

$$e(S,T) \ \leq \ |S| 20\sigma^2 n \ \leq \ 20e\sigma^2 |S||T|$$

or equivalently

$$\frac{e(S,T)}{|S||T|\sigma^2} \ \leq \ 20e$$

which is condition (3.18) of Lemma 3 with $\widetilde{d} = \sigma^2 n$.

Let us now assume $|T| \leq \frac{n}{e}$ (this part of the proof follows [9]).
From Chernoff 1 it follows that, for any $k \geq 4$,

$$\mathbb{P}(X > k\sigma^2 |S||T|) \leq e^{-\mu} \left(\frac{e\mu}{k\sigma^2 |S||T|} \right)^{k\sigma^2 |S||T|}$$

$$\leq \left(e^3 k^{-k}\right)^{\sigma^2|S||T|}$$

$$\leq \exp\left(3\sigma^2|S||T|\left(1-\frac{k\log k}{3}\right)\right)$$

$$\leq \exp\left(-\frac{\sigma^2|S||T|k\log k}{3}\right). \tag{3.32}$$

We would like to find a minimal $k \geq 4$ such that the event $\left\{e(S,T) \leq k\sigma^2|S||T|\right\}$ holds with high probability for every possible S and T. Let us say we want that

$$\exp\Big(-\frac{1}{3}(k\log k)\sigma^2|S||T|\Big)\binom{n}{|S|}\binom{n}{|T|} \leq \frac{1}{n^3}. \tag{3.33}$$

Observe that (3.33) holds if and only if

$$\exp\left(-\frac{1}{3}(k\log k)\sigma^2|S||T|\right)\left(\frac{ne}{|S|}\right)^{|S|}\left(\frac{ne}{|T|}\right)^{|T|}\exp(3\log n) \;\leq\; e^0 \tag{3.34}$$

which is equivalent to

$$\frac{1}{3}(k\log k)\sigma^2|S||T| \;\geq\; |S|\left(1+\log\frac{n}{|S|}\right)+|T|\left(1+\log\frac{n}{|T|}\right)+3\log n. \tag{3.35}$$

We know that $|S| \leq |T| \leq \frac{n}{e}$, and in particular this implies that $\log\frac{n}{|T|} \geq 1$.
Let us consider $f(x) := x\log\frac{n}{x}$: this is an increasing function in $[1, \frac{n}{e}]$, in fact

$$f'(x) = \log\frac{n}{x} + x\frac{x}{n}(-\frac{n}{x^2}) = \log\frac{n}{x} - 1$$

and it is greater than zero if and only if $x \leq \frac{n}{e}$.
It follows that

$$|T|\log\frac{n}{|T|} \geq |S|\log\frac{n}{|S|}. \tag{3.36}$$

Then, instead of (3.35), it is sufficient to require that

$$\frac{1}{3}(k\log k)\sigma^2|S||T| \geq 4|T|\log\frac{n}{|T|} + 3\log n. \tag{3.37}$$

Again, for the monotonicity of $x\log\frac{n}{x}$, it holds that $|T|\log\frac{n}{|T|} \geq \log n$, so it is not restrictive to substitute (3.37) with

$$\frac{1}{3}(k\log k)\sigma^2|S||T| \geq 7|T|\log\frac{n}{|T|}. \tag{3.38}$$

Highlighting the dependence on k, we obtain that

$$k\log k \geq \frac{21|T|}{\sigma^2|S||T|}\log\frac{n}{|T|}. \tag{3.39}$$

Therefore, let k be the smallest number such that $k \log k \geq \frac{C}{\sigma^2 |S||T|} |T| \log \frac{n}{|T|}$. If it happens that $k \leq 4$, since we want the Chernoff bound (3.32) to hold, we set $k := 4$. So, let

$$k' := \max\{k, 4\}.$$

Using the union bound, we get that

$$\begin{aligned}
&\mathbb{P}\{e(S,T) \geq k'\sigma^2|S||T| \text{ for every possible choice of } S \text{ and } T\} \\
&\leq \sum_{|S|=0}^{n} \sum_{|T|=0}^{n} \binom{n}{|S|} \binom{n}{|T|} \exp\Big(-\frac{1}{3}(k' \log k')\sigma^2|S||T|\Big) \\
&\leq n^2 \frac{1}{n^3} = \frac{1}{n} \qquad (3.40)
\end{aligned}$$

thereby, for every choice of S and T,

$$e(S,T) \leq k'\sigma^2|S||T|$$

with probability at least $1 - \frac{1}{n}$.

Finally, if $k' = 4$, we easily recognize condition (3.18) of Lemma 3:

$$\frac{e(S,T)}{\sigma^2|S||T|} \leq 4.$$

At the contrary, if $k \log k = \frac{C}{\sigma^2 |S||T|} |T| \log \frac{n}{|T|}$, we have

$$e(S,T) \leq \frac{C}{\sigma^2|S||T| \log k} |T| \log\Big(\frac{n}{|T|}\Big) \sigma^2|S||T|$$

which is equivalent to

$$e(S,T) \log k \leq C|T| \log \frac{n}{|T|}$$

and since $k \geq \frac{e(S,T)}{\sigma^2|S||T|}$ we obtain (3.19)

$$e(S,T) \log\left(\frac{e(S,T)}{\sigma^2|S||T|}\right) \leq C|T| \log \frac{n}{|T|}.$$

□

With these preliminary results, we are ready to prove the key lemma for this section:

Lemma 5. *Suppose M is a random symmetric matrix with zero on the diagonal whose entries above the diagonal are independent with the following distribution*

$$M_{i,j} = \begin{cases} 1 - p_{i,j} & \text{with probability } p_{i,j} \\ -p_{i,j} & \text{with probability } 1 - p_{i,j} \end{cases}.$$

Let σ be a quantity such that $p_{i,j} \leq \sigma^2$ and M_1 be a matrix obtained from M by zeroing out all the rows and columns having more than $20\sigma^2 n$ positive entries. Then with probability $1 - o(1)$, $\|M_1\| \leq C\sigma\sqrt{n}$ for some constant $C > 0$.

Proof. We follow the same reasoning adopted in the proofs of Lemma 2 and Lemma 4 with some modifications. Consider the *light* and *heavy* couples defined as before.

Dealing with the light couples, we first bound the norm of a matrix M_S given by the deletion of a fixed set S of rows and corresponding columns from M; then, we will deduce the expected bound for M_1 as the union bound among all the possible choice for S. So, given S, we can apply the same reasoning of Lemma 2 since the only difference is that some entries $(M_S)_{i,j}$ are zero. Then, we get that with probability at least $1 - \exp(-3n)$

$$\Big| \sum_{(i,j)\in L} x_i (M_S)_{i,j} y_j \Big| \leq 10\sigma\sqrt{n} \tag{3.41}$$

for every $x, y \in \mathcal{N}_{1/2}$.
If $k := |S|$, the possible choices for S are at most

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n = \exp(n \ln 2).$$

Consequently, the bound (3.41) holds for every possible S with probability at least $1 - \exp(-(3 - \ln 2)n)$, and in particular it holds that

$$\Big| \sum_{(i,j)\in L} x_i (M_1)_{i,j} y_j \Big| \leq 10\sigma\sqrt{n}.$$

In the case of heavy couples, let us consider the same definitions as for Lemma 2 and let us show that we can bound the sum for A_1 using conditions (3.18) and (3.19), where A_1 is the matrix obtained from M_1. First, we observe that A_1, the adjacency matrix of a certain graph G_{A_1}, has bounded degree: indeed, by construction of M_1, we zeroed out the rows and columns correspondig to vertices with degree $\geq 20\sigma^2 n$. Now, as in the proof of Lemma 4, consider $S, T \subset V$. In the case $|T| \geq \frac{n}{e}$, since the maximum degree is bounded, we obtain condition (3.18). If $|T| < \frac{n}{e}$, since zeroing out rows and columns can only decrease the number of edges between S and T, we obtain the same relation for $e(S,T)$ and get that (3.18) or (3.19) holds for A_1. This implies that we can apply Lemma 3 and conclude.

□

Finally, we go back to our goal, which is to bound $\|E\|$, and prove the following:

Corollary 3. *There exist constants $C_0, C > 0$ such that if $a > b \geq C_0$, and E is the matrix produced by the deletion on $E_0 = A_0 - \bar{A}_0$, then we have*

$$\|E\| \leq C\sqrt{d} \text{ with probability } 1 - o(1).$$

Proof. This result follows directly from Lemma 5. Remember that E is obtain from the deletion of $E_0 = A_0 - \bar{A}_0$, where A_0 is the adjacency matrix of the input random graph and $\bar{A}_0 = \mathbb{E}[A_0]$. In particular,

$$(E_0)_{i,j} = \begin{cases} 1 - \frac{a}{n} & \text{with probability } \frac{a}{n} \\ -\frac{a}{n} & \text{with probability } 1 - \frac{a}{n} \end{cases}$$

if i and j belong to the same community and

$$(E_0)_{i,j} = \begin{cases} 1 - \frac{b}{n} & \text{with probability } \frac{b}{n} \\ -\frac{b}{n} & \text{with probability } 1 - \frac{b}{n} \end{cases}$$

if i and j belong to different communities.
In particular, E_0 is a symmetric matrix with a distribution on the entries as requested and, since we study a graph with no loops, E_0 has zero diagonal.
Moreover, we observe that in this case $p_{i,j} \in \{\frac{a}{n}, \frac{b}{n}\}$, so in particular $p_{i,j} \leq \frac{d}{n}$.
Therefore, we can apply Lemma 5 to this $2n \times 2n$ matrix, with $\sigma^2 = \frac{d}{n}$:

$$\|E\| \leq C\sigma\sqrt{2n} = C\sqrt{\frac{d}{n}}\sqrt{2n} = C'\sqrt{d}.$$

□

Step 3: bounding the angle between the eigenspace of the expected value of the adjacency matrix and the eigenspace of the revised adjacency matrix

In this section, we introduce some new notations.
Let $\bar{v_1}, \bar{v_2}$ and v_1, v_2 be the eigenvectors of $\bar{A_0} = \mathbb{E}[A_0]$ and $A = \bar{A_0} + \Delta + E$, respectively, corresponding to their two largest eigenvalues.
Moreover, set

$$\bar{W} := \text{Span}\{\bar{v_1}, \bar{v_2}\}, \quad W := \text{Span}\{v_1, v_2\}.$$

For any two vector subspaces W_1, W_2 of the same dimension, we use the convention

$$\sin\angle(W_1, W_2) := \|P_{W_1} - P_{W_2}\|, \tag{3.42}$$

where P_{W_i} is the orthogonal projection onto W_i.

At this step of the method, we are going to prove that if

$$\frac{(a-b)^2}{a+b} \geq C_1$$

and $a > b > C_0$ for $C_0, C_1 > 0$, then the angle between the eigenspaces W and $\bar{W}$ is small. In other words, under the conditions of Theorem 5, the eigenspace of $A = \bar{A_0} + \Delta + E$ is close to the eigenspace of $\bar{A_0}$, from which we can deduce the community structure of the related random graph $G(2n, \frac{a}{n}, \frac{b}{n})$ (as seen at the beginning of the chapter).

In order to do so, we are going to apply Davis-Kahan theorem (in a different version with respect to Theorem 2), employing the estimates for the matrices Δ and E obtained from Step 1 and Step 2.

Hence we now prove the following

Lemma 6. *Let $\bar{W} = Span\{\bar{v_1}, \bar{v_2}\}$ and $W = Span\{v_1, v_2\}$, with $\bar{v_i}$ and v_i, for $i = 1, 2$ be the eigenvectors related to the two largest eigenvalues of $\bar{A_0}$ and A respectively. For any constant $c < 1$, we can choose constants $C_2, C_3 > 0$ such that if $a \geq C_3$ and*

$$a - b \;\geq\; C_2\sqrt{a+b} \;=\; C_2\sqrt{d}, \tag{3.43}$$

then

$$\sin\angle(\bar{W}, W) \leq c < 1$$

with probability $1 - o(1)$.

Proof. The lemma follows from Davis - Kahan $\sin\Theta$ theorem as in [2] or [7]. For our matrices $\bar{A_0}$ and $A = \bar{A_0} + \Delta + E$ we find that

$$\sin\angle(\bar{W}, W) \leq \frac{\|A - \bar{A_0}\|}{a - b} = \frac{\|\Delta + E\|}{a - b}. \tag{3.44}$$

Since C_3 is a constant such that $a \geq C_3$, Corollary 2 says that

$$\|\Delta\| \leq 1.$$

Moreover, we know from Corollary 3 that exists C such that

$$\|E\| \leq C\sqrt{d}.$$

This means that, employing also the hypothesis (3.43), we can rewrite (3.44) as

$$\sin\angle(\bar{W}, W) \leq \frac{1 + C\sqrt{d}}{C_2\sqrt{d}} \leq \frac{1}{C_2}\left(\frac{1}{\sqrt{d}} + C\right) < c$$

for C_2 big enough.

□

Notice that the definition of the constants C_i in Lemma 6 and Theorem 5 is just a notation. Indeed, C_0 and C_3 can be identified as the same lower bound for the choice of a and the further condition $b > C_0$ in the Theorem 5 is not a limit for the proof of Lemma 6. Besides, C_1 can be taken as the square of C_2.

Step 4: correctness of the recovery of the community structure

Up to now, we have showed that under the hypothesis of Theorem 5, i.e. when condition (3.7) on a and b holds, the spaces $\bar{W} = \text{Span}\{\bar{v_1}, \bar{v_2}\}$ and $W = \text{Span}\{v_1, v_2\}$ given by the eigenvectors of $\bar{A_0}$ and A are close. The aim of the spectral method is to find a way to extract information about the community structure of the input graph $G(2n, \frac{a}{n}, \frac{b}{n})$ from the adjacency matrix.
Recall that A_0 is the adjacency matrix of the graph, $\bar{A_0} = \mathbb{E}[A_0]$ and A is the revised adjacency matrix obtained from the deletion (see (3.4)).

In this final step we are going to prove the correctness of Spectral Partition. Starting from Lemma 6, we are going to see that there exists a vector in W really close to the eigenvector of $\bar{A_0}$ illustrating the vertex partition (as seen in (2.4)). Such a vector is w_2 as defined into Spectral Partition. Finally, sorting the vertices of the graph in two subsets according to their values in w_2, we prove that the number of mislabeled vertices is small. As a consequence, the output partition given by the algorithm will be γ-correct for a small error rate $\gamma > 0$ (as in Definition 6).

Recall that

$$\bar{W} := \text{Span}\{\bar{v_1}, \bar{v_2}\}, \quad W := \text{Span}\{v_1, v_2\}$$

with $\bar{v_i}$ and v_i, for $i = 1, 2$ eigenvectors related to the two largest eigenvalues of $\bar{A_0} = \mathbb{E}[A_0]$ and $A = \bar{A_0} + \Delta + E$ respectively. The first result we want to prove is the following:

Lemma 7. *If* $\sin\angle(\bar{W}, W) \leq c \leq \frac{1}{4}$, *then we can find a vector* $w \in W$ *such that*

$$\sin\angle(w, \bar{v_2}) \leq 2\sqrt{c} := c'.$$

Proof. Recalling the definition (3.42), our hypothesis is that

$$\sin\angle(\bar{W}, W) = \|P_{\bar{W}} - P_W\| \leq c. \tag{3.45}$$

To prove the result, we will define a certain vector and then we will verify that such a vector satisfies the statement.
Given the eigenvectors $\bar{v_1}, \bar{v_2} \in \bar{W}$, the hypothesis (3.45) implies that, for $i = 1, 2$,

$$\|P_W \bar{v_i} - \bar{v_i}\| \leq c. \tag{3.46}$$

For $i = 1, 2$, let

$$u_i := P_W \bar{v_i} \quad \text{and} \quad x_i := u_i - \bar{v_i}$$

Observe that condition (3.46) implies that $\|x_i\| \leq c$.
Now, let $w \in W$ be a unit vector perpendicular to u_1: we want to prove that this unit vector, orthogonal to the projection in W, satisfies the statement.
Let

$$u_\perp := u_2 - \frac{u_1^T u_2}{\|u_1\|^2} u_1.$$

It is a vector in W such that

$$\langle u_\perp, u_1 \rangle = 0 \quad \text{and} \quad \|u_\perp\| \leq 1. \tag{3.47}$$

Recalling that the symmetry of $\bar{A_0}$ implies the orthogonality of the eigenvectors $\bar{v_1}, \bar{v_2}$, we observe that

$$|u_1^T u_2| = |x_1^T x_2 + \bar{v_1}^T x_2 + x_1^T \bar{v_2}| \leq c^2 + 2c. \tag{3.48}$$

Moreover, since $\|u_i\| \leq \|\bar{v_i}\| = 1$ as a property of orthogonal projections, we deduce that

$$c \geq \|x_i\| = \|u_i - \bar{v_i}\| \geq | \, \|u_i\| - \|\bar{v_i}\| \, | = \, \|u_i\| - 1$$

so in particular

$$\|u_i\| \geq 1 - c. \tag{3.49}$$

Then, let us compute the scalar product of $u_\perp$ and $\bar{v_2}$

$$u_\perp^T \bar{v_2} = u_2^T \bar{v_2} - \frac{(u_1^T u_2)(\bar{v_2}^T u_1)}{\|u_1\|^2}$$

and taking the absolute value we get that

$$|u_\perp^T \bar{v_2}| \geq 1 - c - \frac{(2c + c^2)c}{(1 - c)^2}. \tag{3.50}$$

As hypothesis $c \leq \frac{1}{4}$, so (3.50) can be estimated with

$$|u_{\perp}^{T}\bar{v_2}| \geq 1 - c - c\frac{18}{32}\frac{16}{9} = 1 - 2c. \tag{3.51}$$

Since $\|w\| = 1 \geq \|u_{\perp}\|$, (3.51) gives a bound for the absolute value of the scalar product between w and $\bar{v_2}$ so that

$$|w^{T}\bar{v_2}| \geq |u_{\perp}^{T}\bar{v_2}| \geq 1 - 2c. \tag{3.52}$$

Finally, thanks to the definition of cross product and (3.52), we conclude that for the unitary vectors w and $\bar{v_2}$ holds that

$$\sin\angle(w, \bar{v_2}) = \frac{\sqrt{\|w\|^2\|\bar{v_2}\|^2 - |w^{T}\bar{v_2}|^2}}{\|w\|\|\bar{v_2}\|} \leq \sqrt{1 - (1-2c)^2} \leq 2\sqrt{c}.$$

□

From now on, let w_2 be the vector in W that satisfies Lemma 7. As defined in Spectral Partition and in the proof of Lemma 7, w_2 is the unit vector in W perpendicular to the projection in to W of the first eigenvector of $\bar{A_0}$ (defined as u_1 in the proof of the Lemma and w_1 in the algorithm).

Joining Lemma 6 and Lemma 7, we can state that

Corollary 4. *For any constant $c < 1$, we can choose constants C_2 and C_3 in Lemma 6 and find a vector w_2 in W such that $\sin\angle(\bar{v_2}, w_2) \leq c' < 1$ with probability $1 - o(1)$.*

Since we have proved that w_2 is really close to the eigenvector of $\bar{A_0}$ exhibiting the vertex partition, we now want to exploit the information inside w_2 to extract a candidate partition. Spectral Partition splits the set of $2n$ vertices in two subsets $V_1^{'}$ and $V_2^{'}$ according to the values in w_2. We now prove that the number of misclassified vertices in $(V_1^{'}, V_2^{'})$ with respect to the real partition (V_1, V_2) is small. Namely, we see that if the angle between the spaces $\bar{W}$ and W is really small, then the output candidate partition is γ-correct for a really small error rate $\gamma > 0$, i.e.

$$|V_i \cap V_i^{'}| \geq (1 - \gamma)n.$$

for $i = 1, 2$.

Lemma 8. *Let $\bar{W} = Span\{\bar{v_1}, \bar{v_2}\}$ and $W = Span\{v_1, v_2\}$, with $\bar{v_i}$ and v_i, for $i = 1, 2$ be the eigenvectors related to the two largest eigenvalues of $\bar{A_0} = \mathbb{E}[A_0]$ and $A = \bar{A_0} + \Delta + E$ respectively. If $\sin\angle(\bar{W}, W) \leq c < 1/16$, we can recover a $8c/3$ - correct partition.*

In spite of proving directly Lemma 8, we are going to demonstrate the following equivalent deterministic fact:

Lemma 9. *If $\sin\angle(\bar{v_2}, w_2) < c' \leq \frac{1}{2}$, then we can identify at least a $(1 - \frac{4}{3}c'^2)$ fraction of vertices from each block correctly.*

Proof. Given w_2 the vector that satisfies the hypothesis for the angle, let us define the two sets of vertices

$$V_1^{'} = \{i \,|\, w_2(i) > 0\} \text{ and } V_2^{'} = \{i \,|\, w_2(i) < 0\}.$$

One of the sets will have less than or equal to n vertices, let us suppose $|V_1^{'}| \leq n$. We decompose the vector w_2 as a sum

$$w_2 = c_1 \bar{v_2}' + \mathbf{err},$$

where $\bar{v_2}'$ is the vector parallel to $\bar{v_2}$ with entries $\pm\frac{1}{\sqrt{n}}$ and $\mathbf{err}$ is a vector perpendicular to $\bar{v_2}$ and with $\|\mathbf{err}\| < c'$. Observe that we can obtain a relation for c' and c_1 so that

$$c_1 > \sqrt{1 - c'^2}.$$

Now, starting from the size of $\mathbf{err}$, we bound the number of entries of this vector that can be greater than a certain quantity, namely $\frac{\sqrt{1-c'^2}}{\sqrt{n}}$. In order to do so, we consider the extremum case in which all the other entries are zeros and we deduce the requested bound for $\xi = |\{\, i \,|\, \mathbf{err}(i) > \frac{\sqrt{1-c'^2}}{\sqrt{n}} \,\}|$:

$$c'^2 > \|\mathbf{err}\|^2 \geq \xi \left(\frac{1 - c'^2}{n}\right) \Leftrightarrow \xi < \frac{c'^2}{1 - c'^2} n.$$

Finally, we want to find a lower bound for the number of vertices, among those that realize $\bar{v_2}'(i) = \frac{1}{\sqrt{n}}$, such that $w_2(i) > 0$. Since $w_2 = c_1 \bar{v_2}' + \mathbf{err}$ and thanks to the previous estimate on ξ, there are at least $\left(1 - \frac{c'^2}{1-c'^2}\right)n$ coordinates of w_2 that are positive and that are correctly located in the set $V_1^{'}$.
In particular, since $c' \leq \frac{1}{2}$, we obtain that

$$\left|\left\{i | w_2(i) > 0 \text{ with } \bar{v_2}(i) = \frac{1}{\sqrt{n}}\right\}\right| > \left(1 - \frac{c'^2}{1 - c'^2}\right)n \geq \left(1 - \frac{4}{3}c'^2\right)n.$$

□

Summing up, given a sparse graph $G\left(2n, \frac{a}{n}, \frac{b}{n}\right)$ such that the probabilities satisfy the conditions $a > b > C$ and

$$\frac{(a-b)^2}{a+b} \geq C'$$

for some constant $C, C' > 0$, we have proved that, handling the adjacency matrix of the graph as in the steps of Spectral Partition, we produce a γ-correct vertex partition. Thus, we get a solution for the community detection problem on a SBM using a spectral method. Moreover, the error rate $\gamma > 0$ can be small according to the values of the bounds C and C'.

3.2 Refinement of the method with the addition of a correction algorithm

In this second part, we would like to make some modifications on Spectral Partition algorithm in order to get a good dependence between γ and the quantities

a and b describing the probabilities among vertices, and thus, obtain a better recovery. The idea will be to identify the mislabeled vertices and move them in the correct community.

Recall that, given the sparse random graph $G(2n, \frac{a}{n}, \frac{b}{n})$, Spectral Partition produces a partition $(V_1^{'}, V_2^{'})$ which is γ-correct with respect to the real partition (V_1, V_2), namely

$$|V_i \cap V_i^{'}| \geq (1-\gamma)n$$

for $i = 1, 2$ and $\gamma > 0$.
Let v be a vertex belonging to $V_1^{'} \cap V_2$, i.e. a vertex put in the wrong block. Since $v \in V_2$ and $|V_1|, |V_2| = n$, we expect that

$$\mathbb{E}[\,|\text{ neighbors of } v \text{ in } V_1|\,] = n\frac{b}{n} = b$$
$$\mathbb{E}[\,|\text{ neighbors of } v \text{ in } V_2|\,] = n\frac{a}{n} = a.$$

From now on, suppose that Spectral Partition outputs a 0.1-correct partition; it means that we have a partition $(V_1^{'}, V_2^{'})$ such that

$$|V_i \cap V_i^{'}| \geq 0.9n$$

or equivalently

$$|V_i \setminus V_i^{'}| \leq 0.1n$$

for $i = 1, 2$.
In this case, the vertex $v \in V_1^{'} \cap V_2$ as above is supposed to be such that

$$\mathbb{E}[\,|\text{ neighbors of } v \text{ in } V_1^{'}|\,] \leq 0.1n\frac{a}{n} + 0.9n\frac{b}{n} = 0.1a + 0.9b$$
$$\mathbb{E}[\,|\text{ neighbors of } v \text{ in } V_2^{'}|\,] \geq 0.9n\frac{a}{n} + 0.1n\frac{b}{n} = 0.9a + 0.1b.$$

Since $0.1a+0.9b < \frac{a+b}{2} < 0.9a+0.1b$, it is possible to determine if a certain vertex has been assigned to the right or wrong community looking at a thresholding.

In principle, this kind of thresholding should raise some issues since it is expressed in terms of expectation: in the proof of the main result for this section (Lemma 10), we will see that the majority of mislabeled vertices can be detected this way. Moreover, it is exactly at this step that we are going to determine the relation between γ and a and b.

We now go back to Spectral Partition and we want to point out a little serious problem within it: once the algorithm has run, the neighbors of every vertex are no longer random. It means that within the community detection resolution, there is a lack of independence and randomness.
A solution for this problem is a new algorithm, in which we introduce a step that randomly assigns the edges of the input graph to a *Red graph* or a *Blue graph*:

- on the *Red* edges, we apply **Spectral Partition**;
- for the *Blue* part we introduce a new procedure (the sub-routine **Correction**) that detects the mislabeled vertices.

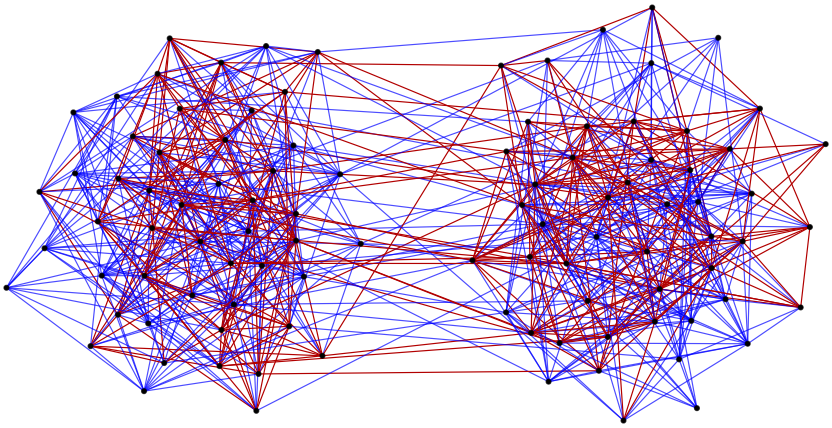


Figure 3.1: The edges of the input random graph are randomly assigned to a *Red* or a *Blue* subgraph. This particular model is generated by $G(100, 15/50, 1/50)$.

Therefore, let us define

Algorithm 3 Partition

1: Input: the adjacency matrix A_0, $d := a + b$.

2: Randomly color the edges with Red and Blue with equal probability.

3: Run **Spectral Partition** on the Red graph, outputting $V_1^{'}, V_2^{'}$.

4: Run **Correction** on the Blue graph.

5: Output the corrected sets $V_1^{'}, V_2^{'}$.

where the sub-routine **Correction** is given by

Algorithm 4 Correction

1: Input: a partition $V_1^{'}, V_2^{'}$ and a Blue graph on $V_1^{'} \cup V_2^{'}$.

2: For any $v \in V_1^{'}$, label v *bad* if the number of neighbors of v in $V_2^{'}$ is at least $\frac{a+b}{4}$ and *good* otherwise.

3: Do the same for any $u \in V_2^{'}$.

4: Correct $V_i^{'}$ by deleting its bad vertices and adding the bad vertices from $V_{3-i}^{'}$.

At last, we have to show that the method given by Partition produces a good output. In section 3.1 we have already proved that Spectral Partition produces a γ-correct vertex partition. Then, what remains to prove is the correctness of Correction. In particular, taking as input a 0.1-correct partition $(V_1^{'}, V_2^{'})$, we

prove that the sub-routine Correction outputs an *optimal* vertex partition in the following sense:

Lemma 10. *Given a* 0.1*-correct partition* $(V_1^{'}, V_2^{'})$ *and a Blue graph on* $V_1^{'} \cup V_2^{'}$ *as input to the sub-routine* ***Correction****, we obtain a* γ*-correct partition with* $\gamma = 2\exp(-0.072\frac{(a-b)^2}{a+b})$.

Proof. We consider a 0.1-correct partition $V_1^{'}, V_2^{'}$ as output of Spectral Partition on the Red graph. Besides, since the graph is sparse, we say that with probability $1 - o(1)$ the maximum Red degree of a vertex u is at most $d(u) = \log^2 n$. Let us focus on the Blue graph. Let $e = (u, v)$ be an edge between the vertices u and v. For $i = 1, 2$

$$\mathbb{P}(e \text{ is Red}) = \mathbb{P}(e \text{ is Blue}) = \begin{cases} \frac{a}{2n} & \text{if } u, v \in V_i \\ \\ \frac{b}{2n} & \text{if } u \in V_i, v \in V_{3-i} \end{cases}.$$

Then, for $i = 1, 2$,

$$\mathbb{P}(e \text{ is Blue} \,|\, e \text{ is not Red}) = \begin{cases} \frac{a/2n}{(1-\frac{a}{n})+\frac{a}{2n}} = \frac{a/2n}{1-\frac{a}{2n}} := \tau & \text{if } u, v \in V_i \\ \\ \frac{b/2n}{(1-\frac{b}{n})+\frac{b}{2n}} = \frac{b/2n}{1-\frac{b}{2n}} := \mu & \text{if } u \in V_i,\ v \in V_{3-i} \end{cases}. \tag{3.53}$$

Define ξ_i^u and ζ_j^u iid indicator random variables with mean μ and τ respectively. Then, for $i = 1, 2$,

$$\left|\left\{\text{Blue neighbors in } V_{3-i}^{'} \text{ for } u \in V_i^{'} \cap V_i\right\}\right| \leq \sum_{i=1}^{0.9n} \xi_i^u + \sum_{j=1}^{0.1n} \zeta_j^u := S(u)$$

$$\left|\left\{\text{Blue neighbors in } V_{3-i}^{'} \text{ for } u \in V_i^{'} \cap V_{3-i}\right\}\right| \geq \sum_{i=1}^{0.9n-d(u)} \zeta_i^u + \sum_{j=1}^{0.1n} \xi_j^u := S'(u).$$

The second step of Correction identifies $u \in V_1^{'}$ as a *bad* vertex if

- $u \in V_1^{'} \cap V_1$ and $S(u) \geq \frac{a+b}{4}$;
- $u \in V_1^{'} \cap V_2$ and $S'(u) \leq \frac{a+b}{4}$.

Thus, putting

$$\rho_1 := \mathbb{P}\left(S(u) \geq \frac{a+b}{4}\right) \qquad \rho_2 := \mathbb{P}\left(S'(u) \leq \frac{a+b}{4}\right), \tag{3.54}$$

the misclassified vertices in $V_1^{'}$ are at most

$$M := \sum_{k=1}^{n} \Gamma_k + \sum_{l=1}^{0.1n} \Lambda_l \tag{3.55}$$

where Γ_k and Λ_l are iid indicator random variables with mean ρ_1 and ρ_2 respectively.
Now our goal becomes to estimate the misclassified vertices M.

Using Chernoff inequality 3, we want to find an estimate for ρ_1 and ρ_2: first observe that

$$\begin{aligned}\mathbb{E}[S(u)] &= 0.9n\mu + 0.1n\tau \\ &= 0.9n\left(\frac{b/2n}{1-\frac{b}{2n}}\right) + 0.1n\left(\frac{a/2n}{1-\frac{a}{2n}}\right) \\ &= 0.9\frac{b}{2} + 0.9\frac{b}{2}\left(\frac{1}{1-\frac{b}{2n}} - 1\right) + 0.1\frac{a}{2} + 0.1\frac{a}{2}\left(\frac{1}{1-\frac{a}{2n}} - 1\right), \quad (3.56)\end{aligned}$$

where in the last passage we have added and subtracted $0.9\frac{b}{2}$ and $0.1\frac{a}{2}$.
Next, set

$$t := \frac{a+b}{4} - \mathbb{E}[S(u)].$$

We equivalently have

$$\begin{aligned}t &= 0.2a - 0.2b - 0.9\frac{b}{2}\left(\frac{1}{1-\frac{b}{2n}} - 1\right) - 0.1\frac{a}{2}\left(\frac{1}{1-\frac{a}{2n}} - 1\right) \\ &= 0.2(a-b) - 0.9\frac{b}{2}\frac{b}{2n-b} - 0.1\frac{a}{2}\frac{a}{2n-a} \qquad (\text{with } a, b < n) \\ &\geq 0.2(a-b) - 0.9\frac{b}{2}\frac{b}{n} - 0.1\frac{a}{2}\frac{a}{n} = a\left(0.2 - 0.05\frac{a}{n}\right) - b\left(0.2 + 0.45\frac{b}{n}\right) \\ &\geq 0.19(a-b) \quad (3.57)\end{aligned}$$

for n sufficiently large.
Then we apply Chernoff 3 and get that

$$\begin{aligned}\rho_1 &= \mathbb{P}\left(S(u) \geq \frac{a+b}{4}\right) \\ &= \mathbb{P}\left(S(u) \geq \mathbb{E}[S(u)] + t\right) \\ &\leq \exp\left(-\frac{(0.19(a-b))^2}{2(0.9n\mu + 0.1n\tau) + 0.19(a-b)}\right). \quad (3.58)\end{aligned}$$

Since we would like to simplify the expression (3.58), we modify the denominator in this way:

$$2(0.9n\mu + 0.1n\tau) + 0.19(a-b) = 1.8n\frac{b/2n}{1-\frac{b}{2n}} + 0.2n\frac{a/2n}{1-\frac{a}{2n}} + 0.19a - 0.19b$$

$$= 0.9b + 1.8\frac{b}{2}\left(\frac{1}{1-\frac{b}{2n}} - 1\right) + 0.1a + 0.2\frac{a}{2}\left(\frac{1}{1-\frac{a}{2n}} - 1\right) + 0.19a - 0.19b$$

$$= 0.29a + 0.71b + o(1)$$

$$\leq \frac{a+b}{2} \tag{3.59}$$

for n sufficiently large and for $a > b$ big enough.
Consequently, we employ the estimate (3.59) into (3.58) and find that

$$\rho_1 \leq \exp\left(-\frac{0.036(a-b)^2}{\frac{a+b}{2}}\right) = \exp\left(-0.072\frac{(a-b)^2}{a+b}\right). \tag{3.60}$$

Proceeding in the same way, we have an equivalent estimate for ρ_2, where the presence of $d(u) = \log^2 n$ is negligible. Therefore, recalling the definition (3.55) of M and using the bound on ρ_1, ρ_2 given by (3.60), we can state that

$$\mathbb{E}[M] = n\rho_1 + 0.1n\rho_2 \leq 1.1n\exp\left(-0.072\frac{(a-b)^2}{a+b}\right). \tag{3.61}$$

Applying Chernoff inequality 3 with $t = 0.9n\exp\left(-0.072\frac{(a-b)^2}{a+b}\right)$, we deduce that

$$\mathbb{P}\left(M \geq \mathbb{E}[M] + t\right) \leq \exp\left(-\frac{\left(0.9n\exp\left(-0.072\frac{(a-b)^2}{a+b}\right)\right)^2}{2(n\rho_1 + 0.1n\rho_2) + 0.9n\exp\left(-0.072\frac{(a-b)^2}{a+b}\right)}\right)$$

$$\leq \exp\left(-Cn\exp\left(-0.072\frac{(a-b)^2}{a+b}\right)\right) \tag{3.62}$$

for a certain constant C. So, finally, we can state that with probability $1-o(1)$,

$$M \leq \mathbb{E}[M] + t = 2n\exp\left(-0.072\frac{(a-b)^2}{a+b}\right). \tag{3.63}$$

In conclusion, these computations lead to affirm that with probability $1-o(1)$, the misclassified vertices in $V_1^{'}$, and equivalently for those in $V_2^{'}$, are

$$|V_1^{'} \setminus V_1| \leq 2n\exp\left(-0.072\frac{(a-b)^2}{a+b}\right).$$

This means that setting $\gamma := 2\exp\left(-0.072\frac{(a-b)^2}{a+b}\right)$, we have proved that, for $i = 1, 2$,

$$|V_i \cap V_i^{'}| = n - |V_i \cap V_{3-i}^{'}| = n - |V_{3-i}^{'} \setminus V_{3-i}| \geq n(1-\gamma).$$

□

With this proof, we have shown that, given a 0.1-correct partition as output of Spectral Partition on the Red graph, the Correction algorithm outputs a corrected partition $(V_1^{'}, V_2^{'})$ which is γ-correct, where γ is related to a and b in this way:

$$\frac{(a-b)^2}{a+b} = \frac{1}{0.072} \log \frac{2}{\gamma} \approx 13.89 \log \frac{2}{\gamma}. \tag{3.64}$$

Notice that Spectral Partition relies on the truthfulness of Theorem 5, in which we required that

$$\frac{(a-b)^2}{a+b} \geq C$$

for a sufficiently large constant C. On the other hand, Correction requires also the relation between γ and a and b exhibited in (3.64). This is not a problem: in fact, if $\gamma < \epsilon$ for a sufficiently small ϵ, the condition (3.64) given by Lemma 10 implies hypothesis (3.7) of Theorem 5.

In general, in order to apply **Partition**, we can substitute condition (3.7) of Theorem 5 with the following hypothesis:

$$\frac{(a-b)^2}{a+b} \geq C_1 \log \frac{1}{\gamma} \tag{3.65}$$

for a certain constant C_1.
In this way, both **Spectral Partition** and **Correction** work and we can run the algorithm obtaining the desired community vertex partition.

Chapter 4

Spectral method for a Stochastic Block Model with k communities

In this final chapter, we broach the crucial subject of this thesis. We are going to display a spectral method to solve the community detection problem in a sparse (as defined in Chapter 1) Stochastic Block Model with $k > 2$ communities. The previous chapters, in particular Chapter 3, are the starting point from wich we will deduce and prove the method for the general case $k > 2$. We should think of this part as an extension of what we already saw for $k = 2$ in Chapter 3, however the general case involves several complications.

Let us consider the SBM $G(n, \frac{a}{n}, \frac{b}{n})$ with $k > 2$ communities. Given a set of $n = |V|$ vertices, this random graph has a community structure described by k subsets $V_1, \dots, V_k$, each of size $\frac{n}{k}$. The probability distribution of the edges is the following:

- an edge between vertices belonging to the same community appears with probability $\frac{a}{n}$;
- an edge between vertices belonging to different communities appears with probability $\frac{b}{n}$,

with $n > a > b > 0$. Observe that we standardize the probabilities and we do not differentiate them according to the different communities as in the more general case of Definition 5.
Like in the previous chapters, we are going to solve the community detection problem for this model $G(n, \frac{a}{n}, \frac{b}{n})$, recovering the partition $(V_1, \dots, V_k)$ by means of a spectral algorithm.

The method we are going to develop for the case of $k > 2$ communities essencially relies on the steps that we exhibited in Chapter 3 for the case $k = 2$. However, there will be some complications: for example, it is not obvious how to approximate $k > 2$ eigenvectors. Hence, we are going to add new steps to

the method, in which we will introduce more random splittings of the edges and also of the vertices of the graph.

The structure of the algorithm will reflect that of the previous seen Partition: given as input the adjacency matrix of the random graph, the algorithm will randomly assign all the edges to a *Red* or a *Blue* subgraph. Then, it will run a spectral method on (a part of) the *Red* graph and will correct the first candidate partition using the left *Blue* (and also *Red*) edges. Since there will be many passages and different random splittings, we now describe the method step by step.

Let A_0 be the $n \times n$ adjacency matrix related to $G\left(n, \frac{a}{n}, \frac{b}{n}\right)$. In order to preserve the randomness of the community detection problem, we split the edges in two sets as in 3.2, namely we randomly assign each edge of the graph to a *Red* or a *Blue* subgraph (see Figure 4.1).

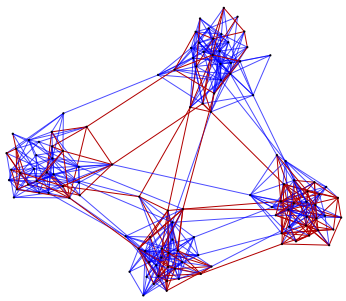


Figure 4.1: The edges of the input random graph are randomly assigned to a *Red* or a *Blue* subgraph. In this representation, we see a SBM with $k = 4$ communities over 120 vertices.

Furthermore, we randomly split the set of vertices as

$$V := Y \cup Z. \tag{4.1}$$

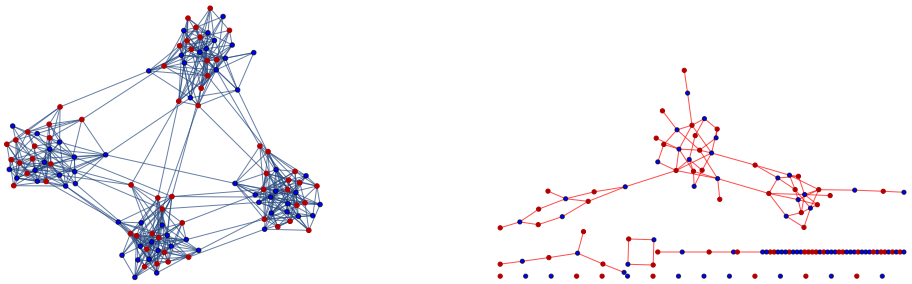
We are going to focus on the vertices in Y and Z separately. In particular, we are going to look for the community structures of the two subsets, so that the global clustering will be given by their union.

Let us define B as the adjacency matrix of the bipartite graph consisting only of the *Red* edges between Y and Z. We index the rows by Z and the columns by Y, then $B \in M_{|Z|\times|Y|}$. Notice that matrix B does not keep track of the edges between vertices in the same subset and of *Blue* edges between Y and Z (see Figure 4.2).

Matrix B becomes the input of the spectral part of the algorithm. The aim of the spectral method is to find an approximation of the eigenvectors of the expected value of the adjacency matrix: as shown for $k = 2$ in (2.4), the entries of such vectors determine the partition in communities. The steps in Spectral Partition are no more sufficient to find a k-community structure: we now need to make other random splittings and compute different kinds of operations.

First, let us randomly split the subset Y:

$$Y := Y_1 \cup Y_2. \tag{4.2}$$



(a) (b)

Figure 4.2: Given a random graph as in Figure 4.1, we randomly assign the vertices to two subsets Y and Z, (a). Then, we exclude all the *Blue* edges and the internal *Red* edges, (b). The above defined matrix B represents only the *Red* edges between the two subsets that remain in (b).

Let A_1 and A_2 be the submatrices of B given by the columns related to vertices belonging to Y_1 and Y_2 respectively. In particular, rearranging the vertices in such a way that $Y = \{Y_1, Y_2\}$, we can think of B as

$$B \sim \left(\begin{array}{c|c} A_1 & A_2 \end{array} \right) . \tag{4.3}$$

Now we do the **deletion**, as showed in (3.4), on the submatrix A_1 for

$$\delta = 20d := 20\,(a + (k-1)b)\,,$$

namely, we zero out the rows and columns of A_1 related to vertices with degree greater than $20d$. Losing part of the information is a cost that we pay in order to find better bounds on the involved matrices.
Then, let A be the resulting matrix after the deletion on A_1 and let W be the space spanned by the k left singular vectors of A (we recall the definition of singular values and singular vectors in A.3.1). The next step consists in taking randomly $m = 2\log n$ columns of A_2, let us say $a_1, \dots, a_m$. Defined $\tilde{a}$ a constant vector such that

$$\tilde{a}(j) := \frac{a+b}{2n} \qquad \text{for every } j \in Z, \tag{4.4}$$

then, project the vectors

$$a_i - \tilde{a} := b_i \tag{4.5}$$

for every $i = 1, \dots, m$ onto W.
Now, as in the methods developed in Chapter 2 and 3, let us look to the values of the vectors: select the higher $\frac{n}{2k}$ coordinates of the $m = 2\log n$ projected vectors. Then, keep half of the subsets of coordinates (i.e. $\log n$), taking those

with higher *Blue* edge density.
Finally, call $U_1^{'}, \ldots, U_k^{'}$ k of them with the following property:

$$|U_i{}' \cap U_j^{'}| < 0.2\frac{n}{2k} \quad \text{for } i \neq j. \tag{4.6}$$

If we say that Z is clustered as

$$Z = U_1 \cup \cdots \cup U_k \tag{4.7}$$

with

$$U_i := Z \cap V_i$$

for every $i = 1, \ldots, k$, then the generated partition of vertices $(U_1^{'}, \ldots, U_k^{'})$ is an approximation of the real partition $(U_1, \ldots, U_k)$ of Z .

We can then summarize the spectral part of the method as

Algorithm 5 k-Spectral Partition

1: Input: B (a matrix of dimensions $|Z| \times |Y|$), a, b and k.

2: Let Y_1 be a random subset of Y by selecting each element with probability $\frac{1}{2}$ independently and let A_1, A_2 be the sub-matrices of B formed by the columns indexed by $Y_1, Y_2 := Y \setminus Y_1$, respectively.

3: Deletion: let $d := a + (k-1)b$. Zero out all the rows and columns of A_1 corresponding to vertices whose degree is bigger than $20d$, and obtain the matrix A.

4: Find the space spanned by k left singular vectors of A, say W.

5: Let $a_1, \ldots, a_m$ be some $m = 2\log n$ random columns of A_2. For each i, project $a_i - \tilde{a}$ onto W, where $\tilde{a}(j) = \frac{a+b}{2n}$ for all j is a constant vector.

6: For each projected vector, identify the top (in value) $\frac{n}{2k}$ coordinates. Of the $2\log n$ sets so obtained, discard half of the sets with the lowest Blue edge density in them.

7: For the remaining subsets, identify some k subsets $U_1^{'}, \ldots, U_k^{'}$ such that $|U_i^{'} \cap U_j^{'}| < 0.2\frac{n}{2k}$, for $i \neq j$.

8: Output $(U_1^{'}, \ldots, U_k^{'})$.

This first sub-routine k-Spectral Partition produces a first candidate partition only for the vertices in the subset Z. Starting from $(U_1^{'}, \ldots, U_k^{'})$, the other part of the method focuses on a correction of such candidate partition of Z and looks for the partition of the remaining vertices in Y. The final result will be given by the union of the two sub-partitions.

The next step already works on the *Red* edges, but now we consider only internal edges between vertices of Z.

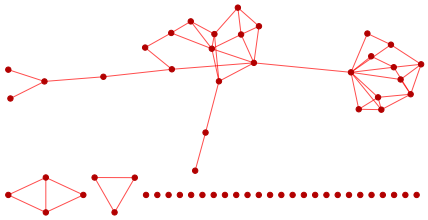


Figure 4.3: Now we just keep the *Red* edges between vertices belonging to the same subset Z.

The goal is to improve the accuracy of the output of k-Spectral Partition. Looking at the degrees of the vertices, every $u \in Z$ is assigned to a revised partition $(U_1^{"}, \ldots, U_k^{"})$. Namely, we define

Algorithm 6 k-Correction

1: Input: a vertex partition $(U_1^{'}, \ldots, U_k^{'})$ of Z and the *Red* edges internal to Z.

2: For every $u \in Z$, if $i \in \{1, \ldots, k\}$ is such that u has maximal neighbors in $U_i^{'}$, then add u to $U_i^{"}$. Break ties arbitrarily.

3: Output $(U_1^{"}, \ldots, U_k^{"})$.

Finally, it remains to find a clustering for Y. We now get the *Blue* edges back, but we just need those between Y and Z.

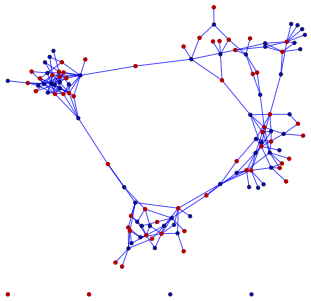


Figure 4.4: In the final step, we focus on *Blue* edges between vertices of the two different subsets Y and Z.

Similarly to what we did in Correction for the case $k = 2$, we classify the vertices of Y by thresholding on their neighbors. In this way, we produce a partition $(X_1^{'}, \ldots, X_k^{'})$ of Y where

$$Y = X_1 \cup \cdots \cup X_k$$

with

$$X_i := Y \cap V_i.$$

Algorithm 7 k-Merging

1: Input: a vertex partition $(U_1^{'}, \ldots, U_k^{'})$ of Z and the *Blue* edges between Y and Z.

2: For every $u \in Y$, label u with "i" if the number of neighbors of u in $U_i^{'}$ is at least $\frac{a+b}{8}$ and add u to $X_i^{'}$. Label the conflicts arbitrarily.

3: Output $(X_1^{'}, \ldots, X_k^{'})$.

Therefore, following the three sub-routines k-Spectral Partition, k-Correction and k-Merging we have produced separately a partition of the random subsets of vertices Y and Z. The final step consists in unifying the vertex clusters according to the same index:

$$(V_1^{'}, \ldots, V_k^{'}) := (U_1^{"} \cup X_1^{'}, \ldots, U_k^{"} \cup X_k^{'}). \tag{4.8}$$

This is our candidate vertex partition for the SBM $G\left(n, \frac{a}{n}, \frac{b}{n}\right)$.

We summarize it all in the following algorithm:

Algorithm 8 k-Partition

1: Input: the adjacency matrix A_0, a and b.

2: Randomly color the edges with Red and Blue with equal probability.

3: Randomly partition V into two subsets Y and Z. Let B be the adjacency matrix of the bipartite graph between Y and Z consisting only of the Red edges, with rows indexed by Z and columns indexed by Y.

4: Run **k-Spectral Partition** on matrix B, and get $(U_1^{'}, \ldots, U_k^{'})$ as output. This part uses only the Red edges that go between vertices in Y and Z and outputs an approximation to the clustering in $Z = U_1 \cup \cdots \cup U_k$, where $U_i := Z \cap V_i$.

5: Run **k-Correction** on the Red graph. This procedure only uses the Red edges that are internal to Z and improves the clustering in Z.

6: Run **k-Merging** on the Blue graph. This part uses only the Blue edges that go between vertices in Y and Z and assigns the vertices in Y to a cluster in $(X_1^{'}, \ldots, X_k^{'})$.

7: Output $(V_1^{'}, \ldots, V_k^{'}) := (U_1^{"} \cup X_1^{'}, \ldots, U_k^{"} \cup X_k^{'})$.

Later on, we are going to show that the output produced by k-Partition is a *good* vertex partition. More precisely, extending Definition 6 to the case $k > 2$, we will say that $(V_1^{'}, \ldots, V_k^{'})$ is γ-correct in the following sense:

Definition 7. *Let $G \sim G\left(n, \frac{a}{n}, \frac{b}{n}\right)$ be a random graph with a community structure given by the partition $(V_1, \dots, V_k)$ of V. Let $(V_1^{'}, \dots, V_k^{'})$ be a candidate partition obtained as output of any method. Then, we define $(V_1^{'}, \dots, V_k^{'})$ a* ***γ-correct partition*** *of V if*

$$|V_i \cap V_i^{'}| \geq (1-\gamma)\frac{n}{k} \tag{4.9}$$

or equivalently

$$|V_i \setminus V_i^{'}| \leq \gamma\frac{n}{k} \tag{4.10}$$

for any $i = 1, \dots, k$.

In the same way we did in Chapter 3, we can treat γ as an error rate, thus we wish to obtain a γ-correct partition with γ small.

In the following part, we will proceed this way:

- Step 1: we analyze the algorithm k-Spectral Partition using the machinery developed in section 3.1. Here we do not directly manage the adjacency matrix of the input graph, but we handle the two blocks A_1 and A_2 of the matrix B as in (4.3). Indeed, we are going to rewrite the submatrix A_1 as a difference of matrices involving its expected value. Bounding the norm of new defined matrices, we will find particular conditions for a, b and k such that it will be possible to apply Davis - Kahan theorem, and thus to go along the steps of section 3.1. Then, we will use the columns of the block A_2 to produce the vectors whose entries will permit us to define the partition.

  The truthfulness of k-Spectral Partition is related to the following

  Theorem 6. *There are constants $C_1, C_2 > 0$ such that for any fixed integer k, $\gamma > 0$ and $a > b > C_1$ satisfying*

  $$\frac{(a-b)^2}{a} \geq C_2 k^2 \tag{4.11}$$

  we can find a γ-correct partition $(U_1^{'}, \dots, U_k^{'})$ of Z with high probability using k-Spectral Partition.

- Step 2: we take a 0.1-correct partition of Z produced by k-Spectral Partition and we try to make a further correction detecting the mislabeled vertices. We prove that the sub-routine k-Correction outputs a modified partition $(U_1^{"}, \dots, U_k^{"})$ which is γ-correct, with

  $$\gamma = 2k \exp\left(-0.04\frac{(a-b)^2}{k(a+b)}\right). \tag{4.12}$$

- Step 3: finally we focus on Y and we try to compute a partition for this remaining part of vertices. Similarly to Step 2, we prove that given a 0.1-correct partition of Z, the sub-routine k-Merging produces a γ-correct partition of Y with an error rate

  $$\gamma = 2k \exp\left(-0.0324\frac{(a-b)^2}{k(a+b)}\right). \tag{4.13}$$

Showing all these steps is equivalent to prove the following main result:

Theorem 7. *There exist constant $C_1, C_2 > 0$ such that if k is any constant as $n \to \infty$ and if*

1. $a > b > C_1$
2. $(a-b)^2 \geq C_2\, k^2\, a \log \frac{1}{\gamma}$,

then we can find a γ-correct partition with probability at least $1 - o(1)$ using a simple spectral algorithm.

Namely to prove Step 2 and Step 3 we fix an error rate $\gamma = 0.1$ for the candidate partition produced by the first algorithm k-Spectral Partition. In this way, we find the explicit relations (4.12) and (4.13). These two expressions satisfy condition *2* of Theorem 7. In fact, up to a constant, it holds that

$$\begin{aligned}
\log \frac{1}{\gamma} &= \log \left(\frac{1}{k \exp\left(-\frac{(a-b)^2}{k(a+b)}\right)} \right) \\
&= \log \left(\frac{\exp\left(\frac{(a-b)^2}{k(a+b)}\right)}{k} \right) \\
&\leq \log \left(\exp\left(\frac{(a-b)^2}{k^2(a+b)} \right) \right) \\
&= \frac{(a-b)^2}{k^2(a+b)} \\
&\leq \frac{(a-b)^2}{k^2 a}. \qquad (4.14)
\end{aligned}$$

Of course, the same computations can be done choosing any small error rate different from 0.1. The resulting relations of the kind of (4.12) and (4.13) will depend on that choice and will already satisfy Theorem 7.

Step 1: correctness of k-Spectral Partition

In this first step, we want to prove that the algorithm k-Spectral Partition permits to recover a large portion of the blocks representing the community structure in Z, given by

$$U_1 = Z \cap V_1, \ldots, U_k = Z \cap V_k.$$

Recall that Z is a randomly chosen subset of vertices of $V = Y \cup Z$ and $(V_1, \ldots, V_k)$ is the real partition of our Stochastic Block Model $G\left(n, \frac{a}{n}, \frac{b}{n}\right)$.

The main result of this part is the following

Theorem 8. *There are constants $C_1, C_2 > 0$ such that for any fixed integer k, $\gamma > 0$ and $a > b > C_1$ satisfying*

$$\frac{(a-b)^2}{a} \geq C_2 k^2, \tag{4.15}$$

we can find a γ-correct partition $(U_1', \ldots, U_k')$ of Z with probability $1-o(1)$ using k-Spectral Partition.

It is clear that Theorem 8 is an extention of Theorem 5 proved in section 3.1. The k-community structure implies a different condition for a and b: now, unlike relation (3.7), it also depends on the number of the communities.
The first part of the proof is quite similar to the proof of Theorem 5, but it deals with different matrices. The last part, on the contrary, follows another path.

Recall that all the edges of the random graph are independently assigned to a *Red* or a *Blue* subgraph (see Figure 4.1). Moreover, we defined

$$\begin{aligned}
A_0 &:= \text{ adjacency matrix of the input random graph } G\left(n, \frac{a}{n}, \frac{b}{n}\right);\\
V &:= Y \cup Z \text{ splitted randomly with equal probability;}\\
Y &:= Y_1 \cup Y_2 \text{ splitted randomly with equal probability;}\\
B &:= \text{ adjacency matrix related to the bipartite graph consisting}\\
&\quad \text{only of the Red edges between } Y \text{ and } Z, \text{ as in Figure 4.2 ;}\\
A_1, A_2 &:= \text{ sub-matrices of } B \text{ formed by the columns indexed}\\
&\quad \text{by } Y_1 \text{ and } Y_2 \text{ respectively, as in (4.3).}
\end{aligned}$$

The k-Spectral Partition algorithm takes as input the $|Z| \times |Y|$ matrix B and focuses on the two blocks A_1 and A_2. With the purpose to simplify the exposition, we say that

Definition 8. *The splitting $Y = Y_1 \cup Y_2$ is **perfect** if it holds that*

$$|Y_1 \cap V_i| = \frac{n}{4k} = |Y_2 \cap V_i| \tag{4.16}$$

for any $i = 1, \ldots, k$.

The splitting $Y = Y_1 \cup Y_2$ will almost always not be perfect, but it suffices to carry throughout an $o(1)$ error and the estimates that we are going to prove will be still true.

Let us now begin the proof of Theorem 8. Consider the sub-matrix A_1 and let A be the resulting matrix after the deletion on A_1: referring to (3.4), with deletion we intend the zeroing out of rows and columns related to vertices with degree greater than

$$20\, d = 20\, (a + (k-1)b). \tag{4.17}$$

Since we want to use the same machinery of section 3.1, let

$$\bar{A_1} = \mathbb{E}[A_1], \qquad \bar{A} = \mathbb{E}[A] \tag{4.18}$$

$$E_1 = A_1 - \bar{A_1}, \qquad E = A - \bar{A} \tag{4.19}$$

$$\Delta = \bar{A} - \bar{A_1} \tag{4.20}$$

so that we can rewrite matrix A as

$$A \;=\; \bar{A} + E \;=\; \bar{A_1} + \Delta + E. \tag{4.21}$$

The reason that leads to rewrite A as in (4.21) is the fact that we can bound the norms of Δ and E and show that A and $\bar{A_1}$ are *close* in a sense that we will define later.
Now we want to apply the estimates that we deduced in section 3.1 for the equivalent matrices Δ and E.

Firstly, notice that we can bound the number of vertices with degree greater than $20\,d$ as in Lemma 1: indeed, the hypothesis (4.15) implies that we can find a constant $d_0 > 0$ such that $d \geq d_0$. Thus, with probability $1 - \exp\left(-\Omega(a^{-2}n)\right)$ there are at most $a^{-3}n$ vertices with high degree. This means that, since $\Delta_{i,j} \leq \frac{a}{n}$ for every entry (i,j) of the matrix, we can repeat the estimate with the Hilbert-Schmidt norm and extend Corollary 2 to

Corollary 5. *Let $d = a + (k-1)b \geq d_0$ and $\Delta = \bar{A} - \bar{A_1} = \mathbb{E}[A] - \mathbb{E}[A_1]$. Then, for d_0 sufficiently large,*

$$\|\Delta\| \leq 1$$

with probability $1 - \exp\left(-\Omega(a^{-2}n)\right)$.

Furthermore, we observe that the entries of matrix $E_1 = A_1 - \bar{A_1}$ are

$$(E_1)_{u,v} = \begin{cases} 1 - \frac{a}{n} & \text{with probability } \frac{a}{n} \\ -\frac{a}{n} & \text{with probability } 1 - \frac{a}{n} \end{cases}$$

if u and v both belong to $Y_1 \cap V_i$ for some $i \in 1, \ldots, k$ and

$$(E_1)_{u,v} = \begin{cases} 1 - \frac{b}{n} & \text{with probability } \frac{b}{n} \\ -\frac{b}{n} & \text{with probability } 1 - \frac{b}{n} \end{cases}$$

if $u \in Y_1 \cap V_i$ and $v \in Y_1 \cap V_j$ with $i \neq j$.
Since the probabilities of the model $\{\frac{a}{n}, \frac{b}{n}\}$ are certainly $\leq \frac{a}{n}$, we can apply Lemma 5 to the deleted version of E_1 with $\sigma^2 = \frac{a}{n} \leq \frac{a+b}{n}$ and get a bound for its norm. In other words, we have the following result:

Lemma 11. *There exist constants $C_1, C > 0$ such that if $a > b \geq C_1$ and E is the matrix produced by the deletion on $E_1 = A_1 - \bar{A_1}$, then we have*

$$\|E\| \leq C\sqrt{a+b} \;\; \textit{with probability } 1 - o(1).$$

In short, Corollary 5 and Lemma 11 permit to bound the norms of matrices Δ and E, and now we would like to use them in an application of Davis - Kahan

theorem as in Step 3.1 of the 2-communities case. Before that, we need to make some considerations on matrix A_1.

Let us recall that we are supposing that the splitting $Y = Y_1 \cup Y_2$ is perfect, as defined in Definition 8. Since $|Y_j \cap V_i| = \frac{n}{4k}$ for $j = 1, 2$ and $i = 1, \dots, k$, in particular $B \in M_{\frac{n}{2} \times \frac{n}{2}}$ and the blocks A_1, A_2 have dimension $\frac{n}{2} \times \frac{n}{4}$. Then, rearranging the columns of A_1 and A_2 collecting the community blocks in the diagonal, we can represent matrix B before and after the splitting as

$$B \simeq \left(\quad \right) \longrightarrow \left(\quad \right) \tag{4.22}$$

where in this representation we have considered $k = 4$. Therefore, looking at the expected value of the first block, $\bar{A_1}$, we can find the same structure in blocks. Moreover, the entries of $\bar{A_1}$ are the expected values of Bernoulli variables of probabilities $\frac{a}{n}$ or $\frac{b}{n}$, according to the membership to the same community or not. With this observation, it is easy to conclude that matrix $\bar{A_1}$ has rank k.

Since we are handling a rectangular matrix, we cannot study the eigenvalues of $\bar{A_1}$. This time, unlike Chapter 3, we are going to look at the singular values (we recall the definition in A.3.1), and in particular at the least non-trivial singular value $\sigma_k\left(\bar{A_1}\right)$.
Let us make some computations in the particular case $n = 12, k = 3, \bar{A_1} \in M_{6\times 3}$. Taking $\bar{A_1}^T \bar{A_1}$, we find that its least eigenvalue is

$$2\left(\frac{a}{n} - \frac{b}{n}\right)^2 = 2\left(\frac{a-b}{12}\right)^2 = \left(\frac{a-b}{3}\right)^2 = \left(\frac{a-b}{k}\right)^2, \tag{4.23}$$

and the related singular value is the square root of (4.23). For a general choice of n and k, it is possible to prove that the least non-trivial singular value is still of the form

$$\sigma_k(\bar{A_1}) = \frac{a-b}{k}. \tag{4.24}$$

Now, let $\sigma_1, \dots, \sigma_k$ be the singular values of $\bar{A_1}$ and $\bar{v}_1, \dots, \bar{v}_k$ be the related singular vectors. Moreover, define $\lambda_1, \dots, \lambda_k$ the singular values of A, i.e. the deleted version of A_1, and $v_1, \dots, v_k$ its singular vectors. Furthermore, define

$$W := \mathrm{Span}\{v_1, \dots, v_k\}, \quad \bar{W} := \mathrm{Span}\{\bar{v}_1, \dots, \bar{v}_k\}. \tag{4.25}$$

Finally recall from (3.42) that given two vector subspaces W_1, W_2 of the same dimension, we defined

$$\sin\angle(W_1, W_2) := \|P_{W_1} - P_{W_2}\|. \tag{4.26}$$

Combining Corollary 5, Lemma 11 and the lower bound for the singular values of $\bar{A_1}$ given by (4.24), we can now apply Davis-Kahan theorem and prove that

Lemma 12. *Let $\bar{W} := Span\{\bar{v}_1, \dots, \bar{v}_k\}$ and $W := Span\{v_1, \dots, v_k\}$, with $\bar{v}_i$ and v_i, for $i = 1, \dots, k$, be the singular vectors of $\bar{A_1}$ and A respectively. For*

any constant $c < 1$, there exist constants $C_1, C_3 > 0$ such that if $a > b \geq C_1$ and

$$(a - b) > C_3\, k\, \sqrt{a}, \tag{4.27}$$

then

$$\sin \angle(\bar{W}, W) \leq c$$

with probability $1 - o(1)$.

Proof. The proof is essencially equal to the proof of Lemma 6: we apply Davis-Kahan theorem as in [2] or [7]. Then, exploiting the bounds given by Corollary 5 and Lemma 11, we get

$$\sin \angle(\bar{W}, W) \leq \frac{\|\Delta + E\|}{a - b} k \leq \frac{1 + C\sqrt{a+b}}{C_3 k \sqrt{a}} k = \frac{1}{C_3} \frac{1 + C\sqrt{a+b}}{\sqrt{a}} < c$$

for C_3 sufficiently large.

□

Notice that we can associate the constants C_2 of Theorem 8 and C_3 of Lemma 12 so that $C_2 = C_3^2$.

Now, the submatrix A_2 comes into play. Let us rewrite this block as a difference of matrices involving its expected value:

$$A_2 := \mathbb{E}[A_2] + E' = \bar{A}_2 + E'. \tag{4.28}$$

Since the splitting of Y in Y_1 and Y_2 is random and the formed blocks A_1 and A_2 have the same inner construction, we expect that the space spanned by the k left singular values of $\bar{A}_2$ matches on average with $\bar{W}$. Thus, we extend the notation $\bar{W}$ also to the related space of $\bar{A}_2$.
As indicated at step 5 of k-Spectral Partition, pick randomly $m = 2 \log n$ indices in Y_2 and $a_1, \ldots, a_m$ the related columns of A_2. Define

$$\bar{a}_1, \ldots, \bar{a}_m \quad \text{and} \quad e_1, \ldots, e_m \tag{4.29}$$

as the columns of $\bar{A}_2$ and E' respectively, corresponding to the chosen indices. Let

$$\tilde{a}(j) := \frac{a+b}{2n} \quad \text{for any } j \in Z$$

be a constant vector and define

$$b_i := \bar{a}_i - \tilde{a}.$$

Observe that if $i \in Y_2 \cap V_{n_i}$,

$$\bar{a}_i(j) = \begin{cases} \frac{a}{n} & \text{if } j \in Z \cap V_{n_i} \\ \\ \frac{b}{n} & \text{otherwise} \end{cases}$$

and in particular

$$b_i(j) = \begin{cases} \frac{a-b}{2n} & \text{if } j \in Z \cap V_{n_i} \\ \\ \frac{b-a}{2n} & \text{otherwise} \end{cases}.$$

Since both $\bar{a}_i$ and $\tilde{a}$ are in the column span of $\bar{A}_2$, for any $i \in \{1, \ldots, m\}$

$$b_i = P_{\bar{W}} b_i. \tag{4.30}$$

Moreover,

$$\|b_i\| = \sqrt{\frac{n}{4}\left(\frac{a-b}{2n}\right)^2 + \frac{n}{4}\left(\frac{b-a}{2n}\right)^2} = \frac{a-b}{2\sqrt{2n}}. \tag{4.31}$$

If we can recover b_i, we can identify the set $Z \cap V_{n_i}$. For this reason, now we show that the projections $P_W(a_i - \tilde{a})$, i.e. the vectors that we will use to recover the partition, are close enough to the b_i. Recall that W is the subspace spanned by the k left singular vectors of the deleted version of A_1.

Rewriting from (4.28)

$$a_i - \tilde{a} \;=\; \bar{a_i} + e_i - \tilde{a} \;=\; b_i + e_i,$$

we have that

$$P_W(a_i - \tilde{a}) = P_W b_i + P_W e_i = P_{\bar{W}} b_i + P_W e_i + \mathbf{err}_i \underset{\text{for (4.27)}}{=} b_i + P_W e_i + \mathbf{err}_i. \tag{4.32}$$

Now we try to bound the norms of the addends in (4.32). Even talking of A_2, we can extend Lemma 12 so that $\sin \angle(W, \bar{W}) \leq \delta_1$ for a small $\delta_1 > 0$. So,

$$\|\mathbf{err}_i\| = \|(P_W - P_{\bar{W}}) b_i\| \leq \delta_1 \|b_i\|. \tag{4.33}$$

Let us focus on $P_W e_i$. Since $\dim W = k$,

$$\mathbb{E}[\,\|P_W e_i\|^2\,] = \sum_{\tau=1}^{k} \mathbb{E}\left[(P_W e_i)_\tau^2\right] \leq \sigma^2 k, \tag{4.34}$$

where $\sigma^2 = \frac{a}{n}$ was an upper bound for the variance of the entries of matrix E that we repropose for E'. Then, by a variant of Markov's inequality (see A.1.1) we get that

$$\mathbb{P}\left(\|P_W e_i\| > 2\sigma\sqrt{k}\right) \leq \frac{\mathbb{E}[\,\|P_W e_i\|^2\,]}{(2\sigma\sqrt{k})^2} \leq \frac{\sigma^2 k}{4\sigma^2 k} = \frac{1}{4}. \tag{4.35}$$

By a simple application of Chernoff bound, it follows that

Lemma 13. *With probability at least $1 - o(1)$, at least $m/2$ of the vectors $e_1, \ldots, e_m$ satisfy*

$$\|P_W e_i\| \leq 2\sigma\sqrt{k}.$$

Let $m_1 \geq m/2$ be the number of vectors that satisfy Lemma 13, so that we define $e_{i_1}, \ldots, e_{i_{m_1}}$ as *good* vectors related to the *good* indices. Then, observe that if

$$(a-b) > C_1\sqrt{ka} \quad \text{for a constant } C_1 > 0 \tag{4.36}$$

and since $\sigma \leq \sqrt{a/n}$, it holds that

$$2\sigma\sqrt{k} \leq 2\sqrt{\frac{a}{n}}\,\frac{(a-b)}{C_1\sqrt{a}} = \frac{2}{C_1}\,\frac{(a-b)}{\sqrt{n}} = \frac{4\sqrt{2}}{C_1}\,\frac{(a-b)}{2\sqrt{2n}} \leq \delta_2\,\|b_{i_j}\|, \tag{4.37}$$

where we can take any $\delta_2 > 0$ small as the constant C_1 grows.

Therefore collecting (4.33), Lemma 13 and (4.37), we deduce that for *good* indices the projection $P_W(a_{i_j} - \tilde{a})$ is close to vertex b_{i_j}:

Lemma 14. *For any $\delta > 0$ exist costants $C_1, C_2 > 0$ such that, if $a > b > C_2$ and*

$$(a-b) > C_1\sqrt{ka},$$

then for all good indices i_j it holds that

$$\|P_W(a_{i_j} - \tilde{a}) - b_{i_j}\| \leq \delta \|b_{i_j}\|.$$

Observe that given the hypothesis of Theorem 8, the requirements for Lemma 14 are satisfied: if we ask $(a-b) > Ck\sqrt{a}$, certainly it holds that $(a-b) > C\sqrt{ka}$, with C a constant.
Going back to the algorithm k-Spectral Partition, after the projection of the $a_i - \tilde{a}$ onto W we take the first $\frac{n}{2k}$ coordinates of higher value of such vectors. Defining $U'_{i_1}, \ldots, U'_{i_{m_1}}$ the sets of selected coordinates, we can choose constants C_1, C_2 in such a way that the intersection of the U_{i_j} with $Z \cap V_{n_{i_j}}$ is large, we can say

$$|U_{i_j} \cap (Z \cap V_{n_{i_j}})| \geq 0.9\frac{n}{2k}.$$

Finally, what remains to prove is step 6 of k-Spectral Partition, namely the fact that discarding half of the sets of coordinates with the lowest *Blue* edge densities is a good choice to detect the vertex partition. For this purpose, we prove that

Lemma 15. *Let X be a subset of Z of size $|X| = \frac{n}{2k}$. Then, there exists a constant $c > 0$ such that the following hold:*

1. *if for all $i \in \{1, \ldots, k\}$*

$$|X \cap V_i| \leq 0.9|X|,$$

then with probability at least $1 - e^{-cn}$ the number of Blue edges in the graph induced by X is at most $an/16k - 0.09(a-b)n/16k$;

2. *if for some $i \in \{1, \ldots, k\}$*

$$|X \cap V_i| \geq 0.95|X|,$$

then with probability at least $1 - e^{-cn}$ the number of Blue edges in the graph induced by X is at least $an/16k - 0.09(a-b)n/16k$.

Proof. We start showing 1. Let $e(X)$ be the number of *Blue* edges in the graph induced by vertices in X. Since we suppose that $|X \cap V_i| \leq 0.9|X|$ for every $i \in \{1, \ldots, k\}$,

$$\begin{aligned}\mathbb{E}[e(X)] \leq \frac{1}{2}k\left(0.9\frac{n}{2k}\right)^2 \frac{a/2n}{1-\frac{a}{2n}} + \frac{1}{2}k\left(0.1\frac{n}{2k}\right)\left(0.9\frac{n}{2k}\right)\frac{b/2n}{1-\frac{b}{2n}} \\ + \frac{1}{2}k\left(0.1\frac{n}{2k}\right)^2 \frac{a/2n}{1-\frac{a}{2n}}\end{aligned}$$

$$= 0.81\frac{an}{16k}\frac{1}{1-\frac{a}{2n}} + 0.09\frac{bn}{16k}\frac{1}{1-\frac{b}{2n}} + 0.01\frac{an}{16k}\frac{1}{1-\frac{a}{2n}}$$

$$= 0.91\frac{an}{16k}\frac{1}{1-\frac{a}{2n}} - 0.09\frac{n}{16k}\left(\frac{a}{1-\frac{a}{2n}} - \frac{b}{1-\frac{b}{2n}}\right)$$

$$\leq \frac{an}{16k} - 0.09(a-b)\frac{n}{8k}. \tag{4.38}$$

In order to bound the desired probability, we apply Chernoff 3 with

$$t := 0.045(a-b)\frac{n}{8k}$$

so that

$$\mathbb{P}\left(e(X) \geq \mathbb{E}[e(X)] + t\right) = \mathbb{P}\left(e(X) \geq \frac{an}{16k} - 0.045(a-b)\frac{n}{8k}\right)$$

$$\leq \exp\left(-\frac{(0.045(a-b)n/8k)^2}{2an/16k + 0.045(a-b)n/8k}\right).$$

The proof of 2. is essencially equal.

□

We have thus concluded the analysis of the sub-routine k-Spectral Partition. Following the steps in the algorithm, we have proved that we can efficiently recover a vertex partition of the random chosen set of vertices Z which is γ-correct, for a small error rate γ. Starting from the output partition $(U_1^{'}, \ldots, U_k^{'})$, the other part of the method will work on a refinement of the $U_i^{'}$ and on the partition on the other set of vertices Y.

Step 2: correctness of k-Correction

In the previous step, we have proved that k-Spectral Partition produces a partition $(U_1^{'}, \ldots, U_k^{'})$ of the random subset of vertices Z which is γ-correct with respect to the real partition $(U_1, \ldots, U_k)$, i.e. for every $i = 1, \ldots, k$

$$|U_i \cap U_i^{'}| \geq (1-\gamma)\frac{n}{2k}.$$

for a small error rate $\gamma > 0$.
In the following part, we want to prove that the sub-routine k-Correction outputs a modified partition $(U_1^{"}, \ldots, U_k^{"})$ wich is optimal in a sense that we will specify later. Recall that k-Correction works on the candidate partition of Z and on the *Red* edges internal to Z (see Figure 4.3).

Assuming that the candidate partition produced by k-Spectral Partition is 0.1-correct, we focus on the mislabeled vertices and we proceed as in 3.2. In the proof of the main result of this part we will also find a relation between γ and a and b that leads to a good candidate partition.

Consider a misclassified vertex $u \in U_1 \cap U_i^{'}$ for $i \neq 1$. Since we suppose that

$|U_j \cap U_j^{'}| \geq 0.9\frac{n}{2k}$ for every $j = 1, \dots, k$, we expect that

$$\mathbb{E}\left[\left|\text{ Red neighbors of } u \text{ in } U_1^{'}\right|\right] \geq 0.9\frac{n}{2k}\frac{a}{2n} + 0.1\frac{n}{2k}\frac{b}{2n} = 0.9\frac{a}{4k} + 0.1\frac{b}{4k}$$

$$\mathbb{E}\left[\left|\text{ Red neighbors of } u \text{ in } U_i^{'}\right|\right] \leq 0.1\frac{n}{2k}\frac{a}{2n} + 0.9\frac{n}{2k}\frac{b}{2n} = 0.1\frac{a}{4k} + 0.9\frac{b}{4k}.$$

Therefore, since $0.1\frac{a}{8k} + 0.9\frac{b}{8k} < \frac{a+b}{8k} < 0.9\frac{a}{8k} + 0.1\frac{b}{8k}$, we can reclassify the mislabeled vertices looking at a thresholding. Namely, we now prove that following the procedure of k-Correction we get an optimal vertex partition of Z in the sense that:

Lemma 16. *Given a 0.1-correct partition $(U_1^{'}, \dots, U_k^{'})$ of $Z = (Z \cap V_1) \cup \dots \cup (Z \cap V_k)$ and the Red graph over Z, the sub-routine **k-Correction** computes a γ-correct partition with $\gamma = 2k \exp\left(-0.04\frac{(a-b)^2}{k(a+b)}\right)$.*

Proof. Let $(U_1^{'}, \dots, U_k^{'})$ be a 0.1-correct partition of Z related to the real partition $(U_1, \dots, U_k)$. Moreover, recall that in this part we only take the *Red* edges between vertices in Z.
Let $e = (u, v)$ be an edge between vertices u and v: for every $i, j \in \{1, \dots, k\}$,

$$\text{if } u \in U_i,\ v \in U_j \text{ with } i \neq j,\ \ \mathbb{P}\left(e \text{ is a Red edge}\right) = \frac{b}{2n} := \mu \tag{4.39}$$

$$\text{if } u, v \in U_i,\ \ \mathbb{P}\left(e \text{ is a Red edge}\right) = \frac{a}{2n} := \tau. \tag{4.40}$$

Let ξ_i^u and ζ_i^u be iid indicator random variables with mean μ and τ respectively. Then, for any $u \in U_1$,

$$\left|\left\{\text{Red neighbors of } u \text{ in } U_j^{'}\right\}\right| \leq \sum_{i=1}^{0.9n/2k} \xi_i^u + \sum_{j=1}^{0.1n/2k} \zeta_j^u := S_{1,j}(u)$$

$$\left|\left\{\text{Red neighbors of } u \text{ in } U_1^{'}\right\}\right| \geq \sum_{i=1}^{0.9n/2k} \zeta_i^u + \sum_{j=1}^{0.1n/2k} \xi_j^u := S_{1,1}(u).$$

After the correction sub-routine, if a vertex $u \in U_1$ is mislabeled, then one of the following holds:

- $S_{1,j}(u) \geq \frac{a+b}{8}$ for $j \neq 1$;
- $S_{1,1}(u) \leq \frac{a+b}{8}$.

Set

$$\rho_1 := \mathbb{P}\left(S_{1,1} \leq \frac{a+b}{8}\right) \qquad \rho_2 := \mathbb{P}\left(S_{1,j} \geq \frac{a+b}{8}\right). \tag{4.41}$$

Applying Chernoff 3 as in the proof of Lemma 10, it is possible to bound ρ_1, and similarly ρ_2, with

$$\rho_1 \leq \exp\left(-0.04\frac{(a-b)^2}{k(a+b)}\right). \tag{4.42}$$

Thus, define ρ as the probability that one of the two conditions of misclassification holds. Furthermore, let us define the number of mislabeled vertices in U_1 after the correction step as

$$M := \sum_{l=1}^{n/2k} \Gamma_l, \tag{4.43}$$

where Γ_l are iid indicator random variables with mean ρ. Since

$$\mathbb{E}[M] \leq \frac{n}{2k} k \exp\left(-0.04 \frac{(a-b)^2}{k(a+b)}\right), \tag{4.44}$$

we use again Chernoff 3 with $t := \frac{n}{2} \exp\left(-0.04 \frac{(a-b)^2}{k(a+b)}\right)$ and get that

$$\mathbb{P}\left(M \geq \mathbb{E}[M] + t\right) \leq \exp\left(-\frac{\left(\frac{n}{2} \exp\left(-0.04 \frac{(a-b)^2}{k(a+b)}\right)\right)^2}{\frac{n}{k} \exp\left(-0.04 \frac{(a-b)^2}{k(a+b)}\right) + \frac{n}{2} \exp\left(-0.04 \frac{(a-b)^2}{k(a+b)}\right)}\right)$$

$$= \exp\left(-\frac{nk}{2(2+k)} \exp\left(-0.04 \frac{(a-b)^2}{k(a+b)}\right)\right).$$

Therefore, with probability $1 - o(1)$, the number of mislabeled vertices in U_1 is

$$M \leq n \exp\left(-0.04 \frac{(a-b)^2}{k(a+b)}\right) \tag{4.45}$$

and we can repeat the same way for each of the k blocks. In conclusion, setting $\gamma := 2k \exp\left(-0.04 \frac{(a-b)^2}{k(a+b)}\right)$, we have proved that for every $i = 1, \ldots, k$

$$|U_i \cap U_i^{'}| = \frac{n}{2k} - |U_i \setminus U_i^{'}| \geq \frac{n}{2k} - n \exp\left(-0.04 \frac{(a-b)^2}{k(a+b)}\right) = \frac{n}{2k}(1-\gamma). \tag{4.46}$$

□

Thanks to Lemma 16, we obtain a relation between the error rate γ and the quantities a and b describing the probabilities in the model. Namely, taking as input a 0.1-correct partition $(U_1^{'}, \ldots, U_k^{'})$ of Z, the algorithm k-Correction with probability $1 - o(1)$ produces a modified γ-correct partition $(U_1^{"}, \ldots, U_k^{"})$ where γ satisfies

$$\frac{(a-b)^2}{k(a+b)} = \frac{1}{0.04} \log \frac{2k}{\gamma} = 25 \log \frac{2k}{\gamma}. \tag{4.47}$$

We have already observed in (4.14) that this expression of the error rate satisfies

$$(a-b)^2 > C_1 \, k^2 \, a \, \log \frac{1}{\gamma},$$

as requested in Theorem 7.

Step 3: correctness of k-Merging

Up to now, working on the *Red* edges we have found a vertex partition $(U_1^{"}, \dots, U_k^{"})$ on Z wich is optimal in the sense given by Lemma 16. What remains to do is to find a partition on the other set of vertices of our graph $G\left(n, \frac{a}{n}, \frac{b}{n}\right)$, i.e. on the subset Y. For this purpose, we are going to consider the *Blue* edges, but in particular only those edges between Y and Z. Once computed the partition for Y, we will obtain the candidate partition on the entire $V = V_1 \cup \cdots \cup V_k$.

We now prove that the sub-routine k-Merging produces a partition of Y with a really small error rate and thus will be optimal in the same sense we defined in Lemma 16. The proof of the following result is similar to the proof of Lemma 16:

Lemma 17. *Given a 0.1-correct partition of $Z = (Z \cap V_1) \cup \cdots \cup (Z \cap V_k)$ and the Blue graph between Y and Z, the sub-routine **k-Merging** computes a γ-correct partition of Y with $\gamma = 2k \exp\left(-0.0324 \frac{(a-b)^2}{k(a+b)}\right)$.*

Proof. Let $(U_1^{'}, \dots, U_k^{'})$ be a 0.1-correct partition of Z. As in step 2 of k-Merging, we are going to label the vertices in Y according to the number of neighbors in the $U_i^{'}$. Recall that the real partition of the random graph $G\left(n, \frac{a}{n}, \frac{b}{n}\right)$ is $(V_1, \dots, V_k)$ and in this proof we only consider *Blue* edges between Y and Z. Moreover, we assume that the maximum *Red* degree of a vertex is at most $d(u) := \log^2 n$.

Let $e = (u, v)$ be an edge between vertices u and v not belonging to the *Red* graph. Recalling what we deduced in (3.53), for $i, j \in \{1, \dots, k\}$,

$$\mathbb{P}(e \text{ is Blue} \mid e \text{ is not Red}) = \begin{cases} \frac{a/2n}{1-\frac{a}{2n}} := \tau & \text{if } u \in V_i \cap Y, v \in V_i \cap Z \\ \\ \frac{b/2n}{1-\frac{b}{2n}} := \mu & \text{if } u \in V_i \cap Y,\ v \in V_j \cap Z \end{cases}. \tag{4.48}$$

Set ξ_i^u and ζ_i^u iid indicator random variables with mean μ and τ respectively. Then, taken $i, j \in \{1, \dots, k\}$, for any $u \in V_i \cap Y$,

$$\left|\left\{\text{Blue neighbors of } u \text{ in } U_j^{'}\right\}\right| \leq \sum_{i=1}^{0.9n/2k} \xi_i^u + \sum_{j=1}^{0.1n/2k} \zeta_j^u := S_j(u)$$

$$\left|\left\{\text{Blue neighbors of } u \text{ in } U_i^{'}\right\}\right| \geq \sum_{i=1}^{0.9n/2k-d(u)} \zeta_i^u + \sum_{j=1}^{0.1n/2k} \xi_j^u := S_i(u).$$

After the correction sub-routine, if a vertex $u \in V_i \cap Y$ is misclassified, then one of the following holds:

- $S_j(u) \geq \frac{a+b}{8k}$;
- $S_i(u) \leq \frac{a+b}{8k}$.

Set

$$\rho_1 := \mathbb{P}\left(S_j(u) \geq \frac{a+b}{8k}\right) \quad \rho_2 := \mathbb{P}\left(S_i(u) \leq \frac{a+b}{8k}\right). \tag{4.49}$$

We follow the same idea as in the proofs of Lemma 10 and Lemma 16: we use Chernoff inequality 3 to bound the probabilities ρ_1 and ρ_2 and to bound the number of mislabeled vertices. Then, we observe that

$$\begin{aligned}\mathbb{E}[S_j(u)] &= 0.9\frac{n}{2k}\frac{b/2n}{1-\frac{b}{2n}} + 0.1\frac{n}{2k}\frac{a/2n}{1-\frac{a}{2n}} \\ &= 0.9\frac{b}{4k} + 0.1\frac{a}{4k} + 0.9\frac{b}{4k}\left(\frac{1}{1-\frac{b}{2n}} - 1\right) + 0.1\frac{a}{4k}\left(\frac{1}{1-\frac{a}{2n}} - 1\right).\end{aligned} \tag{4.50}$$

Defining

$$t := \frac{a+b}{8k} - \mathbb{E}[S_j(u)], \tag{4.51}$$

we further see that

$$\begin{aligned}t &= 0.1\frac{a}{k} - 0.1\frac{b}{k} - 0.1\frac{a}{4k}\left(\frac{1}{1-\frac{a}{2n}} - 1\right) - 0.9\frac{b}{4k}\left(\frac{1}{1-\frac{b}{2n}} - 1\right) \\ &\geq 0.1\frac{a-b}{k} - 0.9\frac{b}{4k}\frac{b}{n} - 0.1\frac{a}{4k}\frac{a}{n} \\ &\geq 0.09\,\frac{a-b}{k}\end{aligned} \tag{4.52}$$

for n sufficiently large. Applying Chernoff 3 we get

$$\rho_1 \leq \exp\left(-\frac{\left(0.09\,\frac{a-b}{k}\right)^2}{2(0.9\frac{n}{2k}\mu + 0.1\frac{n}{2k}\tau) + 0.09\frac{a-b}{k}}\right) \tag{4.53}$$

and it is possible to modify the denominator in such a way that

$$\rho_1 = \mathbb{P}\left(S_j(u) \geq \frac{a+b}{8k}\right) \leq \exp\left(-0.0324\frac{(a-b)^2}{k(a+b)}\right). \tag{4.54}$$

Similarly, we can bound ρ_2. If we set ρ as the probability that at least one of the two conditions of misclassification happens, then the number of mislabeled vertices in Y is at most

$$M := \sum_{k=1}^{n/2} \Gamma_k, \tag{4.55}$$

where the Γ_k are iid indicator random variables with mean ρ. Since

$$\mathbb{E}[M] \leq \frac{n}{2}k\exp\left(-0.0324\frac{(a-b)^2}{k(a+b)}\right),$$

we apply Chernoff 3 with $t := \frac{n}{2}k\exp\left(-0.0324\frac{(a-b)^2}{k(a+b)}\right)$ and conclude that with probability $1-o(1)$ the number of mislabeled vertices in Y is

$$M \leq \mathbb{E}[M] + t = nk\exp\left(-0.0324\frac{(a-b)^2}{k(a+b)}\right). \tag{4.56}$$

□

We have proved that given a 0.1-correct partition of Z, the algorithm k-Merging outputs a γ-correct partition of Y for which we relate γ and a and b this way:

$$\frac{(a-b)^2}{k(a+b)} = \frac{1}{0.0324} \log \frac{2k}{\gamma} \leq 31 \log \frac{2k}{\gamma}. \tag{4.57}$$

As observed at the end of Step 2, this kind of expression satisfies the hypothesis of Theorem 7, so that we can conclude that all the spectral method works when

$$(a-b)^2 \geq C\, k^2\, a\, \log \frac{1}{\gamma}$$

for a constant $C > 0$.

Summarizing, with k-Spectral Partition and k-Correction we have produced a γ-correct partition of the random subset of vertices Z, where the error rate γ is small, as given by Lemma 16. The sub-routine k-Merging outputs a partition on the remaining vertices collected in Y with a small error rate γ deduced by the proof of Lemma 17. The last step consists in gathering the two partitions and generating the clustering over the entire random graph: given $(U_1^{"}, \ldots, U_k^{"})$ and $(X_1^{'}, \ldots, X_k^{'})$ the two semi-partitions, we construct

$$(V_1^{'}, \ldots, V_k^{'}) := (U_1^{"} \cup X_1^{'}, \ldots, U_k^{"} \cup X_k^{'})$$

where we have just joint the local subsets according to the same index. Since the $X_i^{'}$ have been defined looking at the higher densities of edges between vertices of Y and Z, the partition $(V_1^{'}, \ldots, V_k^{'})$ is γ-correct with respect to the real one.

Conclusions

In this thesis we have discussed a spectral method to solve the community detection problem in a sparse Stochastic Block Model with k communities and $n >> 1$ vertices. The method for the general case $k > 2$ follows as an extension of a resolution technique for the simplest case $k = 2$.

The sparsity of the input random graph determines different approaches to the problem. In Chapter 2, we have seen that when the graph is dense, we can exploit the properties induced by the edge densities to recover with high probability the community structure correctly up to a small number of misclassified vertices.

On the contrary, if a random graph with $k = 2$ communities is sparse as in the model we have analyzed, the problem becomes more complicated. Given the SBM $G(2n, \frac{a}{n}, \frac{b}{n})$, we have proved that with high probability the algorithm Partition produces a γ-correct vertex partition (in the sense of Definition 6) with a small error rate $\gamma > 0$ when a, b and γ satisfy the condition

$$\frac{(a-b)^2}{a+b} \geq C \log \frac{1}{\gamma}$$

for a suitable constant $C > 0$.

The generalization to $k > 2$ starts as an extension of the simpler case of two communities. Nevertheless, additional steps and random splittings are needed. Our algorithm k-Partition contains three different sub-routines that work on randomly chosen sets of vertices and edges, keeping the randomness and the independence of the community detection problem. In Chapter 4 we have proved that, taken as input the random graph $G(n, \frac{a}{n}, \frac{b}{n})$ with $k > 2$ communities, k-Partition produces with high probability a γ-correct partition with a small error rate $\gamma > 0$ when

$$\frac{(a-b)^2}{a} \geq C\, k^2\, \log \frac{1}{\gamma}$$

for a suitable constant $C > 0$.

In the development of the method, we have always assumed that the communities, identified as $V_1, \ldots, V_k$, had all the same size. However, the algorithm also works without significant changes when the blocks are not equal, but they must have comparable sizes, let us say $n \leq |V_i| \leq cn$ for some $c \geq 1$. In this case, the relation among a, b and γ also depends on c.
Moreover, the results that we have proved do not necessarily require a and b constant. This means that the algorithm works on denser graphs too.

On the other hand, the SBM is the simplest model of random graphs with community structure. Even if it provides a fertile ground for studying various central questions in machine learning, computer science and statistics, it cannot be extended to many complex models (see for instance [4] or [8]). Another important limit is the fact that spectral methods need to know the number of communities within the graph as a starting information.

The field of community detection is characterized by the presence of really different approaches to the problem. Indeed in literature we can find techniques that consider many points of view. This is certainly a consequence of the fact that we can employ this topic to various scopes. Nowadays, the rapid growth of the Internet, the Web and online interactions lead to new situations and network structures, dealing with a huge quantity of information, that require attention. But there are lots of other open questions dealing with particular graph models (e.g. graphs with labels, with overlaps, dynamic graphs...), accuracy of the recovery or the definition of community itself. The remarkable progress of the last decades suggests that the problem is indeed important and rich, and mathematicians are thus treating this topic with interest.

Appendix A

Appendix

We collect here all the theoretical results and properties that we use in the developement of the thesis.

A.1 Concentration inequalities

We gather some fundamental inequalities describing bounds for the tails of random variables or concentration properties of sums of iid random variables. Notice that, here and in the applications, when the bound refers to the absolute value of a certain object, we add a coefficient 2 ahead.

A.1.1 Markov inequality

Markov. *For any non-negative random variable X and $t > 0$, we have*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Proof. Fixed $t > 0$, the result follows from the fact that for any real number we can apply the identity

$$x = x\,\mathbb{I}_{\{x \geq t\}} + x\,\mathbb{I}_{\{x \leq t\}}.$$

Then, taking the random variable X and applying the expected value both sides we get

$$\mathbb{E}[X] = \mathbb{E}[X\,\mathbb{I}_{\{X \geq t\}}] + \mathbb{E}[X\,\mathbb{I}_{\{X \leq t\}}] \geq \mathbb{E}[t\,\mathbb{I}_{\{X \geq t\}}] + 0 = t\,\mathbb{P}(X \geq t). \tag{A.1}$$

Dividing by t, we find the desired relation.

□

Specifically, in the development of the thesis we need the following version:

Corollary 6. *For any random variable X and $t > 0$, we have for $n > 0$*

$$\mathbb{P}(|X| \geq t) \leq \frac{\mathbb{E}[\,|X|^n\,]}{t^n}.$$

Proof. This variant follows from a quick consideration on the original: we raise to the nth power the event for which we measure the probability and then we apply Markov's inequality:

$$\mathbb{P}\left(|X| \geq t\right) = \mathbb{P}\left(|X|^n \geq t^n\right) \leq \frac{\mathbb{E}[\,|X|^n\,]}{t^n}.$$

□

A.1.2 Chernoff inequality

Chernoff 1. *Let X_i be independent Bernoulli random variables with parameters p_i. Consider their sum $S_N := \sum_{i=1}^{N} X_i$ and denote $\mu = \mathbb{E}[S_N]$. Then, for any $t > \mu$, we have*

$$\mathbb{P}\left(S_N \geq t\right) \leq e^{-\mu}\left(\frac{e\mu}{t}\right)^t.$$

Proof. We are going to bound the probability multiplying both sides of the inequality by a parameter $\lambda > 0$, exponentiating and using Markov inequality as in A.1.1:

$$\mathbb{P}\left(S_N \geq t\right) = \mathbb{P}\left(\exp\left(\lambda \sum_{i=1}^{N} X_i\right) \geq \exp\left(\lambda t\right)\right)$$

$$\leq e^{-\lambda t}\,\mathbb{E}\left[\exp\left(\lambda \sum_{i=1}^{N} X_i\right)\right]$$

$$= e^{-\lambda t} \prod_{i=1}^{N} \mathbb{E}\left[\exp\left(\lambda X_i\right)\right] \tag{A.2}$$

where the last equality in (A.2) holds for the independence of the X_i.
Then, it remains to bound the *moment generating function* $\mathbb{E}\left[\exp\left(\lambda X_i\right)\right]$ for every Bernoulli random variable X_i. Observing that $1 + x \leq e^x$, we have

$$\mathbb{E}\left[\exp\left(\lambda X_i\right)\right] = e^{\lambda} p_i + (1 - p_i) = 1 + p_i(e^{\lambda} - 1) \leq \exp\left((e^{\lambda} - 1)p_i\right). \tag{A.3}$$

Then, (A.2) becomes

$$\mathbb{P}\left(S_N \geq t\right) \leq e^{-\lambda t} \exp\left((e^{\lambda} - 1)\sum_{i=1}^{N} p_i\right) = e^{-\lambda t} \exp\left((e^{\lambda} - 1)\mu\right). \tag{A.4}$$

Since (A.4) holds for every $\lambda > 0$, we substitute $\lambda := \log\left(\frac{t}{\mu}\right)$, which is positive for the hypothesis $t > \mu$. In this way, we conclude that

$$\mathbb{P}\left(S_N \geq t\right) \leq \left(\frac{\mu}{t}\right)^t \exp\left(t - \mu\right) = e^{-\mu}\left(\frac{e\mu}{t}\right)^t.$$

□

Chernoff 2. *Let X_i be independent Bernoulli random variables with parameters p_i. Consider their sum $S_N := \sum_{i=1}^{N} X_i$ and denote $\mu = \mathbb{E}[S_N]$. Then, for $\delta \in (0, 1]$, we have*

$$\mathbb{P}\left(|S_N - \mu| \geq \delta\mu\right) \leq 2\exp\left(-c\mu\delta^2\right),$$

where $c > 0$ is an absolute constant.

Proof. To prove this alternative form of Chernoff inequality, we will use version 1 and we will also show the bound for the lower tail.

For the upper tails, we take $t = (1 + \delta)\mu$:

$$\begin{aligned}\mathbb{P}\left(S_N \geq (1+\delta)\mu\right) &= \mathbb{P}\left(S_N - \mu \geq \delta\mu\right) \\ &\leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu} \\ &= \exp\left(\mu\left(\delta - (1+\delta)\log(1+\delta)\right)\right). \qquad \text{(A.5)}\end{aligned}$$

Using the fact that

$$\log(1+x) \geq \frac{x}{1 + x/2}$$

for every $x > 0$, (A.5) becomes

$$\mathbb{P}\left(S_N \geq (1+\delta)\mu\right) \leq e^{-\frac{\delta^2}{2+\delta}\mu}. \qquad \text{(A.6)}$$

Similarly, take $t = (1 - \delta)\mu$ and see that

$$\begin{aligned}\mathbb{P}\left(S_N \geq (1-\delta)\mu\right) &= \mathbb{P}\left(\mu - S_N \leq \delta\mu\right) \\ &\leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu} \\ &= \exp\left(\mu\left(-\delta - (1-\delta)\log(1-\delta)\right)\right). \qquad \text{(A.7)}\end{aligned}$$

Noticing that for every $x > 0$

$$\log(1-x) \geq -\delta + \frac{\delta^2}{2},$$

(A.7) can be rewritten as

$$\mathbb{P}\left(S_N \geq (1-\delta)\mu\right) \leq e^{-\frac{\delta^2}{2}\mu}. \qquad \text{(A.8)}$$

Combining (A.6) and (A.8) we have the result.

□

Chernoff 3. *Let X_i be independent indicator random variables with mean at most $\rho \leq \frac{1}{2}$. Consider their sum $X := \sum_{i=1}^{n} X_i$. Then, then for any $t > 0$*

$$\begin{aligned}\max\left\{\mathbb{P}(X \geq \mathbb{E}[X] + t), \mathbb{P}(X \leq \mathbb{E}[X] - t)\right\} &\leq \exp\left(-\frac{t^2}{2Var(X) + t}\right) \\ &\leq \exp\left(-\frac{t^2}{2n\rho + t}\right).\end{aligned}$$

A.1.3 Bernstein inequality

We only consider the Bernstein inequality for bounded distributions.

Bernstein. *Let $X_1, \ldots, X_N$ be independent mean-zero random variables such that $|X_i| \leq K$ for all i. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^{N} X_i\right| \geq t\right\} \leq 2\exp\left(-\frac{t^2/2}{\sigma^2 + Kt/3}\right).$$

Here $\sigma^2 = \sum_{i=1}^{N} \mathbb{E}[{X_i}^2]$ is the variance of the sum.

A.2 Binomial coefficient

We make a brief reference to the binomial coefficient just to point out a particular property:

Binomial coefficient bound. *For any $k < n$, it holds*

$$\binom{n}{k} \leq \left(\frac{ne}{k}\right)^k.$$

Proof. Starting from the definition of the binomial coefficient, we have

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\ldots(n-(k-1))}{k!} \leq \frac{n^k}{k!}.$$

Recall that $e^k = \sum_{m=0}^{\infty} \frac{k^m}{m!}$. In particular,

$$e^k > \frac{k^k}{k!} \quad \Rightarrow \quad \frac{1}{k!} < \left(\frac{e}{k}\right)^k.$$

Thus,

$$\binom{n}{k} \leq \frac{n^k}{k!} < \left(\frac{en}{k}\right)^k.$$

□

A.3 Properties on matrices

A.3.1 Eigenvalues and singular values

Firstly, we briefly recall the definitions of eigenvalues and eigenvectors of a (square) matrix A:

Definition 9. *Let V be a vector space with $\dim(V) = n$. Consider en endomorphism ϕ of V and let $A \in M_{n\times n}$ be the related matrix. The **eigenvalues** of A are the roots of the characteristic polynomial*

$$p_A(x) := \det(x\mathbb{I}_n - A) = (-1)^n \det(A - x\mathbb{I}_n).$$

*For any eigenvalue λ_i of A, we define the related **eigenvectors** as the vectors v_{i_j} generating the vector space*

$$\ker(\phi - \lambda_i id_V) = \{v \in V \,|\, \phi(v) = \lambda_i v\}.$$

In the case of a rectangular matrix A, let us say of dimensions $m \times n$, we can no more find its eigenvalues and eigenvectors. Indeed in this case we can describe the matrix with its singular values:

Definition 10. *Given an $m \times n$ matrix A with rank$(A) = k$, the **singular values** $s_1, \ldots, s_k$ of A are the square roots of the eigenvalues λ_i of both AA^T and A^TA:*

$$s_i := \sqrt{\lambda_i(AA^T)} = \sqrt{\lambda_i(A^TA)}. \tag{A.9}$$

For convenience, we extend the sequence by setting $s_i = 0$ for $k < i \leq n$ and we arrange them in a non-increasing order

$$s_1 \geq s_2 \geq \cdots \geq s_n \geq 0.$$

In the particular case A symmetric, the singular values are

$$s_i := |\lambda_i|. \tag{A.10}$$

Moreover, we define $u_1, \ldots, u_k$ the **left singular vectors** of A as the orthonormal eigenvectors of AA^T. Similarly, the **right singular vectors** $v_1, \ldots, v_k$ are the orthonormal eigenvectors of A^TA.

A.3.2 Matrix norm

An $m \times n$ matrix A can be seen as a linear operator from $\mathbb{R}^n$ to $\mathbb{R}^m$.
So, we can define its operator norm as

$$\|A\| := \max_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Ax\|_2}{\|x\|_2} = \max_{x \in S^{n-1}} \|Ax\|_2 \tag{A.11}$$

where $\|\cdot\|_2$ is the Euclidean norm and S^{n-1} is the unit sphere in $\mathbb{R}^n$.
Equivalently, the operator norm of A can be computed by maximizing the quadratic form $\langle Ax, y \rangle$ over all unit vectors x, y:

$$\|A\| := \max_{x \in S^{n-1}, y \in S^{m-1}} \langle Ax, y \rangle. \tag{A.12}$$

In terms of its spectrum, the operator norm of A equals the largest singular value (see Definition 10) of A:

$$\|A\| := s_1(A). \tag{A.13}$$

Definition 11. *The **Hilbert-Schmidt norm**, also called Frobenius norm, of a matrix A with entries A_{ij} is defined as*

$$\|A\|_{HS} := \left(\sum_{1=1}^{m} \sum_{j=1}^{n} |A_{ij}|^2 \right)^{1/2}. \tag{A.14}$$

Thus, the Hilbert-Schmidt norm is the euclidean norm on the space of matrices $\mathbb{R}^{m \times n}$. Its relation with the operator norm is given by

$$\|A\| \leq \|A\|_{HS}. \tag{A.15}$$

A.4 ε-nets

Definition 12. *Let (T, d) be a metric space. Consider a subset $K \subset T$ and let $\varepsilon > 0$. A subset $\mathcal{N} \subseteq K$ is called an ε-**net** of K if every point in K is within a distance ε of some point of $\mathcal{N}$, i.e.*

$$\forall x \in K,\ \exists x_0 \in \mathcal{N}\ :\ d(x, x_0) \leq \varepsilon. \tag{A.16}$$

Equivalently, $\mathcal{N}$ is an ε-net of K if and only if K can be recovered by balls with centers in $\mathcal{N}$ and radii ε.

Definition 13. *The smallest possible cardinality of an ε-net of K is called **covering number** of K and is denoted $\mathcal{N}(K, d, \varepsilon)$. Equivalently, $\mathcal{N}(K, d, \varepsilon)$ is the smallest number of closed balls with centers in K and radii ε whose union covers K.*

Definition 14. *A subset $\mathcal{N}$ of a metric space (T, d, ε) is ε-separated if $d(x, y) > \varepsilon$ for all distinct points $x, y \in \mathcal{N}$. The largest possible cardinality of an ε-separated subset of a given set $K \subset T$ is called the **packing number** of K and is denoted $\mathcal{P}(K, d, \varepsilon)$.*

Let us now focus on the case $T = \mathbb{R}^n$. We take the usual Euclidean metric

$$d(x, y) = \|x - y\|_2$$

for any $x, y \in \mathbb{R}^n$ and we ease the notation writing $\mathcal{N}(K, \varepsilon) = \mathcal{N}(K, d, \varepsilon)$ and $\mathcal{P}(K, \varepsilon) = \mathcal{P}(K, d, \varepsilon)$. Let $|\cdot|$ denote the volume in $\mathbb{R}^n$ and B_2^n be the unit Euclidean ball in $\mathbb{R}^3$. Then, it holds that

Proposition 2. *Let K be a subset of $\mathbb{R}^n$ and $\varepsilon > 0$. Then*

$$\frac{|K|}{|\varepsilon B_2^n|} \leq \mathcal{N}(K, \varepsilon) \leq \mathcal{P}(K, \varepsilon) \leq \frac{|K + (\varepsilon/2)B_2^n|}{|(\varepsilon/2)B_2^n|}.$$

Proof. The middle inequality is true for any metric space (see [16]).

Let us prove the lower bound: K can be covered by $N := \mathcal{N}(K, \varepsilon)$ balls with radii ε. Comparing the volumes we get

$$|K| \leq N|\varepsilon B_2^n|$$

so we just need to divide both sides by $|\varepsilon B_2^n|$ to obtain the relation.

For the upper bound, set $N := \mathcal{P}(K, \varepsilon)$. Then one can construct N closed disjoint balls $B(x_i, \varepsilon/2)$ with centers $x_i \in K$ and radii $\varepsilon/2$. While these balls may not need to fit entirely into K, they do fit into a slightly inflated set, namely $K + (\varepsilon/2)B_2^n$ (see [16]). Comparing the volumes we get

$$N|(\varepsilon/2)B_2^n| \leq |K + (\varepsilon/2)B_2^n|$$

which leads to the upper bound in the proposition.

□

For our purposes, we need in particular the following

Corollary 7. *The covering numbers of the unit Euclidean ball B_2^n satisfy the following for any $\varepsilon > 0$:*

$$\left(\frac{1}{\varepsilon}\right)^n \leq \mathcal{N}(B_2^n, \varepsilon) \leq \left(\frac{2}{\varepsilon} + 1\right)^n.$$

The same upper bound is true for the unit Euclidean ball $\mathcal{S}^{n-1}$.

Proof. The proof follows from Proposition 2. For the lower bound it suffices to observe that the volume in $\mathbb{R}^n$ scales like

$$|\varepsilon B_2^n| = \varepsilon^n |B_2^n|.$$

For the upper bound we see that

$$\mathcal{N}(B_2^n, \varepsilon) \leq \frac{|(1+\varepsilon/2)B_2^n|}{|(\varepsilon/2)B_2^n|} = \frac{(1+\varepsilon/2)^n}{(\varepsilon/2)^n} = \left(\frac{2}{\varepsilon} + 1\right)^n.$$

The upper bound for the sphere can be proved in the same way.

□

A.5 Orders of approximation

We recall here the notation identifying some orders of approximation:

$f(n) = \mathcal{O}(g(n))$	$\exists K > 0$ and $\exists n_0$ s.t. $\lvert f(n)\rvert \leq K g(n) \quad \forall n \geq n_0$	$\limsup_{n\to\infty} \frac{\lvert f(n)\rvert}{g(n)} < \infty$
$f(n) = o(g(n))$	$\forall \varepsilon > 0, \quad \exists n_0$ s.t. $\lvert f(n)\rvert \leq \varepsilon\, g(n) \quad \forall n \geq n_0$	$\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$
$f(n) = \Omega(g(n))$	$\exists K > 0$ and $\exists n_0$ s.t. $f(n) \geq K\, g(n) \quad \forall n \geq n_0$	$\liminf_{n\to\infty} \frac{f(n)}{g(n)} > 0$
$f(n) = \Theta(g(n))$	$\exists K_1, K_2 > 0$ and $\exists n_0$ s.t. $K_1\, g(n) \leq f(n) \leq K_2\, g(n)$ $\forall n \geq n_0$	$f(n) = \mathcal{O}(g(n))$ and $f(n) = \Omega(g(n))$
$f(n) \sim g(n)$	$\forall \varepsilon > 0, \quad \exists n_0$ s.t. $\left\lvert \frac{f(n)}{g(n)} - 1\right\rvert \leq \varepsilon \quad \forall n \geq n_0$	$\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1$

Bibliography

[1] E. Abbe, *Community detection and stochastic block model*, `http://www.princeton.edu/~eabbe/publications/abbe_FNT_2.pdf`, 2017.

[2] R. Bhatia, *Matrix Analysis*, volume 169, Springer, 1997, ISBN 0387948465.

[3] T. N. Bui, S. Chaudhuri, F. T. Leighton, M. Sipser, *Graph bisection algorithms with good average case behaviour*, Combinatorica, 7(2) : 171 − 191, ISSN 0209 − 9683, 1987.

[4] G. Caldarelli, *Scale-free networks: complex webs in nature and technology*, Oxford University Press, 2007.

[5] P. Chin, A. Rao, V. Vu, *Stochastic Block Model and Community Detection in Sparse Graphs: a spectral algorithm with optimal rate of recovery*, arXiv:1501.05021, 2015.

[6] A. Coja-Oghlan, *Graph partitioning via adaptive spectral techniques*, Combinatorics, Probability and Computing, 19 : 227 − 284, ISSN 1469 − 2463, 2010.

[7] C. Davis, *The rotation of eigenvectors by a perturbation*, Journal of Mathematical Analysis and Applications, 6(2) : 159−173, 1963. ISSN 0022−247X.

[8] S. Dorogovtsev, *Lectures on complex networks*, Oxford University Press, 2010.

[9] U. Feige, E. Ofek, *Spectral techniques applied to sparse random graphs*, Random Structures and Algorithms, 27(2) : 251 − 275, 2005. ISSN 1098 − 2418.

[10] S. Fortunato, *Community detection in graphs*, Physics Report, Volume 486, Issues 3 − 5, February 2010, Pages 75 − 174.

[11] S. Fortunato, D. Hrich, *Community detection in complex networks: a user guide*, Physics Report, Volume 659, 11 November 2016, Pages 1 − 44.

[12] F. McSherry, *Spectral partitioning of random graphs*, Foundations of Computer Science, Proceedings, 42nd IEEE Symposium on, 2001.

[13] S. M. Prabhu et al., *Containing COVID-19 Pandemic using Community Detection*, Journal of Physics, Conf. Ser., 1797 012008, 2021.

[14] J. Reichardt, S. Bornholdt, *Statistical mechanics in community detection*, 2006, Phys. Rev. E74 (1), 016110.

[15] M. Rosvall, C. T. Bergstrom, *Maps of random walks on complex networks reveal community structure*, PNAS January 29, 2008, 105(4)1118 – 1123.

[16] R. Vershynin, *High-Dimensional Probability. An Introduction with Applications in Data Science*, Cambridge University Press, 2018.

[17] V. Vu, *A simple SVD algorithm for finding hidden partitions*, arXiv:1404.3918, 2014.

[18] Y. Zhang, H. Zhou, *Minimax rates of community detection in stochastic block model*, Ann. Statist. 44(5) : 2252 – 2280, 2016.