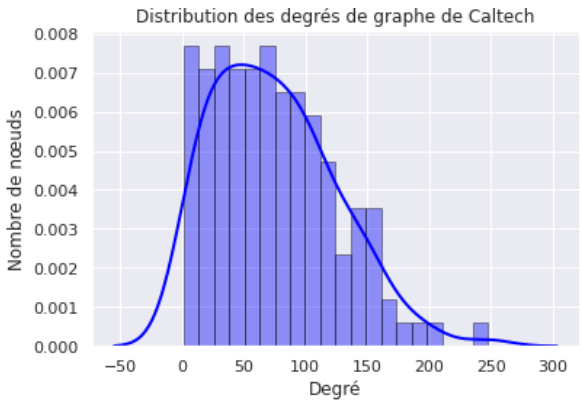
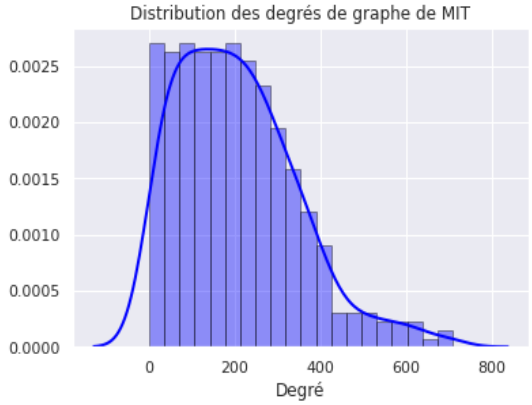
Project Facebook100

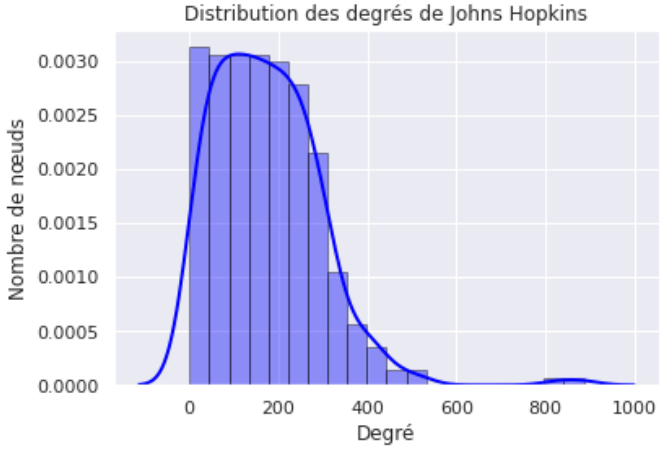
Elaboré par:- DRISS Mohamed Aziz
- KORDOGHLI Oussama

Question 2 : Social Network Analysis with the Facebook100 Dataset

a- Les résultats des trois graphes confirment que le réseau de l'université de Caltech est le plus petit en fonction du nombre de nœuds relié a la plus grande composante connexe. En effet, la plus grande valeur de degré observée est inférieure à 250 pour quelques nœuds dans le premier graphe. Dans le deuxième graphe, la plus grande valeur de degré est d'environ 500, tandis que dans le troisième graphe, elle est supérieure à 600. Cela signifie que, quel que soit le groupe, un faible nombre de nœuds jouent un rôle crucial en tant que "hubs", qui sont essentiels pour les actions de diffusion de l'information et pour connecter les composantes, tandis que les autres nœuds ont une influence plus faible.







b- le coefficient de clustering global (global clustering coefficient) mesure la probabilité que deux voisins d'un même sommet soient également voisins l'un de l'autre.
Dans notre cas, les coefficients de clustering global pour les 3 graphes Caltech36, MIT8 et Johns_Hopkins55 sont respectivement 0.29, 0.18 et 0.19 ce qui est modérément élevé pour les trois graphes. Cela signifie que les 3 graphes ont une connectivité locale relativement forte surtout pour le graphe Caltech36, mais pas nécessairement une connectivité globale forte.

Le coefficient de clustering local moyen calcule la moyenne des coefficients de clustering locaux de tous les sommets dans le graphe.
Dans notre cas, le coefficient de clustering local moyen pour les 3 graphes précédents sont respectivement 0.41, 0.27 et 0.27, ce qui est assez élevé surtout pour le graphe Caltech36. Cela indique que la plupart des sommets ont une forte connectivité locale, ce qui suggère la présence de clusters denses dans les 3 graphes (surtout pour le graphe Caltech36).

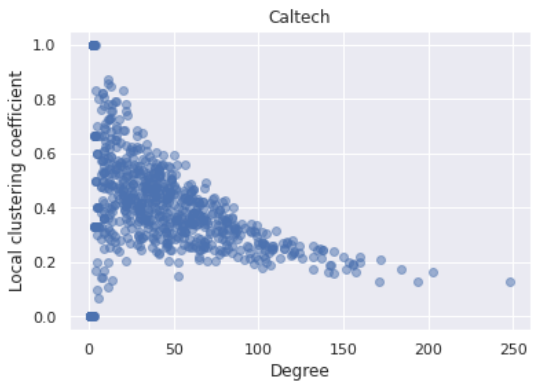
L'edge density (densité d'arêtes) mesure le nombre d'arêtes présentes dans le graphe par rapport au nombre maximum d'arêtes possibles dans le graphe.
Dans notre cas, les edges density pour les 3 graphes précédents sont respectivement 0.05, 0.01 et 0.01. Ces valeurs sont relativement faibles, ce qui suggère que les 3 graphes sont "sparse", c'est-à-dire qu'ils sont relativement peu d'arêtes par rapport au nombre maximum d'arêtes possibles.

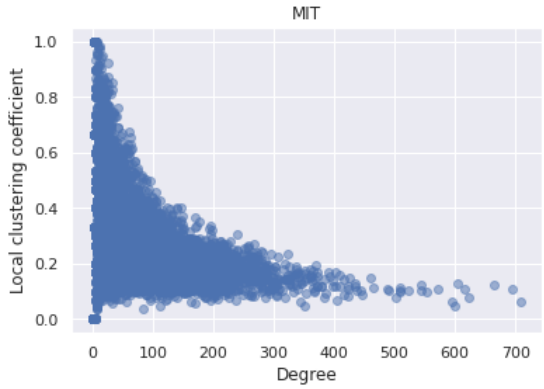
En conclusion, les 3 graphes semblent avoir une forte connectivité locale, avec une présence de clusters denses, mais une connectivité globale modérée. De plus, avec une densité d'arêtes relativement faible, le graphe peut être considéré comme "sparse".

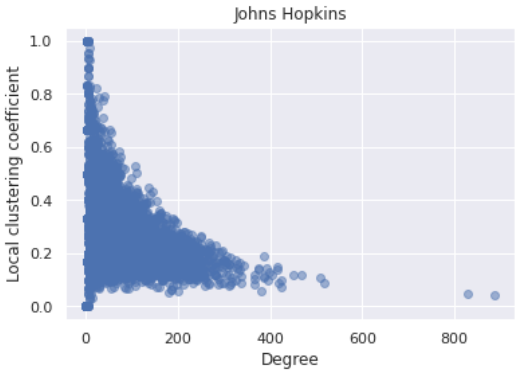
Graphe	Resultats
Caltech	Global clustering coefficient pour le graphe Caltech: 0.2912826901150874 Mean local clustering coefficient pour le graphe Caltech: 0.40929439048517247 Edge density pour le graphe Caltech: 0.05640442132639792

MIT	Global clustering coefficient pour le graphe MIT: 0.18028845093502427 Mean local clustering coefficient pour le graphe MIT: 0.2712187419501315 Edge density pour le graphe MIT: 0.012118119495041378
Johns Hopkins	Global clustering coefficient pour le graphe Johns Hopkins: 0.19316123901594015 Mean local clustering coefficient pour le graphe Johns Hopkins: 0.26839307371293525 Edge density pour le graphe Johns Hopkins: 0.013910200162372396

c- Nous constatons que plus le degré d'un nœud est élevé, plus le coefficient de clustering est faible. Les nœuds ayant un faible degré ont un coefficient de clustering beaucoup plus élevé, cela s'explique par le fait que moins un nœud a de voisins, plus la probabilité que ces voisins soient connectés est élevée. En revanche, les nœuds ayant des degrés élevés ont beaucoup plus de voisins, ce qui réduit la probabilité que les nœuds soient interconnectés. Cela est également dû au fait que les hubs, qui ont des degrés élevés, forment généralement un lien entre les communautés.







Question 3: Social Network Analysis with the Facebook100 Dataset

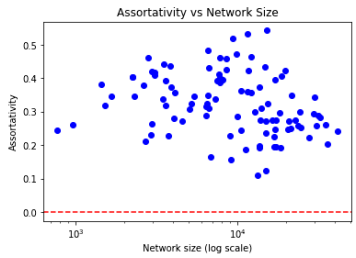
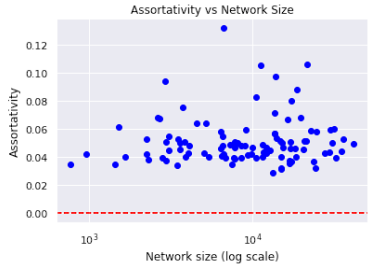
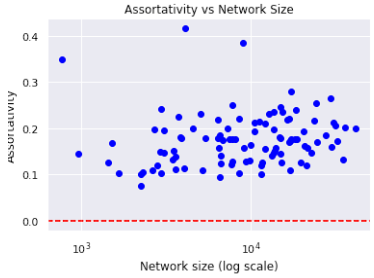
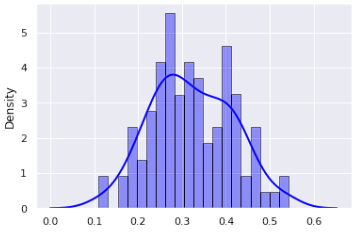
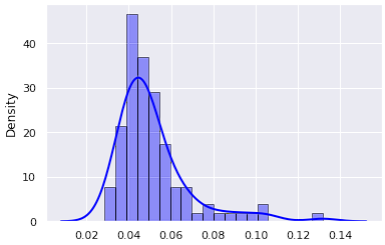
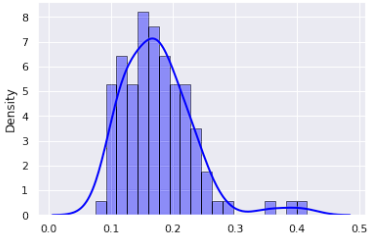
Student/faculty status	Major Index	Dorm
		
		

Tableau 1 : Résultats des algorithmes des prédicteurs de liens

Pour l'attribut statut **student/faculty**, nous pouvons observer que les scores de assortativité sont généralement élevés pour tous les graphes, selon le graphique de densité qui montre une distribution normale. Cela peut impliquer que le statut d'une personne est un facteur important pour déterminer le groupe social auquel elle appartient.

De plus, nous remarquons que pour les graphes de taille moyenne, nous avons une tendance plus élevée à avoir un score de assortativité élevé.

En revanche, pour l'indice **majeur**, nous constatons que la distribution a un pic à la valeur 0,05, ce qui est très faible par rapport au valeur moyenne de statut qui vaut environ 0,3. Cela peut être interprété comme indiquant que l'indice majeur n'est pas un facteur déterminant dans la construction du groupe social d'une personne.

Pour l'attribut **dorm**, nous constatons que la distribution a une moyenne de 0,2, ce qui est un bon indicateur que c'est un facteur important pour choisir le cercle social.
Nous concluons que le statut étudiant/professeur et le dortoir ont des effets plus forts sur la construction du cercle d'amis sur Facebook et que l'indice majeur n'a pas une grande influence sur cela.

Question 4 : Link Prediction

En lançant cette expérience sur les graphes de Caltech et de Johns Hopkins, nous obtenons les résultats suivants :

Université	CommonNeighbors	Jaccard	Adamic/Adar
Caltech	0.87	0.64	0.9
John Hopkins	0.89	0.51	0.88

Tableau 2 : Résultats des algorithmes des prédicteurs de liens

Nous remarquons que les algorithmes CommonNeighbors et Adamic/Adar sont plus performants que l'algorithme Jaccard, avec des scores très proches.
CommonNeighbors et Adamic/Adar ont produit des résultats proches de 90% alors que Jaccard produit des résultats médiocres par rapport à ces deux algorithmes.
Cela est dû à la formule utilisée dans chaque algorithme :

- $CommonNeighbors(u,v) = |N(u) \cap N(v)|$
- $Jaccard(u,v) = |N(u) \cap N(v)| / |N(u) \cup N(v)|$
- $AdamicAdar(u,v) = \sum_{w \in N(u) \cap N(v)} (1 / \log(|N(w)|))$

Pour l'exemple des CommonNeighbors , nous voyons que le score est calculé sur la base du nombre de voisins qu'ils partagent. Et pour Adamic/Adar, le score est calculé en fonction de l'impact des voisins communs qui n'ont pas beaucoup de connexions.
L'algorithme de Jaccard, quant à lui, calcule le pourcentage de voisins communs par rapport au nombre total de voisins des deux nœuds.
Pour cette étude, nous constatons que les voisins communs rares et le nombre de voisins communs ont un impact plus important sur la prédiction des liens que le pourcentage de connexions mutuelles. Cela peut s'expliquer par le fait qu'avoir des amis communs qui ne partagent pas trop de connexions est plus influent que d'avoir le même "grand" cercle d'amis pour le réseau universitaire de Facebook.

Question5 : Find missing labels with the label propagation algorithms

Clé	0.1	0.2	0.3	0.4
dorm	0.881579	0.718954	0.73913	0.716612
year	0.631579	0.718954	0.626087	0.583062
gender	0.605263	0.653595	0.595652	0.563518
major	0.171053	0.150327	0.117391	0.140065

Tableau 3 : Précision de l'algorithme de propagation des étiquettes

Lorsqu'on a appliqué l'algorithme de propagation de labelle en enlevant que ce soit 0.1, 0.2, 0.3 ou 0.4 des noeuds on a remarqué que l'accuracy de modèle est élevé pour certains attributs comme dorm, year et gender et elle est faible pour certains d'autre comme major et ça pour toutes les portions des noeuds qui n'ont pas un effet énorme sur l'accuracy. Cela peut dû aux:

- Nature des attributs: Certains attributs peuvent être plus facilement prévisibles par LPA, car ils peuvent avoir une corrélation plus forte avec les attributs de leurs voisins 'le cas de dorm, year et gender). D'autres attributs peuvent être plus difficiles à prédire car ils peuvent être moins liés aux attributs de leurs voisins (le cas de major).

- La structure du graphe: La structure du graphe peut avoir un impact sur la performance de LPA. Si le graphe est très dispersé, LPA peut avoir des difficultés à propager les étiquettes à travers le graphe, ce qui peut affecter la performance de l'algorithme.