

Hu
Nicolas

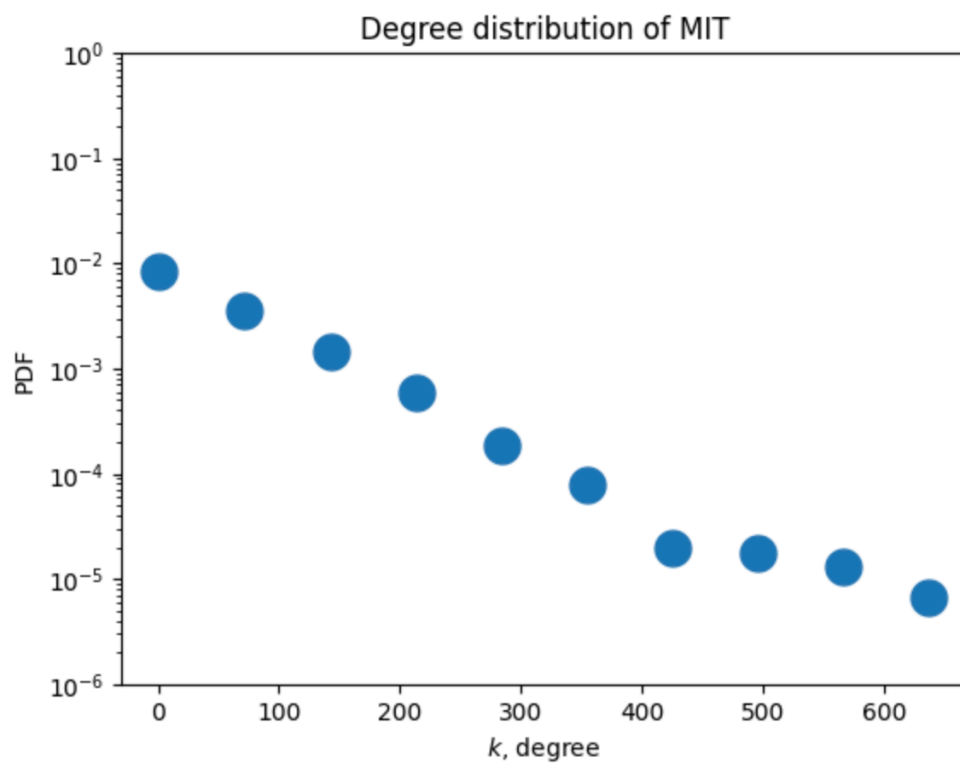
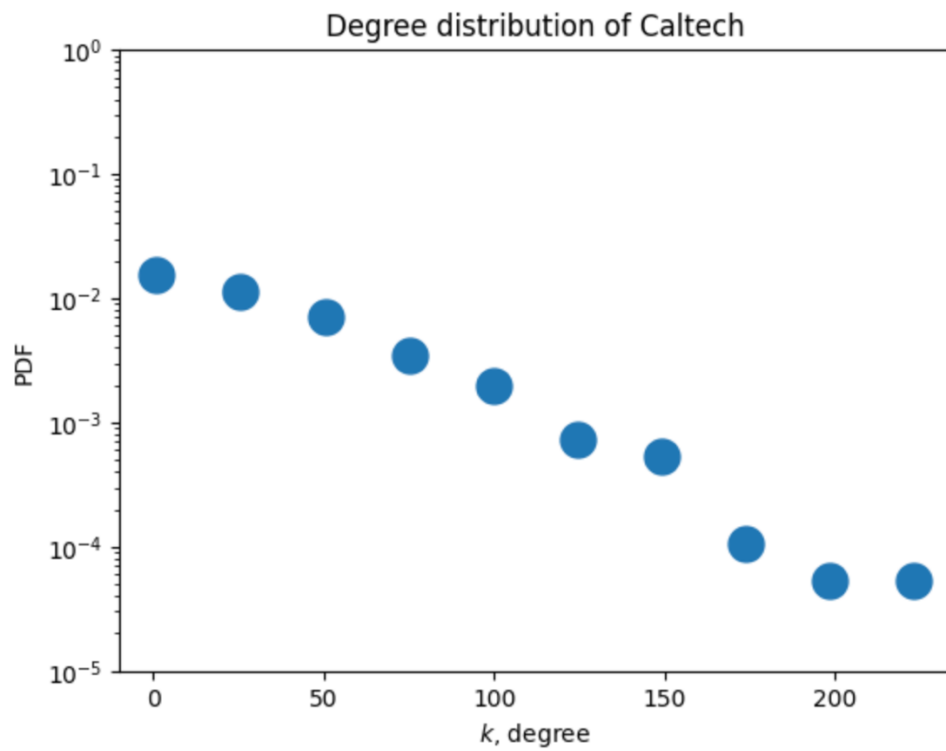
NET4103

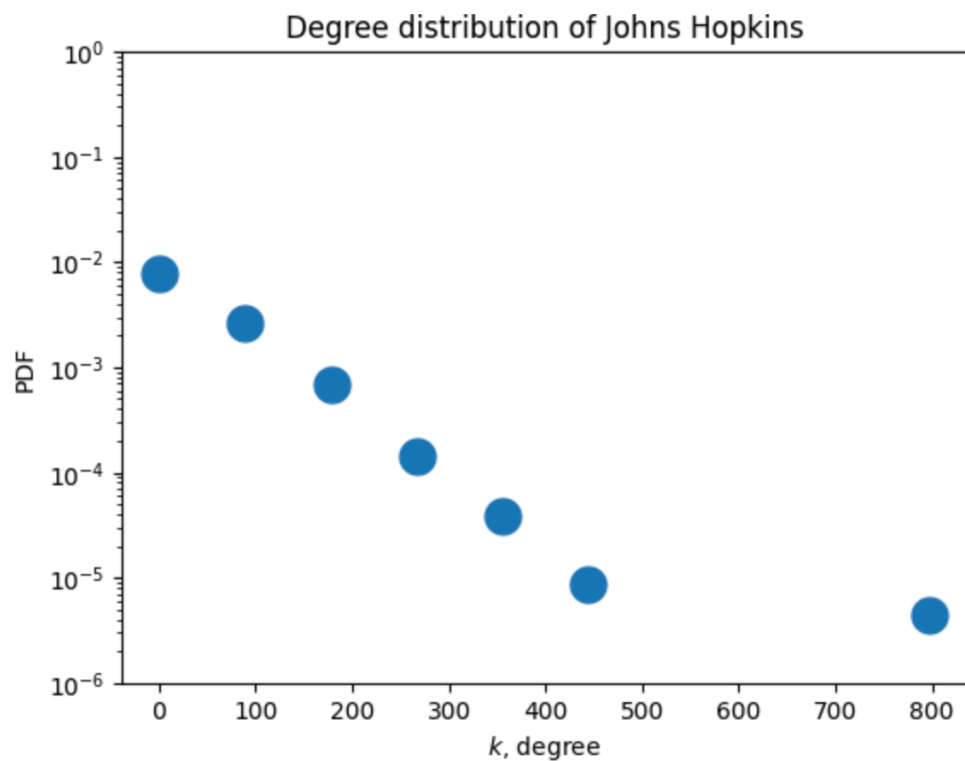
Compte rendu sur le travail à faire sur fb100

Le code du devoir est accessible via le lien suivant :

https://github.com/Raltuld/net4103_fb_homework

Question 2.a





Caltech et le MIT ont à peu près une distribution des degrés en loi de puissance. Johns Hopkins quand lui a une distribution des degrés en loi de puissance de 0 à 500 mais connaît quand même une probabilité non nulle mais extrêmement faible d'avoir 800 degrés sur un des nœuds. Ces distributions correspondent à des graphes qu'on retrouve habituellement avec une structure hiérarchique.

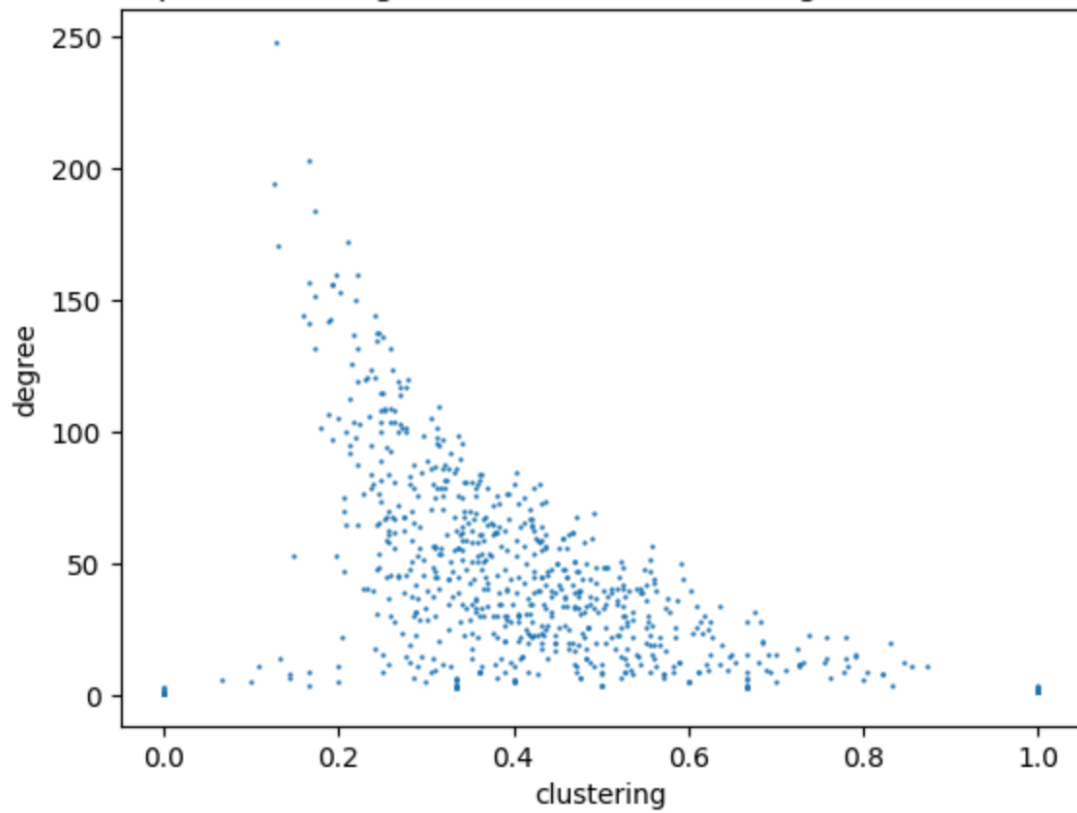
Question 2.b

	Caltech	MIT	Johns Hopkins
Global clustering coefficient (calculé avec <code>nx.transitivity()</code>)	0.2912826901150874	0.18028845093502427	0.19316123901594015
Mean local clustering coefficient (calculé avec <code>nx.average_clustering()</code>)	0.40929439048517247	0.2712187419501315	0.26839307371293525
Edge density (calculé avec <code>nx.density()</code>)	0.05640442132639792	0.012118119495041378	0.013910200162372396

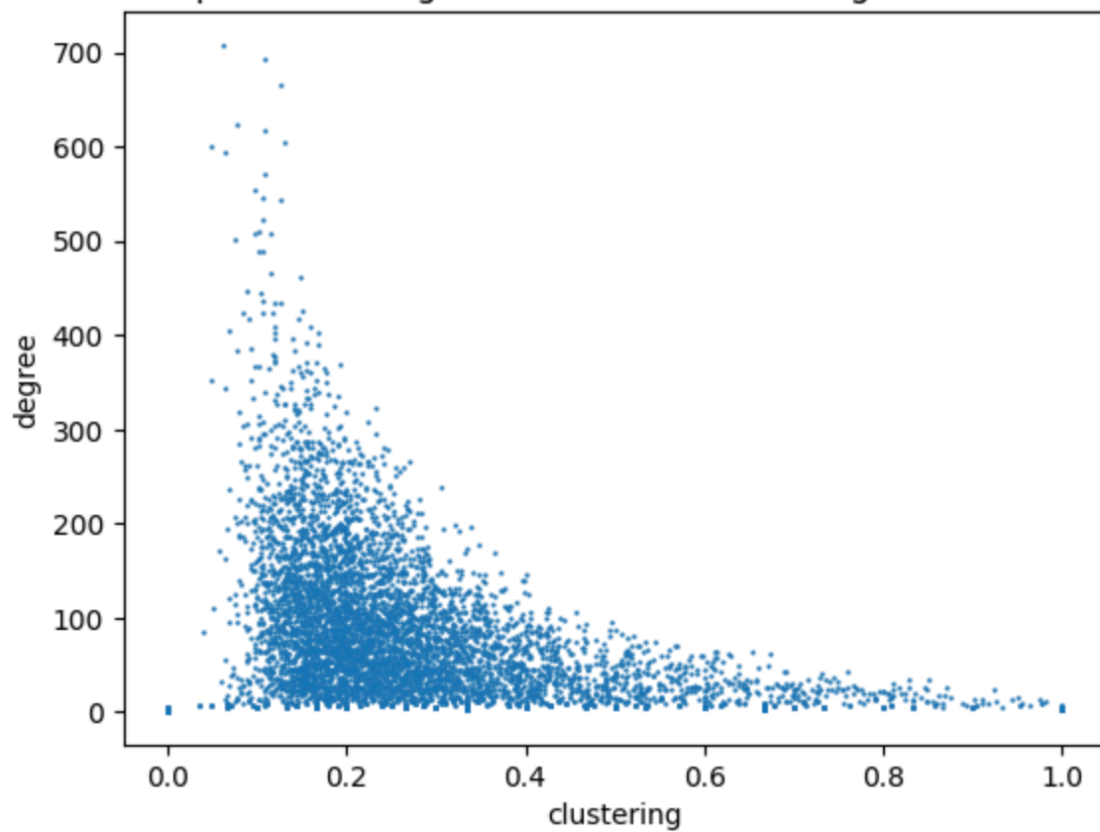
Ayant des densités de connections inférieures à 5%, on peut considérer ses graphes comme étant éparses. Cependant on peut noter que Caltech l'est moins que MIT et Johns Hopkins, et qu'il a plus de clusters dans son graphe que dans les deux autres. On peut cependant noter que ce sont des graphes avec des degrés de clustering assez élevé mais qu'on pourrait justifier par le fait que ce sont des universités et que les gens ont une chance élevée de connaître beaucoup de gens au sein de l'université. Néanmoins, on peut quand même noter que la différence entre Caltech et, MIT et Johns Hopkins, semble assez cohérente car plus un graphe a de nœuds, plus il est difficile d'être 'proche' des autres nœuds.

Question 2.c

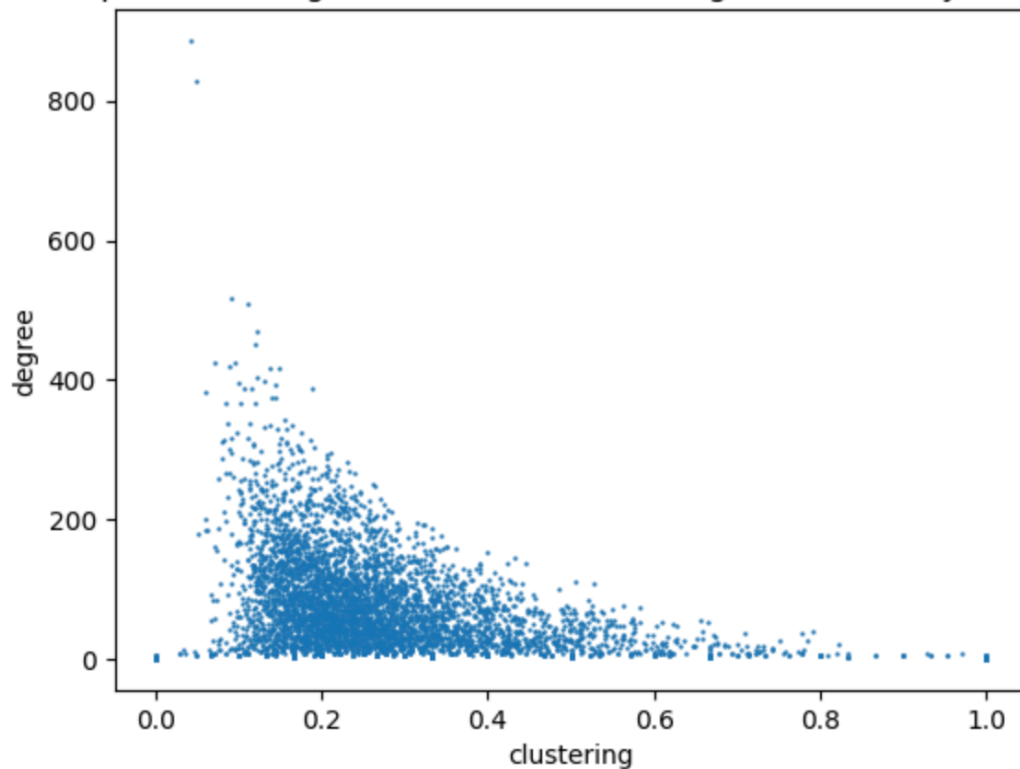
Scatter plot of the degree versus local clustering coefficient of Caltech



Scatter plot of the degree versus local clustering coefficient of MIT



Scatter plot of the degree versus local clustering coefficient of Johns Hopkins

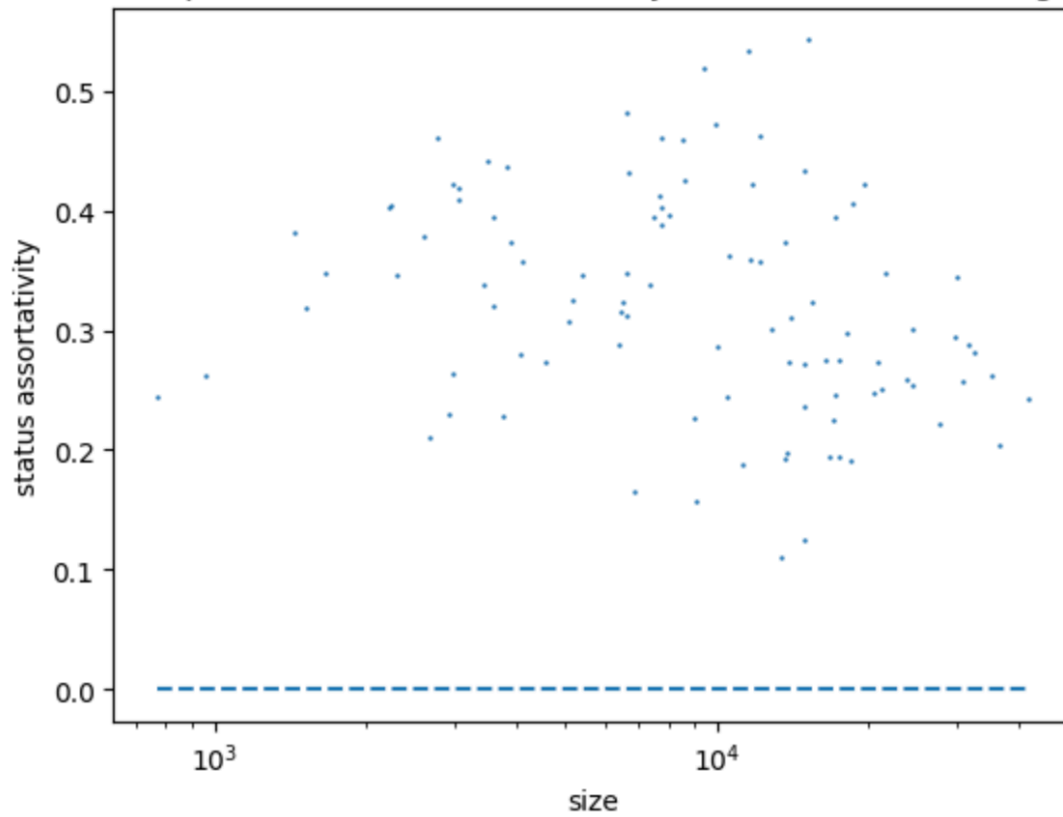


Les graphiques ci-dessus permettent de renforcer l'idée que pour Caltech, étant une université plus petite, les personnes ont plus de chance de former des plus gros clusters que dans le cas du MIT ou de Johns Hopkins ayant plus de personnes au sein de ces universités.

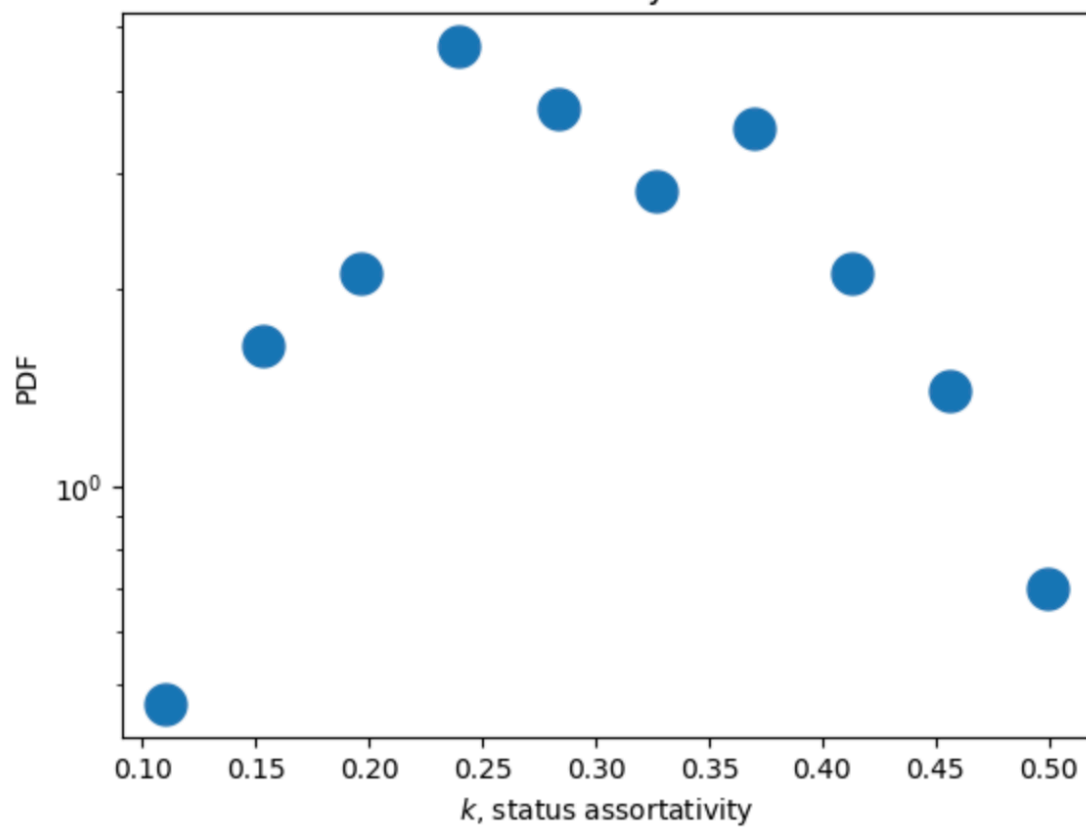
On peut cependant considérer que les trois universités ont des réseaux assez similaires car l'allure et les valeurs des graphes et des calculs précédents le sont.

Question 3

Scatter plot of the status assortativity versus the size of the graph

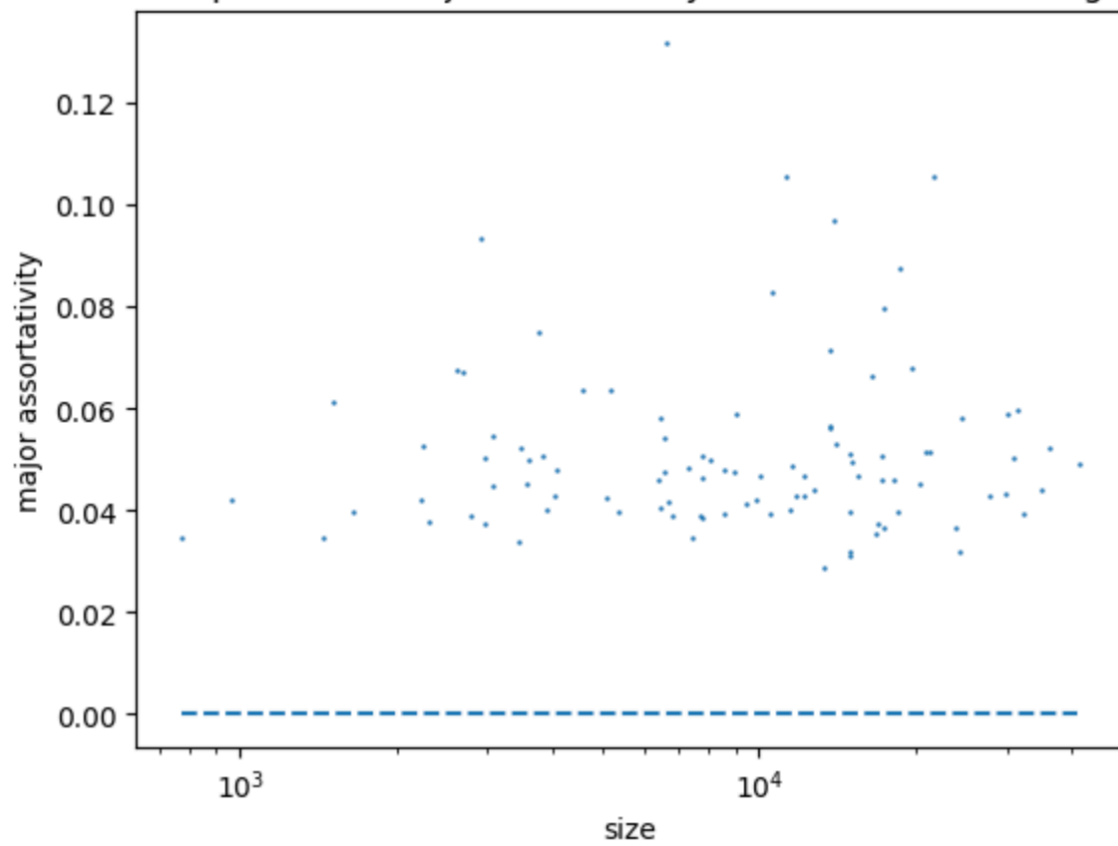


Status assortativity distribution

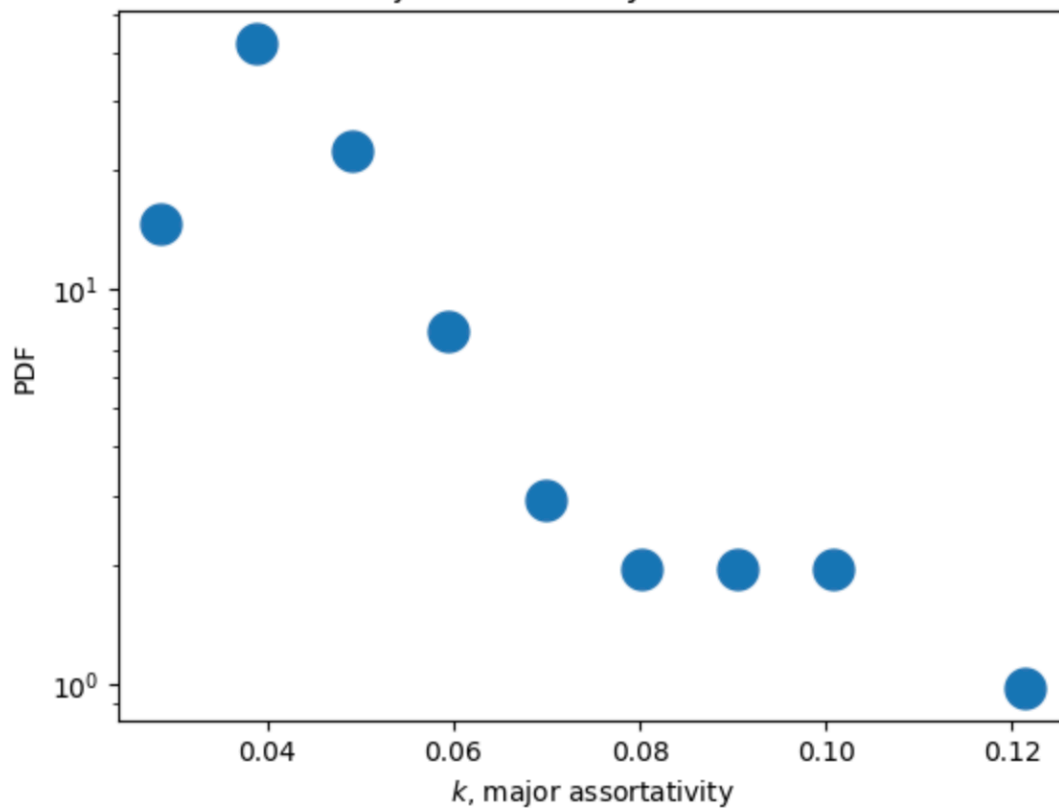


On observe une assortativité moyenne autour de 30% avec aucun point en dessous de la ligne de non-assortativité, ce qui signifie que les gens ont tendance à avoir des liens avec des personnes du même statut, ce qui semble logique.

Scatter plot of the major assortativity versus the size of the graph



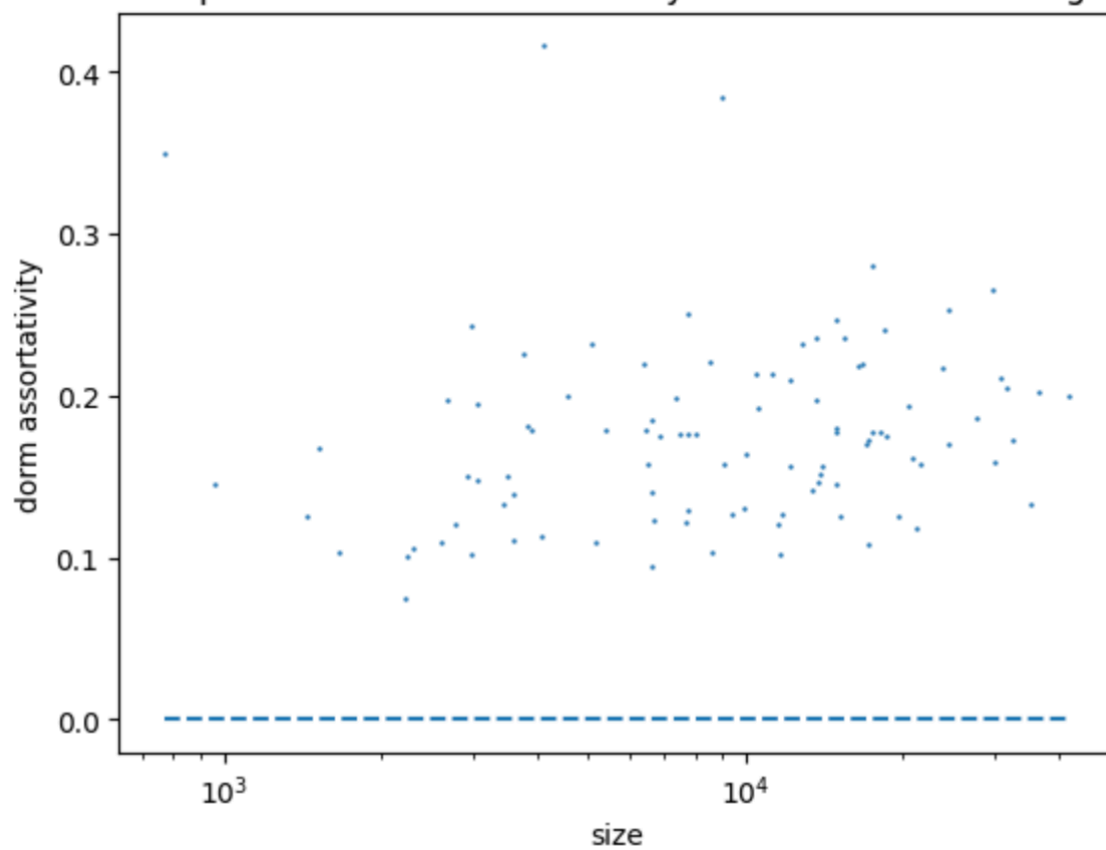
Major assortativity distribution



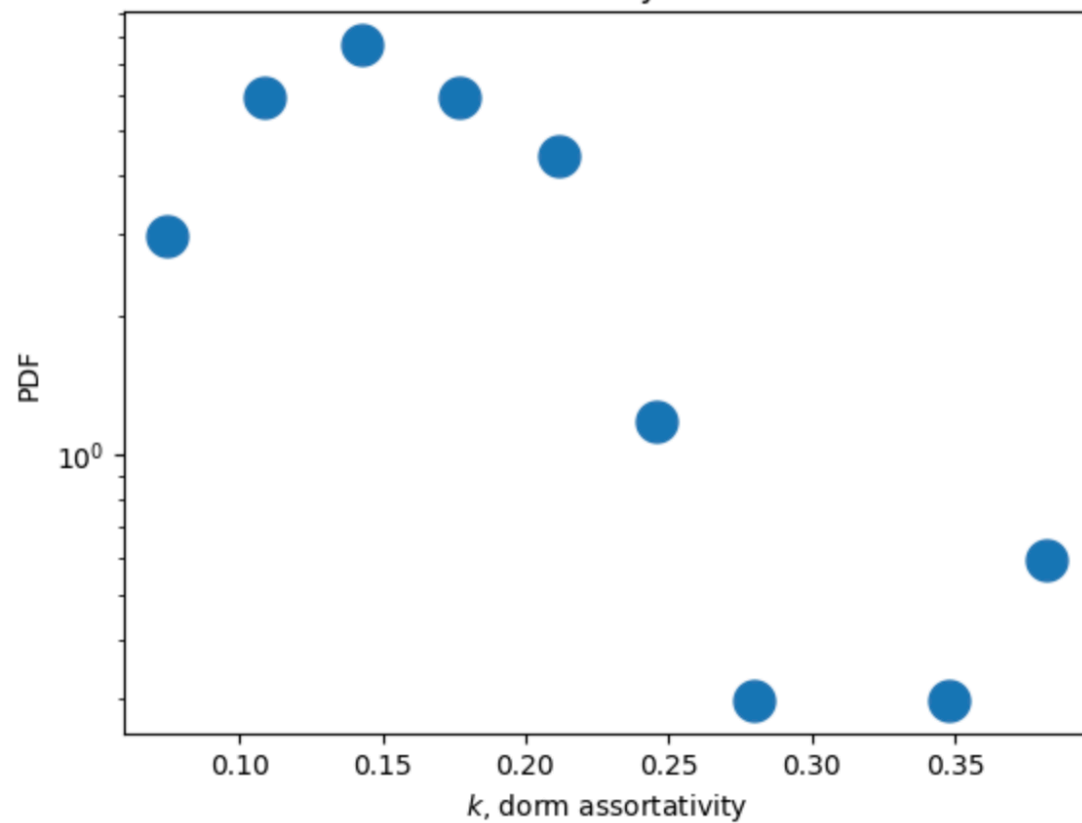
L'assortativité moyenne pour Major est beaucoup plus faible que celle du Statut, sa valeur moyenne se situe autour de 4%, on note qu'il n'y a pas de valeur en dessous de la ligne de

non-assortativité. Ces résultats paraissent logiques dans le sens où tout le monde n'en a pas forcément.

Scatter plot of the dorm assortativity versus the size of the graph



Dorm assortativity distribution

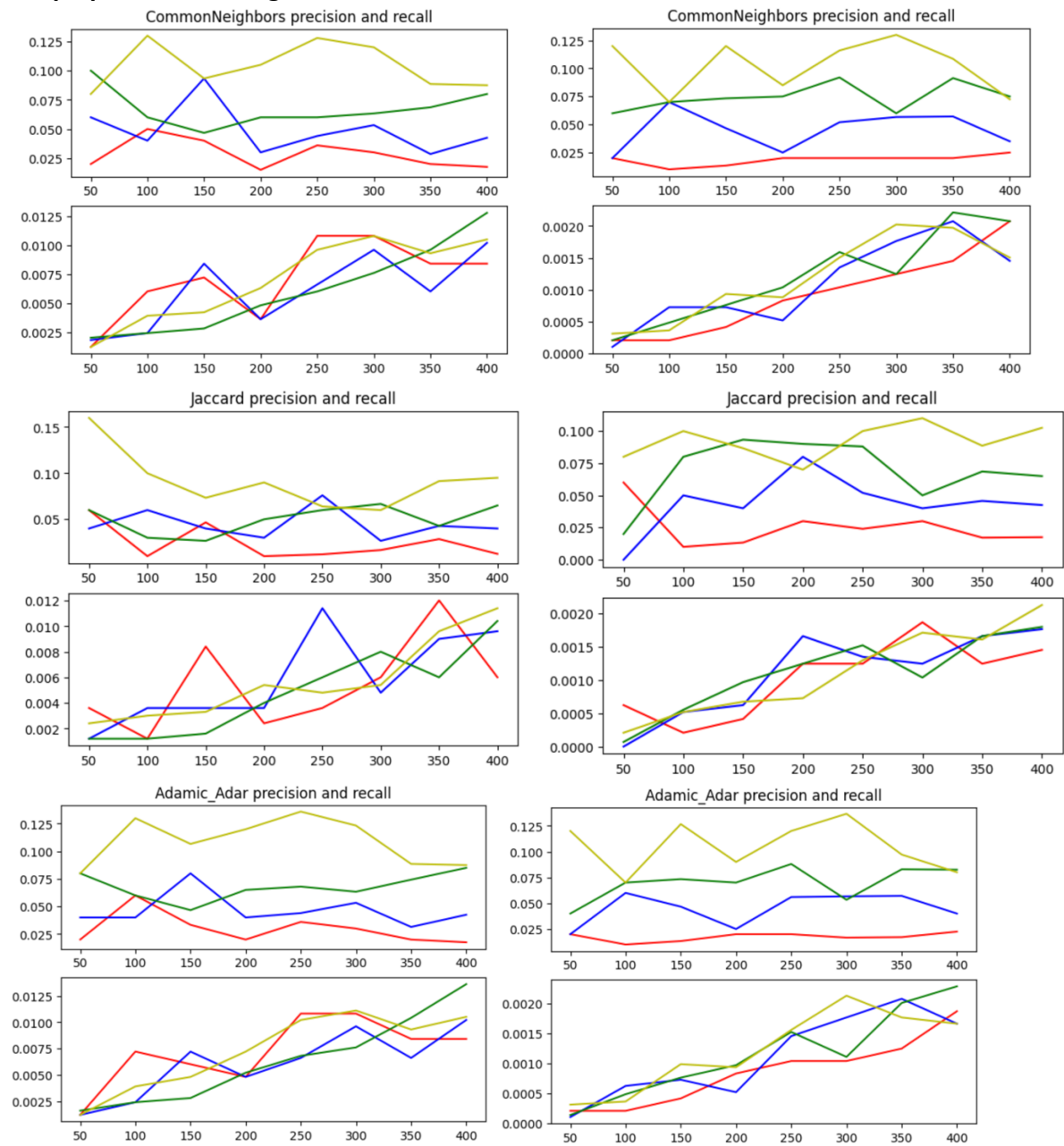


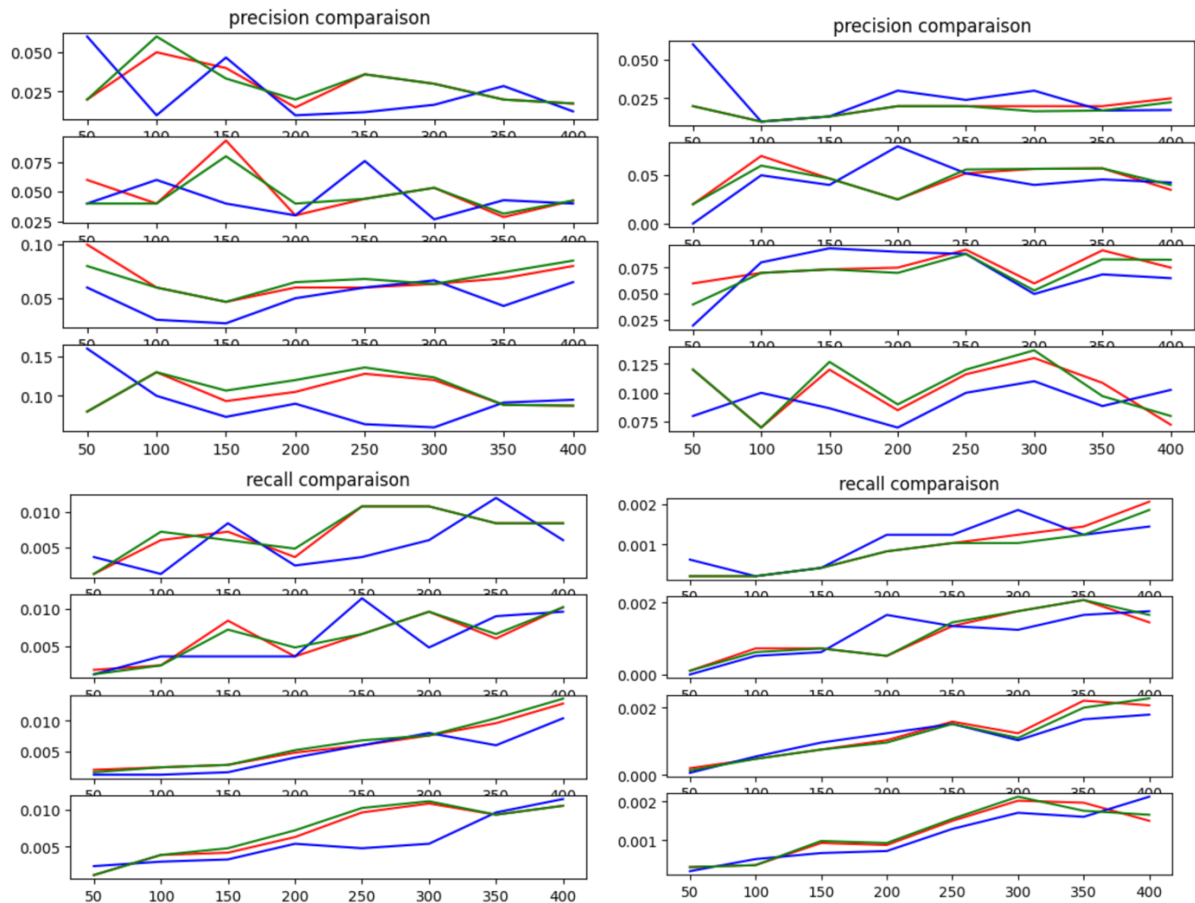
Pour les dortoirs, on observe une assortativité moyenne de 15% et avec un pourcentage faible pour les assortativités plus haute ce qui pourrait paraître logique dans le sens où beaucoup de personnes vivent dans les dortoirs au sein de l'université.

Les attributs Statut, majeurs et dortoirs influencent les liens établies entre les personnes ce qui peut paraître logique au sein d'une université, où les personnes créent des liens avec d'autres personnes qui sont physiquement proches, dans le sens où pour le statut, les personnes du même statut sont assez groupées.

Question 4

Graph pour Caltech à gauche et Hamilton à droite





Les valeurs de précision, topk et recall de Caltech et Hamilton sont dans le fichier Excel Question 4. Elles ont été extraites des dictionnaires créés dans le jupyter notebook.

De manière générale, on note que pour n'importe quel k entre 50 et 400, plus on retire une fraction importante plus les trois critères augmentent. Ce qui est logique car plus on enlève de lien, plus la chance qu'il est une intersection est élevée.

De manière générale aussi, le prédicteur Adamic, Adar et Common Neighbors ont des résultats assez similaires, bien qu'Adamic Adar soit un peu plus précis. On note cependant que Jaccard est ponctuellement plus performant.