Travaux dirigés n°1

Le TD est à réaliser par groupe de 2 (au maximum). A la fin du TD, un compte-rendu est attendu, comportant les explications de la démarche, les résultats (figures si besoin), ainsi que des interprétations (étayés par des résultats chiffrés).

Les réponses aux questions s'appuieront sur des analyses réalisées avec le logiciel R. Le compte-rendu doit être fourni sous l'une des formes suivantes :

1. Rmarkdown (depuis Rstudio)
2. Document .pdf ou docx

Sauf dans les cas où c'est explicitement précisé, vous pouvez utiliser les fonctions haut-niveau fournies dans le paquet igraph, paquet R à installer en début de séance.

I. Analyse empirique des algorithmes de détection de communautés

a) Observer la complexité de calcul de deux algorithmes de détection de communautés (algorithmes au choix) en utilisant les fonctions de génération de graphes aléatoires (algorithmes fournis dans la librairies igraph, parmi les fonctions notées cluster_{edge_betweenness, fast_greedy, …}). Expliquer brièvement le fonctionnement des deux algorithmes choisis.

b) Utiliser un graphe fameux fourni dans igraph (graph.famous("Zachary")) et étudier, en utilisant les mesures classiques d'évaluation des communautés, les deux algorithmes choisis précédemment.

II. Centralité

La centralité est une notion fondamentale pour l'analyse des réseaux. Rappelons, pour définir les notions de centralité, qu'une géodésique est le plus court chemin entre deux nœuds. On distingue :

- *The closeness centrality of a vertex measures how easily other vertices can be reached from it (or the other way: how easily it can be reached from the other vertices). It is defined as the number of the number of vertices minus one divided by the sum of the lengths of all geodesics from/to the given vertex. If the graph is not connected, and there is no path between two vertices, the number of vertices is used instead the length of the geodesic. This is always longer than the longest possible geodesic.*
- *The betweenness centrality of a vertex is the number of geodesics going through it. If there are more than one geodesic between two vertices, the value of these geodesics are weighted by one over the number of geodesics.*
- *Edge degree centrality: node with highest degree.*

Données : On s'intéresse à une version simplifiée du réseau résultant des acteurs de films sortis entre 1995 et 2004. A partir des fichiers fournis, on va créer le graphe non-dirigé tel que deux acteurs sont connectés si et seulement si ils ont tourné un film ensemble. On cherche dans cette

partie à déterminer l'acteur le plus « central », mais on ne sait pas quelle définition sera la bonne. Les fichiers sont fournis : imdb_actor_edges.tsv et imdb_actors_key.tsv

a) A partir des fichiers fournis, traiter les données afin d'obtenir le réseau non dirigé tel qu'une arête existe si et seulement si les acteurs ont tourné ensemble.
b) A l'aide de la fonction component_distribution() de igraph, extraire la plus large composante faiblement connexe du graphe. Les calculs suivants se feront uniquement à partir de ce sous-graphe afin que les calculs soient simplifiés.
c) A partir du sous-graphe précédent, extraire les 20 nœuds ayant les degrés les plus grands. Commenter les résultats.
d) A partir du sous-graphe précédent, calculer la betweenness centrality des nœuds et extraire les 20 les plus importants. Qu'observez-vous par rapport au résultat précédent ?
e) Faire de même pour la closeness centrality et expliquer les résultats, notamment tentez d'interprétez les différences par rapport aux deux autres mesures.