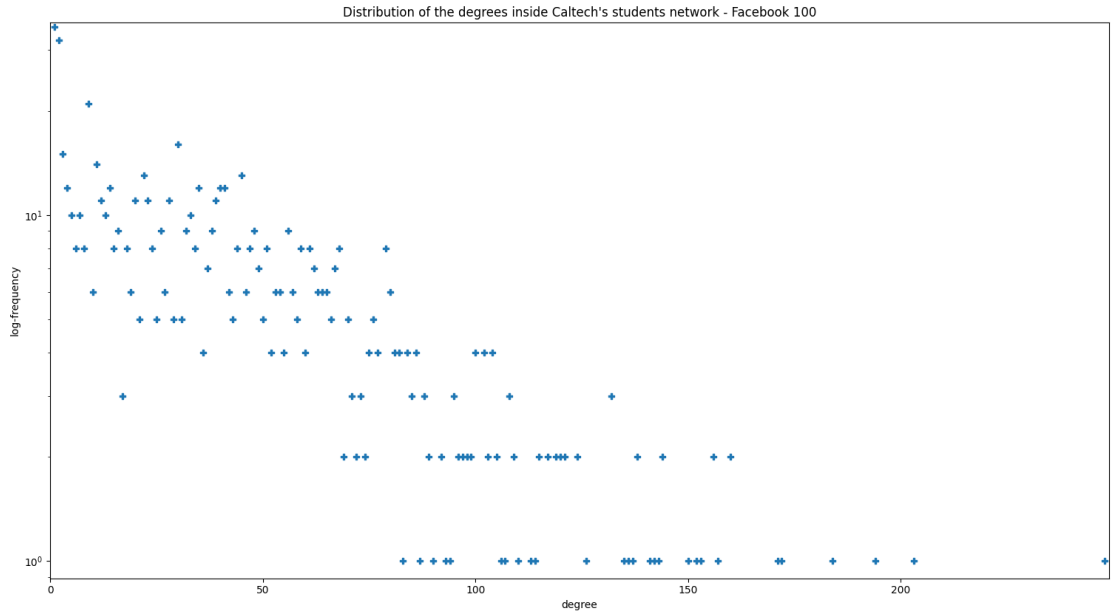
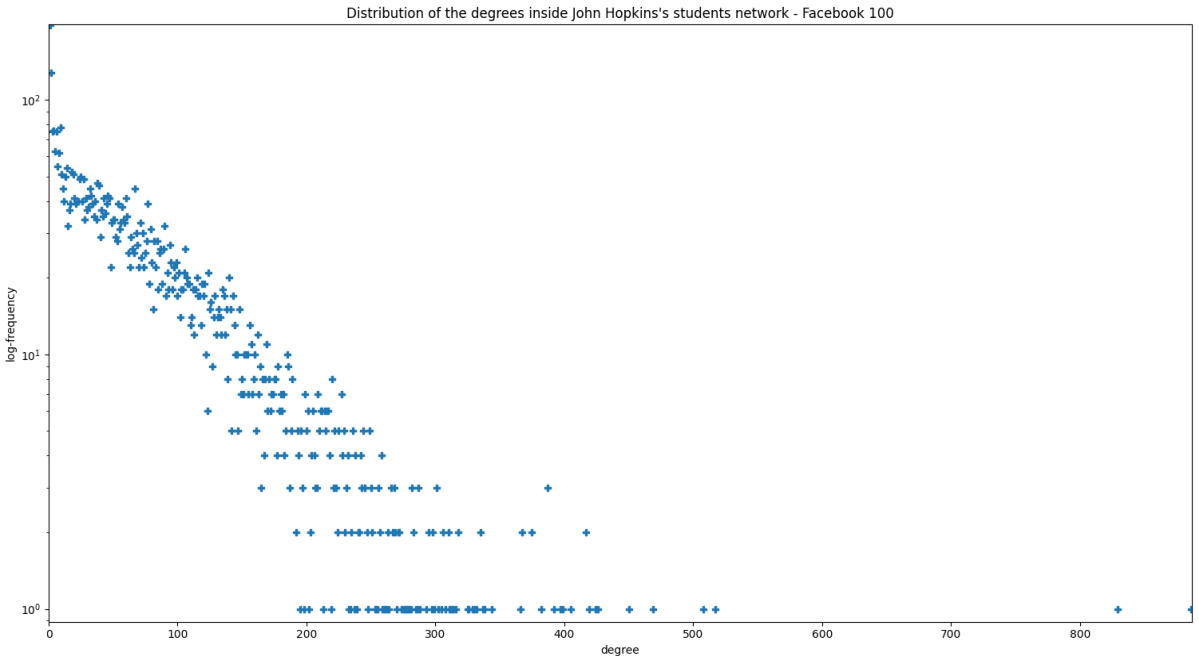
DEGRÉ Mathieu

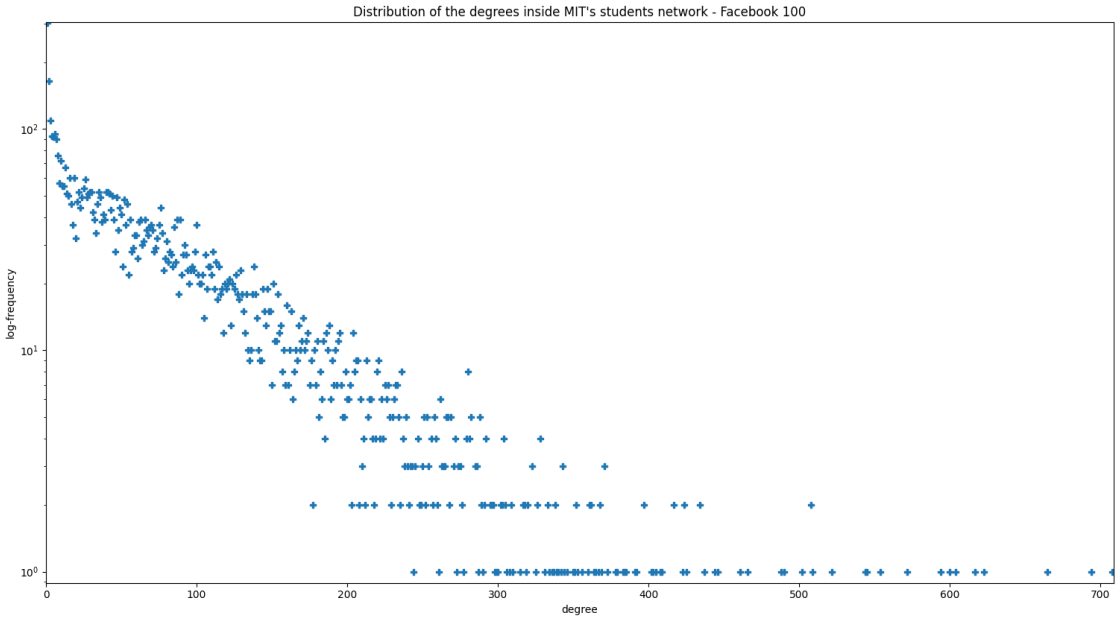
NET 4103/7431 Homework
Network science and Graph Learning

2) a)





DEGRÉ Mathieu



Ces trois distributions de degrés nous montre que les graphes les plus petits (en terme de nombre de nœuds) respectent peu la distribution en loi de puissance, et que plus le graphe est grand plus on se rapproche de cette distribution. Les graphes ici sont cependant trop petits pour pouvoir voir émerger clairement cette loi de distribution.

2) b)

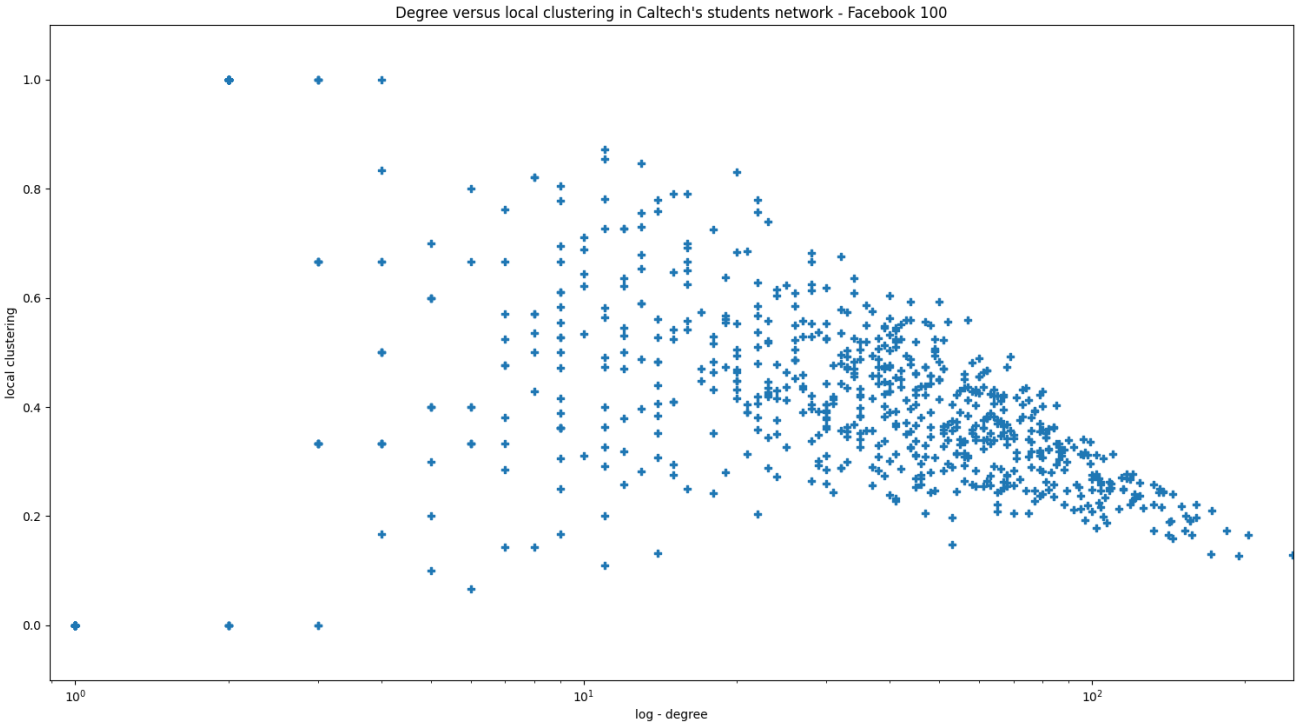
Le résultat de la fonction question2b() est ceci

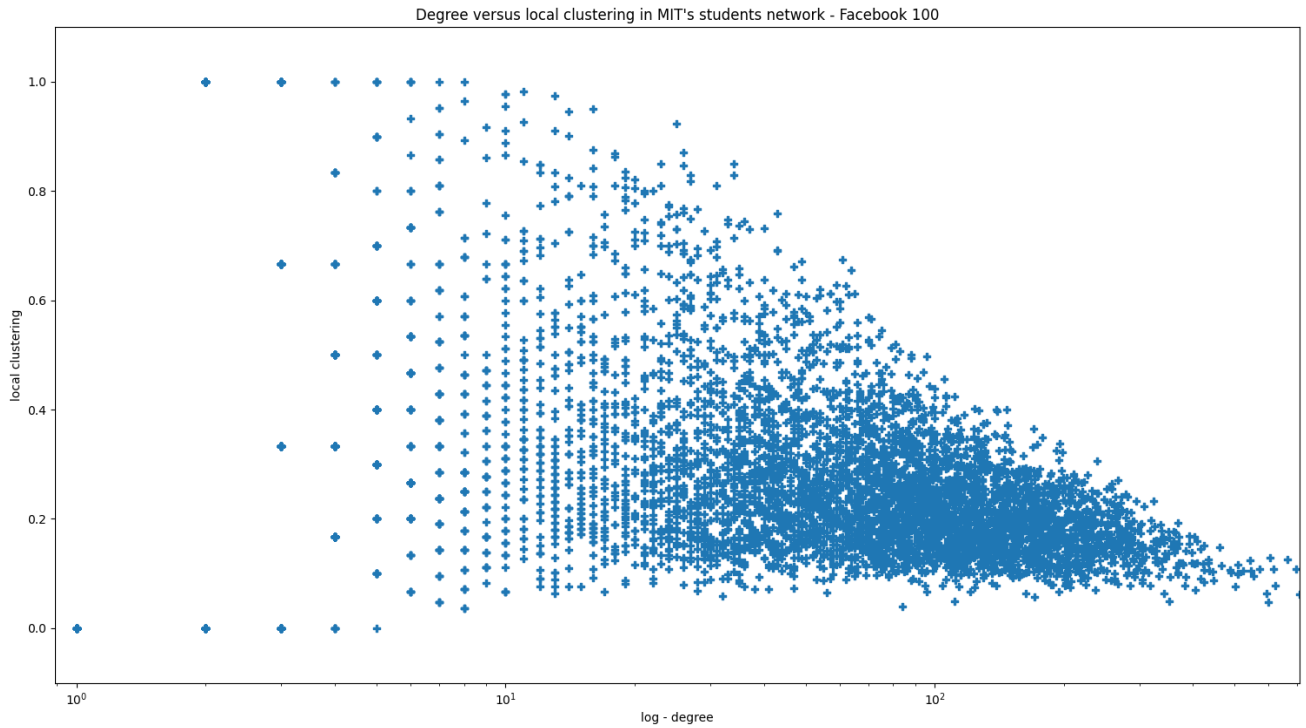
Caltech's GCC: 0.2912826901150874
Caltech's average clustering: 0.40929439048517247
Caltech's edge density: 0.05640442132639792
MIT's GCC: 0.18028845093502427
MIT's average clustering: 0.2712187419501315
MIT's edge density: 0.012118119495041378
John Hopkinks' GCC: 0.19316123901594015
John Hopkinks' average clustering: 0.26839307371293525
John Hopkinks' edge density: 0.013910200162372396

On peut remarquer que le GCC du réseau de chaque université est relativement faible, nettement plus que le coefficient local. Cela veut donc dire que les graphes ne sont pas très dense et plutôt relativement clairsemés, les étudiants étant nettement plus connectés aux amis de leurs amis qu'aux autres personnes du graphe. Cela est confirmé par la densité de chaque graphe, très faible.

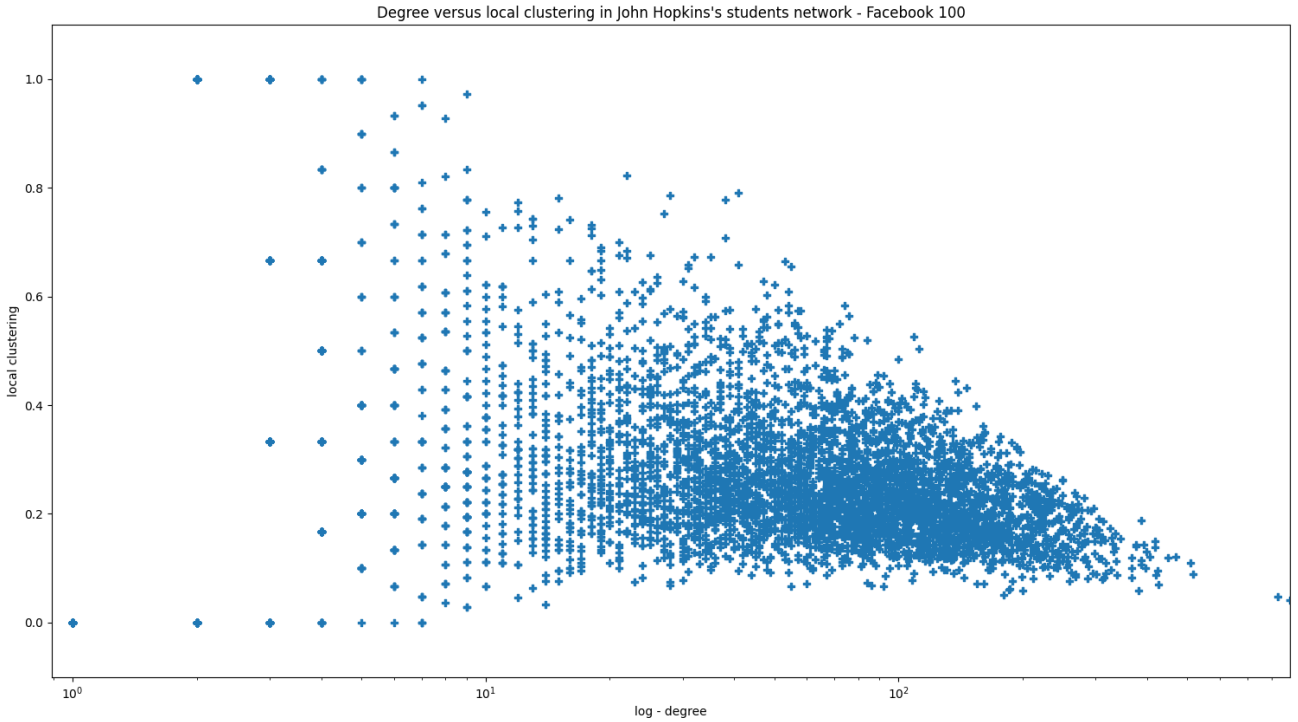
DEGRÉ Mathieu

2) c)



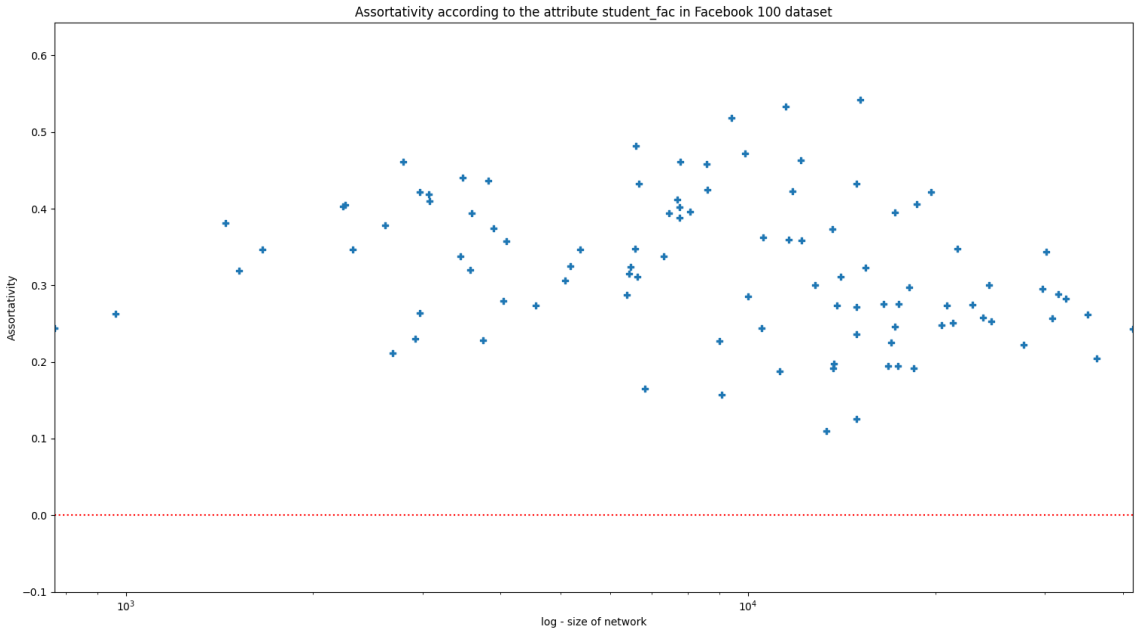


DEGRÉ Mathieu

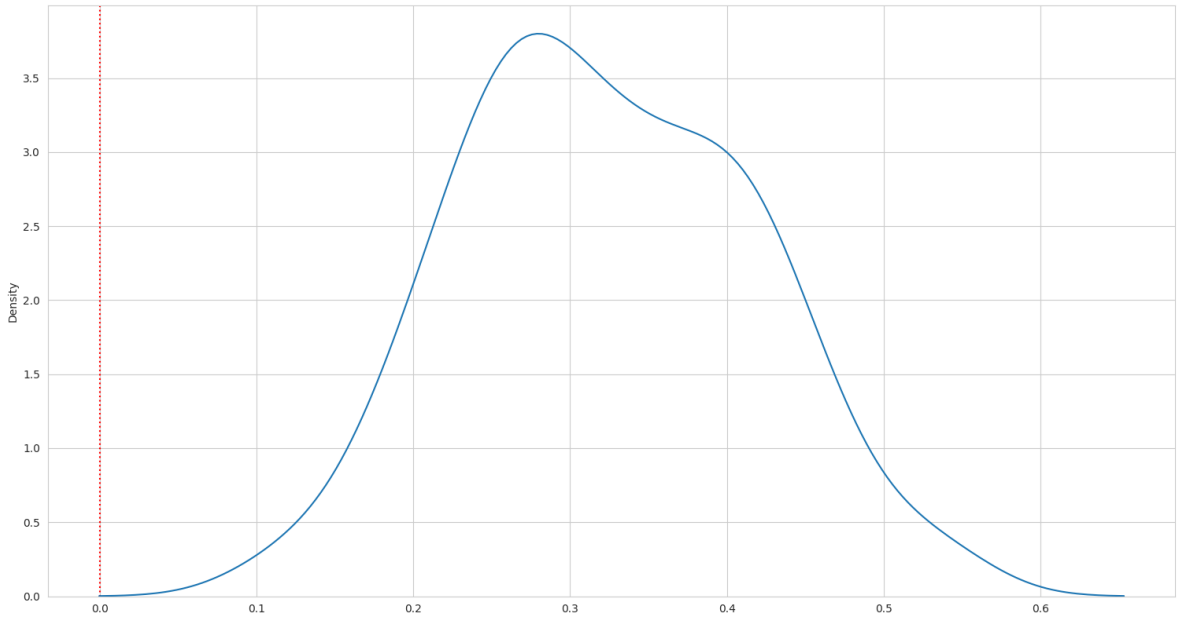


On observe dans ces trois graphes la même tendance: plus le degré d'un nœud est élevé, moins il peut atteindre un local clustering élevé. Cela confirme notre analyse précédente, à savoir que ces graphes ne sont pas denses, puisque les nœuds ayant un degré élevé n'ont qu'une faible probabilité de voir leurs amis être amis. Même si cette tendance se note sur les 3 graphes, on peut quand même constater que, pour celui de Caltech, qui est le graphe avec le plus faible nombre de noeuds, les noeuds à degré élevé ont un local clustering minimum sensiblement plus élevé (~0.2) que ceux des deux autres universités (~0.1).

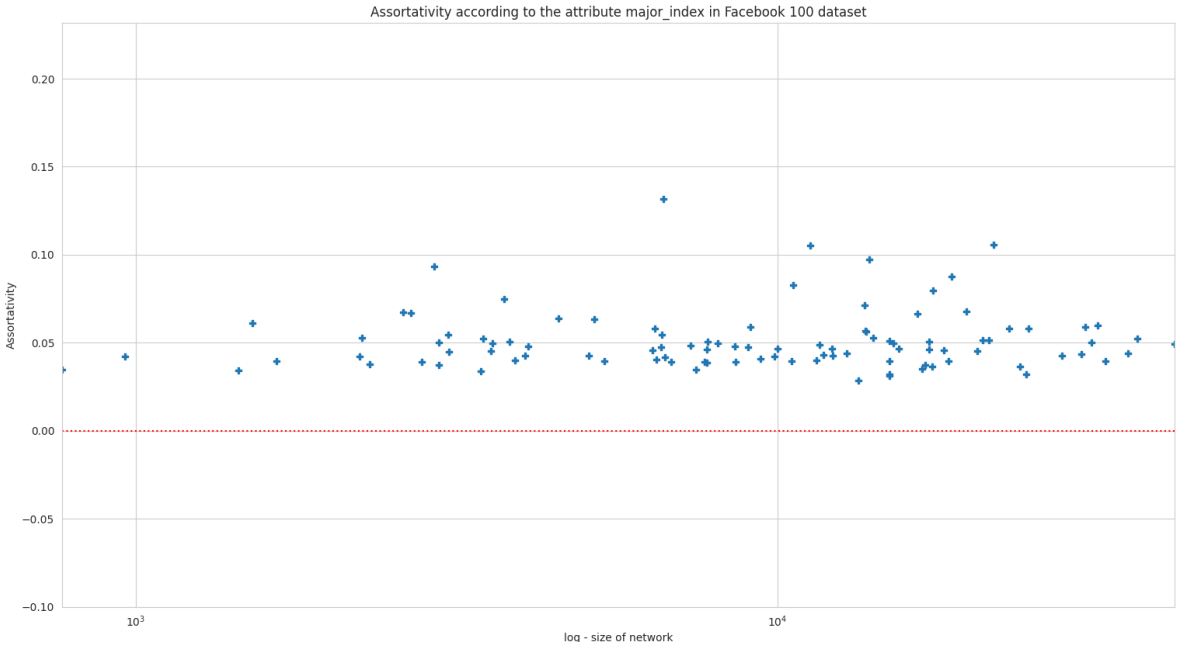
3) a)



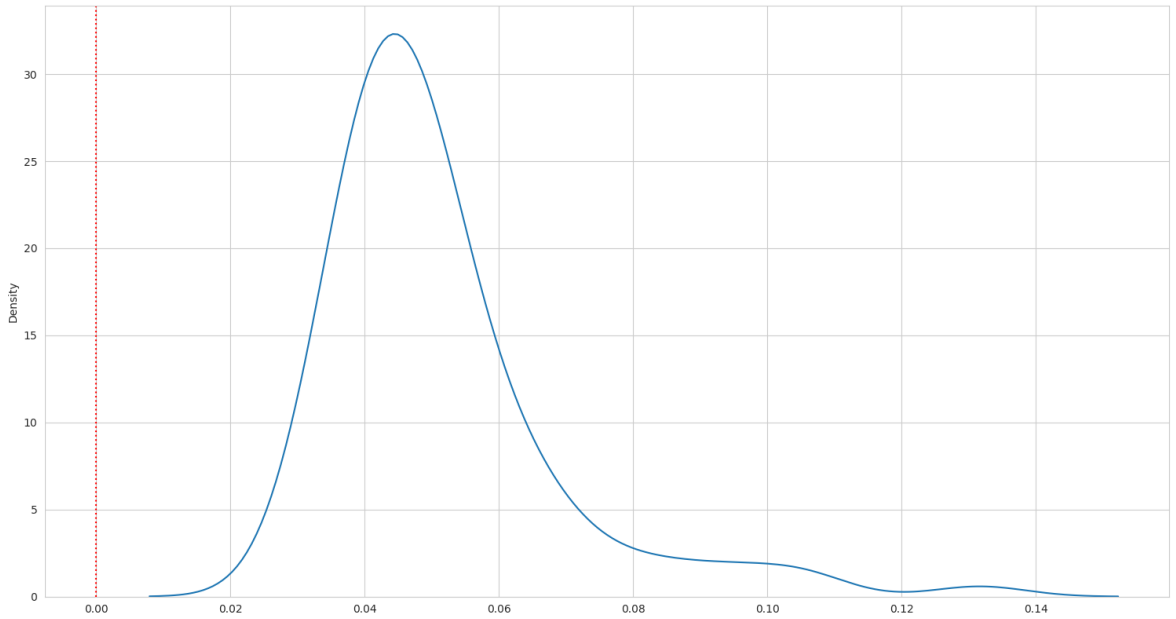
DEGRÉ Mathieu



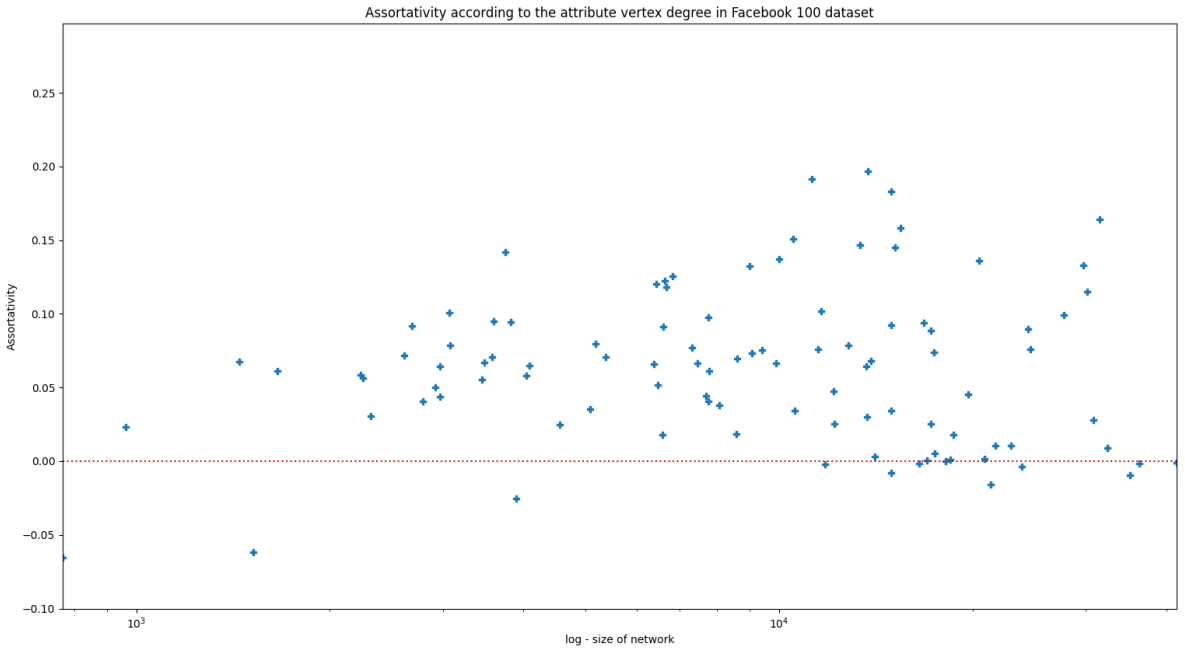
Densité de répartition de l’assortativité pour le facteur “student_fac”



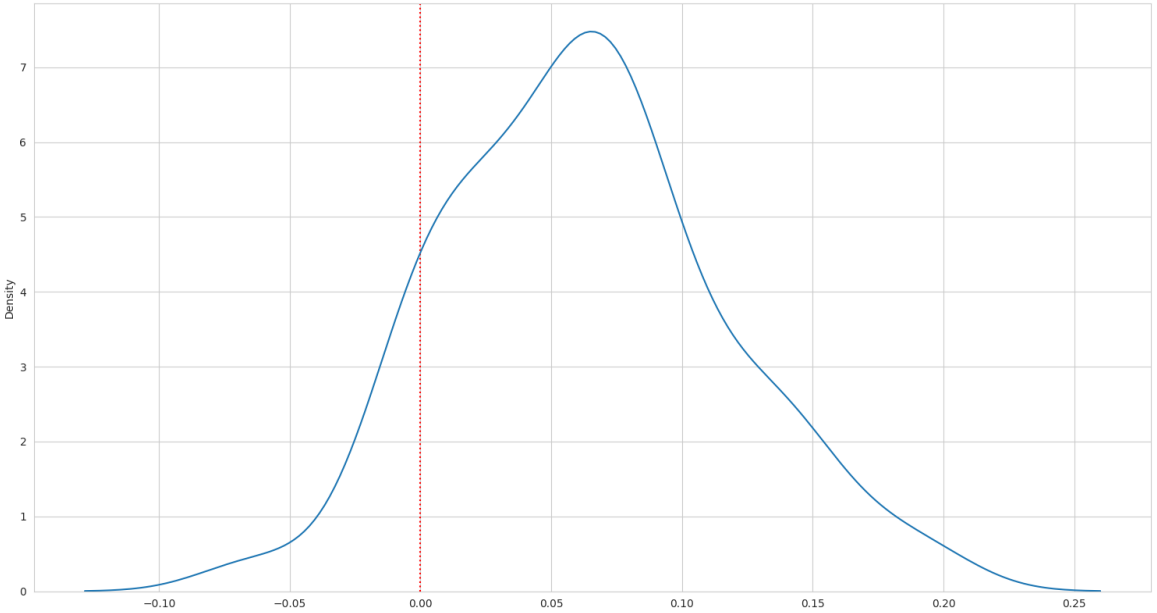
DEGRÉ Mathieu



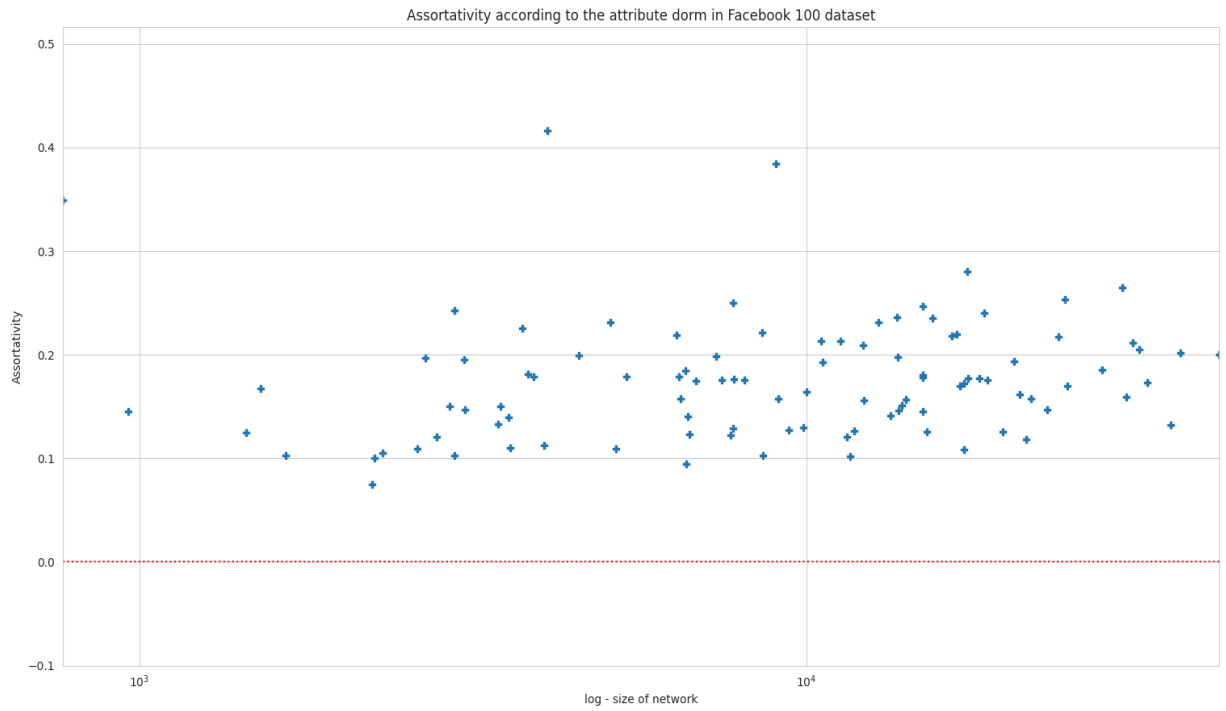
Densité de répartition de l’assortativité pour le facteur “major_index”

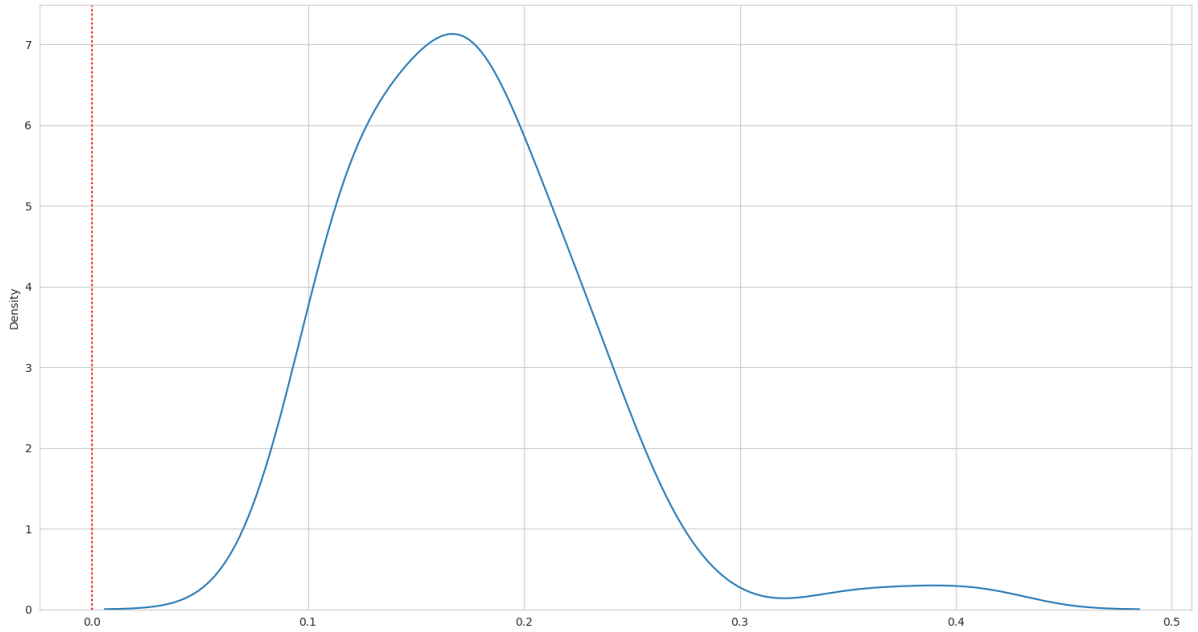


DEGRÉ Mathieu



Densité de répartition de l’assortativité pour le facteur “vertex degree”





Densité de répartition de l'assortativité pour le facteur "dorm"

La première chose qui saute aux yeux en analysant ces données est à quel point le paramètre student_fac est plus assortatif que les autres, avec un pic à 0.3. Cela peut facilement s'expliquer par le fait que les élèves qui viennent d'une même faculté se connaissent ou ont eu l'occasion de se connaître depuis nettement plus longtemps et sont ainsi plus amènes à forger des liens d'amitié. De même, des élèves présents dans le même internat ont une probabilité plus élevée d'avoir forgé des liens d'amitié, d'où le fait que le facteur dorm soit très assortatif lui aussi, avec un pic à 0.2. Les deux autres facteurs apparaissent eux aussi légèrement assortatif, même si nettement moins que les deux autres. Cela peut s'expliquer par le fait que ces deux facteurs sont beaucoup moins spécifiques que les autres: un major spécifique va regrouper beaucoup plus d'élèves qu'un dortoir spécifique, et donc favorise moins l'apparition de liens d'amitié, même s'il les facilite légèrement de manière assez logique, et on peut faire la même analyse pour le paramètre vertex_degree.

4)

Les tests ont été réalisés avec le paramètre n=500 sur les 11 universités possédant les plus petits graphes afin d'éviter un temps de calcul trop long, ce calcul ayant déjà nécessité environ 5 heures.
Les données brutes sont disponibles ici:
Ce que l'on peut constater, c'est que l'algorithme Adamic / Adar semble être le plus efficace: c'est lui qui, en moyenne, possède le meilleur score. L'algorithme Common Neighbors semble également assez bien adapté à ce type de réseaux, puisqu'il possède toujours des scores équivalents à ceux de Adamic / Adar bien que généralement légèrement inférieur. Enfin, l'algorithme de Jacquart semble être le moins adapté, bien que sur certains cas

particuliers il donne des résultats très supérieurs aux deux autres algorithmes (par exemple sur ce test particulier:
Swarthmore42 with n = 500, frac = 0.2
CN: 292
J: 345
AA: 298)
Il semble que de manière générale, il soit moins adapté au cas des réseaux sociaux.
Enfin, on peut également remarquer que l'efficacité des trois algorithmes augmente significativement avec la fraction de liens retirés. Il est probable que cette tendance continue jusqu'au point d'effondrement du système.

5)

En faisant tourner l'algorithme sur John Hopkins, on trouve les résultats suivant:

Accuracy of dorm attribute predictions with 10.0 % of removed labels: 0.4266409266409266
Accuracy of dorm attribute predictions with 20.0 % of removed labels: 0.44787644787644787
Accuracy of dorm attribute predictions with 30.0 % of removed labels: 0.4202059202059202
Accuracy of major_index attribute predictions with 10.0 % of removed labels: 0.15057915057915058
Accuracy of major_index attribute predictions with 20.0 % of removed labels: 0.1167953667953668
Accuracy of major_index attribute predictions with 30.0 % of removed labels: 0.12355212355212356
Accuracy of year attribute predictions with 10.0 % of removed labels: 0.3532818532818533
Accuracy of year attribute predictions with 20.0 % of removed labels: 0.3127413127413127
Accuracy of year attribute predictions with 30.0 % of removed labels: 0.32625482625482627
Accuracy of gender attribute predictions with 10.0 % of removed labels: 0.4420849420849421
Accuracy of gender attribute predictions with 20.0 % of removed labels: 0.416988416988417
Accuracy of gender attribute predictions with 30.0 % of removed labels: 0.4247104247104247

On note une très nette correspondance entre l'assortativité d'un attribut et sa capacité à être efficacement prédis. En effet, Nous avions calculé l'assortativité des paramètres dorm (très élevée) et major_index (plutôt faible) et on note ici que les résultats des prédictions de l'algorithme correspondent bien à ces résultats. Cela s'explique assez bien: si les étudiants ont en effet tendance à se regrouper par dortoir, alors en observant les dortoirs des amis d'un étudiant, on a de bonnes chances de trouver le sien. Je pense que les résultats de l'attribut élevé gender sont principalement dûs au fait que cet attribut à beaucoup moins de valeures potentielles que les autres, et qu'il est donc plus facile à prédire, ou en tout cas à donner une bonne impression de prédiction. Les résultats concernant l'attribut year sont également explicables en raisonnant de la même façon que pour l'assortativité: les élèves d'une même promotion ont plus tendances à se parler qu'entre promotion, mais comme une promotion est beaucoup plus grande qu'un dortoir, il est logique que ces liens d'amitiés soient moins fréquents. Je pense que ses résultats sont même un peu 'surestimé' de la même façon que pour l'attribut genre, puisque le nombre d'année possible est assez limité.

DEGRÉ Mathieu