

UNIVERSITÉ PARIS-SACLAY

Master 2 TRIED

Analyse des Réseaux Sociaux Universitaires

Étude comparative sur le dataset Facebook100

Mahouna Vayssières

Sous la direction de

Vincent Gauthier

Année universitaire 2025–2026

Table des matières

1	Introduction	3
2	Caractérisation topologique des réseaux	3
2.1	Effet d'échelle sur la structure du graphe	3
2.2	Distribution des degrés et structure hiérarchique	4
3	Homophilie et ségrégation sociale	5
4	Prédiction de liens	7
4.1	Cadre expérimental	7
4.2	Influence de la topologie sur les performances	7
4.3	Étude de sensibilité : impact de la fraction retirée	9
5	Inférence d'attributs par propagation de labels	10
6	Détection de communautés	11
7	Conclusion	12

Dépôt GitHub : <https://github.com/mikeydev/graph-learning-facebook100>

1. Introduction

Les réseaux sociaux numériques constituent un terrain d'observation privilégié pour comprendre les dynamiques relationnelles au sein de populations structurées. Le dataset Facebook100, collecté en 2005, offre une photographie unique des graphes d'amitié de 100 campus universitaires américains avant l'ouverture de la plateforme au grand public. Chaque réseau encode non seulement les connexions entre utilisateurs, mais également des attributs socio-démographiques tels que le dortoir de résidence, l'année de promotion, la filière d'études ou encore le genre.

L'objectif de cette étude est d'explorer les mécanismes de formation des liens sociaux au sein de ces communautés universitaires. Notre méthodologie combine deux niveaux d'analyse complémentaires. D'une part, nous conduisons une analyse statistique systématique sur l'intégralité des 100 réseaux pour caractériser les propriétés topologiques (densité, clustering) et étudier les phénomènes d'homophilie et d'assortativité. D'autre part, nous restreignons l'analyse des distributions et l'évaluation algorithmique (prédiction de liens, détection de communautés) à un échantillon de 15 réseaux représentatifs, couvrant un spectre large de tailles et de densités. Enfin, nous approfondissons certains cas contrastés pour mettre en évidence l'influence de la topologie locale sur les performances des algorithmes.

Les 15 campus sélectionnés couvrent l'ensemble du spectre des universités américaines : des petits instituts d'élite comme Caltech36 (762 nœuds) ou Reed98 (962 nœuds) aux grandes universités publiques comme Texas84 (36 364 nœuds) ou UGA50 (24 380 nœuds), en passant par des campus de taille intermédiaire comme MIT8 (6 402 nœuds) ou Duke14 (9 885 nœuds). Cette diversité garantit la robustesse de nos conclusions.

2. Caractérisation topologique des réseaux

2.1. Effet d'échelle sur la structure du graphe

L'analyse comparée de réseaux de tailles différentes met en évidence un phénomène d'échelle. Le Tableau 1 synthétise les métriques topologiques fondamentales pour trois campus représentatifs.

TABLE 1 – Métriques topologiques des réseaux analysés

Réseau	Nœuds	Arêtes	Densité	Clustering Global	Clustering Local
Caltech36	762	16 651	5.74%	0.291	0.409
Johns Hopkins55	5 157	186 572	1.40%	0.193	0.269
MIT8	6 402	251 230	1.23%	0.180	0.272

On observe que la densité s'effondre drastiquement avec la taille du réseau, passant de 5.7% pour Caltech à seulement 1.2% pour MIT. Ce résultat s'interprète aisément : le nombre de relations qu'un individu peut maintenir est cognitivement limité (nombre de Dunbar), tandis que le

nombre de paires possibles croît quadratiquement avec la population. En revanche, le coefficient de clustering local demeure uniformément élevé, oscillant entre 0.27 et 0.41 indépendamment de la taille du campus.

Cette combinaison caractéristique (faible densité globale associée à un fort clustering local) constitue la signature des réseaux dits *Small World*. Les étudiants forment des cliques locales très soudées (groupes d'amis, clubs, équipes sportives) qui sont interconnectées par quelques liens longs, assurant ainsi une faible distance moyenne entre n'importe quelle paire d'individus.

2.2. Distribution des degrés et structure hiérarchique

L'examen de la distribution des degrés révèle une structure hiérarchique prononcée au sein des réseaux étudiés. La Figure 1 présente ces distributions en échelle logarithmique.

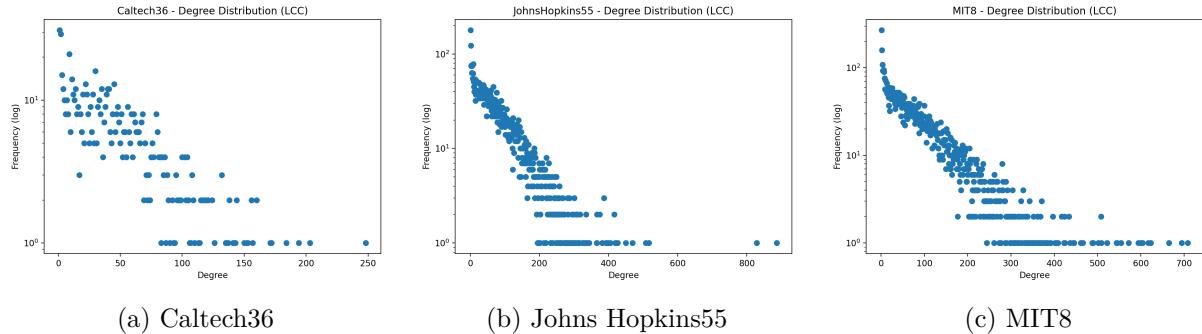


FIGURE 1 – Distribution des degrés en échelle log-log

La décroissance linéaire observée en échelle log-log est caractéristique des lois de puissance, signature des réseaux *Scale-Free*. Cette propriété indique l'existence de *hubs*, c'est-à-dire d'individus exceptionnellement connectés qui jouent un rôle structurant dans le réseau. Ces étudiants populaires constituent des points de passage obligés pour la circulation de l'information sociale.

La Figure 2 illustre la relation entre le degré d'un noeud et son coefficient de clustering local. On y observe une corrélation négative systématique : plus un individu possède de connexions, moins ses contacts sont interconnectés entre eux. Ce phénomène s'explique par le fait que les *hubs* agrègent des contacts issus de cercles sociaux distincts qui n'ont pas vocation à interagir.

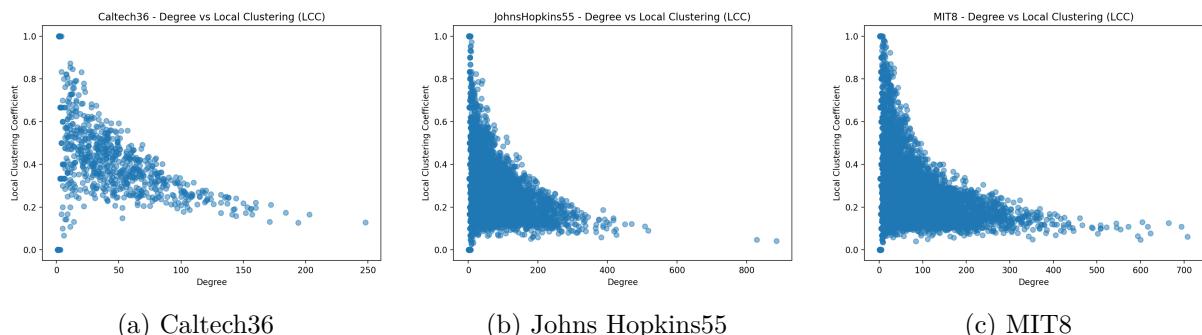


FIGURE 2 – Coefficient de clustering local en fonction du degré

3. Homophilie et ségrégation sociale

L'homophilie désigne la tendance des individus à nouer des liens avec des personnes qui leur ressemblent selon certains attributs. Pour quantifier ce phénomène, nous utilisons le coefficient d'assortativité de Newman, qui varie de -1 (hétérophilie parfaite) à $+1$ (homophilie parfaite), avec une valeur nulle indiquant une absence de corrélation.

L'analyse systématique sur l'ensemble des 100 réseaux du dataset révèle une hiérarchie sociale claire dans les mécanismes de formation des liens. La Figure 3 présente la distribution des coefficients d'assortativité pour deux attributs clés.

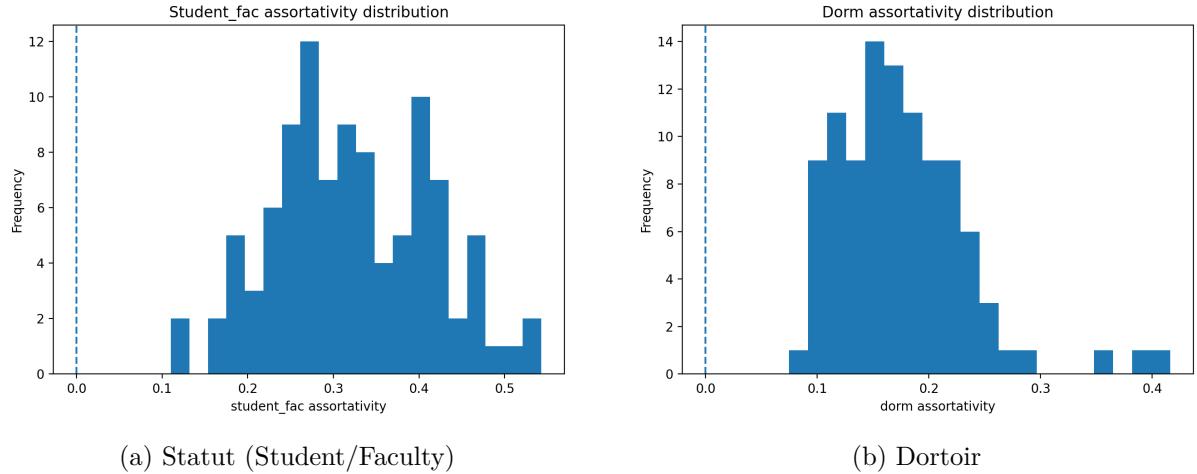


FIGURE 3 – Distribution de l'assortativité pour les attributs structurants

Le statut institutionnel (étudiant versus personnel académique) génère la ségrégation la plus marquée, avec une médiane d'assortativité de 0.32. Cette barrière hiérarchique apparaît comme structurelle : les étudiants tissent leurs liens quasi-exclusivement entre eux, tout comme le corps enseignant. Le dortoir de résidence constitue le second facteur d'homophilie, avec une médiane de 0.17, confirmant l'importance de la proximité géographique immédiate dans la création des amitiés.

Une analyse complémentaire sur la filière d'études (*Major*) et le degré (*Degree*) permet de compléter cette hiérarchie sociale. La Figure 4 illustre ces distributions.

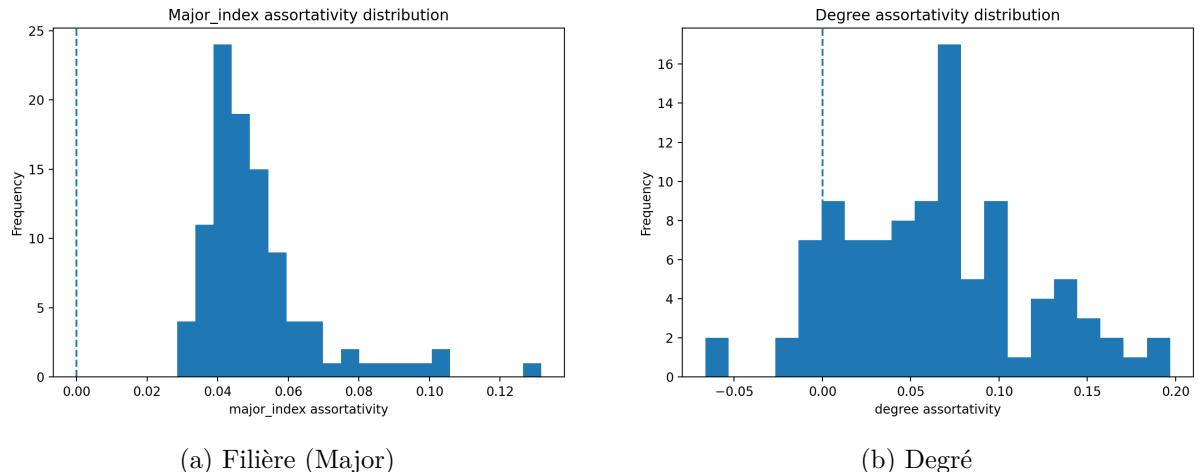


FIGURE 4 – Distribution de l’assortativité pour les attributs secondaires

Contrairement au dortoir, la filière d’études joue un rôle étonnamment faible dans la formation des liens, avec une assortativité médiane autour de 0.05. Cela suggère que sur les campus américains de 2005, la vie sociale s’organise davantage autour du lieu de vie que du lieu d’études. Quelques exceptions existent néanmoins, comme Carnegie49 (un institut technologique) où l’assortativité par filière atteint 0.13, indiquant une culture plus centrée sur les spécialités académiques.

L’assortativité par degré est quant à elle quasiment nulle (moyenne proche de 0.06). Ce résultat est important : il indique l’absence de phénomène de “club des riches” (*Rich Club*). Les étudiants très populaires ne se connectent pas exclusivement entre eux, mais tissent des liens à travers toute la hiérarchie de popularité, jouant ainsi leur rôle de ponts sociaux.

En revanche, le genre ne constitue pas un facteur structurant des réseaux d’amitié universitaires. Comme l’illustre la Figure 5, la distribution de l’assortativité par genre est concentrée dans l’intervalle [0, 0.10], avec une moyenne légèrement positive (environ 0.02). Cette valeur, bien que techniquement non nulle, reste négligeable comparée aux coefficients observés pour le statut (0.32) ou le dortoir (0.17). On peut donc considérer que le genre n’exerce qu’une influence marginale sur la formation des amis.

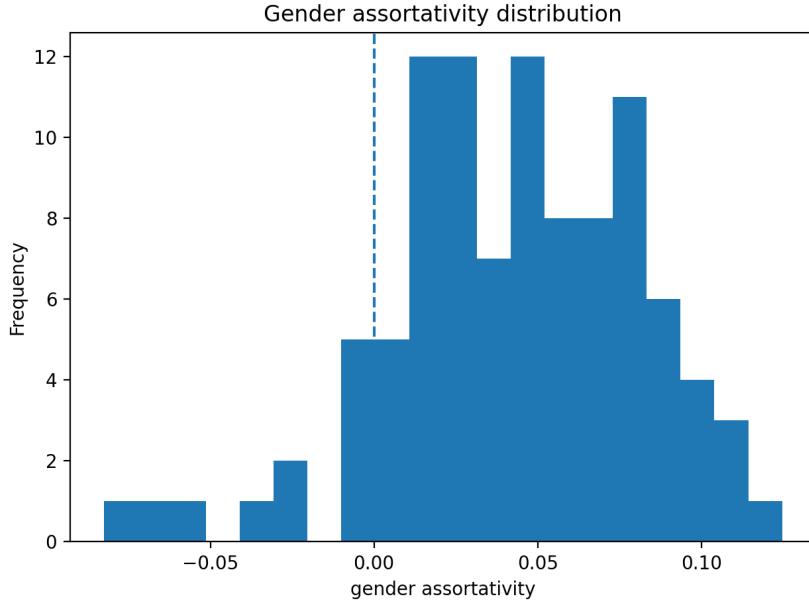


FIGURE 5 – Distribution de l'assortativité par genre (valeurs faiblement positives, effet négligeable)

4. Prédiction de liens

4.1. Cadre expérimental

La prédiction de liens vise à identifier les connexions manquantes ou futures dans un graphe à partir de sa structure locale. Nous évaluons trois métriques de similarité classiques. Common Neighbors (CN) compte simplement le nombre de voisins partagés entre deux nœuds : $|N(u) \cap N(v)|$. Jaccard normalise ce score par la taille de l'union des voisinages : $\frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}$. Enfin, Adamic-Adar (AA) pondère chaque voisin commun par l'inverse du logarithme de son degré, accordant plus d'importance aux connexions rares :

$$AA(u, v) = \sum_{w \in N(u) \cap N(v)} \frac{1}{\log |N(w)|} \quad (1)$$

Le protocole expérimental consiste à supprimer aléatoirement 10% des arêtes du graphe, puis à évaluer la capacité de chaque métrique à les retrouver parmi l'ensemble des paires candidates (nœuds situés à distance 2). Pour les grands graphes comme UNC28, nous avons implémenté un échantillonnage limitant l'évaluation à 100 000 paires candidates afin de maintenir un temps de calcul raisonnable.

4.2. Influence de la topologie sur les performances

Les résultats obtenus sur l'ensemble des 15 réseaux révèlent une dépendance marquée entre la densité du réseau et la métrique optimale. Le Tableau 2 présente les performances moyennes sur l'ensemble du dataset, tandis que le Tableau 3 détaille les résultats pour des réseaux représentatifs de différents régimes topologiques.

TABLE 2 – Precision@50 moyenne sur les 15 réseaux analysés

Métrique	Precision@50 moyenne
Common Neighbors	39.3%
Jaccard	35.8%
Adamic-Adar	42.5%

À l'échelle globale, Adamic-Adar domine légèrement avec 42.5% de précision moyenne, suivi de Common Neighbors (39.3%) et Jaccard (35.8%). Cependant, cette moyenne masque une réalité plus nuancée qui se révèle dans l'analyse par type de réseau.

TABLE 3 – Precision@50 selon le réseau et la métrique : analyse comparative ($f=0.1$)

Réseau	Nœuds	CN	Jaccard	AA
Caltech36 (dense)	762	66%	40%	74%
Rice31 (dense)	4 083	52%	36%	75%
Swarthmore42	1 657	42%	76%	42%
MIT8	6 402	48%	20%	42%
UNC28 (dispersé)	18 158	18%	26%	36%
Texas84 (très grand)	36 364	34%	34%	28%

Sur Caltech36, réseau dense et communautaire, Adamic-Adar domine avec 74% de précision, capitalisant sur la structure en cliques interconnectées. Common Neighbors suit de près (66%), tandis que Jaccard décroche (40%), pénalisé par sa normalisation excessive dans les zones très denses. Cette tendance se confirme sur Rice31, où AA surclasse ses concurrents (75% contre 52% pour CN).

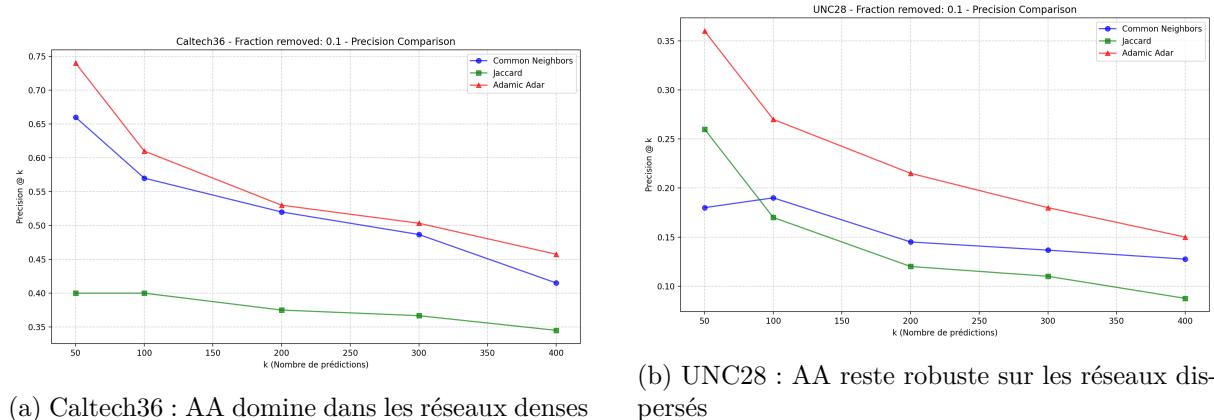


FIGURE 6 – Comparaison des performances de précision

Sur UNC28, un réseau bien plus grand et dispersé, la hiérarchie évolue. Adamic-Adar maintient sa supériorité (36%), prouvant la robustesse de sa pondération logarithmique face aux variations de degré. En revanche, Jaccard (26%) surpassé ici Common Neighbors (18%), car il parvient mieux à filtrer les faux positifs générés par les “super-hubs” omniprésents dans les grands graphes.

Le cas de Swarthmore42 (1 657 nœuds) constitue une anomalie remarquable : Jaccard y écrase la concurrence avec 76% de précision (contre 42% pour AA/CN). De même, sur le géant Texas84 (36 000+ nœuds), Jaccard et CN font jeu égal (34%) devant AA (28%). Ces exceptions confirment que la densité globale n'est pas le seul prédicteur de performance ; la microstructure locale joue un rôle décisif.

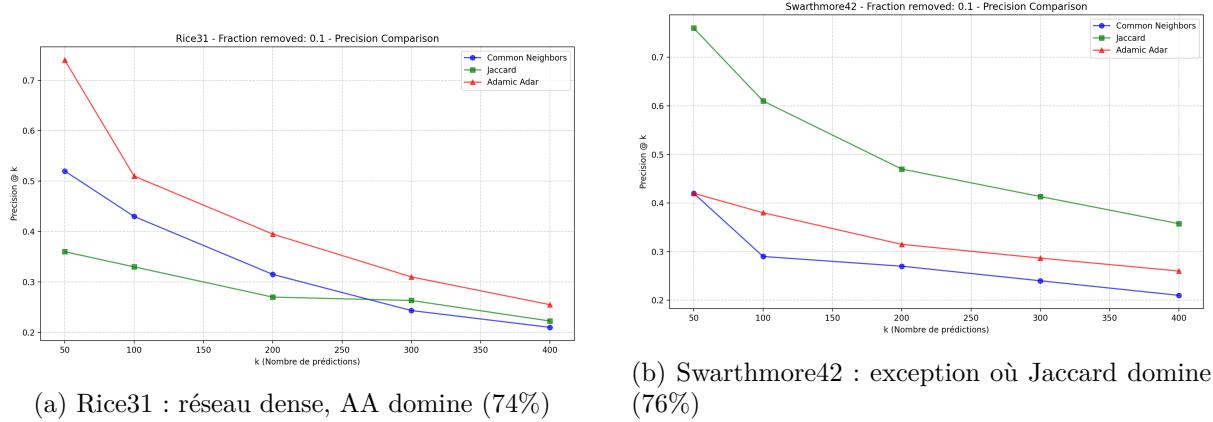


FIGURE 7 – Variabilité des performances selon la structure locale du réseau

Cette analyse démontre qu'il n'existe pas de métrique universellement optimale pour la prédiction de liens. Le choix de l'algorithme doit être guidé par les caractéristiques topologiques du réseau étudié : Adamic-Adar et Common Neighbors pour les communautés denses à fort clustering, Jaccard pour les graphes dispersés à grande échelle où les hubs introduisent des biais systématiques.

4.3. Étude de sensibilité : impact de la fraction retirée

Afin de valider la robustesse de nos conclusions, nous avons conduit une analyse de sensibilité sur le réseau MIT8 (6 402 nœuds) en faisant varier la fraction de liens retirés $f \in \{0.05, 0.10, 0.15, 0.20\}$. Le Tableau 4 synthétise les résultats.

TABLE 4 – Precision@50 sur MIT8 selon la fraction de liens retirés

Métrique	$f = 0.05$	$f = 0.10$	$f = 0.15$	$f = 0.20$
Common Neighbors	28%	48%	46%	70%
Jaccard	26%	20%	50%	50%
Adamic-Adar	24%	42%	70%	72%

TABLE 5 – Recall@50 sur MIT8 selon la fraction de liens retirés

Métrique	$f = 0.05$	$f = 0.10$	$f = 0.15$	$f = 0.20$
Common Neighbors	0.11%	0.10%	0.06%	0.07%
Jaccard	0.10%	0.04%	0.07%	0.05%
Adamic-Adar	0.10%	0.08%	0.09%	0.07%

L’analyse des Tableaux 4 et 5 révèle un paradoxe apparent : la précision augmente avec la fraction de liens retirés, passant de 24-28% à $f = 0.05$ jusqu’à 70-72% à $f = 0.20$ pour les meilleures métriques. Cet effet s’explique par un artefact méthodologique : lorsque le nombre de liens à retrouver croît de 1 256 à 5 025, la probabilité qu’un lien cible figure parmi les top-50 candidats augmente mécaniquement. Le rappel, qui mesure la couverture réelle, nuance drastiquement ce constat : à $f = 0.20$, Adamic-Adar atteint 72% de précision mais seulement 0.07% de rappel, ne retrouvant que 36 liens sur 5 025. La précision élevée traduit la qualité des prédictions émises, mais l’algorithme demeure aveugle à 99% des connexions latentes.

Sur MIT8, Adamic-Adar domine pour $f \geq 0.10$ avec un plateau de 70-72%, capitalisant sur la pondération logarithmique des voisins communs rares dans ce réseau de densité intermédiaire (1.23%). Jaccard présente en revanche un comportement erratique, chutant à 20% pour $f = 0.10$ avant de rebondir à 50% pour $f \geq 0.15$, suggérant une sensibilité particulière aux perturbations locales du graphe. Pour $f = 0.20$, Common Neighbors et Adamic-Adar convergent à 70%, indiquant qu’au-delà d’un seuil de dégradation, la sophistication algorithmique n’apporte plus d’avantage significatif. Ces résultats confirment que la hiérarchie des métriques observée à $f = 0.10$ sur les 15 réseaux n’est pas un artefact : les tendances (dominance de CN sur réseaux denses, supériorité de Jaccard sur graphes dispersés) restent qualitativement valides sur l’ensemble du spectre testé.

5. Inférence d’attributs par propagation de labels

L’algorithme de Label Propagation exploite l’homophilie du réseau pour inférer les attributs manquants. Son principe est simple : chaque nœud non-étiqueté adopte itérativement l’attribut majoritaire parmi ses voisins jusqu’à convergence. Nous évaluons cette approche sur les 15 réseaux en masquant aléatoirement 10%, 20% et 30% des labels, puis en mesurant la précision de reconstruction.

TABLE 6 – Précision de Label Propagation sur 8 réseaux (fraction masquée = 10%)

Réseau	Nœuds	Dortoir	Classes	Filière	Classes	Genre	Classes
Caltech36	762	94.9%	8	27.9%	30	68.1%	2
Rice31	4 083	92.1%	9	31.8%	46	61.5%	2
American75	6 370	75.5%	25	28.8%	75	58.4%	2
MIT8	6 402	70.6%	63	34.6%	43	66.3%	2
Duke14	9 885	53.5%	135	29.6%	66	70.5%	2
UNC28	18 158	52.6%	93	26.5%	73	63.1%	2
Texas84	36 364	56.9%	92	30.6%	107	61.2%	2
Simmons81	1 510	40.0%	10	25.9%	53	99.3%	2

Les résultats confirment et amplifient les observations faites sur l’homophilie. L’attribut dortoir se propage remarquablement bien à travers le graphe sur les campus organisés en *Residential Colleges*. Caltech36 atteint 94.9% de précision malgré 8 dortoirs possibles, et Rice31 culmine à 92.1% avec 9 résidences. Ces deux universités fonctionnent selon un système de “maisons” où les étudiants partagent repas et activités, créant une homophilie résidentielle quasi-parfaite.

L’analyse de la filière d’études (*Major*) révèle un signal intermédiaire. Avec une précision

moyenne oscillant entre 25% et 35%, la propagation de labels fait nettement mieux que le hasard (qui serait inférieur à 2% vu le nombre élevé de filières, souvent supérieur à 50). Cependant, ce score reste bien en deçà des performances obtenues pour le dortoir. Cela indique que si les étudiants tendent à avoir des amis dans leur filière, ce n'est pas le facteur déterminant de leurs interactions sociales.

Sur les grandes universités comme Duke14 (135 dortoirs) ou UNC28 (93 dortoirs), la précision du dortoir reste honorable (53-55%) malgré le nombre élevé de classes. Ramené au hasard (qui donnerait moins de 1% sur Duke), ce score démontre que la structure du graphe encode efficacement l'information résidentielle.

Le genre présente une précision modeste mais constante à travers les campus, oscillant entre 58% et 70%. Cette performance, à peine supérieure à un classifieur naïf pondéré par les fréquences de classes, confirme quantitativement que le genre ne structure pas les réseaux d'amitié universitaires.

Le cas de Simmons81 mérite une attention particulière. Sur cette université historiquement féminine, la précision pour le genre atteint 99.3%, un score apparemment exceptionnel. Ce résultat paradoxal ne traduit cependant pas une structure genrée du réseau, mais plutôt la sensibilité de l'algorithme aux déséquilibres de classes. Dans un campus presque exclusivement féminin, prédire systématiquement "femme" constitue une stratégie gagnante mais dénuée de valeur prédictive réelle. Cette observation souligne l'importance d'interpréter les métriques de performance à la lumière de la distribution des classes.

6. Détection de communautés

La détection de communautés vise à partitionner le graphe en groupes de noeuds densément connectés entre eux. Nous testons l'hypothèse selon laquelle les communautés détectées algorithmiquement correspondent à des attributs socio-démographiques réels. L'algorithme de Louvain, qui optimise la modularité de manière gloutonne, est évalué via le score NMI (Normalized Mutual Information) mesurant la correspondance entre la partition obtenue et les attributs connus.

TABLE 7 – Scores NMI entre communautés Louvain et attributs réels

Réseau	Dortoir	Année	Major	Statut	Genre
Caltech36	0.702	0.082	0.076	0.033	0.009
Rice31	0.751	0.024	0.049	0.011	0.001
Reed98	0.126	0.477	0.062	0.124	0.004
Smith60	0.509	0.181	0.077	0.029	0.006

Les résultats montrent deux tendances principales concernant l'organisation des communautés.

D'un côté, le dortoir reste le facteur dominant pour Caltech et Rice ($NMI > 0.70$). L'algorithme de Louvain retrouve presque exactement les groupes d'étudiants logeant ensemble. De l'autre, pour Reed, c'est l'année de promotion qui prime ($NMI \approx 0.48$), ce qui indique une vie sociale organisée par tranche d'âge plutôt que par lieu de résidence.

Concernant la filière (*Major*), son faible score NMI (autour de 0.06) confirme qu'elle ne suffit

pas à créer des groupes d'amis distincts. Les étudiants de mêmes filières sont connectés, mais restent mélangés aux autres au sein du graphe global.

Le cas du statut (Étudiant/Professeur) est intéressant. Malgré une forte assortativité, son score NMI est bas car l'algorithme de Louvain découpe le réseau en une multitude de petites communautés (groupes d'amis), alors que le statut ne divise le graphe qu'en deux grands blocs. L'algorithme ne peut donc pas s'aligner sur cet attribut trop général, sauf partiellement à Reed (NMI = 0.12) où la séparation est plus marquée.

Enfin, le genre n'a aucun impact sur la formation des communautés (NMI proche de 0), confirmant que les groupes d'amis sont mixtes.

7. Conclusion

Cette étude comparative, menée sur un échantillon représentatif de 15 réseaux issus du dataset Facebook100, a permis de caractériser la structure topologique et sociale des interactions universitaires à grande échelle.

Nos analyses confirment l'universalité des propriétés de *Small World* et l'existence d'une structure hiérarchique au sein de ces graphes, tout en soulignant l'impact déterminant de la taille du réseau sur sa densité. La chute drastique de connectivité observée, passant de 5.7% pour les petits campus à 1.2% pour les grandes universités, illustre empiriquement les limites cognitives de sociabilisation théorisées par le nombre de Dunbar. Sur le plan algorithmique, nous avons démontré la robustesse générale de la métrique *Adamic-Adar* (performante sur Caltech, Rice et UNC28), tout en notant des exceptions notables comme Swarthmore où *Jaccard* s'impose (76%), confirmant l'importance d'adapter l'outil à la topologie locale.

L'apport principal de ce travail réside dans la hiérarchisation des déterminants sociaux. L'analyse croisée de l'homophilie et de la détection de communautés révèle que la sociabilité étudiante est avant tout régie par la proximité physique et temporelle. Le lieu de résidence s'impose comme le vecteur fondamental de formation des communautés sur la majorité des campus (Caltech, Rice), parfois relayé par l'année de promotion dans des environnements spécifiques (Reed). Si le statut institutionnel érige une barrière macroscopique étanche entre étudiants et personnel, il ne suffit pas à définir la granularité fine des groupes sociaux. À l'inverse, les clivages académiques (filière) ou de genre apparaissent négligeables, dépeignant des campus américains de 2005 où l'intégration sociale transcende les catégories disciplinaires et identitaires.