NET 4103/7431 Homework
Network science and Graph Learning

HOSTETTLER Maël
mael.hostettler@telecom-sudparis.eu

Projet de fin de module

Télécom SudParis Année 2024/2025

Table des matières

1 Introduction

L'analyse de graphe est un sujet central à l'étude de plusieurs phénomènes, en particulier les communautés, les réseaux sociaux et les moteurs de recherche. Grâce aux outils acquis lors du cours `NET 4103/7431` nous allons analyser le corpus de données Facebook100. Ce corpus de données décrit les liens d'amitié de différents élèves de 100 universités américaines. L'échelle est donc significative, sans pour autant être tellement grande que l'on ne puisse conduire une telle analyse sur un ordinateur personnel.

L'analyse de ce corpus de données s'articule alors en 5 parties plutôt distinctes :

— L'analyse des propriétés structurelles du graphe comme le clustering global/local
— L'analyse détaillée de l'assortativité par degré
— La prédiction de lien via trois métriques différentes
— La propagation d'étiquettes afin de retrouver/prédire certains attributs
— La détection ainsi que l'influence des concentrateurs de relations (hubs) à l'échelle locale est globale

Le code source de ce projet est disponible en ligne sur le dépôt suivant : `https://github.com/maelhos/TSP-NET4103`. Ce projet vise principalement à faire de l'analyse, donc à implémenter, exécuter et interpréter les résultats. Cependant, dans la dernière partie, nous proposons un algorithme de détection de concentrateurs basé sur l'échantillonnage afin d'étudier la corrélation entre les concentrateurs et les individus avec beaucoup d'amis.

2 Analyse structurelle

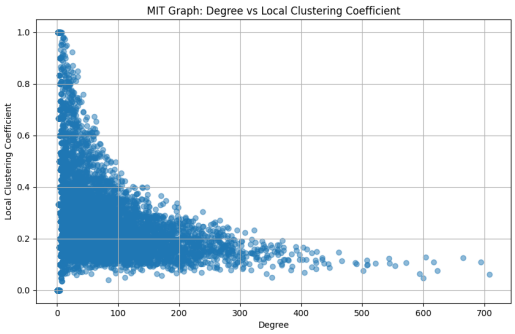
Coefficiant de regroupement			
Université	Global Clustering	Local Clustering	Edge density
Caltech	0.29	0.41	0.06
MIT	0.18	0.27	0.01
Johns Hopkins	0.19	0.27	0.01

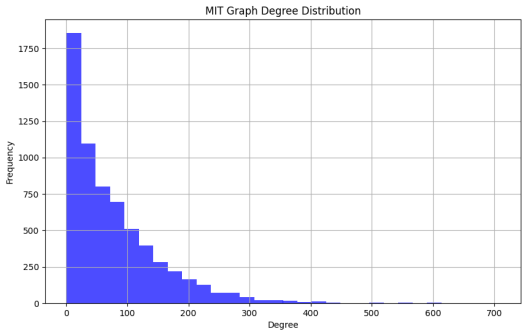
On a ici deux cas de figures très flagrants : les étudiants de Caltech forment des communautés plus soudées, alors que MIT et Johns Hopkins forment un réseau moins connexe. Il est important de noter que Caltech ont aussi un coefficient de clustering global plus élevé, ce qui pourrait indique que les communautés de Caltech ne sont pas plus soudées comme le laisserait penser le coefficient de clustering local et l'edge density, mais simplement plus nombreuses/denses. Cependant, la différence de global clustering entre Caltech et les deux autres et bien plus faible que la différence de local clustering entre Caltech et les deux autres. Il serait intéressant d'avoir une métrique qui analyse le clustering local indépendamment (au sens de la corrélation) du clustering global.

Dans les trois cas les graphes doivent être représenté de manière creuse, un ordre de grandeur donnée souvent dans la littérature (**"Graph Theory and Its Applications"** by Jonathan Gross and Jay Yellen par exemple) est :

— $D < 0.1$: Représentation creuse
— $D > 0.5$: Représentation matricielle dense
— $0.1 \leq D \leq 0.5$: À voire, en fonction de l'application

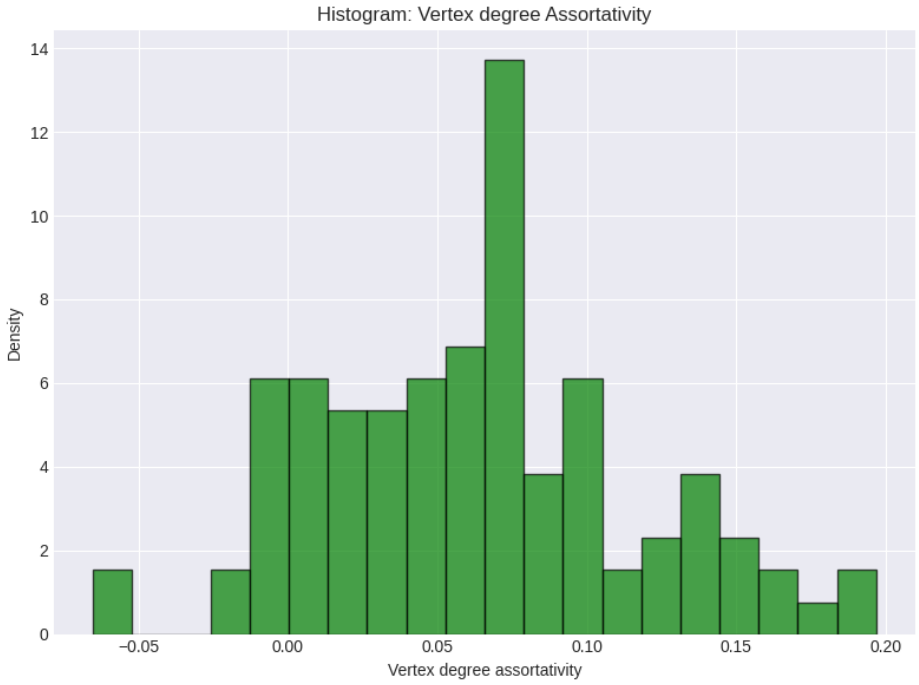
Dans notre cas, comme la densité ne dépasse pas les $D = 0.06$, tout doit être représenté de manière creuse. Une autre observation intéressante est la distribution des degrés. Celle-ci révèle qu'il y a exponentiellement moins d'individus avec de plus en plus d'amis. Ce n'est pas étonnant, mais cela a un rôle important par la suite pour l'étude de concentrateurs (hubs).





3 Analyse d'assortativité

De manière générale, les individus ont une assortativité faible avec une concentration proche de 0. L'analyse d'assortativité nous mène une fois de plus vers les concentrateurs. En effet, on remarque que les différents liens d'amitiés ne sont pas très corrélés au nombre d'amis. Autrement dit, quelqu'un qui a beaucoup d'amis a plutôt tendance à être ami avec des gens qui ont moins d'amis que lui. Plus généralement, les nœuds avec des degrés proches ne sont pas particulièrement plus amis les uns avec les autres.



L'assortativité n'est donc pas tant corrélée au degré, elle est plutôt corrélée aux attributs sociaux et personnels.

4 Prédiction de lien

Nous nous concentrerons sur le graph de Caltech afin de tester les trois métriques différents : Common Neighbors, Adamic-Adar, et Jaccard.

Prédiction en fonction de la métriques choisie					
Métrique	k	Fraction enlevée	Précision	Recall	top-k
Common neighboor	50	0.05	0.20	0.01	0.20
	50	0.2	0.38	0	0.38
	100	0.05	0.21	0.02	0.21
	100	0.2	0.41	0.01	0.41
	200	0.05	0.19	0.04	0.19
	200	0.2	0.38	0.02	0.38
	400	0.05	0.15	0.07	0.15
	400	0.2	0.36	0.04	0.36
Adamic-Adar	50	0.05	0.28	0.02	0.28
	50	0.2	0.48	0.01	0.48
	100	0.05	0.24	0.03	0.24
	100	0.2	0.42	0.01	0.42
	200	0.05	0.20	0.05	0.20
	200	0.2	0.39	0.02	0.39
	400	0.05	0.17	0.08	0.17
	400	0.2	0.36	0.04	0.36
Jaccard	50	0.05	0.14	0.01	0.14
	50	0.2	0.12	0.00	0.12
	100	0.05	0.11	0.01	0.11
	100	0.2	0.22	0.01	0.22
	200	0.05	0.13	0.03	0.13
	200	0.2	0.29	0.02	0.29
	400	0.05	0.11	0.05	0.11
	400	0.2	0.26	0.03	0.26

Pour la métrique Common neighboor, on remarque que celle-ci est assez sensible à la fraction de liens retirés plus qu'au nombre d'itérations. Cette métrique donne donc une meilleure information globale, mais est moins bonne quand on veut des informations précises sur quelques liens manquants. En moyenne, Adamic-Ada est plus performante que Common neighboor avec des caractéristiques semblables à Common neighboor. Cette métrique est aussi moins sensible que Common neighboor. Jaccard est intéressante, mais décevante. Cette métrique semble moins sensible à la proportion de liens enlevés, mais est trop peu performante pour être utile, et cela, peu importe le cas dans le cadre de cette étude.

Les métriquees Common Neighbors et Adamic-Adar sont plus efficaces et Adamic-Adar parait être la meilleure dans ce contexte. Cela n'est pas étonnant, car elle prend en compte la structure locale du réseau, ce qui est crucial dans le cas d'une étude de réseau social.

5 Propagation d'étiquettes

Nous allons désormais nous intéresser au problème de propagation d'étiquettes afin de retrouver des attributs manquants. Tout les résultats seront sur le graph : MIT.

Prédiction en fonction de la métriques choisie			
Attribut	Fraction enlevée	Précision	Erreur moyenne
Dorm	0.1	0.34	0.99
	0.2	0.32	1
	0.3	0.32	1
Major index	0.1	0.24	5.35
	0.2	0.21	5.14
	0.3	0.21	5.49
Gender	0.1	0.55	0.54
	0.2	0.53	0.56
	0.3	0.53	0.55

On remarque que les attributs "Dorm" et "Gender" sont bien mieux retrouvés que "Major index". Cela n'est pas très étonnant non plus, car notre score ne prend pas en compte le biais d'erreur. Il n'y a qu'une chance sur deux de se tromper sur le genre (enfin... presque...) alors qu'il n'y a qu'un seul "Major index" juste parmi une liste non binaire. Il faudrait donc une métrique plus adaptée afin de comparer plus justement.

6 Sectes / communautées fermées

Nous allons nous pencher sur les communautés fermées (gated communities en anglais). Ce sujet a a priori un réel intérêt, car il permettrait potentiellement de détecter des communautés "potentiellement problématiques" purement algorithmiquement sans avoir à analyser le contenu.

L'hypothèse que l'on émet dans le cadre de ce dataset est assez simple : une "gated community" est un ensemble de personnes qui ont beaucoup de liens entre eux, mais peu de lien vers "les autres". On s'attend à ce que ces communautés aient beaucoup de caractéristiques en commun, notre but est donc de vérifier/infirmer cela. En particulier, on s'attend à ce que des caractéristiques comme année ou dortoir soient celles qui forment le plus de gate community.

Il existe un outil pour mesurer ce que l'on souhaite : la conductance. La définition formelle de la conductance implique de parler de stabilité pour des chaînes de Markov, mais concrètement, dans notre cas, cet outil va nous permettre de mesurer le taux de "fuites" d'une communauté : plus celui-ci est bas, plus elle est gated.

On se propose alors de simplement calculer les communautés (dans ce cas-ci, j'ai

choisi l'algorithme de Louvain avec une résolution à 2 pour préférer les plus petites communautés), puis leur conductance. On choisit alors expérimentalement un threshold de conductance pour déclarer une communauté comme "gated" et on regarde le trait dominant de cette communauté.

Universitée	**Rang**	**Conductance**	**Taille**	**Caractéristique en commun**	**Pureté**
Caltech	1	0.413	101	Dorm 169	0.96
	2	0.437	76	Dorm 168	0.95
	3	0.453	93	Dorm 166	0.88
MIT	1	0.263	916	Year 2009	0.95
	2	0.286	5	Major 6	0.75
	3	0.309	40	Dorm 236	0.89
Johns Hopkins	1	0.173	880	Year 2009	0.98
	2	0.227	236	Major 217	0.90
	3	0.435	523	Year 2006	0.29

On remarque effectivement que les caractéristiques en commun sont souvent soit une année soit un dortoir. Notre hypothèse est donc plutôt vérifiée. On remarque pour autant qu'à Caltech les gated communities se font plus par dortoir que par année. Au MIT, mis à part l'année 2009 (pas très intéressante), la majeure 6 qui est en fait "Electrical Engineering and Computer Science (EECS)"…Pas si étonnant. Et pour John Hopkins, la majeure 217 "Biomedical Engineering (BME)" est connue pour être la meilleure aux US en biomédical…Les tailles varient aussi beaucoup, mais cela est probablement lié au fait que, dans les écoles plus petites, on connaît une plus grande proportion des gens (biais communautaire).