Réseaux Parcimonieux et Détection de Communautés
Projet de *Compressed Sensing*

Gautier Appert
gautier.appert@ensae.fr

Guillaume Salha
guillaume.salha@ensae.fr

Résumé

L'objectif de ce projet est de proposer une revue de l'article *Community Detection in Sparse Networks via Grothendieck's Inequality* de Guédon et Vershynin (2015). Dans cet article, les auteurs présentent une méthode flexible pour prouver la précision de certains problèmes d'optimisation SDP (pour *semidefinite programming*) pour la résolution de problèmes stochastiques variés. La méthode, basée notamment sur l'inégalité de Grothendieck, est générale. Nous l'appliquons ici au problème de détection de communautés et notamment au sein de réseaux parcimonieux. Après un tour d'horizon des principaux résultats mathématiques de l'article, nous proposons une implémentation sur des graphes construits via le classique *Stochastic Block Model*. Ces implémentations illustrent la capacité de la méthode à détecter la structure de communauté même au sein de réseaux parcimonieux, avec un taux d'erreur arbitrairement fixé. Nous comparons enfin la méthode avec le clustering spectral. Cette approche alternative permet d'obtenir des résultats satisfaisant pour des graphes relativement denses, mais s'avère moins adaptée que celle de Guédon et Vershynin (2015) pour des graphes parcimonieux.

1 Introduction

Les graphes sont des modèles utiles lorsqu'il est nécessaire de travailler sur les relations entre des entités. En effet, une variété de problèmes concrets peuvent être représentés par des noeuds reliés entre eux par des arêtes, allant des réseaux - sociaux, biologiques, informatiques - aux pixels d'une image. La détermination de clusters pertinents parmi ces noeuds fut l'objet d'une recherche importante en apprentissage. Les applications incluent la segmentation d'images, la catégorisation de pages web, ou encore la reconstitution de groupes sociaux au sein de Facebook ou Twitter.

Pour ce problème de détection de communauté au sein de réseaux, sur lequel nous nous concentrons au sein de ce projet, les résultats de recherche sont souvent basés sur le classique *Stochastic Block Model* [Holland *et al.*, 1983], un modèle génératif pour les graphes aléatoires présenté en Section 2. Ainsi, de nombreuses approches algorithmiques ont été proposées pour reconstituer des communautés dans un graphe, basées sur des techniques combinatoires [Bui *et al.*, 2011 ; Dyer et Frieze, 1989], sur des méthodes variationnelles [Airoldi *et al.*, 2008 ; Celisse *et al.*, 2012], sur des méthodes MCMC [Snijders et Nowicki, 1997 et 2001], sur du clustering spectral [Alon, 1998 ; Newman, 2006 ; Ailon *et al.*, 2014] ou encore sur de l'optimisation convexe et notamment sur la résolution de problèmes SDP [Jalali *et al.*, 2011 ; Chen *et al.*, 2014 ; Cai et Li, 2014]. Toutefois, la plupart des résultats rigoureux connus ont été prouvés pour des réseaux relativement denses, dont l'espérance des degrés des noeuds tend vers l'infini avec n, à une certaine vitesse. Guédon et Vershynin, à l'inverse, proposent dans leur article une méthode permettant d'obtenir, sous certaines conditions développées plus loin, des résultats théoriques y compris pour des graphes plus parcimonieux, que nous définissons plus précisement en Section 2.

Plus généralement, Guédon et Vershynin présentent une méthode générale pour montrer la consistance de problèmes d'optimisation SDP sur des graphes aléatoires. Mettons de côté la détection de communautés pour un instant. Supposons simplement que nous observons une matrice aléatoire A, de taille $n \times n$ et d'espérance inconnue $\bar{A}$. Nous voudrions estimer la solution du problème d'optimisation suivant :

$$\max_{x \in \{-1,1\}^n} x^T \bar{A} x. \tag{1}$$

Comme $\bar{A}$ est inconnue, nous espérons pouvoir estimer la solution $\bar{x}$ de (1) via la résolution de :

$$\max_{x \in \{-1,1\}^n} x^T A x. \tag{2}$$

Dans le cadre de la détection de communautés, que nous détaillerons en Section 2, A représentera la matrice d'adjacence d'un graphe aléatoire, et $\bar{x}$ un vecteur précisant la répartition des n noeuds du graphe en deux communautés. Si nous supposons l'existence de deux communautés de même taille, alors le vecteur doit contenir autant de 1 que de -1. La pertinence de la résolution d'un tel problème d'optimisation pour une bonne reconstitution des communautés sera expliquée avec davantage de détails dans ce projet.

Comme l'expliquent Guédon et Vershynin, une difficulté majeure provient du fait que le problème (2) soit NP-hard en général. Plusieurs relaxations de problèmes d'optimisation de ce type ont été proposées dans la littérature [Goemans et Williamson, 1995 ; Nesterov, 1998 ; Alon et Naor, 2006]. Ces relaxations conduisent à des niveaux de précision constants[1].

Dans leur article, Guédon et Vershynin montrent comment leur relaxation semi-définie de (2) permet d'approcher la solution de (1) *à tout niveau de précision fixé*, en se basant notamment sur l'inégalité de Grothendieck. Contrairement à d'autres méthodes présentées dans la littérature, les auteurs appliquent l'inégalité de Grothendieck, non pas sur la matrice A, mais sur la matrice des erreurs $A - \bar{A}$, ce qui conduit à cette précision fixée arbitrairement. Nous soulignons aussi que la méthode, contrairement à d'autres, est adaptée à l'étude de réseaux parcimonieux i.e. dont la moyenne des degrés des noeuds est bornée.

Dans la suite de ce projet, nous proposons un tour d'horizon des éléments clés de l'approche de Guédon et Vershynin. En Section 2, nous rappelons tout d'abord le principe du *Stochastic Block Model* pour deux communautés de tailles identiques, et nous énonçons le résultat principal de l'article concernant la reconstruction de la structure de communautés dans un tel modèle, y compris pour des réseaux parcimonieux. En Section 3, nous synthétisons la démarche mathématique des auteurs pour aboutir à leur résultat. Nous expliquons plusieurs résultats intermédiaires prouvés dans l'article, en nous attardant notamment sur l'inégalité de Grothendieck et sur la déviation en *cut-norm*. Puis, à partir de ces résultats intermédiaires, nous démontrons le résultat énoncé en Section 2.

Plusieurs implémentations sont ensuite proposées en Section 4, en mettant l'accent sur les réseaux parcimonieux et en comparant notre méthode au clustering spectral. Enfin, nous étendons nos résultats au problème de détection de communautés dans un cadre plus général à k communautés de tailles inconnues en Section 5, et nous concluons en Section 6.

2 Stochastic Block Model et Détection de Communautés

Dans cette section, nous présentons rapidement le *Stochastic Block Model*, pour la génération de graphes aléatoires à deux communautés de même taille. Nous définissons plus précisément le problème de détection de communautés dans un tel modèle ainsi que la problématique des graphes parcimonieux, que nous définissons plus précisement. Enfin, l'un des deux théorèmes principaux de l'article de Guédon et Vershynin est énoncé.

2.1 Génération de graphes aléatoires

En théorie des graphes, le modèle le plus connu pour générer des graphes aléatoires est sans doute le modèle de Erdős–Rényi [Erdős et Rényi, 1959], noté $G(n, p)$. Ce modèle permet de générer un graphe à n noeuds, en plaçant une arête entre chaque paire de noeuds avec une probabilité $p \in [0, 1]$. Une généralisation de ce modèle, adapté au problème de détection de communautés, est le classique *Stochastic Block Model* [Holland *et al.*, 1983 ; Mossell *et al.*, 2012]. Jusqu'à la Section 6 de ce rapport, nous nous situons dans le cadre simple où il n'y a que deux communautés $\mathcal{C}_1$ et $\mathcal{C}_2$, de même taille $n/2$. Nous étendrons nos résultats au cadre plus général à k communautés de tailles inconnues en fin de projet.

Dans ce cadre simple, le *Stochastic Block Model* noté $G(n, p, q)$ permet de générer des graphes aléatoires de taille n en procédant ainsi : pour chaque paire de noeuds, nous plaçons une arête entre les noeuds avec une probabilité $p \in [0, 1]$ s'ils appartiennent à la même communauté, et avec une probabilité $q \leq p$ s'ils appartiennent à des communautés différentes. Lorsque $p = q$, nous retrouvons le modèle de Erdős–Rényi. La Figure 1, ci-contre, représente un graphe - considéré comme "dense" - de 200 noeuds généré à partir d'un *Stochastic Block Model*, avec $p = 0.3$ et $q = 0.005$. Pour des raisons mathématiques illustrées par la suite, les auteurs incluent les boucles dans les graphes générées : chaque noeud est relié à lui-même avec probabilité 1.

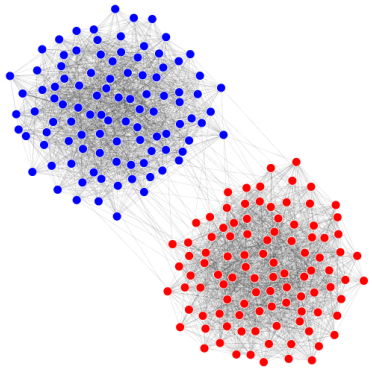

Figure 1: Graphe généré à partir de $G(200, 0.3, 0.005)$

[1] Dans [Alon et Naor, 2006] par exemple, la relaxation conduit à une solution $x_0 \in \{-1, 1\}^n$ telle que $x_0^T A x_0 \geq 0.56 \max_{x \in \{-1,1\}^n} x^T A x$ pour toute matrice A semi-definie positive.

2.2 Communautés et réseaux parcimonieux

Le problème de détection de communautés dans un graphe se formule ainsi : il s'agit de retrouver les communautés $\mathcal{C}_1$ et $\mathcal{C}_2$, i.e. déterminer quels noeuds appartiennent à chacun des ensembles, à partir d'une réalisation aléatoire de $G(n, p, q)$. La plupart des résultats obtenus pour résoudre ce problème, dans les articles évoqués en introduction, sont valables pour des graphes relativement denses, c'est-à-dire ceux dont l'espérance des degrés des noeuds - leur nombre de connexions - est $\Omega(log\ n)$. Bien moins de résultats sont disponibles pour des graphes parcimonieux, dont l'espérance des degrés des noeuds est bornée, ou bien croît à une vitesse arbitrairement plus petite que $log\ n$.

L'étude des réseaux parcimonieux a pourtant un intérêt pratique, comme l'illustrent [Lelarge *et al.*, 2013]. Dans le cadre des systèmes de recommandation, nous pourrions considérer une situation dans laquelle les utilisateurs ne donnent des évaluations qu'à un nombre restreint d'articles. De même, dans un réseau biologique, nous pourrions imaginer n'avoir accès qu'à un faible nombre d'interactions protéine-protéine, à cause de contraintes de coûts. L'article de Guédon et Vershynin que nous étudions, précisément, se consacre tout particulièrement à la détection de communautés dans les réseaux parcimonieux.

2.3 Détection de communautés : Théorème 1

Mathématiquement, retrouver les deux communautés $\mathcal{C}_1$ et $\mathcal{C}_2$ consiste à estimer le vecteur de communauté $\bar{x}$, définit ainsi :

$$\bar{x} \in \{-1, 1\}^n, \qquad \bar{x}_i := \begin{cases} 1 & \text{si } i \in \mathcal{C}_1 \\ -1 & \text{si } i \in \mathcal{C}_2 \end{cases} \tag{3}$$

Soit A la matrice d'adjacence de taille $n \times n$ associée au graphe, i.e. la matrice telle que :

$$A_{i,j} = \mathbf{1}_{\{\text{les noeuds } i \text{ et } j \text{ sont reliés}\}}.$$

La démarche de l'article de Guédon et Vershynin consiste à résoudre le problème d'optimisation SDP suivant :

$$\begin{aligned} &\max_{Z} \ \langle A, Z\rangle - \lambda\langle E_n, Z\rangle \\ &\text{s.c. } Z \succeq 0, \ \text{diag}(Z) \preceq I_n. \end{aligned} \tag{4}$$

Nous avons repris ici les notations de l'article. Nous définissons le produit scalaire matriciel $\langle A, B\rangle := tr(A^T B) = \sum_{i,j} A_{i,j} B_{i,j}$, la notation $Z \succeq 0$ signifie que la matrice Z est semi-définie positive, et $\text{diag}(Z) \preceq I_n$ signifie que les valeurs de la diagonale de Z sont bornées par 1. Par ailleurs, $E_n := \mathbf{1}_n \mathbf{1}_n^T$ et le scalaire λ correspond au degré moyen parmi les noeuds du graphe, une fois les boucles retranchées :

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

L'un des enjeux principaux de l'article de Guédon et Vershynin consiste à prouver le théorème suivant :

Théorème 1. *(Détection de deux communautés dans le Stochastic Block Model : estimation de $\bar{x}\bar{x}^T$)*
Soit $\varepsilon \in (0, 1)$ et $n \geq 10^4 \varepsilon^{-2}$. Soit A la matrice d'adjacence du graphe aléatoire tiré du Stochastic Block Model $G(n, p, q)$ avec $\max\{p(1-p), q(1-q)\} \geq \frac{20}{n}$. Supposons que $p = \frac{a}{n} > q = \frac{b}{n}$, et :

$$(a-b)^2 \geq 10^4 \epsilon^{-2}(a+b). \tag{5}$$

Soit $\hat{Z}$ une solution du problème (4). Alors, avec probabilité au moins $1 - e^3 5^{-n}$, nous avons :

$$\|\hat{Z} - \bar{x}\bar{x}^T\|_2^2 \ \leq \varepsilon n^2 = \varepsilon \|\bar{x}\bar{x}^T\|_2^2. \tag{6}$$

La notation $\|.\|_2$ désigne, dans ce projet, la norme de Frobenius lorsqu'il s'agit de matrices et la norme Euclidienne lorsqu'il s'agit de vecteurs. À partir de l'estimation $\hat{Z}$ de la matrice de rang un $\bar{x}\bar{x}^T$ via le Théorème 1, le vecteur de communauté $\bar{x}$ s'estime facilement, de cette manière :

Corollaire 1. *(Détection de deux communautés dans le Stochastic Block Model : estimation de $\bar{x}$)*
Dans le cadre du Théorème 1, soit $\hat{x}$ le vecteur propre de $\hat{Z}$ associé à sa plus grande valeur propre, et avec $\|\hat{x}\|_2 = \sqrt{n}$. Alors:

$$\min_{\alpha=\pm 1} \|\alpha\hat{x} - \bar{x}\|_2^2 \leq \varepsilon n = \varepsilon \|\hat{x}\|_2^2. \tag{7}$$

En particulier, les signes des coefficients de $\hat{x}$ estiment correctement la partition des noeuds en deux communautés, avec au plus εn erreurs parmi ces noeuds.

Nous verrons au cours de la section suivante comment aboutir à un tel résultat. Notons que ce n'est pas la première fois que de tels problèmes d'optimisation ont été utilisés dans le cadre de la détection de communautés [Amini et Levina, 2014]. Toutefois, la question de savoir si de telles méthodes pouvaient être adaptées à des graphes

parcimonieux est longtemps restée en suspens. Le Théorème 1, énoncé au dessus, montre que c'est effectivement le cas : il peut en particulier être appliqué pour des graphes dont l'espérance des degrés des noeuds - qui correspond ici à a et b - n'augmente pas avec n.

Bien évidemment, le problème reste bien plus difficile pour des graphes parcimonieux. Comme l'expliquent Guédon et Vershynin, si les degrés des noeuds augmentent à une vitesse inférieure à $log\ n$, alors avec une probabilité élevée certains noeuds du graphe seront isolés. Il sera alors impossible de les classifier correctement - ou du moins, pas mieux que via une classification aléatoire. La proportion de noeuds isolés tendant vers zero avec n, nous pouvons toutefois espérer d'obtenir un résultat pertinent pour une majorité des noeuds du graphe, comme cela sera illustré lors de nos implémentations.

3 Méthodologie mathématique

Dans cette section, nous présentons la démarche mathématique des auteurs pour aboutir au résultat que nous venons d'énoncer. Nous nous concentrons ici sur les intuitions et sur la méthodologie. Démontrer l'ensemble des lemmes intermédiaires aurait en effet été beaucoup trop long pour le format de ce projet. Certains résultats de cette section sont donc admis : nous invitons le lecteur intéressé à se reporter à l'article de Guédon et Vershynin pour les détails techniques. Ces lemmes nous serviront en fin de section, pour démontrer le Théorème 1 lui-même, ainsi que son corollaire.

Revenons tout d'abord au problème (2), très général, énoncé en introduction:

$$\max_{x\in\{-1,1\}^n} x^T A x \quad \Leftrightarrow \quad \max_{x\in\{-1,1\}^n} \langle A, xx^T\rangle. \tag{8}$$

Comme nous l'avons déjà annoncé, un tel problème est NP-hard[2]. Ceci amène les auteurs à considérer une relaxation convexe du problème:

$$\max_{Z\in\mathcal{M}_{opt}} \langle A, Z\rangle, \tag{9}$$

pour un certain sous-ensemble convexe $\mathcal{M}_{opt}$ de matrices semi-définies positives. L'objectif est de montrer que la solution $\hat{Z}$ de (9) peut permettre d'estimer avec une certaine précision la matrice $\bar{x}\bar{x}^T$ puis le vecteur $\bar{x}$ solution de (1). Le choix de $\mathcal{M}_{opt}$ dépend du problème considéré. Dans le cadre de la détection de deux communautés de même taille, Guédon et Vershynin considèrent l'ensemble des matrices semi-définies positives dont les entrées des diagonales sont bornées par 1 (en valeur absolue) :

$$\mathcal{M}_{opt} = \mathcal{M}_G^+ := \Big\{ Z \in \mathcal{M}_{n\times n} : Z \succeq 0,\ \text{diag}(Z) \preceq I_n \Big\}.$$

Par ailleurs, la matrice A de (9) correspond, pour ce même problème, à la matrice $B := A - \lambda E_n$ en reprenant les notations de (4).

Afin que la solution $\hat{Z}$ de (9) puisse effectivement permettre d'estimer $\bar{x}$ avec une certaine précision, plusieurs conditions sont à vérifier au sein de cette relaxation convexe. Nous les présentons une par une.

3.1 Maximisation de la fonction-objectif de réference

Tout d'abord, il faut s'assurer que l'ensemble $\mathcal{M}_{opt}$ est *tight* – "sérré" - autrement dit que la solution $\bar{Z}$ du programme :

$$\max_{Z\in\mathcal{M}_{opt}} \langle \bar{A}, Z\rangle, \tag{10}$$

soit telle que :

$$\bar{Z} = \bar{x}\bar{x}^T. \tag{11}$$

Cette condition est vérifiée pour de nombreux exemples. En particulier, Guédon et Vershynin prouvent dans leur article que cette condition est vérifiée dans le cadre de la détection de deux communautés de même taille, ainsi que dans le problème plus général présenté en Section 5.

[2]Il est intéressant de remarquer que, si nous avions maximisé $\langle A, xx^T\rangle$ sur la boule Euclidienne $\mathcal{B}(0, \sqrt{n})$, alors le problème aurait été beaucoup plus simple : la solution aurait été le vecteur propre de A associé à sa plus grande valeur propre. Une telle approche est liée à une autre méthode, nommée *clustering spectral*, que nous présentons lors de nos implémentations en Section 4. Le remplacement de $\mathcal{B}(0, \sqrt{n})$ par $\{-1, 1\}^n$ conduit à exclure des solutions où la masse de x est concentrée sur un faible nombre de coordonnées - i.e. où x est localisé. Or, comme les vecteurs propres des matrices parcimonieuses ont justement tendance à être localisés [Bordenave et Guionnet, 2013], nous verrons en Section 5 que le clustering spectral classique est souvent bien moins recommandable pour l'étude de réseaux parcimonieux.

3.2 Inégalité de Grothendieck et déviations uniformes

Ensuite, nous avons besoin d'une inégalité de déviation uniforme, afin de nous assurer qu'avec une grande probabilité :

$$\max_{Z \in \mathcal{M}_{opt}} |\langle A - \bar{A}, Z \rangle| \leq \varepsilon, \tag{12}$$

pour un certain ε. Aboutir à un tel résultat nous permettrait en effet de montrer que la solution $\hat{Z}$ de (9) approxime la solution $\bar{Z}$, puisque :

$$\begin{aligned} \langle \bar{A}, \hat{Z} \rangle &\geq \langle A, \hat{Z} \rangle - \varepsilon \quad \text{en remplaçant } \bar{A} \text{ par } A \text{ en utilisant } (12) \\ &\geq \langle A, \bar{Z} \rangle - \varepsilon \quad \text{car } \bar{Z} \text{ maximise } (9) \\ &\geq \langle \bar{A}, \bar{Z} \rangle - 2\varepsilon \quad \text{en remplaçant } \bar{A} \text{ par } A \text{ en utilisant } (12) \end{aligned}$$

Un tel résultat signifierait que $\hat{Z}$ maximise "quasiment" la fonction-objectif $\langle \bar{A}, Z \rangle$ de (10). Sa validité, prouvée par Guédon et Vershynin, repose sur l'utilisation de l'inégalité de Grothendieck, que nous présentons dans la sous-partie suivante.

3.2.1 Inégalité de Grothendieck

L'inégalité de Grothendieck [Grothendieck, 1953 ; Lindenstrauss et Pelczynski, 1968], dont les domaines d'applications sont très larges [Pisier, 2012], s'énonce ainsi.

Théorème 2. *(Inégalité de Grothendieck)*
Soit une matrice $n \times n$ composée de nombre réels $B = (b_{ij})$. Supposons que $|\sum_{i,j} b_{i,j} s_i t_j| \leq 1$ pour tout $s_i, t_i \in \{-1, 1\}$. Alors:

$$|\sum_{i,j} b_{i,j} \langle X_i, Y_j \rangle| \leq K_G,$$

pour tout vecteurs $X_i, Y_i \in \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$ la boule unité pour la norme Euclidienne. K_G est une constante absolue et inconnue, nommée constante de Grothendieck.

Considérons de nouveau l'ensemble $\mathcal{M}_G^+$ définit précédemment. Guédon et Vershynin, dans leur article, montre qu'une conséquence de l'inégalité de Grothendieck est que :

$$\max_{Z \in \mathcal{M}_G^+} |\langle B, Z \rangle| \leq K_G \|B\|_{\infty \to 1}, \tag{13}$$

où la notation :

$$\|B\|_{\infty \to 1} := \max_{\|s\|_\infty \leq 1} \|Bs\|_1$$

est équivalente à la *cut norm*, étudiée notamment au sein de problèmes algorithmiques [Alon et Naor, 2006].

Par ailleurs, notons :

$$\hat{Z} := \arg \max_{Z \in \mathcal{M}_{opt}} \langle B, Z \rangle \quad \text{et} \quad Z_{ref} := \arg \max_{Z \in \mathcal{M}_{opt}} \langle \bar{B}, Z \rangle.$$

Où $\bar{B}$ est l'espérance de B. Dans le cadre de la détection de deux communautés, nous avons $B = A - \lambda E_n$ et $\mathcal{M}_{opt} = \mathcal{M}_G^+$, mais nous gardons ici des notations - et donc un résultat - plus général. Alors, Guédon et Vershynin présentent le lemme suivant :

Lemme 1. *Nous avons :*

$$\langle \bar{B}, Z_{ref} \rangle - 2K_G \|B - \bar{B}\|_{\infty \to 1} \leq \langle \bar{B}, \hat{Z} \rangle \leq \langle \bar{B}, Z_{ref} \rangle. \tag{14}$$

L'inégalité de droite découle de la définition de Z_{ref}. Quant à celle de gauche, d'après l'inégalité de Grothendieck:

$$\forall Z \in \mathcal{M}_{opt}, |\langle B - \bar{B}, Z \rangle| \leq K_G \|B - \bar{B}\|_{\infty \to 1} := \varepsilon.$$

En remplaçant $\bar{B}$ par B via ce résultat, puis $\hat{Z}$ par Z_{ref} à partir de sa définition, et en remplaçant de nouveau B par $\bar{B}$, nous obtenons :

$$\langle \bar{B}, \hat{Z} \rangle \geq \langle B, \hat{Z} \rangle - \varepsilon \geq \langle B, Z_{ref} \rangle - \varepsilon \geq \langle \bar{B}, Z_{ref} \rangle - 2\varepsilon.$$

Ce qui prouve le lemme.

3.2.2 *Cut Norm* et déviations

Ce Lemme 1 nous permet de nous approcher du résultat que nous souhaitions en début de section 3.2. Le résultat obtenu s'interprète en effet comme une preuve que $\hat{Z}$ maximise "quasiment" la fonction-objectif de (10). Afin de pouvoir l'utiliser en pratique, il nous reste toutefois à expliquer de manière plus rigoureuse ce que nous mettons derrière ce terme "quasiment". Autrement dit, il nous faut montrer la possibilité de borner le terme $\|B - \bar{B}\|_{\infty\to 1}$ du Lemme 1, avec une forte probabilité. Cette démarche nous conduit au Lemme 2, présenté ci-dessous.

Lemme 2. *(Déviations de la norme $l_\infty \to l_1$)*
Soit $A = (A_{i,j}) \in \mathbb{R}^{n\times n}$ une matrice symétrique dont les entrées de la diagonale sont égales à 1, et dont les entrées au dessus de cette-ci sont des variables aléatoires indépendantes vérifiant $0 \leq A_{i,j} \leq 1$. De plus, supposons que :

$$\bar{p} := \frac{2}{n(n-1)} \sum_{i<j} \mathit{Var}(A_{i,j}) \geq \frac{9}{n}. \tag{15}$$

Alors, avec une probabilité d'au moins $1 - e^3 5^{-n}$, nous avons:

$$\|A - \mathbb{E}A\|_{\infty\to 1} \leq 3\bar{p}^{1/2} n^{3/2}. \tag{16}$$

Nous admettons ici ce Lemme 2, prouvé dans l'article de Guédon et Vershynin, et dont la démonstration repose sur l'utilisation de l'inégalité de Bernstein suivie d'une borne d'union.

Avant de passer à la section suivante, interprétons rapidement ce résultat via un exemple issu de l'article. Si nous considérons un graphe aléatoire parcimonieux généré d'un modèle de Erdős-Rényi $G(n,p)$ avec $p = a/n$ $(a \geq 1)$, alors $\bar{p} = p(1-p) \leq p = a/n$. D'après le Lemme 2, $\|A - \mathbb{E}A\|_{\infty\to 1} \leq 3a^{1/2}n$, à comparer avec $\|\mathbb{E}A\|_{\infty\to 1} = (1 + p(n-1))n \geq an$. Ce qui nous mène à :

$$\|A - \mathbb{E}A\|_{\infty\to 1} \leq 3a^{-1/2}\|\mathbb{E}A\|_{\infty\to 1}.$$

L'inégalité est intéressante lorsque a est une constante relativement grande. Mais, puisque $a = pn$ correspond à l'espérance des degrés des noeuds du graphe, nous obtenons que cette inégalité est applicable aux graphes dont l'espérance des degrés des noeuds est bornée. Ce qui est bien évidemment un résultat intéressant, dans le cadre de notre étude de réseaux parcimonieux.

3.3 Borne sur l'erreur de fonction-objectif

Le dernier élément nécessaire avant de passer à la preuve du Théorème 1 et de son Corollaire, dès la partie suivante, est un résultat sur l'erreur de fonction-objectif. Autrement dit, nous souhaiterions montrer que la matrice $\hat{Z}$, maximisant la fonction objectif utilisable en pratique, est relativement proche de la matrice Z_{ref}, définie plus haut comme la matrice maximisant la fonction objectif de référence, celle reliée à l'espérance de notre matrice aléatoire.

Le Lemme 3, que nous admettons ici mais qui est démontré dans l'article de Guédon et Vershynin, nous conduit à un tel résultat. Sa démonstration s'appuie directement sur le Lemme 2, présenté ci-dessus. Nous le présentons ici dans le cadre du problème étudié en Section 2 - c'est-à-dire la détection de deux communautés de même taille dans un graphe - et nous avons donc $B = A - \lambda E_n$ et $\mathcal{M}_{opt} = \mathcal{M}_G^+$. Toutefois, notons qu'un résultat dans le même esprit que celui-ci mais pour le problème de détection de k communautés de tailles arbitraires, présenté en Section 5, est également proposé par Guédon et Vershynin.

Lemme 3. *($\hat{Z}$ et Z_{ref} sont proches - Dans le cadre de la Section 2)*
Supposons que $\bar{p}$ satisfait la condition (15) du Lemme 2. Alors, avec une probabilité d'au moins $1 - e^3 5^{-n}$, nous avons :

$$\|\hat{Z} - Z_{ref}\|_2^2 \leq \frac{116\bar{p}^{1/2}n^{3/2}}{p-q}. \tag{17}$$

3.4 Preuve du Théorème 1 et de son Corollaire

Nous reportons désormais la preuve du Théorème 1 de notre Section 2, puis celle du Corollaire 1, telles que présentées par Guédon et Vershynin dans leur article. Nous avons déjà éclairci la démarche mathématique développée par les auteurs au cours de cette Section 3. La conclusion du Théorème 1 découle du Lemme 3.

Vérifions tout d'abord que la condition (15) du Lemme 2 est vérifiée. Dans le cadre du problème de détection de deux communautés, nous avons :

$$\bar{p} = \frac{2}{n(n-1)} \sum_{i<j} \mathrm{Var}(A_{i,j}) = \frac{p(1-p)(n-2)}{2(n-1)} + \frac{q(1-q)n}{2(n-1)}. \tag{18}$$

Puisque $p(1-p) \leq 1/4$, nous déduisons :

$$\bar{p} \geq \frac{1}{2}\max\{p(1-p), q(1-q)\} - \frac{1}{8(n-1)} > \frac{9}{n},$$

cette dernière inégalité étant obtenue à partir des hypothèses du Théorème 1. La condition (15) étant donc vérifiée, nous pouvons appliquer le Lemme 3 :

$$\|\hat{Z} - Z_{ref}\|_2^2 \leq \frac{116\bar{p}^{1/2}n^{3/2}}{p-q},$$

avec une probabilité supérieure ou égale à $1 - e^3 5^{-n}$. Par ailleurs, nous remarquons à partir de (18) que $\bar{p} \leq \frac{p+q}{2}$. En substituant dans la dernière inégalité ci-dessus, et en notant $p = a/n$ et $q = b/n$:

$$\|\hat{Z} - Z_{ref}\|_2^2 \leq \frac{116\sqrt{(a+b)/2}}{a-b}n^2. \tag{19}$$

En réarrangeant les termes, nous concluons que cette expression est bornée par εn^2 si :

$$(a-b)^2 \geq 7 \times 10^3 \varepsilon^{-2}(a+b) \tag{20}$$

Mais cette inégalité découle de (5). Terminons par rappeler que, d'après la Section 3.1. et en notant une nouvelle fois $\bar{x}$ le vecteur de communauté, nous avons :

$$Z_{ref} = \bar{x}\bar{x}^T.$$

Le Théorème 1 est donc prouvé.

En ce qui concerne le Corollaire 1, la démonstration s'appuie sur le Théorème de Davis-Kahan [Davis et Kahan, 1970] sur la stabilité des vecteurs propres à des perturbations matricielles. La plus grande valeur propre de $\bar{x}\bar{x}^T$ est n, les autres sont 0 (écart de n, donc). De plus, $\hat{Z} = \hat{Z} - \bar{x}\bar{x}^T + \bar{x}\bar{x}^T$ et $\|\hat{Z} - \bar{x}\bar{x}^T\|_2 \leq \sqrt{\varepsilon}n$. D'après le Théorème de Davis-Kahan :

$$\|\hat{v} - \bar{v}\|_2 = 2|\sin(\theta/2)| \leq C\sqrt{\varepsilon},$$

où $\hat{v}$ et $\bar{v}$ représentent les vecteurs propres de norme 1 associés aux plus grandes valeurs propres de $\hat{Z}$ et de $\bar{x}\bar{x}^T$, and où $\theta \in [0, \pi/2]$ représente l'angle entre ces deux vecteurs. Par définition, on a :

$$\hat{x} = \sqrt{n}\hat{v} \quad \text{et} \quad \bar{x} = \sqrt{n}\bar{v},$$

ce qui conclut la preuve.

4 Applications

Dans cette section, nous présentons quelques applications concrètes de la méthode présentée auparavant. Nous avons utilisé le logiciel R pour les implémentations : l'ensemble de notre code est fourni avec ce rapport[3]. Nous appliquons la méthode de l'article de Guédon et Vershynin à plusieurs graphes générés à partir de *Stochastic Block Models* de paramètres différents. Nous proposons aussi un parallèle entre cette méthode et la détection de communautés par Clustering Spectral, que nous présentons brièvement dans la partie suivante.

4.1 Clustering Spectral : quelques rappels

Le Clustering Spectral [Shi et Malik, 2000; Ng *et al.*, 2001] est une méthode populaire dans la communauté du Machine Learning pour réaliser des clustering de noeuds au sein de graphes. En effet, cette méthode a prouvé sa capacité à reconstruire des structures complexes dans des problèmes variés, permettant souvent d'obtenir des résultats bien plus précis que d'autres algorithmes plus classiques, comme le k-means lorsque les noeuds sont associés à des features [Ng *et al.*, 2001; Luxburg, 2007]. Ce n'est pas surprenant dans la mesure où le Clustering Spectral prend en compte la structure du graphe et la connectivité des noeuds, alors que le k-means cherche uniquement à reconstruire des communautés "compactes".

Supposons que nous avons à notre disposition la matrice d'adjacence $A = (A_{ij})_{i,j=1,...,n}$ associée au graphe[4]. La matrice des degrés D est définie comme une matrice diagonale avec $D_{ii} = \sum_{j=1}^{n} A_{ij}$ pour $i = 1, ..., n$. De A et D, nous définissons le Laplacien - non-normalisé - du graphe :

$$L := D - A.$$

Une revue de ses propriétés est proposée dans [Mohar, 1997] : la matrice L est notamment symétrique, semi-définie positive et, si le graphe est connecté, sa plus petite valeur propre est 0, associée au vecteur propre constant $\mathbf{1}_n$.

[3] Il est aussi disponible sur GitHub : `https://github.com/GuillaumeSalha/SparseNetworks`

[4] Il y a deux possibilités. Soit cette matrice est obtenue directement à partir d'un graphe déjà construit. Soit elle est déduite de données brutes, par exemple en reconstruisant un graphe via la méthode des k plus proches voisins sur les features. [Luxburg,2007] propose une revue complète, ce problème n'est pas détaillé davantage ici.

Nous présentons ici l'algorithme de Clustering Spectral, tel qu'expliqué par [Luxburg, 2007] :

- la première étape consiste à calculer le Laplacien L du graphe ;
- puis, nous calculons les k vecteurs propres de L associés aux k plus petites valeurs propres, où k désigne le nombre de communautés à reconstruire ;
- si $U \in \mathbb{R}^{n \times k}$ est la matrice contenant les vecteurs propres en colonne, soit $y_i \in \mathbb{R}^k$ le vecteur correspondant à la i-ème ligne de U pour $i = 1, ..., n$. Nous classons les $(y_i)_{i=1,...,n}$ en k clusters $C_1, ..., C_k$ via l'algorithme du k-means ;
- nous obtenons finalement les communautés $A_1, ..., A_k$ avec $A_i = \{j | y_j \in C_i\}$.

L'objectif de cette partie est simplement de donner quelques rappels sur la méthode. Nous n'entrerons donc pas dans les nombreux détails techniques à son sujet, notamment la pertinence de parfois remplacer L par des versions "normalisées", ou encore le lien avec les solutions des problèmes relaxés de partition de graphes nommés RatioCut et NCut. Nous renvoyons le lecteur intéressé vers [Luxburg, 2007] ou [Salha et Appert, 2016] pour davantage de détails techniques.

L'essentiel à retenir est que le Clustering Spectral est une méthode directement comparable à celle développée dans l'article de Guédon et Vershynin. Nous les comparons ici, à la fois pour des graphes relativement denses et pour des graphes plus parcimonieux, générés à partir de *Stochastic Block Models*. L'intuition est que, le Clustering Spectral étant focalisé sur l'étude de la connectivité des noeuds du graphe, celui-ci devient moins efficace lorsque le graphe est parcimonieux. Au contraire, la méthode de Guédon et Vershynin, comme nous l'avons vu, propose des résultats théoriques adaptés à l'étude de tels graphes.

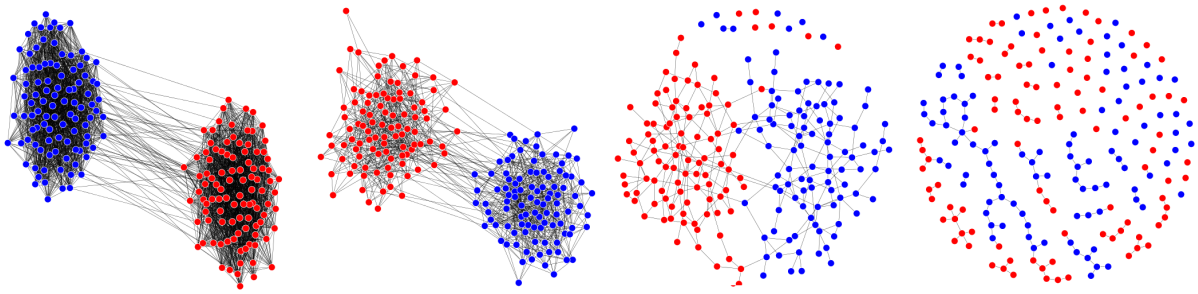

Figure 2: Plusieurs exemples de graphes, de relativement dense à très parcimonieux : $G(200, 0.3, 0.05)$, $G(200, 0.1, 0.005)$, $G(200, 0.03, 0.001)$ et $G(200, 0.01, 0.001)$

4.2 Comparaisons Monte Carlo

La démarche adoptée pour nos implémentations est la suivante:

- nous fixons à l'avance des choix de paramètres n, p et q pour le *Stochastic Block Model* $G(n, p, q)$, vérifiant les conditions imposées par le Théorème 1 ;
- nous générons un graphe aléatoire à partir du modèle $G(n, p, q)$, et nous essayons de reconstruire les deux communautés, à partir des deux méthodes :
  - la méthode de l'article de Guédon et Vershynin ;
  - à titre de comparaison, le Clustering Spectral ;
- nous vérifions quelle méthode renvoie la proportion de noeuds bien classés la plus élevée, autrement dit nous vérifions quelle méthode est la plus précise ;
- étant donné que la génération du graphe à partir de $G(n, p, q)$ est *aléatoire*, les résultats obtenus le sont aussi, et les différences de précision obtenues ne sont pas nécessairement significatives. Par conséquent, nous répétons les trois étapes précédentes N fois, et nous reportons la précision moyenne parmi les N simulations ainsi que l'écart-type correspondant.

D'un point de vue statistique, cette démarche consiste à réaliser des *simulations de type Monte Carlo* sur les graphes aléatoires d'un modèle $G(n, p, q)$ donné. Dans les exemples ci-dessous, nous prenons $N = 100$ ce qui signifie que, pour chaque modèle, nous calculons la précision moyenne et l'écart-type à partir de 100 graphes différents générés d'un même modèle.

La Table 1 ci-dessous reporte les résultats obtenus pour différentes combinaisons possibles de n, p, et q. Nous avons débuté par des graphes relativement denses. Puis, nous avons diminué progressivement le nombre - espéré - de

connexions entre les noeuds, pour obtenir des graphes de plus en plus parcimonieux. Toutefois, nous avons fait en sorte de toujours être en conformité avec les hypothèses du Théorème sur les paramètres (sauf pour les derniers), ce qui est facilement vérifiable. Enfin, nous reportons aussi la précision minimale théorique de la méthode de Guédon et Vershynin, c'est-à-dire le $(1-\varepsilon)\%$ déduit du Corollaire 1[5].

Paramètres des Stochastic Block Models			**Clustering Spectral**		**Méthode de l'article**		
			Précision moyenne	*Écart-Type parmi les*	*Précision moyenne*	*Écart-Type parmi les*	*Précision théorique*
n	p	q	*en %*	*100 graphes*	*en %*	*100 graphes*	*minimale (en %)*
200	0.3	0.05	**100.00**	-	**100.00**	-	83.18
1000	0.3	0.05	**100.00**	-	**100.00**	-	93.04
200	0.1	0.01	99.12	0.052	**99.82**	0.003	76.67
1000	0.1	0.01	**100.00**	-	**100.00**	-	89.57
200	0.05	0.01	60.65	0.126	**90.66**	0.031	56.16
1000	0.05	0.01	99.62	0.016	**99.96**	< 0.001	80.40
200	0.03	0.001	51.82	0.045	**92.33**	0.038	56.87
1000	0.03	0.001	**99.99**	< 0.001	**99.99**	< 0.001	80.71
200	0.005	0.0005	54.84	0.047	**60.04**	0.062	x
1000	0.005	0.0005	50.94	0.007	**69.45**	0.061	x

Table 1: Comparaison des deux méthodes - 100 simulations Monte Carlo

Nous vérifions aisément que, pour des graphes relativement denses tels que les deux premiers du tableau, les deux méthodes conduisent à des résultats comparables. Nous avons obtenu la même conclusion pour d'autres couples de paramètres (p, q). Par ailleurs, le résultat est souvent parfait ou très bon même pour des petits graphes (seule la valeur théorique minimale diminue pour la méthode de l'article). Les deux méthodes semblent donc adaptées au problème de détection de communautés au sein de tels graphes. En particulier, nous notons que la méthode de l'article obtient des résultats aussi bons que ceux du Spectral Clustering, une des méthodes "state-of-the-art" pour l'étude de graphes relativement denses.

En revanche, comme annoncé, le Clustering Spectral semble être de moins en moins précis lorsque le graphe devient plus parcimonieux. Ce résultat n'est pas surprenant dans la mesure où il devient plus difficile de tirer de l'information de la connectivité des noeuds. De plus, une justification plus mathématique serait que, comme nous l'avons déjà noté en Section 3, les vecteurs propres des matrices parcimonieuses ont tendance à être localisés [Bordenave et Guionnet, 2013]. Or, le Clustering Spectral fait notamment appel à un k-means sur les vecteurs propres d'une matrice L parcimonieuse. Sans surprise, un tel k-means peut conduire à des résultats décevants.

La méthode de l'article de Guédon et Vershynin, quant à elle, semble effectivement plus recommandable lorsque le graphe est parcimonieux. Le Théorème 1 et son Corollaire, présentés en Section 2, énonçaient théoriquement des résultats applicables à des graphes dont les degrés espérés des noeuds sont constants - ou augmentent à une très faible vitesse - et nous vérifions cela en pratique. Bien évidemment, le problème de détection de communautés est plus difficile pour des graphes parcimonieux, et par conséquent les précisions obtenues sont moins bonnes que pour des graphes denses. Toutefois, ces précisions restent supérieures à celles du Clustering Spectral dans une majorité de cas, et notamment dans ceux reportés dans la Table 1.

Pour finir, notons que nous aurions souhaité baisser encore davantage les valeurs de p et de q mais que cela aurait supposé, en contrepartie, d'augmenter la taille n du graphe pour satisfaire les hypothèses[6] du Théorème 1. Pour des raisons de temps de calcul, nous nous sommes limités dans ce projet à des graphes de tailles moyennes. Nous illustrons toutefois dans les deux derniers graphes - qui auraient eu besoin de davantage de noeuds pour vérifier les hypothèses du théorème - que la méthode de l'article semble retourner des prédictions qui restent intéressantes. Prolonger notre analyse en plus grande dimension serait une extension possible de notre travail.

5 Détection de k communautés de tailles arbitraires

Comme annoncé en introduction, la démarche proposée par les auteurs est assez générale, et ne se limite pas à la détection de deux communautés de même taille dans des graphes. D'autres problèmes de détection de communautés, et même d'autres problèmes de nature différente pouvant être formulés sous la forme d'un problème optimisation SDP tels que ceux étudiés jusqu'ici, peuvent être considérés. Au cours de cette section, nous proposons une extension de l'analyse réalisée dans les sections précédentes, en étudiant le problème de la détection de k communautés de tailles arbitraires au sein de graphes, denses ou parcimonieux. Le modèle considéré est une extension du *Stochastic Block Model* introduit en Section 2.

[5]Pour être rigoureux, la précision minimale théorique, en pourcentage, serait $\max\{1-\varepsilon, 0\}$, mais $\varepsilon < 1$ dans toutes nos applications. Par ailleurs, les résultats de Guédon et Vershynin ne sont valables qu'avec une probabilité d'au moins $1 - e^3 5^{-n}$, mais nous ne nous attardons pas sur celle-ci car, au sein de nos implémentations, elle est extrêmement proche de 1.

[6]Et notamment l'hypothèse $\max\{p(1-p), q(1-q)\} \geq \frac{20}{n}$.

5.1 General Stochastic Block Model

Afin de définir une telle extension, nommée *General Stochastic Block Model*, nous considérons un graphe composé de n noeuds, répartis en k communautés $\mathcal{C}_1, ..., \mathcal{C}_k$ de taille arbitraire. En particulier, la taille d'une communauté peut être d'un seul noeud, afin de prendre en compte la présence d'éventuels outliers. Ensuite, pour chaque paire de noeud, nous plaçons une arête entre les noeuds avec une probabilité p_{ij}, qui vérifie $p_{ij} \geq p$ si les noeuds appartiennent à la même communauté, et $p_{ij} \leq q$ s'ils appartiennent à des communautés différentes. Nous avons $p, q \in [0, 1]$ avec $p \geq q$. Comme au sein de la Section 2, nous incluons les boucles au modèle. La Figure 3 ci-contre représente un graphe de 200 noeuds, généré à partir d'un *General Stochastic Block Model* avec $k = 4$ communautés de même taille. Les noeuds sont reliés entre eux avec la probabilité 0.3 s'ils appartiennent à la même communauté, et avec la probabilité 0.05 autrement.

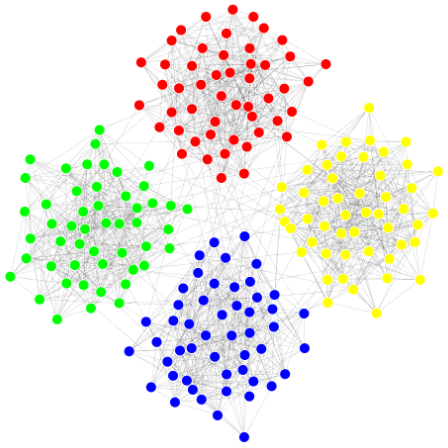

Figure 3: General Stochastic Block Model, avec $k = 4$

Le problème de détection de communauté consiste, comme auparavant, à reconstruire le plus précisément possible les communautés $\mathcal{C}_1, ..., \mathcal{C}_k$, à partir à partir d'une réalisation aléatoire issue du *General Stochastic Block Model*. D'un point de vue plus formalisé, le problème est équivalent à celui de la reconstruction de la matrice des communautés $\bar{Z} \in \{0, 1\}^{n \times n}$, définie ainsi :

$$\bar{Z}_{i,j} := \mathbf{1}_{\{\text{les noeuds } i \text{ et } j \text{ appartiennent à la même communauté}\}}$$

5.2 Problème d'optimisation et Détection de Communautés

Guédon et Vershynin proposent d'estimer $\bar{Z}$ par la solution du problème d'optimisation SDP suivant :

$$\begin{aligned} &\max_{Z} \ \langle A, Z \rangle \\ &\text{s.c. } Z \succeq 0, \ Z \geq 0, \ \text{diag}(Z) \preceq I_n, \ \sum_{i,j=1}^{n} Z_{i,j} = \lambda \end{aligned} \tag{21}$$

Les notations ont les mêmes significations qu'en Section 2. Nous ajoutons la contrainte $Z \geq 0$ qui précise que toutes les entrées de la matrice Z doivent être positives, et la constante λ est désormais fixée comme le nombre d'éléments non nuls dans la matrice des communautés :

$$\lambda = \sum_{i,j=1}^{n} \bar{Z}_{i,j} = \sum_{i=1}^{k} |\mathcal{C}_i|^2.$$

L'ensemble de l'approche précédente, avec notamment la considération d'un problème référent, l'utilisation de l'inégalité de Grothendieck et celle de la *cut norm*, peut être reconduite ici. En adoptant une méthodologie similaire à celle présentée au cours de la Section 3, et en redémontrant un par un des lemmes équivalents pour ce nouveau problème, Guédon et Vershynin parviennent dans leur article à prouver le Théorème 3, présenté ci-dessous. Celui-ci met en avant que, sous certaines conditions, la résolution du problème (21) conduit à une estimation correcte de la matrice des communautés, à un certain nombre d'erreurs près, ce nombre étant fixé et connu. Une nouvelle fois, ce théorème peut s'appliquer aux graphes parcimonieux, ce qui est un résultat important.

Théorème 3. *(Détection de communautés dans le General Stochastic Block Model)*
Soit $\varepsilon \in (0, 1)$. Soit A la matrice d'adjacence du graphe aléatoire généré du General Stochastic Block Model présenté plus haut. Notons $\bar{p}$ l'espérance de la variance des arêtes :

$$\bar{p} = \frac{2}{n(n-1)} \sum_{i<j} p_{ij}(1 - p_{ij}).$$

Supposons que $p = a/n > q = b/n$, que $\bar{p} = g/n$ avec $g \geq 9$, et que :

$$(a - b)^2 \geq 484 \varepsilon^{-2} g.$$

Soit $\hat{Z}$ une solution du problème d'optimisation SDP (21). Alors, avec une probabilité d'au moins $1 - e^3 5^{-n}$, nous avons :

$$\|\hat{Z} - \bar{Z}\|_2^2 \leq \|\hat{Z} - \bar{Z}\|_1 \leq \varepsilon n^2. \tag{22}$$

La notation $\|.\|_2$ désigne une nouvelle fois la norme de Frobenius, et la notation $\|.\|_1$ désigne la norme l_1 des matrices considérées comme des vecteurs, i.e. $\|A\|_1 := \sum_{i,j} |A_{ij}|$.

5.3 Discussion du résultat

Au sein de ce projet, nous admettons ce Théorème 3. La démonstration n'est pas beaucoup plus technique que celle du Théorème 1, mais elle requiert un nombre significatif de résultats intermédiaires qu'il serait trop long de détailler pour le format consacré au projet. Nous invitons donc le lecteur intéressé par les détails mathématiques permet d'aboutir à ce Théorème 3 à se reporter à l'article de Guédon et Vershynin. Ici, nous proposons malgré tout une interprétation de quelques aspects intéressants liés à ce résultat.

Tout d'abord, notons que la puissance de ce Théorème ne dépend pas de la taille ni du nombre de communautés. Ce résultat pourrait à première vue sembler surprenant. Il s'explique en réalité de manière naturelle : les communautés les plus petites - et notamment, les outliers - vont être absorbées dans le terme d'erreur. Elle ne pourront donc pas être retrouvées.

Le problème des réseaux parcimonieux reste quant à lui le même. Comme avant, si les degrés des noeuds augmentent à une vitesse inférieure à $log\ n$, alors avec une probabilité élevée certains noeuds du graphe seront isolés, et il sera impossible de les classifier de manière pertinente. La proportion de ces noeuds tendant vers zéro avec n, nous pouvons une nouvelle fois espérer obtenir un résultat correct pour une majorité des noeuds du graphe.

Par ailleurs, il peut être intéressant de visualiser le problème de reconstruction de la matrice $\bar{Z}$ comme un problème de reconstruction d'un graphe dit *graphe des communautés*, dont la matrice d'adjacence serait précisément la matrice $\bar{Z}$. Au sein d'un tel graphe, tous les noeuds d'une même communauté sont reliés entre eux. À l'inverse, les noeuds de communautés différentes ne sont jamais reliés. Le problème d'optimisation (21) prend compte en entrée un graphe aléatoire généré du *General Stochastic Block Model*, et renvoie un autre graphe rendu plus dense au sein des communautés, et plus parcimonieux entre les communautés.

Enfin, notons que le Théorème 3 suppose, pour le choix de λ, que la taille réelle des communautés soit connue. Cette hypothèse peut sembler restrictive pour un certain nombre de cas pratiques. Un choix "arbitraire" du paramètre $\lambda > 0$ a la conséquence suivante, qui est mise en avant dans la preuve du théorème :

- lorsque $\lambda < \lambda_0 := \sum_{i=1}^{k} |\mathcal{C}_i|^2$, la matrice $\hat{Z}$ estime un sous-graphe du graphe complet - i.e. du graphe des communautés, pour reprendre la remarque précédente. Ce sous-graphe correspond au graphe des communautés, auquel ont été retranchés au plus $\lambda_0 - \lambda$ arêtes.
- lorsque $\lambda > \lambda_0$, nous estimons au contraire un graphe correspondant au graphe des communautés auquel ont été ajoutés au plus $\lambda_0 - \lambda$ arêtes supplémentaires.

6 Conclusion

Au cours de ce projet, nous avons réalisé un tour d'horizon des principaux résultats de l'article *Community Detection in Sparse Networks via Grothendieck's Inequality* de Guédon et Vershynin (2015). Les auteurs ont présenté une méthode flexible pour prouver la précision de certains problèmes d'optimisation SDP avec une application au problème de détection de communautés. La méthode permet notamment d'obtenir des résultats intéressant au sein de réseaux parcimonieux, ce qui est un résultat important vis à vis des autres méthodes développées dans la littérature, majoritairement concentrées sur l'étude de réseaux relativement denses.

Notre objectif, au sein de ce court rapport, n'était pas être exhaustif sur l'ensemble des résultats mathématiques de l'article, mais plutôt de fournir une vue d'ensemble claire et fidèle de la démarche développée par Guédon et Vershynin, et des principaux enjeux derrière chaque lemme ou théorème. Nous avons également tenu à proposer une implémentation concrète de la méthode, qui est absente de l'article, en considérant en parallèle une autre approche pour la détection de communautés, le Clustering Spectral. Nos expériences illustrent la capacité de la méthode de Guédon et Vershynin à détecter la structure de communauté même au sein de réseaux parcimonieux, avec un taux d'erreur maximal connu.

Enfin, notons que plusieurs travaux portant sur l'étude des graphes parcimonieux ont été publiés plus récemment que l'article de Guédon et Vershynin, par exemple [Bordenave *et al.*, 2015] et [Chin *et al.*, 2015]. Une mise en comparaison de l'ensemble de ces méthodes pourrait constituer une extension intéressante de notre travail au sein de ce projet.

Remerciements

Ce travail est notre projet final pour le cours *Introduction mathématique au Compressed Sensing* de M. Guillaume Lecué, à l'ENSAE ParisTech (mars 2016). Nous tenons à le remercier pour l'ensemble des enseignements de ce semestre, et pour nous avoir permis de travailler sur le thème des graphes parcimonieux, qui nous intéressait beaucoup.

Références

O. Guédon, R. Vershynin (2014), Community detection in sparse networks via Grothendieck's inequality. *arXiv preprint, ArXiv:1411.4686*.

****

N. Ailon, Y. Chen, H. Xu (2014), Breaking the small cluster barrier of graph clustering. *Journal of Machine Learning Research, arXiv:1302.4549*.

E. Airoldi, D. Blei, S. Fienberg, E. Xing (2008), Mixed membership stochastic blockmodels. *J. Machine Learning Research 9*, 1981-2014.

N. Alon (2008), Spectral techniques in graph algorithms. *LATIN'98: theor. informatics*, 206-215, Lect. Notes in Comput. Sci.

N. Alon, A. Naor (2006), Approximating the cut-norm via Grothendieck's inequality, *SIAM J. Comput. 35*, 787-803.

A. Amini, E. Levina (2014), On semidefinite relaxations of the block model. *Arxiv: 1406.5647*.

C. Bordenave, A. Guionnet (2013), Localization and delocalization of eigenvectors for heavy-tailed random matrices. *Probab. Theory Related Fields 157*, 885-953.

C. Bordenave, M. Lelarge, L. Massoulié (2015), Non-backtracking spectrum of random graphs: community detection and non-regular Ramanujan graphs, *ArXiv: 1501.06087*.

T. N. Bui, S. Chaudhuri, F. T. Leighton, M. Sipser (1987), Graph bisection algorithms with good average case behavior. *Combinatorica*, 7, 171-191.

T. Cai, X. Li (2014), Robust and computationally feasible community detection in the presence of arbitrary outlier vertices. *ArXiv: 1404.6000*.

A. Celisse, J.-J. Daudin, L. Pierre (2012), Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics 6*, 1847-1899.

Y. Chen, A. Jalali, S. Sanghavi, H. Xu (2014), Clustering partially observed graphs via convex optimization. *Journal of Machine Learning Research*, Arxiv: 1104.4803.

P. Chin, A. Rao, V. Vu (2015), Stochastic block model and community detection in sparse graphs: a spectral algorithm with optimal rate of recovery. *ArXiv: 1501.05021*.

C. Davis, W. M. Kahan (1970), The rotation of eigenvectors by a perturbation. *III. SIAM J. Numer. Anal.*, 7, 1-46.

M. E. Dyer, A. M. Frieze (1989), The solution of some random NP-hard problems in polynomial expected time. *Journal of Algorithms*, 10, 451-489.

P. Erdős, A. Rényi (1959). On random graphs. *Publicationes Mathematicae Debrecen*, 6, 290-297.

M. Goemans, D. Williamson (1959), Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming, *J. Assoc. Comput. Mach. 42*, 1115-1145.

A. Grothendieck (1953), Résumé de la théorie métrique des produits tensoriels et topologiques, *Bol. Soc. Mat. Sao Paulo 8*, 1-79.

P. W. Holland, K. B. Laskey, S. Leinhardt (1983), Stochastic blockmodels: first steps. *Social Networks*, 5, 109-137.

A. Jalali, Y. Chen, S. Sanghavi, H. Xu (2011), Clustering partially observed graphs via convex optimization. *ICML*, ArXiv: 1104.4803.

M. Lelarge, L. Massoulié, J. Xu (2013), Reconstruction in the labeled stochastic block model. *Proceedings of the Information Theory Workshop*, ArXiv: 1502.03365.

U., Luxburg, (2007), A Tutorial on Spectral Clustering. *Statistics and Computing*, 17(4):395-416.

E. Mossel, J. Neeman, A. Sly (2014), Stochastic Block Models and Reconstruction. *Probability Theory and Related Fields*.

Y. Nesterov (1998), Semidefinite relaxation and nonconvex quadratic optimization, *Optim. Methods Softw. 9*, 141-160.

M. Newman (2006), Modularity and community structure in networks. *Proc. Natl. Acad. Sci.*, USA 103, 8577-8582.

A., Ng, M. I., Jordan, Y. Weiss, (2001), On spectral clustering: Analysis and an algorithm. *Advances in neural information processing system (NIPS)*, 849-856.

G. Pisier (2012), Grothendieck's theorem, past and present. *Bull. Amer. Math. Soc. (N.S.)*, 49, no. 2, 237-323.

G. Salha, G. Appert (2016), Large-Scale Spectral Clustering on Graphs. *Master Project, Pr. Valko's Graphs in Machine Learning course*, M2 M.V.A., École Normale Supérieure de Cachan.

T. Snijders, K. Nowicki (1997), Estimation and prediction for stochastic block-structures for graphs with latent block structure, *Journal of Classification*, 14, 75-100.