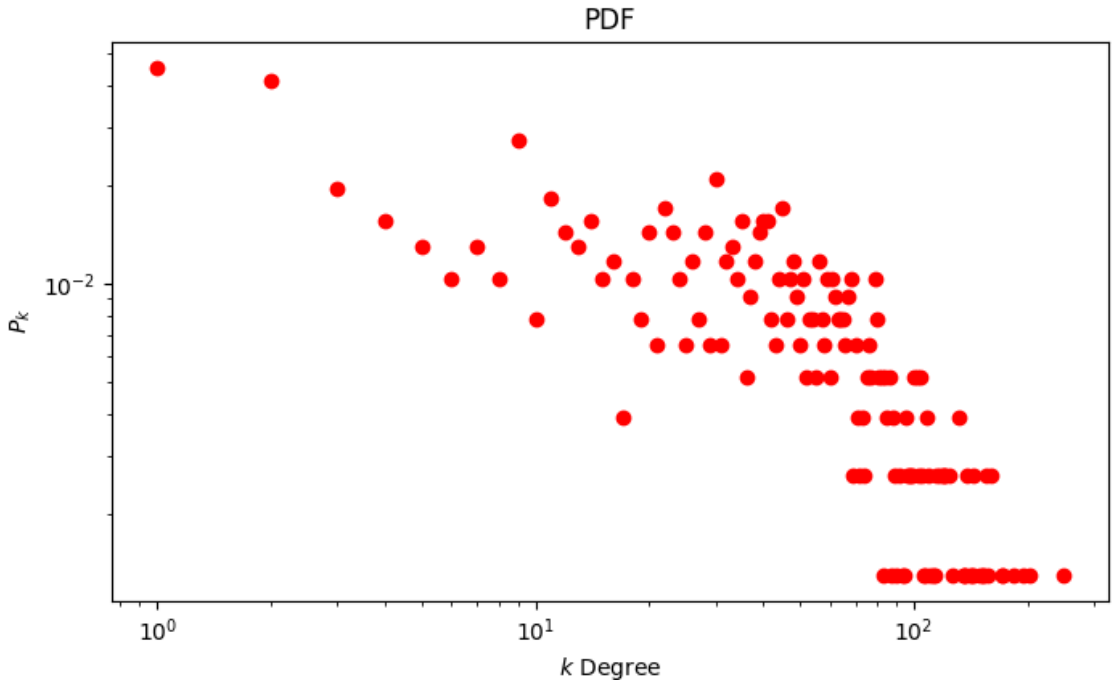
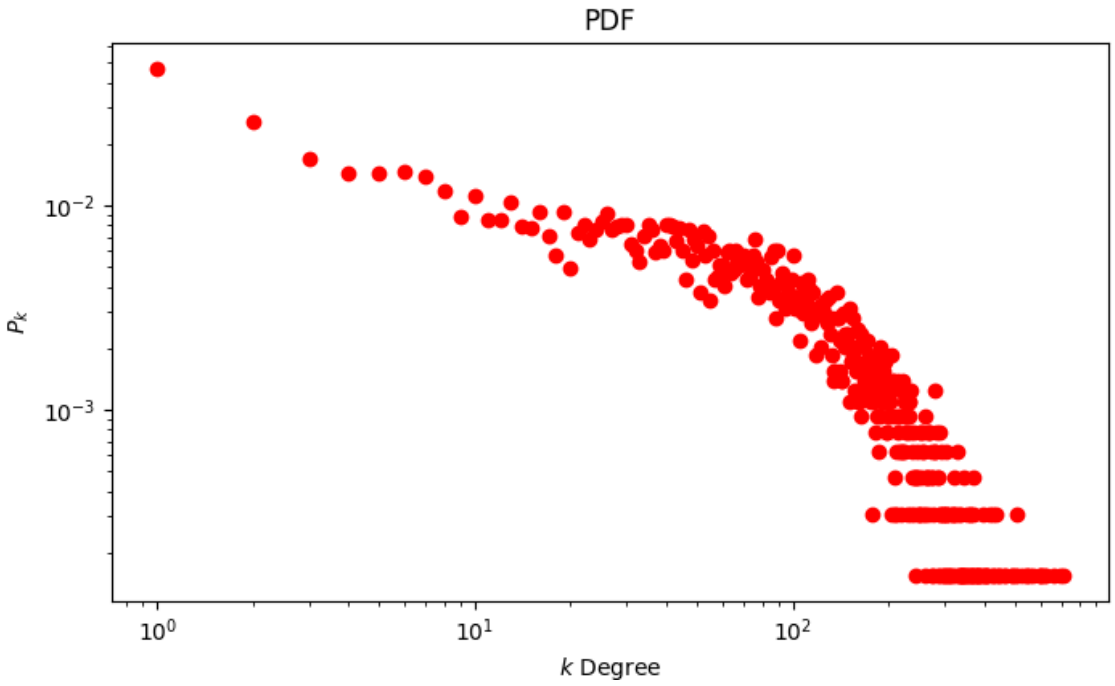
Rendu des résultats du projet de réseaux complexes

Question 2 :

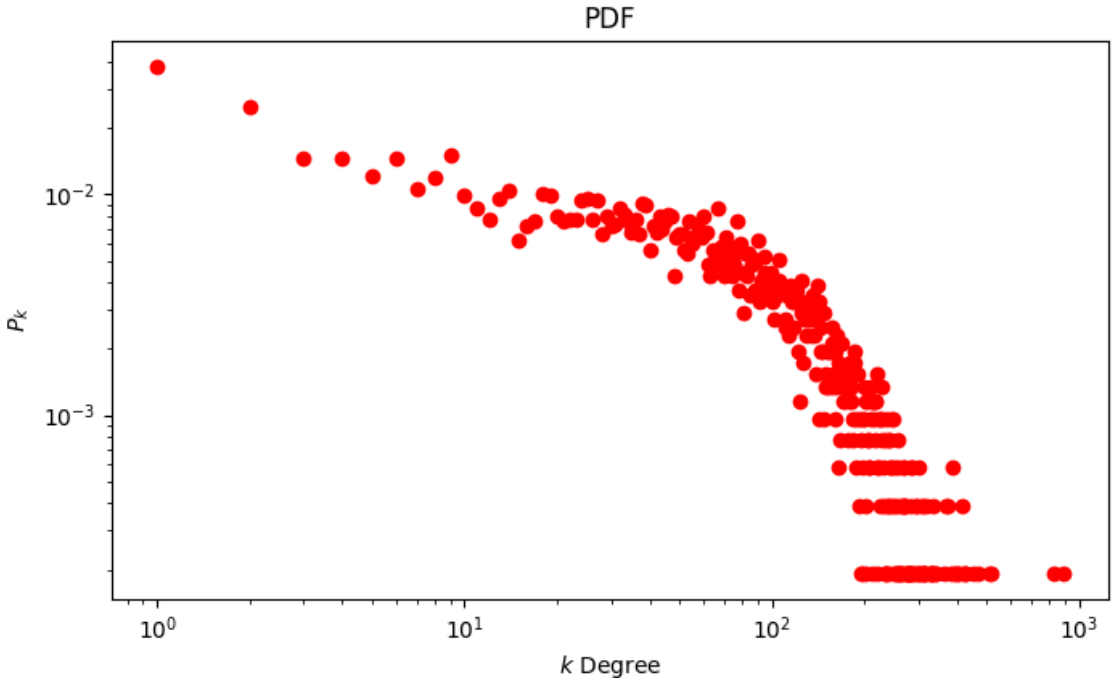
(a)



Degree distribution – CalTech



Degree distribution – MIT



Degree distribution – Johns Hopkins

On constate déjà que Caltech compte bien moins de noeuds que les deux autres universités ce qui confirme l'énoncé (environ 1/10).
Chaque distribution de degré semble suivre une même loi. La probabilité d'avoir un noeud de degré k décroit selon une loi de puissance avec le degré k du noeud.

(b)

Clustering moyen Caltech : 0.40929439048517247
Densité Caltech : 0.05640442132639792

Clustering moyen MIT : 0.2712187419501315
Densité MIT : 0.012118119495041378

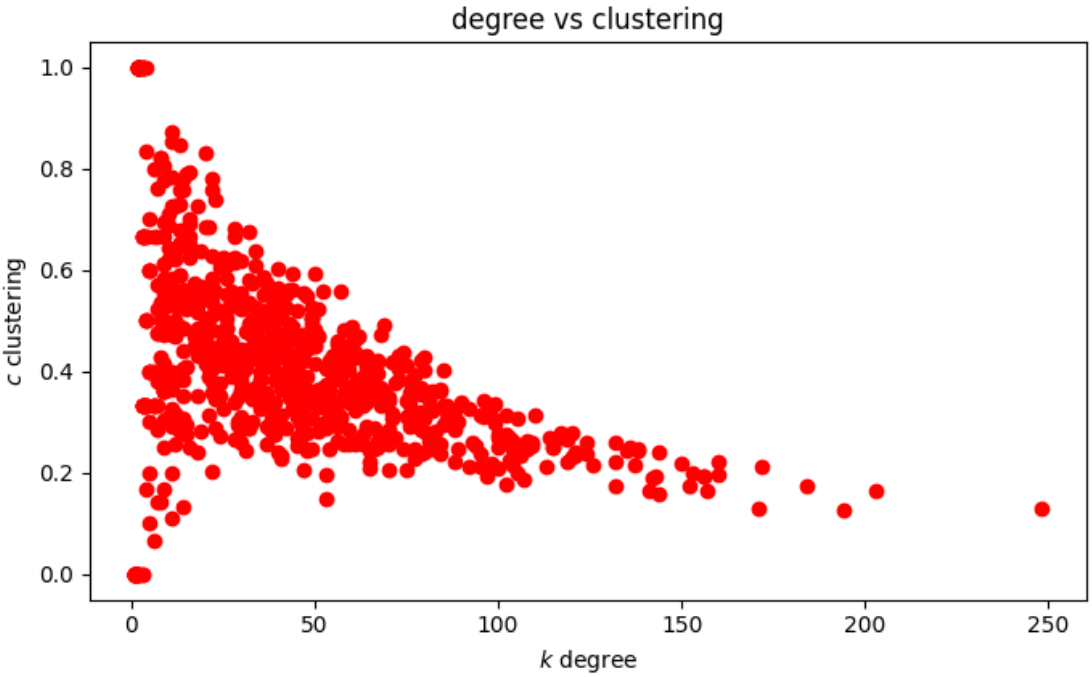
Clustering moyen Johns Hopkins : 0.26839307371293525
Densité Johns Hopkins : 0.013910200162372396

De nouveau, les résultats pour CalTech diffèrent significativement relativement aux deux autres universités.

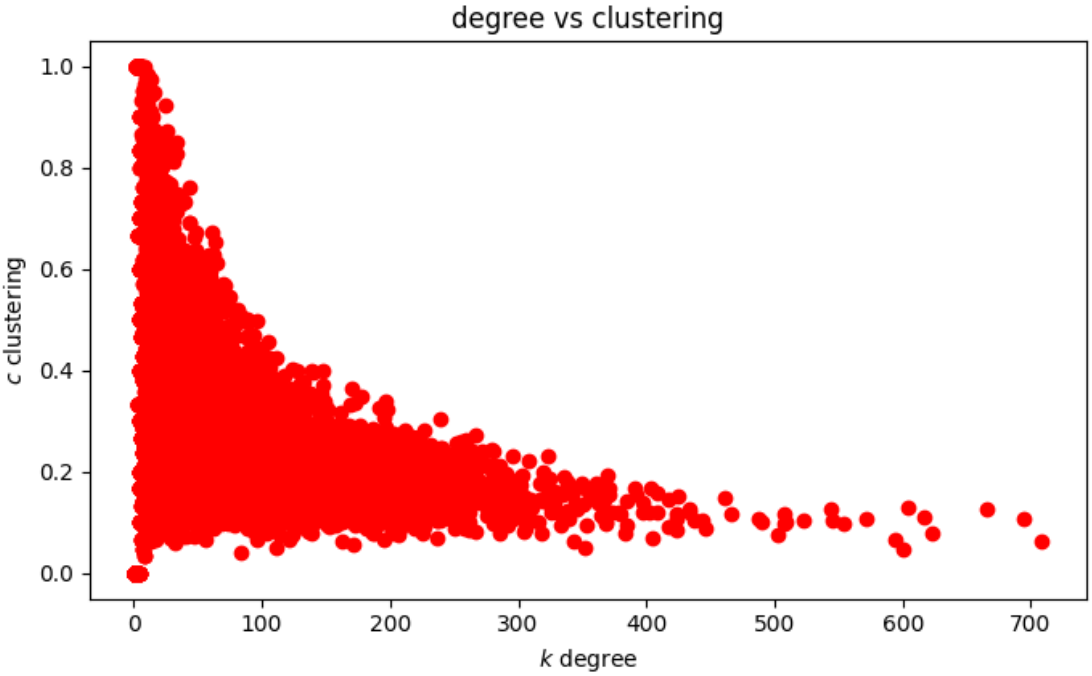
CalTech admet une densité de liens plus faible que les deux autres réseaux (densité de liens < 6%). Ce réseau peut donc être considéré comme peu dense.

De plus, le clustering moyen est plus élevé. Cela peut s'expliquer par le fait que les users à CalTech sont davantage connectés ensemble. Peu d'users mais tous se connaissent probablement.

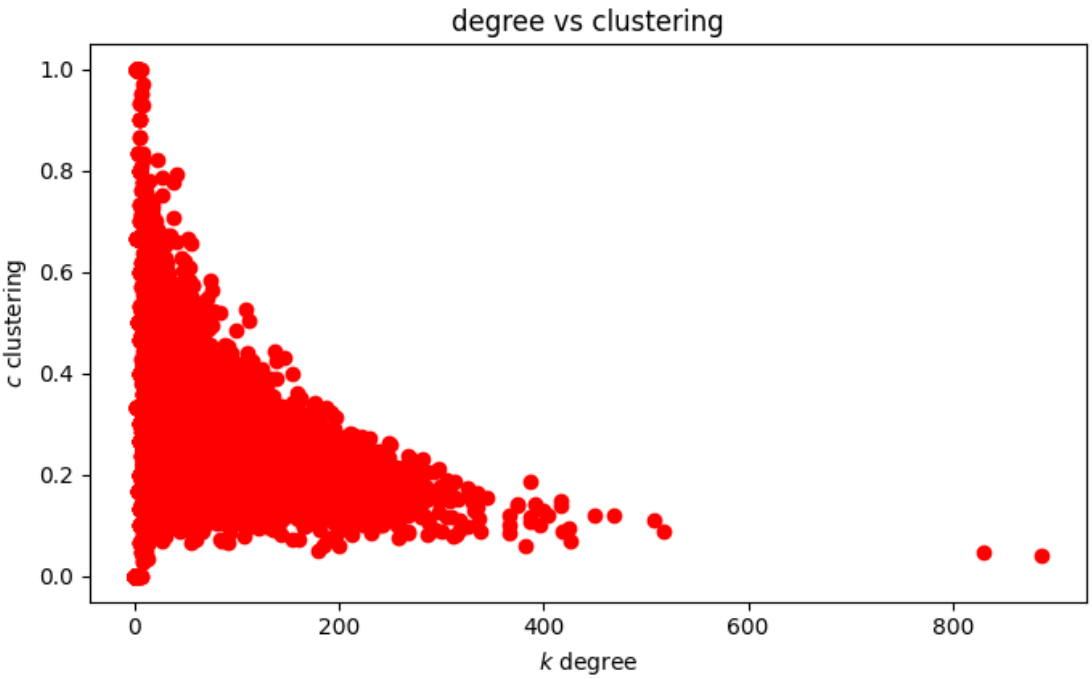
c)



Degree vs clustering – Caltech



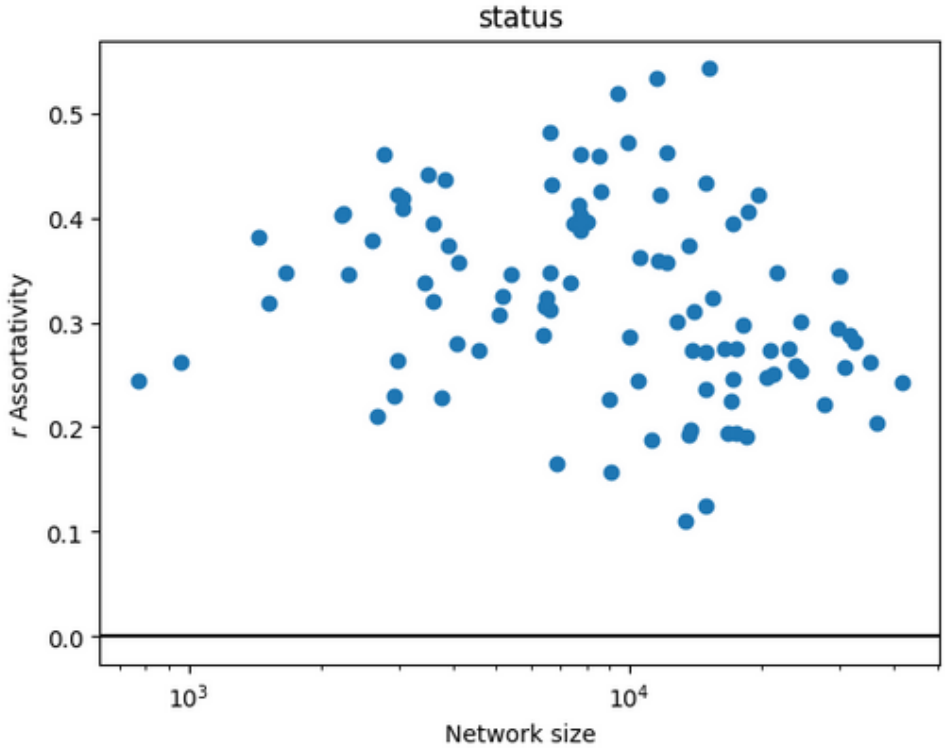
Degree vs clustering – MIT

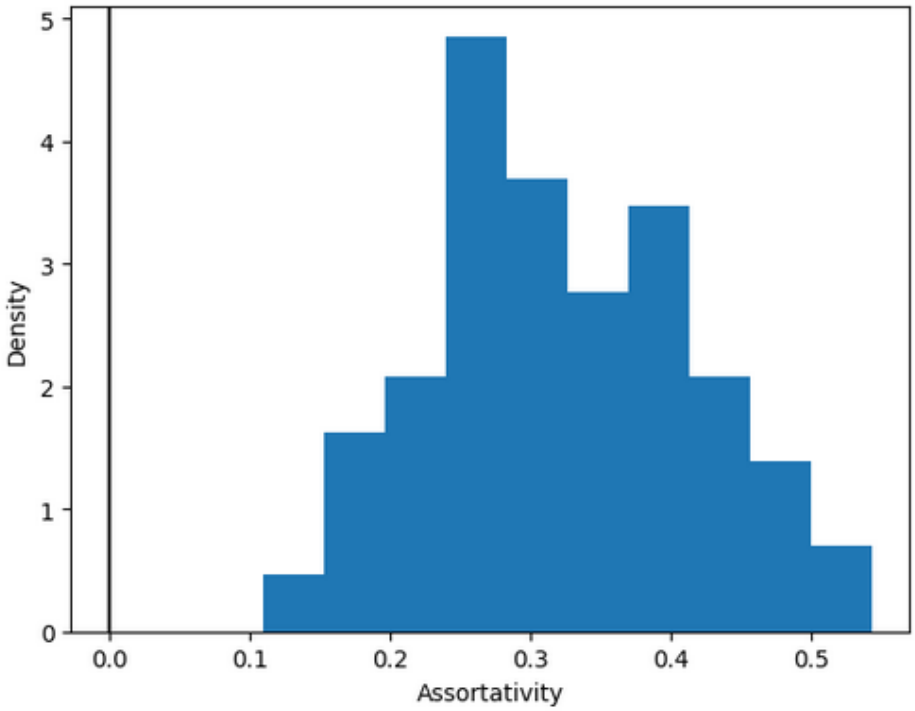


Degree vs clustering – Johns Hopkins

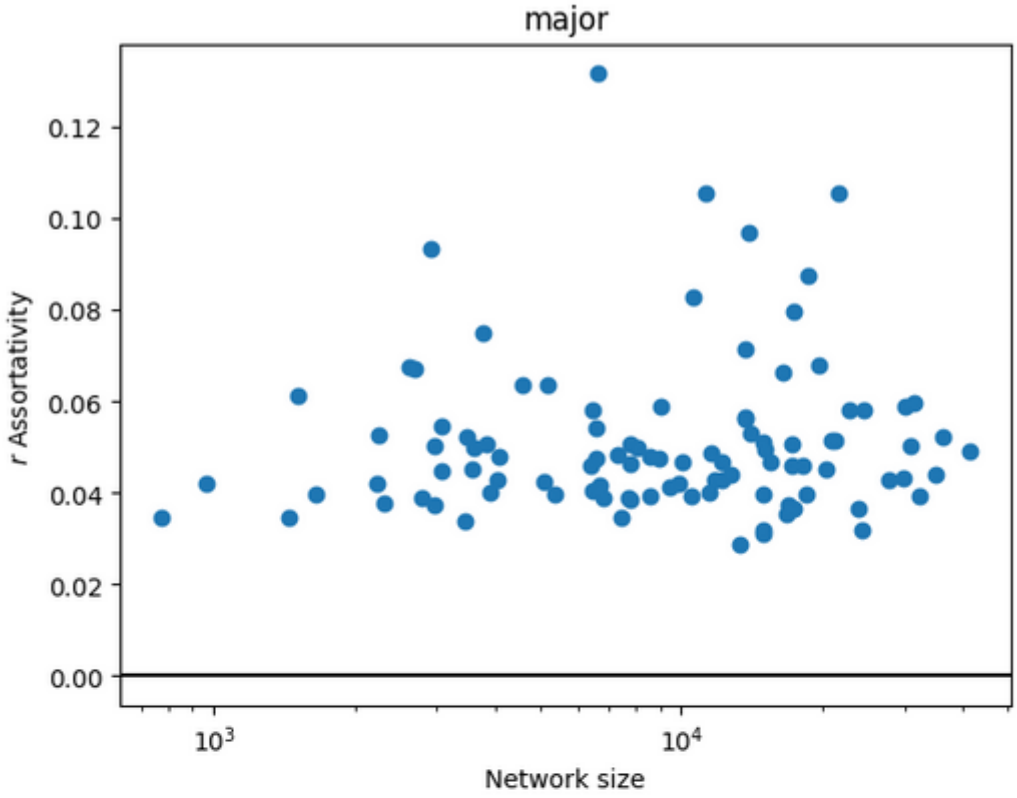
De nouveau, on remarque que Gcaltech possède moins de noeuds que Gjohnshopkins et Gmit. Plus les noeuds sont connectés, plus leur clustering tend vers une valeur asymptotique autour de c = 0.1 pour les 2 graphes les plus peuplés et autour de c = 0.25 pour le graphe de caltech.

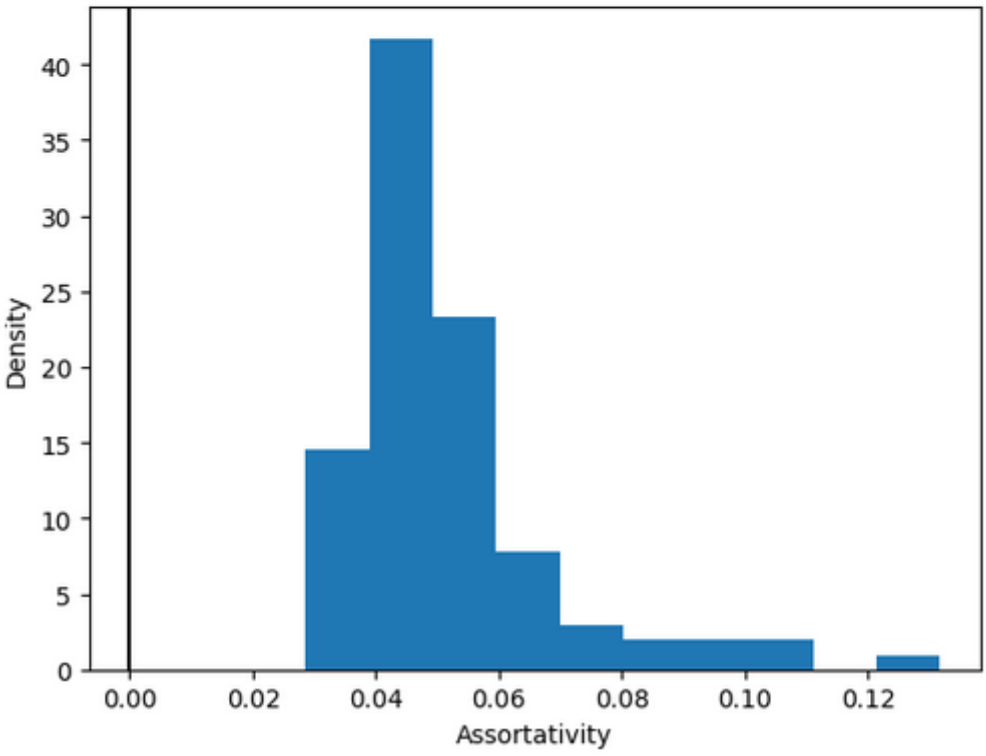
Question 3 :
Status :
On observe que l'assortatvité moyenne tourne autour de 0.25-0.3 ce qui signifie que dans les universités les gens ont des liens qui dépendent à environ 1/4 - 1/3 de leur statut (student ou faculty) ce qui semble naturel. De plus l'assortativité minimale est de 11 %. Donc, le statut joue un rôle « important » dans les liens sociaux même pour les extrema de l'échantillon.



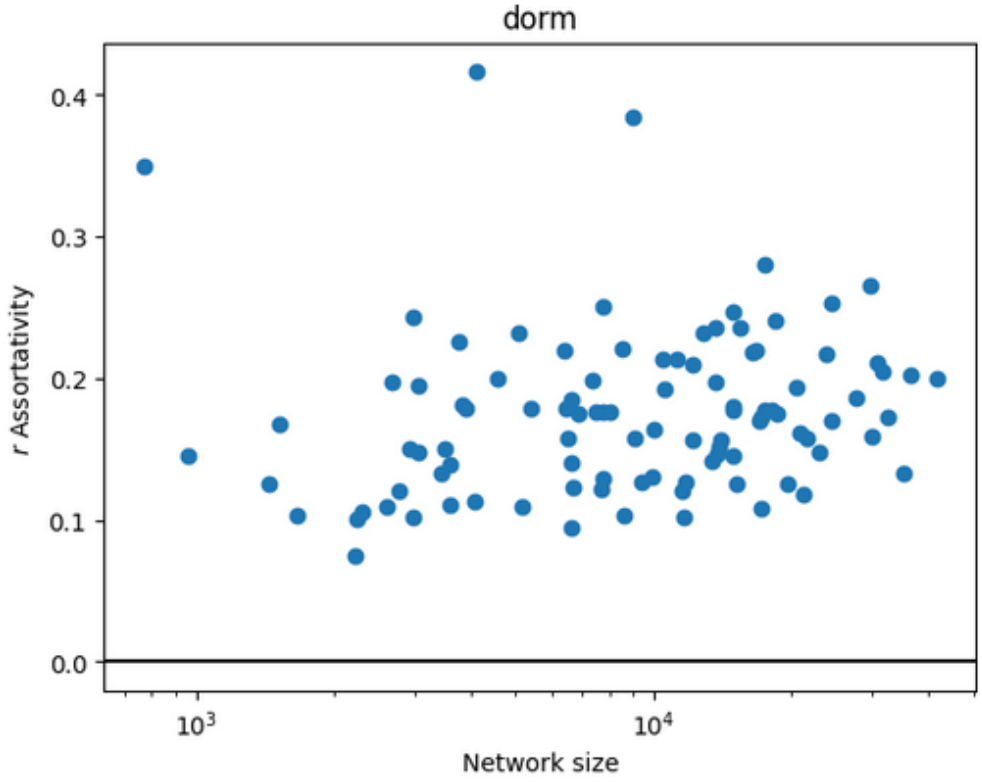


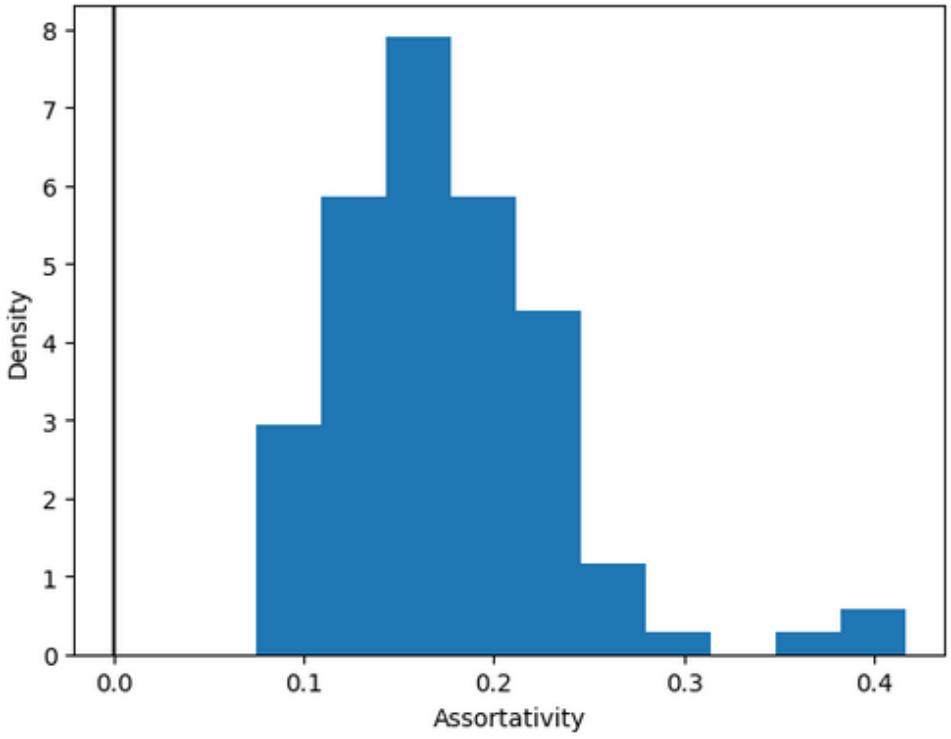
Major : On observe que l’assortatvité moyenne tourne autour de 0.04 ce qui signifie que dans les universités les gens ont des liens qui dépendent à environ 4 % de leur major ce qui semble naturel. De plus, l’écart-type semble relativement faible. Donc, globalement, quelque soit la major, les gens ont des liens sociaux qui dépendent faiblement de leur major.



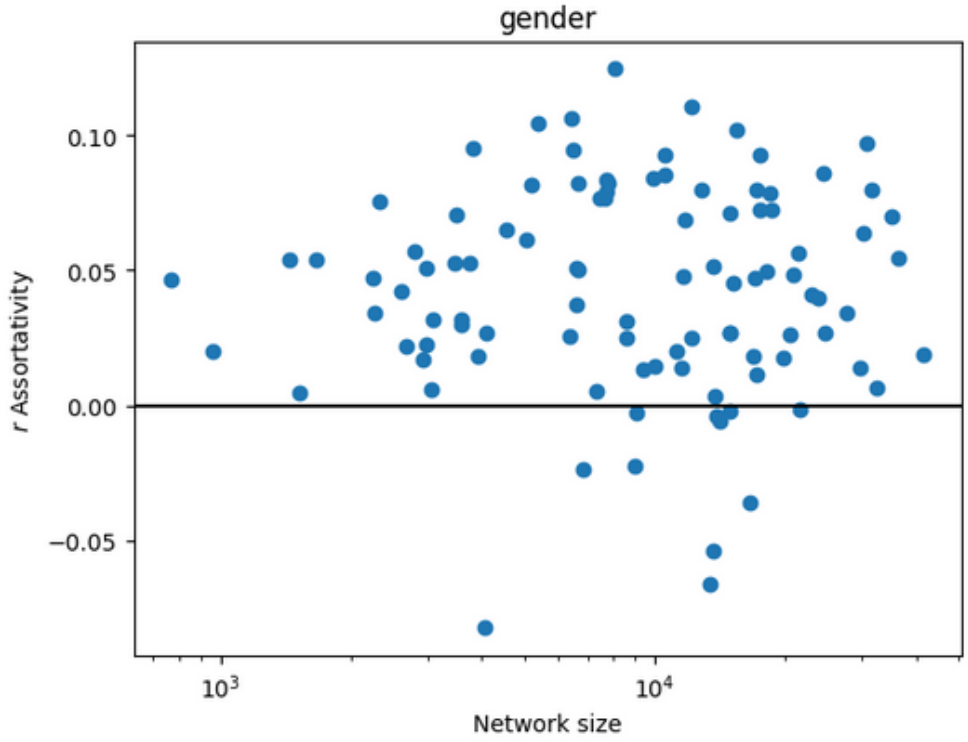


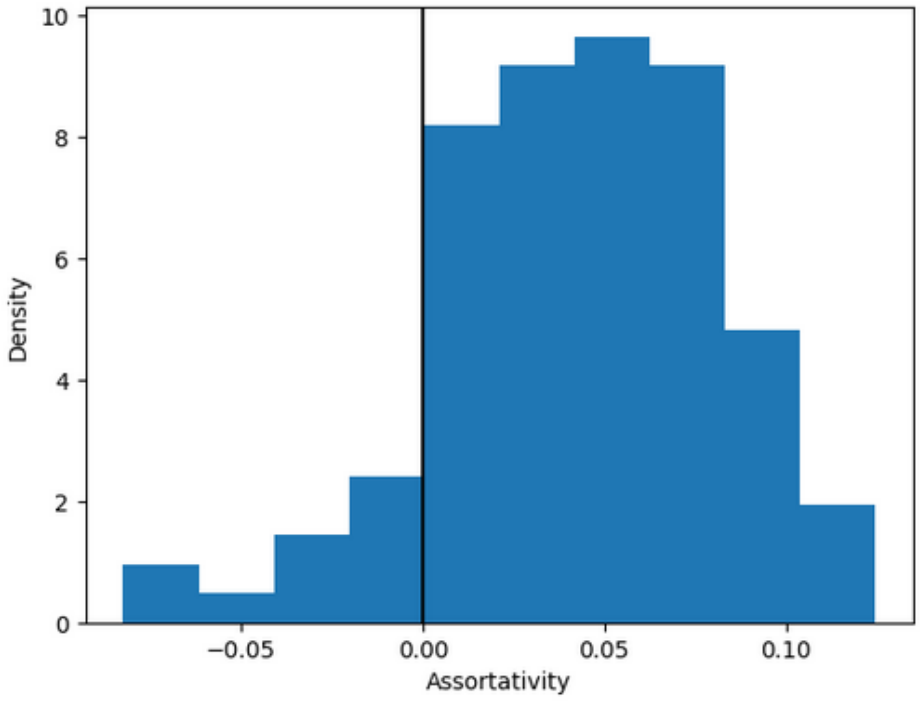
Dorm : On observe que l’assortatvité moyenne tourne autour de 0.15 – 0.2 ce qui signifie que dans les universités les gens ont des liens qui dépendent à environ de 1/6 de leur dorm ce qui semble naturel. En efffet, les gens ont plus tendance à avoir des liens sociaux avec leurs voisins.





Gender : On observe que l’assortatvité moyenne tourne autour de 4 ou 5 % ce qui signifie que dans les universités les gens ont des liens qui dépendent peu de leur genre comparé à leur dorm et status ce qui semble naturel. De plus, les points d’assortativité négative correspondent peut-être à des personnes ayant beaucoup de liens sociaux avec des individus du genre opposé.





Question 4 :

(d)

Il est venu à ma compréhension que les métriques topk et precisionk désignent après simplification ensembliste la même métrique de liens. Ce sont tous deux le rapport du nombre de liens dans l'intersection de Epredictk et Eremoved et du nombre de liens dans Epredictk.
Recallk désigne, quant à elle, le nombre de liens bien prédits par rapport au nombre de liens qu'il aurait fallu prédire (ceux enlevés). Recallk s'assimile à un taux de succès de la métrique de prédiction.

Common Neighbors :

Gcaltech :

top50 = 8 % top100 = 10 % top200 = 9 % top400 = 9.5 %

recall50 = 0.0012008405884118883
recall100 = 0.003002101471029721
recall200 = 0.005403782647853498
recall400 = 0.011407985589912939

precision50 = 0.08 precision100 = 0.1 precision200 = 0.09 precision400 = 0.095

On observe que la précision oscille autour de 9 %. Recallk (le taux de réussite) augmente quasiment linéairement avec k. (fois 2 entre recall(k-1) et recallk). Recallk tourne autour de 5-10 %.

Gamerican :

top50 = 12 % top100 = 15% top200 = 10.5% top400 = 8.75 %

recall50 = 0.0018012608826178324
recall100 = 0.004503152206544582
recall200 = 0.006304413089162414
recall400 = 0.010507355148604023

precision50 = 0.12 precision100 = 0.15 precision200 = 0.105 precision400 = 0.0875

On observe que la précision oscille autour de 11 %. Recallk (le taux de réussite) augmente quasiment linéairement avec k entre recall50 et recall100. Cependant, l'augmentation est plus faible ensuite (fois 1,5). Recallk tourne autour de 6 %.

Jaccard :

Gcaltech :

top50 = 4 % top100 = 10 % top200 = 6.5 % top400 = 8.5 %

recall50 = 0.0006004202942059442
recall100 = 0.003002101471029721
recall200 = 0.003902731912338637
recall400 = 0.01020714500150105

precision50 = 0.04 precision100 = 0.1 precision200 = 0.065 precision400 = 0.085

On observe que la précision oscille autour de 7%. Recallk tourne autour de 6 %. Recall400 = max des recallk

Gamerican :

top50 = 12 % top100 = 7.000000000000001% top200 = 9% top400 = 10.25 %

recall50 = 0.00013782964256179363
recall100 = 0.00016080124965542588
recall200 = 0.0004134889276853809
recall400 = 0.000941835890838923

precision50 = 0.12 precision100 = 0.07000000000000001 precision200 = 0.09
precision400 = 0.1025

On observe que la précision oscille autour de 10 %. Recallk (le taux de réussite) augmente avec k. Recallk tourne autour de 1-10 %.

Adamic/ Adar:

Gcaltech :

top50 = 12 % top100 = 13 % top200 = 10 % top400 = 8.5 %

recall50 = 0.0018012608826178324
recall100 = 0.003902731912338637
recall200 = 0.006004202942059442
recall400 = 0.01020714500150105

precision50 = 0.12 precision100 = 0.13 precision200 = 0.1 precision400 = 0.085

On observe que la précision oscille autour de 11 %. Recallk (le taux de réussite) augmente avec k similairement à la métrique Jaccard. Recallk tourne autour de 6 %.

Gamerican :

top50 = 24 % top100 = 22% top200 = 21.5% top400 = 18.5 %

recall50 = 0.00027565928512358727
recall100 = 0.0005053753560599099
recall200 = 0.0009877791050261877
recall400 = 0.001699898924928788

precision50 = 0.24 precision100 = 0.22 precision200 = 0.215 precision400 = 0.185

On observe que la précision oscille autour de 20%. Recallk (le taux de réussite) augmente avec k similairement à la métrique Jaccard. Recallk tourne autour de 8 %.

Quelque soit la métrique la précision tourne autour de 10 %, excepté pour la métrique Adamic/ Adar sur le le plus gros graphe Gamerican. (sachant que Gamerican est environ 10 fois plus gros que Gcaltech).

Concernant le taux de résussite (recallk), il augmente avec k.
Pour la métrique Common Neighbors, le taux de résussite augmente a peu près linéairement avec k.
Pour les deux autres métriques, le taux de réussite augmente faiblement avec k « faible » (k = 50, 100) mais augmente vite (plus que linéaire) quand k augmente.