

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](#_Toc477005592)

[Introduction 3](#_Toc477005593)

[Génération de graphes aléatoires 4](#_Toc477005594)

[Algorithmes spectraux 4](#_Toc477005595)

[Analyse des résultats 4](#_Toc477005596)

[Conclusion 4](#_Toc477005597)

[Références 5](#_Toc477005598)

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, et 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 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.

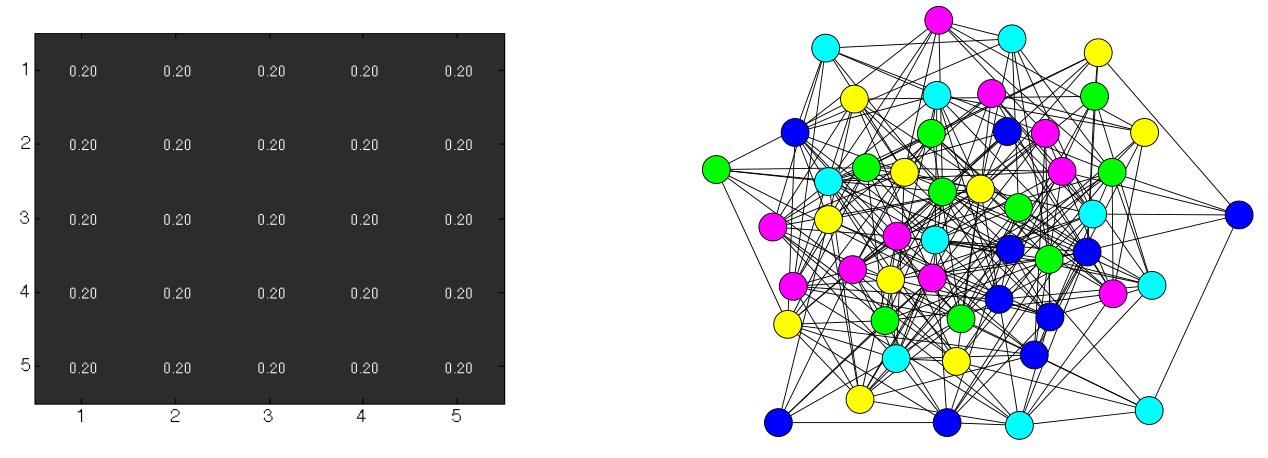


Figure : Graphe purement aléatoire, distribution uniforme

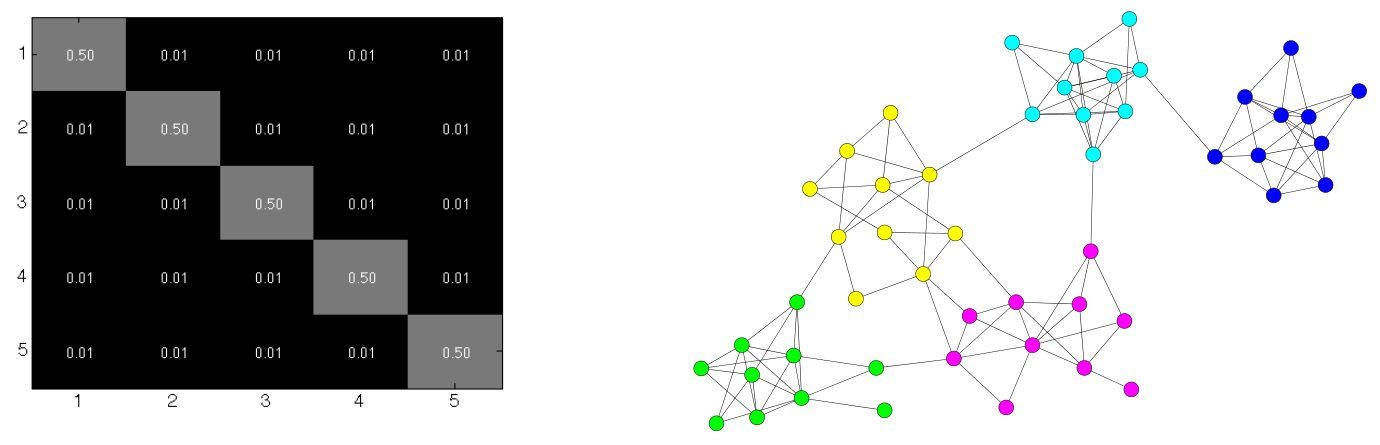


Figure : « Assortative model », on distincte bien les communautés, les nœuds intra-communautaires sont fortement connectés entre eux

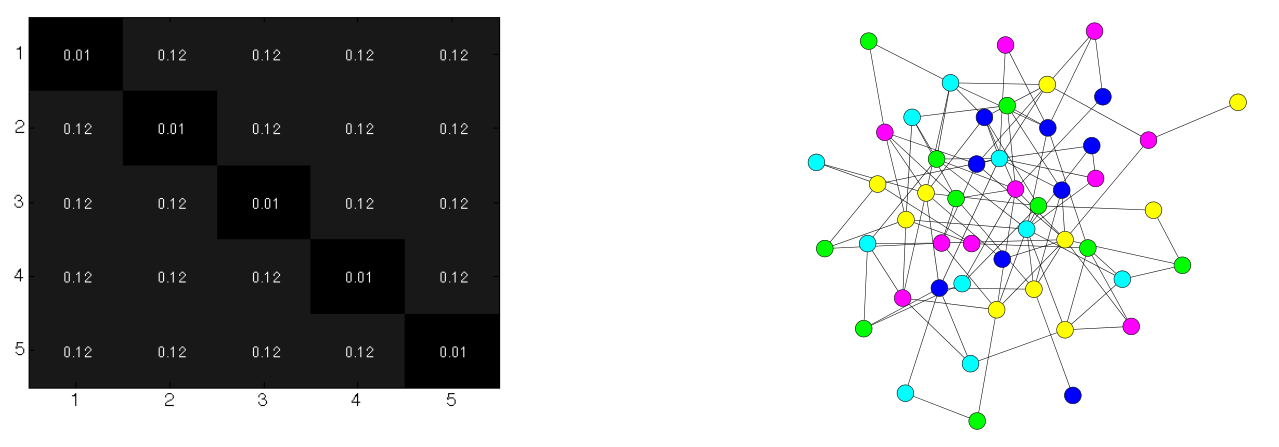


Figure : « Dissortative model », les communautés sont difficilement distinguables, les nœuds intra-communautaires sont faiblement connectés entre eux

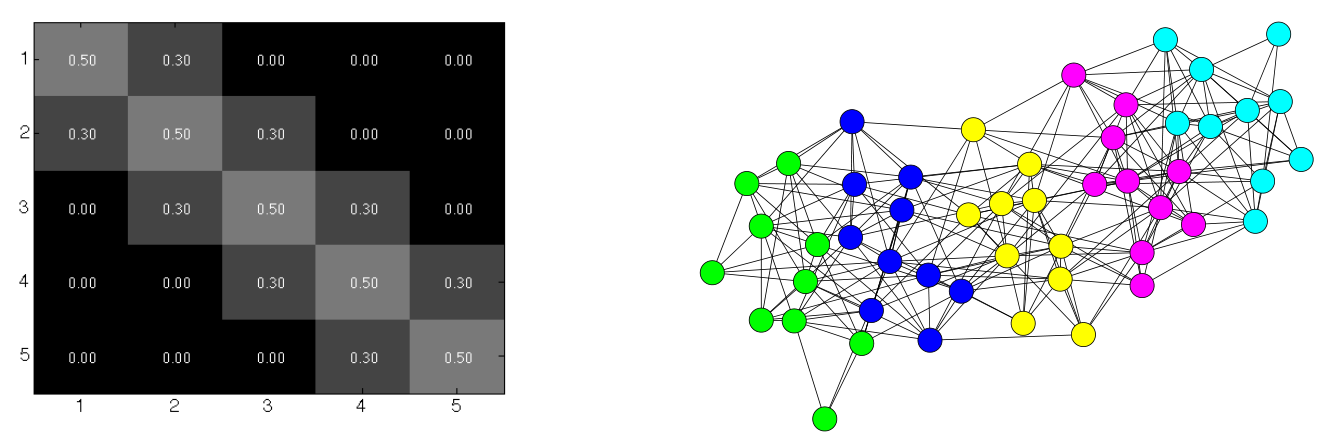


Figure : Communautés ordonnées

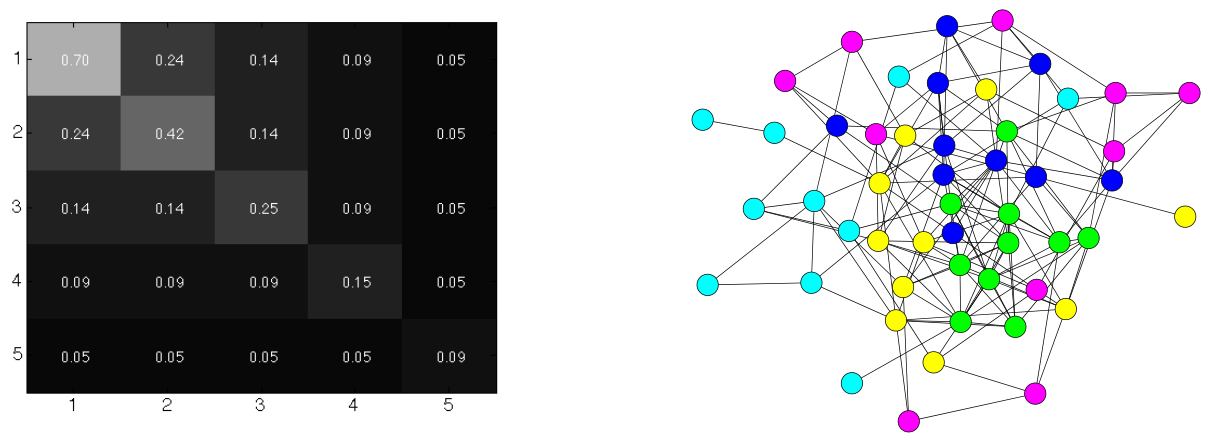


Figure : « Core-periphery structure », avec une communauté centrale (verte)

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.

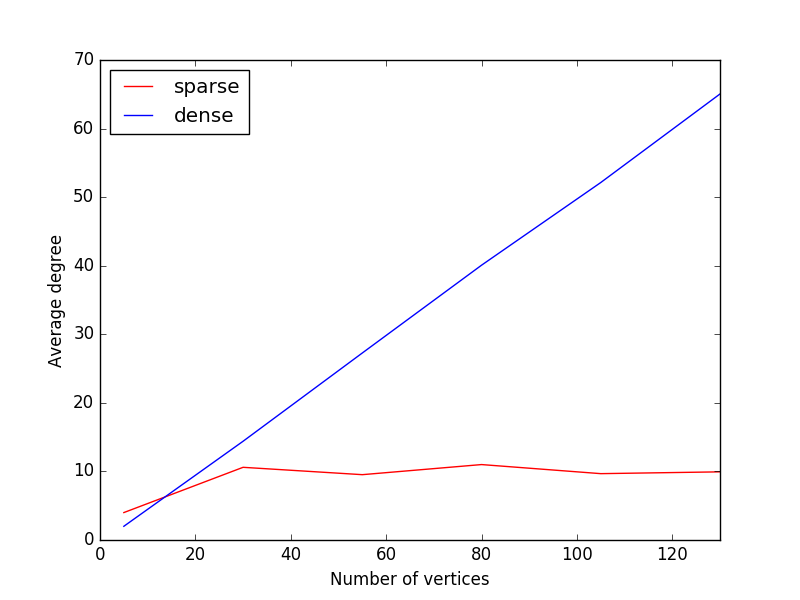


Figure : Comparaison entre graphe sparse et graphe dense

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 quel objet.

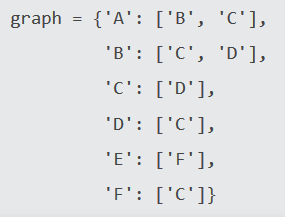
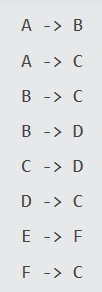


Figure : Représentation symbolique d’un graphe à gauche et équivalent sous forme dictionnaire Python à droite

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. On stockera donc la matrice d’adjacence directement dans la mémoire.

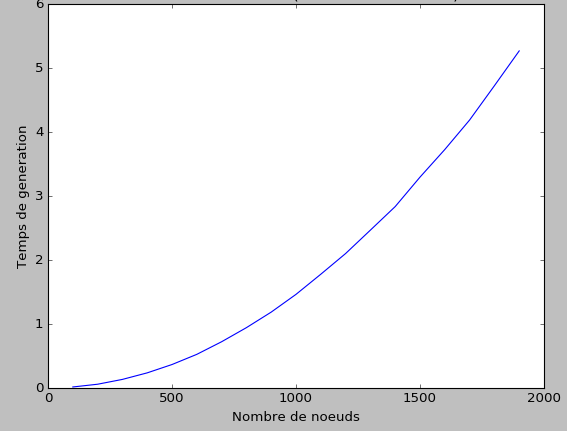


Figure : Temps de génération (sec) d’un graphe par le stochastic block model par rapport au nombre de nœuds

On obtient finalement une complexité quadratique.

Algorithmes spectraux

Matrices utilisées, explication du clustering spectral, histogramme valeurs propre, kmeans, bethe hessian, choix de « r » bethe hessian

Analyse des résultats

Générateur de permutation, mutual information

Conclusion

Références