Facebook100 Dataset Analysis

Thomas Winninger
HackademINT, Télécom SudParis, Évry, France
thomas.winninger@telecom-sudparis.eu

GitHub: https://github.com/Sckathach/graphs-homework

1 Analyse générale

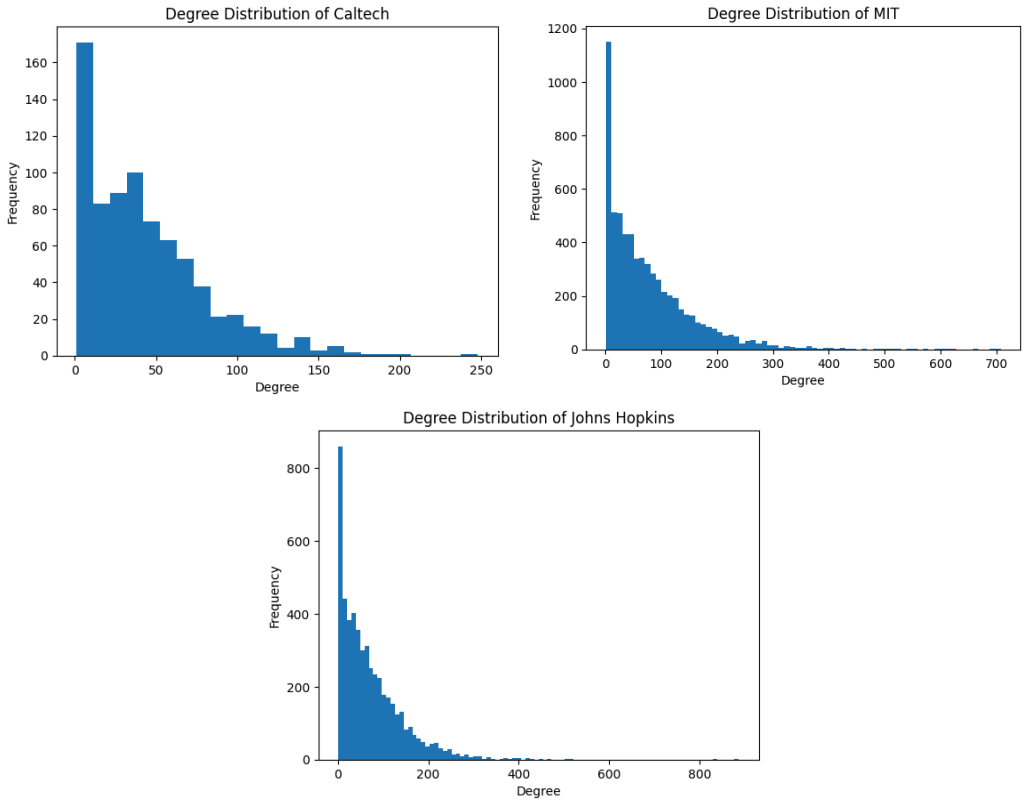


Figure 1: Distribution des degrés des trois universités (Caltech, MIT et Johns Hopkins)

Plusieurs attributs classiques des réseaux sociaux sont visibles. Il y a une grande proportion de nœuds à faible degré. Cela est typique des graphes à échelle libre. Il y a tout de même quelques nœuds à degré très élevés, ce sont des points de concentration forte qui permettent de connecter un grand nombre de personnes. On peut par exemple penser à des étudiants populaires, des comptes d'information, ou des professeurs.

Les graphes sont clairement épars, $|E| \ll |V|(|V|-1)$, et l'on peut voir une loi de puissance.

2 Coefficients de groupe

University	Global Clustering Coefficient	Mean Local Clustering Coefficient	Edge Density
Caltech	0.40	0.29	0.05
MIT	0.27	0.18	0.012
Johns Hopkins	0.26	0.19	0.01

Table 1: Clustering coefficient comparaison

Le graphe de Caltech a un coefficient de groupe bien plus élevé que ceux du MIT et de Johns Hopkins. Cela peut signifier qu'il y a plus de petits groupes au sein de l'université (un ami d'un ami est un ami), alors que les deux autres universités, bien que possédant des nœuds à grandes concentrations, sont majoritairement constitués d'individus dispersés.

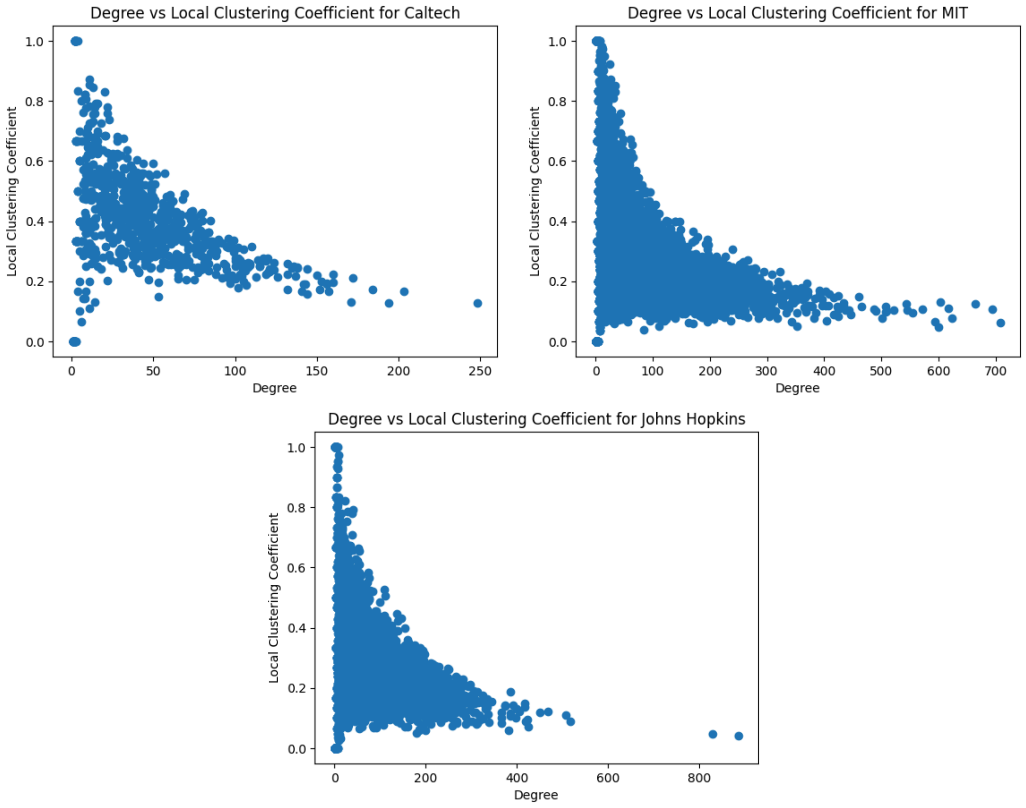


Figure 2: Degré comparé au coefficient local de groupe

On retrouve le phénomène de : un ami d'un ami est un ami, car les nœuds à faible degré ont tendance à avoir un grand coefficient de groupe. De plus, l'université de Johns Hopkins n'a pas de continuité dans les degrés de s im 500 à s im 800, cela peut être la marque d'une hiérarchie différente.

3 Assortativité

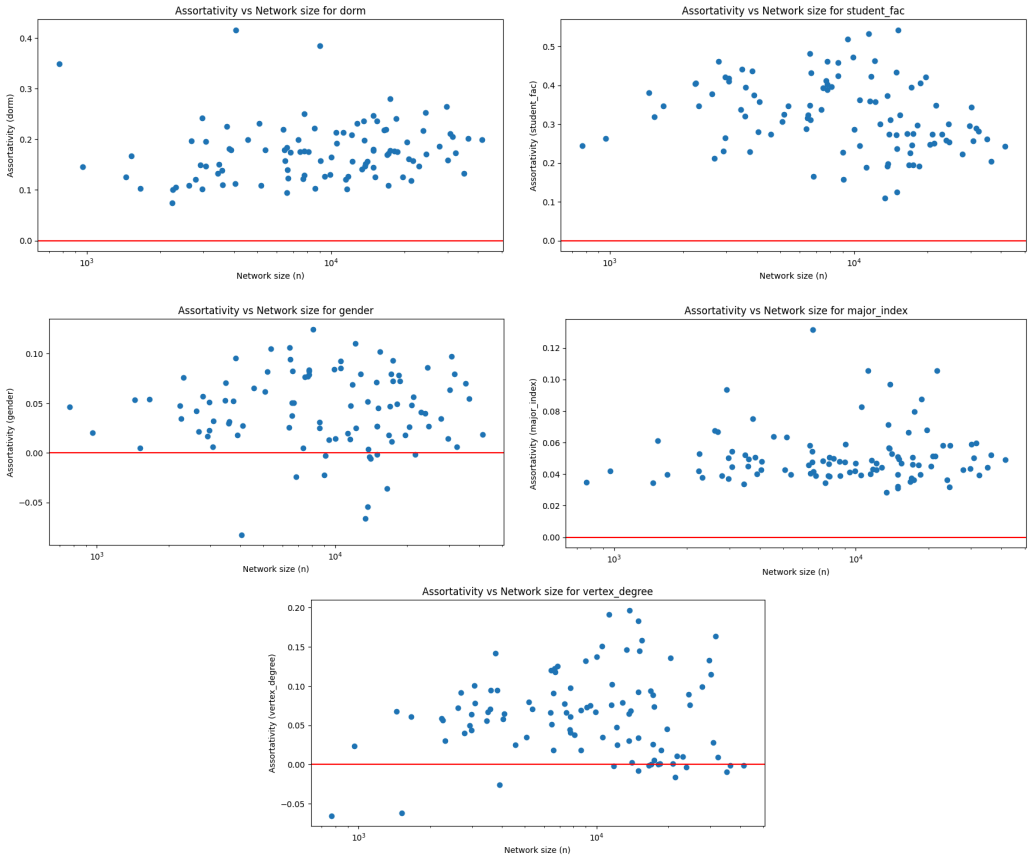
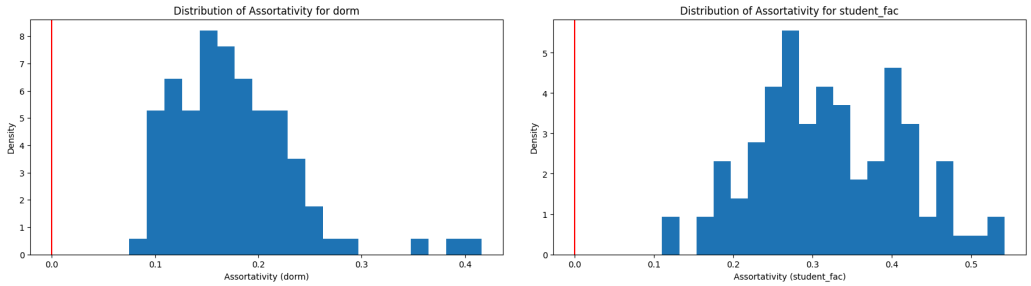


Figure 3: Assortativité comparée à la taille du réseau

Les universités très grandes ou très petites ont tendance à posséder des groupes homogènes, tandis que les universités de taille moyenne ont tendance à avoir plus de mixité. Certains points extrêmes sont explicables par la structure de l'école, par exemple pour le sexe, les universités ayant une assortivité très basse sont celles qui sont unisexe.



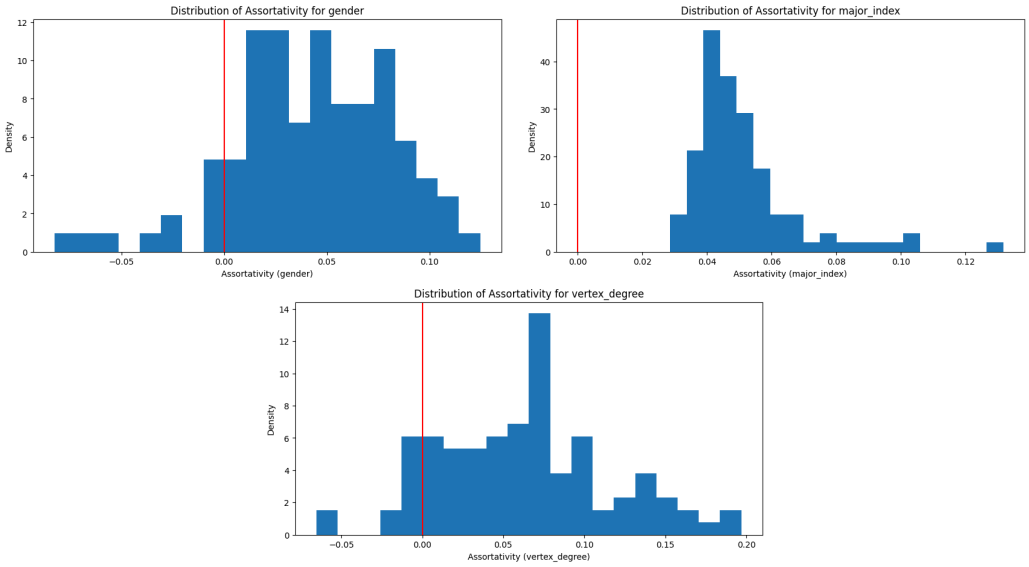


Figure 4: Distribution des assortivités

On remarque que les attributs de sexe et de majeure sont ceux qui ont tendance à favoriser l'homogénéité : des garçons/ filles ont tendance à rester entre eux, alors que le statut de l'étudiant importe peu.

4 Prédiction des arêtes

$f \setminus k$	50	100	200	400
0.05	0.3	0.32	0.25	0.2025
	0.46	0.45	0.35	0.23
	0.46	**0.47**	**0.365**	**0.245**
0.1	0.38	0.36	0.345	0.37
	0.7	**0.6**	0.525	0.45
	0.72	0.59	**0.535**	**0.4625**
0.15	0.42	0.52	0.46	0.4875
	0.78	**0.79**	0.665	0.5475
	0.82	0.76	**0.665**	**0.555**
0.2	0.42	0.55	0.505	0.5175
	0.88	0.84	0.715	0.68
	0.88	**0.84**	**0.72**	**0.685**

Table 2: Clustering coefficient comparaison (Jaccard, CommonNeighbours, AdamicAdar)

La précision a été calculée avec $\mathrm{PR} = \frac{|\mathrm{TP}|}{|\mathrm{TP}|+|\mathrm{FP}|}$. Cependant, je n'ai pas calculé le "recall", car je n'ai pas trouvé comment calculer le nombre de faux négatifs pour un k donné. Si l'on demande les 50 prédictions les plus probables et qu'on a enlevé 2000 arêtes, forcément qu'il y aura au moins 1950 faux négatifs.

Avec les données récoltées, la méthode du coefficient de Jaccard est, dans tous les cas, beaucoup moins efficace que les deux autres, qui sont quasiment équivalentes. Il peut être intéressant de les utiliser en même temps, car elles ne trouvent pas les mêmes arêtes (voir papier).

La deuxième observation que l'on peut faire, c'est que les prédictions sont plus élevées quand le nombre d'arêtes à trouver est élevé, et que l'on en demande peu. Ce qui parait crédible, à chaque choix réussi, le suivant est plus complexe, car il y a moins à trouver.

5 Prédiction des labels

J'ai utilisé l'algorithme de propagation de label de de Zhou et al. (2004), moyenné sur 10 essais pour avoir des résultats stables.

	fraction removed			
	0.1	0.2	0.3	0.4
Caltech				
Major	0.23	0.23	0.21	0.19
Dorm	0.64	0.65	0.66	0.67
Year	0.62	0.60	0.58	0.60
Gender	0.59	0.58	0.60	0.60

Table 3: Accuracy of the label propagation algorithm

La performance de l'algorithme varie grandement en fonction du label recherché. Cela est dû au fait que l'algorithme dépend énormément de la structure du graphe. Si les étudiants forment des groupes homogènes en fonction des attributs recherchés, alors l'algorithme pourra facilement prédire un attribut manquant au sein du groupe. Par contre, si les groupes sont très hétérogènes, la probabilité de succès diminue. En fait, le succès de l'algorithme est inversement proportionnel à l'entropie.

Les résultats peuvent aussi être comparés avec les assortativités. Cependant, avec le peu de mesure que j'ai fait, je n'arrive pas à voir de lien clair, par exemple, qu'une faible assortativité induise un grand succès de l'algorithme.

6 Détection de communautés

En s'inspirant du travail de [1] sur l'analyse des maisons de Caltech, j'ai décidé de regarder la structure des communautés de Caltech et du MIT en fonction de l'année des étudiants. Je me suis demandé si l'on pouvait avoir une idée de la vie étudiante en regardant le mélange d'étudiants d'une année à l'autre.

Je suppose que, en choisissant deux années a et b, et en regardant un groupe d'étudiant appartenant à la même communauté, alors, si la proportion d'étudiant de a est du même ordre de grandeur que celle de b, il y a un mélange et donc une sorte de transmission de la culture étudiante d'année en année.

À l'inverse, si en découpant le graphe en communautés, on s'aperçoit que les groupes sont homogènes, alors je suppose qu'il y a moins de transmissions, et par extension, une culture étudiante moins riche.

Je me base évidemment sur l'expérience française, qui a une forte culture de transmission des traditions (sup $\rightarrow$ spé, 1A $\rightarrow$ 2A).

J'en ai aussi profité pour regarder quel était l'écart maximal d'années ayant un mélange significatif. Cependant, les données du jeu Facebook100 ne permettent pas de remonter assez loin. On se retrouve très vite avec des échantillons très réduits. Par exemple pour Caltech, il y a 173 étudiants de 2008, contre moins de 10 étudiants pour les années inférieures à 2003.

Pour les algorithmes de détection de communauté, comme expliqué dans [2], dés que l'on passe à grande échelle, les résultats deviennent très vite incohérents. Pour pallier ce problème, j'ai réduit la taille des graphes en prenant le sous-graphe $G'(V', E')$ avec V' définit par

$$V' = \{v \in V \mid v_{\text{year}} \in \{a, b\}\}. \tag{1}$$

Les trois algorithmes que j'ai utilisés sont : l'algorithme de Newman Girvan, l'algorithme de Kernighan Lin, et l'algorithme de Louvain. Celui de Newman Girvan est très vite inutilisable, en pratique, je n'ai réussi qu'a récupérer une composante et quelques individus éparses. L'algorithme de Kernighan Lin est pratique, vu qu'il force une section en 2 du graphe qui maximise la modularité (vu dans [1]). Mais étant donné qu'il implique un biais important, j'ai aussi regardé les résultats de l'algorithme de Louvain qui ont l'air très concluants.

	Size	Percentage of year a	Percentage of year b	Modularity
Kernighan Lin				
Group 1	163	54	46	0.23
Group 2	163	52	48	0.23
Louvain				
Group 1	73	55	45	0.35
Group 2	60	43	57	0.35
Group 3	58	40	60	0.35
Group 4	43	60	40	0.35

Table 4: Caltech communities structure, year 2008 and 2006

	Size	Percentage of year a	Percentage of year b	Modularity
Kernighan Lin				
Group 1	31	0	100	0.48
Group 2	31	74	26	0.48
Louvain				
Group 1	20	0	100	0.55
Group 2	17	76	24	0.55
Group 3	13	0	100	0.55
Group 4	8	100	0	0.55

Table 5: Caltech communities structure, year 2009 and 2004

		Size	Percentage of year a	Percentage of year b	Modularity
Kernighan Lin					
	Group 1	936	8	92	0.36
	Group 2	937	99	1	0.36
Louvain					
	Group 1	761	1	99	0.41
	Group 2	394	97	3	0.41
	Group 3	300	94	6	0.41
	Group 4	161	97	3	0.41

Table 6: MIT communities structure, year 2008 and 2009

Les données de Table 4 et de Table 5 illustrent une vie étudiante similaire aux écoles d'ingénieur françaises. Les groupes sont très hétérogènes si l'on considère une fenêtre de moins de 3 ans, par contre, quand il y a plus de 3 ans d'écart, il y a moins de mélange. Le coefficient de modularité très faible peut aussi indiquer qu'il n'y a pas réellement de groupe indépendant du reste, ou en tout cas, qu'il n'est pas déterminé par l'année.

Le *Groupe 2* trouvé par l'algorithme de Kernighan Lin pour les années 2009 et 2004 est clairement faussé. Vu que le premier est composé à 100% d'étudiants de 2004, l'algorithme à forcé les étudiants restants de 2004 à se mélanger avec ceux de 2009 pour avoir deux groupes de taille égale. On peut donc considérer qu'il n'y a aucun mélange.

Les données de Table 6 indiquent une claire différence de structure sociale avec Caltech. Au MIT, même des étudiants d'années très proches (2008-2009) ne se mélangent pas, ou en tout cas, ne forment pas de groupe mixte. Les résultats sont d'autant plus précis que la taille des échantillons est plus grande que les groupes de Caltech.

Les différences peuvent s'expliquer par la taille de l'université, son ouverture aux étudiants internationaux qui ne restent qu'un semestre ou un an, et par la pluralité des disciplines. Caltech est une université centrée sur la technologie, tandis que le MIT est ouvert à pluridisciplinaire.

Il faudrait néanmoins avoir des données d'écoles françaises pour pouvoir réellement comparer les différences de structure. La taille restreinte des échantillons et la particularité de Caltech sont aussi des éléments pouvant altérer les résultats.

[1] A. L. Traud, E. D. Kelsic, P. J. Mucha, and M. A. Porter, "Comparing community structure to characteristics in online collegiate social networks," *SIAM Review*, vol. 53, pp. 526–543, 2011.

[2] C. Lee and P. Cunningham, "Community detection: effective evaluation on large social networks," *Journal of Complex Networks*, vol. 2, pp. 19–37, 2013.