A supprimer pour les figures

Projet NSGL — Network Science and Graph Learning

Tanguy CESAR

Introduction

Ce projet vise à analyser des réseaux sociaux issus du jeu de données *Facebook100*, en mobilisant des outils classiques de **network science** ainsi que des méthodes de **graph learning**. Les expériences portent sur l'analyse structurelle des graphes, l'étude de l'homophilie, la prédiction de liens, la propagation de labels et la détection de communautés.

Question 1 — Analyse descriptive des réseaux sociaux

Cette première analyse porte sur les réseaux sociaux issus du jeu de données *Facebook100*, en se limitant à la plus grande composante connexe de chaque graphe. L'objectif est de caractériser leurs propriétés structurelles globales et de vérifier s'ils présentent des caractéristiques typiques des réseaux sociaux réels.

Les distributions de degré mettent en évidence une forte hétérogénéité : la majorité des sommets possède un faible nombre de connexions, tandis qu'une minorité est très fortement connectée. Les représentations en échelle log-log des histogrammes et des CCDF révèlent des queues lourdes, indiquant que ces réseaux ne peuvent pas être assimilés à des graphes aléatoires homogènes.

Par ailleurs, les coefficients de clustering, tant globaux que locaux, sont élevés, traduisant une forte fermeture triadique. L'analyse du clustering local en fonction du degré montre une relation décroissante, les nœuds faiblement connectés appartenant à des groupes très cohésifs, tandis que les nœuds de fort degré jouent davantage un rôle d'interconnexion entre communautés.

Enfin, l'assortativité par degré est globalement positive, ce qui suggère que les individus très connectés ont tendance à se lier préférentiellement entre eux. Ces résultats sont cohérents avec la littérature sur les réseaux Facebook universitaires et confirment que les graphes étudiés présentent les propriétés structurelles classiques des réseaux sociaux.

Question 2 — Analyse de réseaux sociaux

Les réseaux étudiés dans cette partie correspondent aux universités de *Caltech*, *MIT* et *Johns Hopkins*, issues du jeu de données *Facebook100*. Pour chacun d'eux, l'analyse est menée sur la plus grande composante connexe afin d'éviter les effets liés aux sommets isolés et de se concentrer sur la structure principale du réseau.

Distribution des degrés

Les distributions de degré mettent en évidence une forte hétérogénéité dans les trois réseaux considérés. La majorité des sommets possède un nombre limité de connexions, tandis qu'un petit nombre de nœuds présente des degrés très élevés. La représentation en échelle log-log de la fonction de distribution cumulative complémentaire (CCDF) révèle la présence de queues lourdes, indiquant que ces réseaux ne suivent pas une distribution homogène du type graphe aléatoire d'Erdős–Rényi.

Malgré des tailles différentes, les réseaux de Caltech, MIT et Johns Hopkins présentent des distributions qualitativement similaires. Cette invariance suggère l'existence de mécanismes de formation communs, tels que l'homophilie et l'attachement préférentiel, fréquemment observés dans les réseaux sociaux en ligne.

Clustering et densité

Les réseaux sont caractérisés par une densité très faible, ce qui est attendu pour des graphes sociaux de grande taille où le nombre de relations effectives reste faible comparé au nombre de relations possibles.

Question 3 — Assortativité et homophilie

Nous mesurons l'assortativité des graphes selon différents attributs : statut, spécialité (major), résidence (dorm), genre et degré.

FIGURE 1 – Assortativité en fonction de la taille du réseau

Les résultats montrent une forte homophilie pour les attributs sociaux (dorm, major), tandis que l'assortativité de degré est généralement faible ou légèrement négative.

Question 4 — Prédiction de liens

Nous implémentons trois métriques classiques :

— Common Neighbors

— Jaccard

— Adamic-Adar

Une fraction f des arêtes est supprimée aléatoirement, puis les liens manquants sont prédits à partir des scores.

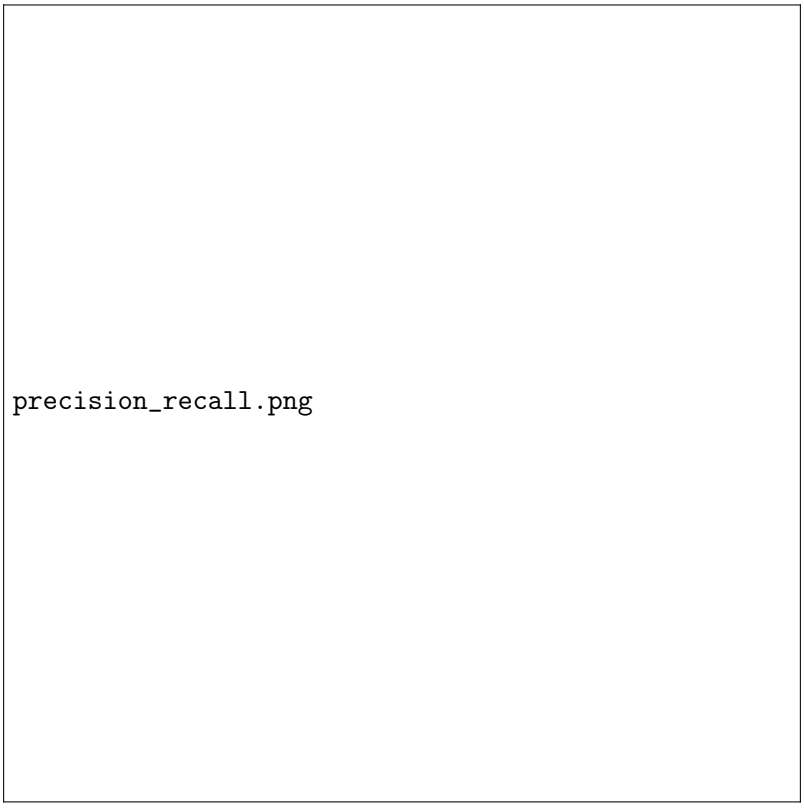

FIGURE 2 – Précision et rappel en fonction de k

Les résultats montrent que les méthodes basées sur les voisins communs sont efficaces pour les petites valeurs de k, Adamic-Adar offrant généralement les meilleures performances.

Question 5 — Propagation de labels

Nous appliquons un algorithme de propagation de labels pour prédire des attributs manquants (dorm, major, gender). Une fraction de 10%, 20% et 30% des labels est masquée aléatoirement.

Attribut	Accuracy	F1-macro	MAE
Dorm			
Major			
Gender			

TABLE 1 – Performances de la propagation de labels

La résidence (*dorm*) est généralement mieux prédite, ce qui s'explique par sa forte corrélation avec la structure communautaire du graphe.

Question 6 — Détection de communautés

Question de recherche : Les communautés détectées correspondent-elles principalement aux résidences étudiantes ?

Nous utilisons les algorithmes de Louvain et de maximisation gloutonne de la modularité.

Méthode	Attribut	NMI	ARI
Louvain	Dorm		
Greedy	Dorm		

Table 2 – Comparaison communautés / attributs

Les scores élevés de NMI et ARI pour l'attribut *dorm* confirment que la structure communautaire reflète largement l'organisation résidentielle.

Conclusion

Ce projet met en évidence les propriétés classiques des réseaux sociaux universitaires : hétérogénéité des degrés, fort clustering, homophilie marquée et structure communautaire significative. Les méthodes simples de graph learning se révèlent efficaces, tant pour la prédiction de liens que pour la récupération d'attributs manquants.