

Encadrant :

MAI Xiaoyi

MEJEAN Alexandre

MAISON Jonas

|  |
| --- |
| CentraleSupélec Campus de Gif, Année 2016-2017  Rapport projet de conception N°SS403 :  Méthodes spectrales pour la détection de groupes d’affinités dans les grands réseaux |

GitHub : <https://github.com/Jonas1312/CommunityDetection>

Sommaire

[Sommaire 2](#_Toc477173736)

[Introduction 3](#_Toc477173737)

[Génération de graphes aléatoires 4](#_Toc477173738)

[Algorithmes spectraux 7](#_Toc477173739)

[Choix de la matrice d’affinité 8](#_Toc477173740)

[Matrice laplacienne 8](#_Toc477173741)

[Matrice de modularité 8](#_Toc477173742)

[Matrice de Bethe-Hess 9](#_Toc477173743)

[Analyse des résultats 10](#_Toc477173744)

[Conclusion 10](#_Toc477173745)

[Références 11](#_Toc477173746)

Introduction

Les graphes permettent de décrire une multitude de systèmes complexes issus de divers domaines : réseaux sociaux, moteurs de recherche, réseaux biologiques, circuits électroniques, jeux en réseaux et encore bien d’autres applications. On modélise en général ces graphes par un ensemble de nœuds qui représentent les entités du réseau ainsi que par des arêtes qui représentent les relations entre ces entités. L’existence de communautés dans un réseau correspond à la présence de certains groupes de nœuds qui sont très fortement connectés entre eux par rapport aux autres nœuds du graphe. On appelle partitionnement de graphe la détection de communautés non chevauchantes.

Il existe actuellement dans la littérature une multitude de travaux portant sur la détection de communautés dans les graphes et ce domaine d’étude est en pleine expansion depuis la création des premiers réseaux sociaux. La plupart de ces méthodes consistent à déterminer une partition des sommets du graphe optimisant un certain critère de qualité d’un partitionnement. Ce problème d’optimisation est NP-complet, et il est donc difficile d’obtenir des résultats en un temps raisonnable sur des graphes de plusieurs milliers de nœuds.

Dans le cadre de notre projet nous avons abordé les algorithmes spectraux, qui sont aussi très utilisés pour la détection de communauté. Le partitionnement spectral utilise les vecteurs et les valeurs propres d’une matrice d’affinité définie à partir d’un graphe et réalise ensuite de la classification non-supervisée des vecteurs propres dominants pour trouver des partitions.

Nous aborderons donc dans un premier temps la génération de graphes aléatoires à l’aide du « stochastic block model », puis nous présenterons les diverses matrices d’affinités que nous avons utilisé et plus particulièrement la matrice appelée Bethe Hessian, et qui a la particularité de donner de meilleurs résultats sur les graphes sparses. Enfin, nous étudierons les performances de nos algorithmes et comparerons les résultats des diverses matrices d’affinités.

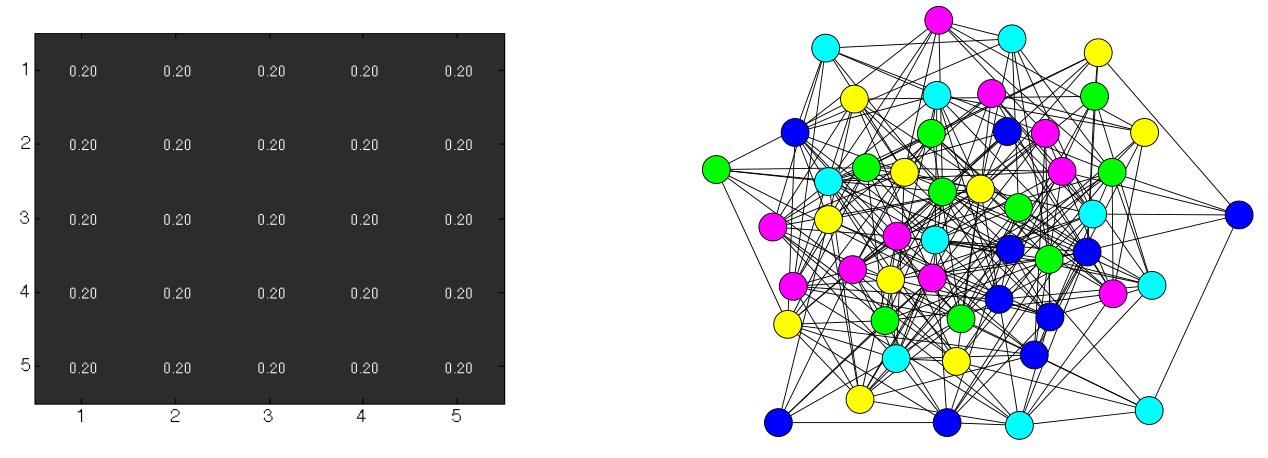
Génération de graphes aléatoires

Le « stochastic block model » est un modèle de génération de graphes aléatoires très connu et utilisé dans la littérature. Il prend en paramètre :

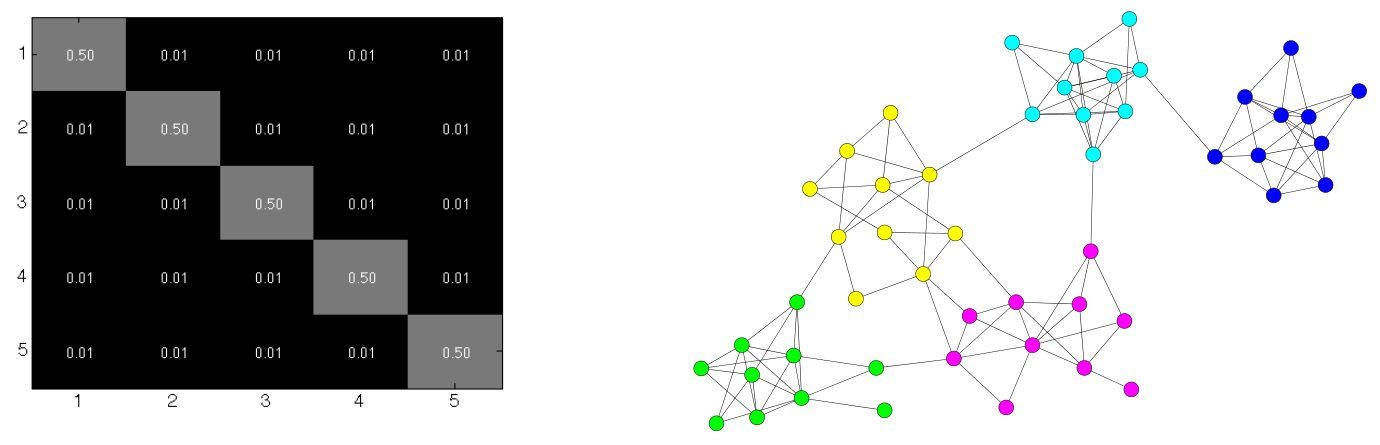
* Le nombre de nœuds du graphe à générer
* Une matrice symétrique de probabilités de taille
* Une partition des nœuds du graphe (où est la i-ème communauté) telle que :

La probabilité pour que deux nœuds et soient reliés par une arête est donc définie comme étant égale à .

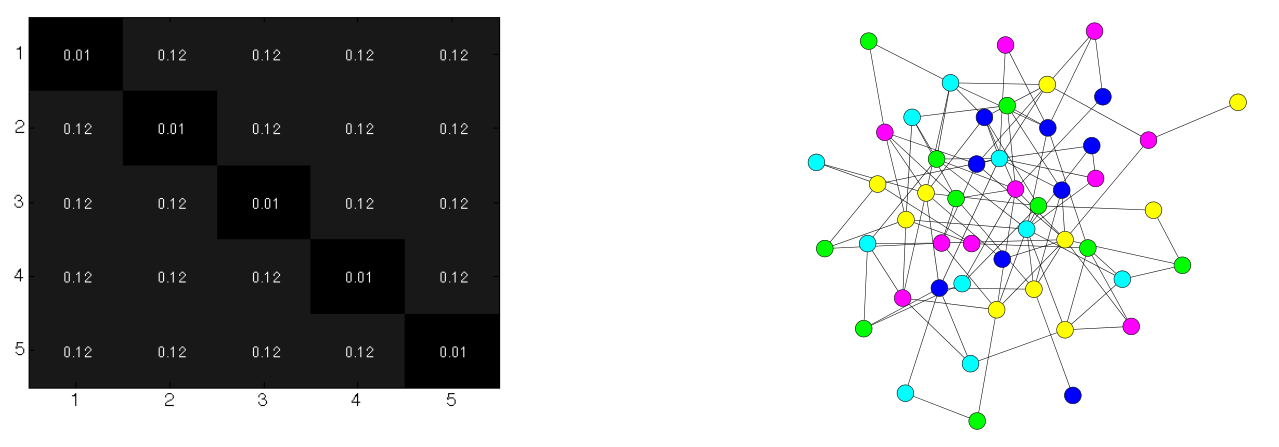
En modifiant la matrice de probabilité on peut donc obtenir des graphes avec certaines propriétés.



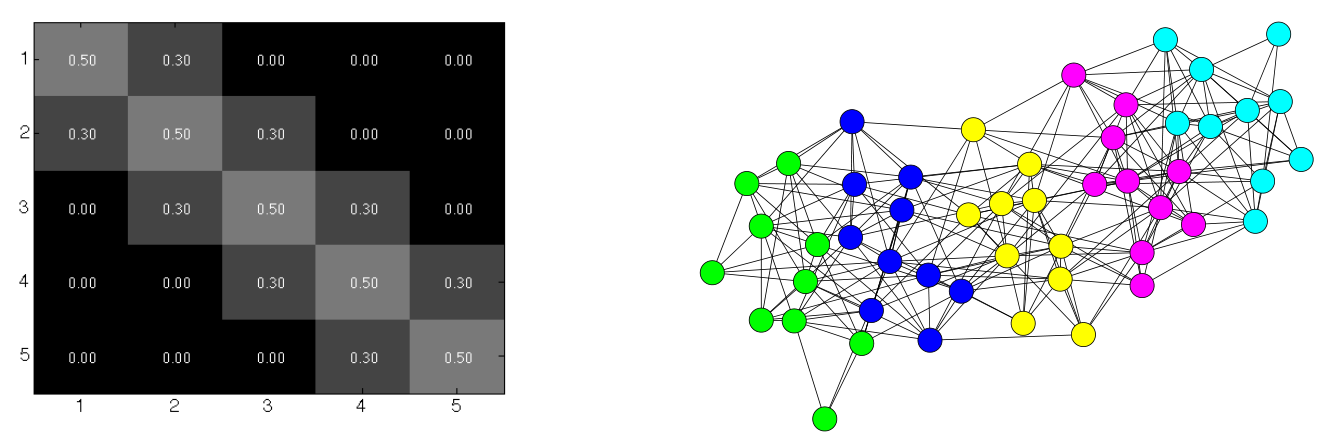
*Figure 1 : Graphe purement aléatoire, distribution uniforme*



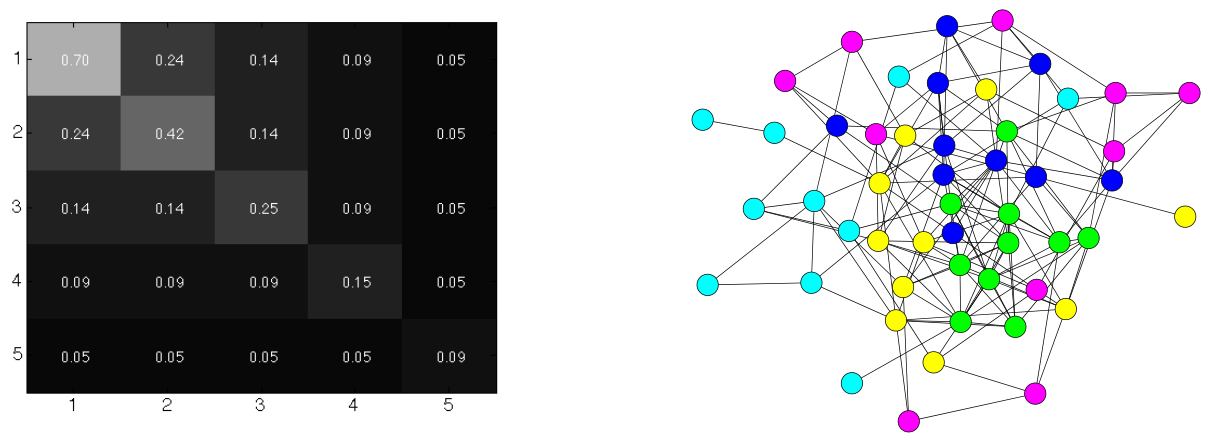
*Figure 2 : « Assortative model », on distincte bien les communautés, les nœuds intra-communautaires sont fortement connectés entre eux*



*Figure 3 : « Dissortative model », les communautés sont difficilement distinguables, les nœuds intra-communautaires sont faiblement connectés entre eux*



*Figure 4 : Communautés ordonnées*



*Figure 5 : « Core-periphery structure », avec une communauté centrale (verte)*

Pour la suite du rapport nous allons préciser quelques définitions sur la théorie des graphes :

* La matrice d’adjacence d’un graphe de taille est définie telle que :

Si le graphe n’est pas orienté la matrice d’adjacence est alors symétrique. Nous aborderons dans ce rapport uniquement le cas des graphes non orientés.

* Le degré d’un nœud est défini comme étant le nombre d’arêtes qui lui sont incidents. On a alors :
* Dans le cas des graphes générés par le stochastic block model, on montre que l’espérance du nœud de degré appartenant à la communauté est égale à :

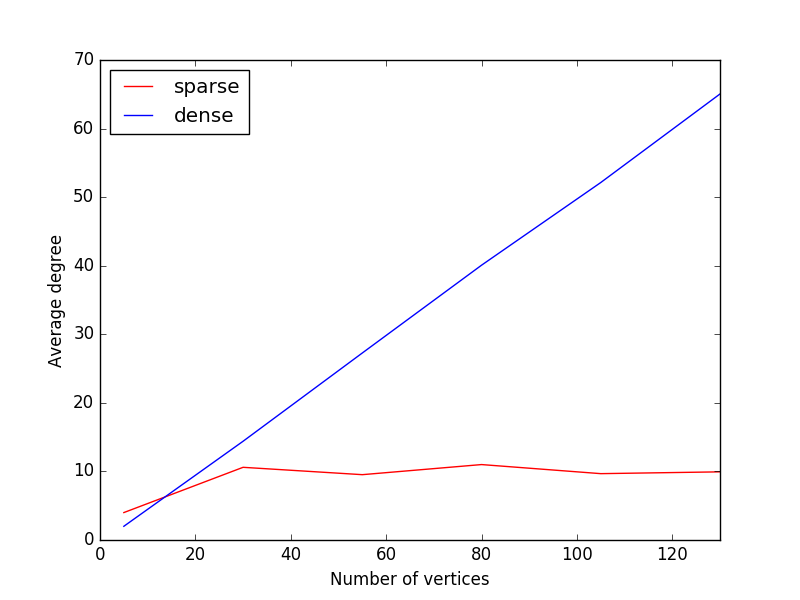
Sous forme matricielle on a :

Où est le nombre de nœud dans le communauté .

On distingue en général deux types de graphe :

* Les graphes denses, pour lesquels le degré moyen des nœuds augmente lorsque le nombre de nœuds augmente
* Les graphes « sparses », pour lesquels le degré moyen des nœuds reste constant même si le nombre de nœuds augmente.

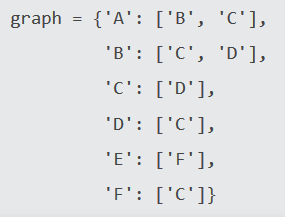
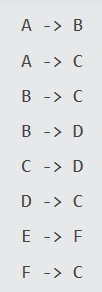
Pour générer un graphe sparse à partir du stochastic block model il suffit de diviser la matrice de probabilité par le nombre de nœud du graphe désiré.



*Figure 6 : Comparaison entre graphe sparse et graphe dense*

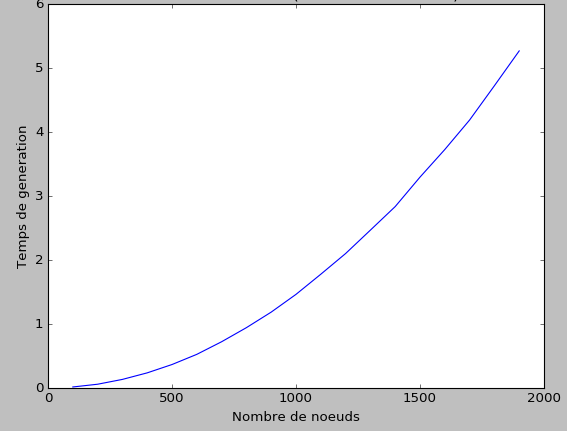
Nous avons fait le choix d’implémenter nos algorithmes en langage Python, qui présente l’avantage d’être très accessible et qui possède un grand nombre de bibliothèques permettant la visualisation de graphes.

La matrice d’adjacence d’un graphe pouvant être très grande et aussi très creuse, on utilise en général le format de donnée « dictionnaire » propre au langage Python, qui utilise des tableaux non pas indexés par des numéros mais par des clés pouvant être de n’importe quelle nature, par exemple un nœud d’un graphe.



*Figure 7 : Représentation symbolique d’un graphe à gauche et équivalent sous forme dictionnaire Python à droite*

Utiliser ce type de structure de données à l’avantage d’être très rapide à générer et consomme peu de mémoire. Cependant, pour les besoins de notre projet nous n’avons pas utilisé ce type de structure car les matrices d’affinités qui nous serons utiles par la suite doivent être calculées à partir de la matrice d’adjacence, il est donc plus simple de stocker la matrice d’adjacence directement dans la mémoire sous forme « d’array » classique.



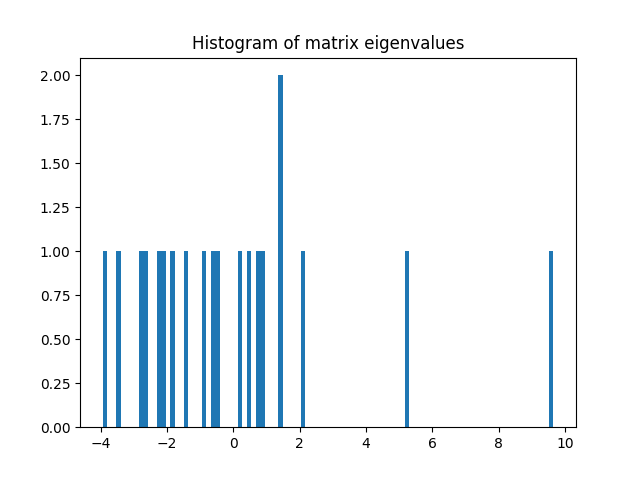
*Figure 8 : Temps de génération (sec) d’un graphe par le stochastic block model en fonction du nombre de nœuds*

On obtient finalement une complexité quadratique.

Algorithmes spectraux

Les algorithmes spectraux de détection de communautés permettent de créer un partitionnement d’un graphe en temps polynomial. Ces algorithmes se basent sur une méthode de classification non supervisée appliquée sur les vecteurs propres d’une matrice d’affinité définie à partir de la matrice d’adjacence d’un graphe. Dans le cadre de notre projet nous étudierons la matrice Laplacienne, la matrice de modularité, la matrice d’adjacence ainsi que la matrice Bethe-Hessian.

Lorsque l’on trace l’histogramme des valeurs propres d’une matrice d’affinité on constate que certaines valeurs propres sont isolées par rapport aux autres. En réalité, ce sont les K plus grandes ou les K plus petites valeurs propres (selon comment sont définies les matrices), K étant le nombre de communautés dans le graphe.

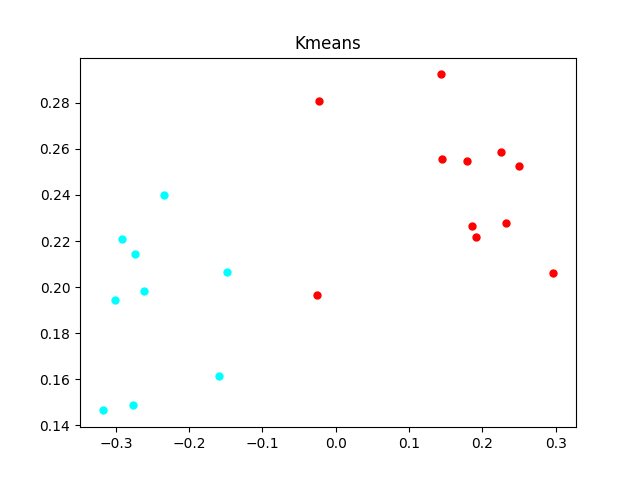


Valeurs propres isolées

*Figure 9 : Histogramme des valeurs propres d’une matrice d’affinité*

On forme alors la matrice X avec les K vecteurs propres associés à ces K valeurs propres :

On applique alors une méthode de classification non supervisée sur les entrées de ces vecteurs propres. Dans notre projet, nous utilisons la méthode la plus simple et la plus répandue dans la littérature : la méthode des K-moyennes.



*Figure 10 : Tracé des deux vecteurs propres et résultat de la classification*

Les lignes de X étant maintenant classées en K communautés, on associe au nœud n du graphe la communauté correspondant à la nième ligne de X.

# Choix de la matrice d’affinité

La littérature propose plusieurs formes de matrices d’affinité pour les algorithmes spectraux. Nous étudierons la matrice Laplacienne, la matrice de modularité, la matrice Bethe-Hessian et la matrice d’adjacence.

## Matrice Laplacienne

La matrice Laplacienne est définie par :

Où :

* A est la matrice d’adjacence du graphe
* I est la matrice identité de même taille que A
* D est la matrice diagonale où le coefficient est égal au degré du sommet i.

## Matrice de modularité

La matrice de modularité est définie par :

Où :

* A est la matrice d’adjacence du graph
* D est la matrice diagonale telle que est égal au degré du sommet i
* narcs est le nombre d’arcs du graphe.

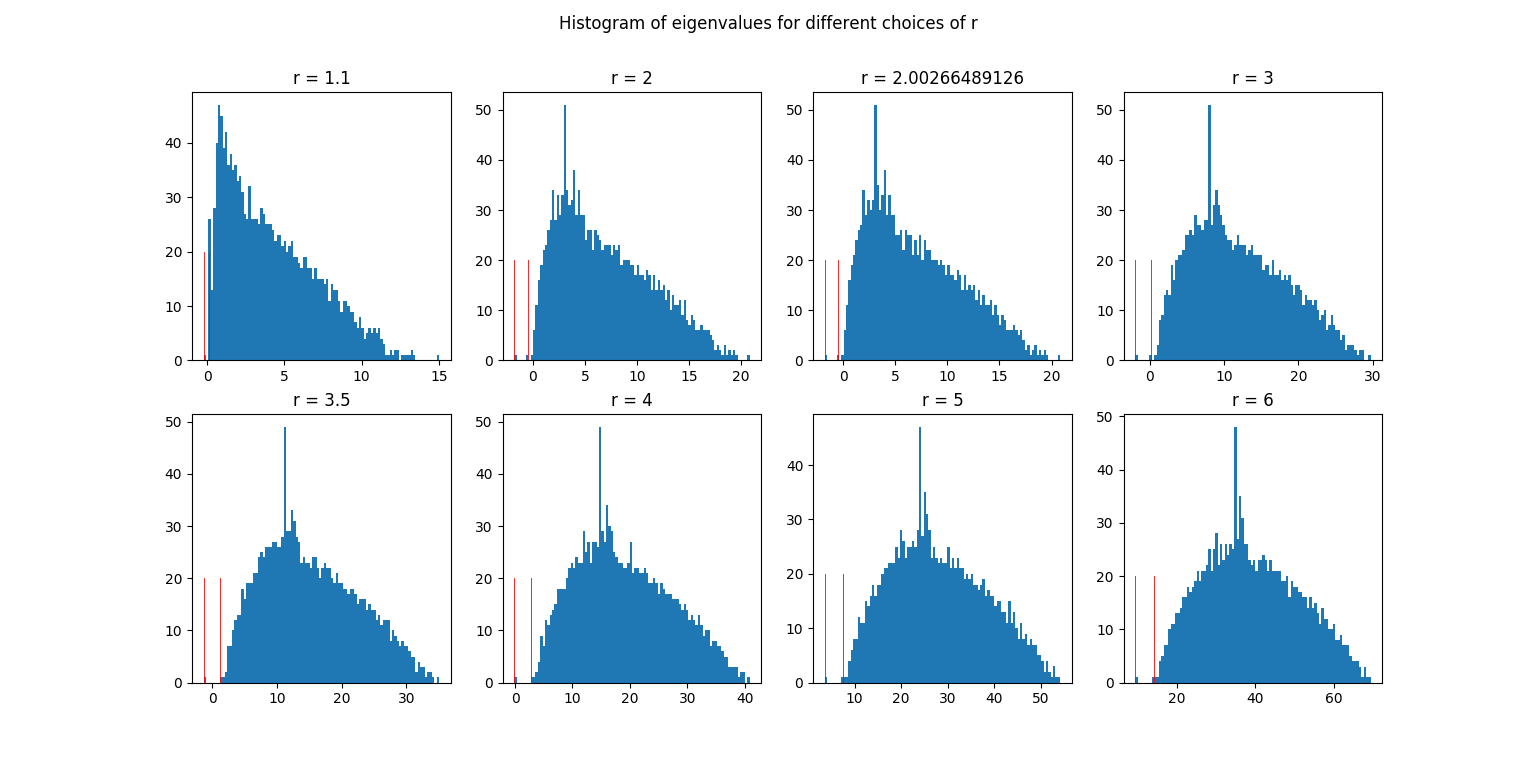
## Matrice de Bethe-Hess

La matrice de Bethe-Hess (ou Bethe-Hessian Matrix) est la matrice définie par :

Où :

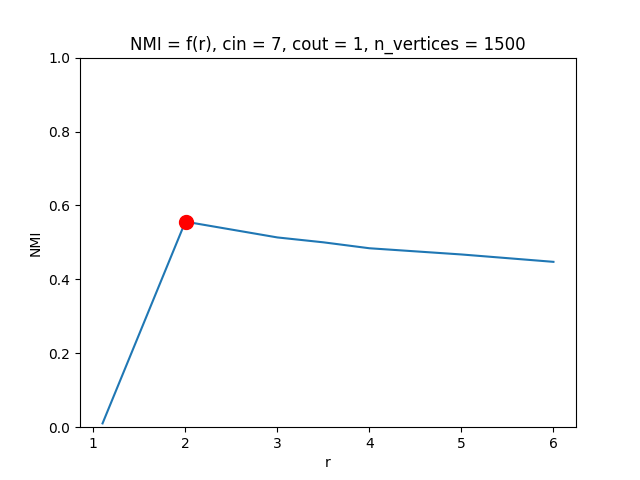
* A est la matrice d’adjacence du graphe
* I est la matrice identité de même taille que A
* D est la matrice diagonale où est égal au degré du sommet i
* r est un paramètre à choisir.

Les histogrammes ci-dessous ont été générés à partir d’un graphe à deux communautés. On cherche donc à ce que les deux valeurs propres les plus faibles, représentées par des traits rouges, soient donc le plus isolées possible. Pour r petit, il est très difficile de distinguer la deuxième valeur propre à isoler, la détection de communautés ne peut donc pas être efficace. Lorsque r augmente, les deux plus petites valeurs propres se distinguent mieux. La publication [4] conseille de prendre r comme étant la racine carrée du degré moyen du graphe et on constate que ce choix est effectivement le meilleur sur les figures 11 et 12, où la racine carré du degré moyen est égale à 2.0027.



*Figure 11 : Histogrammes des valeurs propres pour plusieurs valeurs de r*

L’information mutuelle normalisée (NMI) est une mesure de la dépendance statistique de deux vecteurs aléatoires. Cette quantité permet donc de mesurer la qualité d’un partitionnement en communautés d’un graphe dont on connait un partitionnement de référence (par exemple si le graphe a été généré à l’aide du stochastic block model).



*Figure 12 : Normalized mutual information en fonction de r*

La figure ci-dessus est le tracé de l’information mutuelle normalisée en fonction du choix de r pour l’application de l’algorithme spectral utilisant la matrice de Bethe-Hess sur une matrice générée avec le stochastic bloc model. Le point rouge correspond au r idéal égal à la racine du degré moyen des sommets. Ce tracé confirme donc le choir de r basé sur l’histogramme des valeurs propres. Dans la suite nous n’utiliserons que ce r.

Analyse des résultats

Générateur de permutation, mutual information, tests sur graphs denses et sparses, influence de cin-cout

Conclusion

Comment connaitre n\_cluster et n\_communities

Références

[1] AARON CLAUSET. Network Analysis and Modelling course.

[2] SANTO FORTUNATO. Community detection in graphs. Physics Reports.

[3] RAJ RAO NADAKUDITI AND MARK EJ NEWMAN. Graph spectra and the detectability of community structure in networks. Physical review letters.

[4] ALAA SAADE, FLORENT KRZAKALA, AND LENKA ZDEBOROVA. Spectral clustering of graphs with the Bethe hessian. In Advances in Neural Information Processing Systems.

[5] ROMAIN COUILLET. A random matrix approach to machine learning.

[6] LAURENT MASSOULIE. Community detection with spectral methods.

[7] LUCA DONETTI, MIGUEL A. MUNOZ. Detecting network communities: a new systematic and efficient algorithm.

[8] CAN M. LE. Estimating community structure in networks by spectral methods.

[9] XIAO ZHANG, M. E. J. NEWMAN. Multiway spectral community detection in networks.

[10] MARC LELARGE. Algorithme des réseaux sociaux.

[11] HAFIZ TIOMOKO ALI, ROMAIN COUILLET. Random matrix improved community detection in heterogeneous networks.

[12] EMMANUEL ABBE, COLIN SANDON. Recovering communities in the general stochastic block model without knowing the parameters.

[13] PAN ZHANG. Robust spectral detection of global structures in the data by learning a regularization.

[14] M. E. J. NEWMAN. Spectral methods for network community detection and graph partitioning.

[15] KEEGAN GO, KENJI HATA. Statistical Physics of community detection.