A supprimer pour les figures

Projet NSGL — Network Science and Graph Learning

Tanguy CESAR

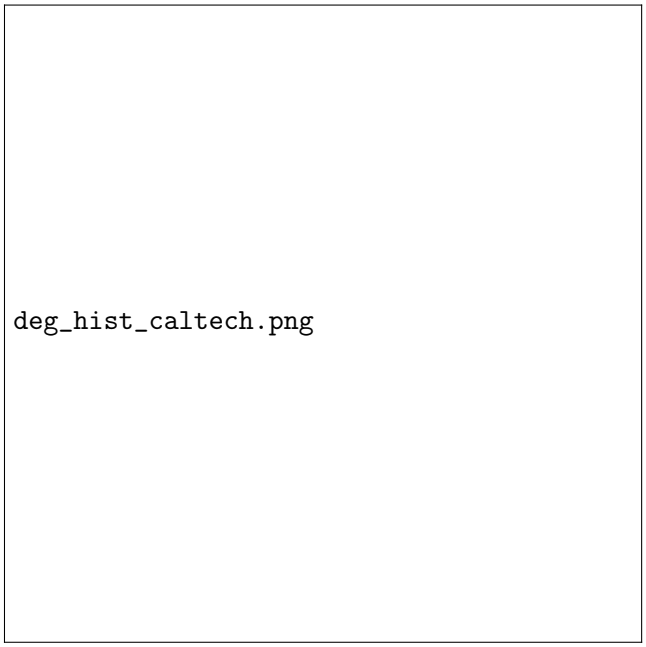
Introduction

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 2 — Analyse de réseaux sociaux

Les réseaux étudiés dans cette partie sont ceux de *Caltech*, *MIT* et *Johns Hopkins*. Pour chacun, nous considérons la plus grande composante connexe du graphe.

Distribution des degrés



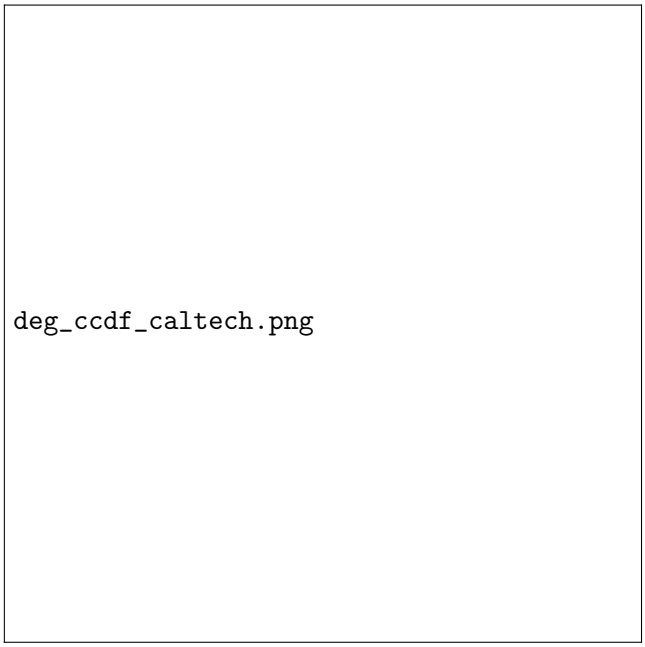


FIGURE 1 – Histogramme et CCDF des degrés — Caltech

Les distributions de degrés présentent une forte hétérogénéité, avec une majorité de sommets faiblement connectés et quelques nœuds très centraux. La représentation en échelle log-log met en évidence une queue lourde, typique des réseaux sociaux.

Clustering et densité

Réseau	n	Densité	Clustering global	Clustering moyen
Caltech				
MIT				
Johns Hopkins				

Lien degré – clustering local

FIGURE 2 – Clustering local en fonction du degré

On observe une relation décroissante entre le degré et le clustering local, traduisant le fait que les nœuds très connectés relient souvent des communautés différentes.

Question 3 — Assortativité et homophilie

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

FIGURE 3 – 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.

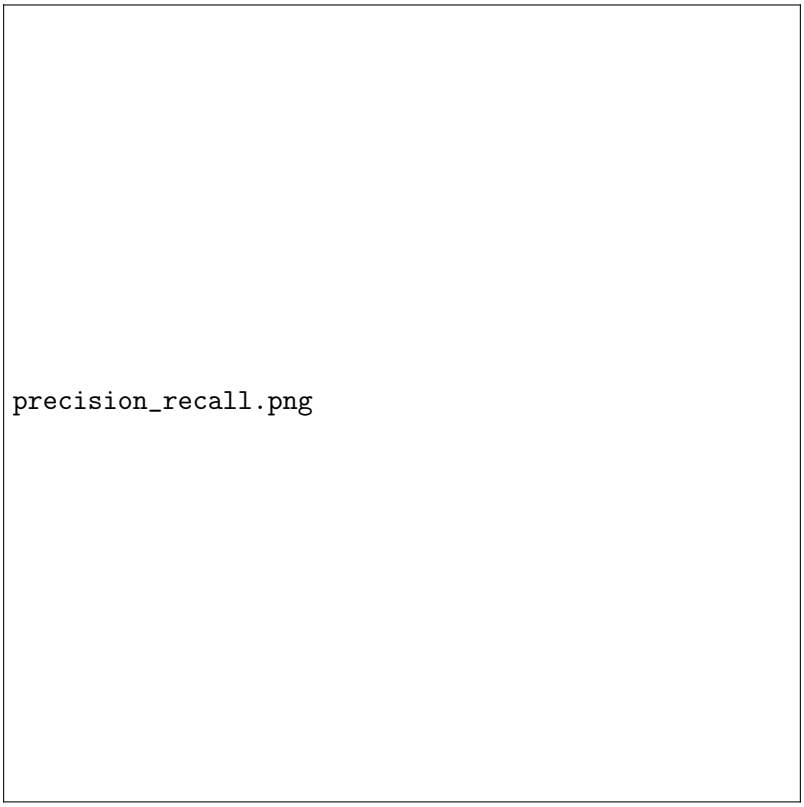


FIGURE 4 – Précision et rappel en fonction de k

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 2 – 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 3 – 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.