Grands Réseaux d'Interaction
TP n° 4 : Exploitation de k-cœurs

Règles Générales

- Les règles générales en TP restent valides, voir feuille du TP 1
- Ce TP est à rendre sur Moodle pour le **27 octobre**

Rôle des k-cœurs

Les k-cœurs présents dans un réseau signalent des sous-ensembles de sommets de ce réseau intéragissant plus spécifiquement entre eux. Identifier les k-cœurs pour de grandes valeurs de k permet donc d'isoler des parties du réseau intéressantes à analyser. C'est ce que nous allons faire aujourd'hui.

But de la séance

Le but de la séance est de mettre à l'épreuve tout ce que vous avez programmé jusqu'à maintenant. Si vos codes ne tournent pas encore parfaitement, c'est le moment de finir de débugguer, d'avoir une version fonctionnelle et stable.
Ensuite, le travail à réaliser est d'analyser les jeux de données proposés et d'en tirer des conclusions et interprétations.
Votre soumission doit contenir un répertoire par graphe (jeu de données). Dans ce répertoire

- un fichier par composante fortement connexe extraite du graphe,
- un fichier par k-cœur "intéressant" également extrait,
- un fichier texte (text brut, texte enrichi, PDF, mais **pas** de Word/ODT) expliquant pourquoi les k-cœurs extraits sont "intéressants",
- quelques tentatives d'explication, par exemple : ce que représentent les entités faisant partie des k-cœurs les plus denses, quelques commentaires sur les entités présentes pour certains k mais pas pour les k supérieurs.

La longueur des explications est libre, pas la peine d'écrire un roman cependant.
La note du TP portera principalement sur le fait que vous ayez bien réussi à extraire des composantes fortement connexes ainsi que des k-cœurs. Les justifications et interprétations seront prises en compte dans leur aspect "raisonnement" et non exactitude : l'essentiel est de trouver les k-cœurs importants présents dans le réseau, essayer d'identifier également quels sommets semblent avoir de l'importance, de décrire les relations entre zones (très denses).
Les explications par rapport à la sémantique du réseau peuvent être erronées, ce n'est pas ce qui sera évalué.

I) Protcole expérimental

Le protocole expérimental suivant est suggéré. Néanmoins, vous êtes libres de l'adapter comme bon vous semble, tant que vous en tirez des résultats intéressants.
Il est à réaliser sur chaque graphe.

1. Convertir le graphe en format *dot*, ou le parser directement vers votre programme.

2. Identifiez toutes les composantes fortement connexes du graphe.
3. Choisissez les composantes connexes à évaluser.
   Par exemple, si le graphe est contient une composante connexe géantes ($> 95\%$ voire 99% de la taille totale), comme c'est souvent le cas dans les réseaux sociaux, vous pouvez vous concentrer uniquement sur l'analyse de cette composante. Sinon, décomposez les graphe et analysez chacune des composantes séparément.
4. Trouvez les k-cœurs intéressants dans la/les composante(s) identifiée(s).
   Vous êtes libres de définir ce qui est potentiellement "intéressant" (cf. *But de la séance*), bien évidemment le cœur le plus dense (cf. exercice 3 du TP 3) est un bon début, mais il ne faut pas s'y limiter, essayer d'explorer les *coreness* successives tant qu'elles restent "intéressantes" par rapport à la structure du graphe (vous pouvez penser à utiliser la densité par exemple).
5. Essayez d'apporter des justifications/interpétations (des éléments dans les descriptions des graphes sont là pour vous aider. Voir aussi *But de la séance*).

II) Conversion de format

Les graphes disponibles sur Internet ne sont pas toujours sous format *dot*. La première étape est donc de les convertir vers *dot*. Vous pouvez écrire un petit script qui convertit chaque graphe en format donc. Vous pouvez également écrire à la place un code (le plus générique possible) pour importer le graphe directement dans votre programme.

III) Graphes proposés

Voici des graphes à traiter ci-dessous, d'abord très petit puis de plus en plus gros. Vous pouvez en trouver d'autres sur Internet [1] et les rajouter si vous le souhaitez !

a) Les poules (`poules.dat`)

Le premier graphe décrit les relations de dominance au sein d'un groupe de 32 poules.
Essentiellement pour voir si vos codes écrits jusqu'ici fonctionnent !

Format du graphe : matrice 32×32. Soit deux poules u (ligne) et v (colonne). Une arête $(u \to v)$ signifie que la poule u domine v.
Pistes d'exploitation : quelles sont les poules les plus influentes ?

b) Les élections Wikipédia (`wiki-vote.edges`)

Pour devenir validateur chez Wikipédia, il faut candidater et être élu par les utilisateurs de Wikipédia, qui vous accordent donc leur confiance pour participer à la modération de cette grande œuvre encyclopédique.

Format du graphe : liste d'arêtes. Une arête $(u \to v)$ signifie que u vote pour v.
Pistes d'exploitation : quels sont les modérateurs qui ont été le plus largement élus ? Quelles sont les relations entre les votants (dont beaucoup d'administrateurs déjà en place) et les élus ? Vous pouvez vous aider des données disponibles ici : `https://snap.stanford.edu/data/wiki-Elec.html` pour affiner votre interprétation.

1. Exemple : `http://konect.uni-koblenz.de/networks/`

c) Aéroports américains (`aeroports-us_2010.edges`)

Les réseaux de transports sont des réseaux très étudiés : routiers, ferroviaires, aériens... Le graphe proposé ici est tiré des aéroports américains. Il répertorie les liaisons aériennes existant entre eux.

Format du graphe : liste d'arêtes. Une arête $(u \to v)$ signifie qu'au moins un vol entre l'aéroport u et l'aéroport v a eu lieu en 2010. Les identifiants de u et v correspondent aux codes d'aéroport contenus dans le fichier `aeroports-us_2010_codes.txt` La troisième colonne représente le nombre de places total offertes sur le vol en question, mais nous ne nous y intéresseront pas ici (elle est à ignorer lors de votre conversion de fichier).
Pistes d'exploitation : quels sont les plus gros hubs (plateforme de correspondance aéroportuaire) des Etats-Unis ? Quels sont les aéroports réalisant le plus de traffic. Cela peut-il être corrélé avec des données démographiques concernant les Etats-Unis ?