Spectral methods and SDP relaxations for community detection in graphs

Guillaume Lecué*

Abstract

In this chapter, we present two types of procedures for the community detection problem in a graph: a spectral method and SDP relaxations. Both arise from a convex relaxation of an NP-hard problem. The spectral method is obtained by convexifying the constraint set. The SDP procedure implements the techniques of *matrix lifting* and *SDP relaxation*. In both cases, the Laplacian of the graph plays a central role, since its spectrum provides information on the connectivity properties of the graph. The first sections are devoted to its study in the ideal case of disjoint communities; one then speaks of connected components. From a statistical viewpoint, we study the estimation and exact recovery properties of two SDP relaxations. As in the previous chapters, this study highlights the role of randomness in this approach, and we introduce on this occasion a random graph model called the *stochastic block model*. From a technical viewpoint, Grothendieck's inequality and the construction of a dual certificate are used to carry out the statistical analysis of the SDP relaxations.

1 Introduction

An example of a community detection problem: We are given a graph in which each node represents a person and an edge of the graph between two nodes represents a connection between these two people, established, for example, from a social network of message exchanges. The goal is to identify, in this graph, groups of nodes that are particularly well connected. These groups may represent groups of friends, or people who share common interests and therefore tend to exchange messages frequently. We are thus given a graph, and the community detection problem is to find groups of nodes that are more densely connected to one another than to the rest of the graph. Once the communities have been detected, one may look for the influencers within them and send them advertisements, or approach them directly.

Definition 1.1. *A* ***graph*** *is a pair* (V, E) *where* V *is a set whose elements are called the* ***nodes of the graph*** *and* $E \subset V \times V$ *is a set whose elements are called the* ***edges of the graph****. A* ***weighted graph*** *is a triple* (V, E, W) *such that* (V, E) *is a graph and* $W \in \mathbb{R}^{|V|\times|V|}$*. The entries of* W *are called the* ***weights of the graph****. A graph* (V, E) *is said to be* ***undirected*** *when* $(i, j) \in E$ *implies* $(j, i) \in E$*, and a weighted graph* (V, E, W) *is said to be* ***undirected*** *when* (V, E) *is an undirected graph and* $W = W^\top$*.*

There are several types of graphs, according to whether they are directed (or not) and weighted (or not), as represented in Figure 3. We also give a few classical examples of graphs in Figures 1 and 2.

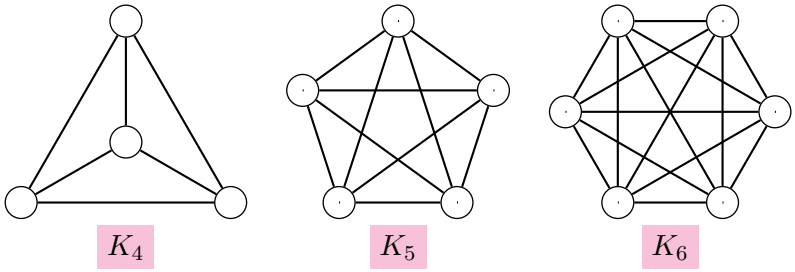


Figure 1: Three examples of complete graphs

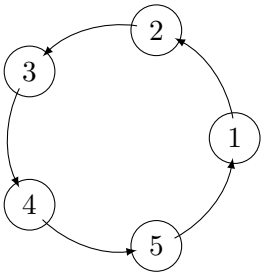


Figure 2: Example of a ring graph

The connection properties of a graph are often represented by means of a matrix.

Definition 1.2. *Let $G = (V, E)$ be a graph. The **adjacency matrix** of G is given by $A \in \{0,1\}^{|V|\times|V|}$ where, for every $(i,j) \in \{1,\dots,|V|\}^2$, one has $A_{ij} = 1$ when $(i,j) \in E$ and $A_{ij} = 0$ when $(i,j) \notin E$. The adjacency matrix of a weighted graph (V, E, W) is $A \in \mathbb{R}^{|V|\times|V|}$, where $A_{ij} = W_{ij}$ when $(i,j) \in E$ and $A_{ij} = 0$ when $(i,j) \notin E$.*

If a graph is undirected, then its adjacency matrix is symmetric. For the community detection problem, one is interested in the connectivity properties of a graph. The first definition concerns connectivity properties within a graph.

Definition 1.3. *A graph $G = (V, E)$ is said to be **connected** when one can find a path of edges linking all the nodes of the graph. A **connected component** of G is a connected subgraph of G with no edge linking it to its complement. A **connected partition** of G is a finite collection of subgraphs of G forming a partition of the nodes of G such that the subgraphs associated with this partition are connected components.*

This definition can also be extended to weighted graphs. In that case, one recovers the previous definition for an unweighted graph by adjoining to it the constant weight matrix $W = (1)_{(i,j)\in V\times V}$.

Definition 1.4. *A weighted graph $G = (V, E, W)$ is said to be **connected** when one can find a path of edges linking all the nodes of the graph and $W_{ij} \neq 0$ for every edge (i,j) of this path. A **connected component** of G is a connected subgraph of G with no edge of nonzero weight linking it to its complement. A **connected partition** of a graph G is a finite collection of subgraphs of G forming a partition of the nodes of G and such that these subgraphs are connected components.*

*ESSEC, Business school. Email: lecue@essec.edu.

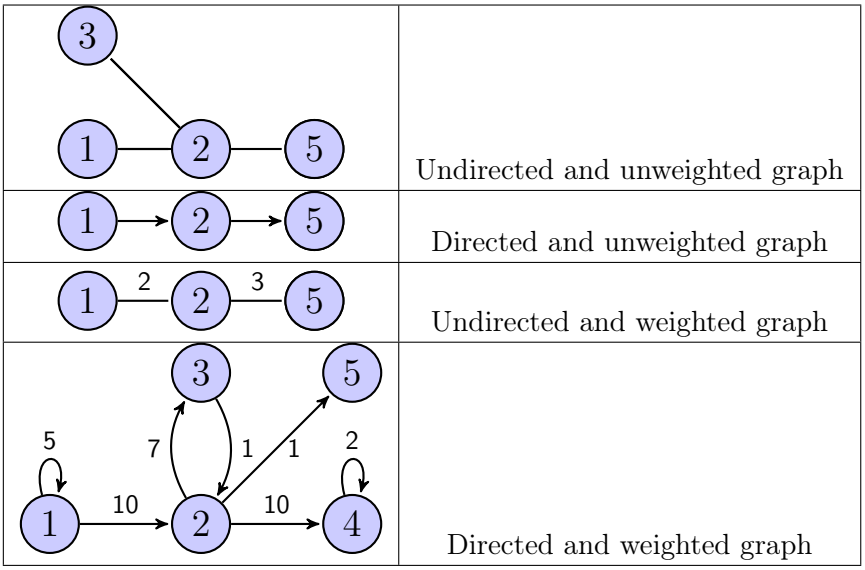


Figure 3: Examples of graphs, directed or not and weighted or not

A graph that can be decomposed into a nontrivial connected partition (that is, a partition not reduced to a single element) is, in a sense, an extremal case for the community detection problem, because in this case the communities do not interact. This is a rather rare situation in practice. Nevertheless, it can guide the design of community detection algorithms, because it is an ideal model for this problem.

2 The Laplacian of a graph and some properties of its spectrum

In this section, we study the link between the connectivity properties of a graph and the spectrum of its Laplacian. We first define the Laplacian of a graph.

Definition 2.1. *Let G be a graph (weighted or not). We denote by V the set of nodes of G and by A its adjacency matrix. The **Laplace matrix**, or **Laplacian**, of G is given by $L = D - A$, where $D = \mathrm{diag}(d)$, $d = (d_i)_{i \in V}$ and $d_i = \sum_{j \in V} A_{ij}$ for every $i \in V$ is the degree of the i-th node.*

There exist other definitions of the Laplacian of a graph. We shall use that of Definition 2.1 throughout the chapter. If G is an undirected graph, then its Laplacian is symmetric. One may therefore apply the spectral theorem to L in this case: the eigenvalues of L are all real, and L is diagonalizable in an orthonormal basis (that is, there exist an orthonormal matrix $P \in \mathcal{O}(|V|)$ and a diagonal matrix $D = \mathrm{diag}(\lambda)$, where $\lambda = (\lambda_i)_{i \in V}$ is the vector of eigenvalues of L, such that $L = PDP^\top$). We shall see below that the spectrum of L (that is, the set of its eigenvalues) plays a key role in the spectral method for community detection.

Proposition 2.2. *Let $G = (V, E, W)$ be an undirected weighted graph. Assume that $W_{ij} \geq 0$ for every $(i, j) \in V \times V$. Then:*

1) for every $f \in \mathbb{R}^{|V|}$, $f^\top L f = \frac{1}{2} \sum_{(i,j) \in E} W_{ij} (f_i - f_j)^2$

2) L is symmetric positive semidefinite

3) 0 is an eigenvalue of L and $L\mathbf{1} = 0$, where $\mathbf{1} = (1)_{i \in V}$.

Proof. **1)** We denote by $A = (W_{ij}\mathbf{1}_{(i,j)\in E})$ the adjacency matrix of G. Since G is undirected, we have $W_{ij} = W_{ji}$ for every $(i,j) \in V \times V$, and hence for every $f \in \mathbb{R}^{|V|}$,

$$\begin{aligned} f^\top L f = f^\top (D - A) f &= \sum_{i\in V} d_i f_i^2 - \sum_{(i,j)\in E} f_i W_{ij} f_j \\ &= \sum_{i\in V}\sum_{j\in V} W_{ij}\mathbf{1}_{(i,j)\in E} f_i^2 - \sum_{(i,j)\in E} f_i W_{ij} f_j = \frac{1}{2}\sum_{(i,j)\in E} W_{ij}(f_i^2 - 2f_i f_j + f_j^2) \\ &= \frac{1}{2}\sum_{(i,j)\in E} W_{ij}(f_i - f_j)^2. \end{aligned}$$

2) For every $f \in \mathbb{R}^{|V|}$,

$$\langle f, Lf\rangle = f^\top L f = \frac{1}{2}\sum_{(i,j)\in E} W_{ij}(f_i - f_j)^2 \geq 0.$$

We therefore have $L \succeq 0$.

3) By definition of the degree of a node, for every $i \in V$,

$$(L\mathbf{1})_i = \sum_{j\in V} L_{ij} = d_i - \sum_{j\in V} A_{ij} = 0.$$

Hence 0 is indeed an eigenvalue of L, and $\mathbf{1}$ is an eigenvector associated with the eigenvalue 0. ■

The following result is a first link between the spectrum of the Laplacian and the connectivity properties of the underlying graph. Indeed, the multiplicity of the eigenvalue 0 of L gives the number of connected components of G. Moreover, the indicators of the nodes of these connected components are eigenvectors spanning this eigenspace (that is, the kernel) of L.

Proposition 2.3. *Let $G = (V, E, W)$ be an undirected weighted graph with nonnegative weights. The multiplicity of the eigenvalue 0 of the Laplacian of G is equal to the number of connected components of G. Moreover, the eigenspace associated with the eigenvalue 0 of L, that is, its kernel $\mathrm{Ker}(L)$, is spanned by $\mathbf{1}_{V_1}, \ldots, \mathbf{1}_{V_k}$, where $V_1 \sqcup \cdots \sqcup V_k$ is the partition of V into the connected components of G.*

Proof. We begin with the case of a single connected component, that is, when G is connected. The kernel of L consists of all elements $f \in \mathbb{R}^{|V|}$ such that $Lf = 0$. If $Lf = 0$, then $f^\top L f = 0$, and hence, by Proposition 2.2, one has $\sum_{(i,j)\in E} W_{ij}(f_j - f_i)^2 = 0$. Moreover, the weights W_{ij} are nonnegative, so if $\sum_{(i,j)\in E} W_{ij}(f_j - f_i)^2 = 0$, then for every $(i,j) \in E$ one has $W_{ij}(f_j - f_i)^2 = 0$, and hence either $W_{ij} = 0$ or $f_j = f_i$. Furthermore, G is connected, so if one takes any pair of nodes $i, j \in V$, there exists a path represented by node indices $i_1, \ldots, i_p \in V$ such that $(i, i_1) \in E, (i_1, i_2) \in E, \ldots, (i_p, j) \in E$ and $W_{ii_1} > 0, W_{i_1 i_2} > 0, \cdots, W_{i_k j} > 0$. Along this path, one has either $W_{pq} = 0$ or $f_p = f_q$ for $(p,q) \in \{(i,i_1), (i_1,i_2), \cdots, (i_p, j)\}$; but since $W_{pq} \neq 0$, one has $f_p = f_q$. Since this holds for every edge of the path, one deduces that the endpoints f_i and f_j are equal. Hence f is a constant vector. We have therefore shown that $\mathrm{Ker}(L)$ is spanned by $\mathbf{1} = (1)_{i\in V}$, the indicator of the unique connected component of G.

We now consider the general case in which G has k connected components. Up to reordering the nodes of V, one may write the Laplacian of G in block-matrix form:

$$L = \left[\begin{array}{c|c|c|c} L_1 & 0 & \cdots & 0 \\ \hline 0 & L_2 & \cdots & 0 \\ \hline \cdots & \cdots & \ddots & 0 \\ \hline 0 & 0 & \cdots & L_k \end{array}\right]$$

where, for every $i = 1, \dots, k$, L_i is the Laplacian of the i-th connected component $G_i = (V_i, E_i, W^{(i)})$ of G, where $V_i \subset V$ is the set of nodes of the i-th connected component of G, $E_i = \{(p, g) : p, q \in V_i, (p, q) \in E\}$ is the set of its edges, and $W^{(i)} = (W_{pq} : p, q \in V_i)$ is its weight matrix.

We denote by $\mathbf{1}_{V_i} \in \mathbb{R}^{|V|}$ the indicator of the nodes of G_i. We wish to show that $\mathrm{Ker}(L)$ is spanned by $\mathbf{1}_{V_i}, i = 1, \cdots, k$. For every $i = 1, \dots, k$, L_i is the Laplacian of G_i, so 0 is an eigenvalue of L_i and $(1)_{p \in V_i}$ spans the kernel of L_i. Hence $\mathbf{1}_{V_i}$ lies in the kernel of L. We therefore have $\mathrm{vect}(\mathbf{1}_{V_i}, i = 1, \cdots, k) \subset \mathrm{Ker}(L)$. Moreover, given the block structure of L, one sees that $Lf = 0$ if and only if $L_i f = 0$ for every $i = 1, \dots, k$, and since $\ker(L_i) = \mathrm{vect}((\mathbf{1})_{\mathbf{p} \in \mathbf{V_i}})$, one has $f_{|V_i} = \alpha_i (\mathbf{1})_{\mathbf{p} \in \mathbf{V_i}}$ for some $\alpha_i \in \mathbb{R}$. One deduces that $f = \sum_i \alpha_i \mathbf{1}_{V_i} \in \mathrm{vect}(\mathbf{1}_{V_i}, i = 1, \cdots, k)$. ∎

Remark 2.4. *Proposition 2.3 also applies to unweighted undirected graphs. It suffices to apply it to the weight matrix* $W = (1)_{(i,j) \in V \times V}$.

Example: Let us compute the multiplicity of the eigenvalue 0 of the Laplacian of a complete graph, and check that it is indeed equal to 1, since this graph is connected. We denote by K_n the complete graph on n vertices. The Laplacian of K_n is given by

$$L = \begin{bmatrix} n-1 & -1 & \cdots & -1 \\ -1 & n-1 & \cdots & -1 \\ \cdots & \cdots & \ddots & -1 \\ -1 & -1 & \cdots & n-1 \end{bmatrix} = \mathrm{diag}(n, \cdots, n) - (1)_{n \times n}.$$

If $f \in \mathbb{R}^n$ is such that $Lf = 0$, then $nf = (\langle f, (1)_1^n \rangle)_{i=1,\dots,n} = \langle f, (1)_1^n \rangle (1)_1^n$. Hence $f \in \mathrm{vect}((1)_1^n)$ and 0 has multiplicity 1 for L. As a consequence, K_n has only one connected component (as expected).

The Laplacian of a graph and Newton's law of cooling. We have seen that the spectrum of the Laplacian $L = D - A$ of a graph plays an essential role in the connectivity properties of the graph. It also appears when one studies other properties of a graph. One may ask about the origin of its name, and in particular whether it is related to the Laplacian encountered in physics when one studies the heat equation. It turns out that there is indeed a link between the two notions, and that the Laplacian of a graph may be viewed as a discrete version of the Laplacian introduced in physics. The link between the two approaches is Newton's law of cooling, or of heat transfer, which states that *heat is transferred from one point to another in proportion to the temperature difference between the two points.*

To make this link formal, imagine that the nodes of the graph $G = (V, E)$ have temperatures given by the family $(T_i)_{i \in V}$, with T_i the temperature of node i at time t. We let the heat transfers on this graph evolve according to Newton's law: "the temperature gradient (that is, what is commonly called heat) between two connected points is proportional to the temperature difference between these two points": for every node $i \in V$, the temperature at this node evolves as follows

$$\frac{dT_i}{dt} = -\kappa \sum_{j \in V} A_{ij}(T_i - T_j). \tag{1}$$

In other words, node i exchanges heat only with its neighbors (that is, the j such that $A_{ij} = 1$); it receives heat from the neighboring node j only if $T_j > T_i$, and it gives heat to the neighboring node j only if $T_j < T_i$. The parameter $\kappa > 0$ is a parameter controlling the speed of this transfer; it is a thermal coefficient.

One may then look at the evolution of the vector of temperatures of the nodes of the graph, $T = (T_i)_{i\in V}$, and expand this identity so as to bring out the Laplacian:

$$\frac{dT}{dt} = -\kappa \left(\sum_{j\in V} A_{ij}(T_i - T_j) \right)_{i\in V} = -\kappa \left(T_i \sum_{j\in V} A_{ij} - \sum_{j\in J} A_{ij} T_j \right)_{i\in V} = -\kappa(D - A)T.$$

One thus obtains a differential equation for the evolution of the temperatures of the nodes of the graph, of the form $dT/dt + \kappa LT = 0$, where $L = D - A$ is the Laplacian of the graph. Recalling the heat equation $\partial u/\partial t = \alpha \Delta u$, where $u : (t, x) \in \mathbb{R} \times \mathbb{R}^d \to u(t, x)$ is the temperature at time t at location x and Δ is the Laplacian $\partial_1^2 + \cdots + \partial_d^2$, one may identify L with $-\Delta$. It is from this analogy that L takes its name.

3 Principle of spectral methods in community detection

We give here two ideas that introduce, more or less formally, the spectral method for finding communities in graphs. We assume that we are given an undirected weighted graph $G = (V, E, W)$.

3.1 Ideal setting of connected components

Ideal setting: When it comes to detecting communities in a graph, the ideal setting occurs, in a sense, when these communities are not connected to one another, that is, when the graph admits a partition into connected components and each connected component constitutes a community. This is an ideal setting because it almost never occurs in practice, since there are always a few inter-community links.

Even though the ideal setting almost never occurs in practice, it is a good guide for understanding how to solve the community detection problem in more general situations.

We therefore place ourselves in the ideal setting (as a first approximation). The Laplacian of $G = (V, E, W)$ and its kernel then have particular forms. Indeed, by Proposition 2.3, the kernel of L is spanned by $\mathbf{1}_{V_1}, \ldots, \mathbf{1}_{V_k}$, where $V_1 \sqcup \cdots \sqcup V_k$ is the partition of the nodes of V into the k connected components of G.

It is, however, not easy to find the indicators $\mathbf{1}_{V_i}, i = 1, \ldots, k$ from L. What is easier is to find an orthonormal basis of $\mathrm{Ker}(L)$: $u_1, \ldots, u_k \in \mathbb{R}^{|V|}$ (one may in fact take $u_1 = (1)_{i=1}^n$). The **spectral method** (see the pseudo-code in Section 1) then proceeds in two steps:

1) one clusters the $|V|$ row vectors $(y_i)_{i\in V}$ of the matrix $[u_1|u_2|\cdots|u_k]$ of size $|V| \times k$ into k clusters $C_1, \cdots, C_k$

2) one returns the partition $V_1 \sqcup \cdots \sqcup V_k$ of the nodes of V, where $V_p = \{i \in V : y_i \in C_p\}$ for $p = 1, \ldots, k$. These are the communities of nodes detected by the spectral method.

One expects the partition $V_1 \sqcup \cdots \sqcup V_k$ of the nodes of V to give the k connected components of G. Indeed, $\{u_1, \ldots, u_k\}$ and $\{\mathbf{1}_{V_1}/\sqrt{|V_1|}, \ldots, \mathbf{1}_{V_k}/\sqrt{|V_k|}\}$ are two orthonormal bases of $\mathrm{Ker}(L)$. There then exists a rotation matrix (orthogonal matrix) $R \in \mathcal{O}(|V|)$ such that $u_i = R\mathbf{1}_{V_i}/\sqrt{|V_i|}, i = 1, \ldots, k$. If one had performed this clustering of the rows of the matrix $[\mathbf{1}_{V_1}/\sqrt{|V_1|}|\cdots|\mathbf{1}_{V_k}/\sqrt{|V_k|}]$ into k clusters, the clusters obtained would consist of rows that are

all equal to one another (those having a $1/\sqrt{|V_i|}$ in the same position and 0 elsewhere), because there are exactly k of them and one performs a clustering into k classes. One would therefore recover exactly the partition of G into its connected components. Moreover, one also expects the rotation R to preserve this clustering of the rows: one has

$$\begin{bmatrix} y_1^\top \\ y_2^\top \\ \vdots \\ y_{|V|}^\top \end{bmatrix} = [u_1|u_2|\cdots|u_k] = R[\mathbf{1}_{V_1}/\sqrt{|V_1|}|\cdots|\mathbf{1}_{V_k}/\sqrt{|V_k|}]$$

and hence clustering the row vectors $(y_i)_{i\in V}$ of the matrix $[u_1|u_2|\cdots|u_k]$ should also recover the connected components of G, just as this clustering does on the rows of $[\mathbf{1}_{V_1}/\sqrt{|V_1|}|\cdots|\mathbf{1}_{V_k}/\sqrt{|V_k|}]$.

This is the intuition behind a spectral method commonly used in community detection. In practice, G will not be partitionable into connected components; thus, one will not look at the kernel of L, but rather at its eigenspace associated with its k smallest eigenvalues. One will then look for a basis of k eigenvectors $u_1, \ldots, u_k$ of this eigenspace. It is then the $|V|$ rows of the matrix $[u_1|\cdots|u_k]$ that one clusters into k groups. From the row indices of these k groups one extracts a clustering of the nodes of the graph, that is, one obtains an estimator of the communities. The choice of the parameter k is made either by cross-validation, if one is given in advance a criterion to minimize—in general, the criteria in community detection are called **modularity functions**—or by plotting the spectrum of L and identifying an "elbow" among its smallest eigenvalues. Both methods are heuristic.

3.2 The "graph cut" viewpoint

In this section we give another viewpoint on the spectral methods used in community detection. The principle used here is that detecting communities is in fact equivalent to finding a partition of the nodes that minimizes the sum of the weights carried by the inter-community edges, while maximizing the size of these communities.

To formalize this approach, we introduce certain functions, called **modularity functions**, which quantify the notions of intra- and inter-community weight mass that one seeks to optimize over all the "**cuts**", that is, partitions, of the graph.

Mincut problem: Given two sets $A, B \subset V$ of nodes, one defines the total mass between A and B by

$$W(A,B) = \sum_{\substack{(i,j)\in E \\ (i,j),(j,i)\in A\times B}} W_{ij} = \sum_{\substack{(i,j)\in E \\ (i,j)\in A\times B}} W_{ij} + W_{ji} = 2 \sum_{\substack{(i,j)\in E \\ (i,j)\in A\times B}} W_{ij}$$

where the symmetry of W was used in the last equality. One may thus define a function associating with each partition $V_1 \sqcup \cdots \sqcup V_k$ of the nodes of the graph the total mass of the intra-community weights:

$$(V_1, \ldots, V_k) \in \mathcal{P}_k(V) \longrightarrow \mathrm{cut}(V_1, \ldots, V_k) = \frac{1}{2}\sum_{i=1}^{k} W(V_i, V_i^c).$$

where $\mathcal{P}_k(V)$ is the set of partitions of V having at most k elements and V_i^c is the complement of V_i in V.

For the two-community problem, one may rewrite the problem in the following form. Each partition of V into two communities $V_1 \sqcup V_2$ (here $V_2 = V_1^c$) is described by a membership vector

$x \in \{0,1\}^{|V|}$ such that $x_i = 1$ if $i \in V_1$ and $x_i = 0$ if $i \in V_2$. One then has

$$\mathrm{cut}(V_1, V_2) = W(V_1, V_2) = \sum_{\substack{i:x_i=1 \\ j:x_j=0 \\ (i,j)\in E}} W_{ij} + W_{ji} = 2 \sum_{(i,j)\in E} x_i W_{ij}(1 - x_j).$$

The two-class mincut problem may therefore be stated as follows: find $x^* \in \{0,1\}^{|V|}$ solving the problem

$$x^* \in \operatorname*{argmin}_{x\in\{0,1\}^{|V|}} \sum_{(i,j)\in E} x_i W_{ij}(1 - x_j).$$

This two-class min-cut problem can be solved efficiently by the Stoer–Wagner algorithm (see Appendix 9). In practice, however, the algorithm often returns a partition in which one class is reduced to a single node (the point of smallest degree, that is, the one least connected to the graph). This is not really the idea one has of a community. One will therefore force the communities to be of sufficient size by introducing a new modularity function.

Ratiocut problem: To avoid the trivial solutions of the mincut problem, one introduces a new modularity function:

$$(V_1, \ldots, V_k) \in \mathcal{P}_k(V) \longrightarrow \mathrm{RatioCut}(V_1, \ldots, V_k) = \frac{1}{2} \sum_{i=1}^{k} \frac{W(V_i, V_i^c)}{|V_i|} \tag{2}$$

which forces the size of the communities not to be too small when one minimizes this ratiocut. The ratio $W(V_i, V_i^c)/|V_i|$ is meant to strike a balance between the connectivity of V_i with its complement and the size of V_i. The RatioCut problem is therefore the problem of minimizing over $\mathcal{P}_k(V)$ the function $\mathrm{RatioCut}(V_1, \ldots, V_k)$.

In general, minimizing the ratiocut (2) over all partitions in $\mathcal{P}_k(V)$ is NP-hard. One will then rewrite this problem in vector form (for $k = 2$) or matrix form (general k) and show that a convex relaxation of this problem yields the spectral method for community detection seen in the previous section.

Ratiocut for $k = 2$: We first consider the ratiocut problem in the case of two communities. One then seeks a solution to the problem

$$\begin{aligned} \min_{V_1 \subset V} \mathrm{RatioCut}(V_1, V_1^c) &= \min_{V_1 \subset V} \left(\frac{W(V_1, V_1^c)}{|V_1|} + \frac{W(V_1^c, V_1)}{|V_1^c|} \right) \\ &= \min_{V_1 \subset V} \left(\frac{2}{|V_1|} \sum_{(i,j)\in E : i\in V_1, j\in V_1^c} W_{ij} + \frac{2}{|V_1^c|} \sum_{(i,j)\in E : i\in V_1^c, j\in V_1} W_{ij} \right). \end{aligned} \tag{3}$$

We rewrite this problem in vector form. For every $V_1 \subset V$, one constructs a vector $f^{V_1} = (f_i^{V_1})_{i\in V} \in \mathbb{R}^{|V|}$ defined for every $i \in V$ by

$$f_i^{V_1} = \begin{cases} \sqrt{\frac{|V_1^c|}{|V_1|}} & \text{if } i \in V_1 \\ -\sqrt{\frac{|V_1|}{|V_1^c|}} & \text{if } i \notin V_1. \end{cases} \tag{4}$$

Proposition 3.1. *Let L be the Laplacian of an undirected weighted graph. Then:*

1) $\left\langle f^{V_1}, \mathbf{1} \right\rangle = 0$ *and* $\left\| f^{V_1} \right\|_2^2 = n$

2) $(f^{V_1})^\top L f^{V_1} = n\mathrm{RatioCut}(V_1, {V_1}^c)$.

Proof. One has

$$
\begin{aligned}
(f^{V_1})^\top L f^{V_1} &= \frac{1}{2} \sum_{(i,j)\in E} W_{ij} (f_i^{V_1} - f_j^{V_1})^2 \\
&= \frac{1}{2} \sum_{\substack{(i,j)\in E \\ i\in V_1, j\in {V_1}^c}} W_{ij} \left(\sqrt{\frac{|{V_1}^c|}{|V_1|}} + \sqrt{\frac{|V_1|}{|{V_1}^c|}} \right)^2 + \frac{1}{2} \sum_{\substack{(i,j)\in E \\ j\in V_1, i\in {V_1}^c}} W_{ij} \left(-\sqrt{\frac{|V_1|}{|{V_1}^c|}} - \sqrt{\frac{|{V_1}^c|}{|V_1|}} \right)^2 \\
&= \frac{1}{2} \sum_{\substack{(i,j)\in E \\ i\in V_1, j\in {V_1}^c}} W_{ij} \left(\frac{n}{|V_1|} + \frac{n}{|{V_1}^c|} \right) + \frac{1}{2} \sum_{\substack{(i,j)\in E \\ j\in V_1, i\in {V_1}^c}} W_{ij} \left(\frac{n}{|{V_1}^c|} + \frac{n}{|V_1|} \right) \\
&= \frac{n}{2|V_1|} \sum_{\substack{(i,j)\in E \\ (i,j),(j,i)\in V_1\times {V_1}^c}} W_{ij} + \frac{n}{2|{V_1}^c|} \sum_{\substack{(i,j)\in E \\ (i,j),(j,i)\in V_1\times {V_1}^c}} W_{ij} = n\frac{W(V_1, {V_1}^c)}{|V_1|} + n\frac{W({V_1}^c, V_1)}{|{V_1}^c|} \\
&= n\mathrm{RatioCut}(V_1, {V_1}^c)
\end{aligned}
$$

where we have used the computation

$$
\begin{aligned}
&\left(\sqrt{\frac{|{V_1}^c|}{|V_1|}} + \sqrt{\frac{|V_1|}{|{V_1}^c|}} \right)^2 = \frac{|{V_1}^c|}{|V_1|} + \frac{|V_1|}{|{V_1}^c|} + 2\sqrt{\frac{|{V_1}^c||V_1|}{|V_1||{V_1}^c|}} \\
&= \frac{|{V_1}^c|}{|V_1|} + \frac{|V_1|}{|{V_1}^c|} + 2 = \frac{|{V_1}^c| + |V_1|}{|V_1|} + \frac{|V_1| + |{V_1}^c|}{|{V_1}^c|} = \frac{n}{|{V_1}^c|} + \frac{n}{|V_1|}.
\end{aligned}
$$

Moreover, one has

$$
\left\| f^{V_1} \right\|_2^2 = \sum_{i\in V_1} \frac{|{V_1}^c|}{|V_1|} + \sum_{i\in {V_1}^c} \frac{|V_1|}{|{V_1}^c|} = |V_1| + |{V_1}^c| = n
$$

and also $\left\langle f^{V_1}, \mathbf{1} \right\rangle = 0$, because

$$
\sum_{i\in V} f_i^{V_1} = \sum_{i\in V_1} \sqrt{\frac{|{V_1}^c|}{|V_1|}} - \sum_{i\in {V_1}^c} \sqrt{\frac{|V_1|}{|{V_1}^c|}} = \sqrt{|V_1||{V_1}^c|} - \sqrt{|V_1||{V_1}^c|} = 0.
$$

■

There is therefore an equivalence between the three problems:

$$
\min_{V_1\subset V} (f^{V_1})^\top L f^{V_1}, \tag{5}
$$

$$
\min_{V_1\subset V} \left((f^{V_1})^\top L f^{V_1} : \left\| f^{V_1} \right\|_2^2 = n, \left\langle f^{V_1}, \mathbf{1} \right\rangle = 0 \right) \tag{6}
$$

and the ratiocut problem

$$
\min_{V_1\subset V} \mathrm{RatioCut}(V_1, {V_1}^c). \tag{7}
$$

The three problems are discrete optimization problems. Since (7) is NP-hard in general, so are the other two, (5) and (6). We shall therefore look for a relaxation of (6).

(None convex) relaxation: The objective function of (5) or (6), namely $f \in \mathbb{R}^n \to f^\top L f$, raises no difficulty from the optimization viewpoint, because it is a function whose Hessian is $2L$, which is positive semidefinite, so the objective function is convex. We are therefore minimizing a convex function in the problems (5) and (6). The main computational difficulty of the problems (5) and (6) is that the search space "$A \subset V$" is discrete. We shall then "convexify" it simply by replacing it with $\mathbb{R}^{|V|}$. We then consider the "relaxed problem of (6)":

$$\min_{f \in \mathbb{R}^{|V|}} \left(f^\top L f : \|f\|_2^2 = 1, \langle f, \mathbf{1} \rangle = 0 \right). \tag{8}$$

Let us show that (8) is indeed the spectral method introduced in the previous section. Since L is symmetric, one can find an orthonormal basis of eigenvectors. Now we know that $\mathbf{1} = (1)_{i \in V}$ is an eigenvector associated with the eigenvalue 0 of L and that, since L is positive semidefinite (see Proposition 2.2), 0 is the smallest eigenvalue of L. Hence (8) amounts to looking for an eigenvector associated with the second smallest eigenvalue of L (which may also be 0 if 0 has multiplicity greater than 2 in the spectrum of L). Such an eigenvector appears often in the study of graphs. It has therefore been given a name.

Definition 3.2. *Let L be the Laplacian of an undirected weighted graph. The **Fiedler vector** of L is an eigenvector associated with the second smallest eigenvalue of L. The second smallest eigenvalue of L is also called the **algebraic connectivity**.*

Proposition 3.3. *The Fiedler vectors of L are the solutions of the problem (8).*

Proof. Since L is symmetric, one may write $L = UDU^\top$, where U is an orthogonal matrix whose column vectors are the eigenvectors of L, denoted $(u_i)_{i \in V}$, and $D = \mathrm{diag}(\lambda)$, where $\lambda = (\lambda_i)_{i \in V}$ is the spectrum of L. We write $V = \{1, \dots, n\}$ and $\lambda = (\lambda_i)_{i=1}^n$ such that $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. One has $u_1 = \mathbf{1}$ and $\mathrm{vect}(u_2, \cdots, u_n) = \mathrm{vect}(u_1)^\perp$.

For every $f \in \mathbb{R}^{|V|}$ such that $\|f\|_2^2 = 1$ and $\langle f, \mathbf{1} \rangle = 0$, one has

$$\begin{aligned} f^\top L f &= f^\top \left(\sum_{i \in V} \lambda_i u_i u_i^\top \right) f = \sum_{i \in V} \lambda_i \langle u_i, f \rangle^2 = \sum_{i=2}^n \lambda_i \langle u_i, f \rangle^2 \\ &\geq \lambda_2 \sum_{i=2}^n \langle u_i, f \rangle^2 = \lambda_2 \|f\|_2^2 = \lambda_2 \end{aligned} \tag{9}$$

because $\lambda_1 = 0$ and, since $\langle f, \mathbf{1} \rangle = 0$, one has $\langle f, u_1 \rangle = 0$, hence $1 = \|f\|_2^2 = \sum_{i=2}^n \langle u_i, f \rangle^2$. Now for $f = u_2$, one has $f^\top L f = \lambda_2$, so this lower bound is attained by a second eigenvector of L. Only these normalized eigenvectors can attain this bound; otherwise it is easy to see that $f^\top L f > \lambda_2$ by inspecting the equality case in (9). ∎

One therefore approximates a solution of the RatioCut problem by the search for a second eigenvector, a Fiedler vector, of L. Once a Fiedler vector u_2 has been computed, one must deduce communities for G. Since u_2 is meant to be an approximate value of the community membership vector $f^{V_1{}^*}$, where $V_1{}^* \subset V$ is a solution of the RatioCut problem, one hopes that the coordinates of u_2 will concentrate around two values (because $f^{V_1{}^*}$ takes exactly 2 values). One then performs a two-class clustering C_1, C_2 on $\mathbb{R}$ of the coordinates of u_2, in order to deduce the communities $\{i : u_{2i} \in C_1\}$, $\{i : u_{2i} \in C_2\}$ of G. **This is exactly the spectral method** in the case $k = 2$ when one takes as first eigenvector $u_1 = \mathbf{1}$ (indeed, clustering the rows of $[u_1|u_2]$ is equivalent to clustering the rows, that is, the coordinates, of $[u_2]$, since u_1 is a vector with all coordinates equal).

RatioCut for k communities: In the case $k = 2$, to each partition $V_1 \sqcup {V_1}^c = V$ one has associated a function f^{V_1} defined in (4). Here, in the general case, with each partition one associates a matrix in the following way. Let $V_1 \sqcup \cdots \sqcup V_k$ be a partition of V; one constructs $H := H(V_1, \ldots, V_k) \in \mathbb{R}^{|V| \times k}$ such that its k column vectors are given by $h_1 = \mathbf{1}_{V_1}/\sqrt{|V_1|}, \ldots, h_k = \mathbf{1}_{V_k}/\sqrt{|V_k|}$. In other words, the membership matrix H is defined for every $i \in V$, $j = 1, \ldots, k$ by

$$h_{ij} = \begin{cases} \frac{1}{\sqrt{|V_j|}} & \text{if } i \in V_j \\ 0 & \text{otherwise.} \end{cases}$$

We give an analogue of Proposition 3.1 in this case.

Proposition 3.4. *The matrix H defined above from a partition $V_1 \sqcup \cdots \sqcup V_k$ of V satisfies:*

1) $H^\top H = I_k$

2) $\mathrm{RatioCut}(V_1, \ldots, V_k) = \mathrm{Tr}(H^\top L H)$.

Proof. **1)** For every $p, q = 1, \ldots, k$, one has

$$(H^\top H)_{pq} = \sum_{i \in V} H_{ip} H_{iq} = \sum_{i \in V} \frac{1}{\sqrt{|V_p||V_q|}} I(i \in V_p) I(i \in V_q).$$

Then if $p \neq q$, since $V_p \cap V_q = \emptyset$, one has $I(i \in V_p) I(i \in V_q) = 0$ for every $i \in V$, and hence $(H^\top H)_{pq} = 0$. When $p = q$, one has $I(i \in V_p) I(i \in V_q) = 1$ for every $i \in V_p$ and otherwise $I(i \in V_p) I(i \in V_q) = 0$, so

$$(H^\top H)_{pp} = \sum_{i \in V_p} \frac{1}{|V_p|} I(i \in V_p) = 1.$$

We therefore have $H^\top H = I_k$.

2) Recall that $h_1, \ldots, h_k$ are the column vectors of H. One has

$$\mathrm{Tr}(H^\top L H) = \sum_{p=1}^{k} (H^\top L H)_{pp} = \sum_{p=1}^{k} h_p^\top L h_p.$$

For every $p = 1, \ldots, k$, one has

$$\begin{aligned} h_p^\top L h_p &= \frac{1}{2} \sum_{(i,j) \in E} W_{ij} (h_{ip} - h_{jp})^2 = \frac{1}{2} \sum_{(i,j) \in E} W_{ij} \left(\frac{I(i \in V_p)}{\sqrt{|V_p|}} - \frac{I(j \in V_p)}{\sqrt{|V_p|}} \right)^2 \\ &= \frac{1}{2} \sum_{\substack{(i,j) \in E \\ i,j \in V_p}} W_{ij} \left(\frac{1}{\sqrt{|V_p|}} - \frac{1}{\sqrt{|V_p|}} \right)^2 + \frac{1}{2} \sum_{\substack{(i,j) \in E \\ i \in V_p, j \notin V_p}} \frac{W_{ij}}{|V_p|} + \frac{1}{2} \sum_{\substack{(i,j) \in E \\ j \in V_p, i \notin V_p}} \frac{W_{ij}}{|V_p|} \\ &= \sum_{\substack{(i,j) \in E \\ i \in V_p, j \notin V_p}} \frac{W_{ij}}{|V_p|} = \frac{W(V_p, V_p^c)}{2|V_p|}. \end{aligned}$$

We therefore have

$$\mathrm{Tr}(H^\top L H) = \sum_{p=1}^{k} \frac{W(V_p, V_p^c)}{2|V_p|} = \mathrm{RatioCut}(V_1, \cdots, V_k).$$

■

Thus, by Proposition 3.4, there is an equivalence between the three problems:

$$\min_{\substack{V_1 \sqcup \cdots \sqcup V_k = V \\ H = H(V_1, \ldots, V_k)}} \operatorname{Tr}(H^\top L H), \tag{10}$$

$$\min_{\substack{V_1 \sqcup \cdots \sqcup V_k = V \\ H = H(V_1, \ldots, V_k) \\ H^\top H = I_k}} \operatorname{Tr}(H^\top L H) \tag{11}$$

and

$$\min_{V_1 \sqcup \cdots \sqcup V_k = V} \operatorname{RatioCut}(V_1, \ldots, V_k). \tag{12}$$

These problems are in general NP-hard. The difficulty does not come from the objective function $F : H \in \mathbb{R}^{n \times k} \to \operatorname{Tr}(H^\top L H)$, which is also convex because its Hessian is constant and equal to $V \in \mathbb{R}^{n \times k} \to 2L$, which is positive semidefinite. The difficulty is (as in the case $k = 2$) due to the constraint, which is of combinatorial type. We shall then perform a convex relaxation of the constraint of (11) by

$$\min_{\substack{H \in \mathbb{R}^{|V| \times k} \\ H^\top H = I_k}} \operatorname{Tr}(H^\top L H). \tag{13}$$

We have simply removed the combinatorial aspect of the constraint, which consists in considering only matrices of the form $H = H(V_1, \ldots, V_k)$.

To recover a solution of (11) from a solution of the relaxed problem, one performs a clustering into k clusters of the rows of the solution matrix of (12), because a solution of (11) is an $n \times k$ matrix having only k distinct rows, and if one runs a k-class clustering on the n rows of this matrix one recovers the k communities. This is therefore indeed the spectral method, provided one proves that the column vectors of a solution of the relaxed problem (12) span a k-dimensional eigenspace associated with the smallest eigenvalues of L. This is what we do now, which will conclude the link between the spectral method and a convex relaxation of the RatioCut problem.

Writing the SVD of $L = UDU^\top$, where $D = \operatorname{diag}(\lambda)$, $\lambda = (\lambda_i)_{i \in V}$ is the spectrum of L such that $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and $U = [u_1 | \cdots | u_n]$ when $V = \{1, \ldots, n\}$, one has

$$\operatorname{Tr}(H^\top L H) = \sum_{i \in V} \lambda_i \, \|a_i\|_2^2$$

where, for every $i \in V$, a_i is the i-th row vector of $U^\top H$, that is, $a_i = (\langle u_i, h_j \rangle)_{j=1,\ldots,k}$. Moreover, since $(h_1, \ldots, h_k)$ is a family of orthonormal vectors (because $H^\top H = I_k$), one has

$$\sum_{i \in V} \|a_i\|_2^2 = \sum_{j=1}^{k} \sum_{i \in V} \langle u_i, h_j \rangle^2 = \sum_{j=1}^{k} \|h_j\|_2^2 = k,$$

$\|a_i\|_2^2 \leq \|u_i\|_2^2 = 1$ for every $i \in V$, and

$$\min_{0 \leq c_i \leq 1 : \sum_i c_i = k} \sum_{i \in V} \lambda_i c_i = \sum_{i=1}^{k} \lambda_i$$

which is attained when $\operatorname{vect}(h_1, \ldots, h_k) = \operatorname{vect}(u_1, \ldots, u_k)$. One therefore sees that the solutions of (13) are the orthogonal matrices H whose columns span the k-dimensional subspace associated with the k smallest eigenvalues of L.

One thus recovers the problem of finding k eigenvectors of L associated with the k smallest eigenvalues of L, taking their multiplicities into account. One then clusters the n row vectors of H into k clusters, from which one deduces the communities of L. This exactly coincides with the spectral method for community detection.

3.3 Pseudo-algorithm of the spectral method for k communities

Let G be an undirected graph, weighted or not. We denote by A its adjacency matrix and by $L = D - A$ its Laplacian, where D is the diagonal matrix of the degrees of the nodes of G. The spectral method for the detection of k communities in G has the following pseudo-code:

Input: L, the Laplacian of the graph G with n nodes, and k, the number of communities
Output: Partition of the nodes of G into k communities
Find an orthogonal basis of k eigenvectors of L associated with the k smallest eigenvalues of L (taking their multiplicity into account).
Denote by H the $n \times k$ matrix whose column vectors are the k eigenvectors of L computed above.
Cluster into k clusters the n row vectors $y_1, \ldots, y_n \in \mathbb{R}^k$ of H: one obtains k clusters $C_1, \ldots, C_k$ forming a partition of the rows $y_1, \ldots, y_n$.
Return the k communities $\{i \in \{1, \ldots, n\} : y_i \in C_p\}, p = 1, \ldots, k$.

Algorithm 1: Spectral method for community detection based on the spectrum of the Laplacian of the graph.

This algorithm is justified either by the convex relaxation of the RatioCut problem introduced in Section 3.2, or by the ideal setting of connected components of Section 3.1. The choice of the number k of communities may be made by an elbow method on the spectrum of the Laplacian (starting from the smallest eigenvalue of L, which equals 0, and moving toward the larger ones, until one meets an elbow, that is, a more pronounced inflection), or by evaluating a modularity function such as the RatioCut function at each solution as a function of k.

4 Probabilistic graph models, spectral method and SDP relaxation for community detection

Idea: In the previous chapters, we assumed that all the edges of the graph are observed. There are, however, cases in which certain links could not be observed, or could not be established with certainty (links between genes, companies, and so on). In that case, one possibility is to assume a probabilistic model underlying our observations. In a sense, one observes the adjacency matrix of a graph only partially, but one still wishes to identify a community structure within this graph. We shall use here the spectral method of the previous sections, but applied only to the Laplacian of the partially observed adjacency matrix.

4.1 Probabilistic graph models and the maximum likelihood estimator.

In this section we present two probabilistic models of random graphs.

The Erdös-Rényi model: One says that the random graph G follows the Erdös-Rényi model on n nodes with parameter p, and one writes $G \sim G(n, p)$, when G is an undirected unweighted graph on n nodes whose adjacency matrix is given by $A = (\delta_{ij})_{1 \leq i,j \leq n}$, where $\delta_{ij} = \delta_{ji}$, $\delta_{ii} = 1$ and $(\delta_{ij} : j > i)$ are independent Bernoulli random variables with parameter p.

Stochastic Block Model (SBM): One says that a random graph follows the SBM on n

nodes with parameters p, q, where $0 \leq q < p \leq 1$, and one writes $G \sim G(n, p, q)$, when $G = (V, E)$ is an undirected unweighted graph with adjacency matrix $A = (\delta_{ij})_{1 \leq i,j \leq n}$ such that there exists a partition $V_1 \sqcup V_1^c = V$ of the vertices of G for which, for every $i < j$,

$$\delta_{ij} \sim \begin{cases} \text{Bern}(p) & \text{if } i, j \in V_1 \text{ or } i, j \in V_1^c \text{ and } i \neq j \\ \text{Bern}(q) & \text{if } (i, j) \in V_1 \times V_1^c \text{ or } (i, j) \in V_1^c \times V_1 \end{cases} \tag{14}$$

and $\delta_{ij} = \delta_{ji}$ and $\delta_{ii} = 1$.

In other words, in an SBM, nodes belonging to the same community are connected with probability p, and if they do not belong to the same community then they are connected with probability q. Since $p > q$, one expects a higher density of nodes within communities than between communities.

The SBM therefore models graphs organized into two communities in the sense defined at the beginning: a community has a higher density of links internally than externally with its complement. The problem we shall seek to solve is the following: given a graph G drawn according to the SBM with parameter (n, p, q), how can one recover the two communities V_1 and V_1^c of G?

Maximum likelihood estimator in the two-community SBM with communities of equal size. The advantage of having a statistical model for our data is that one can use the likelihood of the model to deduce an estimation procedure for the communities V_1 and V_1^c. This approach is that of maximum likelihood: one chooses the parameter V_1 that maximizes "the probability of having observed what one has observed", that is, the likelihood. We begin by giving this likelihood: given the observed adjacency matrix A, for every $V_1 \subset V$ and $0 < p < q < 1$, one has

$$\mathcal{L}(V_1, p, q) = f_{V_1,p,q}^{SBM}(A) = \prod_{(i,j) \in V_1 \times V_1 \cup V_1^c \times V_1^c, i<j} p^{A_{ij}}(1-p)^{1-A_{ij}} \prod_{(i,j) \in V_1 \times V_1^c \cup V_1^c \times V_1, i<j} q^{A_{ij}}(1-q)^{1-A_{ij}}$$

where $f_{V_1,p,q}^{SBM}$ is the density on the adjacency matrices of undirected graphs with diagonal equal to $(1)_1^n$ following the SBM with two communities $V_1 \sqcup V_1^c$ and parameters p, q. We write $\mathcal{C} = \{(i, j) \in V_1 \times V_1 \cup V_1^c \times V_1^c, i < j\}$ and $\mathcal{C}^c = \{(i, j) \in V_1 \times V_1^c \cup V_1^c \times V_1, i < j\}$. The log-likelihood at (V_1, p, q) is therefore given by

$$\log \mathcal{L}(V_1, p, q) = |\mathcal{C}| \log(1 - p) + N_{\mathcal{C}} \log\left(\frac{p}{1-p}\right) + |\mathcal{C}^c| \log(1 - q) + N_{\mathcal{C}^c} \log\left(\frac{q}{1-q}\right)$$

where $N_{\mathcal{C}} = \sum_{(i,j) \in \mathcal{C}} A_{ij}$, $N_{\mathcal{C}^c} = \sum_{(i,j) \in \mathcal{C}^c} A_{ij}$, and $|\mathcal{C}|$ (resp. $|\mathcal{C}^c|$) is the cardinality of $\mathcal{C}$ (resp. $\mathcal{C}^c$). The maximum likelihood estimator consists in maximizing this quantity in (V_1, p, q) under the constraint that $1 > p \geq q > 0$. To this end, one maximizes this quantity in (p, q) under the constraint that $p \geq q$, with V_1 held fixed, and then one maximizes in V_1. We carry out this maximization in the case where the two communities have the same size. In that case, one has $|\mathcal{C}| = |\mathcal{C}^c| = (n^2 - n)/4 := N/2$.

Let $V_1 \subset \{1, \ldots, n\}$ be such that $|V_1| = n/2$. We wish to solve the problem

$$\max\left((1 - P_{\mathcal{C}}) \log(1 - p) + P_{\mathcal{C}} \log(p) + (1 - P_{\mathcal{C}^c}) \log(1 - q) + P_{\mathcal{C}^c} \log(q) : 0 < q \leq p < 1\right) \tag{15}$$

where $P_{\mathcal{C}} = N_{\mathcal{C}}/|\mathcal{C}|$ and $P_{\mathcal{C}^c} = N_{\mathcal{C}^c}/|\mathcal{C}^c|$. One can solve this problem using the KKT theorem (see the end of the section). One finds that if $P_{\mathcal{C}} > P_{\mathcal{C}^c}$, then $p = P_{\mathcal{C}}$ and $q = P_{\mathcal{C}^c}$ is the unique solution, and in this case the value taken by the objective function at this solution equals

$$(1 - P_{\mathcal{C}}) \log(1 - P_{\mathcal{C}}) + P_{\mathcal{C}} \log P_{\mathcal{C}} + (1 - P_{\mathcal{C}^c}) \log(1 - P_{\mathcal{C}^c}) + P_{\mathcal{C}^c} \log P_{\mathcal{C}^c}. \tag{16}$$

Otherwise, when $P_{\mathcal{C}} \leq P_{\mathcal{C}^c}$, the unique solution of (15) is $p = q = (P_{\mathcal{C}} + P_{\mathcal{C}^c})/2$. In this case, and since $P_{\mathcal{C}} + P_{\mathcal{C}^c} = (N_{\mathcal{C}} + N_{\mathcal{C}^c})/(N/2) = (2/N)\sum_{i<j} A_{ij}$ is constant (that is, independent of V_1), the value taken by the objective function at its optimum is independent of V_1 and also equals (16) in the case where $P_{\mathcal{C}} = P_{\mathcal{C}^c}$. One concludes that the optimization problem over $V_1 \subset \{1, \ldots, n\}$ such that $|V_1| = n/2$ that remains to be solved is that of maximizing the optimal value obtained in (16) under the constraint that $P_{\mathcal{C}} \geq P_{\mathcal{C}^c}$, that is,

$$\max_{\substack{V_1 \subset \{1,\ldots,n\} \\ |V_1| = n/2, P_{\mathcal{C}} \geq P_{\mathcal{C}^c}}} \left[(1 - P_{\mathcal{C}}) \log(1 - P_{\mathcal{C}}) + P_{\mathcal{C}} \log P_{\mathcal{C}} + (1 - P_{\mathcal{C}^c}) \log(1 - P_{\mathcal{C}^c}) + P_{\mathcal{C}^c} \log P_{\mathcal{C}^c}\right]. \quad (17)$$

Once this problem has been solved in V_1, one obtains the MLE $\hat{p} = N_{\mathcal{C}}/|\mathcal{C}| = P_{\mathcal{C}}$ and $\hat{q} = N_{\mathcal{C}^c}/|\mathcal{C}^c| = P_{\mathcal{C}^c}$ for the optimal choice of V_1. One may interpret the constraint "$P_{\mathcal{C}} \geq P_{\mathcal{C}^c}$" as an empirical version of the constraint $p \geq q$, since the quantities $P_{\mathcal{C}}$ and $P_{\mathcal{C}^c}$ are the MLEs of p and q for the optimal choice of V_1.

It remains to solve the optimization problem (17). We show at the end of this section that the objective function of the problem (17) is maximal at $(P_{\mathcal{C}}, P_{\mathcal{C}^c}) \in [0,1]^2$ under the constraint that $P_{\mathcal{C}} + P_{\mathcal{C}^c} = (2/N)\sum_{i<j} A_{ij}$ and $P_{\mathcal{C}} \geq P_{\mathcal{C}^c}$ when $P_{\mathcal{C}}$ is maximal, or equivalently when $P_{\mathcal{C}^c}$ is minimal (since their sum is constant). It then remains to find the sets $V_1 \subset \{1, \cdots, n\}$ such that $|V_1| = n/2$ for which $P_{\mathcal{C}^c}$ is minimal. Moreover, one observes that for every $V_1 \subset \{1, \ldots, n\}$ such that $|V_1| = n/2$, one has

$$P_{\mathcal{C}^c} = \frac{2}{N} \sum_{(i,j) \in \mathcal{C}^c} A_{ij} = \frac{W(V_1, V_1^c)}{N} = \mathrm{RatioCut}(V_1, V_1^c).$$

One must therefore minimize, over $V_1 \subset V$ such that $|V_1| = n/2$, the ratiocut $\mathrm{RatioCut}(V_1, V_1^c)$. One thus recovers the problem of minimizing RatioCut as introduced in Section 3.2, in the case where there are two communities of equal size. **One therefore sees that the problem of minimizing the RatioCut in the case of two communities of equal size is also a maximum likelihood problem in the stochastic block model**.

Remark 4.1. *In the case of the detection of two communities of equal size, the minimal RatioCut problem and the MLE are equivalent, and amount to solving*

$$\min\left(x^\top L x : \langle x, 1\rangle = 0, x \in \{-1,1\}^n\right). \quad (18)$$

because for every $V_1 \subset [V]$ such that $|V_1| = |V_1{}^c|$, one has $f^{V_1} \in \{-1,1\}^n$. Moreover, one also has, for every $x \in \{-1,1\}^n$, that $x^\top L x = -x^\top A x + \sum_{i,j} A_{ij}$, because $x^\top D x = \sum_{i,j} D_{ij} x_i x_j = \sum_i d_i x_i^2 = \sum_i d_i$, where d_i is the degree of node i. Since the term $\sum_{i,j} A_{ij}$ is constant (independent of x), one sees that (18) is equivalent to

$$\max\left(x^\top A x : \langle x, 1\rangle = 0, x \in \{-1,1\}^n\right). \quad (19)$$

It is on this problem that we shall perform an SDP relaxation in Section 6.

We conclude this section with the resolution of the two optimization problems (15) and (17). We begin with (15). One sees that the objective function $f : (p,q) \in]0,1[^2 \to (1 - P_{\mathcal{C}}) \log(1-p) + P_{\mathcal{C}} \log(p) + (1 - P_{\mathcal{C}^c}) \log(1-q) + P_{\mathcal{C}^c} \log(q)$ is strongly concave and that its level sets are compact, so (15) admits a unique solution. Let (p,q) be a solution of the problem. By KKT, there exists

$\mu \geq 0$ such that $-\nabla f(p,q) + \mu \nabla h(p,q) = 0$ and $\mu h(p,q) = 0$, where $h : (p,q) \to q - p$. One then has $\mu \geq 0$, $\mu(p-q) = 0$, $p \geq q$ and

$$\frac{p - P_{\mathcal{C}}}{p(1-p)} = \mu = \frac{P_{\mathcal{C}^c} - q}{q(1-q)}. \tag{20}$$

If $p \neq q$, then $\mu = 0$ and hence, by (20), $p = P_{\mathcal{C}}$ and $q = P_{\mathcal{C}^c}$ is indeed the unique solution, provided that $P_{\mathcal{C}} > P_{\mathcal{C}^c}$. If $p = q$, then by (20) one has $p = q = (P_{\mathcal{C}} + P_{\mathcal{C}^c})/2$, which is the unique solution provided that $P_{\mathcal{C}} < P_{\mathcal{C}^c}$ (otherwise $\mu < 0$), and when $P_{\mathcal{C}} = P_{\mathcal{C}^c}$ one has $p = q = P_{\mathcal{C}}$, which is the unique solution. In fact, the problem (15) is a differentiable convex optimization problem, and one can see that when $P_{\mathcal{C}} > P_{\mathcal{C}^c}$, then $((P_{\mathcal{C}}, P_{\mathcal{C}^c}), 0)$ is a saddle point of the Lagrangian, so there is strong duality and $(P_{\mathcal{C}}, P_{\mathcal{C}^c})$ is the unique solution of the primal (that is, of (15)) (and 0 is a solution of the dual problem). When $P_{\mathcal{C}} \leq P_{\mathcal{C}^c}$, then $((P_{\mathcal{C}} + P_{\mathcal{C}^c})/2, (P_{\mathcal{C}} + P_{\mathcal{C}^c})/2), (P_{\mathcal{C}^c} - P_{\mathcal{C}})/[(P_{\mathcal{C}} + P_{\mathcal{C}^c})(1 - (P_{\mathcal{C}} + P_{\mathcal{C}^c})/2)]$ is a saddle point of the Lagrangian, so $p = q = (P_{\mathcal{C}} + P_{\mathcal{C}^c})/2$ is the unique solution of the problem (15).

We now treat the problem (17). We give here the proof that the objective function of the problem (17) is maximal at $(P_{\mathcal{C}}, P_{\mathcal{C}^c}) \in]0,1[^2$ under the constraint that $P_{\mathcal{C}} + P_{\mathcal{C}^c} = (2/N) \sum_{i<j} A_{ij} := s$ is constant and $P_{\mathcal{C}} \geq P_{\mathcal{C}^c}$, when $P_{\mathcal{C}}$ is maximal, or equivalently when $P_{\mathcal{C}^c}$ is minimal (since their sum is constant). Since $P_{\mathcal{C}} + P_{\mathcal{C}^c} = s$ is constant, it suffices to minimize $P_{\mathcal{C}} \to \mathrm{Ent}(P_{\mathcal{C}}) + \mathrm{Ent}(s - P_{\mathcal{C}})$ under the constraint that $s/2 \leq P_{\mathcal{C}} \leq \min(s,1)$. Now the function $F : p \in [0,1] \to \mathrm{Ent}(p) + \mathrm{Ent}(s-p)$ is concave and maximal at $s/2$, so its minimum on $s/2 \leq p \leq \min(s,1)$ is attained at $\min(s,1)$, and hence $P_{\mathcal{C}} \to F(P_{\mathcal{C}})$ is minimal over the set of values taken by $P_{\mathcal{C}}$ when $P_{\mathcal{C}}$ is maximal.

Detectability threshold in the two-community SBM with communities of equal size. We consider a graph obtained from an SBM with parameter (n,p,q), where $p > q$, with two communities V_1, V_1^c of equal size $n/2$. We wish to recover the two communities from an observation of this graph. One may observe that if there is a point of the community V_1 that is isolated from the other nodes of V_1, then one will have no reason to believe that this node is indeed in the community V_1. In that case, recovery of the communities is not possible. In the following result, we therefore give a minimal "signal level" on p so that there is no isolated point in an SBM graph.

Proposition 4.2. *If $p \leq \log(n/2)/[n/2 - 2]$ then the probability that there is an isolated node in the community V_1 is greater than $1/2$.*

Preuve. For every $i \in V_1$, we denote by A_i the event that node i is isolated from the other nodes of the community V_1. Formally, one has

$$A_i = \{\forall j \in V_1, j \neq i, \delta_{ij} = 0\}.$$

We write $N = |V_1| - 1$. One therefore has $\mathbb{P}[A_i] = (1-p)^N$. In particular, $\mathbb{P}[A_i] \geq 1/2$ if and only if $p \leq 1 - \exp(-(\log 2)/N)$, and hence for $p \leq \log 2/(2N)$ one will indeed have that i is an isolated node with probability at least $1/2$. One obtains the result up to a logarithmic factor.

To obtain the result, we shall use the fact that we wish to have an isolated node among the $|V_1|$ possible ones. It is this aspect that will give us the extra $\log N$ in the threshold. We are then interested in the probability $\mathbb{P}[\cup_{i \in V_1} A_i]$. We show by induction that

$$\mathbb{P}[\cup_{i \in V_1} A_i] \geq \sum_{i \in V_1} \mathbb{P}[A_i] - \sum_{i,j \in V_1 : i \neq j} \mathbb{P}[A_i \cap A_j].$$

One sees that if $i \neq j$, then

$$\mathbb{P}[A_i \cap A_j] = \mathbb{P}\left[\delta_{ij} = 0 \& \forall k \in V_1 \backslash \{i,j\}, \delta_{ik} = 0, \delta_{jk} = 0\right] = (1-p)^{2N-1}$$

and hence

$$\mathbb{P}[\cup_{i \in V_1} A_i] \geq (N+1)(1-p)^N - (N+1)N(1-p)^{2N-1}.$$

One then seeks to determine a condition on p which implies that $(N+1)(1-p)^N - (N+1)N(1-p)^{2N-1} \geq 1/2$. We set $X = (N+1)(1-p)^N$. We then wish to solve $X - X^2 N/[(N+1)(1-p)] \geq 1/2$. This inequality holds when

$$2N(1-p)^{N-1} \leq 1 + \left(1 + \frac{2N}{(N+1)(1-p)}\right)^{1/2}.$$

In particular, when $2N(1-p)^{N-1} \leq 2$, this inequality holds. One then sees that for $p \leq (\log N)/(N-1)$, one indeed has $\mathbb{P}[\cup_{i \in V_1} A_i] \geq 1/2$. ∎

A consequence of Proposition 4.2 is that if one considers the problem of exact recovery of the communities underlying a graph drawn according to the SBM, then one must assume that $p \geq \alpha \log(n)/n$. Since $p > q$, one must have $q < \alpha \log(n)/n$. The difficult case is then to consider q as close as possible to p in this regime. One will then also take a q of the form $\beta \log(n)/n$, where $\beta < \alpha$. In what follows, we are interested in the conditions on α and β that ensure exact detection of the communities with high probability by a procedure that is feasible in practice.

If one used the maximum likelihood estimator $\hat{V}_1 \in \operatorname{argmin}_{\{V_1 \subset V : |V_1| = n/2\}} \operatorname{RatioCut}(V_1, V_1^c)$, then one could attain the information-theoretic impossibility threshold for exact recovery, as stated in the following theorem (which we do not prove).

Theorem 4.3 (Phase transition in the two-class equal-size SBM)**.** *If $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$, then with high probability $\hat{V}_1$ and $\hat{V}_1^c$ are the underlying communities V_1 and V_1^c. If $\sqrt{\alpha} - \sqrt{\beta} < \sqrt{2}$, then no algorithm can recover the communities V_1 and V_1^c with probability at least $1/2$.*

This theorem is a result giving a very precise phase transition: if $\sqrt{\alpha} - \sqrt{\beta}$ is above $\sqrt{2}$, then one can recover the communities, and this is impossible if one is below this threshold. Nevertheless, this result is interesting from a theoretical viewpoint, because in practice one will not use the maximum likelihood estimator, but a numerically computable procedure: either a spectral method or an SDP relaxation. One is then interested in identifying a threshold on α and β that makes it possible to recover the underlying communities.

4.2 Spectral methods for the two-class equal-size SBM based on the search for a Fiedler vector

We consider a graph $G = (V, E)$ drawn according to the SBM with parameter (n, p, q), with $p > q$ and $V = \{1, \ldots, n\}$. We denote by A the (observed) adjacency matrix of G as defined in (14). As seen in the previous section, the spectral method here consists in finding a Fiedler vector of the Laplacian of G, that is, of $L = D - A$, where $D = \operatorname{diag}(d_1, \ldots, d_n)$ and $d_i = \sum_{j=1}^n A_{ij}$ is the degree of the i-th node of G. Once a Fiedler vector has been obtained, one clusters its coordinates into two groups, which will ideally give us the two communities of G. This is what we would like to prove here with high probability.

We rewrite the spectral method more simply when the two communities V_1 and V_1^c have the same cardinality, thanks to the following lemma.

Lemma 4.4. *Let $A \in \mathbb{R}^{n\times n}$ be the adjacency matrix of a graph $G = (V, E)$ drawn according to the SBM with parameter (n, p, q), where n is even. Assume that the two communities V_1, V_1^c underlying the SBM have the same size $n/2$. There is an equivalence between:*

1) u_2 is a Fiedler vector of $\mathbb{E}L = \mathbb{E}(D - A)$

2) u_2 is an eigenvector associated with the largest eigenvalue of

$$\mathbb{E}A - \left[n\left(\frac{p+q}{2}\right) + (1-p)\right]\left(\frac{\mathbf{1}}{\sqrt{n}}\right)\left(\frac{\mathbf{1}}{\sqrt{n}}\right)^\top \tag{21}$$

where $\mathbf{1} = (1)_{i=1}^n$.

Proof. Up to reordering the nodes of V, one may write $\mathbb{E}A$ in the form

$$\mathbb{E}A = \left[\begin{array}{cccc|cccc} 1 & p & \cdots & p & q & q & \cdots & q \\ p & 1 & \cdots & p & q & q & \cdots & q \\ \cdots & \cdots & \ddots & \cdots & \cdots & \cdots & \cdots & \cdots \\ p & p & \cdots & 1 & q & \cdots & \cdots & q \\ \hline q & q & \cdots & q & 1 & p & \cdots & p \\ q & q & \cdots & q & p & 1 & \cdots & p \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \ddots & \cdots \\ q & \cdots & \cdots & q & p & p & \cdots & 1 \end{array}\right].$$

One sees that the mean degree matrix is then proportional to the identity, because for every node $i \in \{1, \dots, n\}$, the mean degree of node i is $\mathbb{E}d_i = 1 + (n/2 - 1)p + (n/2)q$. One therefore has $\mathbb{E}D = (1 + (n/2 - 1)p + (n/2)q)I_n$.

Since $\mathbb{E}D$ is proportional to the identity, one sees that f is an eigenvector of $\mathbb{E}L$ with eigenvalue λ if and only if it is an eigenvector of $\mathbb{E}A$ with eigenvalue $[1 + (n/2 - 1)p + (n/2)q] - \lambda$. Thus, looking for a Fiedler vector of $\mathbb{E}L$ is equivalent to looking for an eigenvector of $\mathbb{E}A$ for its second eigenvalue.

Moreover, since 0 is the smallest eigenvalue of $\mathbb{E}L$ (because it is the Laplacian of the graph whose adjacency matrix is given by $\mathbb{E}A$), associated with the eigenvector $\mathbf{1} = (1)_1^n$, the quantity $1 + (n/2 - 1)p + (n/2)q$ is the largest eigenvalue of $\mathbb{E}A$, associated with the same eigenvector $\mathbf{1} = (1)_1^n$. Writing the singular value decomposition of $\mathbb{E}A = \sum_i ([1 + (n/2 - 1)p + (n/2)q] - \lambda_i)u_i \otimes u_i$, where $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ are the eigenvalues of $\mathbb{E}L$ associated with the respective eigenvectors $u_1, \dots, u_n$ (forming an orthonormal basis of $\mathbb{R}^n$), one sees that $\mathbb{E}A - [1 + (n/2 - 1)p + (n/2)q]u_1 \otimes u_1$ has largest eigenvalue $[1 + (n/2 - 1)p + (n/2)q] - \lambda_2$, which is the second largest eigenvalue of $\mathbb{E}A$. It is therefore equivalent to look for an eigenvector of $\mathbb{E}A$ associated with its second largest eigenvalue and to look for an eigenvector associated with the largest eigenvalue of $\mathbb{E}A - [1 + (n/2 - 1)p + (n/2)q]u_1 \otimes u_1$. Now, one may take $u_1 = \mathbf{1}/\|\mathbf{1}\|_2$. This concludes the proof. ∎

In the case of two communities of equal size, the spectral method therefore consists, by Lemma 4.4, in looking for an eigenvector associated with the largest eigenvalue of the matrix (21) (and then in clustering its coordinates).

However, one knows neither $\mathbb{E}A$, nor p, nor q. We shall therefore estimate these quantities: one uses A to estimate $\mathbb{E}A$, and for $(p + q)/2$ one observes that for

$$\lambda = \frac{2}{n(n-1)} \sum_{i<j} A_{ij} \tag{22}$$

one has

$$\mathbb{E}(\lambda) = \frac{2}{n(n-1)}\left[\left(\binom{n}{2}^2 - \binom{n}{2}\right)p + \binom{n}{2}^2 q\right] = \frac{p+q}{2} - \frac{p-q}{n-1}.$$

Thus λ estimates $(p+q)/2$ up to a term of order $1/n$ (λ is also the average of the maximum likelihood estimators of p and q obtained in Section 4.1). One will then estimate a leading eigenvector of (21) by a leading eigenvector of $A - n\lambda(\mathbf{1}/\sqrt{n})(\mathbf{1}/\sqrt{n})^\top$, that is, a solution of the problem

$$\max_{x:\|x\|_2\leq 1} \langle x, (A - \lambda J_n)x\rangle \tag{23}$$

where $J_n = (1)_{n\times n}$. This type of problem is called a **spectral method**, because in the end one is led to compute an eigenvector of a matrix in order to solve the problem.

4.3 *Matrix lifting* and SDP relaxation for the search for communities in the two-class equal-size SBM.

The problem of detecting two communities is written, initially, as a problem of finding a solution of the discrete problem

$$\min_{V_1\subset V}\left((f^{V_1})^\top \mathbb{E}L f^{V_1} : \left\|f^{V_1}\right\|_2^2 = 1, \langle f^{V_1}, \mathbf{1}\rangle = 0\right) \tag{24}$$

as introduced in (6), where $\mathbb{E}L$ was entirely known (and not only through the single observation L). This problem was relaxed by removing the constraint that f had to be of the form f^{V_1} for some $V_1 \subset V$ (where f^{V_1} is defined in (4)). This relaxation motivated the introduction of the search for a Fiedler vector of the Laplacian. One may, however, propose other convex relaxations.

For example, in the case of two classes of equal size, we have seen that looking for a Fiedler vector of $\mathbb{E}L$ is equivalent to looking for a leading eigenvector of (21), which can be estimated by $A - \lambda J_n$, where λ is defined by (22) and $J_n = (1)_{n\times n}$. One may then hope that a solution of the problem

$$\max_{V_1\subset V}\left((f^{V_1})^\top (A - \lambda J_n) f^{V_1}\right) \tag{25}$$

can be a good approximate solution of a solution of the initial problem (24). Moreover, under the assumption of communities of equal size, one sees that $f^A \in \{-1,1\}^n$ is such that $(f^A)_i = 1$ if $i \in V_1$ and $(f^A)_i = -1$ if $i \notin V_1$. The problem (25) may therefore be rewritten as

$$\max_{x\in\{-1,1\}^n}\left(x^\top (A - \lambda J_n)x\right) \tag{26}$$

which is itself equivalent to

$$\max_{x\in\{-1,1\}^n}\langle A - \lambda J_n, xx^\top\rangle \tag{27}$$

where one uses here the inner product between two matrices given by $\langle A, B\rangle = \sum_{i,j} A_{ij}B_{ij} = \mathrm{Tr}(AB^\top)$. The passage from an optimization problem with an objective function quadratic in x to an optimization problem with an objective function linear in $xx^\top$ is known as **matrix lifting**, and is a classic of convex relaxation (see Section 7 for other examples).

The problem (27) is a combinatorial problem, and hence potentially difficult to solve directly. We shall use a convex relaxation to approximate it. This approach, which simply consists in removing a rank constraint, is called an **SDP relaxation**. To this end, one views $Z = xx^\top$ in the problem (27) as a matrix variable having the following properties:

i) Z is symmetric

ii) Z is positive semidefinite, because $\langle xx^\top y, y\rangle = \langle x, y\rangle^2 \geq 0$ for every $y \in \mathbb{R}^n$

iii) $\mathrm{diag}(Z) = \mathrm{diag}(x_1^2, \ldots, x_n^2) \preceq I_n$, because $x_i^2 = 1$ for every $n = 1, \ldots, n$ (we denote by $\mathrm{diag}(Z)$ the diagonal matrix of $\mathbb{R}^{n\times n}$ whose diagonal entries coincide with those of Z).

iv) $\mathrm{rang}(Z) = 1$.

One may rewrite the constraint in the following form:

$$\left\{xx^\top : x_i^2 = 1\right\} \subset \left\{Z \in \mathbb{R}^{n\times n} : Z \succeq 0, \mathrm{diag}(Z) \preceq I_n, \mathrm{rang}(Z) = 1\right\}$$

We shall then perform a convex relaxation of (27) simply by removing the rank constraint in the set above (which is the cause of the nonconvexity of the constraint). In doing so, only affine constraints remain ($\mathrm{diag}(Z) \preceq I_n$ being equivalent to $Z_{ii} \leq 1, \forall i = 1, \ldots, n$, that is, $\langle E_{ii}, Z\rangle \leq 1, \forall i$) and the SDP constraint $Z \succeq 0$. Thus, since the objective function is linear, one obtains an SDP problem, a *semidefinite program*. This is why one speaks of an SDP relaxation. One will then look for a solution of the problem

$$\max_{\substack{Z\succeq 0\\ \mathrm{diag}(Z)\preceq I_n}} \langle A - \lambda J_n, Z\rangle \tag{28}$$

where we recall that $Z \succeq 0$ means that Z is symmetric and that $\langle Zy, y\rangle \geq 0$ for every $y \in \mathbb{R}^n$. One also uses that $A \succeq B$ when $A - B \succeq 0$.

If $\hat{Z}$ is a solution of (28), one then takes a leading eigenvector (an eigenvector associated with its largest eigenvalue) of $\hat{Z}$, and one takes its sign. The coordinates of sign 1 form one community, and the others, of sign -1, form the other community. One hopes that this procedure can estimate well a solution $x^* \in \{-1, 1\}^n$ of the problem (26), but also, and above all, of the initial problem (24), with high probability. Proving this is the object of the following sections.

We first check that the community membership vector $\bar{x} \in \{-1, 1\}^n$ defined by $\bar{x}_i = 1$ if $i \in V_1$ and $\bar{x}_i = -1$ if $i \notin V_1$ is indeed such that $\bar{x}(\bar{x})^\top$ is the unique solution of the relaxed problem in which A and λ are replaced by the "oracle" values $\mathbb{E}A$ and α:

$$\max_{\substack{Z\succeq 0\\ \mathrm{diag}(Z)\preceq I_n}} \langle \mathbb{E}A - \alpha u_1 \otimes u_1, Z\rangle. \tag{29}$$

where $\alpha = [(p+q)/2]n + (1-p)$ and $u_1 = \mathbf{1}/\sqrt{n}$. We also recall that $u \otimes v = uv^\top = (u_i v_j)_{1\leq i,j\leq n} \in \mathbb{R}^{n\times n}$.

Proposition 4.5. *The problem (29) admits a unique solution, given by $\bar{x} \otimes \bar{x}$.*

Proof. We write $B = \mathbb{E}A - \alpha u_1 \otimes u_1$. We begin by studying the spectrum of B. One has

$$B = \left[\begin{array}{c|c} p & q \\ \hline q & p \end{array}\right] - \left[n\left(\frac{p+q}{2}\right) + (1-p)\right] u_1 \otimes u_1 + (1-p) I_n.$$

We write $u_2 = \bar{x}/\sqrt{n}$. One has

$$\left[\begin{array}{c|c} p & q \\ \hline q & p \end{array}\right] = n\left(\frac{p+q}{2}\right) u_1 \otimes u_1 + n\left(\frac{p-q}{2}\right) u_2 \otimes u_2.$$

Since $\langle u_1, u_2 \rangle = 0$ and $\|u_1\|_2 = \|u_2\| = 1$, the pair (u_1, u_2) forms the beginning of an orthonormal basis that one can complete: let $(u_i)_{i=3}^n$ be such that $(u_i)_{i=1}^n$ forms an orthonormal basis of $\mathbb{R}^n$. We write

$$(1-p)I_n = (1-p)u_1 \otimes u_1 + (1-p)u_2 \otimes u_2 + (1-p)\sum_{i=3}^n u_i \otimes u_i.$$

One deduces that

$$B = \left[n\left(\frac{p-q}{2}\right) + (1-p)\right] u_2 \otimes u_2 + (1-p)\sum_{i=3}^n u_i \otimes u_i.$$

One then sees, thanks to the spectral theorem, that there is an equivalence between the two problems:

1)
$$\bar{Z} \in \underset{Z \succeq 0, \mathrm{diag}(Z) \preceq I_n}{\mathrm{argmax}} \langle B, Z \rangle \tag{30}$$

2) $\bar{Z} = \bar{X}(\bar{X})^\top$ and
$$\bar{X} \in \underset{\substack{X \in \mathbb{R}^{n\times n} \\ X_{i\bullet} \in B_2^n}}{\mathrm{argmax}} \langle B, XX^\top \rangle \tag{31}$$

where, for every $i = 1, \ldots, n$, $X_{i\bullet}$ is the i-th row vector of X and $X_{i\bullet} \in B_2^n$ means that $\|X_{i\bullet}\|_2 \leq 1$.

Let $X \in \mathbb{R}^{n\times n}$ be such that for every $i = 1, \ldots, n$ one has $X_{i\bullet} \in B_2^n$. We denote by $X_{\bullet j}$ the j-th column vector of X. One has

$$XX^\top = \left(\sum_{k=1}^n X_{ik}X_{jk}\right)_{1\leq i,j\leq n} = \sum_{k=1}^n X_{\bullet k} \otimes X_{\bullet k}$$

because $X_{\bullet k} \otimes X_{\bullet k} = (X_{ik}X_{jk})_{1\leq i,j\leq n}$. One observes that $\langle u \otimes u, v \otimes v \rangle = \mathrm{Tr}(uu^\top vv^\top) = \langle u, v \rangle^2$ for every $u, v \in \mathbb{R}^n$. Since $p > q$ and $(u_i)_{i=1}^n$ is an orthonormal basis of $\mathbb{R}^n$, one has

$$\begin{aligned}
&\langle B, XX^\top \rangle \\
&= \left\langle \left[n\left(\frac{p-q}{2}\right) + (1-p)\right] u_2 \otimes u_2 + (1-p)\sum_{i=3}^n u_i \otimes u_i, \sum_{k=1}^n X_{\bullet k} \otimes X_{\bullet k} \right\rangle \\
&= \left[n\left(\frac{p-q}{2}\right) + (1-p)\right] \sum_{k=1}^n \langle u_2, X_{\bullet k} \rangle^2 + (1-p)\sum_{i=3}^n \sum_{k=1}^n \langle u_i, X_{\bullet k} \rangle^2 \\
&\leq \left[n\left(\frac{p-q}{2}\right) + (1-p)\right] \sum_{i=2}^n \sum_{k=1}^n \langle u_i, X_{\bullet k} \rangle^2 \\
&\leq \left[n\left(\frac{p-q}{2}\right) + (1-p)\right] \sum_{i=1}^n \sum_{k=1}^n \langle u_i, X_{\bullet k} \rangle^2 \\
&\leq \left[n\left(\frac{p-q}{2}\right) + (1-p)\right] \sum_{k=1}^n \|X_{\cdot k}\|_2^2 = \left[n\left(\frac{p-q}{2}\right) + (1-p)\right] \sum_{i=1}^n \|X_{i\bullet}\|_2^2 \\
&= n\left[n\left(\frac{p-q}{2}\right) + (1-p)\right]
\end{aligned} \tag{32}$$

Moreover, by inspecting the equality case in the upper bound on $\langle B, XX^\top \rangle$ in (32), one sees that necessarily $\langle u_i, X_{\bullet k} \rangle = 0$ for every $i = 1$ and $i = 3, \dots, n$ and $k = 1, \dots, n$. One must also have $\|X_{i\bullet}\|_2 = 1$ for every $i = 1, \dots, n$. Hence the columns $X_{\bullet k}$ of X are necessarily supported on u_2. Writing X as being of the form $[a_1 u_2 | a_2 u_2 | \cdots | a_n u_2]$ for real numbers $a_1, \dots, a_n$, one has $\|X_{i\bullet}\|_2 = 1$ for every $i = 1, \dots, n$ if and only if $\left(\sum_{i=1}^n a_i^2\right)^{1/2} |u_{2i}| = 1$ for every $i = 1, \dots, n$, and since $|u_{2i}| = 1/\sqrt{n}$, one necessarily has $\left((1/n)\sum_{i=1}^n a_i^2\right)^{1/2} = 1$.

On the other hand, one sees that for $\bar{X}$ having column vectors $a_i u_2, i = 1, \dots, n$ for real numbers $a_1, \dots, a_n$ such that $\left((1/n)\sum_{i=1}^n a_i^2\right)^{1/2} = 1$, one has $\bar{X}_{i\bullet} = (a_1 u_{2i}, a_2 u_{2i}, \cdots, a_n u_{2i})^\top \in \mathbb{R}^n$ such that $\left\|\bar{X}_{i\bullet}\right\|_2 = \left(\sum_{i=1}^n a_i^2\right)^{1/2} |u_{2i}| = 1$ for every $i = 1, \dots, n$, because $|u_{2i}| = 1/\sqrt{n}$. Hence $\bar{X}$ does lie in the constraint set of (31). Moreover, $\bar{X}(\bar{X})^\top = (\sum_i a_i^2) u_2 \otimes u_2 = n u_2 \otimes u_2$, so

$$\langle B, \bar{X}(\bar{X})^\top \rangle = n \left[n \left(\frac{p-q}{2} \right) + (1-p) \right].$$

because $\langle u_2 \otimes u_2, u_2 \otimes u_2 \rangle = \langle u_2, u_2 \rangle^2 = 1$. Hence the bound $n\left[n\left((p-q)/2\right) + (1-p)\right]$ is attained by $\bar{X}$ in (32).

Hence the solutions of (31) are all of the form $\bar{X} = [a_1 u_2 | \cdots | a_n u_2]$, where $\sum_i a_i^2 = n$. In this case, the solutions of (30) are of the form $\bar{Z} = \bar{X}(\bar{X})^\top = (\sum_i a_i^2) u_2 \otimes u_2 = n u_2 \otimes u_2 = \bar{x} \otimes \bar{x}$, which is therefore unique. ∎

Since $\bar{x} \otimes \bar{x}$ has rank 1 and its unique eigenspace associated with the eigenvalue n is spanned by $\bar{x}$, one sees that by taking the sign of $\lambda \bar{x}$ (a leading eigenvector of $\bar{x} \otimes \bar{x}$) for every $\lambda \in \mathbb{R}$, one obtains $\bar{x}$ or $-\bar{x}$. One can therefore recover the communities of the graph by taking the sign of a leading eigenvector of a solution of the problem (29). This motivates the spectral method which consists in taking the sign of a leading eigenvector of a solution of the problem (29) in which $\mathbb{E}A - \alpha u_1 \otimes u_1$ has been replaced by its estimate $A - \lambda J_n$. We write this spectral method in pseudo-code:

1 **Input:** A: adjacency matrix of a graph distributed according to the SBM with two communities of equal size.
2 **Output:** Partition of the nodes of G into 2 communities of equal size
3 Computation of λ given in (22)
4 Solution of the SDP problem (28)
5 Search for a leading eigenvector $\hat{x}$ of $\hat{Z}$, a solution of (28)
6 Take the sign of $\hat{x}$
7 Return the 2 communities $\hat{V}_1 = \{i \in \{1, \dots, n\} : \mathrm{sign}(\hat{x}_i) = 1\}$ and its complement $\hat{V}_1^c$.

Algorithm 2: Spectral method for community detection based on an SDP relaxation.

5 Statistical analysis of an SDP relaxation: estimation rate and Grothendieck's inequality.

Problem: One observes the adjacency matrix A of a graph distributed according to an SBM with two communities of equal size. One wishes to recover these two communities from A. In the previous section we introduced a procedure obtained by matrix lifting / SDP relaxation of the

original problem (see Algorithm 2). The objective of this section is to prove that this procedure provides a good estimator of the membership vector $\bar{x}$.

We recall a few notations: λ is defined in (22),

$$\hat{Z} \in \underset{\substack{Z \succeq 0 \\ \operatorname{diag}(Z) \preceq I_n}}{\operatorname{argmax}} \langle A - \lambda J_n, Z \rangle \tag{33}$$

where $J_n = (1)_{n \times n}$ and $\hat{x}$ is a leading eigenvector of $\hat{Z}$. We want to show that $\hat{x}$ is close to $\bar{x}$, the membership vector of the two communities V_1 and V_1^c: $\bar{x}_i = 1$ when $i \in V_1$ and $\bar{x}_i = -1$ when $i \notin V_1$.

Theorem 5.1. *Let $\epsilon \in (0,1)$ be such that $n \geq 10^4 \epsilon$. Let A be the adjacency matrix of a graph distributed according to the SBM with parameter (n,p,q), where $p > q$. Assume that $\max(p(1-p), q(1-q)) \geq 20/n$. Assume that $p \geq a/n$, $q \geq b/n$ and $(a-b)^2 \geq 10^4 \epsilon^{-2}(a+b)$. Let $\hat{Z}$ be a solution of (33). With probability at least $1 - e^3 5^{-n}$, one has*

$$\left\| \hat{Z} - \bar{x}(\bar{x})^\top \right\|_2^2 \leq \epsilon n^2 = \epsilon \left\| \bar{x}(\bar{x})^\top \right\|_2^2 .$$

Next, one passes from the estimation of the matrix $\bar{x}(\bar{x})^\top$ in the 2-norm to the estimation of a leading eigenvector of $\bar{x}(\bar{x})^\top$ in the following way.

Corollary 5.2. *Under the assumptions of Theorem 5.1. If $\hat{x}$ is a leading eigenvector of $\hat{Z}$ such that $\|\hat{x}\|_2 = \sqrt{n}$, then*

$$\min_{\alpha \in \{\pm 1\}} \|\alpha \hat{x} - \bar{x}\|_2^2 \leq 8\epsilon n = 8\epsilon \|\bar{x}\|_2^2 .$$

The proof of Corollary 5.2 relies on the Davis–Kahan theorem, also known as the "$\sin\theta$-theorem", which we now recall.

Theorem 5.3 (Davis–Kahan)**.** *Let A and B be two symmetric matrices in $\mathbb{R}^{n \times n}$. Let $i \in \{1, \dots, n\}$. Assume that the i-th largest eigenvalue $\lambda_i(A)$ of A is well separated from the rest of the spectrum of A:*

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

If $u_i(A)$ is a unit eigenvector of A associated with the eigenvalue $\lambda_i(A)$ and $u_i(B)$ is a unit eigenvector associated with the i-th largest eigenvalue of B, then

$$\min_{\alpha \in \{\pm 1\}} \|u_i(A) - \alpha u_i(B)\|_2 \leq 2^{3/2} \frac{\|A - B\|_2}{\delta}.$$

Proof of Corollary 5.2. We apply Davis–Kahan to $A = \bar{x}(\bar{x})^\top$ and $B = \hat{Z}$. Since A has rank 1 and its largest eigenvalue is $\lambda_1(A) = n$, the spectral gap of A is $\delta = n > 0$. One then has, by the Davis–Kahan theorem,

$$\min_{\alpha \in \pm 1} \|u_1(A) - \alpha u_1(B)\|_2 \leq 2^{3/2} \frac{\|A - B\|_2}{\delta}.$$

where $u_1(A) = \bar{x}/\sqrt{n}$ and $u_1(B) = \hat{x}/\sqrt{n}$. In other words,

$$\min_{\alpha = \pm 1} \|\alpha \hat{x} - \bar{x}\|_2 \leq 2^{3/2} \sqrt{n} \left\| \hat{Z} - \bar{x}(\bar{x})^\top \right\|_2 \leq 2^{3/2} \sqrt{n\epsilon}$$

where we have used Theorem 5.1 in the last inequality. ∎

5.1 Outline of the proof of Theorem 5.1

We write

$$\mathcal{M}_{opt} = \left\{ Z \in \mathbb{R}^{n\times n} : Z \succeq 0, \operatorname{diag}(Z) \preceq I_n \right\}$$

for the constraint set of the SDP procedure (33). We want to show that

$$\hat{Z} \in \operatorname*{argmax}_{Z\in\mathcal{M}_{opt}} \langle A - \lambda J_n, Z\rangle$$

is close to $\bar{Z} = \bar{x}(\bar{x})^\top$.

We shall proceed in 3 steps:

1) one shows that $\bar{Z} \in \operatorname{argmax}_{Z\in\mathcal{M}_{opt}} \langle \mathbb{E}(A - \lambda J_n), Z\rangle$ by modifying very slightly the proof of Proposition 4.5, since $\mathbb{E}\lambda J_n = [(p+q)/2]J_n$ is almost equal to $\alpha u_1 \otimes u_1 = [(p+q)/2 + (1-p)/n]J_n$.

2) One shows that $(A - \lambda J_n)$ is close to $\mathbb{E}(A - \lambda J_n)$ uniformly on $\mathcal{M}_{opt}$, that is, with probability $1 - e^3 5^{-n}$,

$$\sup_{Z\in\mathcal{M}_{opt}} \left| \langle (A - \lambda J_n) - \mathbb{E}(A - \lambda J_n), Z\rangle \right| \le c_0 \epsilon \tag{34}$$

where c_0 is an absolute constant.

3) One shows a curvature inequality for the objective function at the optimum: for every $Z \in \mathcal{M}_{opt}$,

$$\langle \mathbb{E}(A - \lambda J_n), \bar{Z}\rangle - \langle A - \lambda J_n, Z\rangle \ge c_1 \left\| Z - \bar{Z} \right\|_1 \tag{35}$$

The two important points are **2)** and **3)**. The whole probabilistic aspect of the problem lies in point **2)**. We begin by developing that point.

5.2 Probabilistic aspect of the proof

In this section we prove point **2)** of the outline of the proof of Theorem 5.1. This argument relies on Grothendieck's inequality, which we first recall in its general form and then in its matrix form.

Theorem 5.4 (Grothendieck's inequality)**.** *Let $B = (b_{ij})_{i,j}$ be a matrix in $\mathbb{R}^{p\times q}$. Assume that*

$$\left| \sum_{i,j} b_{ij} s_i t_j \right| \le 1$$

for every $s_i, t_j \in \{-1, 1\}$. Then, for every Hilbert space H and all vectors $u_i, v_j \in H$ such that $\|u_i\|_2 \le 1$, $\|v_j\|_2 \le 1$, one has

$$\left| \sum_{i,j} b_{ij} \langle u_i, v_j\rangle \right| \le K_G$$

where $K_G \le 1.783$.

Theorem 5.4 is the best-known form of Grothendieck's inequality. There exist multiple formulations and generalizations, such as Nesterov's theorem, which is widely used to find approximate solutions of combinatorial problems by means of convex relaxations leading to SDPs. The form of Grothendieck's inequality that we shall use is a matrix reformulation of Theorem 5.4 in the square case $p = q = n$ and for $H = \mathbb{R}^n$. We write

$$\mathcal{M}_1 = \{st^\top : s, t \in \{-1, 1\}^n\} \text{ and } \mathcal{M}_G = \{XY^\top \in \mathbb{R}^{n\times n} : \text{ rows } X_{i\bullet}, Y_{i\bullet} \in B_2^n\}$$

where B_2^n is the Euclidean unit ball of $\mathbb{R}^n$. For every $B \in \mathbb{R}^{n\times n}$, one has

$$\sum_{i,j} b_{ij} s_i t_j = \langle B, st^\top \rangle \text{ and } \langle B, XY^\top \rangle = \sum_{i,j} b_{ij} \langle X_{i\bullet}, Y_{j\bullet} \rangle$$

Clearly, $\mathcal{M}_1 \subset \mathcal{M}_G$, so

$$\sup_{Z\in\mathcal{M}_1} \langle B, Z \rangle \leq \sup_{Z\in\mathcal{M}_G} \langle B, Z \rangle.$$

Grothendieck's inequality shows that the reverse inequality is also true up to a constant.

Corollary 5.5. *There exists an absolute constant $K_G \leq 1.783$ such that for every matrix $B \in \mathbb{R}^{n\times n}$, one has*

$$\sup_{Z\in\mathcal{M}_G} \langle B, Z \rangle \leq K_G \sup_{Z\in\mathcal{M}_1} \langle B, Z \rangle.$$

One therefore has, by Grothendieck's inequality, that for every $B \in \mathbb{R}^{n\times n}$,

$$\sup_{Z\in\mathcal{M}_1} \langle B, Z \rangle \leq \sup_{Z\in\mathcal{M}_G} \langle B, Z \rangle \leq K_G \sup_{Z\in\mathcal{M}_1} \langle B, Z \rangle.$$

One can write the supremum over $\mathcal{M}_1$ as an operator norm, because for every $B \in \mathbb{R}^{n\times n}$

$$\sup_{Z\in\mathcal{M}_1} \langle B, Z \rangle = \sup_{s,t\in B_\infty^n} \langle B, st^\top \rangle = \sup_{s,t\in B_\infty^n} \langle Bt, s \rangle = \sup_{t\in B_\infty^n} \|Bt\|_1 = \|B\|_{\infty\to 1}.$$

For the problem we seek to solve—controlling the supremum of an empirical process indexed by $\mathcal{M}_{opt}$—we look only at the symmetric positive semidefinite matrices in $\mathcal{M}_G$, and one can prove the following result.

Proposition 5.6. *One has $\mathcal{M}_{opt} \subset \mathcal{M}_G$.*

Proof. If $Z \in \mathcal{M}_{opt}$, since $Z \succeq 0$, there exists $X \in \mathbb{R}^{n\times n}$ such that $Z = XX^\top$. Moreover, $\operatorname{diag}(Z) \preceq I_n$. Now $\operatorname{diag}(Z) = \operatorname{diag}(\|X_{1\bullet}\|_2^2, \cdots, \|X_{n\bullet}\|_2^2)$, so $\|X_{i\bullet}\|_2^2 \leq 1$ for every $i = 1, \ldots, n$. Hence $Z \in \mathcal{M}_G$. ∎

One deduces from Grothendieck's inequality and from Proposition 5.6 that

$$\begin{aligned}\sup_{Z\in\mathcal{M}_{opt}} \left|\langle (A - \lambda J_n) - \mathbb{E}(A - \lambda J_n), Z \rangle\right| &\leq K_G \sup_{Z\in\mathcal{M}_1} \left|\langle (A - \lambda J_n) - \mathbb{E}(A - \lambda J_n), Z \rangle\right| \\ &\leq K_G \|A - \mathbb{E}A\|_{\infty\to 1} + K_G |\lambda - \mathbb{E}\lambda| \, \|J_n\|_{\infty\to 1}. \end{aligned} \tag{36}$$

It therefore remains to bound $\|A - \mathbb{E}A\|_{\infty\to 1}$ and $|\lambda - \mathbb{E}\lambda|$ with high probability. To this end, we shall use Bernstein's concentration inequality followed by a "union bound". We recall Bernstein's inequality here.

Theorem 5.7. *Let $Z_1, \ldots, Z_m$ be independent centered random variables such that $|Z_i| \leq b$ for every $i = 1, \ldots, m$, almost surely. Write $\sigma^2 = (1/m)\sum_{i=1}^m \mathbb{E}Z_i^2$. For every $t > 0$,*

$$\mathbb{P}\left[\frac{1}{m}\sum_{i=1}^m Z_i \geq t\right] \leq \exp\left(\frac{-mt^2}{2\sigma^2 + 2bt/3}\right).$$

Bernstein's inequality and the union bound give the following result.

Proposition 5.8. *Set $\bar{p} = [2/(n(n-1))]\sum_{i<j} \operatorname{var}(A_{ij})$. If $\bar{p} > 9/n$, then with probability at least $1 - e^3 5^{-n}$, $\|A - \mathbb{E}A\|_{\infty\to 1} \leq 6n(n-1)\sqrt{\bar{p}/n}$.*

Proof. One has

$$\|A - \mathbb{E}A\|_{\infty\to 1} = \max_{s_i,t_j\in\{-1,1\}} \sum_{i,j} (A_{ij} - \mathbb{E}A_{ij}) s_i t_j.$$

Let $s, t \in \{-1,1\}^n$. Since $A_{ii} = 1$ and $A_{ij} = A_{ji}$, one has

$$\sum_{i,j} (A_{ij} - \mathbb{E}A_{ij}) s_i t_j = \sum_{i<j} (A_{ij} - \mathbb{E}A_{ij})(s_i t_j + s_j t_i)$$

■

6 Statistical analysis of an SDP relaxation: proof of exact recovery by construction of a dual certificate

In this section, we propose to show that it is possible to recover exactly the membership vector by SDP relaxation in the case of two communities of equal size in the stochastic block model, and under conditions on the "strength" of the signal (that is, for p and q sufficiently large). To prove such a result we shall use the technique of **construction of a dual certificate**. Before implementing this technique, we recall a result on SDP optimization problems.

Weak duality and dual certification for SDP problems. We consider a problem of SDP type: let $C \in \mathbb{R}^{n\times n}$, $A_1 \ldots, A_m \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^m$. We write $\mathcal{A} : X \in \mathbb{R}^{n\times n} \to (\langle A_i, X\rangle)_{i=1}^m$ and $\mathcal{A}^\top : \lambda \in \mathbb{R}^m \to \sum_i \lambda_i A_i$ for its adjoint operator. An SDP problem is of the form

$$\min\left(\langle C, X\rangle : X \succeq 0, \mathcal{A}(X) = b\right) \tag{37}$$

where we recall that $X \succeq 0$ means that X is positive semidefinite. The set of symmetric positive semidefinite matrices is a nonempty closed convex cone. One thus recovers a problem of conic LP type. The associated Lagrangian is

$$\mathcal{L} : (X, (\lambda, Y)) \in \mathbb{R}^{n\times n} \times \mathbb{R}^m \times \mathcal{S}_n^+ \to \langle C, X\rangle + \langle X, -Y\rangle + \langle \lambda, b - \mathcal{A}(X)\rangle$$

which has dual function

$$\psi : (\lambda, Y) \in \mathbb{R}^m \times \mathcal{S}_n^+ \to \begin{cases} -\infty & \text{if } C - \mathcal{A}^\top(\lambda) - Y \neq 0 \\ \langle \lambda, b\rangle & \text{otherwise.} \end{cases}$$

The associated dual problem is then

$$\max\left(\langle \lambda, b\rangle : C - \mathcal{A}^\top(\lambda) \succeq 0\right). \tag{38}$$

Weak duality says that for every $\lambda \in \mathbb{R}^m$ such that $C - \mathcal{A}^\top(\lambda) \succeq 0$ and for every $X \succeq 0$ such that $\mathcal{A}(X) = b$, one has $\langle C, X\rangle \geq \langle \lambda, b\rangle$. For the construction of a dual certificate, one needs only weak duality: it is indeed this inequality that allows one to say that a λ^* feasible for the dual problem and such that $\langle C, X^*\rangle = \langle \lambda^*, b\rangle$ certifies X^* as a solution of the primal problem, that is, here of (37) (provided one has indeed taken X^* to be feasible for the primal). However, a dual certificate can exist only if there is strong duality; indeed, if $\langle C, X^*\rangle = \langle \lambda^*, b\rangle$ for λ^* feasible for the dual and X^* feasible for the primal, then there is strong duality. It is therefore useful to know in advance (even if it is not necessary to prove it) whether there is strong duality or not. We therefore recall a strong-duality result for SDP problems. In the following theorem, we thus identify a situation in which one can have the equality $\langle C, X^*\rangle = \langle \lambda^*, b\rangle$ for a primal solution X^* and a dual solution λ^*.

Theorem 6.1. *Assume that there exists $X_0 \succ 0$ such that $\mathcal{A}(X_0) = b$ and a $\lambda_0 \in \mathbb{R}^m$ such that $C - \mathcal{A}^\top(\lambda_0) \succ 0$. Then the primal SDP problem (37) admits a solution, the dual SDP problem (38) admits a solution, there is strong duality, and for every primal solution X^* and every dual solution λ^*, one has $\langle C, X^* \rangle = \langle b, \lambda^* \rangle$.*

This type of theorem is classical in convex optimization. The "strict feasibility" assumption is sometimes called Slater's condition. It is in general "easy" to check, and can therefore assure us, or not, of the existence of a dual certificate before one embarks on its construction. However, as indicated above, the result one needs for the construction of a dual certificate is in fact more a converse to Theorem 6.1, and in fact requires only weak duality in order to be proved (the proof of Theorem 6.1 is more complicated).

Proposition 6.2. *If X^* is feasible for the primal (37) and λ^* is feasible for the dual (38), and they are such that $\langle C, X^* \rangle = \langle b, \lambda^* \rangle$, then X^* is a solution of the primal and λ^* is a solution of the dual.*

Preuve. By weak duality, for every X feasible for the primal and every λ feasible for the dual, one has $\langle X, C \rangle \geq \langle \lambda, b \rangle$. In particular, for every X feasible for the primal, $\langle X, C \rangle \geq \langle \lambda^*, b \rangle = \langle X^*, C \rangle$, and hence, since X^* is feasible for the primal, X^* is a solution of the primal. Likewise, for every λ feasible for the dual, one has $\langle b, \lambda \rangle \leq \langle C, X^* \rangle = \langle b, \lambda^* \rangle$, and hence λ^*, being feasible for the dual, is a solution of the dual. ∎

It is Proposition 6.2 that makes it possible to develop the method of the **dual certificate**: indeed, given a point X^* feasible for the primal, if one can find a λ^* feasible for the dual problem (that is, such that $C - \mathcal{A}^\top(\lambda^*) \succeq 0$) such that $\langle C, X^* \rangle = \langle b, \lambda^* \rangle$, then necessarily X^* is a solution of the primal. This idea is very practical when one believes that an element X^* feasible for the primal problem (that is, such that $X^* \succeq 0$ and $\mathcal{A}(X^*) = b$) is probably a solution of the primal, and one wishes to prove it.

We now explain how we shall use Proposition 6.2 in our setting. The idea is that we have an initial problem (P_0) that we do not know how to solve in general (for example, it is NP-hard)—for instance a problem such as (26). We "convexify" this problem by convex relaxation. This gives us another problem, denoted (P), such as (28). Sometimes this relaxation yields a problem of SDP type. In that case, one may consider the dual problem (D) associated with this SDP. The goal is to show that the set of solutions of the initial problem (P_0) is also the set of solutions of the relaxed problem (P). One will thus be able to solve (P_0) by means of (P). To do so, one may take a solution of the initial problem; in general it is a unique vector x_0 (such as the community membership vector f^{V_1} that one wishes to reconstruct in community detection) and show that $X_0 = x_0 x_0^\top$ is the unique solution of the relaxed problem (P). To this end, one seeks to construct an element λ_0 feasible for the dual and such that $\langle C, X_0 \rangle = \langle b, \lambda_0 \rangle$. One will then know that X_0 is a solution of (P): one says in this case that λ_0 has certified X_0 as a solution of (P). This is the principle of the **dual certificate**. The whole technical difficulty lies in the construction of λ_0. However, one generally proceeds by "reverse engineering", starting from the fact that if λ_0 is a dual certificate, then it must both be feasible for the dual and also satisfy $\langle C, X_0 \rangle = \langle b, \lambda_0 \rangle$. One can also construct λ_0 by seeking to solve the dual problem, since one is sure that if λ_0 is feasible for the dual and $\langle C, X_0 \rangle = \langle b, \lambda_0 \rangle$, where X_0 is feasible for the primal, then necessarily λ_0 is a solution of the dual. It is this technique of construction of a dual certificate that we shall implement in order to show exact recovery by SDP for the community detection problem.

SDP relaxation. The SDP relaxation proposed in (28) is not the only one one can make. There exist many others. In this section, we introduce another one that will allow us to recover

the communities exactly.

The problem of detecting two communities is written, initially, as a problem of finding a solution of the discrete problem

$$\min_{V_1 \subset V} \left((f^{V_1})^\top L f^{V_1} : \left\| f^{V_1} \right\|_2^2 = n, \left\langle f^{V_1}, \mathbf{1} \right\rangle = 0 \right) \tag{39}$$

(see (6)), where $f^{V_1} = (f_i^{V_1})_{i \in V} \in \mathbb{R}^{|V|}$ is defined for every $i \in V$ by

$$f_i^{V_1} = \begin{cases} \sqrt{\frac{|V_1{}^c|}{|V_1|}} & \text{if } i \in V_1 \\ -\sqrt{\frac{|V_1|}{|V_1{}^c|}} & \text{if } i \notin V_1. \end{cases}$$

We motivated the study of this problem by introducing the modularity function *RatioCut* in Section 3.2, which one wishes to minimize. One can also motivate this problem as a maximum likelihood problem, as introduced in Section 4.1, in the case where the two communities one seeks to detect have the same size. Nevertheless, in our stochastic setting the matrix L that one observes is not the "true" Laplacian matrix of the "true" underlying graph: one has only observed a noisy version (coming from the SBM) of the "true" graph, whose "true" Laplacian is $\mathbb{E}L$. The problem one would ideally wish to solve is in fact

$$\min_{V_1 \subset V} \left((f^{V_1})^\top \mathbb{E}L f^{V_1} : \left\| f^{V_1} \right\|_2^2 = n, \left\langle f^{V_1}, \mathbf{1} \right\rangle = 0 \right) \tag{40}$$

where the Laplacian is the "theoretical" Laplacian $\mathbb{E}L$ and not the observed one. Of course, since one does not know $\mathbb{E}L$, the problem (40) is itself purely theoretical. One can nevertheless show that the community vectors one seeks to recover, that is, f^{V_1} and $f^{V_1^c}$, are the only two solutions of the problem (40). This gives another justification for considering the problem (39), which one then views as the empirical version of (40), since one has simply replaced the theoretical quantity $\mathbb{E}L$ by its empirical counterpart L.

Proposition 6.3. *If $L = D - A$ is the Laplacian of a graph drawn according to the SBM with parameter (n, p, q), where $0 < q \leq p < 1$, and according to the partition $V = V_1 \sqcup V_1^c$, then f^{V_1} and $f^{V_1^c}$ are the only two solutions of the problem (40).*

Preuve. We write $i \sim j$ when i and j are in the same community (that is, $i, j \in V_1$ or $i, j \in V_1^c$) and $i \not\sim j$ when i and j are not in the same community (that is, $[i \in V_1$ and $j \in V_1^c]$ or $[j \in V_1$ and $i \in V_1^c]$). Let $V_0 \subset V$. We write $i \sim_{V_0} j$ when $i, j \in V_0$ or $i, j \in {V_0}^c$, and $i \not\sim_{V_0} j$ otherwise. One has

$$(f^{V_0})^\top \mathbb{E}L f^{V_0} = \frac{1}{2} \sum_{(i,j) \in V \times V} \mathbb{E}A_{ij} (f_i^{V_0} - f_j^{V_0})^2 = \frac{1}{2} \sum_{i \sim j} p (f_i^{V_0} - f_j^{V_0})^2 + \frac{1}{2} \sum_{i \not\sim j} q (f_i^{V_0} - f_j^{V_0})^2$$

$$= \frac{1}{2} \left(\frac{n}{|V_0|} + \frac{n}{|V_0{}^c|} \right) \left(\sum_{i \sim j, i \not\sim_{V_0} j} p + \sum_{i \not\sim j, i \not\sim_{V_0} j} q \right) = \frac{1}{2} \left(\frac{n}{|V_0|} + \frac{n}{|V_0{}^c|} \right) \left(\sum_{i \not\sim_{V_0} j} q + \sum_{i \not\sim j, i \not\sim_{V_0} j} (p - q) \right)$$

$$\geq \frac{1}{2} \left(\frac{n}{|V_0|} + \frac{n}{|V_0{}^c|} \right) (2q |V_0| |V_0{}^c|) = n^2 q$$

because $p \geq q$. One has equality above if and only if the term

$$\sum_{i \not\sim j, i \not\sim_{V_0} j} (p - q)$$

vanishes. This occurs if and only if the sets $\{(i,j) : i \sim j\}$ and $\{(i,j) : i \not\sim_{V_0} j\}$ are disjoint, that is, when $V_0 = V_1$ or $V_0 = V_1^c$. Moreover, one also has

$$(f^{V_1})^\top \mathbb{E}L f^{V_1} = (f^{V_1^c})^\top \mathbb{E}L f^{V_1^c} = \frac{q}{2} \sum_{i \not\sim j} \left(\frac{n}{|V_1|} + \frac{n}{|V_1^c|} \right) = n^2 q.$$

Hence f^{V_1} and $f^{V_1^c}$ are indeed the only two solutions of the problem (40). ∎

Proposition 6.3 tells us that the optimization problem has as its unique solution the membership vector (up to sign) f^{V_1}, which is indeed the vector one wishes to find when one wants to solve the community detection problem in a two-class SBM.

Problem (6.3) is, however, not available to us, since $\mathbb{E}L$ is unknown to us: one observes only L. Problem (39) is therefore the one we must consider; in it one has only replaced the unknown quantity $\mathbb{E}L$ by the observed one, L. We turn to the effective resolution of (39), which will rely on a convex relaxation.

The practical difficulty with (39) is that it is NP-hard in general; one may therefore consider it from a theoretical viewpoint, but not a practical one. We shall then perform a convex relaxation of this problem in order to obtain an optimization problem that can be solved (at least approximately) numerically. There are therefore two difficulties in the problem (39) that we now consider:

- one does not have access to the "true" graph, but only to a "noisy" version (certain inter-community links appear) and a "masked" version (certain intra-community links are not given): one may speak of partial information when one observes L rather than $\mathbb{E}L$,
- the problem one wishes to solve is NP-hard in general: this is a difficulty of a computational nature.

We shall nevertheless show that a convex relaxation of (39) makes it possible to resolve these two problems: we shall indeed show that with high probability f^{V_1} and $f^{V_1^c}$ will be the only two solutions of an SDP problem constructed from L, the observed Laplacian (and not $\mathbb{E}L$). The convex relaxation that we shall implement here to solve this problem is based on the techniques of matrix lifting and SDP relaxation (see Section 7 for other examples), and the statistical analysis that we shall carry out to show the exact recovery properties of this procedure is based on the construction of a dual certificate.

We carry out this work in the case of the detection of two communities of equal size. In this case, problem (39) may be written equivalently as

$$\max \left(x^\top A x : \langle x, 1 \rangle = 0, x \in \{-1,1\}^n \right) \tag{41}$$

(see Remark 4.1), where A is the observed adjacency matrix. We note in passing that (41) is also the maximum likelihood estimator in the SBM obtained in Section 4.1, in the case of the detection of two communities of equal size. It is the problem (41) that we shall now relax.

The matrix lifting approach, as presented in the first two examples of Section 7, consists in rewriting the functions quadratic in x as linear functions of $xx^\top$. In this scheme, linear constraints such as "$\langle x, 1 \rangle = 0$" are not handled. One can, however, also handle linear terms (present either in the constraint or in the objective function), but instead of lifting only x to $xx^\top$, one must lift $\binom{1}{x}$ to $\binom{1}{x}\begin{pmatrix}1 & x^\top\end{pmatrix}$. We refer to the third paragraph of Section 7, where this approach, which handles linear terms, is proposed. In this section, we shall simply remove this linear constraint, but beforehand we shall include it as a penalization term in the objective function.

If one removes the constraint "$\langle x, 1\rangle = 0$" in (41), one sees that $\mathbf{1} = (\mathbf{1})_{\mathbf{i=1}}^{\mathbf{n}}$ (and its opposite) is a solution of the problem. This is a solution one wishes to avoid, because it proposes no nontrivial community structure, but only that consisting of the whole graph. We shall then introduce a penalization term in the objective function in order to avoid this solution; we then consider an objective function of the form $x^\top Ax - \lambda\langle x, \mathbf{1}\rangle^2$, where $\lambda \geq 0$ is the regularization parameter. We shall also add a term $\lambda \|x\|_2^2$, which changes nothing, since it is constant on the vectors of $\{-1, 1\}^n$, and which will allow us to write the objective function easily (this term is of no importance and in fact depends only on the convention one chooses for the adjacency matrix on the diagonal terms; here we have chosen to set $A_{ii} = 1$, but one could have made another choice, such as putting 0). Since $\langle x, \mathbf{1}\rangle^2 = x^\top \mathbf{1}\mathbf{1}^\top x$ and $\|x\|_2^2 = x^\top x$, one has, for $\lambda = 1/2$ (and up to multiplying the objective function by 2), that $2x^\top Ax - \langle x, \mathbf{1}\rangle^2 - \|x\|_2^2 = x^\top(2A - \mathbf{1}\mathbf{1}^\top - I_n)x$. One then introduces the matrix

$$B = 2A - \mathbf{1}\mathbf{1}^\top - I_n \text{ that is, } B_{ij} = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{if } (i,j) \in V \\ -1 & \text{if } (i,j) \notin V \end{cases} \tag{42}$$

One then considers the following regularized problem:

$$\max\left(x^\top Bx : x \in \{-1, 1\}^n\right). \tag{43}$$

In particular, $\mathbf{1}$ is no longer a solution of this problem, thanks to the penalization $x \to -\langle x, \mathbf{1}\rangle^2$ added to the objective function, which penalizes $\mathbf{1}$ the most, because it equals $-n^2$ at this vector. The regularization will therefore disqualify partitions that propose communities of unbalanced size.

We apply the matrix lifting method to problem (43). We rewrite the objective function, which is quadratic in x, as a linear function of $Z = xx^\top$: $x^\top Bx = \langle B, xx^\top\rangle = \langle B, Z\rangle$, and we rewrite the constraints, which are also quadratic in x, as linear constraints on Z: $x_i^2 = 1$ is equivalent to $Z_{ii} = 1$. We then consider the "lifted" problem of (43), followed by an SDP relaxation (one removes the rank constraint on Z), in order to obtain

$$\max\left(\langle B, Z\rangle : Z \succeq 0, Z_{ii} = 1, \forall i \in \{1, \ldots, n\}\right). \tag{44}$$

One indeed obtains a problem of SDP type. To make the link between a solution of (44) and the original problem (43), one may either take a leading eigenvector of a solution of (44) and cluster into two clusters the entries of this vector, or take $G \sim \mathcal{N}(0, \hat{Z})$, where $\hat{Z}$ is a solution of (44), and take the sign of the coordinates of G—this latter method is called matrix rounding and is used, for example, to prove Grothendieck's inequality. Our objective in what follows is to show that $f^{V_1}(f^{V_1})^\top$ is the unique solution of (44), where f^{V_1} is indeed the membership vector of the communities underlying the SBM. To this end we shall construct a (strict) dual certificate certifying that $f^{V_1}(f^{V_1})^\top$ is indeed the unique solution of (44).

Construction of a dual certificate. We wish to prove that $gg^\top$, where $g := f^{V_1}$, is the unique solution of (44). To this end, we shall construct a strict dual certificate. We must then first identify the dual problem of (44). One may use the problem introduced in (38), or recover it directly.

The Lagrangian associated with (44) is

$$\mathcal{L} : (Z, (\lambda, X)) \in \mathbb{R}^{n\times n} \times (\mathbb{R}^n \times \mathcal{S}_n^+) \to -\langle B, Z\rangle + \langle -X, Z\rangle + \langle \lambda, \mathbf{1} - \mathrm{diag}(Z)\rangle$$

where $\mathcal{S}_n^+$ is the cone of symmetric positive semidefinite $n \times n$ matrices and $\mathrm{diag}(Z) = (Z_{ii})_{i=1}^n$. One can make sure that one has defined the right Lagrangian for the problem (44) by checking the equality between the original problem and the primal problem in Z associated with this Lagrangian:

$$\inf_{Z \in \mathbb{R}^{n \times n}} \max_{(\lambda, X) \in \mathbb{R}^n \times \mathcal{S}_n^+} \mathcal{L}(Z, (\lambda, X)) = \min(-\langle B, Z \rangle : Z \in \mathcal{S}_n^+, Z_{ii} = 1).$$

The dual function is

$$\psi : (\lambda, X) \in \mathbb{R}^n \times \mathcal{S}_n^+ \to \inf_{Z \in \mathbb{R}^{n \times n}} \mathcal{L}(Z, (\lambda, X)) = \begin{cases} -\infty & \text{if } -B - X - \mathrm{Diag}(\lambda) \neq 0 \\ \langle \lambda, \mathbf{1} \rangle & \text{otherwise.} \end{cases}$$

where $\mathrm{Diag}(\lambda)$ is the diagonal $n \times n$ matrix whose diagonal entries are the coordinates of λ. To write $\mathcal{L}$ and ψ we have used that the dual cone of $\mathcal{S}_n^+$ (to which the primal variable Z belongs) is $-\mathcal{S}_n^+$ (to which the dual variable $-X$ associated with the constraint "$X \succeq 0$" belongs). The dual problem is therefore $\max(\psi(\lambda, X) : (\lambda, X) \in \mathbb{R}^n \times \mathcal{S}_n^+)$, that is,

$$\max\left(\langle \lambda, \mathbf{1} \rangle : B + X + \mathrm{Diag}(\lambda) = 0, X \succeq 0\right) \tag{45}$$

which is equivalent to

$$\max\left(\langle \lambda, \mathbf{1} \rangle : -B - \mathrm{Diag}(\lambda) \succeq 0\right)$$

which is also equivalent to

$$\min\left(\mathrm{Tr}(Y) : Y - B \succeq 0, Y \text{ is diagonal}\right). \tag{46}$$

There is indeed an equivalence between (45) and (46), because if (λ, X) is a solution of (45), then $-\mathrm{Diag}(\lambda)$ is a solution of (46), and if Y is a solution of (46), then $(-\mathrm{diag}(Y), -B + Y)$ is a solution of (45).

Weak duality is given here by the following inequality: for every $Z \succeq 0$ such that $Z_{ii} = 1$ and every $\lambda \in \mathbb{R}^n$ such that $-B - \mathrm{Diag}(\lambda) \succeq 0$, one has $-\langle B, Z \rangle \geq \langle \lambda, \mathbf{1} \rangle$. This may be rewritten in the terms of the problem (46) by: for every $Z \succeq 0$ such that $Z_{ii} = 1$ and every diagonal matrix Y such that $Y - B \succeq 0$, one has $-\langle B, Z \rangle \geq -\mathrm{Tr}(Y)$. One can recover this inequality directly by observing that

$$-\mathrm{Tr}(Y) + \langle B, Z \rangle = -\langle Y, \mathrm{Diag}(\mathbf{1}) \rangle + \langle B, Z \rangle \leq -\langle Y, \mathrm{Diag}(\mathbf{1}) \rangle + \langle Y, Z \rangle = \langle Y, Z - \mathrm{Diag}(\mathbf{1}) \rangle = 0$$

because $Y \succeq B$ and hence, since $Z \succeq 0$, one has $\langle B, Z \rangle \leq \langle Y, Z \rangle$, and also $Z - \mathrm{Diag}(\mathbf{1})$ has zero diagonal and Y is diagonal, so $\langle Y, Z - \mathrm{Diag}(\mathbf{1}) \rangle = 0$.

One deduces that in order to certify that a Z, feasible for the primal (that is, such that $Z \succeq 0$ and $Z_{ii} = 1$), is indeed a solution of the problem (44), it suffices to construct a matrix $Y \in \mathbb{R}^{n \times n}$ such that:

a) Y is diagonal and $Y \succeq B$ (that is, Y is feasible for the dual)

b) $\langle B, Z \rangle = \mathrm{Tr}(Y)$ (that is, the primal and dual objective functions coincide at Z and Y respectively), which is equivalent to $\langle B - Y, Z \rangle = 0$, because Y is diagonal and Z has a diagonal of ones.

In this case Y is called a dual certificate of Z for the problem (44).

We wish to show that $gg^\top$ is the unique solution of (44). It then suffices to construct a diagonal matrix $Y \in \mathbb{R}^{n \times n}$ such that $Y \succeq B$ and $\langle B - Y, gg^\top \rangle = 0$. This will prove that $gg^\top$ is indeed a solution of (44). For uniqueness afterwards, it suffices to prove that the kernel of $B - Y$ (where Y is the dual certificate of $gg^\top$ that we shall construct below) has dimension 1. This will indeed prove that $gg^\top$ is the unique solution, by the following result.

Lemma 6.4. *Let Y be a dual certificate of $gg^\top$ for the problem (44). If the kernel of $B - Y$ has dimension 1, then $gg^\top$ is the unique solution of (44).*

Preuve. Let Z be a solution of (44). We write the eigenvalue decomposition of Z: one has $Z = \sum_i \lambda_i u_i u_i^\top$, where $0 \leq \lambda_1 \leq \ldots \leq \lambda_n$ and $(u_1, \ldots, u_n)$ is an orthonormal basis of $\mathbb{R}^n$. By Proposition 6.9, we know that Y is also a certificate for Z: $\langle B - Y, Z \rangle = 0$. One therefore has $\sum_i u_i^\top (B - Y) u_i = 0$, and since $Y \succeq B$, one has $u_i^\top (Y - B) u_i \geq 0$ for every i. One therefore deduces that $[u_i^\top (Y - B) u_i = 0$ or $\lambda_i = 0]$, and hence $[(Y - B)u_i = 0$ or $\lambda_i = 0]$, that is, u_i lies in the kernel of $B - Y$, or $\lambda_i = 0$. But $(B - Y)g = 0$ (because $\langle B - Y, gg^\top \rangle = 0$) and the kernel of $B - Y$ has dimension 1, so $\mathrm{Ker}(B - L) = \mathrm{vect}(g)$. One then deduces that u_i lies in $\mathrm{vect}(g)$, or $\lambda_i = 0$. Since $(u_1, \ldots, u_n)$ is an orthonormal basis and the kernel has dimension 1, one deduces that only one of the u_i can lie in the kernel, and then it is proportional to g, and for all the others one has $\lambda_j = 0$. Hence Z is of the form $Z = \lambda gg^\top$. Moreover, $Z_{ii} = 1 = g_i g_i$, so $\lambda = 1$, that is, $Z = gg^\top$. Hence $gg^\top$ is indeed the unique solution of (44). ∎

A dual certificate that makes it possible to certify the uniqueness of a solution is called a **strict dual certificate**. This is what we wish to construct here for $gg^\top$. To summarize, we wish to construct a matrix $Y \in \mathbb{R}^{n \times n}$ such that

a) Y is diagonal and $Y \succeq B$

b) $(B - Y)g = 0$

c) $\lambda_2(Y - B) > 0$, that is, the second smallest eigenvalue of $Y - B$ is strictly positive (and hence its kernel has dimension 1).

Conditions *a)* and *b)* do not leave us much choice: one has $Bg = Yg$, and since Y is diagonal one must necessarily take $Y_{ii} = (Bg)_i / g_i$ for every $i \in \{1, \ldots, n\}$. Moreover, one has $Bg = (2A - \mathbf{1}\mathbf{1}^\top - I_n)g = 2Ag - g$, because $\langle \mathbf{1}, g \rangle = 0$, and hence for every $i, j \in \{1, \ldots, n\}$,

$$Y_{ii} = \frac{2(Ag)_i}{g_i} - 1 \text{ and } Y_{ij} = 0 \text{ when } i \neq j. \tag{47}$$

It remains to show that $Y \succeq B$. We know that $(Y - B)g = 0$ by construction of g, so if one proves that the second smallest eigenvalue of $Y - B$ is strictly positive, one will be done, because this will prove both that $Y \succeq B$ and point *c)* (by construction Y is diagonal and $Bg = Yg$). One may therefore establish the following result.

Proposition 6.5. *If the second smallest eigenvalue of $Y - B$ is strictly positive, then $gg^\top$ is the unique solution of the SDP problem (44).*

We denote by $\lambda_2(Y - B)$ the second smallest eigenvalue of $Y - B$. By Proposition 6.5, if one has $\lambda_2(Y - B)$, then $gg^\top$ is the unique solution of (44), and hence by looking for the leading eigenvector of $gg^\top$ one will obtain $\pm g / \|g\|_2$, and one will therefore be able to deduce the communities $V_1 = \{i : gi > 0\}$ and $V_i^c = \{g_i < 0\}$. It then remains to prove that $\lambda_2(Y - B) > 0$.

Probabilistic aspect: lower bound on $\lambda_2(Y - B)$ in the SBM. To show that $\lambda_2(Y - B) > 0$, it is at this stage that one uses the random aspect of the problem. We shall therefore show that if A is the adjacency matrix of an SBM with parameter (n, p, q), where $p = \alpha(\log n)/n$ and $q = \beta(\log n)/n$ with $\alpha > \beta$, one will have $\lambda_2(Y - B) > 0$ under a certain condition on α and β that one hopes is not too far from the optimal phase transition given in Theorem 4.3, that is, $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$.

The strategy we follow to prove that $\lambda_2(Y - B) > 0$ is based on a matrix concentration argument. We shall prove the following two results:

1) $\lambda_2(\mathbb{E}(Y - B)) := \square$

2) with high probability $\|(Y - B) - \mathbb{E}(Y - B)\|_{op} < \square$

where $\square = (\alpha - \beta)\log n$ and $\|\cdot\|_{op}$ is the operator norm. These two results will be sufficient to prove that $\lambda_2(Y - B) > 0$, since by the Courant–Fisher formula (in particular, its corollary on the interlacing properties of singular values), one has for every symmetric positive semidefinite $n \times n$ matrices A and B that

$$\max_{1 \leq i \leq n} |\lambda_i(A) - \lambda_i(B)| \leq \|A - B\|_{op}$$

for $0 \leq \lambda_1(A) \leq \cdots \leq \lambda_n(A)$ the eigenvalues of A in increasing order (and likewise for B). One therefore sees that if *1)* and *2)* hold, then, with high probability,

$$\lambda_2(Y - B) \geq \lambda_2(\mathbb{E}(Y - B)) - \|(Y - B) - \mathbb{E}(Y - B)\|_{op} > 0.$$

It then remains to prove points *1)* and *2)*. We begin with the first point.

For the computation of the second smallest eigenvalue of $\mathbb{E}(Y - B)$, one uses the definitions of B and Y (see (42) and (47)) to obtain, for every $i \in V$,

$$\mathbb{E}(Ag)_i = \begin{cases} \left(1 - p + \frac{n(p-q)}{2}\right) & \text{if } i \in V_1 \\ -\left(1 - p + \frac{n(p-q)}{2}\right) & \text{if } i \notin V_1 \end{cases} \quad \text{and hence } \mathbb{E}Y = (1 - 2p + n(p - q))\, I_n$$

and

$$\mathbb{E}B = (1 - 2p)I_n + (p - q)\, gg^\top + (p + q - 1)\, \mathbf{1}\mathbf{1}^\top.$$

One then has

$$\mathbb{E}(Y - B) = n(p - q)I_n - (p - q)\, gg^\top + (1 - (p + q))\, \mathbf{1}\mathbf{1}^\top.$$

One indeed recovers that $g^\top \mathbb{E}(Y - B)g = 0$ (one constructed Y so that $Yg = Bg$), so g is an eigenvector associated with the eigenvalue 0. The other eigenvectors are to be sought in $\mathrm{vect}(g)^\top$: let $v \in \mathbb{R}^n$ be such that $\|v\|_2 = 1$ and $\langle v, g \rangle = 0$; one has

$$v^\top \mathbb{E}(Y - B)v = n(p - q) + (1 - (p + q))\, \langle \mathbf{1}, v \rangle^2.$$

Since we place ourselves in the regime where $p = \alpha \log(n)/n$ and $q = \beta \log(n)/n$, one may assume that $1 \geq p + q$. Thus $v^\top \mathbb{E}(Y - B)v \geq n(p - q)$, and this bound is attained by a v orthogonal to $\mathbf{1}$. One therefore has that

$$\lambda_2(\mathbb{E}(Y - B)) = n(p - q) = (\alpha - \beta)\log n \tag{48}$$

(one sees that g is an eigenvector associated with the smallest eigenvalue 0, a v such that $\langle v, g \rangle = \langle v, \mathbf{1} \rangle = 0$ is a Fiedler vector of $\mathbb{E}(Y - B)$ associated with the eigenvalue $n(p - q)$, there are $n - 2$ of them, and $\mathbf{1}/\sqrt{n}$ is an eigenvector associated with the eigenvalue $n(1 - 2q)$).

We turn to point *2)*. We wish to show that $\|(Y - B) - \mathbb{E}(Y - B)\|_{op} < (\alpha - \beta)\log n$ with high probability when $p = \alpha \log(n)/n$ and $q = \beta \log(n)/n$ (and $p + q < 1$). This is a concentration result for the matrix $Y - B$ about its mean with respect to the operator norm that we wish to prove here. To prove such a result, we shall use a concentration result on the largest eigenvalue of an average of symmetric independent random matrices, which we call the **matrix Bernstein inequality** and which we now state.

Theorem 6.6. *Let $(X_k)_{k=1}^n$ be independent symmetric random matrices of size $d \times d$. Assume that for every $k = 1, \dots, n$, one has*

$$\mathbb{E}X_k = 0 \text{ and } \lambda_{max}(X_k) \leq R \text{ a.s.}.$$

One then has, for every $t > 0$, with probability at least $1 - d\exp(-t)$,

$$\lambda_{max}\left(\frac{1}{n}\sum_{k=1}^n X_k\right) \leq \sigma\sqrt{\frac{2t}{n}} + \frac{2Rt}{3n}$$

where

$$\sigma^2 = \left\|\frac{1}{n}\sum_{k=1}^n \mathbb{E}(X_k^2)\right\|_{op}$$

and $\|\cdot\|_{op}$ is the operator norm and $\lambda_{max}(\cdot)$ is the largest eigenvalue of a symmetric matrix. One obtains a bound on the operator norm of $(1/n)\sum_{k=1}^n X_k$ by applying this result to the $-X_k$ and observing that $\|M\|_{op} = \max(\lambda_{max}(M), \lambda_{max}(-M)$ for every symmetric matrix M.

The matrix Bernstein inequality looks very much like the "classical" Bernstein inequality, that is, the one concerning the concentration of a sum of independent real random variables. One must, however, pay close attention to the presence of the dimension d in the deviation "$1 - d\exp(-t)$" in Theorem 6.6. The presence of this dimension is sometimes responsible for a logarithmic loss in the rates of convergence that one obtains by using this inequality. This will not be our case here, because we shall work at a deviation level $t = (\alpha - \beta)\log n$, which is of the order of the logarithm of the dimension, and hence the dimensional term in d in Theorem 6.6 will at worst be responsible for a loss at the level of the constants.

We shall therefore use the matrix Bernstein inequality to prove point *2)*. The first thing to do is to bring out sums of independent random matrices in $Y - B$. To this end, we shall consider a partition of the graph $G = (V, E)$ into the two subgraphs $G^{\sim} = (V, E^{\sim})$ and $G^{\not\sim} = (V, E^{\not\sim})$, where $E^{\sim}$ is the set of all the edges of E linking two points in the same community and $E^{\not\sim}$ is the set of all the edges of E linking two points of two different communities. These two subgraphs have adjacency matrices denoted respectively $A^{\sim}$ and $A^{\not\sim}$, and (diagonal) degree matrices denoted respectively $D^{\sim}$ and $D^{\not\sim}$. Formally, for every $i, j \in V$,

$$A_{ij}^{\sim} = \begin{cases} A_{ij} & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases}, \quad D^{\sim} = \mathrm{Diag}(d_1^{\sim}, \dots, d_n^{\sim}) \text{ where } d_i^{\sim} = \sum_{j:i\sim j} A_{ij} = \sum_{j\in V} A_{ij}^{\sim}.$$

and

$$A_{ij}^{\not\sim} = \begin{cases} A_{ij} & \text{if } i \not\sim j \\ 0 & \text{otherwise} \end{cases}, \quad D^{\not\sim} = \mathrm{Diag}(d_1^{\not\sim}, \dots, d_n^{\not\sim}) \text{ where } d_i^{\not\sim} = \sum_{j:i\not\sim j} A_{ij} = \sum_{j\in V} A_{ij}^{\not\sim}.$$

In particular, one has $A = A^{\sim} + A^{\not\sim}$, as represented in Figure 4.

One also has, for every $i \in V$,

$$Y_{ii} = 2g_i(Ag)_i - 1 = 2\left(\sum_{j\in V} A_{ij}g_ig_j\right) - 1 = 2\left(\sum_{j:i\sim j} A_{ij} - \sum_{j:i\not\sim j} A_{ij}\right) - 1 = 2(d_i^{\sim} - d_i^{\not\sim}) - 1$$

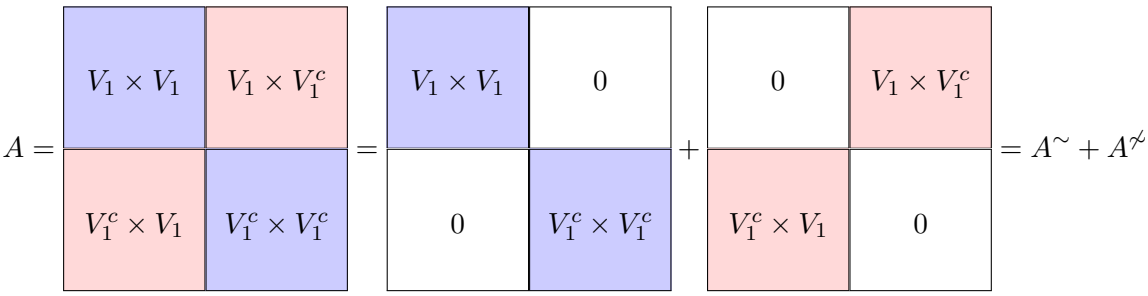


Figure 4: Decomposition of the adjacency matrix A of G into the sum of the two adjacency matrices of the subgraphs $G^{\sim}$ and $G^{\not\sim}$.

(one uses here that $g_i \in \{-1, 1\}$, so $1/g_i = g_i$, and if $i \sim j$ then $g_i g_j = 1$, otherwise $g_i g_j = -1$). One therefore has $Y = 2(D^{\sim} - D^{\not\sim}) - I_n$, and since $A = A^{\sim} + A^{\not\sim}$ and $B = 2A - \mathbf{1}\mathbf{1}^{\top} - I_n$, one obtains that

$$Y - B = 2[(D^{\sim} - A^{\sim}) - (D^{\not\sim} - A^{\not\sim})] + \mathbf{1}\mathbf{1}^{\top}.$$

One thus sees appearing the difference of the Laplacians of $G^{\sim}$ and $G^{\not\sim}$ (this difference is sometimes called the Laplacian of the SBM; one must nevertheless observe that it is not the Laplacian of a graph, since this difference is not necessarily a positive semidefinite matrix). One then writes the two Laplacians $D^{\sim} - A^{\sim}$ and $D^{\not\sim} - A^{\not\sim}$ as sums of independent symmetric random matrices.

One introduces a "basis" of symmetric matrices in which one can write $D^{\sim} - A^{\sim}$ and $D^{\not\sim} - A^{\not\sim}$ as sums of independent symmetric random matrices: for every $i, j \in V$,

$$\Delta_{ij} = E_{ii} + E_{jj} - E_{ij} - E_{ji}$$

where $E_{pq} \in \mathbb{R}^{n \times n}$ has entries 0 everywhere except at (p, q), where it equals 1. One checks that

$$D^{\sim} - A^{\sim} = \sum_{i<j} A^{\sim}_{ij} \Delta_{ij} \text{ and } D^{\not\sim} - A^{\not\sim} = \sum_{i<j} A^{\not\sim}_{ij} \Delta_{ij}.$$

Indeed, in general, if $L = D - A$ is the Laplacian of an undirected unweighted graph, one has $D - A = \sum_{i<j} A_{ij} \Delta_{ij}$. Indeed, $(D - A)$ and $\sum_{i<j} A_{ij} \Delta_{ij}$ are symmetric, so it suffices to check that the entries above the diagonal are equal. For $i < j$, one has

$$(D - A)_{ij} = -A_{ij} \text{ and } \left(\sum_{p<q} A_{pq} \Delta_{pq} \right)_{ij} = \sum_{p<q} A_{pq} (\Delta_{pq})_{ij} = -A_{ij}$$

since for $i < j$ and $p < q$ one has $(\Delta_{pq})_{ij} = 1$ if $i = p$ and $p = j$, and 0 otherwise. For every $i \in V$, $(D - A)_{ii} = \sum_{j \in V} A_{ij} - A_{ii}$ and

$$\begin{aligned}
&\left(\sum_{p<q} A_{pq} \Delta_{pq} \right)_{ii} = \sum_{p<q} A_{pq} (\Delta_{pq})_{ii} = \sum_{p<q} A_{pq} \left(I(p = i) + I(q = i) \right) \\
&= \sum_{p=1}^{n} \sum_{q=p+1}^{n} A_{pq} I(p = i) + \sum_{p=1}^{n} \sum_{q=p+1}^{n} A_{pq} I(q = i) = \sum_{q=i+1}^{n} A_{iq} + \sum_{p=1}^{i-1} A_{pi} \\
&= \sum_{p=1:p \neq i}^{n} A_{ip} = \sum_{p=1}^{n} A_{ip} - A_{ii}.
\end{aligned}$$

One then has

$$Y - B = 2[(D^{\sim} - A^{\sim}) - (D^{\not\sim} - A^{\not\sim})] + \mathbf{1}\mathbf{1}^\top = 2\left[\sum_{i<j} A^{\sim}_{ij}\Delta_{ij} - \sum_{i<j} A^{\not\sim}_{ij}\Delta_{ij}\right] + \mathbf{1}\mathbf{1}^\top$$

and hence

$$(Y - B) - \mathbb{E}(Y - B) = 2\left[\sum_{i<j:i\sim j} (A^{\sim}_{ij} - \mathbb{E}A^{\sim}_{ij})\Delta_{ij} - \sum_{i<j:i\not\sim j} (A^{\not\sim}_{ij} - \mathbb{E}A^{\not\sim}_{ij})\Delta_{ij}\right]. \tag{49}$$

One observes that summing $A^{\sim}_{ij}\Delta_{ij}$ over $i < j$ is equivalent to summing over $i < j$ such that $i \sim j$, since $A^{\sim}_{ij} = 0$ when $i \not\sim j$. The point of bringing out this summation is that the two sums in (49) are disjoint and, since $(A^{\sim}_{ij})_{i<j:i\sim j}$ and $(A^{\sim}_{ij})_{i<j:i\not\sim j}$ are two independent families of independent variables, one indeed has a sum of independent symmetric random matrices in (49), to which one can then apply the matrix Bernstein inequality: for every $t > 0$, with probability at least $1 - n\exp(-t)$,

$$\lambda_{max}\left(\sum_{i<j:i\sim j} (A^{\sim}_{ij} - \mathbb{E}A^{\sim}_{ij})\Delta_{ij} - \sum_{i<j:i\not\sim j} (A^{\not\sim}_{ij} - \mathbb{E}A^{\not\sim}_{ij})\Delta_{ij}\right) \leq \sigma\sqrt{2t} + \frac{2Rt}{3}$$

where R is such that for every $i < j$,

$$\lambda_{max}\left((A^{\sim}_{ij} - \mathbb{E}A^{\sim}_{ij})\Delta_{ij}\right) \leq R \text{ a.s. and } \lambda_{max}\left((A^{\not\sim}_{ij} - \mathbb{E}A^{\not\sim}_{ij})\Delta_{ij}\right) \leq R \text{ a.s.}$$

and

$$\sigma^2 = \left\| \sum_{i<j:i\sim j} \mathbb{E}(A^{\sim}_{ij} - \mathbb{E}A^{\sim}_{ij})^2\Delta^2_{ij} + \sum_{i<j:i\not\sim j} \mathbb{E}(A^{\not\sim}_{ij} - \mathbb{E}A^{\not\sim}_{ij})^2\Delta^2_{ij} \right\|_{op}.$$

One has $\lambda_{max}(\Delta_{ij}) = 2$, so one may take $R = 2$. Moreover $\Delta^2_{ij} = 2\Delta_{ij}$, and hence

$$\sum_{i<j:i\sim j} \Delta^2_{ij} = nI_n - (\mathbf{1}\mathbf{1}^\top + gg^\top) \text{ and } \sum_{i<j:i\not\sim j} \Delta_{ij} = nI_n - (\mathbf{1}\mathbf{1}^\top - gg^\top)$$

because $|V_1| = |V_1^c| = n/2$. One therefore has

$$\begin{aligned}
\sigma^2 &= \left\| p(1-p)[nI_n - (\mathbf{1}\mathbf{1}^\top + gg^\top)] + q(1-q)[nI_n - (\mathbf{1}\mathbf{1}^\top - gg^\top)] \right\|_{op} \\
&= \left\| [np(1-p) + nq(1-q)]I_n - [p(1-p) + q(1-q)]\mathbf{1}\mathbf{1}^\top + [q(1-q) - p(1-p)]gg^\top \right\|_{op} \qquad (50) \\
&= [q(1-q) - p(1-p)] + [np(1-p) + nq(1-q)] \leq n(p+q)
\end{aligned}$$

because $\langle \mathbf{1}, g\rangle = 0$, and hence if one completes $(\mathbf{1}/\sqrt{n}, q/\sqrt{n})$ to an orthonormal basis $(u_1 := \mathbf{1}/\sqrt{n}, u_2 := q/\sqrt{n}, u_3, \ldots, u_n)$ of $\mathbb{R}^n$, one will have $I_n = \sum_i u_i u_i^\top$. Thus $[q(1-q) - p(1-p)] + [np(1-p) + nq(1-q)]$ is the largest eigenvalue, of multiplicity 1, with eigenvector $g/\sqrt{n}$; $[np(1-p) + nq(1-q)]$ is an eigenvalue of multiplicity $(n-2)$ for the eigenvectors $u_3, \ldots, u_n$; and $[np(1-p) + nq(1-q)]I_n - [p(1-p) + q(1-q)]$ is the smallest eigenvalue, of multiplicity 1, for $\mathbf{1}/\sqrt{n}$, of the matrix appearing in (50).

One then obtains that for every $t > 0$, with probability at least $1 - 2n\exp(-t)$,

$$\|(Y - B) - \mathbb{E}(Y - B)\|_{op} \leq \sqrt{n(p+q)}\sqrt{2t} + \frac{4t}{3} = \sqrt{(\alpha+\beta)\log(n)}\sqrt{2t} + \frac{4t}{3}.$$

For example, for $t = (1+\epsilon)\log n$ where $\epsilon > 0$, one obtains with probability at least $1 - 2/n^{\epsilon}$ that $\|(Y - B) - \mathbb{E}(Y - B)\|_{op} < (\alpha - \beta)\log n$ as soon as

$$\sqrt{2(1+\epsilon)(\alpha+\beta)} + \frac{4(1+\epsilon)}{3} < \alpha - \beta.$$

We have therefore proved the following result thanks to the construction of a strict dual certificate.

Theorem 6.7. *Let A be the adjacency matrix of a graph drawn according to the SBM with parameter (n, p, q) such that $p = \alpha\log(n)/n$ and $q = \beta\log(n)/n$ for $0 < \beta < \alpha$ such that $p + q < 1$. We denote by $V_1 \sqcup V_1^c = V$ the community structure underlying this graph, and we assume that $|V_1| = |V_1^c| = n/2$. We consider the community membership vector $g \in \mathbb{R}^n$ such that $g_i = 1$ if $1 \in V_1$ and $g_i = -1$ if $i \in V_1^c$. One has, for $B = 2A - \mathbf{1}\mathbf{1}^\top - I_n$, that*

$$\{gg^\top\} = \operatorname{argmax}\left(\langle B, Z\rangle : Z \succeq 0, Z_{ii} = 1, \forall i = 1, \ldots, n\right)$$

with probability at least $1 - 2/n^{\epsilon}$ for every $\epsilon > 0$, as soon as

$$\sqrt{2(1+\epsilon)(\alpha+\beta)} + \frac{4(1+\epsilon)}{3} < \alpha - \beta.$$

Theorem 6.7 therefore ensures that it is possible to recover exactly the underlying community structure of a graph when the latter is a random graph drawn according to an SBM having inter- and intra-community probabilities according to the assumptions of this theorem. Moreover, and this is the most important point, one can do so efficiently by solving an SDP optimization problem. What is particularly remarkable is that this SDP procedure is obtained by convex relaxation (following the method of matrix lifting / SDP relaxation) of a problem that is NP-hard in general. This computational aspect of the original problem (43) is therefore particularly pessimistic, since we have just seen that for most graphs obtained by an SBM, this problem is solved in linear time.

Concerning the threshold "$\sqrt{2(1+\epsilon)(\alpha+\beta)} + \frac{4(1+\epsilon)}{3} < \alpha - \beta$" obtained in Theorem 6.7, it does not coincide with the phase transition $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$ obtained in Theorem 4.3, and one can be sure that the threshold of Theorem 6.7 is more restrictive than this phase transition (otherwise this would contradict Theorem 4.3). It is, however, possible to improve the analysis carried out above on $\lambda_2(Y - B)$ in order to recover exactly this threshold.

Theorem 6.8. *Let A be the adjacency matrix of a graph drawn according to the SBM with parameter (n, p, q) such that $p = \alpha\log(n)/n$ and $q = \beta\log(n)/n$ for $0 < \beta < \alpha$ such that $p + q < 1$. We denote by $V_1 \sqcup V_1^c = V$ the community structure underlying this graph, and we assume that $|V_1| = |V_1^c| = n/2$. We consider the community membership vector $g \in \mathbb{R}^n$ such that $g_i = 1$ if $1 \in V_1$ and $g_i = -1$ if $i \in V_1^c$. One has, for $B = 2A - \mathbf{1}\mathbf{1}^\top - I_n$, that*

$$\{gg^\top\} = \operatorname{argmax}\left(\langle B, Z\rangle : Z \succeq 0, Z_{ii} = 1, \forall i = 1, \ldots, n\right)$$

with probability at least $1 - 2$ as soon as $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$.

One therefore sees, thanks to Theorem 6.8, that it is possible to attain the phase transition in community detection by an SDP problem. One then says that **there is no computational gap for the problem of detecting two communities of equal size for random graphs drawn according to an SBM**.

Conclusion: One sees once again, on the community detection problem, that **"randomness does things well"** by placing in favorable positions problems that are NP-hard in general, which can in these cases be solved in linear time. One also sees that the convex relaxation of a problem, even an NP-hard one, is a good thing to do, whether by computing convex envelopes of nonconvex objective functions (such as the "ℓ_0 norm" in Compressed Sensing, or the rank function in matrix completion) or by convexification of the constraint (as for the spectral method in community detection) or by matrix lifting / SDP relaxation (for the SDP method in community detection).

Dual certification in Compressed Sensing. The construction of a dual certificate is not specific to the SDP problem. In fact, one can apply this approach to any problem as soon as one believes that there is strong duality (without strong duality, a dual certificate does not exist). One can, for example, use this approach for the Basis Pursuit problem that we encountered at the beginning of the course in Compressed Sensing. It is this method that we develop in this paragraph.

The optimization problem posed for the construction of the procedure called "basis pursuit" is the following:

$$\min\left(\|t\|_1 : At = y\right) \tag{51}$$

where $A \in \mathbb{R}^{m\times N}$ and $y \in \mathbb{R}^m$ are respectively the measurement matrix and the vector of measurements.

We propose to determine what a dual certificate is for the problem (51), in two ways. In the first approach, one brings out a subdifferential. We write $\Omega = \{t \in \mathbb{R}^N : At = y\}$ and i_Ω for the indicator (in the "optimization" sense) of Ω, that is, for every $x \in \mathbb{R}^N$, $i_\Omega(x) = +\infty$ if $x \notin \Omega$ and $i_\Omega(x) = 0$ when $x \in \Omega$. One then sees that solving (51) is equivalent to finding a minimum over $\mathbb{R}^N$ of $f : x \in \mathbb{R}^N \to \|x\|_1 + i_\Omega(x)$, which is itself equivalent to finding an $x^* \in \mathbb{R}^N$ such that $0 \in \partial^- f(x^*)$, where $\partial^- f(x)$ is the subdifferential of f at x given by

$$\partial^- f(x) = \{g \in \mathbb{R}^N : f(x+h) \geq f(x) + \langle g, h\rangle, \forall h \in \mathbb{R}^N\}.$$

For every $x \in \mathbb{R}^N$ with support $I = \mathrm{supp}(x)$, one has

$$\partial^- \|\cdot\|_1 (x) = \{g \in \mathbb{R}^N : g_I = \mathrm{sgn}(x_I), \|g\|_\infty \leq 1\}$$

and $\partial^- i_\Omega(x) = \mathrm{Im}(A^\top)$.

Thus $0 \in \partial^- f(x^*)$ if and only if $\partial^- i_\Omega(x^*) \cap \partial^- \|\cdot\|_1 (x^*) \neq \emptyset$ (one uses here that $\partial^- f(x^*) = \partial^- i_\Omega(x^*) + \partial^- \|\cdot\|_1 (x^*)$), that is, if there exists $\beta^* \in \mathbb{R}^m$ such that $A^\top\beta^* \in \partial^- \|\cdot\|_1 (x^*)$, that is, such that $\left\|A^\top\beta^*\right\|_\infty \leq 1$ and $(A^\top\beta^*)_I = \mathrm{sgn}(x_I)$. This vector β^* is a dual certificate: given an $x^* \in \mathbb{R}^N$ such that $Ax^* = y$, if one can construct a $\beta^* \in \mathbb{R}^m$ such that $\left\|A^\top\beta^*\right\|_\infty \leq 1$ and $(A^\top\beta^*)_I = \mathrm{sgn}(x_I)$, then one will have certified that x^* is a solution of (51).

One can see that β^*, as introduced above, is indeed a dual certificate in the sense in which we introduced it above for SDP problems: that is, as a feasible point of the dual problem for which the primal and dual objective functions coincide respectively at x^* and β^*. It is this construction that makes it possible to show the optimality of an x^* on the basis of the construction of a

dual certificate. The first thing to do is to determine the dual problem of (51), and hence the Lagrangian:

$$\mathcal{L} : \left\{ \begin{array}{ccc} \mathbb{R}^N \times \mathbb{R}^m & \longrightarrow & \mathbb{R} \\ (t, \beta) & \longrightarrow & \|t\|_1 + \langle \beta, y - At \rangle. \end{array} \right. \tag{52}$$

The dual function is

$$\psi : \left\{ \begin{array}{ccc} \mathbb{R}^m & \longrightarrow & \mathbb{R} \\ \beta & \longrightarrow & \inf_{t \in \mathbb{R}^N} \mathcal{L}(t, \beta) = \mathcal{L}(t^*_\beta, \beta) \end{array} \right.$$

where t^*_β is a minimizer of $t \to \mathcal{L}(t, \beta)$. By convexity of the objective function, one sees that $t^*_\beta \in \operatorname{argmin}_{t \in \mathbb{R}^N} \mathcal{L}(t, \beta)$ if and only if $A^\top \beta \in \partial^- \|\cdot\|_1 (t^*)$, that is, if and only if β and t^*_β are such that $\left\|A^\top \beta\right\|_\infty \leq 1$ and $(A^\top \beta)_I = \operatorname{sgn}((t^*_\beta)_I)$, where $I = \operatorname{supp}(t^*_\beta)$. One then obtains, for every $\beta \in \mathbb{R}^m$,

$$\psi(\beta) = \mathcal{L}(t^*_\beta, \beta) = \left\|t^*_\beta\right\|_1 - \left\langle A^\top \beta, t^*_\beta \right\rangle + \langle \beta, y \rangle = \langle \beta, y \rangle + \sum_{i \in I} |(t^*_\beta)_i| - \sum_{i \in I} (A^\top \beta)_i (t^*_\beta)_i = \langle \beta, y \rangle$$

because $(A^\top \beta)_i (t^*_\beta)_i = |(t^*_\beta)_i|$ for every $i \in I$. The dual problem is then

$$\max \left(\langle \beta, y \rangle : \left\|A^\top \beta\right\|_\infty \leq 1 \right) \tag{53}$$

Weak duality says that for every $\beta \in \mathbb{R}^m$ such that $\left\|A^\top \beta\right\|_\infty \leq 1$ and $t \in, \mathbb{R}^N$ such that $At = y$, one has $\|t\|_1 \geq \langle \beta, y \rangle$. One can check it directly here, since

$$\langle \beta, y \rangle = \langle \beta, At \rangle = \left\langle A^\top \beta, t \right\rangle \leq \left\|A^\top \beta\right\|_\infty \|t\|_1 \leq \|t\|_1 .$$

The dual certification approach says the following: let $t^* \in \mathbb{R}^N$ be such that $At^* = y$; in order to certify that t^* is indeed a solution of (51), it suffices to find $\beta^* \in \mathbb{R}^m$ such that $\left\|A^\top \beta^*\right\|_\infty \leq 1$ and $\left\langle y, \beta^* \right\rangle = \|t^*\|_1$. Such a β^* will indeed certify that t^* is a solution of (51), because for every $t \in \mathbb{R}^N$ one has $\|t\|_1 \geq \left\langle \beta^*, y \right\rangle$ by strong duality, and hence if $\left\langle y, \beta^* \right\rangle = \|t^*\|_1$ one will also have $\|t\|_1 \geq \|t^*\|_1$, and since t^* is feasible for the primal problem, one indeed concludes that t^* is a solution of (51). To recover exactly the conditions announced above via the subdifferential, it suffices to observe that the strong duality condition at (t^*, β^*), that is, $\left\langle y, \beta^* \right\rangle = \|t^*\|_1$, may also be written $\left\langle t^*, A^\top \beta^* \right\rangle = \|t^*\|_1$, since $At^* = y$. But since $\left\|A^\top \beta^*\right\|_\infty \leq 1$, one has $\left\langle t^*, A^\top \beta^* \right\rangle = \|t^*\|_1$ if and only if $(A^\top \beta^*)_I = \operatorname{sgn}(t^*_I)$, where I is the support of t^*.

One indeed recovers, in both approaches (the one based on the subdifferential and the one based on Lagrangian duality), that the construction of a dual certificate for a candidate $t^* \in \mathbb{R}^N$ such that $At^* = y$ for the problem (51) is given by the problem of finding a $\beta^* \in \mathbb{R}^m$ such that $(A^\top \beta^*)_I = \operatorname{sgn}(t^*_I)$, where I is the support of t^*, and $\left\|A^\top \beta^*\right\|_\infty \leq 1$.

Strict dual certification: uniqueness of a solution. Whether for the SDP problem that we encountered above in community detection, or for the basis pursuit problem, one can certify that a candidate solution of the primal problem is indeed a solution of this problem by the construction of a dual certificate. This approach does not, however, make it possible to prove uniqueness. However, one can also certify the uniqueness of the solution by the construction of a certificate that one calls strict. This is the approach we followed in community detection.

Proposition 6.9. *Given an optimization problem whose primal objective function is φ and whose dual objective function is ψ. If one assumes that there is strong duality and that λ^* is a solution of the dual problem, then one has an equivalence between:*

1) x^* is a solution of the primal problem

2) $\varphi(x^*) = \psi(\lambda^*)$.

Proposition 6.9 assures us that in order to certify a primal solution, it suffices to check that $\varphi(x^*) = \psi(\lambda^*)$ for any solution λ^* of the dual problem. In particular, if one has constructed a dual certificate for a candidate solution x_0^* of the primal (which is therefore feasible for the primal), then this certificate can also serve us to certify other solutions of the primal problem, if there are any. Indeed, if one has shown the existence of a dual certificate λ^*, then there is strong duality, and also λ^* is a solution of the dual. One may therefore apply Proposition 6.9. In particular, if one shows that whenever x^* is feasible for the primal and $x^* \neq x_0^*$, one has $\varphi(x^*) > \psi(\lambda^*)$, one will be able to deduce that x_0^* is the unique solution of the primal problem. The construction of a dual certificate that permits such a strict inequality with all the feasible elements of the primal different from x_0^* are called **strict dual certificates**. It is they that make it possible to certify that a feasible point of the primal is the **unique** solution of the primal. This is the type of construction that we carried out in community detection above.

7 *Matrix lifting* and SDP relaxation

In the previous sections, we used a convex relaxation method mainly to "convexify" constraints that had a combinatorial "nature". In compressed sensing or matrix completion, one had rather convexified the objective function. In the case of the relaxation obtained in (28) for the community detection problem, one speaks of an SDP relaxation, because the problem obtained in the end is a *semidefinite programming (SDP)* problem. One also speaks of *matrix lifting*, because one has rewritten the objective function $x^\top(A - \lambda J_n)x$, which is quadratic in x, as an objective function $\langle A - \lambda J_n, xx^\top \rangle$, which is linear in $xx^\top$. Thus, in passing from the variable x to the variable $Z = xx^\top$, one passes from a vector problem of degree 2 to a linear matrix problem. This device of "*matrix lifting*" makes it possible to convexify the constraint simply by removing the constraint that Z must have rank 1. It is in fact a convex relaxation that one encounters often when one works with quadratic functions of a vector. We give two other examples in which this matrix lifting device can be used.

The *matrix lifting* method and SDP relaxation for the search for a leading eigenvector. In this first example, one implements the matrix lifting method for a simple problem, which is to find a leading eigenvector of a symmetric matrix: let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix; one seeks a vector $u_1 \in \mathbb{R}^n$ such that

$$u_1 \in \operatorname*{argmax}_{\|x\|_2=1} \|Ax\|_2 \,. \tag{54}$$

The *power method* is probably one of the fastest algorithms for approximating a solution of the problem (54); it consists in choosing a point x_0 at random in $\mathbb{R}^n$ and iterating $x_{k+1} = (Ax_k)/\|Ax_k\|_2$. If one has particular factorizations of A, one can also use them to find a solution of (54). The approach we use here is based on *matrix lifting* and constitutes a good introductory example of this method. Nevertheless, it turns out that in the case of (54), this relaxation is in fact *exact*, which is not classical in matrix lifting. We shall come back to this point later.

In the matrix lifting approach, the first step is to bring out a function linear in $xx^\top$. Here, for every $x \in \mathbb{R}^n$, one has

$$\|Ax\|_2^2 = \langle Ax, Ax \rangle = \langle A^2, xx^\top \rangle$$

where one uses the matrix inner product in the last identity (i.e. $\langle A, B \rangle = \mathrm{Tr}(A^\top B)$). One has therefore indeed rewritten (the square of) the objective function of (54) as a linear map of $xx^\top$. The second step is to reinterpret the constraints on x as constraints on $xx^\top$. Here one constrains x to be such that $\|x\|_2 = 1$. This constraint is equivalent to requiring that the trace of $xx^\top$ be equal to 1, because $\mathrm{Tr}(xx^\top) = \sum_i x_i^2 = \|x\|_2^2$. One has therefore lifted the problem (54) into the form:

$$\max_{x:\|x\|_2=1} \|Ax\|_2^2 = \max_{Z\succeq 0:\mathrm{rang}(Z)=1,\mathrm{Tr}(Z)=1} \langle A^2, Z \rangle. \tag{55}$$

The constraints $Z \succeq 0$ and $\mathrm{rang}(Z) = 1$ are here to ensure that the solutions of the problem on the right-hand side of (55) are of the form $xx^\top$ and not $xy^\top$. Indeed, one has

$$\left\{ Z \in \mathbb{R}^{n\times n} : Z \succeq 0, \mathrm{rang}(Z) = 1, \mathrm{Tr}(Z) = 1 \right\} = \left\{ xx^\top : \|x\|_2 = 1 \right\}.$$

There is therefore indeed equality in (55).

The rewriting carried out in (55), of a quadratic vector problem into a linear matrix problem, is at the heart of the matrix lifting approach. To recover a solution u_1 from a solution Z_1 of the matrix problem, it suffices to take the first column of Z_1 and normalize it.

Nevertheless, the matrix problem on the right-hand side of (55) still has the disadvantage of being a problem under a nonconvex constraint, because of the rank constraint. The idea in SDP relaxation is simply to remove the rank constraint in order to obtain a problem of SDP type:

$$\max_{Z\succeq 0:\mathrm{Tr}(Z)=1} \langle A^2, Z \rangle.$$

This is indeed a problem of SDP type, because the constraint bears on the cone of SDP matrices (positive semidefinite) such that $\mathrm{Tr}(Z) = 1$, which is indeed an affine constraint (one can rewrite it as $\langle Z, I_n \rangle = 1$), and the objective function $Z \to \langle A^2, Z \rangle$ is linear. Since one has removed a constraint, one has

$$\max_{Z\succeq 0:\mathrm{rang}(Z)=1,\mathrm{Tr}(Z)=1} \langle A^2, Z \rangle \leq \max_{Z\succeq 0:\mathrm{Tr}(Z)=1} \langle A^2, Z \rangle. \tag{56}$$

In general, this inequality is not an equality, and one then has a "gap" between the original problem and the relaxed problem. However, it turns out that in the particular case of the problem considered here, there is in fact equality. This is why one says that the relaxation is exact here.

To show that one indeed has equality in (56), one considers the eigenvalue decomposition of $A^2 = \sum_i \lambda_i^2 u_i u_i^\top$, and one has for every $Z \succeq 0$ such that $\mathrm{Tr}(Z) = 1$,

$$\langle A^2, Z \rangle = \sum_i \lambda_i^2 \langle u_i u_i^\top, XX^\top \rangle = \sum_i \lambda_i^2 \left\| X^\top u_i \right\|_2^2 \leq \lambda_1^2 \sum_i \left\| X^\top u_i \right\|_2^2 = \lambda_1^2 \mathrm{Tr}(XX^\top) = \lambda_1^2 \tag{57}$$

where one has written $Z = XX^\top$ for $X \in \mathbb{R}^{n\times r}$ and $r = \mathrm{rang}(Z)$ (one can always write a matrix $Z \succeq 0$ in this form thanks to its singular value decomposition). But one sees that if u_1 is a top eigenvector of A, then one has $\langle A^2, u_1 u_1^\top \rangle = \lambda_1^2$, so

$$\lambda_1^2 = \langle A^2, u_1 u_1^\top \rangle \leq \max_{Z\succeq 0:\mathrm{rang}(Z)=1,\mathrm{Tr}(Z)=1} \langle A^2, Z \rangle \leq \max_{Z\succeq 0:\mathrm{Tr}(Z)=1} \langle A^2, Z \rangle \leq \lambda_1^2.$$

One therefore deduces that one indeed has equality in (56) and that the relaxation is exact. One can also see that if one denotes by $E_1 = \mathrm{argmax}_{x:\|x\|_2=1} \|Ax\|_2^2$ the set of solutions of the initial problem, then

$$\operatorname*{argmax}_{Z\succeq 0:\mathrm{Tr}(Z)=1} \langle A^2, Z \rangle = \left\{ xx^\top : x \in \mathrm{conv}(E_1) \right\}$$

where $\mathrm{conv}(E_1)$ is the convex hull of E_1. In particular, if A has a largest eigenvalue of multiplicity 1, then $E_1 = \{-x^*, x^*\}$, where x^* is a leading eigenvector of A. In this case $\mathrm{argmax}_{Z\succeq 0:\mathrm{Tr}(Z)=1}\langle A^2, Z\rangle = \{x^*(x^*)^\top\}$, and hence one obtains x^* or $-x^*$ from $x^*(x^*)^\top$ by taking the first column of $x^*(x^*)^\top$ and normalizing it by its ℓ_2^n norm.

The *matrix lifting* method and SDP relaxation for the MAX-CUT problem, and the bidual Lagrangian relaxation. There exist many optimization problems on graphs. These problems have undergone important developments since the emergence of the internet and of social networks. We give here an example of an optimization problem on a graph, known as the MAX-CUT problem (in French: coupure maximale), which makes it possible to implement the devices of matrix lifting and SDP relaxation.

We are given a graph $G = (V, E)$, where V is the set of vertices; in general, one takes $V = \{1, \ldots, n\}$, and E is the set of edges of the graph: it is a subset of $V \times V$ ($(i, j) \in V$ if and only if there is an edge between node i and node j). We assume that the graph is undirected, which is equivalent to saying that if $(i, j) \in V$ then $(j, i) \in V$.

A **CUT** of G is a partition of these vertices into two parts: $S \sqcup S^c = V$ (where S^c is the complement of S in V). MAX-CUT is the name given to the problem which consists in finding a CUT of G—that is, a subset $S \subset V$—such that the number of edges between S and its complement S^c is maximal. One writes this MAX-CUT problem formally as

$$\max_{S\subset V} \sum_{(i,j)\in V} I(i \in S \text{ and } j \in S^c). \tag{58}$$

It is often useful to transform optimization problems on graphs into optimization problems on matrices. To this end, one introduces the **adjacency matrix** of a graph $G = (V, E)$, where $V = \{1, \ldots, n\}$, by

$$A_{ij} = \begin{cases} 1 & \text{if } (i, j) \in V \\ 0 & \text{otherwise.} \end{cases}$$

Using this notation, one then sees that the MAX-CUT problem may be rewritten as the following optimization problem:

$$\max_{x\in\{-1,1\}^n} \frac{1}{4} \sum_{i,j=1}^{n} A_{ij}(1 - x_j x_i). \tag{59}$$

Both problems (58) and (59) are equivalent: they achieve the same maximum value and if x^* is a solution of (59), then setting $S^* = \{i \in \{1, \ldots, n\} : x_i^* = 1\}$, one obtains a solution of (58). To see this, one can write for every $i, j \in \{1, \ldots, n\}$, $I(i \in S \text{ and } j \in S^c) = I(i \in S)I(j \in S^c) = \frac{(1+x_i)}{2}\frac{(1-x_j)}{2}$, where $x = (x_i)_{i\in V} \in \{-1, 1\}^n$ is such that $x_i = 1$ if $i \in S$ and $x_i = -1$ if $i \in S^c$. Moreover, one has $\sum_{i,j} A_{ij}(x_i - x_j) = 0$ and summing over V is equivalent to summing over $\{1, \ldots, n\}$ if one multiplies the summand by A_{ij}.

Dropping the $1/4$ constant factor, one can also rewrite the problem (59) in the form of a problem of minimizing a quadratic functional under constraints:

$$\min_{x\in\mathbb{R}^n} \left(x^\top A x : x_i^2 = 1\right). \tag{60}$$

It is therefore a constrained optimization problem; the objective function is a quadratic function $x \to x^\top A x$, and the constraint is given by n equality constraints "$x_i^2 = 1$". The difficulty here is that the constraint is not a convex set. One possible approach is then to "convexify" this set.

The "difficult" constraints here are all the $x_i^2 = 1, i = 1, \ldots, n$, because they force the x_i to be discrete (here $x_i^2 = 1$ if and only if $x_i \in \{-1, 1\}$). It is these constraints that make MAX-CUT

a combinatorial problem. We shall then convexify these constraints by using the matrix lifting / SDP relaxation approach. We then show that this approach is here equivalent to a double Lagrangian dualization of the problem, called the bidual Lagrangian relaxation.

For the matrix lifting approach, one writes the objective function as a linear function of $xx^\top$: one has $x^\top Ax = \langle A, xx^\top \rangle$. One writes the constraints on x as constraints on $Z = xx^\top$: $x_i^2 = 1$ if and only if $Z_{ii} = 1$. One therefore has

$$\min_{x\in\mathbb{R}^n} \left(x^\top Ax : x_i^2 = 1 \right) = \min \left(\langle A, Z \rangle : Z \succeq 0, \operatorname{rang}(Z) = 1, Z_{ii} = 1, i = 1, \dots, n \right). \tag{61}$$

The SDP relaxation then tells us simply to remove the rank constraint. One thus obtains the optimization problem

$$\min \left(\langle A, Z \rangle : Z \succeq 0, Z_{ii} = 1, i = 1, \dots, n \right). \tag{62}$$

This SDP relaxation of MAX-CUT is quite famous and bears the name of the Goemans–Williamson relaxation. It is possible to find a solution approximating the solution of the MAX-CUT problem (58) from a solution Z_1 of (62). To this end, one takes $G \sim \mathcal{N}(0, Z_1)$ and $\hat{u} = \operatorname{sign}(G)$. Then the Goemans–Williamson theorem shows that

$$\mathbb{E}\left(\frac{1}{4} \sum_{i,j=1}^n A_{ij}(1 - \hat{u}_j \hat{u}_i) \right) \geq \alpha_{GW} \max_{x\in\{-1,1\}^n} \frac{1}{4} \sum_{i,j=1}^n A_{ij}(1 - x_j x_i),$$

where $\alpha_{GW} = \frac{2}{\pi} \min_{0\leq\theta\leq\pi} \frac{\theta}{1-\cos(\theta)} \approx 0.878$. The proof (see Appendix 8) uses the same technique as that used in the proof of Grothendieck's inequality. This technique bears the name of *rounding technique* and is found in many theorems, such as Nesterov's theorem.

We now show the equivalence of the matrix lifting / SDP relaxation approach and the bidual Lagrangian relaxation. The Lagrangian obtained by Lagrangian relaxation of the difficult constraints associated with the MAX-CUT problem is

$$\mathcal{L} : (x, u) \in \mathbb{R}^n \times \mathbb{R}^n \to x^\top Ax - \sum_{i=1}^n u_i(x_i^2 - 1) = x^\top (A - \operatorname{Diag}(u))x + \langle e, u \rangle$$

where $\operatorname{Diag}(u)$ is the diagonal matrix of size $n \times n$ whose diagonal entries are given by the coordinates of u, and $e = (1)_1^n$. The dual function is

$$\psi : u \in \mathbb{R}^n \to \min_{x\in\mathbb{R}^n} \mathcal{L}(x, u) = \begin{cases} \langle e, u \rangle & \text{when } A - \operatorname{Diag}(u) \succeq 0 \\ -\infty & \text{otherwise.} \end{cases}$$

The dual problem is therefore the problem

$$\max_{u\in\mathbb{R}^n} \psi(u) = \max_{u\in\mathbb{R}^n} \left(\langle e, u \rangle : A - \operatorname{Diag}(u) \succeq 0 \right). \tag{63}$$

Setting $F(u) = A - \operatorname{Diag}(u)$ and denoting by $\mathcal{S}_+^n$ the cone of symmetric positive semidefinite matrices, one sees that the dual problem is a problem under a conic constraint. One can then dualize this problem as we have seen above. Here, (63) is already a dual problem (it is the dual of MAX-CUT); one then speaks of a **bidual problem**. The dual function of (63) is

$$\mathcal{L}' : (u, X) \in \mathbb{R}^n \times (\mathcal{S}_+^n)^\circ \to \langle e, u \rangle - \langle X, A - \operatorname{Diag}(u) \rangle = \langle e + \operatorname{Diag}(X), u \rangle - \langle X, A \rangle$$

where $\mathrm{Diag}(X)$ is the diagonal vector of size n whose coordinates are given by the diagonal entries of X, and $(\mathcal{S}_+^n)^\circ$ is the dual cone of $\mathcal{S}_+^n$. One can show that it is the cone of symmetric negative semidefinite matrices. One then obtains that the dual function of (63) is

$$\psi' : X \preceq 0 \to \max_{u\in\mathbb{R}^n} \mathcal{L}'(u,X) = \begin{cases} +\infty & \text{if } e + \mathrm{Diag}(X) \neq 0 \\ -\langle X, A\rangle & \text{otherwise.} \end{cases}$$

Replacing X by $-X$, one sees that the dual problem is

$$\min_{X\succeq 0} \psi'(-X) = \min_{X\succeq 0} \big(\langle X, A\rangle : X_{ii} = 1, i = 1,\ldots,n\big) \tag{64}$$

One therefore obtains here, by bidual Lagrangian relaxation, exactly the same optimization problem as by the matrix lifting and SDP relaxation approach. Using this last remark and weak duality (of the first dualization), one obtains the following sandwich for the SDP relaxation of MAX-CUT:

$$\max_{u\in\mathbb{R}^n} \big(\langle e, u\rangle : A - \mathrm{Diag}(u) \succeq 0\big) \leq \min_{X\succeq 0} \big(\langle X, A\rangle : X_{ii} = 1\big) \leq \min_{x\in\mathbb{R}^n} \Big(x^\top A x : x_i^2 = 1\Big).$$

Matrix lifting for linear and quadratic constraints in community detection. In the detection of two communities of equal size, we encountered the optimization problem

$$\min \Big(x^\top L x : \langle x, 1\rangle = 0, x \in \{-1,1\}^n\Big). \tag{65}$$

In the approach of Section 6, we introduced a matrix lifting method based only on the lifting of x to $xx^\top$. In doing so, one was able to rewrite the terms of degree exactly two, because there are only terms of degree exactly two in the matrix $xx^\top$. However, in the problem (65), there is also a linear constraint, "$\langle x, 1\rangle = 0$", which cannot be rewritten as a linear function of $xx^\top$. We therefore simply removed this constraint in Section 6, but beforehand we introduced a penalty term in the objective function; it is by this device that we were able to show that the problem obtained by removing the constraint "$\langle x, 1\rangle = 0$" still permitted exact recovery of the communities.

In this paragraph, we propose another matrix lifting method that makes it possible to handle linear terms (in the constraint and/or in the objective function). The idea is to perform the following lifting:

$$\begin{pmatrix} 1 \\ x \end{pmatrix} \rightsquigarrow Z = \begin{pmatrix} 1 \\ x \end{pmatrix} \begin{pmatrix} 1 & x^\top \end{pmatrix} = \left(\begin{array}{c|c} 1 & x^\top \\ \hline x & xx^\top \end{array}\right). \tag{66}$$

By increasing the dimension by just one unit, adding a coordinate equal to 1 to x, one sees linear terms appear in the lifted matrix Z. One can then rewrite the terms linear in x as terms linear in Z (in fact, in the entries of the first row or column of Z). One thus obtains the following lifted problem of the initial problem (65):

$$\min \big(\langle L, Z\rangle : Z \succeq 0, \mathrm{rang}(Z) = 1, Z_{00} = 1, Z_{ii} = 1, i = 1,\ldots,n, \langle Z, A\rangle = 0\big) \tag{67}$$

where $A \in \mathbb{R}^{(n+1)\times(n+1)}$ is given by $A_{ij} = 1$ if $i = 1$ and $j \in \{1,\ldots,n\}$. One indexes here the entries of the matrices Z and A by $\{0,1,\cdots,n\}^2$ (rather than by $\{1,\ldots,n+1\}^2$). The two problems (65) and (67) are equivalent: $\hat{x}$ is a solution of (65) if and only if $\begin{pmatrix} 1 \\ \hat{x} \end{pmatrix} \begin{pmatrix} 1 & \hat{x}^\top \end{pmatrix}$ is a solution of (67). The next step is to perform an SDP relaxation by removing the rank constraint. One then obtains

$$\min \big(\langle L, Z\rangle : Z \succeq 0, Z_{00} = 1, Z_{ii} = 1, i = 1,\ldots,n, \langle Z, A\rangle = 0\big) \tag{68}$$

which is indeed an SDP problem.

One can therefore perform matrix lifting that takes linear terms into account, provided one adds one dimension to the problem and performs the lifting on $\begin{pmatrix} 1 \\ x \end{pmatrix}$ rather than on x, as indicated in (66).

8 Appendix: Proof of the Goemans-Williamson Theorem

Grothendieck's Identity (Sheppard's Lemma)

Lemma 8.1 (Sheppard's Lemma)**.** *Let (X, Y) be a bivariate normal random vector with mean zero, variance 1, and covariance $\rho \in [-1, 1]$. Then the probability that X and Y have opposite signs is given by:*

$$\mathbb{P}(\text{sign}(X) \neq \text{sign}(Y)) = \frac{1}{\pi} \arccos(\rho).$$

Proof. Because the covariance matrix of (X, Y) is positive semi-definite and the marginal variances are 1, we can geometrically represent X and Y as inner products of a standard normal vector in $\mathbb{R}^2$ with two unit vectors. Specifically, let $w \sim \mathcal{N}(0, I_2)$ be a standard 2D Gaussian vector. Let $u, v \in \mathbb{R}^2$ be unit vectors such that $\langle u, v \rangle = \rho$. We can write $(X, Y) \stackrel{d}{=} (\langle w, u \rangle, \langle w, v \rangle)$.

The variables X and Y will have opposite signs if and only if the random vector w falls into the region of the plane where $\langle w, u \rangle$ and $\langle w, v \rangle$ have different signs. This region corresponds exactly to the two wedges formed by the hyperplanes (lines in 2D) orthogonal to u and v.

Since $\langle u, v \rangle = \cos(\theta) = \rho$, the angle between the vectors u and v is $\theta = \arccos(\rho)$. The angle of the wedge separating their orthogonal complements is also θ. Because $w \sim \mathcal{N}(0, I_2)$ is spherically symmetric, the probability that w falls into these alternating wedges is strictly proportional to their angular measure. The total angle of the two wedges is 2θ out of 2π, thus:

$$\mathbb{P}(\text{sign}(X) \neq \text{sign}(Y)) = \frac{2\theta}{2\pi} = \frac{\theta}{\pi} = \frac{1}{\pi} \arccos(\rho).$$

■

The Goemans-Williamson Theorem

Theorem 8.2 (Goemans-Williamson, 1995)**.** *Let Z^* be an optimal solution to the SDP relaxation (62). Let $G \sim \mathcal{N}(0, Z^*)$ and $\hat{u} = \text{sign}(G)$. Then,*

$$\mathbb{E}\left[\frac{1}{4} \sum_{i,j=1}^{n} A_{ij}(1 - \hat{u}_j \hat{u}_i)\right] \geq \alpha_{GW} \max_{x \in \{-1,1\}^n} \frac{1}{4} \sum_{i,j=1}^{n} A_{ij}(1 - x_j x_i),$$

where $\alpha_{GW} = \frac{2}{\pi} \min_{0 \leq \theta \leq \pi} \frac{\theta}{1 - \cos(\theta)} \approx 0.87856$.

Proof. Since Z^* is a symmetric positive semi-definite matrix ($Z^* \succeq 0$), it admits a Cholesky or spectral decomposition. Therefore, there exists a set of vectors $v_1, v_2, \ldots, v_n \in \mathbb{R}^n$ such that $Z^*_{ij} = \langle v_i, v_j \rangle$ for all i, j. The SDP constraint $Z^*_{ii} = 1$ implies that $\langle v_i, v_i \rangle = \|v_i\|^2 = 1$. Thus, the SDP relaxation naturally maps each node i to a unit vector v_i on the unit sphere in $\mathbb{R}^n$.

The objective value of the SDP relaxation is given by:

$$\begin{aligned} \text{SDP-VAL} &:= \max \left(\frac{1}{4} \sum_{i,j=1}^{n} A_{ij}(1 - Z^*_{ij}) : Z \succeq 0, Z_{ii} = 1 \right) \\ &= \frac{1}{4} \sum_{i,j=1}^{n} A_{ij}(1 - Z^*_{ij}) = \frac{1}{4} \sum_{i,j=1}^{n} A_{ij}(1 - \langle v_i, v_j \rangle). \end{aligned}$$

By the definition of the random vector $G \sim \mathcal{N}(0, Z^*)$, we can equivalently generate G by sampling a standard normal vector $w \sim \mathcal{N}(0, I_n)$ and setting $G_i = \langle w, v_i \rangle$. The rounded solution

is then $\hat{u}_i = \text{sign}(\langle w, v_i \rangle)$. Geometrically, w defines a random hyperplane $w^\top = \{v : \langle v, w \rangle = 0\}$ through the origin, and $\hat{u}_i$ assigns $+1$ or -1 depending on which side of the hyperplane the vector v_i lies.

The size of the cut produced by $\hat{u}$ is defined by

$$\text{CUT} := \frac{1}{4} \sum_{i,j=1}^{n} A_{ij}(1 - \hat{u}_i \hat{u}_j) = \sum_{i<j} A_{ij} \frac{1 - \hat{u}_i \hat{u}_j}{2}$$

because $A = A^\top$ and $\hat{u}_i^2 = 1$. Hence, its expected value is

$$\mathbb{E}[\text{CUT}] = \mathbb{E}\left[\sum_{i<j} A_{ij} \frac{1 - \hat{u}_i \hat{u}_j}{2}\right] = \sum_{i<j} A_{ij} \mathbb{P}(\hat{u}_i \neq \hat{u}_j).$$

Notice that (G_i, G_j) is a bivariate normal vector with mean zero, variance $\langle v_i, v_i \rangle = 1$, and covariance $\langle v_i, v_j \rangle$. Applying Lemma 8.1 (Sheppard's Lemma), the probability that v_i and v_j are separated by the random hyperplane is exactly:

$$\mathbb{P}(\hat{u}_i \neq \hat{u}_j) = \frac{1}{\pi} \arccos(\langle v_i, v_j \rangle).$$

Let $\theta_{ij} = \arccos(\langle v_i, v_j \rangle)$ be the angle between the vectors v_i and v_j. We can write the expected cut as:

$$\mathbb{E}[\text{CUT}] = \sum_{i<j} A_{ij} \frac{\theta_{ij}}{\pi}.$$

We want to compare this expected cut term-by-term against the optimal SDP value, whose terms are proportional to $(1 - Z^*_{ij}) = (1 - \cos\theta_{ij})$. For any non-zero angle $\theta_{ij} \in (0, \pi]$, we evaluate the ratio:

$$\frac{\text{Expected Cut}_{ij}}{\text{SDP-VAL}_{ij}} = \frac{\theta_{ij}/\pi}{(1 - \cos\theta_{ij})/2} = \frac{2}{\pi} \frac{\theta_{ij}}{1 - \cos\theta_{ij}}.$$

To establish a global lower bound, we find the worst-case angle that minimizes this ratio. Calculus reveals that the minimum of $f(\theta) = \frac{\theta}{1-\cos(\theta)}$ on $(0, \pi]$ occurs at an angle $\theta^* \approx 2.3311$ radians (the non-zero root of $\cos\theta + \theta \sin\theta = 1$). Substituting this back yields the famous constant:

$$\alpha_{GW} = \min_{0<\theta\leq\pi} \frac{2}{\pi} \frac{\theta}{1 - \cos(\theta)} \approx 0.87856.$$

Since $A_{ij} \geq 0$ for all valid weights in MAX-CUT, we can bound the sum term-by-term:

$$\begin{aligned}
\mathbb{E}[\text{CUT}] &= \sum_{i<j} A_{ij} \frac{\theta_{ij}}{\pi} \\
&\geq \sum_{i<j} A_{ij} \left(\alpha_{GW} \frac{1 - \cos\theta_{ij}}{2}\right) \\
&= \alpha_{GW} \left(\sum_{i<j} A_{ij} \frac{1 - Z^*_{ij}}{2}\right) \\
&= \alpha_{GW} \cdot \text{SDP-VAL}.
\end{aligned}$$

Because the SDP relaxation is a valid upper bound on the true integer MAX-CUT, we know that SDP-VAL $\geq$ OPT where OPT is the value of the optimal solution to the Max-Cut problem. Thus,

$$\mathbb{E}[\text{CUT}] \geq \alpha_{GW} \cdot \text{SDP-VAL} \geq \alpha_{GW} \cdot \text{OPT}.$$

This concludes the proof. ∎

9 Appendix: The Stoer–Wagner algorithm for the minimum cut problem

In this appendix, we present the algorithm of Stoer and Wagner [M. Stoer and F. Wagner, *A simple min-cut algorithm*, Journal of the ACM, 44(4), 1997], which solves exactly, in polynomial time and without any flow computation, the two-class mincut problem introduced in Section 3.2.

Setting and notation

Let $G = (V, E, W)$ be an undirected weighted graph with $n = |V| \geq 2$ nodes and m edges (an edge is an unordered pair $\{i, j\}$ such that $W_{ij} > 0$). We assume that $W_{ij} = W_{ji} \geq 0$ for all $i, j \in V$, that $W_{ii} = 0$ and that $W_{ij} = 0$ when $\{i, j\}$ is not an edge. For two disjoint sets $A, B \subset V$, we denote by

$$w(A, B) = \sum_{i \in A} \sum_{j \in B} W_{ij}$$

the total weight of the edges between A and B, each edge being counted once, and we write $w(A, v) = w(A, \{v\})$ for $v \notin A$. A **cut** of G is a set S such that $\emptyset \neq S \subsetneq V$; its weight is $c(S) = w(S, S^c)$. Note that $c(S) = c(S^c)$. The **global minimum cut** problem is

$$\lambda(G) = \min_{\emptyset \neq S \subsetneq V} c(S). \tag{69}$$

For $x = \mathbf{1}_S \in \{0, 1\}^{|V|}$, it follows from Proposition 2.2 that

$$c(S) = \sum_{(i,j) \in E} x_i W_{ij} (1 - x_j) = \mathbf{1}_S^\top L \mathbf{1}_S.$$

Problem (69) is therefore exactly the two-class mincut problem of Section 3.2, in which the two trivial vectors $x = 0$ and $x = \mathbf{1}$ (which do not define a partition and have value 0) are excluded.

Given two distinct nodes $s, t \in V$, a cut S **separates** s and t when S contains exactly one of them. The **minimum s–t cut** value is

$$\lambda_{s,t}(G) = \min \left(c(S) : S \text{ separates } s \text{ and } t \right).$$

Clearly, for any fixed $s \in V$, $\lambda(G) = \min_{t \neq s} \lambda_{s,t}(G)$.

Remark 9.1 (Mincut, degrees and spectrum). *1) For every $i \in V$, $c(\{i\}) = d_i$. Hence $\lambda(G) \leq \min_{i \in V} d_i$. This explains the phenomenon mentioned in Section 3.2: as soon as one node is weakly connected to the rest of the graph, the minimum cut tends to isolate it. This is a property of the objective function (69), not of the algorithm.*

2) $\lambda(G) = 0$ if and only if G is not connected, that is, by Proposition 2.3, if and only if the second smallest eigenvalue $\lambda_2(L)$ of L is 0. More quantitatively, for every cut S, the vector $f = \mathbf{1}_S - (|S|/n)\mathbf{1}$ is orthogonal to $\mathbf{1}$, satisfies $f^\top L f = \mathbf{1}_S^\top L \mathbf{1}_S = c(S)$ (since $L\mathbf{1} = 0$) and $\|f\|_2^2 = |S||S^c|/n \geq (n-1)/n$. By the Courant–Fischer theorem, one obtains Fiedler's bound

$$\lambda_2(L) = \min_{f \perp \mathbf{1}, f \neq 0} \frac{f^\top L f}{\|f\|_2^2} \leq \frac{n}{n-1} \lambda(G).$$

Intuition: a recursion based on contraction

Given two distinct nodes $s, t \in V$, the **contracted graph** $G/\{s,t\}$ is obtained by replacing s and t by a single new node u and by setting $W_{uv} = W_{sv} + W_{tv}$ for every $v \in V \setminus \{s,t\}$ (the weight W_{st} is discarded). The cuts of $G/\{s,t\}$ are in one-to-one, weight-preserving correspondence with the cuts of G that do *not* separate s and t.

The whole algorithm rests on the following dichotomy. Let S^* be a global minimum cut of G. Either S^* separates s and t, in which case $\lambda(G) = \lambda_{s,t}(G)$, or it does not, in which case S^* survives the contraction and $\lambda(G) = \lambda(G/\{s,t\})$. Since, moreover, $\lambda(G) \le \lambda_{s,t}(G)$ and $\lambda(G) \le \lambda(G/\{s,t\})$, one has, for *every* pair of distinct nodes s, t,

$$\lambda(G) = \min\big(\lambda_{s,t}(G), \lambda(G/\{s,t\})\big). \tag{70}$$

Iterating (70) $n-1$ times reduces the graph to a single node and computes $\lambda(G)$, *provided* that one can compute $\lambda_{s,t}$ cheaply at each step. For an arbitrary pair (s,t), this requires a maximum flow computation. The idea of Stoer and Wagner is not to choose the pair (s,t), but to let a greedy procedure produce a pair (s,t) for which a minimum s–t cut is trivial to identify: it is the singleton $\{t\}$.

Maximum adjacency ordering and the minimum cut phase

Definition 9.2. *An ordering $v_1, \dots, v_n$ of the nodes of G is a **maximum adjacency ordering** if, for every $i = 2, \dots, n$,*

$$v_i \in \operatorname*{argmax}_{v \notin A_i} w(A_i, v), \qquad \text{where } A_i = \{v_1, \dots, v_{i-1}\}.$$

In words, starting from an arbitrary node v_1, one adds at each step the node that is the most tightly connected to the set of nodes already selected. This is the same greedy scheme as in Prim's algorithm for minimum spanning trees, the key "distance to the current set" being replaced by the "total weight towards the current set".

Lemma 9.3 (Stoer–Wagner; Frank; Nagamochi–Ibaraki)**.** *Let $v_1, \dots, v_n$ be a maximum adjacency ordering of G, and let $s = v_{n-1}$ and $t = v_n$. Then*

$$c(\{t\}) = w(V \setminus \{t\}, t) = \lambda_{s,t}(G).$$

That is, $\{t\}$ is a minimum s–t cut of G.

Proof. Since $\{t\}$ separates s and t, one has $c(\{t\}) \ge \lambda_{s,t}(G)$. Conversely, let S be any cut separating s and t. For $i = 1, \dots, n$, denote by $c_i(S)$ the total weight of the edges crossing (S, S^c) whose two endpoints belong to $\{v_1, \dots, v_i\}$, so that $c_n(S) = c(S)$. We say that v_i ($i \ge 2$) is **active** if v_i and v_{i-1} lie on different sides of the cut (S, S^c). We show by induction over the active nodes that

$$w(A_i, v_i) \le c_i(S) \quad \text{for every active node } v_i. \tag{71}$$

Base case. Let v_i be the first active node. Then $v_1, \dots, v_{i-1}$ all lie on the same side of the cut and v_i lies on the other side. All the edges between A_i and v_i therefore cross the cut, and $w(A_i, v_i) \le c_i(S)$ (in fact with equality).

Induction step. Let v_i be an active node and let v_j, $j < i$, be the last active node before v_i. We decompose

$$w(A_i, v_i) = w(A_j, v_i) + w(A_i \setminus A_j, v_i).$$

For the first term, since $v_i \notin A_j$ was a candidate when v_j was selected, the maximum adjacency property and the induction hypothesis give

$$w(A_j, v_i) \le w(A_j, v_j) \le c_j(S).$$

For the second term, since none of $v_{j+1}, \dots, v_{i-1}$ is active, the nodes of $A_i \setminus A_j = \{v_j, \dots, v_{i-1}\}$ all lie on the same side as v_{i-1}, whereas v_i is active and thus lies on the other side. Hence all the edges between $A_i \setminus A_j$ and v_i cross the cut. These edges have v_i as an endpoint, so they are counted in $c_i(S)$ but not in $c_j(S)$. Since all the weights are nonnegative, one gets

$$w(A_i, v_i) \le c_j(S) + w(A_i \setminus A_j, v_i) \le c_i(S),$$

which proves (71).

Finally, $s = v_{n-1}$ and $t = v_n$ are separated by S, so t is active and (71) with $i = n$ gives $c(\{t\}) = w(A_n, t) \le c_n(S) = c(S)$. Taking the minimum over all cuts S separating s and t yields $c(\{t\}) \le \lambda_{s,t}(G)$. ∎

Note that the nonnegativity of the weights is used in a crucial way in the last inequality of the induction step: the edges of $c_i(S)$ that are counted neither in $c_j(S)$ nor in $w(A_i \setminus A_j, v_i)$ must contribute nonnegatively.

The algorithm

In the course of the algorithm, each node v of the current (contracted) graph $G' = (V', E', W')$ represents a set of original nodes $[v] \subset V$, and the contraction rule guarantees the invariant $W'_{uv} = w([u], [v])$ for all distinct $u, v \in V'$. In particular, the weight of a cut of G' equals the weight of the corresponding cut of G. We denote by w' the function w computed in G'.

Input: A weighted graph $G' = (V', E', W')$ with $|V'| \ge 2$ and nonnegative weights, a starting node $a \in V'$
Output: Two nodes $s, t \in V'$ and the value $c'(\{t\})$ of the *cut-of-the-phase*

$A \leftarrow \{a\}$;
while $A \neq V'$ **do**
  choose $v \in \operatorname{argmax}_{u \in V' \setminus A} w'(A, u)$;
  $A \leftarrow A \cup \{v\}$;
end
let s and t be the last two nodes added to A (in this order);
return $\big(s, t, w'(V' \setminus \{t\}, t)\big)$

Algorithm 3: Minimum cut phase (maximum adjacency ordering).

Input: An undirected weighted graph $G = (V, E, W)$ with $n \geq 2$ nodes and nonnegative weights
Output: The value $\lambda(G)$ and a global minimum cut S^*
$G' \leftarrow G$, $[v] \leftarrow \{v\}$ for every $v \in V$, $c^* \leftarrow +\infty$;
for $k = 1, \ldots, n-1$ **do**
  choose an arbitrary node $a \in V'$;
  $(s, t, c) \leftarrow$ output of Algorithm 3 applied to (G', a);
  if $c < c^*$ **then**
    $c^* \leftarrow c$ and $S^* \leftarrow [t]$;
  end
  replace s and t in G' by a new node u with $[u] = [s] \cup [t]$ and $W'_{uv} = W'_{sv} + W'_{tv}$ for every $v \in V' \setminus \{s, t\}$;
end
return (c^*, S^*)

Algorithm 4: Stoer–Wagner algorithm for the global minimum cut.

Theorem 9.4. *Let G be an undirected graph with nonnegative weights and $n \geq 2$ nodes. Algorithm 4 returns $c^* = \lambda(G)$ and a set S^* which is a global minimum cut of G.*

Proof. By the invariant $W'_{uv} = w([u], [v])$, at each phase the value c returned by Algorithm 3 is the weight $c([t])$ in G of the cut $[t]$ of G. Hence $c^* = c(S^*) \geq \lambda(G)$, and it remains to show that $c^* \leq \lambda(G)$. We proceed by induction on n. If $n = 2$, there is a single phase, and $c = c(\{t\})$ is the weight of the unique cut of G (up to complement). If $n \geq 3$, the first phase produces a pair (s, t) with $c = \lambda_{s,t}(G)$ by Lemma 9.3, since Algorithm 3 constructs a maximum adjacency ordering. The remaining $n-2$ phases are exactly Algorithm 4 run on $G/\{s, t\}$, which has $n-1$ nodes, so by the induction hypothesis their best value is $\lambda(G/\{s, t\})$. Therefore $c^* = \min\big(\lambda_{s,t}(G), \lambda(G/\{s, t\})\big) = \lambda(G)$ by (70). ∎

Observe that the cut-of-the-phase is a minimum s–t cut for the pair (s, t) produced by the phase, but it is in general *not* a global minimum cut of the current graph. It is only the minimum over the $n-1$ phases which is guaranteed to be globally optimal.

Computational complexity

Proposition 9.5. *Algorithm 4 can be implemented to run in time*

- $O(nm + n^2 \log n)$ *using a Fibonacci heap,*
- $O(nm \log n)$ *using a binary heap,*
- $O(n^3)$ *using an adjacency matrix and an array of keys, which is the best option for dense graphs ($m \asymp n^2$),*

with memory $O(n + m)$ (adjacency lists) or $O(n^2)$ (adjacency matrix).

Proof. There are $n-1$ phases. Since contraction never increases the number of nodes nor the number of edges, each phase is run on a graph with at most n nodes and m edges. In a phase, one maintains for each $v \notin A$ the key $w'(A, v)$ in a max-priority queue. Each node is extracted once (at most n `extract-max` operations), and when a node v is added to A, the key of each neighbor $u \notin A$ of v is increased by W'_{vu} (at most m `increase-key` operations in total). With

a Fibonacci heap, `extract-max` costs $O(\log n)$ and `increase-key` costs $O(1)$ in amortized time, giving $O(m + n \log n)$ per phase; with a binary heap, both cost $O(\log n)$, giving $O(m \log n)$ per phase; with an array, `extract-max` costs $O(n)$ and `increase-key` costs $O(1)$, giving $O(n^2)$ per phase. The contraction and the update of $[u]$ cost $O(n)$ per phase. Multiplying by $n - 1$ phases gives the result. ∎

Further facts

History and comparison with other algorithms.

- The naive approach computes $\lambda(G) = \min_{t \neq s} \lambda_{s,t}(G)$ with $n - 1$ maximum flow computations (by the max-flow/min-cut theorem). Hao and Orlin (1994) showed that the global minimum cut can be obtained at the cost of essentially a single push-relabel maximum flow computation, in time $O(nm \log(n^2/m))$.
- Nagamochi and Ibaraki (1992) gave the first deterministic algorithm that does not use flows, with the same complexity $O(nm + n^2 \log n)$ as Stoer–Wagner. Stoer and Wagner (conference version 1994, journal version 1997), and independently Frank (1994), greatly simplified both the algorithm and its proof, the key ingredient being Lemma 9.3.
- Karger's randomized contraction algorithm (1993) contracts edges chosen at random with probability proportional to their weight until two nodes remain. A fixed global minimum cut survives with probability at least $2/(n(n-1))$. A nice combinatorial consequence is that a graph has at most $\binom{n}{2}$ global minimum cuts (a bound attained by the cycle on n nodes). Karger and Stein (1996) improved this approach to $O(n^2 \log^3 n)$, and Karger (2000) obtained a randomized algorithm running in near-linear time $O(m \log^3 n)$. Deterministic near-linear time algorithms have been obtained more recently (Kawarabayashi and Thorup (2015) for simple unweighted graphs, Henzinger, Li, Rao and Wang (2024) for weighted graphs). Stoer–Wagner nevertheless remains a reference algorithm thanks to its simplicity, its determinism and the fact that it does not rely on flows.

Why mincut is easy while maxcut is hard. The global minimum cut problem and the maxcut problem of Section 7 are both optimization problems of the quadratic form $x \mapsto x^\top L x$ over the hypercube. However, mincut can be solved exactly in polynomial time, whereas maxcut is NP-hard and only admits the α_{GW}-approximation of Goemans and Williamson. The difference comes entirely from the sign of the weights: minimizing the cut function associated with $-W$ is the same as maximizing the cut function associated with W, so the minimum cut problem with weights of arbitrary sign is NP-hard. The nonnegativity of W is exactly what is used in the proof of Lemma 9.3. At a more structural level, when $W \geq 0$ the cut function is **symmetric submodular**: $c(S) = c(S^c)$ and $c(S \cup T) + c(S \cap T) \leq c(S) + c(T)$ for all $S, T \subset V$. Queyranne (1998) showed that the maximum adjacency ordering idea extends to the minimization of an arbitrary symmetric submodular function over nonempty proper subsets, using $O(n^3)$ evaluations of the function.

An exact convex relaxation. Submodularity also explains why convex relaxation is *exact* for cut problems with nonnegative weights, in contrast with the relaxations of the RatioCut and maxcut problems studied in this chapter. The convex function

$$x \in \mathbb{R}^{|V|} \longrightarrow \mathrm{TV}(x) = \sum_{\{i,j\} \in E} W_{ij} |x_i - x_j|$$

(a weighted ℓ_1-norm of the discrete gradient of x, also called the graph total variation; it is the Lovász extension of c) satisfies $\mathrm{TV}(\mathbf{1}_S) = c(S)$, and it is convex because $W \geq 0$.

Proposition 9.6. *Let $s \neq t$ be two nodes of G. Then*

$$\lambda_{s,t}(G) = \min\Big(\mathrm{TV}(x) : x \in [0,1]^{|V|}, x_s = 1, x_t = 0\Big).$$

Proof. For every cut S containing s but not t, $x = \mathbf{1}_S$ is feasible and $\mathrm{TV}(\mathbf{1}_S) = c(S)$, so the minimum is at most $\lambda_{s,t}(G)$ (recall that $c(S) = c(S^c)$, so one may assume $s \in S$). Conversely, let x be feasible, let θ be uniformly distributed on $(0,1)$ and let $S_\theta = \{i \in V : x_i > \theta\}$. Almost surely, $s \in S_\theta$ (since $x_s = 1 > \theta$) and $t \notin S_\theta$ (since $x_t = 0 \leq \theta$), so S_θ separates s and t and $c(S_\theta) \geq \lambda_{s,t}(G)$. Moreover, an edge $\{i,j\}$ crosses (S_θ, S_θ^c) if and only if $\min(x_i, x_j) \leq \theta < \max(x_i, x_j)$, an event of probability $|x_i - x_j|$. Hence

$$\lambda_{s,t}(G) \leq \mathbb{E}\left[c(S_\theta)\right] = \sum_{\{i,j\} \in E} W_{ij}\, \mathbb{P}\left(\{i,j\} \text{ crosses } (S_\theta, S_\theta^c)\right) = \mathrm{TV}(x).$$

Minimizing over feasible x concludes the proof. ∎

This *threshold rounding* argument is the cut analogue of the random hyperplane rounding used in the proof of the Goemans–Williamson theorem, except that here no loss occurs: the relaxation gap is zero. Since $|x_i - x_j|$ can be linearized with auxiliary variables, the relaxation of Proposition 9.6 is a linear program, and its dual is the maximum s–t flow problem: this is one way to prove the max-flow/min-cut theorem by LP duality. It is also instructive to compare it with the spectral method: replacing the ℓ_1-type penalty $|x_i - x_j|$ by the ℓ_2-type penalty $(x_i - x_j)^2$ gives the quadratic form $x^\top L x$ (Proposition 2.2), whose minimizer under the same constraints is the electrical potential of the graph and is in general not an indicator vector; that relaxation is not exact.

Balance constraints destroy tractability. As explained in Section 3.2 and in Remark 9.1, the global minimum cut tends to isolate small sets of nodes. Forcing the communities to be balanced makes the problem hard: the minimum bisection problem (mincut under the constraint $|S| = n/2$), as well as the RatioCut and normalized cut problems, are NP-hard. This is the reason why, in the core of this chapter, these problems are attacked through convex relaxations (spectral and SDP) rather than solved exactly. Similarly, the minimum k-cut problem (removing a minimum weight set of edges so as to obtain at least k connected components) is polynomial for every fixed k (Goldschmidt and Hochbaum, 1994), but NP-hard when k is part of the input.

In practice. The algorithm is available, for instance, in the Python library `networkx` through the function `networkx.stoer_wagner(G, weight="weight")`, which returns the value $\lambda(G)$ and the corresponding partition $(S^*, (S^*)^c)$; its default implementation uses a binary heap, hence runs in $O(nm \log n)$.