Mémoire Master 2 : Probabilités et Statistiques
Analyse spectrale des graphes et détection de structures de communauté dans les réseaux

Hugues-Vincent Ropert

Mai 2018

Table des matières

Première partie

Introduction

Ce mémoire a été réalisé en collaboration avec Jamal Najim au sein de LIGM (Laboratoire d'Informatique Gaspard Monge) à l'Université de Marne La Vallée. L'objectif de ce stage était d'étudier les méthodes de détection de communauté du point de vue des matrices aléatoires.

La détection de communautés est un important domaine de recherche. Depuis l'arrivé des nouvelles technologies et notamment depuis la propagation du Web, une quantité gigantesque d'information est apparue donnant ainsi naissance à une abondance de graphes et de réseaux à étudier. Il y a un nombre substantiel d'applications à l'étude de ces graphes. En effet la structure des graphes se retrouve partout :

— Ensemble des sites web ;
— Ensemble des articles scientifiques ;
— Organes du corps humain ;
— Théorèmes mathématiques ;
— Images ;
— Flux économiques.

Une propriété des réseaux que l'on souhaite naturellement dégager est leur structure de communauté. Dit autrement, nous souhaitons trouver une partition des nœuds du graphe afin d'étudier les connections ou les similarités les éléments du réseaux en questions.

Dans ce contexte, la théorie des matrices aléatoires a une application direct, le "spectral graph clustering". L'intérêt de ces méthodes de détection sont leur rapidité d'exécution. C'est domaine de recherche récent qui date des années 2000 et qui à eu un regain de vigueur suite à la publication de l'article "Graph spectra and the detectability of community structure in networks" [7] en 2012.

Dans ce mémoire, nous allons précisément étudier cet article. Il y a une erreur théorique dans ce papier et nous allons essayer de redémontrer les résultats en dégageant l'erreur commise. Ensuite nous étudierons des algorithmes spectral plus sophistiqués et performants.

Deuxième partie

Vue d'ensemble de la théorie du clustering spectral de graphes

1 La détection de communauté

1.1 Motivations

La théorie des réseaux, a pour but d'analyser les graphes correspondant à des réseaux tels que celui de l'Internet, de la politique ou de la classification en biologie. Chacun de ces graphes ont des propriétés spécifiques telles que :

- Centrality : mesure de cohésion du graphe ;
- "small world effect" : distance moyenne entre deux nœuds est proportionnelle au log du nombre de nœuds dans le graphe ;
- Clustering : mesure du regroupement des nœuds du graphe ;
- Efficiency : mesure la résistance du graphe ;
- Degree distribution ;
- Community Structure.

C'est cette dernière propriété qui va nous intéresser, à savoir la structure de communauté d'un graphe. L'idée de communauté correspond à l'intuition selon laquelle il y a des nœuds qui ont un lien étroit et forment des sous ensembles de nœuds de ce graphe. Par exemple dans le graphe des représentants politique Français, les hommes politiques d'un même partie appartiendraient à une même communauté. Cependant, dans cet exemple les choses peuvent être plus subtiles que ça, c'est là qu'intervient la détection de communauté.

1.2 Méthodes

Il y a différentes classes de méthodes pour la détection de communauté, en voici une liste non exhaustive. Pour une liste fournies des méthodes existantes voir [5].

- Hierarchical clustering
- Graph partitioning
- Partitional clustering
- Spectral clustering

1.3 Comparaison des structures de communauté

La notion de structure de communauté dans un graphe n'a pas de définition claire et consensuelle. En effet, lorsque l'on veut faire de la détection de communauté nous cherchons d'hypothétiques lien entre les nœuds d'un réseau. Cependant, ce lien ne peut pas avoir de sens a priori. Dans l'exemple du graphe du personnel politique français, si on trouve une certaine structure de communauté, quel sens peut-on donner au fait que deux personnes appartiennent à la même communauté. Pour trouver un sens aux liens extraits par les algorithmes, on peut essayer de partir du sens de l'information à partir de laquelle le graphe a été construit, mais cela reste de l'ordre de l'interprétation.

De plus, le résultat obtenu dépend entièrement du choix de la méthode utilisée. La non unicité des résultats des algorithmes de détection de communauté rajoute une variabilité substantielle à l'interprétation d'une structure de communauté. En substance, cela veut dire qu'en plus de choisir comment construire le graphe (i.e à partir de quelle information), il faut aussi choisir une métrique qui mesure la notion de proximité entre deux éléments du réseau.

Une conséquence négative de la non unicité des résultats est le fait qu'il n'existe pas de référence à partir de laquelle comparer les différentes structures de communautés obtenues. Le seul moyen de juger le véracité des résultats obtenus est de les comparer entre eux et de faire une synthèse. Il existe une abondante littérature sur la comparaison des méthodes.

On peut cependant essayer de comparer théoriquement les méthodes utilisés en analysant des graphes que l'on génère de manière aléatoire et dont on connaît les communautés. On appelle ces graphes, "graphe aléatoire" ou 'random graph' Cette manière de procéder à l'avantage de posséder une référence à partir de laquelle comparer les algorithmes. En revanche, les graphes générés ainsi ne correspondent pas forcement au graphes que l'on trouve dans la "nature". Par conséquent des algorithmes qui ont de très bonnes performances sur un certain type de graphe, ne seront pas forcement pertinents sur des graphes générées à partir de données réelles.

Une grande part de la recherche dans ce domaine est de rajouter des propriétés et des contraintes aux algorithmes de génération de graphes afin qu'ils satisfassent au mieux les propriétés des graphes que l'on observe dans la réalité (réseaux sociaux ...). On peut donc ainsi trouver le meilleur algorithme pour un ensemble de graphes satisfaisant un certain nombre de propriétés.

Pour comparer les résultat a partir d'un graphe aléatoire il faut choisir une métrique. Il existe une très grand nombre de métrique qui donnent là aussi des résultats divergeant. On ne sait pas jusqu'à présent quel algorithme est le plus fiable.

Dans le cas où l'on utilise un graphe aléatoire, il existe plusieurs moyens de comparer les partitions obtenues avec la partition de référence. Pour une liste fournies des méthodes existantes voir [5, p.77-79]. Ci-dessous une liste non exhaustive des métriques qui seront utilisées ultérieurement.

- Fraction of correctly classified vertices ;
- Fowlkes and Mallows metric ;
- Rand Index ;
- Normalized Mutal Information ;

1.4 Comparaison des performances

Une autre problématique des algorithmes de détection de communautés est la performance. En effet, beaucoup de méthodes considérées comme les plus efficaces du point de vue des communautés obtenues ont une complexité polynomiale voire exponentielle. Or beaucoup de réseaux du vrai-monde ont un nombre de nœuds d'un ordre largement supérieur au milliard. On se retrouve donc dans le trade-off classique entre la performance et l'erreur d'approximation. Dans ces situations, il faut dont trouver des algorithmes à la fois performants et consistants.

C'est dans cette perspective que se place les papiers que nous étudions dans ce mémoire. Nous nous placerons toujours dans le contexte d'un Stochastic Block Model, ou SBM, à 2 communautés de même taille et à deux probabilités p_{in} et p_{out}.

Un SBM est un graphe aléatoire dans lequel on choisi un nombre de nœuds n, le nombre de communautés q. On génère ainsi une graphe à n nœuds sans arrêtes, et ensuite pour chaque pair de nœuds on simule une variable de Bernoulli pour décider si l'arrête existe ou non. Le paramètre de la loi de Bernoulli est p_{in} si les deux nœuds appartiennent à la même communauté et p_{out} sinon.

Dans ce contexte, d'après le paragraphe [2, Numerical results], l'algorithme "belief propagation" est réputé pour être optimal sur les partitions rendues. Ainsi, pour comparer les performances des algorithmes dans ce contexte, la littérature utilise cet algorithme comme un benchmark du point de vue de la répartition obtenue. Ensuite ils comparent la performance et le temps d'exécution par rapport à celui-ci. Ci-dessous une figure provenant de [4] qui synthétise ce paragraphe.

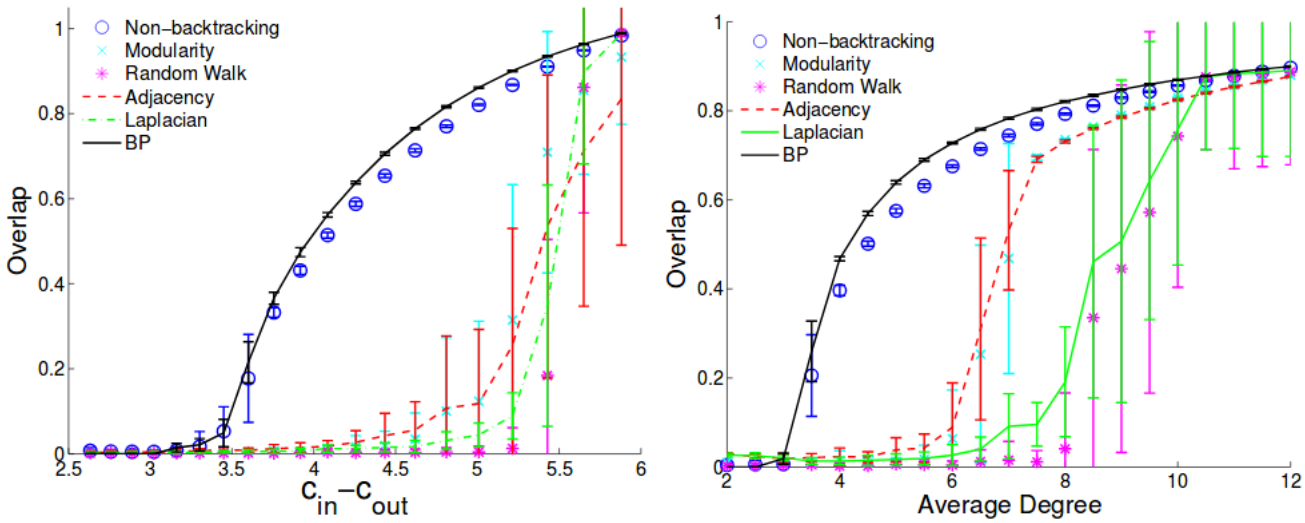


FIGURE 1 – Précision des algorithmes de détection spectral basées sur différents opérateurs linéaires par rapport l'algorithme "belief propagation". [4]

1.5 Perspectives

Les limites des modèles spectraux qui sont en développement pour la détection de communauté sont les suivantes :

1- Raffinement des modèles génératifs :

Comme dit précédemment, l'idéal, serait d'avoir un modèle génératif de graphe qui correspond au mieux aux réseaux obtenus sur des exemples concrets. Cependant, les graphes théoriques sont loin d'êtres représentatif de la réalité et il y a encore beaucoup de propriétés à découvrir sur les graphes du vrai-monde.

2- Temps de calcul sur de grands graphes :

Les graphes étant de plus en plus grands, beaucoup de recherches se tournent sur la performance algorithmique des algorithmes de détection de communautés et en particulier sur les méthodes spectrales.

3- Seuil à partir duquel la méthode spectrale n'est plus en mesure de détecter le communauté :

Certains algorithmes ont des limites quand à la détectabilité de la structure de communauté. Le but ici est de trouver des modèles qui permettent de trouver les partitions optimums et qui fonctionnent sur le plus large ensemble de graphes possibles.

D'autres perspectives de recherche existe également dans le domaine de la détection de communautés indépendamment des matrices aléatoires.

1- Détection de communautés avec chevauchement (i.e nœuds appartenant à plusieurs communautés à la fois) ;
2- Détection de communautés sur des réseaux multi-dimensionnel.

2 Spectral clustering

L'idée centrale des techniques de clustering spectral est qu'un graphe est représentable par une matrice à partir de laquelle on peut utiliser les techniques d'analyse de l'algèbre linéaire.

Un graphe G est la donnée d'un couple (V, E) tel que V est une ensemble de nœuds et E et un ensemble d'arêtes (i.e un couple (i, j) où $i, j \in V$). A partir de cette définition on peut représenter le graphe par une matrice dont les éléments correspondent à certaines données de G. Il existe tout un ensemble de matrices représentant le graphe :

- Adjencency Matrix = A ;
- Laplacian Matrix = L ;
- Modularity Matrix = M ;
- Bethe Hessian Matrix = H ;
- "α-normalized" Adjacency Matrix = D.

Par exemple, la matrice d'adjacence d'un graphe, notée A, est définie telle que $\forall i, j \in V,\ A_{ij} = \mathbb{1}_{(i,j)\in E}$. Une colonne i représente le nœud i dont chaque composante, j, est égale à 1 si il existe une arête entre i et j et 0 sinon.

Ci-dessous un graphe qui permettant d'avoir une vue d'ensemble sur l'avancement des méthodes de spectrales de détection de communautés. Cette liste n'est pas exhaustive.

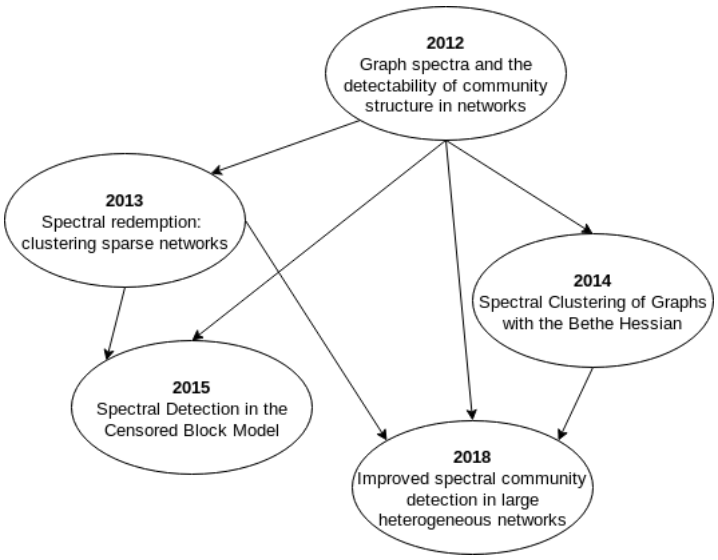


FIGURE 2 – graphe des différentes méthodes de "graph spectral clustering"

2.1 Algorithme de clustering spectral de graphe

La procédure pour associer une communauté aux nœuds de G via une méthode spectrale est la suivante :

1- Calcul des vecteurs propres, v_i, d'une matrice représentant le Graphe G ;
2- Sélection des l vecteurs propres portant l'information de la structure de communauté ;
3- Construction de la matrice $W = [v_1, \cdots, v_l] \in \mathbb{R}^{n\times l}$;
4- Projection des vecteur lignes r_j de W sur l'espace de dimension l, où chaque r_j correspond au nœud j ;
5- Catégorisation des vecteurs r_j dans une communauté via des algorithmes de clustering : *K-Means*, *Expetation-Maximization*, *Support vector machine* , etc.

Étape 1 : La matrice A est carrée, par conséquent elle peut être interprétée comme la représentation d'un endomorphisme dans un espace X de dimension n (nombre de nœuds dans G). Soit la base canonique $(e_i)_{i=1:n}$, chaque e_i correspond au nœud i. Les vecteurs propres de A sont donc des combinaisons linéaires des nœuds de G. Les nœuds concernés ont donc une certaine dépendance.

Étape 2 : Les entrées de la Matrice A sont modélisées par des variables aléatoires vérifiant certaines hypothèses (e.g centrées, moments finis, indépendantes ...). La théorie des matrices aléatoires nous dit qu'asymptotiquement la mesure spectrale de A converge vers une loi déterministe μ (loi qui dépend de la matrice étudiée). Les valeurs propres de A distribuées selon μ peuvent être interprétées comme du bruit lié aux fluctuations aléatoires des simulations. Elles nous renseigne en rien sur la structure non aléatoire des entrées de A .
Cependant, lorsque le graphe admet une structure de communauté, les entrées ne sont plus tout à fait indé-

pendantes. Dans ce cas de figure, la théorie prévoit que des valeurs propres sortent du support de la mesure spectrale μ. Ce sont ces valeurs propres qui contiennent l'information de la structure de communauté de G. Autrement dit, si la structure de communauté de G n'est pas assez "explicite", les valeurs propres censées porter l'information des communautés ne sortiront pas du support de la mesure spectrale théorique, et donc, seront interprétées comme du simple bruit. Dans cette situation, la méthode spectrale est incapable de déceler la structure de communauté de G.

Étape 4 : Les vecteurs colonnes de W correspondent à des combinaisons linéaires de nœuds. De par la construction de A, les entrées i de ces vecteurs colonnes sont la somme des arêtes entre les nœuds sous-jacent et le nœud i. Les vecteurs propres que l'on a gardés sont ceux qui portent l'information des communautés. Donc chaque vecteur ligne représente un nœud du graphe et l'espace dans lequel il se meut constitue l'information de la structure de communauté.

Les algorithmes de détection de communauté spectraux à partir des opérateurs linéaires tels que la matrice "non-backtracking" et la matrice "Bethe Hessian" sont similaires et sont décrits dans l'article [2, Spectral detection in the censored block model].

Troisième partie

Article : Graph spectra and the detectability of community structure in networks - Raj Rao Nadakuditi and M. E. J. Newman

3 Théorie

3.1 Contexte : Stochastic Block Model à 2 communautés

Le but de ce papier est d'expliquer en détail la méthode de détection de communauté via la théorie spectrale de l'article de RR. Nadakuditi et M. E. J. Newmann. C'est un article qui décrit les limites d'un modèle spectral dans un cadre spécifique à deux communautés.

Nous allons nous placer dans le contexte d'un Stochastic Block Model (SBM) à deux communautés (i.e $\mathbf{q}$=2). C'est un graphe $\mathbf{G}$ non orienté à $\mathbf{n}$ nœuds dont chaque arête entre deux nœuds suit une loi de Bernoulli de paramètre p_{in} si les nœuds appartiennent à la même communauté et p_{out} si les nœuds ne sont pas dans la même communauté.

Nous noterons $\mathbf{A}$ la matrice d'adjacence du graphe $\mathbf{G}$. Elle est symétrique de par le fait que le graphe soit non orienté. Nous supposerons d'ailleurs, que ses éléments sont rangés dans l'ordre de leur communauté. Dans notre cas avec $\mathbf{q}$=2, les $\mathbf{n}/2$ premières lignes correspondent aux noeuds de la communauté 1, et les $\mathbf{n}/2$ dernières la communauté 2. Même chose pour les colonnes, par symétrie de $\mathbf{A}$.
Cette disposition des éléments ne change pas le résultat final de l'analayse. En effet, on sait que toutes les matrices congruentes représentes la même forme bilinéaire dans des bases différentes. Or le but de la procédure qui va suivre va être d'analyser la distribution des valeurs propres de notre matrice. Par conséquent ranger les colonnes d'une matrice selon un ordre arbitraire correspond à la même forme bilinéaire et donc aux mêmes valeurs propres.

Chaque élément de la matrice $\mathbf{A}$ est simulé par une loi de Bernoulli avec :

$$A_{ij} \sim \begin{cases} B(p_{in}) & : (i,j < \frac{n}{2}) \; ou \; (i,j \geq \frac{n}{2}) \\ B(p_{out}) & : else \; where \end{cases}$$

$$A_{ij} = A_{ji}$$

3.2 Analyse spectrale de la matrice d'adjacence A

L'idée générale de l'analyse spectrale qui va suivre est de nous ramener à un régime spectral de grande matrices aléatoires connu. Dans notre cas, nous verrons que le régime associé à notre matrice d'adjacence (SBM $\mathbf{q} = 2$) est celui du théorème de *Wigner* avec perturbation de rang fini. La trame sera la suivante :

1- réécriture de la matrice $A = \langle A \rangle + X$;

2- étude de la mesure spectrale de X ;

3- étude de la mesure spectrale de B où $B = X + P_1$;
4- étude de la mesure spectrale de $A = B + P_2 = X + P_1 + P_2$.
où P_1, P_2 sont des perturbations de rang 1 et $\langle A \rangle$ correspond à la moyenne de A du SBM.

1- Équation de $\langle A \rangle$

En partant de A, on définit $\langle A \rangle$ comme la matrice dont les entrées $\langle A \rangle_{ij} = \mathbb{E}(A_{ij})$ Soit

$$\begin{aligned}
\langle A \rangle &= \begin{bmatrix} p_{in} & p_{out} \\ p_{out} & p_{in} \end{bmatrix} \otimes \mathbf{1}_{\frac{n}{2}} \mathbf{1}_{\frac{n}{2}}^T \times \frac{n}{2} \\
&= \begin{bmatrix} c_{in} & c_{out} \\ c_{out} & c_{in} \end{bmatrix} \otimes \mathbf{1}_{\frac{n}{2}} \mathbf{1}_{\frac{n}{2}}^T \times \frac{1}{2} \\
&= \left(\frac{1}{2} \times \begin{bmatrix} c_{in} + c_{out} & c_{in} + c_{out} \\ c_{in} + c_{out} & c_{in} + c_{out} \end{bmatrix} + \frac{1}{2} \times \begin{bmatrix} c_{in} - c_{out} & -c_{in} + c_{out} \\ -c_{in} + c_{out} & c_{in} - c_{out} \end{bmatrix} \right) \otimes \mathbf{1}_{\frac{n}{2}} \mathbf{1}_{\frac{n}{2}}^T \times \frac{1}{2} \\
&= \left(\frac{1}{2}(c_{in} + c_{out}) \times \mathbf{1}_2 \mathbf{1}_2^T + \frac{1}{2}(c_{in} - c_{out}) \times \mathbf{u}_2 \mathbf{u}_2^T \right) \otimes \mathbf{1}_{\frac{n}{2}} \mathbf{1}_{\frac{n}{2}}^T \\
&= \frac{1}{2}(c_{in} + c_{out}) \mathbf{1}_n \mathbf{1}_n^T + \frac{1}{2}(c_{in} - c_{out}) \mathbf{u}_n \mathbf{u}_n^T
\end{aligned} \tag{1}$$

Avec

$$\begin{aligned}
c_{in} &= n p_{int} \\
c_{out} &= n p_{out} \\
\mathbf{1}_n &= (\underbrace{1, \ldots, 1}_{n\text{-times}})/\sqrt{n} \\
\mathbf{u}_n &= (\underbrace{1, \ldots, 1}_{\frac{n}{2}\text{-times}}, \underbrace{-1, \ldots, 1}_{\frac{n}{2}\text{-times}})/\sqrt{n} \\
\langle \mathbf{1_n} | \mathbf{u_n} \rangle &= 0
\end{aligned}$$

À présent **A** peut être écrite sous la forme $A = \langle A \rangle + X$. La matrice X est interprétable comme la déviation entre la matrice d'adjacence du graphe et sa moyenne. La matrice X est par définition une matrice aléatoire symétrique à entrées indépendantes et de moyenne 0. Essayons d'analyser sa mesure spectrale. On a

$$X = A - \langle A \rangle \tag{2}$$

$$X_{ij} = \begin{cases} B_{ij}(p_{in}) - p_{in} & : (i, j < \frac{n}{2}) \; ou \; (i, j \geq \frac{n}{2}) \\ B_{ij}(p_{out}) - p_{out} & : else \; where \end{cases}$$

Où $B_{ij}(p) \sim B(p)$, $B(p)$ loi de Bernoulli de paramètre p et $B_{ij} = B_{ji}$

2- Recherche de la mesure spectrale de $\frac{X}{\sqrt{n}}$

Nous aimerions modifier la forme des entrées X_{ij} en $\sigma_{ij} Z_{ij}$, où Z_{ij} est une variable aléatoire centrée réduite, afin de nous ramener à des théorèmes connus.

Pour $(i, j < \frac{n}{2})$ *ou* $(i, j \geq \frac{n}{2})$ on a :

$$\begin{aligned}
\mathbb{E}(X_{ij}) &= \mathbb{E}(B(p_{in})) - p_{in} = 0 \\
\sigma_{in}^2 &= \mathbb{V}(X_{ij}) \\
&= \mathbb{V}(B(p_{in})) \\
&= p_{in}(1 - p_{in})
\end{aligned}$$

même raisonnement avec p_{out}

$$\sigma_{out}^2 = p_{out}(1 - p_{out})$$

finalement on obtient

$$X_{ij} \sim \begin{cases} \sigma_{in} Z_{ij} & : (i, j < \frac{n}{2}) \; ou \; (i, j \geq \frac{n}{2}) \\ \sigma_{out} Z_{ij} & : else \; where \end{cases}$$

Où $Z_{ij} = \frac{B(p)-p}{\sqrt{p(1-p)}}$ *avec* $p = p_{in}$ *ou* p_{out}

Le théorème de Wigner ne fonctionne que pour les matrices à entrées iid. Il existe un théorème lorsque le profil de variance de la matrice aléatoire étudié est stochastique.

Theorem 1 *Soit* $W \in M_n(\mathbb{R})$, W_{ij} *sont des variables aléatoires réelles*
On suppose que les $(W_{ij})_{i\leq j}$ *sont indépendantes telles que* $\mathbb{E}(W_{ij}) = 0$, $\sigma_{ij}^2 = \mathbb{E}(W_{ij}^2) < \infty$, $W_{ij} = W_{ji}$
On appelle profil de variance la matrice symétrique

$$\tilde{V}_n = (\sigma_{ij}^2)_{1\leq i,j\leq n}$$

et profil de variance normalisé

$$V_n = \left(\frac{\sigma_{ij}^2}{n}\right)_{1\leq i,j\leq n}$$

On suppose que V_n *est stochastique :* $\forall i = 1 : n$, $\sum_{j=1}^n V_{ij} = \sigma^2$
Alors la mesure spectrale L_n *de* $\frac{W}{\sqrt{n}}$ *vérifie*

$$L_n \xrightarrow[n\to+\infty]{etr} \mathbb{P}_{wig}\ p.s$$

où $\mathbb{P}_{wig}$ *est la loi de Wigner de densité* $f_{wig}(x) = \frac{\sqrt{(4\sigma^2-x^2)_+}}{2\pi\sigma^2}$

Soit V le profil de variance de $\frac{X}{\sqrt{n}}$. $\forall i = 1 : n$ on a :

$$\begin{aligned}\sigma_{i\cdot}^2 &= \sum_{j=1}^n V_{ij} \\ &= \boxed{\frac{\sigma_{in}^2 + \sigma_{out}^2}{2} = \sigma^2}\end{aligned}$$

Si on note $\rho(x)$ la densité spectrale de $\frac{X}{\sqrt{n}}$ on a ;

$$\rho(x) = \frac{\sqrt{(2(\sigma_{in}^2 + \sigma_{out}^2) - x^2)_+}}{\pi(\sigma_{in}^2 + \sigma_{out}^2)} \tag{3}$$

3- Étude de la mesure spectrale de $\frac{X}{\sqrt{n}}$ et recherche de ces valeurs propres maximales

Nous venons de trouver la densité spectrale de $\frac{X}{\sqrt{n}}$. Nous voulons à présent étudier ces valeurs propres. Pour ce faire nous allons séparer l'équation $\frac{X}{\sqrt{n}} = \frac{1}{\sqrt{n}}(A - \langle A\rangle)$ en deux parties, à savoir $\frac{X}{\sqrt{n}} = \frac{1}{\sqrt{n}}(B + \frac{1}{2}(c_{in}+c_{out})\mathbf{1}\mathbf{1}^T)$
Dans la suite de l'étude nous noterons B la matrice de modularité telle que

$$\begin{aligned}B :&= X + \frac{1}{2}(c_{in} - c_{out})\mathbf{u}\mathbf{u}^T \\ \frac{B}{\sqrt{n}} &= \frac{X}{\sqrt{n}} + \frac{1}{2\sqrt{n}}(c_{in} - c_{out})\mathbf{u}\mathbf{u}^T\end{aligned}$$

Le raisonnement qui va suivre est une heuristique. En effet, pour compléter rigoureusement la preuve qui va suivre, il faudrait vérifier la convergence du terme $\frac{B}{\sqrt{n}}$ en entier.

Soit v le vecteur propre de $\frac{B}{\sqrt{n}}$ associé à λ_{max} et soit $\Gamma = \frac{X}{\sqrt{n}}$

$$\begin{aligned}\frac{B}{\sqrt{n}}\mathbf{v} &= \lambda_{max}\mathbf{v} \\ (\Gamma - \lambda_{max}I)\mathbf{v} &= -\frac{1}{2\sqrt{n}}(c_{in} - c_{out})\mathbf{u}\mathbf{u}^T\mathbf{v} \\ \mathbf{u}^\mathbf{T}\mathbf{v} &= -\frac{1}{2\sqrt{n}}(c_{in} - c_{out})\mathbf{u}^\mathbf{T}(\Gamma - \lambda_{max}I)^{-1}\mathbf{u}\mathbf{u}^T\mathbf{v} \\ 1 &= -\frac{1}{2\sqrt{n}}(c_{in} - c_{out})\mathbf{u}^\mathbf{T}(\Gamma - \lambda_{max}I)^{-1}\mathbf{u}\end{aligned} \tag{4}$$

Theorem 2 (Théorème de Wigner isotrope) *Soit $W \in M_n(\mathbb{R})$, telle que W_{ij} sont des variables aléatoires réels, $\mathbb{E}(W_{ij}) = 0$, $\mathbb{E}(W_{ij}^2) < \infty$, $W_{ij} = W_{ji}$, W à un profil de variance V tel que $\forall i = 1 : n$, $\sum_{j=1}^{n} V_{ij} = \sigma^2$ pour $\sigma \in \mathbb{R}$*
Soit $Q(z)$ la résolvante de W

$$Q(z) = (W - zI)^{-1}$$

Soient $\mathbf{u}$, $\mathbf{v}$ des vecteurs déterministes tels que $\|\mathbf{u}\|, \|\mathbf{v}\| < \infty$.
Alors

$$\mathbf{u}^* Q(z)\mathbf{v} - \langle \mathbf{u}, \mathbf{v} \rangle g_{wig}^{\sigma^2}(z) \xrightarrow[n\to+\infty]{} 0$$

où $g_{wig}^{\sigma^2}$ est la transformé de Stieltjes de la loi de Wigner de paramètre σ^2 qui satisfait l'équation

$$\sigma^2 g_{wig}^{\sigma^2}(z)^2 + z g_{wig}^{\sigma^2}(z) + 1 = 0 \tag{5}$$

Reprenons l'équation (4), en appliquant le Théorème de Wigner isotrope et en s'assurant que tous les termes de droite convergent, nous avons :

$$\mathbf{u^T}(\Gamma - \lambda_{max} I)^{-1}\mathbf{u} \xrightarrow[n\to+\infty]{} g_{wig}^{\sigma^2}(\lambda_{max})$$

or $g_{wig}^{\sigma^2}(\lambda_{max})$ satisfait l'équation suivante

$$\sigma^2 g_{wig}^{\sigma^2}(\lambda_{max})^2 + \lambda_{max} g_{wig}^{\sigma^2}(\lambda_{max}) + 1 = 0 \implies g_{wig}^{\sigma^2}(\lambda_{max}) = \frac{-\lambda_{max} \pm \sqrt{(\lambda_{max}^2 - 4\sigma^2)}}{2\sigma^2} \tag{6}$$

Donc

$$\begin{aligned}
(4) &\Leftrightarrow 1 = -\frac{1}{2\sqrt{n}}(c_{in} - c_{out}) g_{wig}^{\sigma^2}(\lambda_{max}) \\
&\Leftrightarrow 1 = -\frac{1}{2\sqrt{n}}(c_{in} - c_{out}) \frac{-\lambda_{max} - \sqrt{(\lambda_{max}^2 - 4\sigma^2)}}{2\sigma^2} \\
&\Leftrightarrow \boxed{\lambda_{max} = \frac{(c_{in} - c_{out})}{2\sqrt{n}} + \sqrt{n}\frac{\sigma_{in}^2 + \sigma_{out}^2}{c_{in} - c_{out}}}
\end{aligned} \tag{7}$$

de plus $\frac{1}{2\sqrt{n}}(c_{in} - c_{out}) = \mathcal{O}(\sqrt{n})$ et $\sqrt{n}\frac{\sigma_{in}^2 + \sigma_{out}^2}{c_{in} - c_{out}} = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$

Comme $A = X + \frac{1}{2}(c_{in} + c_{out})\mathbf{1}\mathbf{1}^T + \frac{1}{2}(c_{in} - c_{out})\mathbf{u}\mathbf{u}^T$ et $\langle \mathbf{u}, \mathbf{1} \rangle = 0$ $\|\mathbf{u}\| = \|\mathbf{v}\| = 1$ on a

$$\begin{aligned}
&\Leftrightarrow (\Gamma + \alpha \mathbf{1}\mathbf{1}^T + \beta \mathbf{u}\mathbf{u}^T)\mathbf{v} = \lambda \mathbf{v} \\
&\Leftrightarrow (\Gamma - \lambda I)\mathbf{v} = -\alpha \mathbf{1}\mathbf{1}^T\mathbf{v} - \beta \mathbf{u}\mathbf{u}^T\mathbf{v} \\
&\Leftrightarrow \mathbf{1^T}\mathbf{v} = -\alpha \mathbf{1^T}(\Gamma - \lambda I)^{-1}\mathbf{1}\mathbf{1}^T\mathbf{v} - \beta \mathbf{1^T}(\Gamma - \lambda I)^{-1}\mathbf{u}\mathbf{u}^T\mathbf{v} \\
&\xrightarrow[n\to+\infty]{} 1 = -\alpha g_{wig}^{\sigma^2}(\lambda) \\
&\Leftrightarrow 1 = -\alpha \frac{-\lambda - \sqrt{(\lambda^2 - 4\sigma^2)}}{2\sigma^2} \\
&\Leftrightarrow \boxed{\lambda = \frac{(c_{in} + c_{out})}{2\sqrt{n}} + \sqrt{n}\frac{\sigma_{in}^2 + \sigma_{out}^2}{c_{in} + c_{out}}}
\end{aligned} \tag{8}$$

On a finalement une matrice A qui est la somme d'une matrice aléatoire X et de deux perturbations de rang 1. De fait, la mesure spectrale de A à deux valeurs propres qui sortent du support de la distribution de Wigner, à savoir λ_{max} et λ. Par la suite nous noterons $z_1 = \lambda_{max}$ et $z_2 = \lambda$, et nous appellerons "bulk" le support de la mesure spectrale $\rho(z)$.

3.3 Interprétation des résultats obtenus

Si les 2 plus petites valeurs propres sont inférieures au bord gauche du bulk (ici $\lambda^- = -2\sigma$) alors le graphe admet une structure de communauté "disassortative". Si les 2 plus grandes valeurs propres sont supérieures au bord droit du bulk (ici $\lambda^+ = 2\sigma$) alors le graphe admet une structure de communauté "assortative". Si 0 ou 1 valeur propre sort du bulk alors la méthode spectrale ne peut rien conclure sur la structure de communauté du graphe G.

Il est important de remarquer que z_2 est toujours supérieure au bord droit du bulk, indépendamment des valeurs de p_{in} et p_{out}. Par conséquent, c'est z_1 qui nous indique si oui ou non il y a une structure de communauté dans le graphe.

Nous cherchons à déterminer $p_{lim} = p_{in} - p_{out}$, qui est la condition limite supérieure sur les paramètres du SBM pour que l'algorithme puisse détecter la structure de communauté "assortative" du graphe G. Dans la mesure où λ^+ est positif ou nul, lorsque $p_{in} - p_{out} \in [0, p_{lim}]$ alors la méthode spectrale est incapable de conclure sur la structure "assortative".

La condition limite naturelle est celle où la valeur propre maximale est égale au bord droit du support de la mesure spectrale de la matrice A. À partir de cette condition nous allons essayer de retrouver p_{lim}.
On a alors

$$\Leftrightarrow \lambda^+ = \lambda_{max}$$
$$\Leftrightarrow \sqrt{2(\sigma_{in}^2 + \sigma_{out}^2)} = \frac{(c_{in} - c_{out})}{2\sqrt{n}} + \sqrt{n}\frac{\sigma_{in}^2 + \sigma_{out}^2}{c_{in} - c_{out}}$$
$$\Leftrightarrow \sqrt{2S} = \alpha + \beta S \qquad avec\ S = \sigma_{in}^2 + \sigma_{out}^2,\ \alpha = \frac{(c_{in} - c_{out})}{2\sqrt{n}},\ \beta = \frac{\sqrt{n}}{c_{in} - c_{out}}$$
$$\Leftrightarrow 2S = \alpha^2 + 2\alpha\beta S + \beta^2 S^2$$
$$\Leftrightarrow 0 = \beta^2 S^2 + 2(\alpha\beta - 1)S + \alpha^2$$
$$\Leftrightarrow p_{in} - p_{out} = \frac{\sqrt{2(\sigma_{in}^2 + \sigma_{out}^2)}}{\sqrt{n}}$$

Donc

$$\boxed{p_{lim} = \frac{\sqrt{2(\sigma_{in}^2 + \sigma_{out}^2)}}{\sqrt{n}} = \frac{2\sigma}{\sqrt{n}} = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)} \tag{9}$$

D'après ce modèle, plus on a de noeuds dans le graphe, plus on a de l'information, et donc moins on a de chance de tomber sur cet intervalle d'indécidabilité.

3.4 Bilan

Ci-dessous un tableau comparatif entre les résultats obtenus et ceux de l'article [7]. L'essence de la divergence intervient lorsque qu'il est défini que $\mathbb{V}(X_{ij}) = \frac{(c_{in}+c_{out})}{2n}$. Nous avons normalisé les équations de l'article original par $\frac{1}{\sqrt{n}}$ afin de pouvoir les comparer avec nos résultats.
Rappel : $\sigma_{in}^2 = p_{in}(1-p_{in})$; $\sigma_{out}^2 = p_{out}(1-p_{out})$; $c_{in} = np_{in}$; $c_{out} = p_{out}$

	Résultats de l'article	Résultats obtenus
$\rho(z)$	$\frac{1}{\pi\sqrt{n}}\frac{\sqrt{(2(c_{in}+c_{out})-x^2)_+}}{(c_{in}+c_{out})}$	$\frac{\sqrt{(2(\sigma_{in}^2+\sigma_{out}^2)-x^2)_+}}{\pi(\sigma_{in}^2+\sigma_{out}^2)}$
z_1	$\frac{c_{in}-c_{out}}{2\sqrt{n}} + \sqrt{n}\frac{c_{in}+c_{out}}{c_{in}-c_{out}}$	$\frac{c_{in}-c_{out}}{2\sqrt{n}} + \sqrt{n}\frac{\sigma_{in}^2+\sigma_{out}^2}{c_{in}-c_{out}}$
z_2	$\frac{c_{in}+c_{out}}{2\sqrt{n}} + \sqrt{n}$	$\frac{c_{in}+c_{out}}{2\sqrt{n}} + \sqrt{n}\frac{\sigma_{in}^2+\sigma_{out}^2}{c_{in}+c_{out}}$
p_{lim}	$\frac{\sqrt{2(c_{in}-c_{out})}}{n}$	$\frac{\sqrt{2(\sigma_{in}^2+\sigma_{out}^2)}}{\sqrt{n}}$

Les résultats quantitatifs empiriques que nous avons obtenu sont différents de ceux de l'article de référence. Au contraire, les simulations corroborent les formules trouvées ci-dessus. Nous le verrons plus précisément dans la section suivante.

4 Simulations

Le but est de générer une matrice d'adjacence A sous les mêmes hypothèses que (1) en faisant varier les différents paramètres, à savoir : n, p_{in}, p_{out}

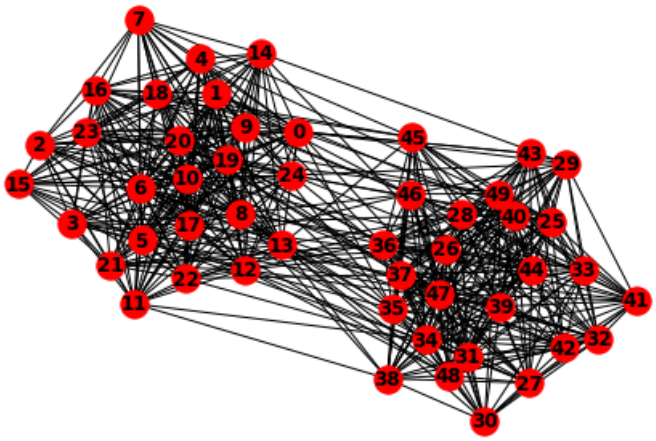


FIGURE 4 – Graphe généré à partir des paramètres : $n = 50$, $p_{in} = 0.8$, $p_{out} = 0.1$

L'élément discriminant du test spectral étant la variable $\Delta p = p_{in} - p_{out}$, nous allons tester 3 valeurs représentatives des différents types de résultats :

1- $\Delta p \in [-1,\ -p_{lim}] \implies$ le graphe comporte de structure de communauté "disassortative" ;
2- $\Delta p \in [-p_{lim},\ p_{lim}] \implies$ la méthode spectrale ne peut rien conclure sur la stucture de communauté du graphe ;
2- $\Delta p \in [p_{lim},\ 1] \implies$ le graphe comporte une structure de communauté "assortative".

Par la suite nous allons tester pour les valeurs $\Delta p = 0.5, 0.08, -0.5$. De plus nous ferons une première batterie de test avec $n = 100$ et une autre avec $n = 1000$. La terminologie utilisée dans les figures ci-dessous est :

- n, p_{in}, p_{out}, p_{lim}, z_1, z_2 : sont identiques aux notations utilisées jusqu'à présent ;
- z_1 *theoric*, z_2 *theoric* : sont les plus grandes valeurs propres z_1 et z_2 calculées via les équations (7) et (8) ;
- p_{in} *estimated*, p_{out} *estimated* : sont les probabilités du SBM calculées a posteriori grâce aux valeurs propres z_1 et z_2.

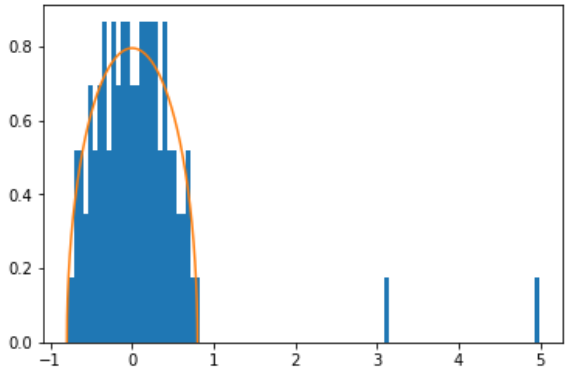
Les p_{in} *estimated*, p_{out} *estimated* sont calculés via un calcul d'optimisation dont les conditions initiales sont $p_{in,0} = 1$ et $p_{out,0} = 0$.

La première observation que l'on peut faire est que les valeurs propres de nos matrices d'adjacence sont bien distribuées selon la loi du demi-cercle de Wigner. De plus, en fonction de Δp la mesure spectrale est perturbée (ou pas) par une ou deux valeurs propres qui sortent du support de la distribution initiale.

Dans les cas avec $\Delta p = 0.5$ (Figure 3a, Figure 3d), on voit très clairement deux valeurs propres qui se détachent du support de la distribution de Wigner. Les valeurs z_1, z_2 correspondent bien aux valeurs théoriques avec un taux d'erreur de l'ordre de 10^{-2}. Par conséquent lorsque l'on obtient une valeur propre négative, le modèle spectral nous permet de conclure qu'il y a une structure de communauté "assortative" dans le graphe.

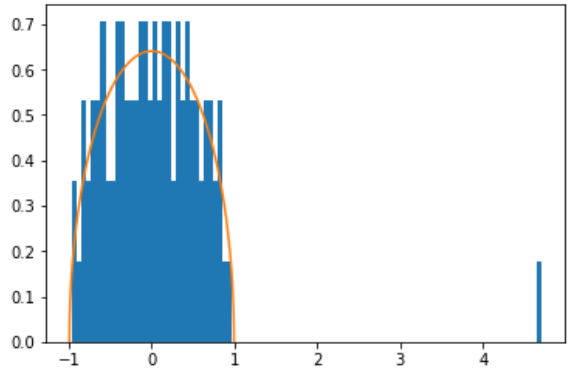
Dans les cas avec $\Delta p = -0.5$ (Figure 3b, Figure 3e), selon l'Équation 8, on ne s'attend à ce qu'une des valeurs propres théorique soit négative. On observe une valeur propre négative en dehors du support de la loi de Wigner et une autre au dessus du support. Ces valeurs observées correspondent on valeurs théoriques avec une erreur de l'ordre de 10^{-2}. Par conséquent lorsque l'on obtient une valeur propre négative, le modèle spectral

```
n = 100
p_in = 0.8 , p_out = 0.2
p_in esitmated = 0.7931 , p_out esitmated = 0.1931
p_lim = 0.08 , pin - pout = 0.6
z1 = 4.98 , z1 theoric = 5.032
z2 = 3.081 , z2 theoric = 3.053
```



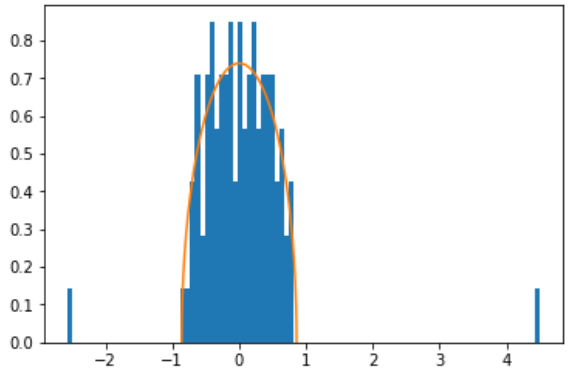
(a) $n = 100$, $\Delta p = 0.6$

```
n = 100
p_in = 0.5 , p_out = 0.42
p_in esitmated = 0.5134 , p_out esitmated = 0.4158
p_lim = 0.099 , pin - pout = 0.08
z1 = 4.698 , z1 theoric = 4.654
z2 = 0.937 , z2 theoric = 1.017
```



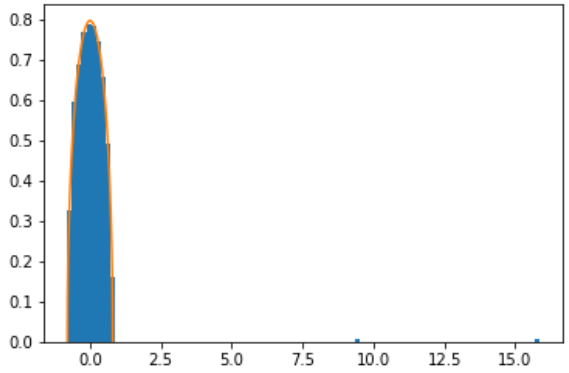
(b) $n = 100$, $\Delta p = 0.08$

```
n = 100
p_in = 0.2 , p_out = 0.7
p_in esitmated = 0.197 , p_out esitmated = 0.6898
p_lim = 0.086 , pin - pout = -0.5
z1 = 4.489 , z1 theoric = 4.541
z2 = -2.563 , z2 theoric = -2.574
```



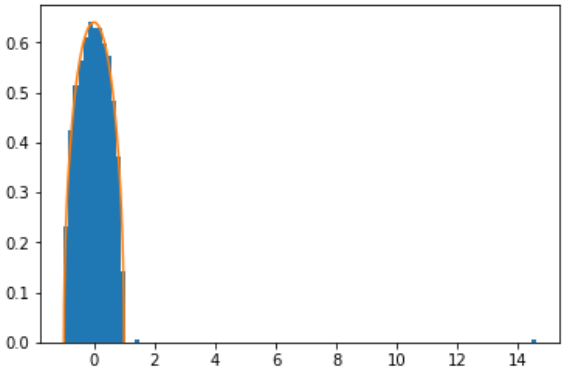
(c) $n = 100$, $\Delta p = -0.5$

```
n = 1000
p_in = 0.8 , p_out = 0.2
p_in esitmated = 0.7993 , p_out esitmated = 0.2012
p_lim = 0.025 , pin - pout = 0.6
z1 = 15.836 , z1 theoric = 15.822
z2 = 9.482 , z2 theoric = 9.504
```



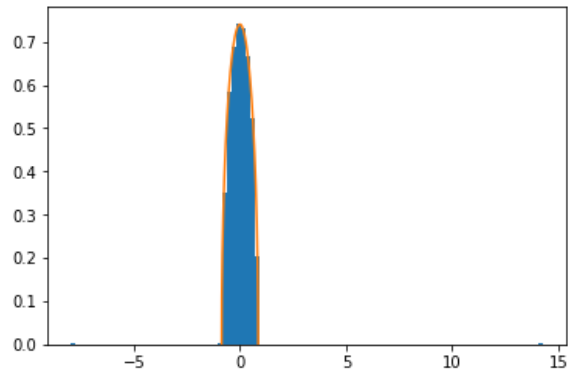
(d) $n = 1000$, $\Delta p = 0.6$

```
n = 1000
p_in = 0.5 , p_out = 0.42
p_in esitmated = 0.5018 , p_out esitmated = 0.4187
p_lim = 0.031 , pin - pout = 0.08
z1 = 14.571 , z1 theoric = 14.563
z2 = 1.499 , z2 theoric = 1.46
```



(e) $n = 1000$, $\Delta p = 0.08$

```
n = 1000
p_in = 0.2 , p_out = 0.7
p_in esitmated = 0.1994 , p_out esitmated = 0.7008
p_lim = 0.027 , pin - pout = -0.5
z1 = 14.252 , z1 theoric = 14.243
z2 = -7.959 , z2 theoric = -7.929
```



(f) $n = 1000$, $\Delta p = -0.5$

nous permet de conclure qu'il y a une structure de communauté "disassortative" dans le graphe.

Enfin, les cas avec $\Delta p = 0.08$ (Figure 3c, Figure 3f). On se trouve dans le cas limite décrit dans sous-section 3.3, dans lequel le modèle ne peut pas interpréter les résultats du modèle spectral. On peut observer sur la Figure 3c qu'il n'y a qu'une seule valeur propre qui est en dehors du support. Par conséquent il n'y a qu'une seule valeur propre extrémale (ici z_1) qui correspond à sa valeur théorique, de fait nous ne pouvons conclure sur la structure de communauté du graphe. Cependant, avec le même Δp mais avec plus de nœud ($n = 1000$) le modèle spectral parvient à détecter la structure de communauté et à retrouver les valeurs z_1 et z_2. Ceci confirme la remarque du paragraphe 3.3 selon laquelle, plus il y a d'information (i.e nœuds) dans le graphe plus la valeur de p_{lim} tend vers 0.

Si on compare les valeurs théoriques de l'article initial [7] avec les valeurs des simulations, on voit très clairement une différence.
Par exemple, pour $p_{in} = 0.8$, $p_{out} = 0.3$, et $n = 100$ les valeurs propres calculées avec les équations Bilan sont $z_1 = 22.785$ et $z_2 = 18.815$. Les valeurs propres empiriques sont $z_1 = 5.538$ et $z_2 = 2.545$.
Ou encore, pour $p_{in} = 0.2$, $p_{out} = 0.7$, et $n = 1000$ les valeurs propres calculées avec les équations Bilan sont $z_1 = -56.732$ et $z_2 = 31.833$. Les valeurs propres empiriques sont $z_1 = 14.299$ et $z_2 = -7.934$.

5 Erreurs de l'article

Dans cette section, allons redémontrer les résultats obtenus dans le Tableau **??** via la combinatoire. Par la même occasion nous allons expliquer les erreurs commises dans l'article [7, Graph spectra and the detectability of community structure in networks].

5.1 Normalisation de la mesure spectrale de X

Dans l'article l'équation (7) donne la mesure spectrale de la matrice X.

$$\rho(z) = \frac{n}{\pi}\frac{\sqrt{2(c_{in}+c_{out})-z^2}}{c_{in}+c_{out}}$$

Calculons l'intégrale de $\rho(z)$. On sait que son support est $[a,b] = [-\sqrt{2(c_{in}+c_{out})}, \sqrt{2(c_{in}+c_{out})}]$.

$$\begin{aligned}
\int_a^b \rho(z)\,\mathrm{d}z &= \int_a^b \frac{n}{\pi}\frac{\sqrt{2(c_{in}+c_{out})-z^2}}{c_{in}+c_{out}}\,\mathrm{d}z \\
&= \int_a^b \frac{1}{\pi}\frac{\sqrt{2n(p_{in}+p_{out})-z^2}}{p_{in}+p_{out}}\,\mathrm{d}z \\
&= \int_a^b \frac{\sqrt{n}}{\pi}\frac{\sqrt{2(p_{in}+p_{out})-\frac{z^2}{n}}}{p_{in}+p_{out}}\,\mathrm{d}z \\
&= n\int_{a'}^{b'} \frac{1}{\pi}\frac{\sqrt{2(p_{in}+p_{out})-u^2}}{p_{in}+p_{out}}\,\mathrm{d}u
\end{aligned}$$

avec $u^2 = \frac{z^2}{n}$, $a' = \sqrt{2(p_{in}+p_{out})}$ et $b' = \sqrt{2(p_{in}+p_{out})}$. Or la fonction sous l'intégrale correspond à la loi du demi-cercle de Wigner de paramètre $\sigma = (p_{in}+p_{out})$, qui est normalisée sur le fermé $[a', b']$. On obtient donc :

$$\int_a^b \rho(z)\,\mathrm{d}z = n$$

Il y a donc un problème de normalisation.

5.2 Choix de la variance des entrées de X

Nous allons réécrire le raisonnement de l'article [7, Graph spectra and the detectability of community structure in networks] lorsque l'on souhaite déterminer la densité spectrale de X.

La transformé de Stieltjes de X est :

$$\rho(z) = \frac{1}{\pi} Im\langle Tr(zI - X)^{-1}\rangle \tag{10}$$

où $\langle \dots \rangle$ indique la moyenne de l'ensemble. On peut réécrire la trace de la moyenne comme ce qui suit :

$$\langle Tr(zI - X)^{-1} \rangle = \frac{1}{z} \sum_{k=0}^{\infty} \frac{Tr\langle X^k \rangle}{z^k} \tag{11}$$

$$Tr\langle X^k \rangle = \sum_{i_1 \dots i_k} \langle X_{i_1 i_2} X_{i_1 i_2} \dots X_{i_k i_1} \rangle \tag{12}$$

D'après l'analyse préliminaire effectué sous-section 3.2, on sait que X est centrée et que les X_{ij} $\forall i \leq j$ sont des variables de Bernoulli indépendantes définies celon (2). Par conséquent $\langle X_{i_1 i_2} X_{i_1 i_2} \dots X_{i_k i_1} \rangle \neq 0$ si $k = 2m$ avec $m \in \mathbb{N}$ et si chaque X_{ij} apparaît exactement deux fois. De plus.

$$\langle X_{ij}^2 \rangle = \begin{cases} \sigma_{in}^2 & si\ (i,j < \frac{n}{2})\ ou\ (i,j \geq \frac{n}{2}) \\ \sigma_{out}^2 & else\ where \end{cases}$$

C'est à ce moment qu'apparaît l'erreur. En effet dans l'article ils choisissent comme variances des entrées de X :

$$\langle X_{ij}^2 \rangle = \frac{c_{in} + c_{out}}{2n} \quad \forall\, i,j \in \{1, \dots, n\}$$

5.3 Détermination de la mesure spectrale de X par la combinatoire

Pour trouver la mesure spectrale de X, l'article utilise une méthode combinatoire. Nous allons refaire le raisonnement en se basant sur la démonstration du théorème de Wigner par la combinatoire dans [3, Introduction aux matrices aléatoires].

L'idée est de montrer que

$$\mathbb{E} \int f d\mu_n \xrightarrow[n \to +\infty]{} \int f d\mu \tag{13}$$

Avec μ_n la densité spectrale de la matrice $\frac{X}{\sqrt{n}}$ construite celons Équation 1 et μ la loi de Wigner de paramètre $\sigma^2 = \frac{\sigma_{in}^2 + \sigma_{out}^2}{2}$.

Presque tout le raisonnement de notre preuve est identique à celui dans [3, Introduction aux matrices aléatoires]. Cependant il diffère à dans la définition de matrice étudié X. En effet, dans leur cas les variances de toutes les entrées de la matrice hermitienne sont centrée réduites, elles sont iid et réelles sur la diagonale et iid et complexes ailleurs.

Il faut reprendre le raisonnement sur les graphes qui contribue ou non à la valeur des moments de μ_n.

type 1 : ceux pour qui chaque arrête est présente dans l'autre sens le même nombre de fois, et le graphe obtenu en effaçant les orientations des arrêtes est un graphe sans cycles, c'est-à-dire un arbre ;

type 2 : ceux pour qui une arrête au moins n'apparaît qu'une seule fois ;

type 3 : ceux qui ne sont pas de type 1 ou de type 2.

La contribution de chaque type reste la même, nous allons donc nous concentrer sur les graphes de type 1. En reprenant la formule des moments on a

$$\begin{aligned} \mathbb{E}[X^2 m] &= \sum_{G de type 1} \frac{n(n-1)\dots(n-t+1)}{n^{1+r/2}} E(H_G) \\ &= \frac{n}{n} \dots \frac{n-m+1}{n} \frac{1}{k+1} \binom{2m}{m} E(H_G) \\ &\xrightarrow[n \to +\infty]{} \frac{1}{1+m} \binom{2m}{m} E(H_G) \end{aligned}$$

or après calcul on a

$$E(H_G) = \frac{1}{2^m} \sum_{k=0}^{m} \binom{m}{k} (\sigma_{out}^2)^{m-k} (\sigma_{in}^2)^k$$

On retrouve la formule du binôme de newton. On agrégeant tous les résultats on a

$$\boxed{\mathbb{E}[X^2 m] \xrightarrow[n \to +\infty]{} \frac{1}{1+m} \binom{2m}{m} (\frac{\sigma_{in}^2 + \sigma_{out}^2}{2})^m} \tag{14}$$

Ce qui correspond exactement au moment d'ordre de $2m$ de la loi du demi-cercle de paramètre $\sigma^2 = \frac{\sigma_{in}^2 + \sigma_{out}^2}{2}$.

6 Généralisation

6.1 Cas avec n communautés

On peut a présent généraliser à un nombre de communautés $q \geq 2$. Nous allons supposer que les communautés sont de même taille, à savoir $n_q = \frac{n}{q}$.

Une première contrainte apparaît, de par l'utilisation des théorèmes 1 et 2, sur les valeurs des probabilités de la matrice d'adjacence A. En effet, d'après le Théorème 1, pour que la matrice X est une mesure spectrale qui tende vers la loi de Wigner il faut que la norme 1 des vecteurs lignes de son profil de variance soient égales ($\| x \|_1 = \sum_{j=1}^{n} |x_j|$). Par conséquent, si on veut augmenter le nombre de communautés q dans le modèle, on est forcé de garder deux probabilités p_{in} et p_{out} qui jouent le même rôle que celles introduites précédemment (cf. 1).

On sait que la matrice d'adjacence du graphe sous le SBM à q communautés est $A = X + \langle A \rangle$. Pour poursuivre l'analyse on va suivre la trame suivante :

1- Trouver l'équation de $\langle A \rangle$;
2- Trouvez l'équation de X et déterminer son profil de variance ;
3- Trouvez les q valeurs propres associées aux perturbations de rang 1 ;
4- Trouver p_{lim}.

1- Équation de $\langle A \rangle$

$\langle A \rangle$ étant symétrique, le théorème spectral nous dis qu'il existe une base orthonormée telle que $\langle A \rangle = \sum_{n}^{i=1} \lambda_i \mathbf{u}_i \mathbf{u}_i^\star$. Après les calculs on trouve (voir Annexe A) :

$$\langle A \rangle := n_q(p_{in} + (q-1)p_{out})\mathbf{u}_1\mathbf{u}_1^\star + n_q(p_{in} - p_{out}) \sum_{i=1}^{q-1} \mathbf{u}_i\mathbf{u}_i^\star \tag{15}$$

$$= \frac{c_{in} + (q-1)c_{out}}{q}\mathbf{u}_1\mathbf{u}_1^\star + \frac{c_{in} - c_{out}}{q} \sum_{i=1}^{q-1} \mathbf{u}_i\mathbf{u}_i^\star \tag{16}$$

où les trois valeurs propres sont 0, $n_q(p_{in} + (q-1)p_{out})$, $n_q(p_{in} - p_{out})$ de multiplicités $q(n_q - 1)$, 1, $q - 1$.

2- Profil de variance de $\frac{X}{\sqrt{n}}$

De la même manière que dans sous-section 3.2 on trouve :

$$X_{ij} \sim \begin{cases} \sigma_{in} Z_{ij} & : (i,j \in P_{in}) \\ \sigma_{out} Z_{ij} & : (i,j \in P_{out}) \end{cases}$$

Où $Z_{ij} = \frac{B_{ij}(p) - p}{\sqrt{p(1-p)}}$ *avec* $p = p_{in}$ *ou* p_{out}, $B_{ij}(p) \sim B(p)$, $B(p)$ loi de Bernoulli de paramètre p
La somme de n'importe quel vecteur ligne (ou colonne) du profil de variance de $\frac{X}{\sqrt{n}}$ est égale à :

$$\sigma^2 = \frac{\sigma_{in}^2 + (q-1)\sigma_{out}^2}{q} \tag{17}$$

Le profil de variance de $\frac{X}{\sqrt{n}}$ est donc une matrice bi-stochastique.

3- Valeurs propres de $\frac{A}{\sqrt{n}}$

Soit λ une valeur propre de $\frac{A}{\sqrt{n}}$ et le vecteur propre associé.

$$\begin{aligned}
&\Leftrightarrow \frac{A}{\sqrt{n}}v = \lambda v \\
&\Leftrightarrow \frac{\langle A\rangle + X}{\sqrt{n}}v = \lambda v \\
&\Leftrightarrow (\frac{X}{\sqrt{n}} - \lambda I)v = -\langle A\rangle v \\
&\Leftrightarrow (\Gamma - \lambda I)v = -\frac{c_{in} + (q-1)c_{out}}{q}\mathbf{u}_1\mathbf{u}_1^\star - \frac{c_{in} - c_{out}}{q}\sum_{i=1}^{q-1}\mathbf{u}_i\mathbf{u}_i^\star
\end{aligned} \tag{18}$$

Pour trouver la valeur propre associée à $\mathbf{u}_1$ on multiplie à gauche par $\mathbf{u}_1^\star(\Gamma - \lambda I)^{-1}$ et on obtient :

$$\begin{aligned}
(18) \Leftrightarrow \mathbf{u}_1^\star v &= -\alpha\mathbf{u}_1^\star(\Gamma - \lambda I)^{-1}\mathbf{u}_1\mathbf{u}_1^\star v - \beta\mathbf{u}_1^\star(\Gamma - \lambda I)^{-1}\sum_{i=1}^{q-1}\mathbf{u}_i\mathbf{u}_i^\star v \\
&\xrightarrow[n\to+\infty]{} 1 = -\alpha g_{wig}^{\sigma^2}(\lambda) \\
&\Leftrightarrow 1 = \alpha\frac{\lambda + \sqrt{\lambda^2 - 4\sigma^2}}{2\sigma^2} \\
&\Leftrightarrow \lambda = \frac{c_{in} + (q-1)c_{out}}{q\sqrt{n}} + \frac{q\sqrt{n}\sigma^2}{c_{in} + (q-1)c_{out}}
\end{aligned} \tag{19}$$

Si on remplace q par 2 on retrouve l'équation (8).
Pour trouver les valeurs propres associée aux $\mathbf{u}_i$, $\forall i \in \{2, \cdots, q-1\}$ on multiplie à gauche par $\mathbf{u}_2^\star(\Gamma - \lambda I)^{-1}$ et on obtient :

$$\begin{aligned}
(18) \Leftrightarrow \mathbf{u}_2^\star v &= -\alpha\mathbf{u}_2^\star(\Gamma - \lambda I)^{-1}\mathbf{u}_1\mathbf{u}_1^\star v - \beta\mathbf{u}_2^\star(\Gamma - \lambda I)^{-1}\sum_{i=1}^{q-1}\mathbf{u}_i\mathbf{u}_i^\star v \\
&\xrightarrow[n\to+\infty]{} 1 = -\beta g_{wig}^{\sigma^2}(\lambda) \\
&\Leftrightarrow 1 = \beta\frac{\lambda + \sqrt{\lambda^2 - 4\sigma^2}}{2\sigma^2} \\
&\Leftrightarrow \lambda = \frac{c_{in} - c_{out}}{q\sqrt{n}} + \frac{q\sqrt{n}\sigma^2}{c_{in} - c_{out}}
\end{aligned} \tag{20}$$

Si on remplace q par 2 on retrouve l'équation (7).
Les valeurs propres de A pour les valeurs propres de $\langle A\rangle$ égales à zéros appartiennent au support de la distribution de Wigner. Elles n'apportent donc aucune information supplémentaire sur la structure de communauté du graphe étudié.

Nous noterons $z_1 =$ (19) et $z_2 =$ (20) pour la suite. On voit immédiatement que $0 \leq z_1$ et que $z_2 \leq z_1$. Le critère discriminant pour détecter la structure de communauté donc la valeur propre $z2$. C'est la même conclusion que dans le modèle avec 2 communautés, à la différence que la valeur propre non discriminante z_1 est de multiplicité $q-1$. On retrouve donc les même règles que dans paragraphe 4

4- Seuil de décidabilité p_{lim}

Nous cherchons maintenant à déterminer p_{lim}. La condition limite naturelle est celle où la valeur propre z_2 qui sort du support de la distribution de Wigner est égale au bord droit du support de la mesure spectrale de la matrice A. On a alors

$$\begin{aligned}
&\Leftrightarrow \lambda^+ = z_1 \\
&\Leftrightarrow 2\sigma = \frac{c_{in} - c_{out}}{q\sqrt{n}} + \frac{q\sqrt{n}\sigma^2}{c_{in} - c_{out}} \\
&\Leftrightarrow 0 = \beta\sigma^2 - 2\sigma + \alpha \\
&\Leftrightarrow p_{in} - p_{out} = \frac{q\sigma}{\sqrt{n}}
\end{aligned}$$

Donc

$$p_{lim} = \frac{\sqrt{q(\sigma_{in}^2 + (q-1)\sigma_{out}^2)}}{\sqrt{n}} = \frac{q\sigma}{\sqrt{n}} \tag{21}$$

6.2 Bilan

Ci-dessous le bilan de la généralisation :

$$\begin{aligned}
\sigma^2 &= \frac{\sigma_{in}^2 + (q-1)\sigma_{out}^2}{q} \\
z_1 &= \frac{c_{in} + (q-1)c_{out}}{q\sqrt{n}} + \frac{q\sqrt{n}\sigma^2}{c_{in} + (q-1)c_{out}} \\
z_2 &= \frac{c_{in} - c_{out}}{q\sqrt{n}} + \frac{q\sqrt{n}\sigma^2}{c_{in} - c_{out}} \\
p_{lim} &= \frac{\sqrt{q(\sigma_{in}^2 + (q-1)\sigma_{out}^2)}}{\sqrt{n}}
\end{aligned}$$

6.3 Simulations

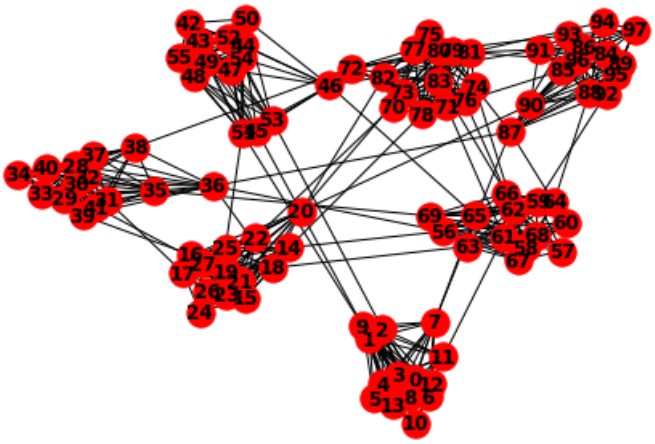


FIGURE 5 – Graphe généré à partir des paramètres : $q = 7$ $n = 100$, $p_{in} = 0.8$, $p_{out} = 0.01$

```
n = 500
p_in = 0.8 , p_out = 0.2
p_in esitmated = 0.7971 , p_out esitmated = 0.201
p_lim = 0.036 , pin - pout = 0.6
z1 = 7.858 , z1 theoric = 7.847
z2 = 3.406 , z2 theoric = 3.402
```

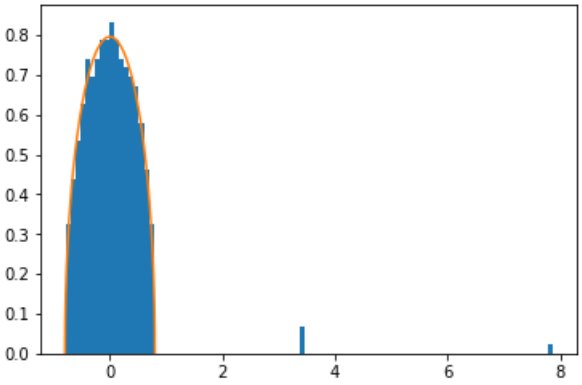


FIGURE 6 – $q = 4$, $n = 500$, $\Delta p = 0.6$

6.4 Limites du modèle

Le première limite de ce modèle est que l'on est cantonné à des communautés de même taille n_q. En effet si on change ce paramètre pour chacune des communautés alors le Théorème 1 n'est plus applicable, la somme des éléments de chaque vecteurs ligne du profil de variance n'est plus constant.

La deuxième contrainte est le fait que l'on doit toujours garder deux paramètres p_{in} et p_{out} indépendamment du nombre de communautés n_q et du nombre de nœuds dans le graphe n. Idéalement nous souhaiterions avoir un paramètre p_{ij} correspondant à la probabilité d'avoir une arrête entre le nœud i et le nœud j et ce $\forall i < j$.
Une manière d'encoder ces paramètres est d'utiliser la relation suivante $p_{ij} = q_i q_j C_\alpha$

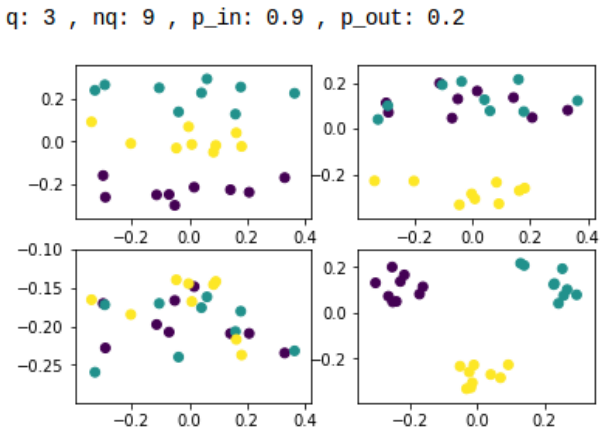
$$p_{ij} = q_i q_j C_{g_i g_j} \tag{22}$$

Où q_i est la probabilité intrinsèque du nœud i à avoir une arrête, g_i est la communauté correspondant au nœud i et $C_{g_i g_j}$ est le facteur de correction par communauté.
Cette formalisation plus générale est très répandu dans la théorie de la détection de communauté spectrale.

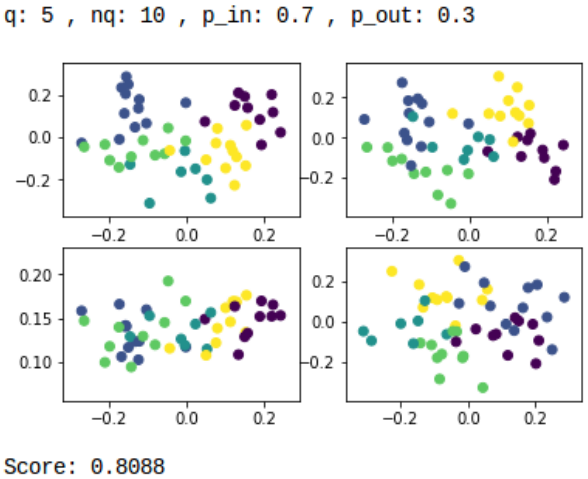
6.5 Test complet de l'algorithme de clustering spectral de graphe

Dans cette section nous allons retrouver les communautés d'un graphe généré à partir d'un SBM. Pour ce faire, nous allons suivre la procédure expliquer en 2.1. Le score utilisé est le *Rand index*, [5, p.78 eq.(88)]. Pour catégoriser les nœuds projetés dans l'espace associé aux vecteurs propres nous avons utilisé l'algorithme *K-means*.

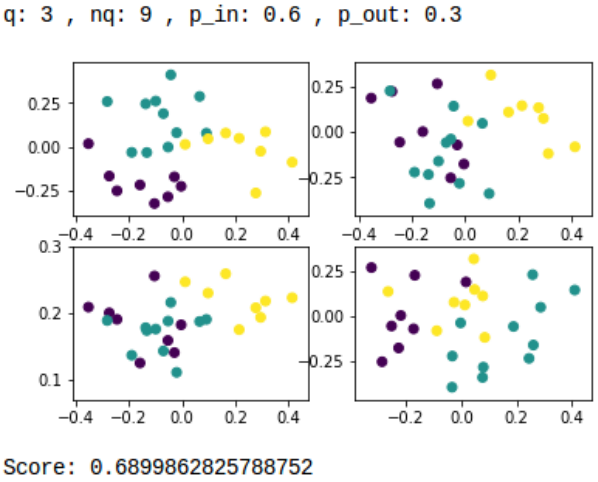
Ci-dessous les batteries de tests, leur score ainsi que les nœuds projetés sur les plans des 4 valeurs propres les plus significatives $\{z_0, z1\}$ $\{z_0, z2\}$ $\{z_0, z3\}$ $\{z_1, z2\}$



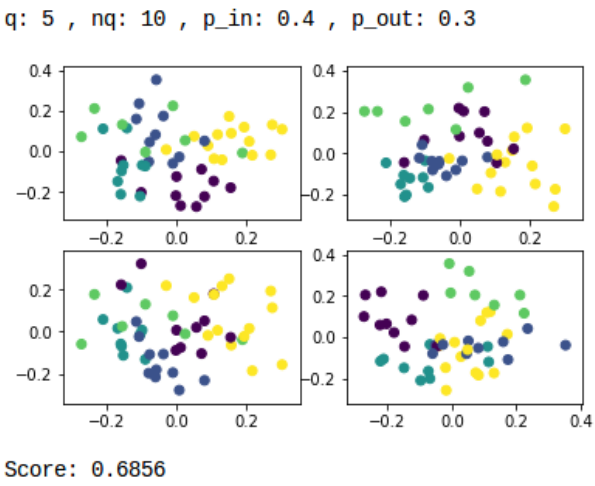
(a) $n = 27$, $p_{in} = 0.9$, $p_{out} = 0.2$



(b) $n = 50$, $p_{in} = 0.7$, $p_{out} = 0.3$



(c) $n = 50$, $p_{in} = 0.6$, $p_{out} = 0.3$



(d) $n = 50$, $p_{in} = 0.4$, $p_{out} = 0.3$

On remarque que même pour des graphes avec un ΔP très faible, la qualité du résultat est de l'ordre de 70%.

Quatrième partie
Analyse d'algorithmes de graph spectral clusturing

7 Bethe Hessian

7.1 Principe

Dans le contexte d'un SBM avec n nœuds et q communautés. Chaque nœud appartient à une communauté $g_v \in \{1, \dots, q\}$ La probabilité d'une arrête entre une paire de nœud (u, v) est $\mathbb{P}[A_{u,v} = 1] = p_{g_u, g_v}$.

L'algorithme est le même que pour la matrice d'adjacence A. Le modèle est suffisamment général pour prendre en compte un nombre arbitraire de communautés avec q^2 probabilités p_{g_u, g_v} d'existence arrêtes différentes entre chaque communauté.

Il y a cependant l'apparition d'un paramètre de régularisation qu'il faut calculer au préalable. Cependant il existe une formule fermée pour le calculer, à savoir $r = \sqrt{\frac{\langle d^2 \rangle}{\langle d \rangle} - 1}$, où $\langle d \rangle$ et $\langle d^2 \rangle$ sont le premier et le deuxième moment de la distribution des dégrée des nœuds dans le graphe.

L'avantage de cet opérateur linéaire par rapport à la matrice d'adjacence A est, qu'avec le bon choix de r,

l'ensemble des q valeurs propres portant l'information de la structure de communauté du graphe sont négatives alors que toutes les valeurs propres du bulk de $H(r)$ sont positives. Il suffit donc de calculer les q valeurs propres négatives et s'arrêter dès que le signe change. Grâce à ce critère discriminant, nous avons instantanément le nombre de communautés dans le graphe. Dans le cas avec l'opérateur d'adjacence, il est impossible de faire une inférence du nombre de communauté dans la graphe.

Dans le cas la matrice Bethe Hessian est définie de la manière suivante :

$$H(r) := (r^2 - 1)I - rA + D \tag{23}$$

Où $D_{ii} = d_i \ \forall i \in V$ avec d_i est le degré du nœud i, et r est le paramètre de régularisation.

7.2 Simulations

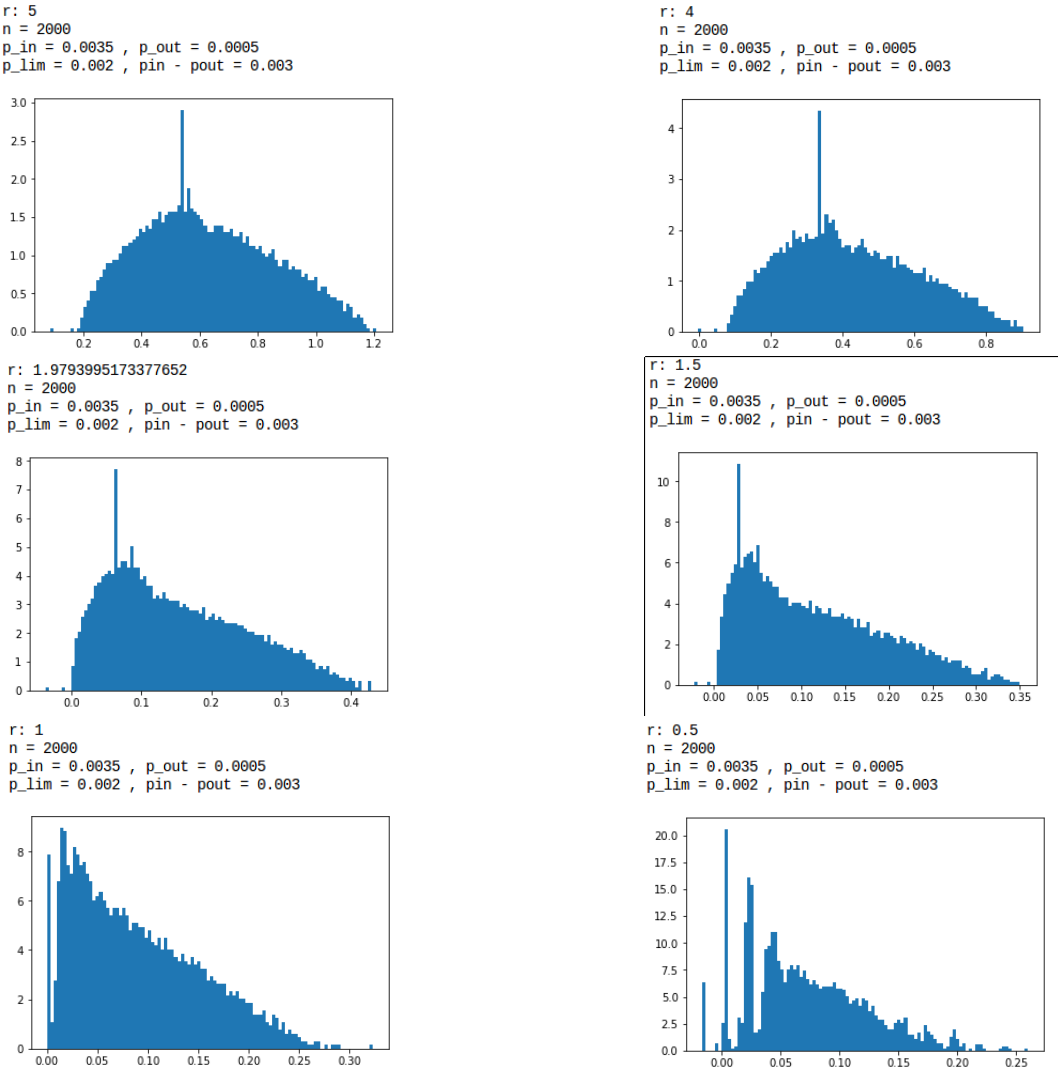
Nous allons dans un premier temps simuler les densités spectrales des matrices Bethe hessiennes avec un SBM avec $q = 2, n_q = 2000, p_{in} = 0.0035, p_{out} = 0.0005$. Avec ces paramètres, le paramètre optimale est $r_{opt} \simeq 2$.

On peut observer que lorsque r est largement supérieur à r_{opt}, on perd la propriété où les valeurs propres contenant l'information sont négatives et celles du bulk sont positives. Par conséquent, on ne peut plus trouver le nombre de communautés dans le graphe. De même lorsque r est largement inférieur à r_{opt}.

On peut voir que le comportement plus la densité spectrale en fonction de r est le suivant :

— plus r est grand plus la moyenne de la densité spectrale augmente

— plus r est proche de 0 plus les valeurs propres portant l'information se rapprochent de 0.



Comme décrite dans la première partie, il faut comparer la partition obtenue de la matrice Bethe Hessian avec celle obtenue via l'algorithme "belief propagation". D'après le figure 2 du papier [1] on a :

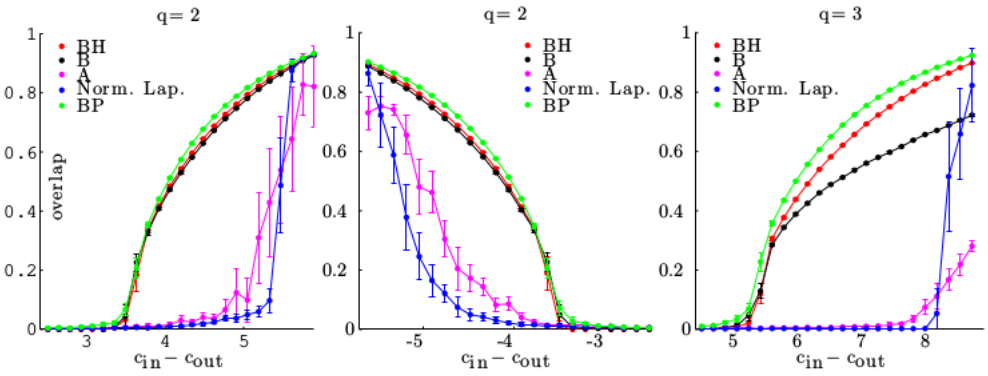


FIGURE 9 – graphe généré à partir d'un SBM : $n = 10^5$

Ce sur cette figure il y a en ordonnée la performance de chaque algorithme via une mesure du chevauchement. La figure à gauche correspond au cas avec 2 communautés et une structure de communauté "assortative". La figure au milieu correspond aux cas avec 2 communautés et une structure de communauté "assortative". La figure à droite correspond au cas avec 3 communautés et une structure de communauté "assortative".

A Valeurs propres obtenues via algèbre linéaire

Dans cette annexe nous allons trouver les valeurs propres de l'Équation 15.
On a

$$\langle A \rangle = \theta_q \otimes \mathbf{1}_{n_q} \mathbf{1}_{n_q}^t$$

avec θ_q une matrice $q \times q$, dont les éléments diagonaux sont égaux à a et les autres sont égaux à b, $\mathbf{1}_{n_q}$ le vecteur colonne rempli de 1 et de taille n_q, et $\otimes$ est le produit de Kronecker.
On cherche les valeurs propres de $\langle A \rangle$:

$$\begin{aligned} &\Leftrightarrow |\langle A \rangle - \lambda I_n| = 0 \\ &\Leftrightarrow det(\theta_q \otimes \mathbf{1}_{n_q} \mathbf{1}_{n_q}^t - \lambda I_n) = 0 \end{aligned}$$

On a

$$det(\langle A \rangle) = \begin{vmatrix} B & C & \dots & \dots & C \\ C & \ddots & & & \vdots \\ \vdots & & & \ddots & C \\ C & \dots & \dots & C & B \end{vmatrix}$$

Où $B = a\mathbf{1}_{n_q} \mathbf{1}_{n_q}^t - \lambda I_{n_q}$ et $C = b\mathbf{1}_{n_q} \mathbf{1}_{n_q}^t$.

On a la formule de déterminant par blocs suivante :

$$det \begin{pmatrix} A & 0 \\ C & D \end{pmatrix} = det(A) det(D)$$

et en utilisant les combinaisons linaires lignes/colonnes suivantes $q - 2$ fois

$$\begin{aligned} (L_n \leftarrow L_n - L_{n-1}) \\ (C_{n-1} \leftarrow C_{n-1} + C_n) \end{aligned} \tag{24}$$

on obtient

$$\begin{aligned} det(\langle A\rangle) &= \begin{vmatrix} B & (q-1)C \\ C & B+(q-2)C \end{vmatrix} |B-C|^{q-2} \\ &= F_{n_q} G_{n_q}^{q-2} \end{aligned}$$

Commençons par factoriser le polynôme $G_{n_q}^{q-2}$. En utilisant la même stratégie que (24), on obtient :

$$G_{n_q} = (-\lambda)^{n_q-1}(\lambda - n_q(a-b))$$

Même stratégie pour F_{n_q} et H_{n_q} qui suit, on obtient :

$$\begin{aligned} F_{n_q} &= (-\lambda)^{n_q-1} H_{n_q} \\ H_{n_q} &= \begin{vmatrix} A & n_q(q-1)\mathbf{1}_{n_q} \\ b\mathbf{1}_{n_q}^t & n_q(a+(q-2)b)-\lambda \end{vmatrix} \\ &= (-\lambda)^{n_q-1} E_{n_q} \\ &= (-\lambda)^{n_q-1}(\lambda - n_q(a+(q-1)b))(\lambda - n_q(a-b)) \end{aligned}$$

On a donc au final

$$\boxed{det(\langle A\rangle - \lambda I_n) = (-\lambda)^{q(n_q-1)}(\lambda - n_q(a-b))^{q-1}(\lambda - n_q(a+(q-1)b))} \tag{25}$$

Donc les trois valeurs propres de la matrice $\langle A\rangle$ sont 0, $n_q(p_{in}+(q-1)p_{out})$, $n_q(p_{in}-p_{out})$ de multiplicités respectives $q(n_q-1)$, 1, $q-1$.

Cinquième partie
Conclusion

Bilan

Nous avons réussi à redémontrer les résultats de l'article [7]. L'intérêt de ces nouveaux résultats réside dans la nouvelle limite de cet algorithme de détection de communauté que nous avons trouvé. En effet, cette limite est largement citée et réutilisée dans les articles de "graphe spectral clustering".

Une perspective serait de réétudier les quelques articles qui utilisent comme base l'ancienne limite et ainsi corriger les résultats dont ils dérivent.

Nous avons aussi étudié d'autres algorithmes spectraux plus fins doté d'un performance supérieur. Les recherches récentes dans ce domaine se tournent principalement sur la "Non-Backtracking Matrix". Ceci est dû à ces propriétés plus intéressantes que celle des autres opérateurs linéaires (e.g matrice d'adjacence). Mais aussi et surtout grâce au fait qu'elle à des connections avec des théories en physique comme les modelés d'Ising.

Remerciement

Je tiens à remercier Jamal Najim pour m'avoir aiguillé et soutenu tout au long de ce stage. Ses conseils ont été essentiels à l'aboutissement de ce mémoire.

Références

[1] Lenka Zdeborov Alaa Saade, Florent Krzakala. Spectral clustering of graphs with the bethe hessian. 2014.

[2] Marc Lelarge Lenka Zdeborov Alaa Saade, Florent Krzakala. Spectral detection in the censored block model. 2015.

[3] Djalil Chafaï. Introduction aux matrices aléatoires. 2013.

[4] Elchanan Mossel Joe Neeman Allan Sly Lenka Zdeborová Florent Krzakala, Cristopher Moore and Pan Zhang. Spectral redemption : clustering sparse networks. 2013.

[5] Santo Fortunato. Community detection in graphs. 2010.

[6] Romain Couillet Hafiz Tiomoko Ali. Improved spectral community detection in large heterogeneous networks. 2018.

[7] Raj Rao Nadakuditi and M. E. J. Newman. Graph spectra and the detectability of community structure in networks. 2012.