NET 4103/7431 Homework
Network science and Graph Learning

Nathan Piatte

31 janvier 2025

1 Introduction

Les réseaux sociaux font aujourd'hui partie de la vie quotidienne de milliards de personnes à travers le monde. Pourtant, leur fonctionnement requière une compréhension assez fine de la structure des réseaux qui le composent : comment les individus interagissent entre eux, comment l'information se propage, comment les communautés se forment. Pour étudier cela, le cadre théorique de l'apprentissage sur les graphes est particulièrement adapté.

Pour illustrer cela, nous étudions dans ce travail les données de *Facebook100*, un jeu de données qui contient des réseaux sociaux des 100 premières universités américaines arrivées sur la plateforme. Le but de cette étude est d'identifier les caractéristiques communes à différents graphes, mais également de mettre en évidence les différences entre eux. Nous nous intéressons en particulier à la structure des réseaux, à l'assortativité des attributs, à la prédiction de liens, à la propagation d'étiquettes, et à la détection de communautés.

2 Description du réseau de données

Question 2

(a) On présente la distribution des degrés du réseau de données sur la figure 1. On observe que la distribution des degrés est très hétérogène, avec une queue de distribution très longue. La plupart des nœuds ont un degré faible, mais quelques nœuds ont un degré très élevé. Dans le contexte de Facebook, cela signifie que la plupart des utilisateurs ont un nombre d'amis faible, mais quelques utilisateurs ont un nombre d'amis très élevé. Pour Caltech par exemple, qui ne comporte que 769 nœuds, certains individus ont plus de 200 amis. Les trois distributions sont très similaires malgré des échelles différentes (moins de 1000 nœuds pour Caltech, plus de 5000 pour le MIT et Johns Hopkins). Nous pouvons en conclure que certains individus sont très populaires dans les trois réseaux. Comme les données sont une image figée de l'état du réseau social à un moment donné, assez tôt dans son histoire, il est difficile de savoir si ces individus sont simplement les premiers arrivés sur le réseau, ou bien si leur degré est lié à leur popularité réelle. Quoi qu'il en soit, ces individus sont des points de passage obligés pour la propagation de l'information dans le réseau. De manière plus générale, les degrés moyens sont tous relativement bas par rapport au nombre de nœuds, ce qui signifie que les réseaux sont peu denses.

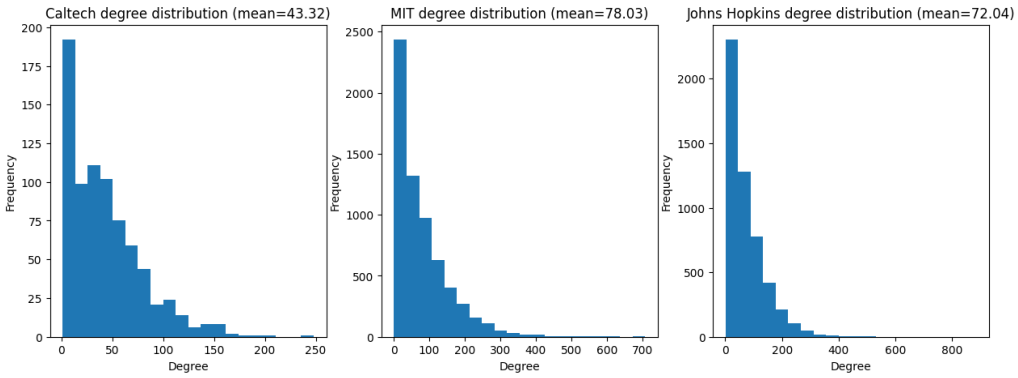

FIGURE 1 – Distribution des degrés du réseau de données

(b) On donne le coefficient de clustering global, le coefficient de clustering local moyen et la densité d'arêtes du réseau dans le tableau 1.

On voit que Caltech a ces 3 indicateurs beaucoup plus élevés que les deux autres réseaux. Cependant, pour les 3 réseaux, la densité d'arêtes est faible, on peut donc considérer que ce sont des graphes creux. Le coefficient de clustering global est plus faible que le coefficient de clustering local moyen pour les 3 réseaux, ce qui peut

TABLE 1 – Coefficient de clustering global, coefficient de clustering local moyen et densité d'arêtes pour les trois réseaux

	Coeff. de clustering global	Coeff. de clustering local moyen	Densité d'arêtes (%)
Caltech	0.291	0.409	5.64
MIT	0.180	0.271	1.21
Johns Hopkins	0.193	0.268	1.39

signifier que les nœuds sont regroupés en communautés, et que ces communautés sont peu connectées entre elles.

(c) On trace le degré en fonction du coefficient local de clustering pour les trois réseaux sur la figure 2.

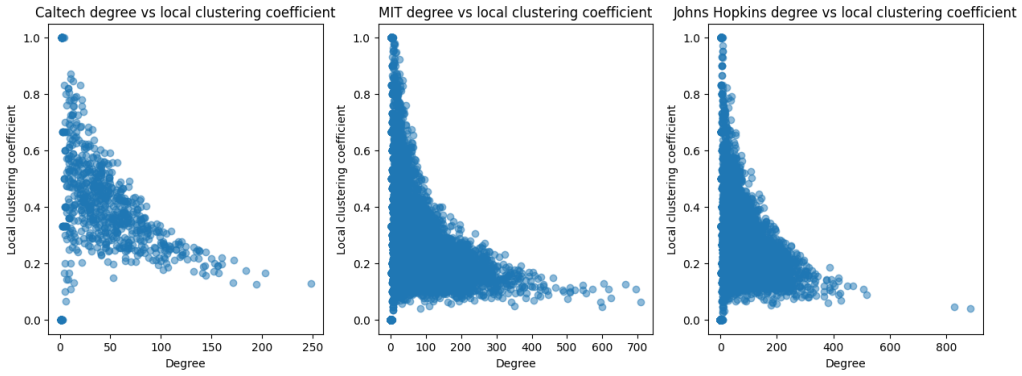

FIGURE 2 – Degré en fonction du coefficient de clustering local pour les trois réseaux

On distingue 2 tendances assez différentes sur ces réseaux, même s'ils suivent tous les trois la règle selon laquelle le coefficient local de clustering est inversement proportionnel au degré. D'un côté, les étdiants de Caltech ont un coefficient de clustering local très élevé, ce qui explique la différence dans le tableau 1 avec les autres écoles. D'un autre côté, pour les étudiants du MIT et de Johns Hopkins, bien que la limite supérieure suive la même tendance que pour Caltech, une grande partie des points se situe en dessous de cette limite. Ainsi, ces deux écoles ont de nombreux nœuds avec un degré faible et un coefficient de clustering local faible. Cela pourrait signifier que les étudiants de ces écoles ont tendance à se regrouper en communautés plus grandes, moins connectées entre elles.

Enfin, cette figure illustre qu'il existe quelques nœuds particuliers avec des valeurs extrêmes. On distingue en particulier les nœuds qui ont un degré très élevé et un coefficient de clustering local très faible, et ceux qui ont un degré très faible et un coefficient de clustering local très élevé. Ces nœuds sont des points de passage obligés pour la propagation de l'information dans le réseau.

Question 3

On trace les assortativités pour différents attributs sur les figures 3 à 7.

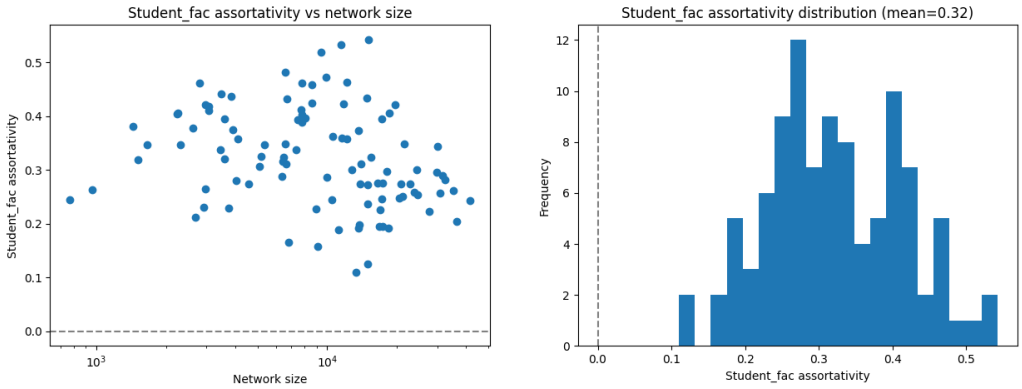

FIGURE 3 – Assortativité pour le statut des individus

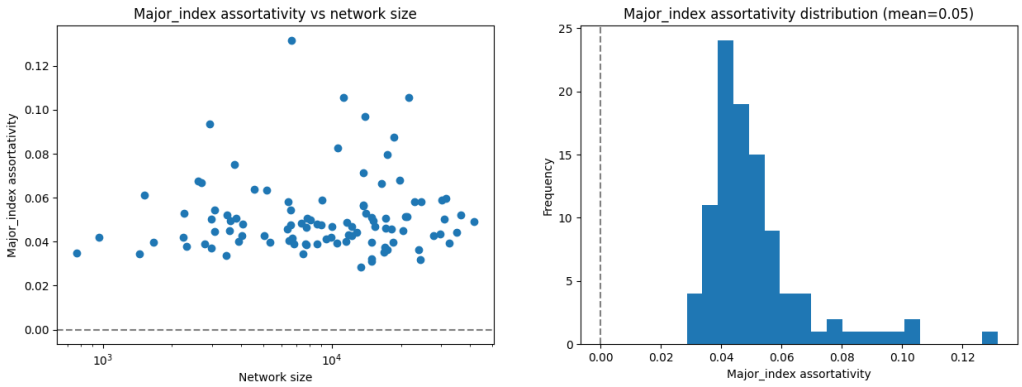

FIGURE 4 – Assortativité pour la majeure des individus

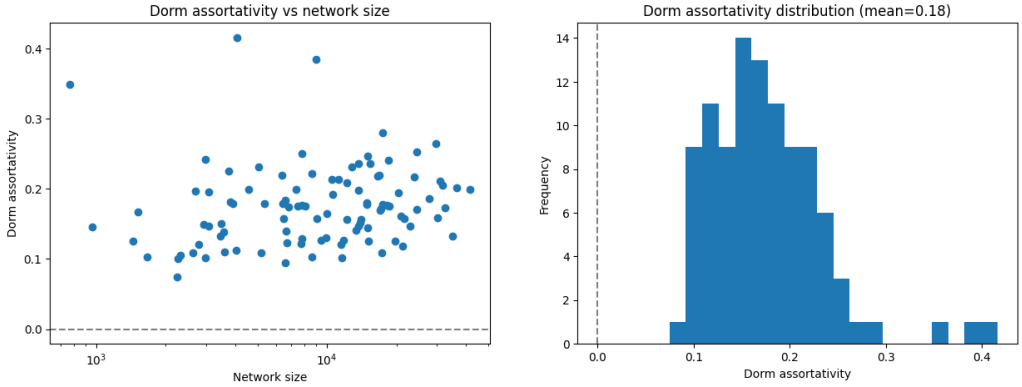

FIGURE 5 – Assortativité pour le dortoir des individus

On voit que toutes ces assortativités ont une tendance positive. Dans le contexte de Facebook, cela indique que les amis se font généralement plutôt entre personnes qui se ressemblent du point de vue des attributs considérés.

Certaines assortativités sont systématiquement positives : c'est le cas pour le statut, la majeure et le dortoir des individus. Cela montre une tendance forte à ce que deux amis partagent les mêmes valeurs pour ces attributs.

L'assortativité est particulièrement élevée pour le statut, avec des valeurs allant de 0.1 à 0.55, et une moyenne autour de 0.32. De manière légèrement plus faible que le statut mais objectivement assez forte, l'assortativité en fonction du dortoir est également élevée, avec une moyenne à 0.18. Cela montre que les communautés au sein des réseaux sont très fortement liées au statut des individus qui composent chaque communauté, et également au dortoir dans lequel ils se trouvent.

Enfin, bien que l'assortativité en fonction du degré ne soit pas systématiquement positive, elle est en moyenne de 0.06, ce qui montre une tendance à ce que les individus avec un degré élevé aient tendance à se lier entre eux. Dans le cadre de Facebook, cela signifie que les individus populaires ont tendance à se lier entre eux, même si cette tendance est faible.

Question 4

On évalue les différentes métriques de prédiction de liens sur deux graphes choisis arbitrairement : Caltech, et le MIT. Ces deux graphes sont choisis car nous avons vu plus tôt qu'ils avaient des cractéristiques assez différentes, et nous voulons voir si cela a un impact sur les résultats des différentes métriques. Les résultats sont présentés dans les tableaux 2 et 3, où p est la proportion d'arêtes à retirer dans le graphe original. En lisant ces tableaux, il faut prendre garde au fait que nous avons décidé de décrire le rappel en pour mille, car les valeurs de rappel sont très faibles à cause de nombreux faux négatifs. Pour chaque combinaison de (métrique, critère de performance), nous avons mis en gras la meilleure valeur obtenue, ce qui nous donne le couple (p, k) le plus pertinent.

On remarque que le rappel est systématiquement beaucoup plus élevé quand k est grand. Cela s'explique par le fait que le nombre de faux négatifs est très grands. Ainsi, même quand la proportion p d'arêtes retirée est grande, le nombre de faux négatifs reste grand, mais c'est le nombre de vrais positifs qui augmente largement avec k.

Pour ce qui est de la précision, on observe que les meilleures valeurs sont observées plutôt pour des p grands.

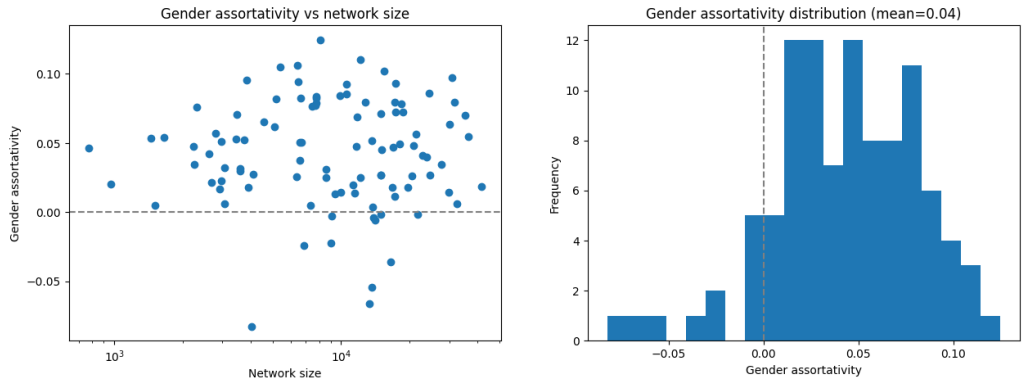

FIGURE 6 – Assortativité pour le genre des individus

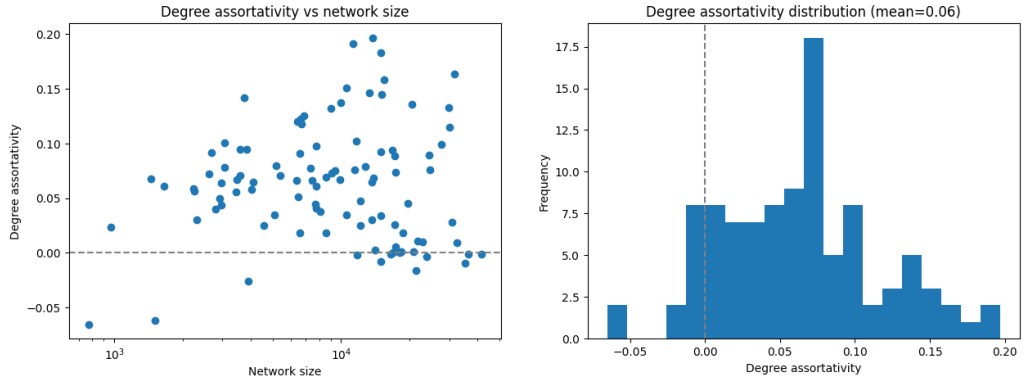

FIGURE 7 – Assortativité en fonction du degré des individus

Cela s'explique par le fait que la précision est le rapport entre le nombre de vrais positifs et le nombre de vrais positifs et de faux positifs. Ainsi, quand p est grand, le nombre de faux positifs est plus faible, ce qui augmente la précision. On pouvait donc s'attendre à ce résultat. Une valeur fait exception à cette règle : pour le MIT avec la métrique du coefficient de Jaccard, la précision est maximale pour $p = 0.05$ et $k = 1600$.

Question 5

(c) On choisit le graphe de Caltech pour illustrer les performances de la propagation d'étiquettes, car il est relativement petit et permettra des calculs plus rapides.

(d) On donne les valeurs de précision, l'erreur absolue moyenne et le *F1-Score* pour les fractions $f \in [0.1, 0.2, 0.3, 0.4]$ de nœuds auxquels on a retiré successivement les étiquettes `major`, `dorm` et `gender` dans le tableau 4.

On observe tout d'abord que les étiquettes `gender` et `dorm` ont presque exactement les mêmes valeurs. Il est difficile de donner une explication certaine pour ce résultat, mais il est possible que le fait que les dortoirs ne soient généralement pas mixtes ait un impact sur la propagation des étiquettes. En effet, on crée ainsi de la redondance dans les informations, ce qui pourrait expliquer que les deux attributs aient des performances similaires.

Ensuite, on observe un résultat étonnant lorsque l'on retire l'étiquette `major_index` : la précision et le *F1-Score* sont meilleurs lorsque l'on retire une grande partie des étiquettes. Cela peut s'expliquer par le fait que les individus de Caltech ont tendance à se regrouper en fonction de leur majeure, et que la propagation d'étiquettes est plus efficace lorsque l'on retire une grande partie des étiquettes. En effet, cela permet de créer des communautés plus homogènes, ce qui facilite la propagation des étiquettes.

Enfin, une dernière remarque peut être faite : quand les performances en retirant l'étiquette `major_index` sont les meilleures, les performances en retirant les étiquettes `gender` et `dorm` sont les plus faibles, et vice-versa. Cela peut s'expliquer par le fait que ces attributs sont corrélés positivement, et que retirer une partie des étiquettes d'un attribut permet de mieux prédire les étiquettes de l'autre attribut.

TABLE 2 – Performances des différentes métriques de prédiction de liens pour Caltech

		Précision (%)				Rappel (‰)			
	$k =$ / $p =$	50	100	400	1600	50	100	400	1600
Voisins communs	0.05	72.0	73.0	75.5	75.6	2.16	4.38	18.1	72.6
	0.10	72.0	78.0	78.0	76.2	2.16	4.68	19.0	**73.2**
	0.15	72.0	76.0	77.0	76.1	2.16	4.56	18.5	73.1
	0.20	70.0	74.0	**78.3**	74.1	2.10	4.44	18.8	71.1
Coeff. de Jaccard	0.05	74.0	69.0	73.8	70.8	2.22	4.14	1.75	68.0
	0.10	46.0	62.0	72.8	71.9	1.38	3.72	17.5	**69.0**
	0.15	64.0	73.0	**75.0**	70.8	1.92	4.38	18.0	68.0
	0.20	46.0	56.0	67.5	69.4	1.38	3.36	16.2	66.6
Indice Adamic-Adar	0.05	66.0	73.0	76.8	77.1	1.98	4.38	18.4	**74.0**
	0.10	66.0	73.0	77.8	76.9	1.98	4.38	18.7	73.9
	0.15	64.0	70.0	76.8	76.4	1.92	4.20	18.4	73.4
	0.20	70.0	75.0	**79.0**	76.9	2.10	4.50	1.90	73.8

TABLE 3 – Performances des différentes métriques de prédiction de liens pour le MIT

		Précision (%)				Rappel (‰)			
	$k =$ / $p =$	50	100	400	1600	50	100	400	1600
Voisins communs	0.05	92.0	92.0	85.3	82.7	0.18	0.37	1.36	5.27
	0.10	90.0	91.0	85.5	83.3	0.18	0.36	1.36	5.30
	0.15	92.0	92.0	85.3	82.6	0.18	0.37	1.36	5.26
	0.20	92.0	**93.0**	85.3	83.4	0.18	0.37	1.36	**5.31**
Coeff. de Jaccard	0.05	38.0	60.0	63.5	**69.6**	0.076	0.24	1.01	**4.43**
	0.10	30.0	56.0	58.8	69.5	0.060	0.22	0.94	**4.43**
	0.15	36.0	52.0	57.5	68.5	0.072	0.21	0.92	4.36
	0.20	24.0	52.0	55.8	64.6	0.048	0.51	0.89	4.12
Indice Adamic-Adar	0.05	90.0	92.0	84.0	82.9	0.18	0.37	1.33	5.28
	0.10	94.0	92.0	85.0	82.4	0.19	0.37	1.35	5.25
	0.15	92.0	92.0	84.0	82.6	0.18	0.37	1.34	5.26
	0.20	**96.0**	91.0	85.3	83.4	0.19	0.36	1.36	**5.31**

TABLE 4 – Performances de la propagation d'étiquettes pour Caltech

	Proportion d'étiquettes retirées (%)	Attribut retiré		
		`dorm`	`gender`	`major_index`
Précision (%)	10	**90.0**	**90.0**	90.0
	20	80.2	80.0	80.0
	30	70.0	70.7	93.4
	40	60.2	60.5	**96.9**
Erreur absolue moyenne (%)	10	10.0	10.0	10.0
	20	19.8	20.0	**20.0**
	30	30.0	29.3	6.63
	40	**39.8**	**39.5**	3.12
F1-Score (%)	10	**90.0**	**90.0**	90.0
	20	80.0	80.0	80.0
	30	70.0	70.7	93.4
	40	60.2	60.5	**96.9**

Question 6

(a) Nous nous intéresserons à répondre à la question suivante :

Est-ce que les majeures et la localisation du dortoir jouent un rôle conjoint dans la création de communautés au sein des des campus universitaires ?

Nous avons vu à travers les assortativités que ces deux attributs étaient corrélés positivement avec la création de liens entre individus. Cependant, l'assortativité de chacun de ces attributs ne nous permet pas de conclure sur leur caractère conjoint ou séparé sur l'influence de la création de communautés. Ainsi, nous proposons l'hypothèse que :

Les communautés au sein des campus universitaires sont formées principalement par des individus qui partagent la même majeure ou le même dortoir, mais pas forcément les deux à la fois.

En effet, la majeure et le dortoir permettent des proximités directes favorisant la création de liens. Mais il nous semble tout à fait probable qu'au sein de chaque dortoir, il y ait des individus de différentes majeures, et que dans chaque majeure, il y ait des individus de différents dortoirs. Cela permettrait ainsi la création de communautés sans que ces deux attributs soient systématiquement conjoints.

(c) Après avoir détecté les communautés grâce à l'algorithme de Louvain, nous avons calculé les proportions au sein de chaque communauté des attributs `major` et `dorm`. Si notre hypothèse est correcte, nous devrions observer que les communautés sont principalement formées par des individus qui partagent la même majeure ou le même dortoir, mais les proportions de ces attributs ne devraient pas être systématiquement conjointes. On peut détecter cela en calculant la corrélation entre les attributs `major` et `dorm` pour chaque communauté. Si la corrélation est faible, cela signifie que les communautés sont formées par des individus qui partagent soit la même majeure, soit le même dortoir, mais pas forcément les deux à la fois.

Tout d'abord, la première partie de l'hypothèse, à savoir le fait que les communautés sont formées principalement par des individus qui partagent la même majeure ou le même dortoir, est confirmée par les résultats. En effet, au sein d'une même communauté, la majeure est la même pour 16.3% des individus, et le dortoir est le même environ 22.7% des individus. Cela montre que les communautés sont formées par des individus qui partagent la même majeure ou le même dortoir.

Cependant, lorsque l'on s'intéresse à la corrélation entre les attributs `major` et `dorm`, on observe que la corrélation est faible, avec une valeur moyenne de 0.081. Cela confirme la seconde partie de notre hypothèse : les communautés sont formées par des individus qui partagent soit la même majeure, soit le même dortoir, mais pas forcément les deux à la fois.

3 Conclusion

Dans ce travail, nous avons étudié les réseaux sociaux de plusieurs universités américaines, pour nous concentrer particulièrement sur 3 aspects : la description des données (attributs, liens), la prédiction des liens, la propagation d'étiquettes, et la détection de communautés. Nous avons pu observer que les réseaux sociaux des universités étudiées sont très hétérogènes, avec des degrés très variables, des coefficients de clustering faibles, et des assortativités positives pour les attributs étudiés. Nous avons également pu observer que les performances de différentes métriques de prédiction de liens varient en fonction de la proportion d'arêtes retirées, et que la propagation d'étiquettes peut s'avérer plus ou moins efficace en fonction de l'étiquette retirée et de la proportion retirée. Enfin, nous avons proposé une hypothèse sur la formation des communautés au sein des campus universitaires, et nous avons pu la confirmer en calculant la corrélation entre les attributs `major` et `dorm` pour chaque communauté.