Méthodes spectrales et relaxations SDP pour la détection de communautés dans les graphes

Guillaume Lecué*

Résumé

Dans ce chapitre, nous présentons deux types de procédures pour le problème de détection de communautés dans un graphe : une méthode spectrale et des relaxations SDP. Toutes deux sont issues d'une relaxation convexe d'un problème NP-hard. La méthode spectrale est obtenue en convexifiant l'espace de contrainte. La procédure SDP mets en oeuvre la technique de *matrix lifting* et de *relaxation SDP*. Dans les deux cas, le Laplacien du graphe joue un rôle central puisque son spectre nous donne des informations sur les propriétés de connectivité du graphe. Les premières sections sont dédiés à son étude dans le cas idéal de communautés disjointes ; on parle alors de composantes connexes. D'un point de vue statistique, on étudie les propriétés d'estimation et de reconstruction exacte de deux relaxations SDP. Comme dans les chapitres précédents cette étude mets en avant le rôle de l'aléatoire dans cette approche et nous introduisons à cette occasion un modèle de graphe aléatoire appelé le *stochastic block model*. D'un point de vue technique, l'inégalité de Grothendieck et la construction d'un certificat dual sont utilisés pour mener l'étude statistique des relaxations SDP.

1 Introduction

Exemple d'un problème de détection de communauté : On dispose d'un graphe où chaque nœud représente une personne et un lien du graphe entre deux nœuds représente une connexion entres ces deux personnes établie, par exemple, à partir d'un réseau social d'échange de messages. On cherche à identifier dans ce graphe des groupes de noeuds particulièrement connectés. Ces groupes peuvent représenter des groupes d'amis ou des personnes partageant des intérêts communs qui ont donc tendance à souvent s'échanger des messages. On dispose donc d'un graphe et le problème de détection de communauté est de trouver des groupes de nœuds plus densément connectés entre eux qu'avec le reste du graphe. Une fois les communautés détectées, on peut chercher les influenceurs en leur sein et leur envoyer des pubs ou les démarcher directement.

Définition 1.1. *Un* ***graphe*** *est un couple (V, E) où V est un ensemble dont les éléments sont appelés les* ***noeuds du graphe*** *et $E \subset V \times V$ est un ensemble dont les éléments sont appelés* ***arêtes du graphe****. Un* ***graphe pondéré*** *est un triplet (V, E, W) tel que (V, E) est un graphe et $W \in \mathbb{R}^{|V|\times|V|}$. Les entrées de W sont appelées les* ***poids du graphe****. On dit qu'un graphe (V, E) est* ***non orienté*** *quand $(i, j) \in E$ implique $(j, i) \in E$ et qu'un graphe pondéré (V, E, W) est* ***non orienté*** *quand (V, E) est un graphe non orienté et $W = W^\top$.*

Il existe plusieurs types de graphes selon qu'ils soient orientés (ou non) et pondérés (ou non) comme représenté dans la Figure 3. On donne aussi quelques exemples classiques de graphes dans les Figures 1 et 2.

*CREST, ENSAE. Bureau 3029, 5 avenue Henry Le Chatelier. 91 120 Palaiseau. Email: guillaume.lecue@ensae.fr.

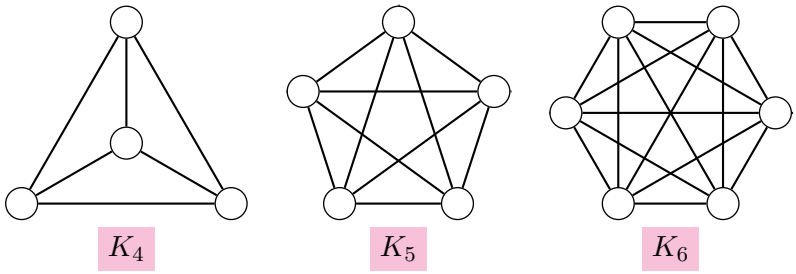

FIGURE 1 – Trois exemples de graphes complets

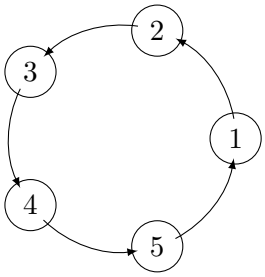

FIGURE 2 – Exemple d'un ring graph

On représente souvent les propriétés de connexions d'un graphe à l'aide d'une matrice.

Définition 1.2. *Soit $G = (V, E)$ un graphe. La* ***matrice d'adjacence*** *de G est donnée par $A \in \{0,1\}^{|V|\times|V|}$ où pour tout $(i,j) \in \{1,\dots,|V|\}^2$, on a $A_{ij} = 1$ quand $(i,j) \in E$ et $A_{ij} = 0$ quand $(i,j) \notin E$. La matrice d'adjacence d'un graphe pondéré (V, E, W) est $A \in \mathbb{R}^{|V|\times|V|}$ où on a $A_{ij} = W_{ij}$ quand $(i,j) \in E$ et $A_{ij} = 0$ quand $(i,j) \notin E$.*

Si un graphe est non orienté alors sa matrice d'adjacence est symétrique. Pour le problème de détection de communautés, on s'intéresse aux propriétés de connectivité dans un graphe. La première définition porte sur les propriétés de connectivités au sein d'un graphe.

Définition 1.3. *Un graphe $G = (V, E)$ est dit* ***connecté*** *ou* ***connexe*** *quand on peut trouver un chemin d'arêtes liant tous les nœuds du graphe. Une* ***composante connexe*** *de G est un sous-graphe de G connecté et sans arête le liant avec son complémentaire. Une* ***partition connexe*** *de G est un ensemble fini de sous-graphes de G formant une partition des nœuds de G telle que les sous-graphes associés à cette partition sont des composantes connexes.*

On peut aussi étendre cette définition aux graphes pondérés. On retrouve dans ce cas la définition précédente pour un graphe non pondéré quand on lui ajoute la matrice de poids constants $W = (1)_{(i,j)\in V\times V}$.

Définition 1.4. *Un graphe pondéré $G = (V, E, W)$ est dit* ***connecté*** *ou* ***connexe*** *quand on peut trouver un chemin d'arêtes liant tous les nœuds du graphe et $W_{ij} \neq 0$ pour toutes les arêtes (i,j) de ce chemin. Une* ***composante connexe*** *de G est un sous-graphe de G connecté et sans arête de poids non nul le liant avec son complémentaire. Une* ***partition connexe*** *d'un graphe G est un ensemble fini de sous-graphes de G formant une partition des nœuds de G et tels que ces sous-graphes sont des composantes connexes.*

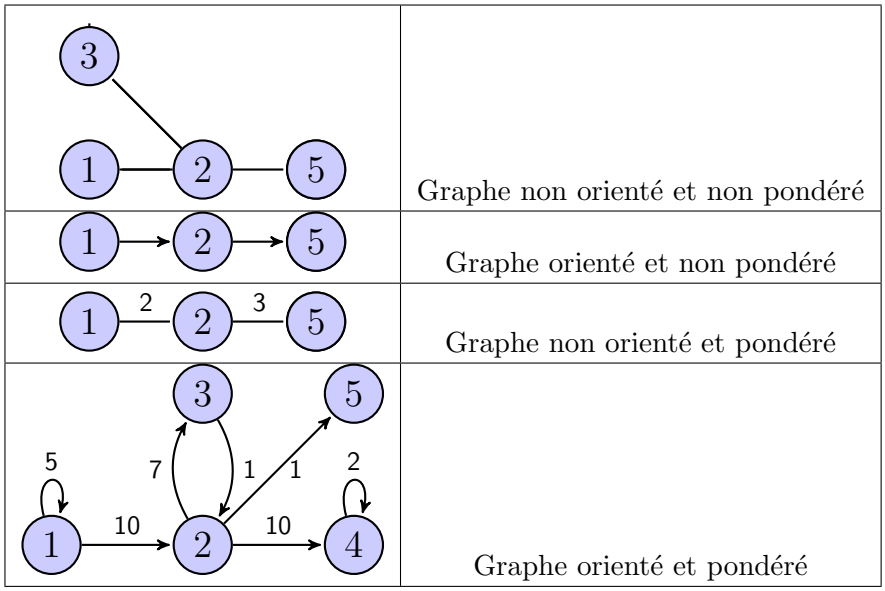

FIGURE 3 – Exemples de graphes orientés ou non et pondérés ou non

Un graphe pouvant se décomposer en une partition connexe non triviale (càd non réduite à un élément) est en quelque sorte un cas extrémal pour le problème de détection de communautés car dans ce cas les communautés n'interagissent pas. C'est un cas assez rare en pratique. Néanmoins, il peut nous guider pour la conception d'algorithmes en détection de communautés car c'est un modèle idéal pour ce problème.

2 Le Laplacien d'un graphe et quelques propriétés de son spectre

Dans cette section, on s'intéresse au lien entre les propriétés de connectivité d'un graphe et le spectre de son Laplacien. On définit d'abord le Laplacien d'un graphe.

Définition 2.1. *Soit G un graphe (pondéré ou non). On note par V l'ensemble des nœuds de G et par A sa matrice d'adjacence. La* ***matrice de Laplace*** *ou* ***Laplacien*** *de G est donnée par $L = D - A$ où $D = \mathrm{diag}(d)$, $d = (d_i)_{i\in V}$ et $d_i = \sum_{j\in V} A_{ij}$ pour tout $i \in V$ est le degré du i-ième noeud.*

Il existe d'autres définition du Laplacien d'un graphe. Nous utiliserons celle de la Définition 2.1 durant tout le chapitre. Si G est un graphe non orienté alors son Laplacien est symétrique. On peut donc appliquer le théorème spectral à L dans ce cas : les valeurs propres de L sont toutes réelles et L est diagonalisable dans une base orthonormale (càd il existe une matrice orthonormale $P \in \mathcal{O}(|V|)$ et une matrice diagonale $D = \mathrm{diag}(\lambda)$ où $\lambda = (\lambda_i)_{i\in V}$ est le vecteur des valeurs propres de L telles que $L = PDP^\top$). On verra dans la suite que le spectre de L (càd l'ensemble de ses valeurs propres) va jouer un rôle clef pour la méthode spectral en détection des communautés.

Proposition 2.2. *Soit $G = (V, E, W)$ un graphe pondéré non orienté. On suppose que $W_{ij} \geq 0$ pour tout $(i, j) \in V \times V$. On a :*

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

2) L est symétrique positive

3) 0 est valeur propre de L et $L\mathbf{1} = 0$ où $\mathbf{1} = (1)_{i\in V}$.

Démonstration. **1)** On note par $A = (W_{ij}\mathbf{1}_{(i,j)\in E})$ la matrice d'adjacence de G. Comme G est non orienté, on a $W_{ij} = W_{ji}$ pour tout $(i,j) \in V \times V$ et donc pour tout $f \in \mathbb{R}^{|V|}$,

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

2) On a pour tout $f \in \mathbb{R}^{|V|}$,

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

On a donc bien $L \succeq 0$.

3) Par définition du degré d'un noeud, on a pour tout $i \in V$,

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

On a donc bien 0 comme valeur propre de L et $\mathbf{1}$ est un vecteur propre associé à la valeur propre 0. ∎

Le résultat suivant est un premier résultat faisant un lien entre le spectre du Laplacien et les propriétés de connectivité du graphe sous-jacent. En effet, la multiplicité de la valeur propre 0 de L donne le nombre de composantes connexes de G. De plus, les indicatrices des nœuds de ces composantes connexes sont des vecteurs propres engendrant cet espace propre (càd le noyau) de L.

Proposition 2.3. *Soit $G = (V, E, W)$ un graphe pondéré non orienté dont les poids sont positifs. La multiplicité de la valeur propre 0 du Laplacien de G est égale au nombre de composante connexes de G. De plus, l'espace propre associé à la valeur propre 0 de L, càd son noyau $\mathrm{Ker}(L)$, est engendré par $\mathbf{1}_{V_1}, \ldots, \mathbf{1}_{V_k}$ où $V_1 \sqcup \cdots \sqcup V_k$ est la partition de V en composantes connexes de G.*

Démonstration. On commence par le cas d'une seule composante connexe, càd quand G est connexe. Le noyau de L est formé de tous les éléments $f \in \mathbb{R}^{|V|}$ tels que $Lf = 0$. Si $Lf = 0$ alors $f^\top Lf = 0$ et donc d'après la Proposition 2.2, on aura $\sum_{(i,j)\in E} W_{ij}(f_j - f_i)^2 = 0$. Par ailleurs, les poids W_{ij} sont positifs ou nuls, alors, si $\sum_{(i,j)\in E} W_{ij}(f_j - f_i)^2 = 0$, on a pour tout $(i,j) \in E$, $W_{ij}(f_j - f_i)^2 = 0$ donc soit $W_{ij} = 0$ soit $f_j = f_i$. Par ailleurs, G est connexe donc, si on prend n'importe quel couple de nœuds $i, j \in V$ il existe un chemin représenté par des indices de noeuds $i_1, \ldots, i_p \in V$ tels que $(i, i_1) \in E, (i_1, i_2) \in E, \ldots, (i_p, j) \in E$ et $W_{ii_1} > 0, W_{i_1i_2} > 0, \cdots, W_{i_kj} > 0$. Le long de ce chemin, on a soit $W_{pq} = 0$ soit $f_p = f_q$ pour $(p,q) \in \{(i,i_1), (i_1,i_2), \cdots, (i_p,j)\}$; mais comme $W_{pq} \neq 0$ on a $f_p = f_q$. Ceci étant vrai pour toutes les arêtes du chemin, on en déduit que les extrémités f_i et f_j sont égales. Donc, f est un vecteur constant. On a donc bien que $\mathrm{Ker}(L)$ est engendré par $\mathbf{1} = (1)_{i\in V}$ l'indicatrice de l'unique composante connexe de G.

On considère le cas général où G a k composantes connexes. Quitte à réordonner les nœuds de V, on peut écrire le Laplacien de G sous forme de matrice par blocs :

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

où pour tout $i = 1, \ldots, k$, L_i est le Laplacien de la i-ième composante connexe $G_i = (V_i, E_i, W^{(i)})$ de G où $V_i \subset V$ est l'ensemble des nœuds de la i-ième composante connexe de G, $E_i = \{(p, g) : p, q \in V_i, (p, q) \in E\}$ est l'ensemble de ses arêtes et $W^{(i)} = (W_{pq} : p, q \in V_i)$ est sa matrice de poids.

On note $\mathbf{1}_{V_i} \in \mathbb{R}^{|V|}$ l'indicatrice des nœuds de G_i. On veut montrer que $\mathrm{Ker}(L)$ est engendré par $\mathbf{1}_{V_i}, i = 1, \cdots, k$. Pour tout $i = 1, \ldots, k$, L_i est le Laplacien de G_i donc 0 est valeur propre de L_i et $(1)_{p \in V_i}$ engendre le noyau de L_i. Donc $\mathbf{1}_{V_i}$ est dans le noyau de L. On a donc $\mathrm{vect}(\mathbf{1}_{V_i}, i = 1, \cdots, k) \subset \mathrm{Ker}(L)$. Par ailleurs, étant donnée la structure par bloc de L, on voit que $Lf = 0$ si et seulement si $L_i f = 0$ pour tout $i = 1, \ldots, k$ et comme $\ker(L_i) = \mathrm{vect}((\mathbf{1})_{\mathbf{p} \in \mathbf{V_i}})$, on a $f_{|V_i} = \alpha_i (\mathbf{1})_{\mathbf{p} \in \mathbf{V_i}}$ pour un certain $\alpha_i \in \mathbb{R}$. On en déduit que $f = \sum_i \alpha_i \mathbf{1}_{V_i} \in \mathrm{vect}(\mathbf{1}_{V_i}, i = 1, \cdots, k)$. ■

Remarque 2.4. *La Proposition 2.3 s'applique aussi aux graphes non pondérés non orientés. Il suffit de l'appliquer à la matrice de poids $W = (1)_{(i,j) \in V \times V}$.*

Exemple : Calculons la multiplicité de la valeur propre 0 du Laplacien d'un graphe complet et vérifions qu'elle est bien égale à 1 étant donné que ce graphe est connexe. On note K_n le graphe complet de n sommets. Le Laplacien de K_n est donné par

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

Si $f \in \mathbb{R}^n$ est tel que $Lf = 0$ alors $nf = (\langle f, (1)_1^n \rangle)_{i=1,\ldots,n} = \langle f, (1)_1^n \rangle (1)_1^n$. Donc $f \in \mathrm{vect}((1)_1^n)$ et 0 est de multiplicité 1 pour L.

Le Laplacien d'un graphe et la loi de refroidissement de Newton. On a vu que le spectre du Laplacien $L = D - A$ d'un graphe joue un rôle essentiel sur les propriétés de connectivité du graphe. Il apparaît aussi lorsqu'on étudie d'autres propriétés d'un graphe. On peut se poser la question sur l'origine de son nom et en particulier s'il a un lien avec le Laplacien qu'on rencontre en physique quand on étudie l'équation de la chaleur. Il se trouve qu'il y a bien un lien entre les deux notions et qu'on peut voir le Laplacien d'un graphe comme une version discrète du Laplacien introduit en physique. Le lien unissant les deux approches est la loi de refroidissement de Newton ou de transfert de chaleur (*Newton's law of cooling* en anglais) disant que *la chaleur se transfert d'un point à un autre proportionnellement à la différence de température entre les deux points.*

Pour faire ce lien formellement, on imagine que les nœuds du graphe $G = (V, E)$ ont des températures données par la famille $(T_i)_{i \in V}$; T_i étant la température du nœud i à la date t. On laisse évoluer les transferts de chaleur sur ce graphe en fonction de la loi de Newton : 'le gradient de température (càd ce qu'on appelle communément la chaleur) entre deux points connectés est

proportionnel à la différence de température entre ces deux points' : pour tout nœud $i \in V$, la température en ce nœud va évoluer de la manière suivante

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

En d'autres termes, le nœud i échange de la chaleur uniquement avec ses voisins (càd les j tels que $A_{ij} = 1$) ; il reçoit de la chaleur du noeud voisin j que si $T_j > T_i$ et il en donne au noeud voisin j que si $T_j < T_i$. Le paramètre $\kappa > 0$ est un paramètre contrôlant la vitesse de ce transfert, c'est un coefficient thermique.

On peut ensuite regarder l'évolution du vecteur des températures des noeuds du graphe $T = (T_i)_{i \in V}$ et développer cette égalité pour faire apparaître le Laplacien :

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

On obtient donc une équation différentielle sur l'évolution des températures des nœuds du graphe de la forme $dT/dt + \kappa LT = 0$ où $L = D - A$ est le Laplacien du graphe. Si on rappelle l'équation de la chaleur $\partial u/\partial t = \alpha \Delta u$ où $u : (t, x) \in \mathbb{R} \times \mathbb{R}^d \to u(t, x)$ est la température au temps t à l'endroit x, Δ est le Laplacien $\partial_1^2 + \cdots + \partial_d^2$, on peut identifier L et $-\Delta$. C'est de cette analogie dont L tire son nom.

3 Principe des méthodes spectrales en détection de communautés

On donne ici deux idées introduisant plus ou moins formellement la méthode spectrale pour trouver des communautés dans des graphes. On suppose qu'on dispose d'un graphe $G = (V, E, W)$ pondéré non-orienté.

3.1 Cadre idéal de composantes connexes

Cadre idéal : Quand il s'agit de détecter des communautés dans un graphe, en quelque sorte, le cadre idéal a lieu quand ces communautés ne sont pas connectées entre elles, càd quand le graphe admet une partition en composantes connexes et que chaque composante connexe constitue une communauté. C'est un cadre idéal car il n'apparaît presque jamais en pratique, vue qu'on a toujours quelques liens inter-communautés.

Même si le cadre idéal est un cadre qui n'a presque jamais lieu en pratique, il est un bon guide pour comprendre comment solutionner le problème de détection de communautés dans des situations plus générales.

On se place alors dans le cadre idéal (en première approximation). Le Laplacien de $G = (V, E, W)$ et son noyau ont donc des formes particulières. En effet, d'après la Proposition 2.3, le noyau de L est engendré par $\mathbf{1}_{V_1}, \ldots, \mathbf{1}_{V_k}$ où $V_1 \sqcup \cdots \sqcup V_k$ est la partition de nœuds de V en les k composantes connexes de G.

Il n'est cependant pas facile de trouver les indicatrices $\mathbf{1}_{V_i}, i = 1, \ldots, k$ à partir de L. Ce qui est plus facile est de trouver une base orthonormale de $\mathrm{Ker}(L)$: $u_1, \ldots, u_k \in \mathbb{R}^{|V|}$ (on peut d'ailleurs prendre $u_1 = (1)_{i=1}^n$). La **méthode spectrale** procède ensuite en deux étapes :

1) on clusterise les $|V|$ vecteurs lignes $(y_i)_{i\in V}$ de la matrice $[u_1|u_2|\cdots|u_k]$ de taille $|V|\times k$ en k clusters $C_1,\cdots,C_k$
2) on retourne la partition $V_1\sqcup\cdots\sqcup V_k$ des nœuds de V où $V_p=\{i\in V : y_i\in C_p\}$ pour $p=1,\ldots,k$. Ceux sont les communautés de nœuds qu'on a détectées par la méthode spectrale.

On s'attend à ce que la partition $V_1\sqcup\cdots\sqcup V_k$ des nœuds de V donne les k composantes connexes de G . En effet, $\{u_1,\ldots,u_k\}$ et $\{\mathbf{1}_{V_1}/\sqrt{|V_1|},\ldots,\mathbf{1}_{V_k}/\sqrt{|V_k|}\}$ sont deux bases orthonormales de $\mathrm{Ker}(L)$. Il existe alors une matrice de rotation (matrice orthogonale) $R\in\mathcal{O}(|V|)$ telle que $u_i=R\mathbf{1}_{V_i}/\sqrt{|V_i|}, i=1,\ldots,k$. Si on avait effectué ce clustering des lignes sur la matrice $[\mathbf{1}_{V_1}/\sqrt{|V_1|}|\cdots|\mathbf{1}_{V_k}/\sqrt{|V_k|}]$ à k clusters, les clusters obtenus seraient formés des lignes qui sont toutes égales entre elles (celles ayant un $1/\sqrt{|V_i|}$ au même endroit et 0 ailleurs) car il y en a exactement k et on fait un clustering à k classes. C'est donc la partition de G en ses composantes connexes qu'on retrouverait exactement. Par ailleurs, on se dit aussi que la rotation R devrait conserver ce clustering des lignes : on a

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

et que donc clusteriser les vecteurs lignes $(y_i)_{i\in V}$ de la matrice $[u_1|u_2|\cdots|u_k]$ devrait aussi redonner les composantes connexes de G comme ce clustering le fait sur les lignes de $[\mathbf{1}_{V_1}/\sqrt{|V_1|}|\cdots|\mathbf{1}_{V_k}/\sqrt{|V_k|}]$.

Voilà pour ce qui est de l'intuition derrière une méthode spectrale couramment utilisée en détection de communautés. En pratique, G ne sera pas partitionable en composantes connexes ; ainsi, on ne regardera pas le noyau de L mais plutôt son espace propre associé à ses k plus petites valeurs propres. On cherchera ensuite une base de k vecteurs propres $u_1,\ldots,u_k$ de cet espace propre. C'est alors les $|V|$ lignes de la matrice $[u_1|\cdots|u_k]$ qu'on clusterisera en k groupes. Des indices de ligne de ces k groupes on extraira un clustering des noeuds du graphe, càd on aura un estimateur des communautés. Le choix du paramètre k se fait soit par validation croisée si on se donne en amont un critère à minimiser – généralement, les critères en détection de communautés sont appelés **fonctions de modularité** – soit en représentant le spectre de L et en identifiant un 'coude' au niveau de ses plus petites valeurs propres. Ces deux méthodes sont heuristiques.

3.2 Point de vue "graph cut"

On donne dans cette section, un autre point de vue sur les méthodes spectrales utilisées en détection de communautés. Le principe utilisé ici est que détecter des communautés est en fait équivalent à trouver une partition des nœuds minimisant la somme des poids portés par les arêtes inter-communautés tout en maximisant la taille de ces communautés.

Pour formaliser cette approche, on introduit quelque fonctions, appelées **fonctions de modularité**, qui quantifie les notions de masse de poids intra et inter communautés qu'on cherchera à optimiser sur tous les "**cuts**", càd partitions, du graphe.

Mincut problem : Étant donné deux ensembles $A,B\subset V$ de nœuds, on définit la masse totale entre A et B par

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

où on a utilisé la symétrie de W dans la dernière égalité. Ainsi on peut définir une fonction associant à chaque partition $V_1 \sqcup \cdots \sqcup V_k$ des nœuds du graphe la masse totale des poids intra-communauté :

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

où $\mathcal{P}_k(V)$ est l'ensemble des partitions de V ayant au plus k éléments et V_i^c est le complémentaire de V_i dans V.

Pour le problème à deux communautés, on peut récrire le problème sous la forme suivante. Chaque partition de V en deux communautés $V_1 \sqcup V_2$ (ici $V_2 = V_1^c$) est décrite par un vecteur $x \in \{0,1\}^{|V|}$ d'appartenance tel que $x_i = 1$ si $i \in V_1$ et $x_i = 0$ si $i \in V_2$. On a alors

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

Le problème du mincut à deux classes peut donc s'énoncer sous la forme suivante : trouver $x^* \in \{0,1\}^{|V|}$ solution du problème

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

Ce problème de min-cut à deux classes peut se résoudre de manière efficace grâce à l'algorithme de Stoer-Wagner. Mais en pratique l'algorithme renvoie souvent une partition dont une classe se réduit à un seul noeud (le point de plus petit degré, càd, celui le moins connecté au graphe). C'est pas vraiment l'idée qu'on se fait d'une communauté. On va donc forcer les communautés à être de taille suffisante en introduisant une nouvelle fonction de modularité.

Ratiocut problem : Pour éviter les solutions triviales du mincut problem, on introduit une nouvelle fonction de modularité :

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

qui force la taille des communautés à ne pas être trop petite lorsque l'on va minimiser ce ratiocut. Le ratio $W(V_i, V_i^c)/|V_i|$ est sensé réaliser une balance entre la connectivité de V_i avec son complémentaire et la taille de V_i. Le problème du RatioCut est donc celui qui consiste à minimiser sur $\mathcal{P}_k(V)$ la fonction $\text{RatioCut}(V_1, \ldots, V_k)$

En général minimiser le ratiocut (2) sur toutes les partitions de $\mathcal{P}_k(V)$ est NP-hard. On va alors récrire ce problème sous forme vectoriel (pour $k = 2$) ou matriciel (k général) et montrer qu'une relaxation convexe de ce problème donne la méthode spectrale de détection de communauté vue au chapitre précédent.

Ratiocut pour $k = 2$: On regarde d'abord le problème du ratiocut dans le cas de deux communautés. On cherche alors une solution au problème

$$\min_{V_1\subset V} \text{RatioCut}(V_1, V_1^c) = \min_{V_1\subset V} \left(\frac{W(V_1, V_1^c)}{|V_1|} + \frac{W(V_1^c, V_1)}{|V_1^c|} \right) \tag{3}$$

$$= \min_{V_1\subset V} \left(\frac{2}{|V_1|} \sum_{(i,j)\in E: i\in V_1, j\in V_1^c} W_{ij} + \frac{2}{|V_1^c|} \sum_{(i,j)\in E: i\in V_1^c, j\in V_1} W_{ij} \right).$$

On réécrit ce problème sous une forme vectorielle. Pour tout $V_1 \subset V$, on construit un vecteur $f^{V_1} = (f_i^{V_1})_{i\in V} \in \mathbb{R}^{|V|}$ définie pour tout $i \in V$ par

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

Proposition 3.1. *Soit L le Laplacien d'un graphe pondéré non orienté. On a :*
1) $\langle f^{V_1}, \mathbf{1}\rangle = 0$ *et* $\|f^{V_1}\|_2^2 = n$
2) $(f^{V_1})^\top L f^{V_1} = n\text{RatioCut}(V_1, {V_1}^c)$.

Démonstration. On a

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

où on a utilisé le calcul

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

De plus, on a

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

et aussi $\langle f^{V_1}, \mathbf{1}\rangle = 0$ car

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

■

Il y a donc équivalence entre les trois problèmes :

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

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

et le ratiocut problem

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

Les trois problèmes sont des problèmes d'optimisation discrète. Comme (7) est NP-hard en général c'est aussi le cas pour les deux autres (5) et (6). On va alors chercher une relaxation convexe pour (6).

La fonction objective de (5) ou (6) est $f \in \mathbb{R}^n \to f^\top L f$ ne pose pas de problème du point de vue de l'optimisation car c'est une fonction dont la Hessienne est $2L$ qui est positive donc la fonction objective est convexe. On est donc en train de minimiser une fonction convexe dans les problèmes (5) ou (6). La principale difficulté computationelle des problèmes (5) et (6) est que l'espace de recherche "$A \subset V$" est discret. On va alors le "convexifier" simplement en le remplaçant par $\mathbb{R}^{|V|}$. On considère alors le problème "relaché de (6)" :

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

Montrons que (8) est bien la méthode spectrale introduite au chapitre précédent. Comme L est symétrique, on peut trouver une base orthonormale de vecteurs propres. Or on sait que $\mathbf{1} = (1)_{i \in V}$ est vecteur propre associé à la valeur propre 0 de L et que, comme L est positive (voir Proposition 2.2), 0 est la plus petite valeur propre de L, (8) revient donc à chercher un vecteur propre associé à la deuxième plus petite valeur propre de L (qui peut être aussi 0 si 0 est de multiplicité plus grande que 2 dans le spectre de L). Un tel vecteur propre apparait souvent dans l'étude des graphes. On lui a alors donné un nom.

Définition 3.2. *Soit L le Laplacien d'un graphe pondéré non orienté. Le* ***Fiedler vector*** *de L est un vecteur propre associé à la deuxième plus petite valeur propre de L. On appelle aussi la deuxième plus petite valeur propre de L la* ***connectivité algébrique****.*

Proposition 3.3. *Les vecteurs de Fiedler de L sont les solutions du problème (8).*

Démonstration. Comme L est symétrique, on peut écrire $L = UDU^\top$ où U est une matrice orthogonale ayant pour vecteurs colonnes les vecteurs propres de L noté $(u_i)_{i \in V}$ et $D = \text{diag}(\lambda)$ où $\lambda = (\lambda_i)_{i \in V}$ est le spectre de L. On note $V = \{1, \dots, n\}$ et $\lambda = (\lambda_i)_{i=1}^n$ tel que $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$. On a $u_1 = \mathbf{1}$ et $\text{vect}(u_2, \cdots, u_n) = \text{vect}(u_1)^\perp$.

Pour tout $f \in \mathbb{R}^{|V|}$ tel que $\|f\|_2^2 = 1$ et $\langle f, \mathbf{1} \rangle = 0$, on a

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

car $\lambda_1 = 0$ et, comme $\langle f, \mathbf{1} \rangle = 0$ on a $\langle f, u_1 \rangle = 0$, donc $1 = \|f\|_2^2 = \sum_{i=2}^n \langle u_i, f \rangle^2$. Or pour $f = u_2$, on a $f^\top L f = \lambda_2$ donc cette borne inférieure est atteinte par un deuxième vecteur propre de L. Seuls ces vecteurs propres normalisés peuvent atteindre cette borne sinon il est facile de voir que $f^\top L f > \lambda_2$ en regardant le cas d'égalité dans (9). ∎

On approche donc une solution du RatioCut problem par la recherche d'un deuxième vecteur propre, Fiedler vector, de L. Une fois calculé un Fiedler vector u_2, on doit en déduire des communautés pour G. Comme u_2 est sensé être une valeur approchante du vecteur d'appartenance aux communautés $f^{V_1{}^*}$ où $V_1{}^* \subset V$ est une solution du RatioCut problem, on espère que les coordonnées de u_2 vont se concentrées autour de deux valeurs (car $f^{V_1{}^*}$ prend exactement 2 valeurs). On fait alors un clustering à deux classes C_1, C_2 sur $\mathbb{R}$ des coordonnées de u_2 pour en déduire les communautés $\{i : u_{2i} \in C_1\}$, $\{i : u_{2i} \in C_2\}$ de G. **C'est exactement la méthode spectrale** dans le cas $k = 2$ quand on prend pour premier vecteur propre $u_1 = \mathbf{1}$ (en effet, clusteriser les lignes de $[u_1|u_2]$ est équivalent à clusteriser les lignes (= coordonnées) de $[u_2]$ vue que u_1 est un vecteur à coordonnées toutes égales).

RatioCut pour k communautés : Dans le cas $k = 2$, à chaque partition $V_1 \sqcup V_1{}^c = V$, on a associé une fonction f^{V_1} définie dans (4). Ici pour le cas général, à chaque partition, on associe une matrice de la manière suivante. Soit $V_1 \sqcup \cdots \sqcup V_k$ une partition de V, on construit $H := H(V_1, \ldots, V_k) \in \mathbb{R}^{|V| \times k}$ tel que ces k vecteurs colonnes sont donnés par $h_1 = \mathbf{1}_{V_1}/\sqrt{|V_1|}, \ldots, h_k = \mathbf{1}_{V_k}/\sqrt{|V_k|}$. Autrement dit la 'membership matrix' H est définie pour tout $i \in V, j = 1, \ldots, k$ par

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

On donne un équivalent à la Proposition 3.1 dans ce cas là.

Proposition 3.4. *La matrice H définie ci-dessus à partir d'une partition $V_1 \sqcup \cdots \sqcup V_k$ de V vérifie :*

1) $H^\top H = I_k$

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

Démonstration. **1)** On a pour tout $p, q = 1, \ldots, k$,

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

Alors si $p \neq q$, comme $V_p \cap V_q = \emptyset$ on a $I(i \in V_p) I(i \in V_q) = 0$ pour tout $i \in V$ et donc $(H^\top H)_{pq} = 0$. Quand $p = q$, on a $I(i \in V_p) I(i \in V_q) = 1$ pour tout $i \in V_p$ et sinon $I(i \in V_p) I(i \in V_q) = 0$ donc

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

On a donc bien $H^\top H = I_k$.

2) On rappelle que $h_1, \ldots, h_k$ sont les vecteurs colonnes de H. On a

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

Pour tout $p = 1, \dots, k$, on a

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

On a donc

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

■

Ainsi d'après la Proposition 3.4, il y a équivalence entre les trois problèmes :

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

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

et

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

Ces deux problèmes sont en général NP-hard. Le problème ne vient pas de la fonction objectif $F : H \in \mathbb{R}^{n\times k} \to \mathrm{Tr}(H^\top L H)$ qui est aussi convexe car sa Hessienne est constante égale à $V \in \mathbb{R}^{n\times k} \to 2L$ qui est positive. Le problème est (comme dans le cas $k = 2$) dû à la contrainte qui est de type combinatoire. On va alors effectuer une relaxation convexe de la contrainte de (11) par

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

On a tout simplement enlevé l'aspect combinatoire de la contrainte qui consiste à ne considérer que des matrices de la forme $H = H(V_1, \dots, V_k)$.

Pour retrouver une solution à (11) à partir d'une solution du problème relaché, on fait un clustering à k clusters des lignes de la matrice solution de (12) parce qu'une solution de (11) est une matrice $n \times k$ ayant seulement k lignes différentes et que si on lance un cluster à k classes sur les n lignes de cette matrice on retrouve les k communautés. C'est donc bien la méthode spectrale à condition qu'on prouve que les vecteurs colonnes d'une solution au problème relaché (12) engendrent un sous-espace propre de dimension k associé aux plus petites valeurs propres de L. C'est ce que nous faisons maintenant et qui concluera sur le lien entre la méthode spectrale et une relaxation convexe du problème de RatioCut.

En écrivant la SVD de $L = UDU^\top$ où $D = \mathrm{diag}(\lambda)$, $\lambda = (\lambda_i)_{i\in V}$ est le spectre de L tel que $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ et $U = [u_1|\cdots|u_n]$ quand $V = \{1, \ldots, n\}$, on a

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

où, pour tout $i \in V$, a_i est le i-ième vecteur ligne de $U^\top H$ càd $a_i = (\langle u_i, h_j\rangle)_{j=1,\ldots,k}$. Par ailleurs, comme $(h_1, \ldots, h_k)$ est une famille de vecteurs orthonormaux (car $H^\top H = I_k$), on a

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

$\|a_i\|_2^2 \leq \|u_i\|_2^2 = 1$ pour tout $i \in V$ et que

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

qui est atteint lorsque $\mathrm{vect}(h_1, \ldots, h_k) = \mathrm{vect}(u_1, \ldots, u_k)$. On voit donc que les solutions de (13) sont les matrices H orthogonales dont les colonnes engendrent le sev de dimension k associé à k plus petites valeurs propres de L.

On retombe bien sur le problème qui consiste à trouver k vecteurs propres de L associés aux k plus petites valeurs propres de L en tenant compte de leurs multiplicité. Ensuite, on clusterise les n vecteurs lignes de H en k clusters dont on déduit les communautés de L.

3.3 Pseudo-algorithme de la méthode spectrale à k communautés

Soit G un graphe non orienté et pondéré ou pas. On note par A sa matrice d'adjacence et par $L = D - A$ son Laplacien où D est la matrice diagonale des degrés des nœuds de G. La méthode spectrale pour la détection de k communautés dans G a pour pseudo-code :

1 **Input :** L Laplacien du graphe G à n nœuds et k le nombre de communautés
2 **Output :** Partition des nœuds de G en k communautés
3 Trouver une base orthogonale de k vecteurs propres de L associés aux k plus petites valeurs propres de L (en tenant compte de leur multiplicté).
4 On note H la matrice de taille $n \times k$ ayant pour vecteurs colonnes les k vecteurs propres de L précédemment calculés.
5 On clusterise en k clusters les n vecteurs lignes $y_1, \ldots, y_n \in \mathbb{R}^k$ de H : on obtient k clusters $C_1, \ldots, C_k$ formant une partition des lignes $y_1, \ldots, y_n$.
6 On retourne les k communautés $\{i \in \{1, \ldots, n\} : y_i \in C_p\}, p = 1, \ldots, k$.

Algorithm 1: Méthode spectrale pour la détection de communautés basée sur l'étude du spectre du Laplacien du graphe.

Cet algorithme se justifie soit par la relaxation convexe du problème de Ratiocut introduit dans la Section 3.2 ou par le cadre idéal des composantes connexes de la Section 3.1. Le choix du nombre k de communautés peut se faire par une méthode du coude sur le spectre du Laplacien (en partant de la plus petite valant propre de L valant 0 et en allant vers les plus grandes,

jusqu'à la rencontre d'un coude, càd une inflection plus prononcée) soit en évaluant une fonction de modularité comme la fonction RatioCut en chaque solution en fonction de k.

4 Modèles probabilistes de graphes, méthode spectrale et relaxation SDP pour la détection de communautés

Idée : Dans les chapitres précédents, on a supposé que toutes les arêtes du graphe sont observées. Il y a cependant des cas où certains liens n'ont pas pu être observés ou établis avec certitude (liens entre gènes, compagnies, etc.). Dans ce cas, une possibilité, est de supposer un modèle probabiliste sous-jacent à nos observations. En quelque sorte, on observe que partiellement la matrice d'adjacence d'un graphe mais on souhaite toujours identifier une structure de communautés au sein de ce graphe. On va ici utiliser la méthode spectrale des sections précédentes mais appliquées seulement au Laplacien de la matrice d'adjacence partiellement observée.

4.1 Modèles probabilistes de graphes et estimateur du maximum de vraisemblance.

On présente dans cette section deux modèles probabilistes de graphes aléatoires.

Le modèle de Erdös-Rényi : On dit que le graphe aléatoire G suit le modèle d'Erdös-Rényi à n nœuds et de paramètre p, et on note $G \sim G(n,p)$, quand G est un graphe non orienté non pondéré sur n nœuds dont la matrice d'adjacence est donnée par $A = (\delta_{ij})_{1\leq i,j\leq n}$ où $\delta_{ij} = \delta_{ji}$, $\delta_{ii} = 1$ et $(\delta_{ij} : j > i)$ sont des variables aléatoires de Bernoulli de paramètre p indépendantes.

Stochastic Block Model (SBM) : On dit qu'un graphe aléatoire suit le SBM sur n noeuds et de paramètres p, q où $0 \leq q < p \leq 1$, et on note $G \sim G(n,p,q)$ quand $G = (V,E)$ est un graphe non orienté et non pondéré de matrice d'adjacence $A = (\delta_{ij})_{1\leq i,j\leq n}$ telle qu'il existe une partition $V_1 \sqcup V_1^c = V$ des sommets de G pour laquelle on a pour tout $i < j$

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

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

Autrement dit, dans un SBM, les nœuds appartenant à la même communauté sont connectés avec proba p et s'ils n'appartiennent pas à la même communauté alors ils sont connectés avec probabilité q. Comme $p > q$ on s'attend à ce qu'il y ait une plus forte densité de nœuds intra-communautés qu'en inter-communautés.

Le SBM modélise donc les graphes organisés en deux communautés au sens où on l'a définit au début : une communauté a une plus forte densité de liens en interne qu'en externe avec son complémentaire. Le problème qu'on va chercher à résoudre est le suivant : étant donné un graphe G tiré selon le SBM de paramètre (n,p,q), comment retrouver les deux communautés V_1 et V_1^c de G ?

Estimateur du maximum de vraisemblance dans le SBM à deux communautés de même taille. L'avantage d'avoir un modèle statistique sur nos données est qu'on peut utiliser la vraisemblance du modèle pour en déduire une procédure d'estimation des communautés V_1 et V_1^c. Cette approche est celle du maximum de vraisemblance : on choisit le paramètre V_1 qui maximise 'la probabilité d'avoir observé ce qu'on a observé', càd la vraisemblance. On commence

par donnée cette vraisemblance : étant donné la matrice d'adjacence observée A, pour tout $V_1 \subset V$ et $0 < p < q < 1$, on a

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

où $f^{SBM}_{V_1,p,q}$ est la densité sur les matrices d'adjacence de graphes non orientés de diagonale égale à $(1)^n_1$ suivant le modèle SBM à deux communautés $V_1 \sqcup V_1^c$ et de paramètres p, q. On note $\mathcal{C} = \{(i,j) \in V_1 \times V_1 \cup V_1^c \times V_1^c, i < j\}$ et $\mathcal{C}^c = \{(i,j) \in V_1 \times V_1^c \cup V_1^c \times V_1, i < j\}$. La log-vraisemblance en (V_1, p, q) est donc donnée par

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

où $N_{\mathcal{C}} = \sum_{(i,j)\in\mathcal{C}} A_{ij}$, $N_{\mathcal{C}^c} = \sum_{(i,j)\in\mathcal{C}^c} A_{ij}$, et $|\mathcal{C}|$ (resp. $|\mathcal{C}^c|$) est le cardinal de $\mathcal{C}$ (resp. $\mathcal{C}^c$). L'estimateur du maximum de vraisemblance consiste à maximiser cette quantité en (V_1, p, q) sous la contrainte que $1 > p \geq q > 0$. Pour cela, on maximise cette quantité en (p, q) sous la contrainte que $p \geq q$ à V_1 fixe et ensuite on maximisera en V_1. On effectue cette maximisation dans le cas où les deux communautés sont de même taille. Dans ce cas, on a $|\mathcal{C}| = |\mathcal{C}^c| = (n^2 - n)/4 := N/2$.

Soit $V_1 \subset \{1, \ldots, n\}$ tel que $|V_1| = n/2$. On souhaite résoudre le problème

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

où $P_{\mathcal{C}} = N_{\mathcal{C}}/|\mathcal{C}|$ et $P_{\mathcal{C}^c} = N_{\mathcal{C}^c}/|\mathcal{C}^c|$. On peut résoudre ce problème en utilisant le thèorème de KKT (voir à la fin de la section). On trouve que si $P_{\mathcal{C}} > P_{\mathcal{C}^c}$ alors $p = P_{\mathcal{C}}$ et $q = P_{\mathcal{C}^c}$ est l'unique solution et dans ce cas la valeur prise par la fonction objective en cette solution vaut

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

Sinon, quand $P_{\mathcal{C}} \leq P_{\mathcal{C}^c}$, l'unique solution de (15) est $p = q = (P_{\mathcal{C}} + P_{\mathcal{C}^c})/2$. Dans ce cas et comme $P_{\mathcal{C}} + P_{\mathcal{C}^c} = (N_{\mathcal{C}} + N_{\mathcal{C}^c})/(N/2) = (2/N) \sum_{i<j} A_{ij}$ est constant (càd indépendant de V_1), la valeur prise par la fonction objective en son optimum est indépendant de V_1 et vaut aussi (16) dans le cas où $P_{\mathcal{C}} = P_{\mathcal{C}^c}$. On en conclut que le problème d'optimisation sur $V_1 \subset \{1, \ldots, n\}$ tel que $|V_1| = n/2$ qui reste à résoudre est celui de maximiser la valeur optimale obtenue en (16) sous la contrainte que $P_{\mathcal{C}} \geq P_{\mathcal{C}^c}$, càd

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

Une fois résolu ce problème en V_1, on obtiendra l'EMV $\hat{p} = N_{\mathcal{C}}/|\mathcal{C}| = P_{\mathcal{C}}$ et $\hat{q} = N_{\mathcal{C}^c}/|\mathcal{C}^c| = P_{\mathcal{C}^c}$ pour le choix optimal de V_1. On peut interpréter la contrainte '$P_{\mathcal{C}} \geq P_{\mathcal{C}^c}$' comme une version empirique de la contrainte $p \geq q$ vue que les quantités $P_{\mathcal{C}}$ et $P_{\mathcal{C}^c}$ sont les EMV de p et q pour le choix optimal de V_1.

Il reste à résoudre le problème d'optimisation (17). On démontre à la fin de cette section que la fonction objectif du problème (17) est maximale en $(P_{\mathcal{C}}, P_{\mathcal{C}^c}) \in [0,1]^2$ sous la contrainte que $P_{\mathcal{C}} + P_{\mathcal{C}^c} = (2/N) \sum_{i<j} A_{ij}$ et $P_{\mathcal{C}} \geq P_{\mathcal{C}^c}$ quand $P_{\mathcal{C}}$ est maximal ou de manière équivalente quant $P_{\mathcal{C}^c}$ est minimal (vu que leur somme est constante). Il reste alors à trouver les ensembles $V_1 \subset \{1, \cdots, n\}$ tels que $|V_1| = n/2$ pour lesquels $P_{\mathcal{C}^c}$ est minimal. Par ailleurs, on remarque que pour tout $V_1 \subset \{1, \ldots, n\}$ tel que $|V_1| = n/2$, on a

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

On doit donc minimiser en $V_1 \subset V$ tel que $|V_1| = n/2$ le ratiocut $\mathrm{RatioCut}(V_1, V_1^c)$. On retrouve donc le problème de minimisation de RatioCut comme introduit dans la Section 3.2 dans le cas où il y a deux communautés de même taille. **On voit donc que le problème de minimisation du RatioCut dans le cas de deux communautés de même taille est aussi un problème de calcul de maximum de vraisemblance dans le Stochastic block model**.

Remarque 4.1. *Dans le cas de la détection de deux communautés de même taille le problème du RatioCut minimal et de l'EMV sont équivalents et reviennent à résoudre*

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

car pour tout $V_1 \subset [V]$ tel que $|V_1| = |V_1{}^c|$, on a $f^{V_1} \in \{-1,1\}^n$. Par ailleurs, on a aussi pour tout $x \in \{-1,1\}^n$ que $x^\top Lx = -x^\top Ax + \sum_{i,j} A_{ij}$, car $x^\top Dx = \sum_{i,j} D_{ij}x_ix_j = \sum_i d_i x_i^2 = \sum_i d_i$ où d_i est le degrés du nœud i. Comme le terme $\sum_{i,j} A_{ij}$ est constant (indépendant de x), on voit que (18) est équivalent à

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

C'est sur ce problème que nous effectuerons une relaxation SDP en Section 6.

On conclut cette section avec la résolution des deux problèmes d'optimisation (15) et (17). On commence avec (15). On voit que la fonction objective $f : (p,q) \in]0,1[^2 \to (1-P_{\mathcal{C}})\log(1-p) + P_{\mathcal{C}}\log(p) + (1-P_{\mathcal{C}^c})\log(1-q) + P_{\mathcal{C}^c}\log(q)$ est fortement concave et que ses ensembles de niveau sont compacts donc (15) admet une unique solution. Soit (p,q) une solution au problème. D'après KKT, il existe $\mu \geq 0$ tel que $-\nabla f(p,q) + \mu \nabla h(p,q) = 0$ et $\mu h(p,q) = 0$ où $h : (p,q) \to q - p$. On a alors $\mu \geq 0$, $\mu(p-q) = 0$, $p \geq q$ et

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

Si $p \neq q$ alors $\mu = 0$ et donc, d'après (20), $p = P_{\mathcal{C}}$ et $q = P_{\mathcal{C}^c}$ est bien l'unique solution à condition que $P_{\mathcal{C}} > P_{\mathcal{C}^c}$. Si $p = q$ alors d'après (20), on a $p = q = (P_{\mathcal{C}} + P_{\mathcal{C}^c})/2$ qui est l'unique solution à condition que $P_{\mathcal{C}} < P_{\mathcal{C}^c}$ (sinon $\mu < 0$) et quand $P_{\mathcal{C}} = P_{\mathcal{C}^c}$, on a $p = q = P_{\mathcal{C}}$ qui est l'unique solution. En fait, le problème (15) est un problème d'optimisation convexe différentiable et on peut voir que quand $P_{\mathcal{C}} > P_{\mathcal{C}^c}$ alors $((P_{\mathcal{C}}, P_{\mathcal{C}^c}), 0)$ est un point-selle de la fonction de Lagrange alors il y a dualité forte et $(P_{\mathcal{C}}, P_{\mathcal{C}^c})$ est l'unique solution du primal (càd de (15)) (et 0 est solution du problème duale). Quand $P_{\mathcal{C}} \leq P_{\mathcal{C}^c}$ alors $((P_{\mathcal{C}}+P_{\mathcal{C}^c})/2, (P_{\mathcal{C}}+P_{\mathcal{C}^c})/2), (P_{\mathcal{C}^c}-P_{\mathcal{C}})/[(P_{\mathcal{C}}+P_{\mathcal{C}^c})(1-(P_{\mathcal{C}}+P_{\mathcal{C}^c})/2)]$ est un point-selle de la fonction de Lagrange donc $p = q = (P_{\mathcal{C}} + P_{\mathcal{C}^c})/2$ est l'unique solution du problème (15).

On traite le problème (17). On donne ici la preuve que la fonction objectif du problème (17) est maximale en $(P_{\mathcal{C}}, P_{\mathcal{C}^c}) \in]0,1[^2$ sous la contrainte que $P_{\mathcal{C}} + P_{\mathcal{C}^c} = (2/N)\sum_{i<j} A_{ij} := s$ est constante et $P_{\mathcal{C}} \geq P_{\mathcal{C}^c}$ quand $P_{\mathcal{C}}$ est maximal ou de manière équivalente quand $P_{\mathcal{C}^c}$ est minimal (vu que leur somme est constante). Comme $P_{\mathcal{C}} + P_{\mathcal{C}^c} = s$ est constant, il suffit de minimiser $P_{\mathcal{C}} \to \mathrm{Ent}(P_{\mathcal{C}}) + \mathrm{Ent}(s - P_{\mathcal{C}})$ sous la contrainte que $s/2 \leq P_{\mathcal{C}} \leq \min(s,1)$. Or la fonction $F : p \in [0,1] \to \mathrm{Ent}(p) + \mathrm{Ent}(s-p)$ est concave et maximale en $s/2$ donc son minimum sur $s/2 \leq p \leq \min(s,1)$ est atteint en $\min(s,1)$ et donc $P_{\mathcal{C}} \to F(P_{\mathcal{C}})$ est minimal sur l'ensemble des valeurs prises par $P_{\mathcal{C}}$ quand $P_{\mathcal{C}}$ est maximal.

Seuil de détectabilité dans le SBM à deux communautés de même taille. On considère un graphe obtenu à partir d'un SBM de paramètre (n, p, q) où $p > q$ à deux communautés V_1, V_1^c de même taille $n/2$. On souhaite retrouver les deux communautés à partir d'une observation de ce graphe. On peut remarquer que s'il y a un point de la communauté V_1 qui est isolé des autres noeuds de V_1, alors on n'aura aucune raison de croire que ce noeud est bien dans la communauté V_1. Dans ce cas, la reconstruction des communautés n'est pas possible. Dans le résultat suivant, on donne donc un 'niveau de signal' minimal sur p pour qu'il n'y ait pas de point isolé dans un graphe SBM.

Proposition 4.2. *Si $p \leq \log(n/2)/[n/2-2]$ alors la probabilité qu'il y ait un nœud isolé dans la communauté V_1 est plus grande que $1/2$.*

Preuve. Pour tout $i \in V_1$, on note par A_i l'événement que le nœud i est isolé des autres nœuds de la communauté V_1. Formellement, on a

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

On note $N = |V_1| - 1$. On a donc $\mathbb{P}[A_i] = (1-p)^N$. On a en particulier, $\mathbb{P}[A_i] \geq 1/2$ si et seulement si $p \leq 1 - \exp(-(\log 2)/N)$ et donc pour $p \leq \log 2/(2N)$ on aura bien que i est un nœud isolé avec probabilité au moins $1/2$. On obtient le résultat à un facteur log près.

Pour obtenir le résultat, on va utiliser le fait qu'on souhaite avoir un noeud isolé parmi les $|V_1|$ possible. C'est cet aspect là qui va nous faire gagner le $\log N$ en plus dans le seuil. On s'intéresse alors à la probabilité $\mathbb{P}[\cup_{i \in V_1} A_i]$. On montre par récurrence que

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

On voit que si $i \neq j$ alors

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

et donc

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

On cherche ensuite à déterminer une condition sur p qui implique que $(N+1)(1-p)^N - (N+1)N(1-p)^{2N-1} \geq 1/2$. On pose $X = (N+1)(1-p)^N$. On souhaite alors résoudre $X - X^2 N/[(N+1)(1-p)] \geq 1/2$. Cette inéquation est vraie quand

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

En particulier, quand $2N(1-p)^{N-1} \leq 2$, cette inégalité est vraie. On voit alors que pour $p \leq (\log N)/(N-1)$, on a bien que $\mathbb{P}[\cup_{i \in V_1} A_i] \geq 1/2$. ∎

Une conséquence de la Proposition 4.2 est que si on se pose le problème de reconstruction exacte des communautés sous-jacentes à un graphe tiré selon le SBM alors on doit supposer que $p \geq \alpha \log(n)/n$. Comme $p > q$, on doit avoir $q < \alpha \log(n)/n$. Le cas difficile est alors de considérer q le plus proche de p dans ce cadre. On va alors aussi prendre un q de la forme $\beta \log(n)/n$ où $\beta < \alpha$. On s'intéresse dans la suite aux conditions sur α et β qui assurent la détection exactes des communautés avec grande probabilité par une procédure réalisable en pratique.

Si on utilisait le maximum de vraisemblance : $\hat{V}_1 \in \operatorname{argmin}_{\{V_1 \subset V : |V_1| = n/2\}} \mathrm{RatioCut}(V_1, V_1^c)$ alors on pourrait atteindre le seuil d'impossibilité de reconstruction exacte obtenue par la théorie de l'information comme énoncé dans le théorème suivant (que nous ne démontrons pas).

Théorème 4.3 (Transition de phase dans le SBM à 2 classes de même taille)**.** *Si* $\sqrt{\alpha}-\sqrt{\beta}>\sqrt{2}$ *alors avec grande probabilité* $\hat{V}_1$ *et* $\hat{V}_1^c$ *sont les communautés sous-jacentes* V_1 *et* V_1^c*. Si* $\sqrt{\alpha}-\sqrt{\beta}<\sqrt{2}$ *alors aucun algorithme ne peut retrouver les communautés* V_1 *et* V_1^c *avec probabilité au moins* $1/2$*.*

Ce Théorème est un résultat donnant une transition de phase très précise : si $\sqrt{\alpha}-\sqrt{\beta}$ est au-dessus de $\sqrt{2}$ alors on peut retrouver les communautés et c'est impossible si on est en dessous de ce seuil. Néanmoins, ce résultat est intéressant d'un point de vue théorique car en pratique, on ne va pas utiliser le maximum de vraisemblance mais une procédure calculable numériquement : soit une méthode spectrale soit une relaxation SDP. On s'intéresse alors à identifier un seuil portant sur α et β permettant de retrouver les communautés sous-jacentes.

4.2 Méthodes spectrales pour le SBM à deux classes de même taille basées sur la recherche d'un vecteur de Fiedler

On considère un graphe $G=(V,E)$ tiré selon le SBM de paramètre (n,p,q) avec $p>q$ et $V=\{1,\dots,n\}$. On note par A la matrice d'adjacence (observée) de G comme définie dans (14). Comme vu au chapitre précédent la méthode spectrale consiste ici à trouver un vecteur de Fiedler du Laplacien de G càd de $L=D-A$ où $D=\mathrm{diag}(d_1,\dots,d_n)$ où $d_i=\sum_{j=1}^n A_{ij}$ est le degré du i-ième nœud de G. Une fois un vecteur de Fiedler obtenu on clusterise ces coordonnées en deux groupes qui vont nous donner les deux communauté de G idéalement. C'est ce qu'on aimerait démontrer ici avec grande probabilité.

On réécrit la méthode spectrale plus simplement quand les deux communautés V_1 et V_1^c ont même cardinal grâce au lemme suivant.

Lemme 4.4. *Soit* $A\in\mathbb{R}^{n\times n}$ *la matrice d'adjacence d'un graphe* $G=(V,E)$ *tiré selon le SBM de paramètre* (n,p,q) *où* n *est pair. On suppose que les deux communautés* V_1,V_1^c *sous-jacentes au modèle SBM sont de même taille* $n/2$*. Il y a équivalence entre :*

1) u_2 *est un vecteur de Fiedler de* $\mathbb{E}L=\mathbb{E}(D-A)$

2) u_2 *est un vecteur propre associé à la plus grande valeur propre de*

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

où $\mathbf{1}=(1)_{i=1}^n$*.*

Démonstration. Quitte à réordonner les noeuds de V, on peut écrire $\mathbb{E}A$ sous la forme

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

On voit que la matrice moyenne des degré est alors proportionnelle à l'identité car pour tout noeud $i\in\{1,\dots,n\}$, le degré moyen du nœud i est $\mathbb{E}d_i=1+(n/2-1)p+(n/2)q$. On a donc $\mathbb{E}D=(1+(n/2-1)p+(n/2)q)I_n$.

Comme $\mathbb{E}D$ est proportionnel à l'identité, on voit que f est un vecteur propre de $\mathbb{E}L$ de valeur propre λ si et seulement si c'est un vecteur propre de $\mathbb{E}A$ de valeur propre $[1+(n/2-1)p+(n/2)q]-\lambda$. Ainsi chercher un vecteur de Fiedler pour $\mathbb{E}L$ est équivalent à chercher un vecteur propre pour $\mathbb{E}A$ pour sa deuxième valeur propre.

Par ailleurs, comme 0 est la plus petite valeur propre de $\mathbb{E}L$ (car c'est le Laplacien du graphe dont la matrice d'adjacence est donnée par $\mathbb{E}A$) associé au vecteur propre $\mathbf{1}=(1)_1^n$, $1+(n/2-1)p+(n/2)q$ est la plus grande valeur propre de $\mathbb{E}A$ associé au même vecteur propre $\mathbf{1}=(1)_1^n$. En écrivant la décomposition en valeur singulières de $\mathbb{E}A=\sum_i([1+(n/2-1)p+(n/2)q]-\lambda_i)u_i\otimes u_i$, où $0=\lambda_1\leq\lambda_2\leq\cdots\leq\lambda_n$ sont les valeurs propres de $\mathbb{E}L$ associées aux vecteurs propres respectifs $u_1,\ldots,u_n$ (formant une base orthonormale de $\mathbb{R}^n$), on voit que $\mathbb{E}A-[1+(n/2-1)p+(n/2)q]u_1\otimes u_1$ a pour plus grande valeur propre $[1+(n/2-1)p+(n/2)q]-\lambda_2$ qui est la deuxième valeur propre la plus grande de $\mathbb{E}A$. Il est donc équivalent de chercher un vecteur propre de $\mathbb{E}A$ associé à sa deuxième plus grande valeur propre que de chercher un vecteur propre associé à la plus grande valeur propre de $\mathbb{E}A-[1+(n/2-1)p+(n/2)q]u_1\otimes u_1$. Hors, on peut prendre $u_1=\mathbf{1}/\|\mathbf{1}\|_2$. Ce qui conclut la preuve. ■

Dans le cas de deux communautés de même taille, la méthode spectrale consiste donc, d'après le Lemme 4.4, à chercher un vecteur propre associé à la plus grande valeur propre de la matrice (21) (et ensuite à clusteriser ses coordonnées).

Cependant, on ne connaît ni $\mathbb{E}A$, ni p ni q. On va donc estimer ces quantités : on utilise A pour estimer $\mathbb{E}A$ et pour $(p+q)/2$, on remarque que pour

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

on a

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

Ainsi λ estime $(p+q)/2$ à un terme en $1/n$ près (λ est aussi la moyenne des estimateurs par maximum de vraisemblance de p et q obtenus dans la Section 4.1). On va alors estimer un plus grand vecteur propre de (21) par un plus grand vecteur propre de $A-n\lambda(\mathbf{1}/\sqrt{n})(\mathbf{1}/\sqrt{n})^\top$ càd une solution au problème

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

où $J_n=(1)_{n\times n}$. Ce type de problème porte le nom de **méthode spectrale** car in fine on est amené à calculer un vecteur propre d'une matrice pour résoudre le problème.

4.3 *Matrix lifting* et relaxation SDP pour la recherche de communautés dans le SBM à deux classes de même taille.

Le problème de détection de deux communautés s'écrit, initialement, comme un problème de recherche d'une solution au problème discret

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

comme introduit dans (6) où $\mathbb{E}L$ était entièrement connu (et pas seulement au travers de la seule observation L). Ce problème a été relaxé en supprimant la contrainte que f devait être de la forme f^{V_1} pour un certain $V_1\subset V$ (où f^{V_1} est défini dans (4)). Cette relaxation a motivé l'introduction de la recherche d'un vecteur de Fiedler du Laplacien. On peut cependant proposer d'autres relaxations convexes.

Par exemple, dans le cas de deux classes de même taille, on a vu que chercher un vecteur de Fiedler de $\mathbb{E}L$ est équivalent à chercher un premier vecteur propre de (21) qui peut être estimé par $A - \lambda J_n$ où λ est défini par (22) et $J_n = (1)_{n\times n}$. On peut alors espérer qu'une solution au problème

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

peut être une bonne solution approchante d'une solution du problème initial (24). Par ailleurs, sous l'hypothèse de communautés de tailles égales, on voit que $f^A \in \{-1, 1\}^n$ est tel que $(f^A)_i = 1$ si $i \in V_1$ et $(f^A)_i = -1$ si $i \notin V_1$. Le problème (25) peut donc se réécrire comme

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

qui est lui-même équivalent à

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

où on utilise ici le produit scalaire entre deux matrices données par $\left\langle A, B \right\rangle = \sum_{i,j} A_{ij} B_{ij} = \mathrm{Tr}(AB^\top)$. Le passage d'un problème d'optimisation avec une fonction objectif quadratique en x à un problème d'optimisation avec une fonction objectif linéaire en $xx^\top$ est connue sous le nom de **matrix lifting** et est un classique de la relaxation convexe (voir la section 7 pour d'autres exemples).

Le problème (27) est un problème combinatoire et donc potentiellement difficile à résoudre directement. On va utiliser une relaxation convexe pour l'approcher. Cette approche qui va simplement consister à supprimer une contrainte de rang porte le nom de **relaxation SDP**. Pour cela, on voit $Z = xx^\top$ dans le problème (27) comme une variable matricielle ayant les propriétés suivantes :

i) Z est symétrique
ii) Z est positive car $\left\langle xx^\top y, y \right\rangle = \left\langle x, y \right\rangle^2 \geq 0$ pour tout $y \in \mathbb{R}^n$
iii) $\mathrm{diag}(Z) = \mathrm{diag}(x_1^2, \ldots, x_n^2) \preceq I_n$ car $x_i^2 = 1$ pour tout $n = 1, \ldots, n$ (on note par $\mathrm{diag}(Z)$ la matrice diagonale de $\mathbb{R}^{n\times n}$ dont les éléments diagonaux coïncident avec ceux de Z).
iv) $\mathrm{rang}(Z) = 1$.

On peut réécrire la contrainte sous la forme suivante :

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

On va alors faire une relaxation convexe de (27) tout simplement en enlevant la contrainte de rang dans l'ensemble ci-dessus (qui est la cause de non-convexité de la contrainte). En faisant cela, il ne reste plus que des contraintes affines ($\mathrm{diag}(Z) \preceq I_n$ étant équivalent à avoir $Z_{ii} \leq 1, \forall i = 1, \ldots, n$, càd $\left\langle E_{ii}, Z \right\rangle \leq 1, \forall i$)) et la contrainte SDP $Z \succeq 0$. Ainsi, vue que la fonction objectif est linéaire, on trouve un problème SDP *semi-definite programming*. C'est la raison pour laquelle on parle de relaxation SDP. On va alors chercher une solution au problème

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

où on rappelle que $Z \succeq 0$ signifie que Z est symétrique et que $\left\langle Zy, y \right\rangle \geq 0$ pour tout $y \in \mathbb{R}^n$. On utilise aussi que $A \succeq B$ quand $A - B \succeq 0$.

Si $\hat{Z}$ est solution de (28), on prend ensuite un plus grand vecteur propre (un vecteur propre associé à sa plus grande valeur propre) de $\hat{Z}$ dont on prend le signe. Les coordonnées de signe 1

forment une communauté et les autres de signe -1 forment l'autre communauté. On espère que cette procédure puisse bien estimer une solution $x^* \in \{-1,1\}^n$ du problème (26) mais aussi et surtout du problème initial (24) avec grande probabilité. C'est l'objet des sections suivantes de le démontrer.

On s'assure d'abord que le vecteur d'appartenance aux communautés $\bar{x} \in \{-1,1\}^n$ défini par $\bar{x}_i = 1$ si $i \in V_1$ et $\bar{x}_i = -1$ si $i \notin V_1$ est bien tel que $\bar{x}(\bar{x})^\top$ est l'unique solution du problème relaxé où A et λ sont remplacés par les valeurs 'oracles' $\mathbb{E}A$ et α :

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

où $\alpha = [(p+q)/2]n + (1-p)$ et $u_1 = \mathbf{1}/\sqrt{n}$. On rappelle aussi que $u \otimes v = uv^\top = (u_i v_j)_{1 \le i,j \le n} \in \mathbb{R}^{n \times n}$.

Proposition 4.5. *Le problème (29) admet une unique solution donnée par $\bar{x} \otimes \bar{x}$.*

Démonstration. On note $B = \mathbb{E}A - \alpha u_1 \otimes u_1$. On commence par étudier le spectre de B. On a

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

On note $u_2 = \bar{x}/\sqrt{n}$. On a

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

Comme $\langle u_1, u_2 \rangle = 0$ et que $\|u_1\|_2 = \|u_2\| = 1$, (u_1, u_2) forme le début d'une base orthonormale qu'on peut compléter : soit $(u_i)_{i=3}^n$ tel que $(u_i)_{i=1}^n$ forme une base orthonormale de $\mathbb{R}^n$. On écrit

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

On en déduit que

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

On voit ensuite grâce au théorème spectral qu'il y a équivalence entre les deux problèmes :

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

2) $\bar{Z} = \bar{X}(\bar{X})^\top$ et
$$\bar{X} \in \operatorname*{argmax}_{\substack{X \in \mathbb{R}^{n \times n} \\ X_{i\bullet} \in B_2^n}} \big\langle B, XX^\top \big\rangle \tag{31}$$

où, pour tout $i = 1, \dots, n$, $X_{i\bullet}$ est le $i-$ième vecteur ligne de X et $X_{i\bullet} \in B_2^n$ signifie que $\|X_{i\bullet}\|_2 \le 1$.

Soit $X \in \mathbb{R}^{n \times n}$ tel que pour tout $i = 1, \dots, n$ on a $X_{i\bullet} \in B_2^n$. On note par $X_{\bullet j}$ le j-ième vecteur colonne de X. On a

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

car $X_{\bullet k} \otimes X_{\bullet k} = (X_{ik}X_{jk})_{1\leq i,j\leq n}$. On remarque que $\langle u \otimes u, v \otimes v\rangle = \mathrm{Tr}(uu^\top vv^\top) = \langle u, v\rangle^2$ pour tout $u, v \in \mathbb{R}^n$. Comme $p > q$ et que $(u_i)_{i=1}^n$ est une base orthonormale de $\mathbb{R}^n$, on a

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

Par ailleurs, en explorant le cas d'égalité dans la majoration de $\langle B, XX^\top\rangle$ dans (32), on voit que nécessairement $\langle u_i, X_{\bullet k}\rangle = 0$ pour tout $i = 1$ et $i = 3, \ldots, n$ et $k = 1, \ldots, n$. On doit aussi avoir $\|X_{i\bullet}\|_2 = 1$ pour tout $i = 1, \ldots, n$. Donc les colonnes $X_{\bullet k}$ de X sont nécessairement portées par u_2. En écrivant X comme étant de la forme $[a_1u_2|a_2u_2|\cdots|a_nu_2]$ pour des réels $a_1, \ldots, a_n$. On a $\|X_{i\bullet}\|_2 = 1$ pour tout $i = 1, \ldots, n$ si et seulement si $\left(\sum_{i=1}^n a_i^2\right)^{1/2} |u_{2i}| = 1$ pour tout $i = 1, \ldots, n$ et comme $|u_{2i}| = 1/\sqrt{n}$, on a donc nécessairement $\left((1/n)\sum_{i=1}^n a_i^2\right)^{1/2} = 1$.

D'un autre côté, on voit que pour $\bar{X}$ ayant pour vecteurs colonnes $a_iu_2, i = 1, \ldots, n$ pour des réels $a_1, \ldots, a_n$ tels que $\left((1/n)\sum_{i=1}^n a_i^2\right)^{1/2} = 1$, on a $\bar{X}_{i\bullet} = (a_1u_{2i}, a_2u_{2i}, \cdots, a_nu_{2i})^\top \in \mathbb{R}^n$ est tel que $\left\|\bar{X}_{i\bullet}\right\|_2 = \left(\sum_{i=1}^n a_i^2\right)^{1/2} |u_{2i}| = 1$ pour tout $i = 1, \ldots, n$ car $|u_{2i}| = 1/\sqrt{n}$. Donc $\bar{X}$ est bien dans l'ensemble de contrainte de (31). Par ailleurs, $\bar{X}(\bar{X})^\top = (\sum_i a_i^2)u_2 \otimes u_2 = nu_2 \otimes u_2$ alors

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

car $\langle u_2 \otimes u_2, u_2 \otimes u_2\rangle = \langle u_2, u_2\rangle^2 = 1$. Donc la borne $n\left[n\left((p-q)/2\right) + (1-p)\right]$ est atteinte par $\bar{X}$ dans (32).

Donc les solutions de (31) sont toutes de la forme $\bar{X} = [a_1u_2|\cdots|a_nu_2]$ où $\sum_i a_i^2 = n$. Dans ce cas, les solution de (30) sont de la forme $\bar{Z} = \bar{X}(\bar{X})^\top = (\sum_i a_i^2)u_2 \otimes u_2 = nu_2 \otimes u_2 = \bar{x} \otimes \bar{x}$ qui est donc unique. ∎

Comme $\bar{x} \otimes \bar{x}$ est de rang 1 et que son unique espace propre associé à la valeur propre n est engendré par $\bar{x}$, on voit bien qu'en prenant le signe de $\lambda\bar{x}$ (un plus grand vecteur propre de $\bar{x} \otimes \bar{x}$) pour tout $\lambda \in \mathbb{R}$ on obtient $\bar{x}$ ou $-\bar{x}$. On peux donc retrouver les communautés du graphes en prenant le signe d'un plus grand vecteur propre d'une solution au problème (29). Ceci motive la méthode spectrale qui consiste à prendre le signe d'un plus grand vecteur propre d'une solution du problème (29) où $\mathbb{E}A - \alpha u_1 \otimes u_1$ a été remplacé par son estimation $A - \lambda J_n$. On écrit cette méthode spectrale en pseudo-code :

Input : A : matrice d'adjacence d'un graphe distribué selon le SBM à deux communautés de même taille.
Output : Partition des nœuds de G en 2 communautés de même taille
Calcul de λ donné dans (22)
Résolution du problème SDP de (28)
Recherche d'un plus grand vecteur propre $\hat{x}$ de $\hat{Z}$ solution de (28)
On prend le signe de $\hat{x}$
On retourne les 2 communautés $\hat{V}_1 = \{i \in \{1, \dots, n\} : \text{sign}(\hat{x}_i) = 1\}$ et son complémentaire $\hat{V}_1^c$.

Algorithm 2: Méthode spectrale pour le détection de communautés basée sur une relaxation SDP.

5 Étude statistique d'une relaxations SDP : vitesse d'estimation et inégalité de Grothendieck.

Problème : On observe la matrice d'adjacence A d'un graphe distribué selon un SBM à deux communautés de même taille. On souhaite retrouver ces deux communautés à partir de A. On a introduit dans la section précédente une procédure obtenue par matrix lifting / relaxation SDP du problème d'origine (voir Algorithme 2). L'objectif de cette section est de prouver que cet procédure fournit un bon estimateur du vecteur d'appartenance $\bar{x}$.

On rappelle quelques notations : λ est défini dans (22),

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

où $J_n = (1)_{n \times n}$ et $\hat{x}$ est un plus grand vecteur propre de $\hat{Z}$. On veut montrer que $\hat{x}$ est proche de $\bar{x}$ le vecteur d'appartenance aux deux communautés V_1 et V_1^c : $\bar{x}_i = 1$ quand $i \in V_1$ et $\bar{x}_i = -1$ quand $i \notin V_1$.

Théorème 5.1. *Soit $\epsilon \in (0, 1)$ tel que $n \geq 10^4 \epsilon$. Soit A la matrice d'adjacence d'un graphe distribué selon le SBM de paramètre (n, p, q) où $p > q$. On suppose que $\max(p(1-p), q(1-q)) \geq 20/n$. On suppose que $p \geq a/n$, $q \geq b/n$ et $(a-b)^2 \geq 10^4 \epsilon^{-2}(a+b)$. Soit $\hat{Z}$ une solution de (33). Avec probabilité au moins $1 - e^3 5^{-n}$, on a*

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

Ensuite, on passe de l'estimation de la matrice $\bar{x}(\bar{x})^\top$ en norme 2 à l'estimation d'un plus grand vecteur propre de $\bar{x}(\bar{x})^\top$ de la manière suivante.

Corollaire 5.2. *Sous les hypothèse du Théorème 5.1. Si $\hat{x}$ est un plus grand vecteur propre de $\hat{Z}$ tel que $\|\hat{x}\|_2 = \sqrt{n}$ alors*

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

La preuve du Corollaire 5.2 s'appuie sur le théorème de Davis-Kahane aussi connu sous le nom de "$\sin\theta$-theorem" qu'on rappelle maintenant.

Théorème 5.3 (Davis-Kahan)**.** *Soit A et B deux matrices symétriques de $\mathbb{R}^{n\times n}$. Soit $i \in \{1,\dots,n\}$. On suppose que la i-ième plus grande valeur propre $\lambda_i(A)$ de A est bien séparée du reste du spectre de A :*

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

Si $u_i(A)$ est un vecteur propre unitaire de A associé à la valeur propre $\lambda_i(A)$ et que $u_i(B)$ est un vecteur propre unitaire associé à la $i-$ième plus grande valeur propre de B, on a

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

Démonstration du Corollaire 5.2. On applique Davis-Kahan à $A = \bar{x}(\bar{x})^\top$ et $B = \hat{Z}$. Comme A est de rang 1 et que sa plus grande valeur propre est $\lambda_1(A) = n$, le spectral gap de A est $\delta = n > 0$. On a alors d'après le Théorème de Davis-Kahan que

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

où $u_1(A) = \bar{x}/\sqrt{n}$ et $u_1(B) = \hat{x}/\sqrt{n}$. Autrement dit,

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

où on a utilisé le Théorème 5.1 dans la dernière inégalité. ∎

5.1 Schéma de la preuve du Théorème 5.1

On note

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

l'ensemble de contrainte de la procédure SDP (33). On veut montrer que

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

est proche de $\bar{Z} = \bar{x}(\bar{x})^\top$.

On va procéder en 3 étapes :

1) on montre que $\bar{Z} \in \mathrm{argmax}_{Z\in\mathcal{M}_{opt}}\langle \mathbb{E}(A-\lambda J_n), Z\rangle$ en modifiant très légèrement la preuve de la Proposition 4.5 vu que $\mathbb{E}\lambda J_n = [(p+q)/2]J_n$ est presque égale à $\alpha u_1 \otimes u_1 = [(p+q)/2 + (1-p)/n]J_n$.

2) On montre que $(A - \lambda J_n)$ est proche de $\mathbb{E}(A-\lambda J_n)$ uniformément sur $\mathcal{M}_{opt}$, càd avec probabilité $1 - e^3 5^{-n}$,

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

où c_0 est une constante absolue.

3) On montre une inégalité de courbure de la fonction objectif à l'optimum : pour tout $Z \in \mathcal{M}_{opt}$,

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

Les deux points importants sont **2)** et **3)**. Tous l'aspect probabiliste du problème se trouve dans le point **2)**. On commence par détailler ce point-là.

5.2 Aspect probabiliste de la preuve

On montre dans cette section le point **2)** du schéma de la preuve du Théorème 5.1. Cet argument s'appuie sur l'inégalité de Grothendieck qu'on rappelle maintenant d'abord sous sa forme générale puis sous sa forme matricielle.

Théorème 5.4 (Inégalité de Grothendieck)**.** *Soit $B = (b_{ij})_{i,j}$ une matrice de $\mathbb{R}^{p\times q}$. On suppose que*

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

pour tout $s_i, t_j \in \{-1, 1\}$ alors, pour tout espace de Hilbert H et tous vecteurs $u_i, v_j \in H$ tels que $\|u_i\|_2 \leq 1, \|v_j\|_2 \leq 1$ on a

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

où $K_G \leq 1.783$.

Théorème 5.4 est la forme la plus connue de l'inégalité de Grothendieck. Il en existe de multiple formulations et des généralisation, comme celle du théorème de Nesterov très utilisé pour trouver des solutions approchantes à des problèmes combinatoires grâce à des relaxation convexes menant à des SDP. La forme de l'inégalité de Grothendieck que nous allons utiliser est une reformulation sous forme matricielle du Théorème 5.4 dans le cas carré $p = q = n$ et pour $H = \mathbb{R}^n$. On note

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

où B_2^n est la boule unité Euclidienne de $\mathbb{R}^n$. Pour tout $B \in \mathbb{R}^{n\times n}$, on a

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

On a clairement, $\mathcal{M}_1 \subset \mathcal{M}_G$ alors

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

L'inégalité de Grothendieck montre que l'inégalité inverse est aussi vraie à constante près.

Corollaire 5.5. *Il existe une constante absolue $K_G \leq 1.783$ telle que pour toute matrice $B \in \mathbb{R}^{n\times n}$, on a*

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

On a donc d'après l'inégalité de Grothendieck que pour tout $B \in \mathbb{R}^{n\times n}$,

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

On peux écrire le sup sur $\mathcal{M}_1$ comme une norme d'opérateur car pour tout $B \in \mathbb{R}^{n\times n}$

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

Pour le problème qu'on cherche à résoudre – contrôler le supremum d'un processus empirique indexé par $\mathcal{M}_{opt}$ – on ne regarde que les matrices symétriques positive de $\mathcal{M}_G$ et on peut démontrer le résultat suivant.

Proposition 5.6. *On a $\mathcal{M}_{opt} \subset \mathcal{M}_G$.*

Démonstration. Si $Z \in \mathcal{M}_{opt}$, comme $Z \succeq 0$, il existe $X \in \mathbb{R}^{n\times n}$ tel que $Z = XX^\top$. Par ailleurs, $\mathrm{diag}(Z) \preceq I_n$. Or $\mathrm{diag}(Z) = \mathrm{diag}(\|X_{1\bullet}\|_2^2, \cdots, \|X_{n\bullet}\|_2^2)$ donc $\|X_{i\bullet}\|_2^2 \leq 1$ pour tout $i = 1, \ldots, n$. Alors $Z \in \mathcal{M}_G$. ∎

On déduit de l'inégalité de Grothendieck et de la Proposition 5.6 que

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

Il reste donc à majorer $\|A - \mathbb{E}A\|_{\infty\to 1}$ et $|\lambda - \mathbb{E}\lambda|$ avec grande probabilité. Pour cela, on va utiliser l'inégalité de concentration de Bernstein suivie d'une "union bound". On rappelle ici l'inégalité de Bernstein.

Théorème 5.7. *Soit $Z_1, \ldots, Z_m$ des variables aléatoires indépendantes centrées telles que $|Z_i| \leq b$ pour tout $i = 1, \ldots, m$ presque sûrement. On note $\sigma^2 = (1/m)\sum_{i=1}^m \mathbb{E}Z_i^2$. Pour tout $t > 0$,*

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

L'inégalité de Bernstein et l'union bound donnent le résultat suivant.

Proposition 5.8. *On pose $\bar{p} = [2/(n(n-1))]\sum_{i<j}\mathrm{var}(A_{ij})$. Si $\bar{p} > 9/n$ alors avec probabilité au moins $1 - e^3 5^{-n}$, $\|A - \mathbb{E}A\|_{\infty\to 1} \leq 6n(n-1)\sqrt{\bar{p}/n}$.*

Démonstration. On a

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

Soit $s, t \in \{-1, 1\}^n$. Comme $A_{ii} = 1$ et $A_{ij} = A_{ji}$, on a

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

∎

6 Étude statistique d'une relaxations SDP : preuve de la reconstruction exacte par construction d'un certificat dual

Dans cette section, on propose de montrer qu'il est possible de reconstruire exactement le vecteur d'appartenance par relaxation SDP dans le cas de deux communautés de même taille dans le stochastic block model et sous des conditions de 'puissance' du signal (càd pour p et q suffisamment grands). Pour démontrer un tel résultat on va utiliser la technique de **construction d'un certificat dual**. Avant de mettre en oeuvre cette technique, on rappelle un résultat sur les problème d'optimisation SDP.

Dualité faible et certification duale pour les problèmes SDP. On considère un problème de type SDP : soit $C \in \mathbb{R}^{n\times n}$, $A_1 \ldots, A_m \in \mathbb{R}^{n\times n}$ et $b \in \mathbb{R}^m$. On note $\mathcal{A} : X \in \mathbb{R}^{n\times n} \to (\langle A_i, X\rangle)_{i=1}^m$ et $\mathcal{A}^\top : \lambda \in \mathbb{R}^m \to \sum_i \lambda_i A_i$ son opérateur adjoint. Un problème SDP est de la forme

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

où on rappelle que $X \succeq 0$ signifie que X est positive. L'ensemble des matrices symétriques positives est un cône convexe fermé non vide. On retrouve donc un problème de type conic LP. La fonction de Lagrange associée est

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

qui a pour fonction duale

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

Le problème dual associé est alors

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

La dualité faible dit que pour tout $\lambda \in \mathbb{R}^m$ tel que $C - \mathcal{A}^\top(\lambda) \succeq 0$ et pour tout $X \succeq 0$ tel que $\mathcal{A}(X) = b$, on a $\langle C, X\rangle \geq \langle \lambda, b\rangle$. Pour la construction d'un certificat dual, on a seulement besoin de la dualité faible : c'est en effet cette inégalité qui permet de dire qu'un λ^* faisable pour le problème dual et tel que $\langle C, X^*\rangle = \langle \lambda^*, b\rangle$ certifie X^* comme étant solution du problème primal càd ici de (37) (à condition qu'on ait bien pris X^* comme faisable pour le primal). Cependant, un certificat dual ne peut exister que s'il y a dualité forte ; en effet, si $\langle C, X^*\rangle = \langle \lambda^*, b\rangle$ pour λ^* faisable pour le dual et X^* faisable pour le primal alors il y a dualité forte. Il est donc bon de savoir en amont (même si ce n'est pas nécessaire de le démontrer) s'il y a dualité forte ou pas. On rappelle alors un résultat de dualité forte pour les problèmes SDP. Dans le théorème suivant, on identifie donc une situation où on peut avoir égalité $\langle C, X^*\rangle = \langle \lambda^*, b\rangle$ pour X^* solution primale et λ^* solution duale.

Théorème 6.1. *On suppose qu'il existe $X_0 \succ 0$ tel que $\mathcal{A}(X_0) = b$ et un $\lambda_0 \in \mathbb{R}^m$ tel que $C - \mathcal{A}^\top(\lambda_0) \succ 0$. Alors le problème SDP primal (37) admet une solution, le problème SDP dual (38) admet une solution, il y a dualité forte et pour toute solution X^* du primal et toute solution du dual λ^*, on a $\langle C, X^*\rangle = \langle b, \lambda^*\rangle$.*

Ce type de théorème est classique en optimisation convexe. L'hypothèse de 'strict feasibility' est parfois appelée condition de Slater. Elle est en général 'facile' à vérifier et peut donc nous assurer ou pas de l'existence d'un certificat dual avant de se lancer dans sa construction. Cependant, comme indiqué plus haut, le résultat dont on a besoin pour la construction d'un certificat dual et en fait plus une réciproque au Théorème 6.1 et ne nécessite en fait que la dualité faible pour être démontré (la preuve du Théorème 6.1 est plus compliquée).

Proposition 6.2. *Si X^* est faisable pour le primal (37) et λ^* est faisable pour le dual (38) et sont tels que $\langle C, X^*\rangle = \langle b, \lambda^*\rangle$ alors X^* est solution du primal et λ^* est solution du duale.*

Preuve. Par dualité faible, on a pour tout X faisable pour le primal et tout λ faisable pour le dual que $\langle X, C\rangle \geq \langle \lambda, b\rangle$. En particulier, pour tout X faisable pour le primal $\langle X, C\rangle \geq \langle \lambda^*, b\rangle = \langle X^*, C\rangle$ et donc X^*, comme X^* est faisable pour le primal, X^* est solution du primal. De même pour tout λ faisable pour le dual, on a $\langle b, \lambda\rangle \leq \langle C, X^*\rangle = \langle b, \lambda^*\rangle$ et donc λ^* étant faisable pour le dual est solution du dual. ∎

C'est la Proposition 6.2 qui permet de développer la méthode du **certificat dual** : en effet, étant donné un point X^* faisable pour le primal, si on peut trouver un λ^* faisable pour le problème dual (càd tel que $C - \mathcal{A}^\top(\lambda^*) \succeq 0$) tel que $\langle C, X^* \rangle = \langle b, \lambda^* \rangle$ alors forcément X^* est solution du primal. Cette idée est très pratique lorsqu'on pense qu'un élément X^* faisable pour le problème primal (càd tel que $X^* \succeq 0$ et $\mathcal{A}(X^*) = b$) est probablement solution du primal et qu'on souhaite le démontrer.

On explique maintenant comment on va employer la Proposition 6.2 dans notre cadre. L'idée est qu'on a un problème initial (P_0) qu'on ne sait pas résoudre en général (par exemple, il est NP-hard) – c'est par exemple un problème comme (26). On 'convexifie' ce problème par relaxation convexe. Ceci nous donne un autre problème, noté (P), comme (28). Parfois cette relaxation donne un problème de type SDP. Dans ce cas, on peut considérer le problème dual (D) associé à ce SDP. Le but est de montrer que l'ensemble des solutions du problème initial (P_0) est aussi l'ensemble des solutions du problème relaxé (P). On pourra ainsi résoudre (P_0) grâce à (P). Pour ce faire, on peut prendre une solution du problème initial, généralement c'est un unique vecteur x_0 (comme le 'membership community vector' f^{V_1} qu'on souhaite reconstruire en détection de communautés) et montrer que $X_0 = x_0 x_0^\top$ est l'unique solution du problème relaxé (P). Pour cela, on cherche à construire un élément λ_0 faisable pour le dual et tel que $\langle C, X_0 \rangle = \langle b, \lambda_0 \rangle$. Ainsi, on saura que X_0 est solution de (P) : on dit dans ce cas que λ_0 a certifié X_0 comme étant solution de (P). C'est le principe du **certificat dual**. Toute la difficulté technique est dans la construction de λ_0. Cependant, on fait en général de la 'retro engineering' en partant du fait que si λ_0 est un certificat dual alors il doit à la fois être faisable pour le dual et aussi vérifier $\langle C, X_0 \rangle = \langle b, \lambda_0 \rangle$. On peut aussi construire λ_0 en cherchant à résoudre le problème dual vu qu'on est sûr que si λ_0 est faisable pour le dual et que $\langle C, X_0 \rangle = \langle b, \lambda_0 \rangle$ où X_0 est faisable pour le primal alors nécessairement λ_0 est solution du dual. C'est cette technique de construction d'un certificat dual que nous allons mettre en œuvre pour montrer la reconstruction exacte par SDP pour le problème de détection de communautés.

Relaxation SDP. La relaxation SDP proposée en (28) n'est pas la seule qu'on puisse faire. Il en existe beaucoup d'autres. Dans cette section, on en introduit une autre qui nous permettra de reconstruire exactement les communautés.

Le problème de détection de deux communautés s'écrit, initialement, comme un problème de recherche d'une solution au problème discret

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

(voir (6)) où $f^{V_1} = (f_i^{V_1})_{i \in V} \in \mathbb{R}^{|V|}$ est défini pour tout $i \in V$ par

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

On a motivé l'étude de ce problème en introduisant la fonction de modularité *RatioCut* en Section 3.2 qu'on souhaite minimiser. On peut aussi motiver ce problème comme un problème de recherche de maximum de vraisemblance comme introduit en Section 4.1 dans le cas où les deux communautés qu'on cherche à détecter sont de même taille. Néanmoins, dans notre cadre stochastique la matrice L qu'on observe n'est pas la 'vraie' matrice Laplacienne du 'vrai' graphe sous-jacent : on a seulement observée une version bruitée (issue du SBM) du 'vrai' graphe dont le 'vrai' Laplacien est $\mathbb{E}L$. Le problème qu'on souhaiterait idéalement résoudre est en fait

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

où le Laplacien est le Laplacien 'théorique' $\mathbb{E}L$ et non l'observé. Bien sûr, vu qu'on ne connait par $\mathbb{E}L$, le problème (40) est lui aussi purement théorique. On peut néanmoins montrer que le vecteur de communauté qu'on cherche à retrouver càd f^{V_1} et $f^{V_1^c}$ sont les deux seules solution du problème (40). Ceci donne une autre justification pour considérer le problème (39) qu'on voit alors comme la version empirique de (40) puisqu'on a simplement remplacer la quantité théorique $\mathbb{E}L$ par sa quantité empirique L.

Proposition 6.3. *Si $L = D - A$ est Laplacien d'un graphe tiré selon le SBM de paramètre (n, p, q) où $0 < q \leq p < 1$ et selon la partition $V = V_1 \sqcup V_1^c$ alors f^{V_1} et $f^{V_1^c}$ sont les deux seules solutions du problème (40).*

Preuve. On note $i \sim j$ quand i et j sont dans la même communauté (càd $i, j \in V_1$ ou $i, j \in V_1^c$) et $i \not\sim j$ quand i et j ne sont dans la même communauté (càd $[i \in V_1$ et $j \in V_1^c]$ ou $[j \in V_1$ et $i \in V_1^c]$). Soit $V_0 \subset V$. On note $i \sim_{V_0} j$ quand $i, j \in V_0$ ou $i, j \in {V_0}^c$ et $i \not\sim_{V_0} j$ sinon. On a

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

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

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

Car $p \geq q$. On a égalité ci-dessus si et seulement si le terme

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

est nul. Celà a lieu si et seulement si les ensembles $\{(i,j) : i \sim j\}$ et $\{(i,j) : i \not\sim_{V_0} j\}$ sont disjoints càd quand $V_0 = V_1$ ou $V_0 = V_1^c$. Par ailleurs, on a aussi

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

Donc f^{V_1} et $f^{V_1^c}$ sont bien les deux seules solutions au problème (40). ∎

La Proposition 6.3 nous dit que le problème d'optimisation a pour unique solution le vecteur d'appartenance (au signe près) f^{V_1} qui est bien le vecteur qu'on souhaite trouver quand on veut résoudre le problème de détection de communauté dans un SBM à deux classes.

Le problème (6.3) n'est cependant pas à notre disposition vu que $\mathbb{E}L$ nous est inconnu : on n'observe que L. Le problème (39) est donc celui que nous devons considérons ; on y a seulement remplacé la quantité inconnue $\mathbb{E}L$ par celle observée L. On passe à la résolution effective de (39) qui va s'appuyer sur une relaxation convexe.

La difficulté pratique avec (39) est qu'il est NP-hard en général ; on peut donc le considérer d'un point de vue théorique mais pas pratique. On va alors faire une relaxation convexe de ce problème pour obtenir un problème d'optimisation qui peut être résolu (au moins approximativement) de manière numérique. Il y a donc deux difficultés dans le problème (39) qu'on considère maintenant :

— on n'a pas accès au 'vrai' graphe mais seulement à une version 'bruitée' (certains liens inter-communautés apparaissent) et 'masquée' (certains liens intra-communautés ne sont pas donnés) : on peut parler d'information partielle quand on observe L plutôt que $\mathbb{E}L$,
— le problème qu'on souhaite résoudre est NP-hard en général : c'est une difficulté d'ordre computationnelle.

On va néanmoins montré qu'une relaxation convexe de (39) permet de résoudre ces deux problème : on va en effet montrer qu'avec grande probabilité f^{V_1} et $f^{V_1^c}$ seront les deux seules et uniques solutions d'un problème SDP construit à partir de L, le Laplacien observé (et non $\mathbb{E}L$). La relaxation convexe que nous allons mettre en œuvre ici pour résoudre ce problème est basée sur les techniques de matrix lifting et relaxation SDP (voir Section 7 pour d'autres exemples) et l'étude statistique que nous allons mener pour montrer les propriétés de reconstruction exacte de cette procédure se base sur la construction d'un certificat dual.

On effectue ce travail dans le cas de détection de deux communautés de même taille. Dans ce cas le problème (39) s'écrit de manière équivalente par

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

(voir la Remarque 4.1) où A est la matrice d'adjacence observée. On note au passage que (41) est aussi l'estimateur du maximum de vraisemblance dans la SBM obtenu dans la Section 4.1 dans le cas de détection de deux communautés de même taille. C'est le problème (41) que nous allons relaxer maintenant.

L'approche de matrix lifting comme présentée dans les deux premiers exemples de la Section 7 consiste à récrire les fonctions quadratiques en x comme des fonctions linéaires de $xx^\top$. Dans ce schéma les contraintes linéaires telles que '$\langle x, 1\rangle = 0$' ne sont pas gérées. On peut cependant aussi gérer les termes linéaires (présents soit dans la contrainte soit dans la fonction objectif) mais au lieu de 'lifter' seulement x en $xx^\top$, il faut 'lifter' $\binom{1}{x}$ en $\binom{1}{x}\begin{pmatrix}1 & x^\top\end{pmatrix}$. On renvoie au troisième paragraphe de la Section 7 où cette approche gérant les termes linéaires est proposée. Dans cette section, on va simplement enlever cette contrainte linéaire mais au préalable on va l'inclure comme terme de pénalisation dans la fonction objectif.

Si on enlève la contrainte '$\langle x, 1\rangle = 0$' dans (41), on voit que $\mathbf{1} = (\mathbf{1})_{\mathbf{i=1}}^{\mathbf{n}}$ (et son opposé) est solution du problème. C'est une solution qu'on souhaite éviter car elle ne propose aucune structure non triviale de communautés mais juste celle faite de tout le graphe. On va alors introduire un terme de pénalisation dans la fonction objective pour éviter cette solution, on considère alors une fonction objective de la forme $x^\top Ax - \lambda\langle x, \mathbf{1}\rangle^2$ où $\lambda \geq 0$ est le paramètre de régularisation. On va aussi ajouter un terme en $\lambda\|x\|_2^2$ qui ne change rien vu qu'il est constant sur les vecteurs de $\{-1,1\}^n$ et qu'il va nous permettre d'écrire facilement la fonction objectif (ce terme n'a pas d'importance et dépend en fait uniquement de la convention qu'on choisit sur la matrice d'adjacence pour les termes diagonaux, ici on a choisit de poser $A_{ii} = 1$ mais on aurait pu faire un autre choix comme mettre 0). Comme $\langle x, \mathbf{1}\rangle^2 = x^\top \mathbf{1}\mathbf{1}^\top x$ et $\|x\|_2^2 = x^\top x$, on a pour $\lambda = 1/2$ (et quitte à multiplier la fonction objectif par 2) que $2x^\top Ax - \langle x, \mathbf{1}\rangle^2 - \|x\|_2^2 = x^\top(2A - \mathbf{1}\mathbf{1}^\top - I_n)x$. On introduit alors la matrice

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

On considère alors le problème régularisé suivant :

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

En particulier, $\mathbf{1}$ n'est plus solution de ce problème grâce à la pénalisation $x \to -\langle x, \mathbf{1}\rangle^2$ ajoutée à la fonction objectif qui pénalise le plus $\mathbf{1}$ car elle vaut $-n^2$ en ce vecteur. La régularisation va donc disqualifier les partitions proposant des communautés de taille déséquilibrées.

On applique la méthode de matrix lifting au problème (43). On réécrit la fonction objectif qui est quadratique en x en une fonction linéaire de $Z = xx^\top$: $x^\top Bx = \langle B, xx^\top\rangle = \langle B, Z\rangle$ et on réécrit les contraintes qui sont aussi quadratique en x, comme contraintes linéaires sur Z : $x_i^2 = 1$ est équivalent à $Z_{ii} = 1$. On considère alors le problème 'lifté' de (43) suivi d'une relaxation SDP (on enléve la contrainte de rang sur Z) pour obtenir

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

On trouve bien un problème de type SDP. Pour faire le lien entre une solution de (44) et le problème d'origine (43), on peut soit prendre un plus grand vecteur propre d'une solution de (44) et clusteriser en deux cluster les lignes de ce vecteur. Soit on prend $G \sim \mathcal{N}(0, \hat{Z})$ où $\hat{Z}$ est une solution de (44) et on prend le signe des coordonnées de G – cette dernière méthode s'appelle le matrix rounding et est utilisé pour démontrer l'inégalité de Grothendieck par exemple. Notre objectif dans la suite est de montrer que $f^{V_1}(f^{V_1})^\top$ est l'unique solution de (44) où f^{V_1} est bien le vecteur d'appartenance des communautés sous-jacentes au SBM. Pour celà on va construire un certificat dual (stricte) certifiant que $f^{V_1}(f^{V_1})^\top$ est bien l'unique solution de (44).

Construction d'un certificat dual. On souhaiter prouver que $gg^\top$ où $g := f^{V_1}$ est l'unique solution de (44). Pour cela, on va construire un certificat dual stricte. On doit alors d'abord identifier le problème dual de (44). On peut utiliser le problème introduit dans (38) ou le retrouver directement.

La fonction de Lagrange associée à (44) est

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

où $\mathcal{S}_n^+$ est le cône des matrices $n\times n$ symétriques positives et $\mathrm{diag}(Z) = (Z_{ii})_{i=1}^n$. On peut s'assurer qu'on a défini la bonne fonction de Lagrange pour le problème (44) en vérifiant l'égalité entre le problème d'origine et le problème primal en Z associé à cette fonction de Lagrange :

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

La fonction duale est

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

où $\mathrm{Diag}(\lambda)$ est la matrice $n \times n$ diagonale dont les éléments diagonaux sont les coordonnées de λ. Pour écrire $\mathcal{L}$ et ψ on a utilisé que le cône dual de $\mathcal{S}_n^+$ (où appartient la variable primale Z) est $-\mathcal{S}_n^+$ (où appartient la variable duale $-X$ associée à la contrainte '$X \succeq 0$'). Le problème dual est donc $\max(\psi(\lambda, X) : (\lambda, X) \in \mathbb{R}^n \times \mathcal{S}_n^+)$ càd

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

qui est équivalent à

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

qui est aussi équivalent à

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

On a bien équivalence entre (45) et (46) car si (λ, X) est solution de (45) alors $-\mathrm{Diag}(\lambda)$ est solution de (46) et si Y est solution de (46) alors $(-\mathrm{diag}(Y), -B+Y)$ est solution de (45).

La dualité faible est ici donnée par l'inégalité suivante : pour tout $Z \succeq 0$ tel que $Z_{ii} = 1$ et tout $\lambda \in \mathbb{R}^n$ tel que $-B - \mathrm{Diag}(\lambda) \succeq 0$, on a $-\langle B, Z\rangle \geq \langle \lambda, \mathbf{1}\rangle$. Qui peut se récrire en les termes du problème (46) par : pour tout $Z \succeq 0$ tel que $Z_{ii} = 1$ et toute matrice diagonale Y telle que $Y - B \succeq 0$, on a $-\langle B, Z\rangle \geq -\mathrm{Tr}(Y)$. On peut retrouver cette inégalité directement en observant que

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

car $Y \succeq B$ et donc comme $Z \succeq 0$, on a $\langle B, Z\rangle \leq \langle Y, Z\rangle$ et aussi $Z - \mathrm{Diag}(\mathbf{1})$ a une diagonale nulle et Y est diagonale donc $\langle Y, Z - \mathrm{Diag}(\mathbf{1})\rangle = 0$.

On en déduit que pour certifier qu'un Z, faisable pour le primal (càd tel que $Z \succeq 0$ et $Z_{ii} = 1$), est bien solution du problème (44), il suffit de construire une matrice $Y \in \mathbb{R}^{n\times n}$ telle que :

a) Y est diagonale et $Y \succeq B$ (càd Y est faisable pour le dual)
b) $\langle B, Z\rangle = \mathrm{Tr}(Y)$ (càd les fonctions objectives primales et duales coïncident en Z et Y respectivement) qui est équivalent à $\langle B - Y, Z\rangle = 0$ car Y est diagonale et Z a une diagonale de 1.

Dans ce cas Y est appelé un certificat dual de Z pour le problème (44).

On souhaite montrer que $gg^\top$ est l'unique solution de (44). Il suffit alors de construire une matrice diagonale $Y \in \mathbb{R}^{n\times n}$ telle que $Y \succeq B$ et $\langle B - Y, gg^\top\rangle = 0$. Ceci prouvera que $gg^\top$ est bien solution de (44). Pour l'unicité ensuite, il suffit de prouver que le noyau de $B - Y$ (où Y est le certificat dual de $gg^\top$ qu'on va construire dans la suite) est de dimension 1. Ceci prouvera bien que $gg^\top$ est l'unique solution d'après le résultat suivant.

Lemme 6.4. *Soit Y un certificat dual de $gg^\top$ pour le problème (44). Si le noyau de $B - Y$ est de dimension 1 alors $gg^\top$ est l'unique solution de (44).*

Preuve. Soit Z une solution de (44). On écrit la décomposition en valeurs propres de Z : on a $Z = \sum_i \lambda_i u_i u_i^\top$ où $0 \leq \lambda_1 \leq \ldots \leq \lambda_n$ et $(u_1, \ldots, u_n)$ est une base orthonormale de $\mathbb{R}^n$. D'après la Proposition 6.9, on sait que Y est aussi un certificat pour Z : $\langle B - Y, Z\rangle = 0$. On a donc $\sum_i u_i^\top (B - Y) u_i = 0$ et comme $Y \succeq B$, on a $u_i^\top (Y - B) u_i \geq 0$ pour tout i. On en déduit donc que $[u_i^\top (Y - B) u_i = 0$ ou $\lambda_i = 0]$ et donc $[(Y - B)u_i = 0$ ou $\lambda_i = 0]$ càd u_i est dans le noyau de $B - Y$ ou $\lambda_i = 0$. Mais $(B - Y)g = 0$ (car $\langle B - Y, gg^\top\rangle = 0$) et le noyau de $B - Y$ est de dimension 1 donc $\mathrm{Ker}(B - L) = \mathrm{vect}(g)$. On en déduit alors que u_i est dans $\mathrm{vect}(g)$ ou $\lambda_i = 0$. Comme $(u_1, \ldots, u_n)$ est une base orthonormale et que le noyau est de dimension 1, on en déduit que seul un des u_i peut être dans le noyau et alors il est proportionnel à g et pour tous les autres on a $\lambda_j = 0$. Donc Z est de la forme $Z = \lambda gg^\top$. Par ailleurs, $Z_{ii} = 1 = g_i g_i$ donc $\lambda = 1$, càd $Z = gg^\top$. Donc $gg^\top$ est bien l'unique solution de (44). ∎

Un certificat dual permettant de certifier l'unicité d'une solution est appelé un **certificat dual stricte**. C'est ce que nous voulons construire ici pour $gg^\top$. Si on résume on souhaite construire une matrice $Y \in \mathbb{R}^{n\times n}$ telle que

a) Y est diagonale et $Y \succeq B$
b) $(B - Y)g = 0$
c) $\lambda_2(Y - B) > 0$ càd la deuxième plus petite valeur propre de $Y - B$ est strictement positive (et donc son noyau est de dimension 1).

Les conditions *a)* et *b)* ne nous laisse pas beaucoup de choix : on a $Bg = Yg$ et comme Y est diagonale on doit forcément prendre $Y_{ii} = (Bg)_i / g_i$ pour tout $i \in \{1, \ldots, n\}$. Par ailleurs, on a

$Bg = (2A - \mathbf{1}\mathbf{1}^\top - I_n)g = 2Ag - g$ car $\langle \mathbf{1}, g \rangle = 0$ et donc pour tout $i, j \in \{1, \dots, n\}$,

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

Il reste à montrer que $Y \succeq B$. On sait que $(Y - B)g = 0$ par construction de g donc si on prouve que la deuxième plus petite valeur propre de $Y - B$ est strictement positive, on aura fini car ceci prouvera à la fois que $Y \succeq B$ et le point *c)* (par construction Y est diagonale et $Bg = Yb$). On peut donc établir le résultat suivant.

Proposition 6.5. *Si la deuxième plus petite valeur propre de $Y - B$ est strictement positive alors $gg^\top$ est l'unique solution du problème SDP (44).*

On note $\lambda_2(Y - B)$ la deuxième plus petite valeur propre de $Y - B$. D'après la Proposition 6.5, si on a $\lambda_2(Y - B)$ alors $gg^\top$ est l'unique solution de (44) et donc en cherchant le plus grand vecteur propre de $gg^\top$, on obtiendra $\pm g/\|g\|_2$ et donc on pourra en déduire les communautés $V_1 = \{i : gi > 0\}$ et $V_i^c = \{g_i < 0\}$. Il reste alors à démontrer que $\lambda_2(Y - B) > 0$.

Aspect probabiliste : minoration de $\lambda_2(Y - B)$ dans le SBM. Pour montrer que $\lambda_2(Y - B) > 0$, c'est à ce stade qu'on va utiliser l'aspect aléatoire du problème. On va donc montrer que si A est la matrice d'adjacence d'un SBM de paramètre (n, p, q) où $p = \alpha(\log n)/n$ et $q = \beta(\log n)/n$ où $\alpha > \beta$ on aura $\lambda_2(Y - B) > 0$ sous certaine condition sur α et β qu'on espère pas trop éloignée de la transition de phase optimale donnée dans le Théorème 4.3, càd $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$.

La stratégie que nous suivons pour prouver que $\lambda_2(Y - B) > 0$ est basé sur un argument de concentration de matrice. On va prouver les deux résultats suivants :

1) $\lambda_2(\mathbb{E}(Y - B)) := \square$
2) avec grande probabilité $\|(Y - B) - \mathbb{E}(Y - B)\|_{op} < \square$

où $\square = (\alpha - \beta)\log n$ et $\|\cdot\|_{op}$ est la norme d'opérateur. Ces deux résultats seront suffisant pour démontrer que $\lambda_2(Y - B) > 0$ puisque par la formule de Courant-Fisher (en particulier, son corollaire sur les propriétés d'entrelacement des valeurs singulières), on a pour toute matrice $n \times n$ symétrique positive A et B que

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

pour $0 \leq \lambda_1(A) \leq \cdots \leq \lambda_n(A)$ les valeurs propres de A en ordre croissant (et de même pour B). On voit donc que si *1)* et *2)* sont vérifiées alors, avec grande probabilité,

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

Il reste alors à démontrer les points *1)* et *2)*. On commence par le premier point.

Pour le calcul de la deuxième plus petite valeur propre de $\mathbb{E}(Y - B)$, on utilise les définitions de B et Y (voir (42) et (47)) pour obtenir pour tout $i \in V$

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

et

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

On a alors

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

On retrouve bien que $g^\top \mathbb{E}(Y-B)g = 0$ (on a construit Y tel que $Yg = Bg$) alors g est un vecteur propre associé à la valeur propre 0. Les autres vecteurs propres sont à chercher dans $\text{vect}(g)^\top$: soit $v \in \mathbb{R}^n$ tel que $\|v\|_2 = 1$ et $\langle v, g\rangle = 0$, on a

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

Vu qu'on se place dans le régime où $p = \alpha\log(n)/n$ et $q = \beta\log(n)/n$, on peut supposer que $1 \geq p+q$. Ainsi $v^\top \mathbb{E}(Y-B)v \geq n(p-q)$ et cette borne est atteinte par un v orthogonal à $\mathbf{1}$. On a donc que

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

(on voit que g est vecteur propre associé à la plus petite valeur propre 0, un v tel que $\langle v, g\rangle = \langle v, \mathbf{1}\rangle = 0$ est un vecteur de Fiedler de $\mathbb{E}(Y-B)$ associé à la valeur propre $n(p-q)$, il y en a $n-2$ et $\mathbf{1}/\sqrt{n}$ est vecteur propre associé à la valeur propre $n(1-2q)$).

On passe au point *2)*. On souhaite montrer que $\|(Y-B) - \mathbb{E}(Y-B)\|_{op} < (\alpha-\beta)\log n$ avec grande probabilité quand $p = \alpha\log(n)/n$ et $q = \beta\log(n)/n$ (et $p+q<1$). C'est un résultat de concentration de la matrice $Y-B$ autour de sa moyenne par rapport à la norme d'opérateur qu'on souhaite ici démontrer. Pour prouver un tel résultat, on va utiliser un résultat de concentration sur la plus grande valeur propre d'une moyenne de matrices aléatoires symétriques et indépendantes qu'on appelle **l'inégalité de Bernstein matricielle** et qu'on énonce maintenant.

Théorème 6.6. *Soit $(X_k)_{k=1}^n$ des matrices aléatoires symétriques indépendantes de taille $d \times d$. On suppose que pour tout $k = 1, \dots, n$, on a*

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

On a alors pour tout $t > 0$, avec probabilité au moins $1 - d\exp(-t)$

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

où

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

et $\|\cdot\|_{op}$ est la norme d'opérateur et $\lambda_{max}(\cdot)$ est la plus grande valeur propre d'une matrice symétrique. On obtient une borne sur la norme d'opérateur de $(1/n)\sum_{k=1}^n X_k$ en appliquant ce résultat aux $-X_k$ et en remarquant que $\|M\|_{op} = \max(\lambda_{max}(M), \lambda_{max}(-M)$ pour toute matrice symétrique M.

L'inégalité de Bernstein matricielle ressemble beaucoup à l'inégalité de Bernstein 'classique', càd celle portant sur la concentration d'une somme de variables aléatoires réelles indépendantes. Il faut cependant bien faire attention à la présence de la dimension d dans la déviation en '$1 - d\exp(-t)$' dans le Théorème 6.6. La présence de cette dimension est parfois responsable d'une perte logarithmique dans les vitesses de convergence qu'on obtient en utilisant cette inégalité. Ce ne sera pas notre cas ici car on va travailler à un niveau de déviation $t = (\alpha-\beta)\log n$ qui est de l'ordre du log de la dimension et donc le terme dimensionnel en d dans le Théorème 6.6 sera au pire responsable d'une perte au niveau des constantes.

On va donc utiliser l'inégalité de Bernstein matricielle pour démontrer que le point *2)*. La première chose à faire est de faire apparaître des sommes de matrices aléatoires indépendantes

dans $Y - B$. Pour cela, on va considérer une partition du graphe $G = (V, E)$ en les deux sous-graphes $G^{\sim} = (V, E^{\sim})$ et $G^{\not\sim} = (V, E^{\not\sim})$ où $E^{\sim}$ est l'ensemble de toutes les arêtes de E reliant deux points dans une même communauté et $E^{\not\sim}$ est l'ensemble de toutes les arêtes de E reliant deux points de deux communautés différentes. Ces deux sous-graphes ont une matrice d'adjacence notée respectivement $A^{\sim}$ et $A^{\not\sim}$ et des matrices (diagonales) de degrés notées respectivement $D^{\sim}$ et $D^{\not\sim}$. On a formellement pour tout $i, j \in V$

$$A^{\sim}_{ij} = \begin{cases} A_{ij} & \text{si } i \sim j \\ 0 & \text{sinon} \end{cases}, \quad D^{\sim} = \mathrm{Diag}(d^{\sim}_1, \dots, d^{\sim}_n) \text{ où } d^{\sim}_i = \sum_{j : i \sim j} A_{ij} = \sum_{j \in V} A^{\sim}_{ij}.$$

et

$$A^{\not\sim}_{ij} = \begin{cases} A_{ij} & \text{si } i \not\sim j \\ 0 & \text{sinon} \end{cases}, \quad D^{\not\sim} = \mathrm{Diag}(d^{\not\sim}_1, \dots, d^{\not\sim}_n) \text{ où } d^{\not\sim}_i = \sum_{j : i \not\sim j} A_{ij} = \sum_{j \in V} A^{\not\sim}_{ij}.$$

En particulier, on a $A = A^{\sim} + A^{\not\sim}$ comme représenté dans la Figure 4.

$$A = \begin{array}{|c|c|}\hline V_1 \times V_1 & V_1 \times V_1^c \\ \hline V_1^c \times V_1 & V_1^c \times V_1^c \\ \hline \end{array} = \begin{array}{|c|c|}\hline V_1 \times V_1 & 0 \\ \hline 0 & V_1^c \times V_1^c \\ \hline \end{array} + \begin{array}{|c|c|}\hline 0 & V_1 \times V_1^c \\ \hline V_1^c \times V_1 & 0 \\ \hline \end{array} = A^{\sim} + A^{\not\sim}$$

FIGURE 4 – Décomposition de la matrice d'adjacence A de G en la somme des deux matrices d'adjacence des sous-graphes $G^{\sim}$ et $G^{\not\sim}$.

On a aussi pour tout $i \in V$,

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

(on utilise ici que $g_i \in \{-1, 1\}$, donc $1/g_i = g_i$ et si $i \sim j$ alors $g_i g_j = 1$ sinon $g_i g_j = -1$). On a donc $Y = 2(D^{\sim} - D^{\not\sim}) - I_n$ et comme $A = A^{\sim} + A^{\not\sim}$ et $B = 2A - \mathbf{1}\mathbf{1}^{\top} - I_n$, on obtient que

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

On voit donc apparaître la différence des Laplacien de $G^{\sim}$ et $G^{\not\sim}$ (cette différence est parfois appelée le Laplacien du SBM, il faut néanmoins remarquer que ce n'est pas le Laplacien d'un graphe vu que cette différence n'est pas forcément une matrice positive). On écrit ensuite les deux Laplaciens $D^{\sim} - A^{\sim}$ et $D^{\not\sim} - A^{\not\sim}$ comme sommes de matrices aléatoires symétriques et indépendantes.

On introduit une 'base' de matrices symétriques dans laquelle on peut écrire $D^{\sim} - A^{\sim}$ et $D^{\not\sim} - A^{\not\sim}$ comme des sommes de matrices aléatoires symétriques indépendantes : pour tout $i, j \in V$,

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

où $E_{pq} \in \mathbb{R}^{n\times n}$ a pour entrées 0 partout sauf en (p,q) où elle vaut 1. On vérifie que

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

En effet, d'une manière générale si $L = D - A$ est le Laplacien d'un graphe non orienté non pondéré, on a $D - A = \sum_{i<j} A_{ij}\Delta_{ij}$. En effet, $(D-A)$ et $\sum_{i<j} A_{ij}\Delta_{ij}$ sont symétriques, il suffit alors de vérifier que les entrées au-dessus de la diagonale sont égales. Pour $i<j$, on a

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

vu que pour $i<j$ et $p<q$ on a $(\Delta_{pq})_{ij} = 1$ si $i=p$ et $p=j$ et 0 sinon. Pour tout $i \in V$, $(D-A)_{ii} = \sum_{j\in V} A_{ij} - A_{ii}$ et

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

On a alors

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

et donc

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

On remarque que sommer $A^{\sim}_{ij}\Delta_{ij}$ sur $i<j$ est équivalent à sommer sur $i<j$ tel que $i\sim j$ vu que $A^{\sim}_{ij} = 0$ quand $i \not\sim j$. L'intérêt de faire apparaître cette sommation est que les deux sommes dans (49) sont disjointes et comme $(A^{\sim}_{ij})_{i<j:i\sim j}$ et $(A^{\sim}_{ij})_{i<j:i\not\sim j}$ sont deux familles indépendantes de variables indépendantes, on a bien une somme de matrices aléatoires indépendante symétriques dans (49) et qu'on peut alors appliquer l'inégalité de Bernstein matricielle : pour tout $t>0$, avec probabilité au moins $1 - n\exp(-t)$,

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

où R est tel que pour tout $i<j$,

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

et

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

On a $\lambda_{max}(\Delta_{ij}) = 2$, on peut alors prendre $R = 2$. De plus $\Delta_{ij}^2 = 2\Delta_{ij}$ et donc

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

car $|V_1| = |V_1^c| = n/2$. On a donc

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

car $\langle \mathbf{1}, g \rangle = 0$ et donc si on complète $(\mathbf{1}/\sqrt{n}, q/\sqrt{n})$ en une base orthonormale $(u_1 := \mathbf{1}/\sqrt{n}, u_2 := q/\sqrt{n}, u_3, \ldots, u_n)$ de $\mathbb{R}^n$, on aura $I_n = \sum_i u_i u_i^\top$. Ainsi $[q(1-q) - p(1-p)] + [np(1-p) + nq(1-q)]$ est la plus grande valeur propre de multiplicité 1 de vecteur propre $g/\sqrt{n}$, $[np(1-p) + nq(1-q)]$ est valeur propre de multiplicité $(n-2)$ pour les vecteurs propres $u_3, \ldots, u_n$ et $[np(1-p) + nq(1-q)]I_n - [p(1-p) + q(1-q)]$ est la plus petite valeur propre de multiplicité 1 pour $\mathbf{1}/\sqrt{n}$ de la matrice apparaissant dans (50).

On obtient alors que pour tout $t > 0$, avec probabilité au moins $1 - 2n\exp(-t)$,

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

Par exemple, pour $t = (1+\epsilon)\log n$ où $\epsilon > 0$, on obtient avec probabilité au moins $1 - 2/n^\epsilon$ que $\|(Y - B) - \mathbb{E}(Y - B)\|_{op} < (\alpha - \beta)\log n$ dès que

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

On a donc démontrer le résultat suivant grâce à la construction d'un certificat dual stricte.

Théorème 6.7. *Soit A la matrice d'adjacence d'un graphe tiré selon le SBM de paramètre (n, p, q) tels que $p = \alpha\log(n)/n$ et $q = \beta\log(n)/n$ pour $0 < \beta < \alpha$ tels que $p+q < 1$. On note par $V_1 \sqcup V_1^c = V$ la structure de communautés sous-jacente à ce graphe et on suppose que $|V_1| = |V_1^c| = n/2$. On considère le 'community membership vector' $g \in \mathbb{R}^n$ tel que $g_i = 1$ si $1 \in V_1$ et $g_i = -1$ si $i \in V_1^c$. On a pour $B = 2A - \mathbf{1}\mathbf{1}^\top - I_n$ que*

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

avec probabilité au moins $1 - 2/n^\epsilon$ pour tout $\epsilon > 0$ dès que

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

Le Théorème 6.7 assure donc qu'il est possible de retrouver exactement la structure des communautés sous-jacente d'un graphe lorsque celui-ci est un graphe aléatoire tiré selon un SBM ayant des probabilités inter et intra communautés selon les hypothèse de ce théorème. De plus,

et ce qui est le plus important, on peut le faire de manière efficace en solutionnant un problème d'optimisation SDP. Ce qui est particulièrement remarquable est que cette procédure SDP est obtenue par relaxation convexe (suivant la méthode de matrix lifting / relaxation SDP) d'un problème qui est NP-hard en général. Cet aspect computationnel du problème d'origine (43) est donc particulièrement pessimiste vue qu'on vient de voir que pour la plupart des graphes obtenus par SBM, ce problème se résout en temps linéaire.

Concernant le seuil '$\sqrt{2(1+\epsilon)(\alpha+\beta)} + \frac{4(1+\epsilon)}{3} < \alpha - \beta$' obtenu dans le Théorème 6.7, il ne coïncide pas avec la transition de phase $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$ obtenue au Théorème 4.3 et on peut être sûr que le seuil du Théorème 6.7 est plus contraignant que cette transition de phase (sinon cela contredirait le Théorème 4.3). Il est cependant possible d'améliorer l'analyse faite plus haut sur $\lambda_2(Y - B)$ pour retrouver exactement ce seuil.

Théorème 6.8. *Soit A la matrice d'adjacence d'un graphe tiré selon le SBM de paramètre (n, p, q) tels que $p = \alpha \log(n)/n$ et $q = \beta \log(n)/n$ pour $0 < \beta < \alpha$ tels que $p+q < 1$. On note par $V_1 \sqcup V_1^c = V$ la structure de communautés sous-jacente à ce graphe et on suppose que $|V_1| = |V_1^c| = n/2$. On considère le 'community membership vector' $g \in \mathbb{R}^n$ tel que $g_i = 1$ si $1 \in V_1$ et $g_i = -1$ si $i \in V_1^c$. On a pour $B = 2A - \mathbf{1}\mathbf{1}^\top - I_n$ que*

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

avec probabilité au moins $1 - 2$ dès que $\sqrt{\alpha} - \sqrt{\beta} > \sqrt{2}$.

On voit donc grâce au Théorème 6.8 qu'il est possible d'atteindre la transition de phase en détection de communauté par un problème SDP. On dit alors qu'**il n'y a pas de gap computationnel pour le problème de détection de deux communautés de même taille pour les graphes aléatoires tirés selon un SBM**.

Conclusion : On voit une fois de plus sur le problème de la détection de communautés que **'l'aléatoire fait bien les choses'** en mettant dans des positions favorables des problèmes NP-hard en général qui peuvent dans ces cas être résolus en temps linéaire. On voit aussi que la relaxation convexe d'un problème même NP-hard est aussi une bonne chose à faire ; que ce soit par des calculs d'enveloppes convexes de fonctions objectives non-convexes (comme la 'norme ℓ_0' en Compressed Sensing ou la fonction de rang en complétion de matrices) ou par convexification de la contrainte (comme pour la méthode spectrale en détection de communautés) ou par matrix lifting / relaxation SDP (pour la méthode SDP en détection de communautés).

Certification duale en Compressed Sensing. La construction d'un certificat dual n'est pas propre au probléme SDP. En fait, on peut appliquer cette approche pour tout problème dès qu'on pense qu'il y a dualité forte (sans dualité forte, un certificat dual n'existe pas). On peut par exemple utiliser cette approche pour le problème de Basis Pursuit qu'on a rencontré au début du cours en Compressed Sensing. C'est cette méthode que nous développons dans ce paragraphe.

Le problème d'optimisation posé pour la construction de la procédure appelée 'basis pursuit' est le suivant :

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

où $A \in \mathbb{R}^{m \times N}$ et $y \in \mathbb{R}^m$ sont respectivement la matrice de mesure et le vecteur de mesures.

On propose de déterminer ce qu'est un certificat dual pour le problème (51) de deux manières. Dans la première approche, on fait apparaître une sous-différentielle. On note $\Omega = \{t \in \mathbb{R}^N : At = y\}$ et i_Ω l'indicatrice (au sens 'optimisation') de Ω càd pour tout $x \in \mathbb{R}^N$, $i_\Omega(x) = +\infty$ si $x \notin \Omega$

et $i_\Omega(x) = 0$ quand $x \in \Omega$. On voit alors qu'il est équivalent de résoudre (51) ou de trouver un minimum sur $\mathbb{R}^N$ de $f : x \in \mathbb{R}^N \to \|x\|_1 + i_\Omega(x)$ qui est lui-même équivalent à trouver un $x^* \in \mathbb{R}^N$ tel que $0 \in \partial^- f(x^*)$ où $\partial^- f(x)$ est la sous-différentielle de f en x donnée par

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

On a pour tout $x \in \mathbb{R}^N$ de support $I = \mathrm{supp}(x)$,

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

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

Ainsi $0 \in \partial^- f(x^*)$ si et seulement si $\partial^- i_\Omega(x^*) \cap \partial^- \|\cdot\|_1 (x^*) \neq \emptyset$ (on utilise ici que $\partial^- f(x^*) = \partial^- i_\Omega(x^*) + \partial^- \|\cdot\|_1 (x^*)$) càd s'il existe $\beta^* \in \mathbb{R}^m$ tel que $A^\top \beta^* \in \partial^- \|\cdot\|_1 (x^*)$ càd tel que $\|A^\top \beta^*\|_\infty \leq 1$ et $(A^\top \beta^*)_I = \mathrm{sgn}(x_I)$. Ce vecteur β^* est un certificat dual : étant donné un $x^* \in \mathbb{R}^N$ tel que $Ax^* = y$, si on peut construire un $\beta^* \in \mathbb{R}^m$ tel que $\|A^\top \beta^*\|_\infty \leq 1$ et $(A^\top \beta^*)_I = \mathrm{sgn}(x_I)$ alors on aura certifié que x^* est solution de (51).

On peut voir que β^* tel qu'introduit précédemment est bien un certificat dual au sens où nous l'avons introduit plus haut pour les problème SDP : càd comme un point faisable du problème dual pour lequel les fonctions objectives primales et duales coïncident respectivement en x^* et β^*. C'est cette construction qui permet de montrer l'optimalité d'un x^* basée sur la construction d'un certificat dual. La première chose à faire est de déterminer le problème dual de (51) et donc la fonction de Lagrange :

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

La fonction duale est

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

où t^*_β est un minimiseur de $t \to \mathcal{L}(t, \beta)$. Par convexité de la fonction objective, on voit que $t^*_\beta \in \mathrm{argmin}_{t \in \mathbb{R}^N} \mathcal{L}(t, \beta)$ si et seulement si $A^\top \beta \in \partial^- \|\cdot\|_1 (t^*)$ càd si et seulement si β et t^*_β sont tels que $\|A^\top \beta\|_\infty \leq 1$ et $(A^\top \beta)_I = \mathrm{sgn}((t^*_\beta)_I)$ où $I = \mathrm{supp}(t^*_\beta)$. On obtient alors pour tout $\beta \in \mathbb{R}^m$,

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

car $(A^\top \beta)_i (t^*_\beta)_i = |(t^*_\beta)_i|$ pour tout $i \in I$. Le problème dual est alors

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

La dualité faible dit que pour tout $\beta \in \mathbb{R}^m$ tel que $\|A^\top \beta\|_\infty \leq 1$ et $t \in, \mathbb{R}^N$ tel que $At = y$, on a $\|t\|_1 \geq \langle \beta, y\rangle$. On peut la vérifier directement ici vu que

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

L'approche par certification duale dit la chose suivante : soit $t^* \in \mathbb{R}^N$ tel que $At^* = y$, pour certifier que t^* est bien solution de (51), il suffit de trouver $\beta^* \in \mathbb{R}^m$ tel que $\|A^\top \beta^*\|_\infty \leq 1$ et $\langle y, \beta^*\rangle = \|t^*\|_1$. Un tel β^* certifiera bien que t^* est solution de (51) car pour tout $t \in \mathbb{R}^N$ on a

$\|t\|_1 \geq \langle \beta^*, y \rangle$ par dualité forte et donc si $\langle y, \beta^* \rangle = \|t^*\|_1$ on aura aussi $\|t\|_1 \geq \|t^*\|_1$ et comme t^* est faisable pour le problème primal, on conclut bien que t^* est solution de (51). Pour retrouver exactement les conditions annoncées plus haut via la sous-différentielle, il suffit de remarquer que la condition de dualité forte en (t^*, β^*), càd $\langle y, \beta^* \rangle = \|t^*\|_1$ s'écrit aussi $\langle t^*, A^\top \beta^* \rangle = \|t^*\|_1$ vu que $At^* = y$. Mais comme $\left\|A^\top \beta^*\right\|_\infty \leq 1$, on a $\langle t^*, A^\top \beta^* \rangle = \|t^*\|_1$ si et seulement si $(A^\top \beta^*)_I = \text{sgn}(t^*_I)$ où I est le support de t^*.

On retrouve bien dans les deux approches (celle basée sur la sous-différentielle et celle basée sur la dualité Lagrangienne) que la construction d'un certificat dual pour un candidat $t^* \in \mathbb{R}^N$ tel que $At^* = y$ au problème (51) est donnée par le problème de recherche d'un $\beta^* \in \mathbb{R}^m$ tel que $(A^\top \beta^*)_I = \text{sgn}(t^*_I)$ où I est le support de t^* et $\left\|A^\top \beta^*\right\|_\infty \leq 1$.

Certification duale stricte : unicité d'une solution. Que ce soit pour le problème SDP qu'on a rencontré plus haut en détection de communautés ou pour le problème du basis pursuit, on peut certifier qu'un candidat solution au problème primal est effectivement solution de ce problème par la construction d'un certificat dual. Cette approche ne permet cependant pas de démontrer l'unicité. Cependant, on peut aussi certifier l'uncité de la solution par la construction d'un certificat qu'on appelle stricte. C'est l'approche qu'on a suivie en détection de communauté.

Proposition 6.9. *Etant donné un problème d'optimisation dont φ est la fonction objective primale et ψ la fonction objective duale. Si on suppose qu'il y a dualité forte et que λ^* est solution du problème dual alors on a équivalence entre :*

1) x^* est solution du problème primal

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

La Proposition 6.9 nous assure que pour certifier une solution primale il suffit de vérifier que $\varphi(x^*) = \psi(\lambda^*)$ pour n'importe quelle solution λ^* du problème dual. En particulier, si on a construit un certificat dual pour un candidat solution x_0^* du primal (qui est donc faisable pour le primal) alors ce certificat peut aussi nous servir à certifier d'autres solutions du problème primal s'il y en a. En effet, si on a montré l'existence d'un certificat dual λ^* alors il y a dualité forte et aussi λ^* est solution du dual. On peut donc appliquer la Proposition 6.9. En particulier, si on montre que dès que x^* est faisable pour le primal et $x^* \neq x_0^*$ alors $\varphi(x^*) > \psi(\lambda^*)$ on pourra en déduire que x_0^* est l'unique solution du problème primal. La construction d'un certificat dual qui permet une telle inégalité stricte avec tous les éléments faisable du primal différent de x_0^* sont appelés des **certificats duaux strictes**. Ce sont eux qui permettent de certifier qu'un point faisable du primal est l'**unique** solution du primal. C'est ce type de construction que nous avons fait en détection de communautés plus haut.

7 *Matrix lifting* et Relaxation SDP

Dans les sections précédentes, nous avons utilisé une méthode de relaxation convexe principalement pour 'convexifier' les contraintes qui avaient une 'nature' combinatoire. En compressed sensing ou matrix completion on avait plutôt convexifier la fonction objectif. Dans le cas de la relaxation obtenue en (28) pour le problème de la détection de communautés, on parle de relaxation SDP, car le problème obtenu en final est un problème *semi-definite programming (SDP)*. On parle aussi de *matrix lifting* car on a réécrit la fonction objectif $x^\top (A - \lambda J_n) x$ qui est quadratique en x en une fonction objectif $\langle A - \lambda J_n, xx^\top \rangle$ qui est linéaire en $xx^\top$. Ainsi en passant de la variable x à la variable $Z = xx^\top$ on passe d'un problème vectoriel de degrés 2 à un problème matriciel linéaire. Cette astuce de '*matrix lifting*' permet de convexifier la contrainte tout simplement en

enlevant la contrainte que Z doit être de rang 1. C'est en fait une relaxation convexe qu'on rencontre souvent quand on travail avec des fonctions quadratiques d'un vecteur. Nous donnons deux autres exemples où cette astuce de matrix lifting peut être utilisée.

Méthode de *Matrix lifting* et Relaxation SDP pour la recherche d'un plus grand vecteur propre. Dans ce premier exemple, on mets en oeuvre la méthode de matrix lifting pour un problème simple qui est de trouver un plus grand vecteur propre d'une matrice symétrique : soit $A \in \mathbb{R}^{n\times n}$ une matrice symétrique, on cherche un vecteur $u_1 \in \mathbb{R}^n$ tel que

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

La *power method* est probablement un des algorithmes les plus rapides pour approcher une solution au problème (54) ; elle consiste à choisir un point x_0 aléatoirement dans $\mathbb{R}^n$ et à itérer $x_{k+1} = (Ax_k)/\|Ax_k\|_2$. Si on dispose de factorisations particulières de A, on peut aussi les utiliser pour trouver une solution à (54). L'approche qu'on utilise ici est basée sur le *matrix lifting* et constitue un bon exemple introductif à cette méthode. Néanmoins, il se trouve que dans le cas de (54), cette relaxation est en faite *exacte* ; ce qui n'est pas classique en matrix lifting. On reviendra sur ce point plus tard.

Dans l'approche de matrix lifting, la première étape est de faire apparaître une fonction linéaire en $xx^\top$. On a ici pour tout $x \in \mathbb{R}^n$,

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

où on utilise le produit scalaire matriciel dans la dernière inégalité (i.e. $\langle A, B\rangle = \mathrm{Tr}(A^\top B)$). On a donc bien réécrit (le carré de) la fonction objectif de (54) comme une application linéaire de $xx^\top$. La deuxième étape est de réinterpréter les contraintes sur x en contraintes sur $xx^\top$. Ici on contraint x à être tel que $\|x\|_2 = 1$. Cette contrainte est équivalente à demander que la trace de $xx^\top$ soit égale à 1, car $\mathrm{Tr}(xx^\top) = \sum_i x_i^2 = \|x\|_2^2$. On a donc 'lifter' le problème (54) sous la forme :

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

Les contraintes $Z \succeq 0$ et $\mathrm{rang}(Z) = 1$ sont ici pour assurer que les solutions au problème de droite de (55) sont de la forme $xx^\top$ et non pas $xy^\top$. On a en effet,

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

Il y a donc bien égalité en (55).

La réécriture effectuée en (55) d'un problème quadratique vectoriel en un problème linéaire matriciel est au coeur de l'approche de matrix lifting. Pour retrouver une solution u_1 à partir d'une solution Z_1 au problème matriciel, il suffit de prendre la première colonne de Z_1 et de la normaliser.

Néanmoins, le problème matricielle à droite de (55) a toujours le désavantage d'être un problème sous contrainte non-convexe à cause de la contrainte de rang. L'idée en relaxation SDP est simplement d'enlever la contrainte de rang pour obtenir un problème de type SDP :

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

C'est bien un problème de type SDP car la contrainte porte sur le cône des matrices SDP (semi-definite positive) telles que $\mathrm{Tr}(Z) = 1$ qui est bien une contrainte affine (on peut la réécrire comme

$\langle Z, I_n \rangle = 1$) et la fonction objective $Z \to \langle A^2, Z \rangle$ est linéaire. Comme on a enlevé une contrainte, on a

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

En général, cette inégalité n'est pas une égalité et on a alors un 'gap' entre le problème d'origine et le problème relâché. Cependant, il se trouve que dans le cas particulier du problème considéré ici, il y a en fait égalité. C'est pour ça qu'on dit que la relaxation est exacte ici.

Pour montrer qu'on a bien égalité dans (56), on considère la décomposition en valeurs propres de $A^2 = \sum_i \lambda_i^2 u_i u_i^\top$ et on a pour tout $Z \succeq 0$ tel que $\mathrm{Tr}(Z) = 1$,

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

où on a écrit $Z = XX^\top$ pour $X \in \mathbb{R}^{n \times r}$ et $r = \mathrm{rang}(Z)$ (on peut toujours écrire une matrice $Z \succeq 0$ sous cette forme grâce à sa décomposition en valeurs singulières). Mais, on voit que si u_1 est un 'top eigenvector' de A alors on a $\langle A^2, u_1 u_1^\top \rangle = \lambda_1^2$ alors

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

On en déduit donc qu'on a bien égalité dans (56) et que la relaxation est exacte. On peut aussi voir que si on note par $E_1 = \mathrm{argmax}_{x : \|x\|_2 = 1} \|Ax\|_2^2$ l'ensemble des solutions du problème initial alors

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

où $\mathrm{conv}(E_1)$ est l'enveloppe convexe de E_1. En particulier, si A a une plus grande valeur propre de multiplicité 1 alors $E_1 = \{-x^*, x^*\}$ où x^* est un plus grand vecteur propre de A. Dans ce cas $\mathrm{argmax}_{Z \succeq 0 : \mathrm{Tr}(Z)=1} \langle A^2, Z \rangle = \{x^*(x^*)^\top\}$ et donc on obtient x^* ou $-x^*$ à partir de $x^*(x^*)^\top$ en prenant la première colonne de $x^*(x^*)^\top$ et en la normalisant par sa norme ℓ_2^n.

Méthode de *Matrix lifting* et Relaxation SDP pour le problème de MAX-CUT et relaxation Lagrangienne biduale. Il existe de nombreux problèmes d'optimisation sur les graphes. Ces problèmes ont connus d'importants développements depuis l'émergence d'internet et de réseaux sociaux. On donne ici un exemple de problème d'optimisation sur un graphe connu sous le nom de problème de MAX-CUT (en français : coupure maximale) qui permet de mettre en œuvre les astuces de matrix lifting et de relaxation SDP

On se donne un graphe $G = (V, E)$ où V est l'ensemble des sommets, généralement, on prends $V = \{1, \ldots, n\}$ et E est l'ensemble des arêtes du graphe : c'est un sous-ensemble de $V \times V$ ($(i, j) \in V$ ssi il y a une arête entre le noeud i et le noeud j). On suppose que le graphe est non orienté, ce qui est équivalent à dire que si $(i, j) \in V$ alors $(j, i) \in V$.

Un **CUT** de G est une partition de ces sommets en deux parties : $S \sqcup S^c = V$ (où S^c est le complémentaire de S dans V). MAX-CUT est le nom qu'on donne au problème qui consiste à trouver un CUT de G – càd un sous-ensemble $S \subset V$ – tel que le nombre d'arêtes entre S et son complémentaire S^c est maximal. On écrit formellement ce problème du MAX-CUT par

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

Il est souvent utile de transformer les problèmes d'optimisation sur les graphes en des problèmes d'optimisation sur les matrices. Pour cela, on introduit la **matrice d'adjacence** d'un

graphe $G = (V, E)$ où $V = \{1, \dots, n\}$ par

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

En utilisant cette notation on voit que pour tout $i, j \in \{1, \dots, n\}$, on a $I(i \in S \text{ et } j \in S^c) = (1 - x_i x_j)/2$ où $x = (x_i)_{i \in V} \in \{-1, 1\}^n$ est tel que $x_i = 1$ si $i \in S$ et $x_i = -1$ si $i \in S^c$ et que sommer sur V est équivalent à sommer sur $\{1, \dots, n\}$ si on multiplie le terme de sommation par A_{ij}. On voit alors que le problème du MAX-CUT peut se réécrire comme le problème d'optimisation suivant :

$$\max_{x \in \{-1,1\}^n} \frac{1}{2} \sum_{i,j=1}^{n} A_{ij}(1 - x_j x_i). \tag{59}$$

En effet, si x^* est solution du problème précédent alors en posant $S^* = \{i \in \{1, \dots, n\} : x_i^* = 1\}$, on obtient une solution au problème initial.

On peut aussi réécrire le problème (59) sous la forme d'un problème de minimisation d'une fonctionnelle quadratique sous-contraintes :

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

C'est donc un problème d'optimisation sous contrainte ; la fonction objectif est une fonction quadratique $x \to x^\top A x$ et la contrainte est donnée par n contraintes d'égalité "$x_i^2 = 1$". La difficulté ici est que la contrainte n'est pas un ensemble convexe. Une approche possible est alors de 'convexifier' cet ensemble.

Les contraintes "difficiles" sont ici toutes les $x_i^2 = 1, i = 1, \dots, n$ car elles imposent aux x_i d'être discrètes (ici $x_i^2 = 1$ ssi $x_i \in \{-1, 1\}$). Ce sont ces contraintes qui font de MAX-CUT un problème combinatoire. On va alors convexifier ces contraintes en utilisant l'approche matrix lifting / relaxation SDP. On montre ensuite que cette approche est ici équivalente à une double dualisation Lagrangienne du problème, appelée relaxation Lagrangienne biduale.

Pour l'approche de matrix lifting, on écrit la fonction objectif comme une fonction linéaire de $xx^\top$: on a $x^\top A x = \langle A, xx^\top \rangle$. On écrit les contraintes sur x en contraintes sur $Z = xx^\top$: $x_i^2 = 1$ ssi $Z_{ii} = 1$. On a donc

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

La relaxation SDP nous dit ensuite de simplement retirer la contrainte de rang. On obtient ainsi le problème d'optimisation

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

Cette relaxation SDP de MAX-CUT est assez célèbre et porte le nom de la relaxation de Goemans et Williamson. Il est possible de trouver une solution approchante la solution du problème du MAX-CUT (58) à partir d'une solution Z_1 de (62). Pour cela, on prend $G \sim \mathcal{N}(0, Z_1)$ et $\hat{u} = \operatorname{sign}(G)$. Alors le Théorème de Goemans et Williamson montre que

$$\mathbb{E}\left(\frac{1}{2} \sum_{i,j=1}^{n} A_{ij}(1 - \hat{u}_j \hat{u}_i) \right) \geq 0.868 \max_{x \in \{-1,1\}^n} \frac{1}{2} \sum_{i,j=1}^{n} A_{ij}(1 - x_j x_i).$$

Le preuve utilise la même technique que celle utilisée dans la démonstration de l'inégalité de Grothendieck. Cette technique porte le nom de *rounding technic* et se retrouve dans beaucoup de théorèmes tel que le théorème de Nesterov.

On montre maintenant l'équivalence de l'approche matrix lifting / relaxation SDP et la relaxation Lagrangienne biduale. La fonction de Lagrange obtenue par relaxation Lagrangienne des contraintes difficiles associée au problème du MAX-CUT est

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

où $\mathrm{Diag}(u)$ est la matrice diagonale de taille $n \times n$ dont les éléments diagonaux sont donnés par les coordonnées de u et $e = (1)_1^n$. La fonction duale est

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

Le problème dual est donc le problème

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

En posant $F(u) = A - \mathrm{Diag}(u)$ et en notant $\mathcal{S}_+^n$ le cône des matrices symétriques semi-définie positives, on voit que le problème dual est un problème sous contrainte conique. On peut alors dualiser ce problème comme nous l'avons vu plus haut. Ici, (63) est déjà un problème dual (c'est le dual de MAX-CUT), on parle alors de **problème bidual**. La fonction dual de (63) est

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

où $\mathrm{Diag}(X)$ est le vecteur diagonal de taille n dont les coordonnées sont données par les éléments diagonaux de X et $(\mathcal{S}_+^n)^\circ$ est le cône dual de $\mathcal{S}_+^n$. On peut montrer que c'est le cône des matrices symétriques semi-définies négatives. On obtient alors que la fonction duale de (63) est

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

En remplaçant X par $-X$, on voit que le problème dual est

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

On obtient donc ici par relaxation Lagrangienne biduale exactement le même problème d'optimisation que par l'approche de matrix lifting et relaxation SDP. En utilisant cette dernière remarque et la dualité faible (de la première dualisation), on obtient l'encadrement suivant de la relaxation SDP de MAX-CUT :

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

Matrix lifting pour des contraintes linéaires et quadratiques en détection de communautés. En détection de deux communautés de même taille, on a rencontré le problème d'optimisation

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

Dans l'approche de la Section 6, on a introduit une méthode de matrix lifting basée uniquement sur le lifting de x en $xx^\top$. En faisant celà, on a pu réécrire les termes seulement de dégrés deux exactement car il n'y a que des termes exactement de degrés deux dans la matrice $xx^\top$. Cependant

dans le problème (65), il y a aussi une contrainte linéaire : '$\langle x, 1\rangle = 0$' qui ne peut pas être réécrite en fonction linéaire de $xx^\top$. On a donc simplement enlever cette contrainte dans la Section 6 mais au préalable on a introduit un terme de pénalité dans la fonction objective, c'est par cette astuce qu'on a pu montrer que le problème obtenu en enlevant la contrainte '$\langle x, 1\rangle = 0$' permettait quand même de faire de la reconstruction exacte des communautés.

Dans ce paragraphe, on propose une autre méthode de matrix lifting qui permet de gérer les termes linéaires (dans la contrainte et/ou dans la fonction objective). L'idée est de faire le 'lifting' suivant :

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

En augmentant juste la dimension d'une unité en ajoutant une coordonnée égale à 1 à x, on voit apparaître des termes linéaires dans la matrice liftée Z. On peut alors réécrire les termes linéaires en x comme des termes linéaires en Z (en fait, en les éléments de la première ligne ou colonne de Z). On obtient ainsi le problème lifté du problème initial (65) suivant :

$$\min\left(\langle L, Z\rangle : Z \succeq 0, \operatorname{rang}(Z) = 1, Z_{00} = 1, Z_{ii} = 1, i = 1, \dots, n, \langle Z, A\rangle = 0\right) \tag{67}$$

où $A \in \mathbb{R}^{(n+1)\times(n+1)}$ est donné par $A_{ij} = 1$ si $i = 1$ et $j \in \{1, \dots, n\}$. On indexe ici les entrées des matrices Z et A par $\{0, 1, \cdots, n\}^2$ (plutôt que par $\{1, \dots, n+1\}^2$). Les deux problèmes (65) et (67) sont équivalents : $\hat{x}$ est solution de (65) ssi $\left(\begin{smallmatrix} 1 \\ \hat{x} \end{smallmatrix}\right)\left(\begin{smallmatrix} 1 & \hat{x}^\top \end{smallmatrix}\right)$ est solution de (67). L'étape suivante est de faire de la relaxation SDP en enlevant la contrainte de rang. On obtient alors

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

qui est bien un problème SDP.

On peut donc faire du matrix lifting qui tient en compte les termes linéaires à condition d'ajouter une dimension au problème et de faire le lifting sur $\left(\begin{smallmatrix} 1 \\ x \end{smallmatrix}\right)$ plutôt que sur x comme indiqué dans (66).