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

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 sociodémographiques 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. Plutôt que d'adopter une approche statistique globale consistant à agréger les résultats sur l'ensemble des graphes, nous privilégions une démarche comparative ciblée. Cette méthodologie permet d'analyser comment la topologie spécifique de chaque réseau influence la performance des algorithmes de prédiction de liens et de détection de communautés.

Notre analyse repose sur trois réseaux aux profils contrastés. Caltech36, avec ses 762 nœuds et une densité de 5.7%, représente une petite communauté soudée caractéristique des instituts technologiques d'élite. À l'opposé, UNC28 (University of North Carolina) compte 18 158 nœuds pour une densité inférieure à 0.01%, illustrant la structure dispersée des grandes universités publiques. Enfin, Reed98 occupe une position intermédiaire avec 962 nœuds et servira à valider nos hypothèses sur la structure communautaire.

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 remarquable. 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

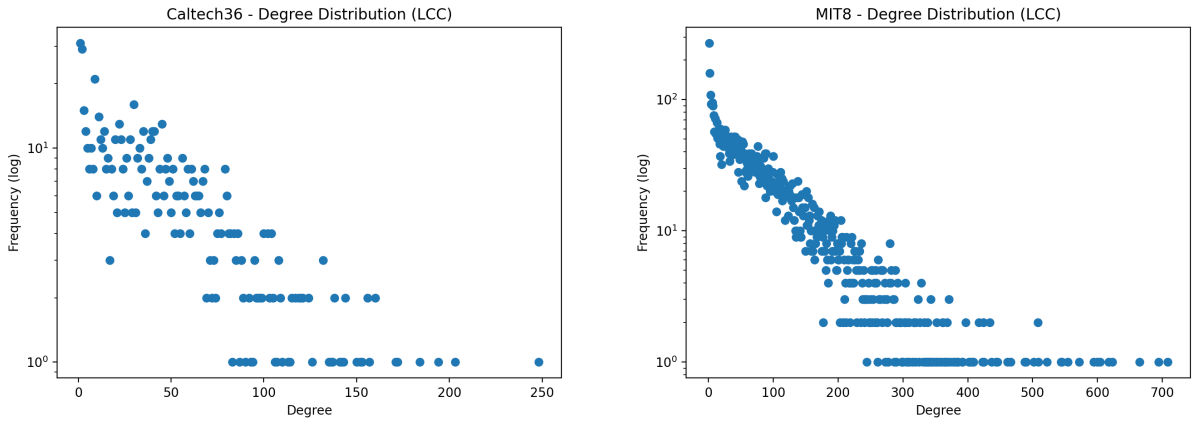
On observe que la densité s'effondre drastiquement avec la taille du réseau, passant de 5.7% pour Caltech à seulement 1.2% pour MIT. Ce résultat s'interprète aisément : le nombre de relations qu'un individu peut maintenir est cognitivement limité (nombre de Dunbar), tandis que le nombre de paires possibles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

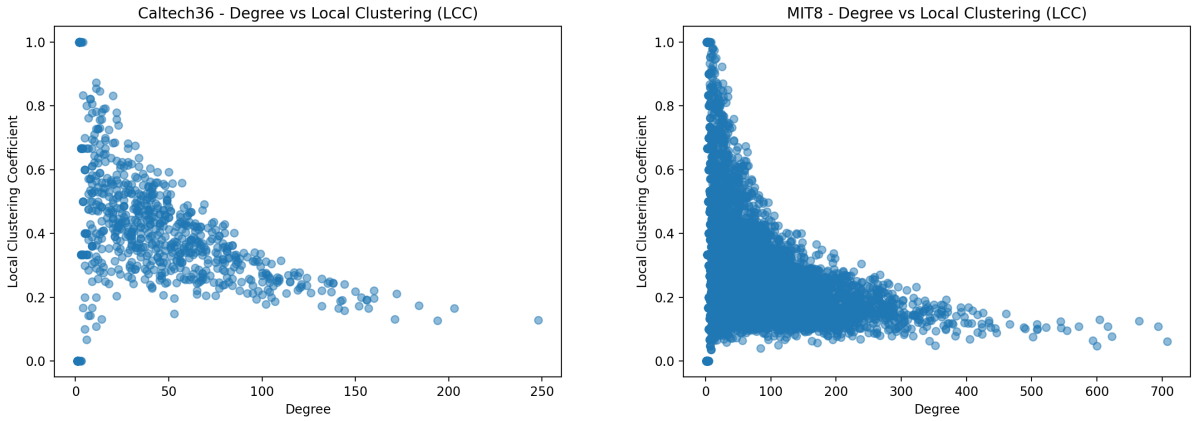


(a) Caltech36 (b) MIT8

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 nœud 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.



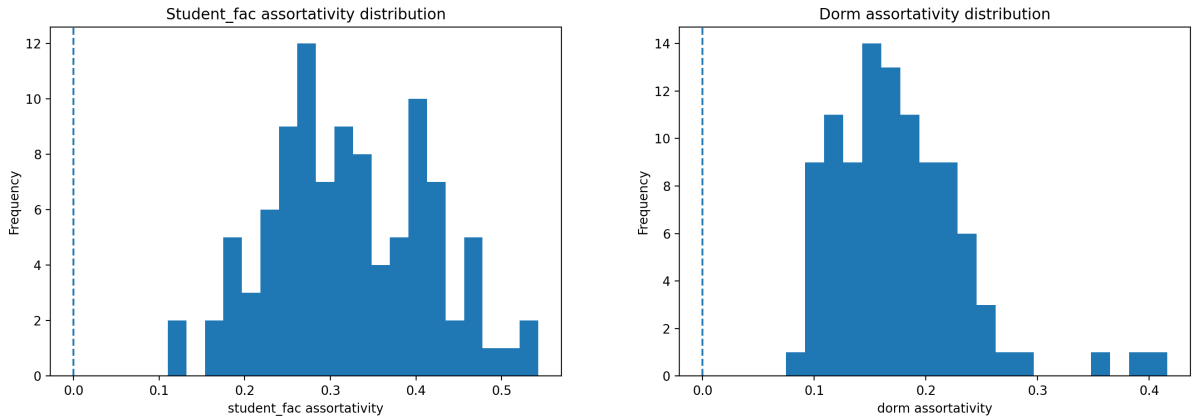
(a) Caltech36 (b) MIT8

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.



(a) Statut (Student/Faculty)

(b) Dortoir

FIGURE 3 – Distribution de l'assortativité par attribut sur les 100 campus

Le statut institutionnel (étudiant versus personnel académique) génère la ségrégation la plus marquée, avec une médiane d'assortativité proche de 0.35. 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 autour de 0.17, confirmant l'importance de la proximité géographique immédiate dans la création des amitiés.

En revanche, le genre ne constitue pas un facteur structurant des réseaux d'amitié universitaires. Comme l'illustre la Figure 4, la distribution de l'assortativité par genre est centrée sur zéro, attestant de la mixité des relations sociales sur les campus américains de 2005.

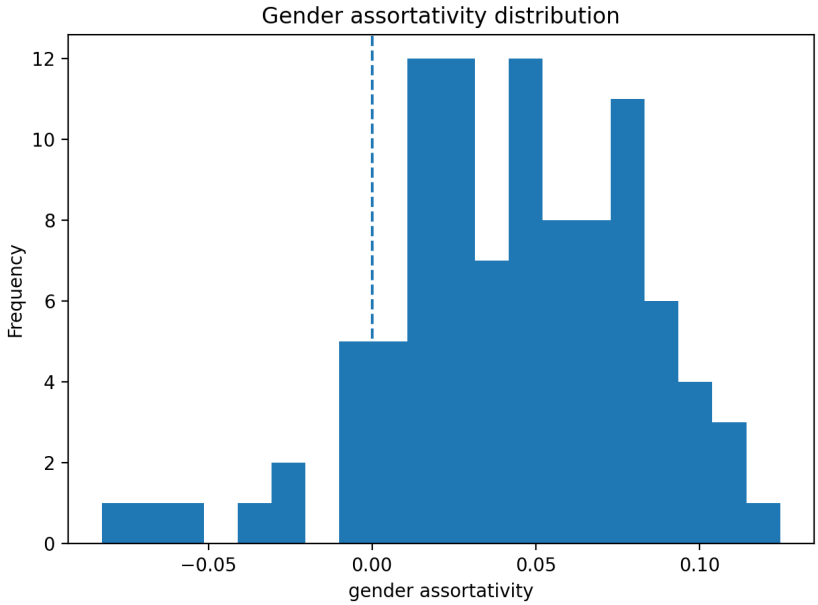


FIGURE 4 – Distribution de l'assortativité par genre — centrée sur zéro

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 :

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

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.

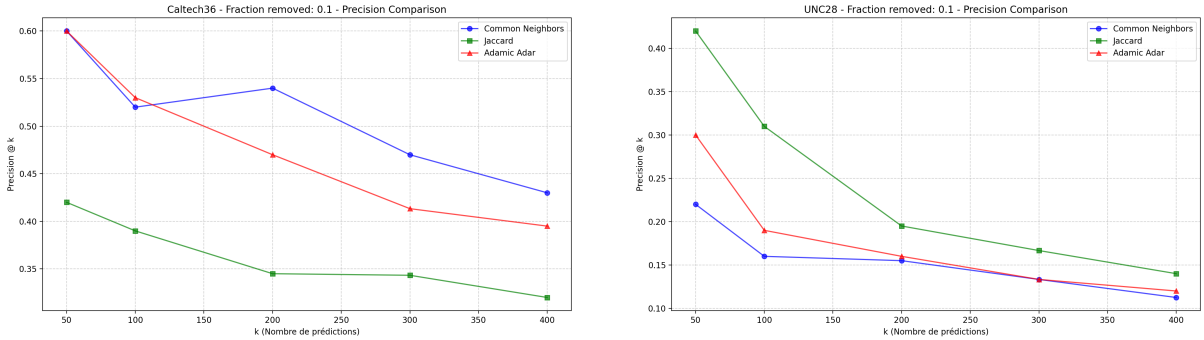
4.2. Influence de la topologie sur les performances

Les résultats obtenus révèlent une dépendance marquée entre la densité du réseau et la métrique optimale. Le Tableau 2 synthétise les précisions obtenues pour les 50 meilleures prédictions.

TABLE 2 – Precision@50 selon le réseau et la métrique

Réseau	**Common Neighbors**	**Jaccard**	**Adamic-Adar**
Caltech36 (dense)	**60%**	42%	**60%**
Reed98 (moyen)	42%	40%	**44%**
UNC28 (dispersé)	22%	**42%**	30%

Sur Caltech36, réseau dense et communautaire, les métriques non-normalisées (CN et AA) dominent nettement avec 60% de précision, contre seulement 42% pour Jaccard. Dans une petite communauté où tout le monde se connaît, le nombre absolu d'amis communs constitue un indicateur fiable de proximité sociale. La normalisation opérée par Jaccard s'avère ici contre-productive.



(a) Caltech36 — réseau dense

(b) UNC28 — réseau dispersé

FIGURE 5 – Courbes Precision@k illustrant l'inversion de performance selon la densité

Sur UNC28, la hiérarchie s'inverse de manière spectaculaire. Jaccard atteint 42% de précision tandis que CN chute à 22% et AA à 30%. Ce phénomène s'explique par la présence de *super-hubs* dans les grands réseaux dispersés. Ces étudiants exceptionnellement populaires partagent de nombreux voisins communs avec une grande partie du campus, conduisant CN à prédire — à tort — des liens entre individus qui ne se connaissent pas. La normalisation de Jaccard neutralise ce biais en divisant par l'union des voisinages.

Reed98 présente un comportement intermédiaire (Figure 6), avec un léger avantage pour Adamic-Adar qui combine le comptage des voisins communs avec une pondération pénalisant les hubs.

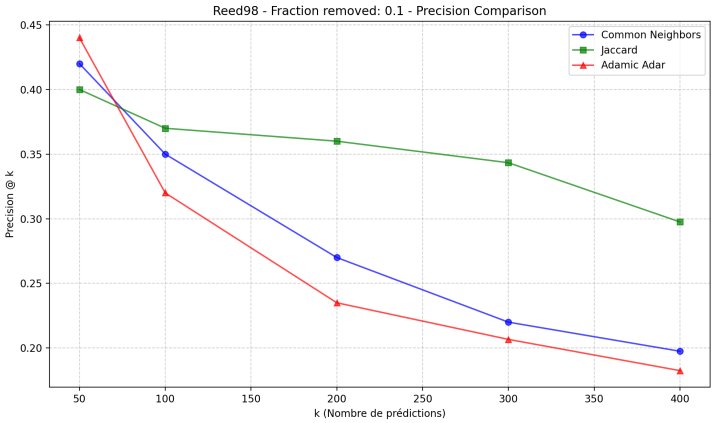


FIGURE 6 – Reed98 — comportement intermédiaire entre les deux régimes

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 pour les communautés denses, Jaccard pour les graphes dispersés à grande échelle.

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 en masquant aléatoirement 10%, 20% et 30% des labels, puis en mesurant la précision de reconstruction.

TABLE 3 – Précision de Label Propagation (fraction masquée = 10%)

Réseau	**Dortoir**	**Classes**	**Genre**	**Classes**
Duke14	53.5%	135	70.5%	2
Caltech36	**94.9%**	8	68.1%	2
MIT8	70.6%	63	66.3%	2
Simmons81	40.0%	10	**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, atteignant 94.9% de précision sur Caltech malgré 8 classes possibles, et 53.5% sur Duke avec 135 dortoirs distincts. Cette performance s'explique par la forte corrélation entre voisinage dans le graphe et proximité résidentielle : les colocataires et voisins de palier sont naturellement surreprésentés parmi les amis Facebook.

Le genre présente une précision plus modeste, oscillant entre 65% et 70% sur la plupart des campus. Cette performance, à peine supérieure à un classifieur naïf pondéré par les fréquences de classes, confirme que le genre ne structure pas les réseaux d'amitié.

Le cas de Simmons81 mérite une attention particulière. Sur cette université historiquement féminine, la précision pour le genre atteint 99.3%. Ce résultat paradoxal ne traduit 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 peu informative.

6. Détection de communautés

La détection de communautés vise à partitionner le graphe en groupes de nœuds 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 4 – Scores NMI entre communautés Louvain et attributs réels

Réseau	**Dortoir**	**Année**	**Genre**
Caltech36	**0.685**	0.104	0.013
Rice31	**0.791**	0.017	0.000
Reed98	0.152	**0.451**	0.007
Smith60	**0.495**	0.187	0.007

Sur Caltech36 et Rice31, les communautés détectées correspondent presque parfaitement aux dortoirs, avec des scores NMI respectifs de 0.685 et 0.791. Ces deux campus fonctionnent selon un système de *Residential Colleges* ou "maisons" qui structure fortement la vie sociale. Les étudiants d'une même résidence partagent repas, activités et espaces communs, créant naturellement des communautés denses.

Reed98 constitue une exception notable. Sur ce campus, c'est l'année de promotion qui structure le graphe social (NMI = 0.451) plutôt que le dortoir (NMI = 0.152). Cette particularité suggère une culture de campus différente où les liens inter-promotions sont plus rares, possiblement en raison de traditions ou d'un curriculum particulièrement exigeant favorisant la solidarité au sein de chaque cohorte.

Quel que soit le campus étudié, le genre n'apparaît jamais comme facteur de structuration communautaire, avec des scores NMI systématiquement proches de zéro. Les amitiés universitaires transcendent la dimension genrée, confirmant la mixité sociale structurelle observée dans l'analyse d'homophilie.

7. Conclusion

Cette étude met en évidence la complexité structurelle des réseaux sociaux universitaires et l'impossibilité de les traiter comme un ensemble homogène. Les graphes Facebook100 présentent une variabilité topologique considérable qui conditionne le choix des méthodes d'analyse.

Sur le plan topologique, tous les réseaux étudiés présentent les caractéristiques des graphes *Small World* : faible densité globale mais fort clustering local, distribution des degrés en loi de puissance avec présence de hubs structurants. Cependant, l'ampleur de ces phénomènes varie considérablement selon la taille du campus.

Pour la prédiction de liens, nos résultats démontrent l'absence de métrique universellement optimale. Les algorithmes non-normalisés comme Adamic-Adar excellent sur les réseaux denses

où le comptage des voisins communs reflète fidèlement la proximité sociale. En revanche, sur les grands graphes dispersés, la normalisation opérée par Jaccard devient indispensable pour filtrer les faux positifs générés par les super-hubs.

Du point de vue sociologique, la structure communautaire des campus apparaît dictée par des contraintes physiques (le dortoir de résidence) ou temporelles (l'année de promotion), mais jamais par le genre. Cette mixité sociale structurelle des campus américains en 2005 constitue un résultat robuste, confirmé tant par l'analyse d'homophilie que par la détection de communautés.

Enfin, il convient de mentionner une limite méthodologique de cette étude. L'échantillonnage des paires candidates (plafonné à 100 000) sur les grands réseaux comme UNC28 introduit une variance dans les résultats de prédiction de liens. Une analyse exhaustive nécessiterait des ressources de calcul significativement plus importantes, mais les tendances observées demeurent qualitativement robustes.