Mini-projet de TP : analyse de reseaux reels

Systemes complexes avances

M1 Informatique - semestre d'automne 2023-2024

Université Côte d'Azur

Christophe Crespelle
christophe.crespelle@univ-cotedazur.fr

Le but de ce mini-projet de TP est d'analyser deux reseaux complexes provenant de donnees reelles : un de petite taille (entre 3K et 12K noeuds) et un de plus grande taille (entre 35K et 105K noeuds). Vous trouverez les deux reseaux qui vous ont ete attribues, ainsi qu'un lien pour telecharger les donnees brutes, sur la page `https://www.i3s.unice.fr/~ccrespelle/enseignements/22-23-aut_M1SC/TP/TP1/affectation-reseaux.html`.

Travail a effectuer.

La premiere chose que vous devez faire de filtrer les donnees brutes de ces deux reseaux et d'en calculer les statistiques de base (comme fait lors du TP1). Ensuite, vous choisirez une ou deux directions d'approfondissement que vous developperez en menant les analyses necessaires. Ces approfondissements doivent tenter de repondre a une ou des questions que vous choisirez de vous poser sur la structure de ces deux reseaux. Je vous propose ci-dessous plusieurs themes d'approfondissement possibles. Vous pouvez en choisir un parmi eux, ou bien mixer et croiser les questions entre plusieurs des themes proposes ou, encore mieux, choisir une direction que vous aurez vous meme imaginee.

Rendu.

Votre rendu sera constitue du rapport que vous rendrez et qui comprendra les resultats de vos analyses. Ce rapport sera obligatoirement au format PDF, il fera au plus 10 pages (tout compris : titre, references, etc.), avec une police de caractere d'au moins 11pt. Il pourra contenir des annexes qui seront lues a la discretion du correcteur. Mon conseil est que vous selectionnez les analyses qui donnent les resultats les plus interessants parmi celles que vous aurez faites et que vous mettiez les autres en annexe. **Vous devez formuler explicitement et tres clairement les questions auxquelles tentent de repondre vos analyses** et decrire ce que contiennent, objectivement, les courbes que vous tracez. Portez un soin particulier a la lisibilite des legendes des axes. Vous devez aussi produire une analyse critique des resultats mis en evidence par ces courbes. Enfin, pour chaque statistique que vous calculez, vous devez donner en annexe les outils, les lignes de commandes ou le code que vous avez utilises pour produire ces statistiques. Si vous avez ecrit du code, envoyez les fichiers sources en plus du rapport.

1 Prise en main et filtrage des donnees

La premiere etape de votre projet consistera a filtrer les donnees brutes de vos deux reseaux et a les mettre au format decrit dans le TP1. Vous realiserez toutes les analyses demandees dans le TP1 : calculez les quatres proprietes fondamentales de vos reseaux et comparez les a celles des modeles d'Erdos-Renyi et de configuration de memes parametres. Vous pouvez, si vous le souhaitez, vous affarnchir du calcul de la distribution du coefficient de clustering local par noeud (Question 9). Si vous rencontrez des difficultes de calcul pour la distance moyenne, pensez que vous pouvez vous contenter de faire de l'echantillonnage (`https://www.youtube.com/watch?v=E0-_aSMkwGw`).

Vous trouverez ci-dessous plusieurs suggestions d'approfondissements possibles. Vous devez explorer au moins une direction d'approfondissement, ou deux. Mais vous n'etes pas obliges de la choisir parmi celles proposees. Vous pouvez a la place croiser les questions des differentes directions proposees ou proposer vous meme une autre direction. N'hesitez pas, l'originalite des analyses menees sera fortement valorisee dans la notation.

2 Distances

Cette partie propose d'approfondir les proprietes metriques du reseau, c'est a dire des distances et des chemins dans le reseau. Vous pourrez, par exemple, vous poser les questions suivantes.

Question 1. Calculez la distribution des distances dans le reseau.

Definition 1 (Excentricite)**.** *L'excentricite d'un sommet u est le maximum des distances entre u et chacun des autres sommets du graphe : $exc(u) = \max_{v \in V \setminus \{u\}} \{dist(u, v)\}$.*

Question 2. Calculez la distribution de l'excentricite d'un sommet.

Question 3. Quelle est l'excentricite minimum d'un sommet, combien de sommets realisent ce minimum ?

Question 4. Combien de couples realisent le diametre ? Combien de sommets sont impliques dans un tel couple ?

Question 5. Quel est le nombre minimum de sommets a enlever du graphe pour faire chuter le diametre de 1 ? de k ?

3 Centralites

Une possibilite est d'etudier et de comparer differentes metriques de centralite dans le reseau.

Question 6. Calculez au moins deux notions de centralite des noeuds et etudiez les correlations entre leurs scores.

Question 7. Etudiez les correlations des classements qu'elles donnent sur les noeuds.

Question 8. Sur quel(s) type(s) de noeuds sont elles le plus d'accord ? le moins d'accord ?

Question 9. Pour les mesures definies aussi sur les liens (betweeness, eigenvector, PageRank), etudiez la correlation entre la centralite d'un lien et celles de ses deux noeuds incidents.

4 Communautes

Une autre direction possible est l'etude de la structure communautaire du reseau.

Question 10. Comparez le resultat de differentes methodes de partition en communautes. Sont elles ressemblantes ? Où sont elles differentes ?

Question 11. Etudiez la stabilite de la methode de Louvain lorsqu'on change l'ordre dans lequel sont traites les sommets. Les differentes partitions obtenues sont-elles ressemblantes ? Où sont elles identiques ? Où sont elles differentes ?

Question 12. Etudiez la structure du reseau entre les communautes. Combien d'aretes y a-t-il entre les communautes ? Combien a l'interieur des communates ? Quelle est la distribution du nombre d'aretes entre deux communautes ? La distribution du degre d'une communaute ?

Question 13. Etudier les interfaces entre les communautes. Quels sont les noeuds a l'interface de plusieurs communautes ? Y en a-t-il beaucoup ? Quelles sont leurs caracteristiques ?

Question 14. Analysez la structure communautaire a l'interieur de la plus grosse communaute.

5 Coeur/peripherie du reseau

La question poursuivie ici est de degager la structure du coeur (k-core en anglais) et de la peripherie du reseau et de savoir en quoi celles-ci different de celle du reseau tout entier.

Definition 2 (voir `https://en.wikipedia.org/wiki/Degeneracy_(graph_theory)`)**.** *Le k-coeur d'un reseau est l'ensemble des noeuds qui restent apres avoir retire iterativement tous les noeuds de degre au plus $k-1$, en retirant aussi ceux qui deviennent de degre au plus $k-1$ au cours de ce procede. Le coeur du reseau est le k-coeur obtenu pour le k maximum tel que le k-coeur est non vide.*

Question 15. Determinez les k-coeurs du reseau. Comment evolue le nombre de noeud dans le k-coeur en fonction de k ?

Question 16. Ou se situent les aretes du reseau par rapport a ses differentes couches de k-coeur ? On pourra par exemple :
— faire la ditribution de l'ecart dans ces couches entre les deux extremites d'une arete,
— tracer l'evolution en fonction de k du nombre d'aretes dans le k-coeur,
— tracer l'evolution en fonction de k du nombre d'aretes incidentes a un noeud du k-coeur.

Le coefficient de *rich club* (voir `https://en.wikipedia.org/wiki/Rich-club_coefficient`) mesure a quel point les noeuds de fort degre sont connectes entre eux.

Question 17. Determinez le coefficient de *rich club* du reseau.

Question 18. Y a-t-il un rapport entre coeur et rich club ?

Question 19. Le coeur ou le rich club ont-ils la meme structure que le reseau en entier ?

Question 20. Quelles sont les proprietes de la peripherie du reseau, c'est a dire la partie qui n'est pas dans le coeur ?

6 Prevision de liens

On pourra egalement s'interesser a la structure des liens les plus predictibles dans le reseau et aux rapports possibles avec d'autres proprietes structurelles du reseau comme ses communautes, son coeur, la centralite de ses neouds, etc.

Question 21. Utilisez une methode de prediction de liens pour identifier les couples de noeuds les plus enclins a former une nouvelle arete. Où sont situes ces couples par rapport aux communautes et au coeur du reseau ?

Question 22. Classez les liens existant dans le reseau par niveau de predictibilite decroissant. Pour cela, retirez un lien du reseau et calculez le score qui lui est attribue par la methode de prevision.

Question 23. Comment evoluent les proprietes du reseau lorsqu'on ajoute ses liens un par un par predictibilite decroissante ? Le reseau des liens les plus predictibles a-t-il la meme structure que le reseau tout entier ?

Question 24. Y a-t-il un rapport entre predictibilite des liens du reseau et importance des noeuds auxquels ils sont rattaches ?