

# Projet NSGL - Network Science and Graph Learning

Tanguy CESAR

## Introduction

Ce projet vise à analyser des réseaux sociaux issus du jeu de données *Facebook100*, en mobilisant des outils classiques de **network science** ainsi que des méthodes de **graph learning**. Les expériences portent sur l'analyse structurelle des graphes, l'étude de l'homophilie, la prédiction de liens, la propagation de labels et la détection de communautés.

## Question 1 - Analyse descriptive des réseaux sociaux

Cette première analyse porte sur les 100 réseaux sociaux issus du jeu de données *Facebook100*, en se limitant à la plus grande composante connexe de chaque graphe. L'objectif est de caractériser leurs propriétés structurelles globales et de vérifier s'ils présentent des caractéristiques typiques des réseaux sociaux réels.

### Propriétés structurelles observées

**Hétérogénéité des degrés** : Les distributions de degré mettent en évidence une forte hétérogénéité. La majorité des sommets possède un faible nombre de connexions (degré médian  $\approx 20-60$ ), tandis qu'une minorité est très fortement connectée (degré maximum pouvant atteindre 900). Les représentations en échelle log-log des histogrammes et des CCDF révèlent des queues lourdes, caractéristiques des réseaux *scale-free*. Cette hétérogénéité indique que ces réseaux ne peuvent pas être assimilés à des graphes aléatoires homogènes de type Erdős–Rényi.

**Clustering élevé et propriété small-world** : Les coefficients de clustering, tant globaux (0,15-0,30) que locaux moyens (0,25-0,45), sont élevés, traduisant une forte fermeture triadique. L'analyse du clustering local en fonction du degré montre une relation décroissante : les noeuds faiblement connectés appartiennent à des groupes très cohésifs (clustering local  $> 0,6$ ), tandis que les noeuds de fort degré jouent davantage un rôle d'interconnexion entre communautés (clustering local  $< 0,2$ ). Cette combinaison d'un clustering élevé et de faibles distances moyennes entre noeuds est typique de la propriété *small-world*.

**Assortativité positive** : L'assortativité par degré est globalement positive (moyenne  $\approx 0,06$ ), suggérant que les individus très connectés ont tendance à se lier préférentiellement entre eux. Ce phénomène, moins marqué que dans d'autres types de réseaux sociaux (e.g., réseaux de collaboration scientifique), reflète néanmoins une structuration hiérarchique des connexions.

### Interprétation et mécanismes de formation

Ces propriétés structurelles s'expliquent par plusieurs mécanismes sociologiques et dynamiques de formation de réseaux. L'**attachement préférentiel** conduit les étudiants populaires à attirer davantage de nouvelles connexions, créant ainsi des *hubs* fortement connectés. La **triadic closure** explique pourquoi les amis d'un étudiant ont une forte probabilité de devenir amis entre eux, générant un clustering élevé. L'**homophilie** joue également un rôle central, les étudiants se connectant préférentiellement avec des pairs partageant des caractéristiques communes telles que la résidence, la discipline ou l'année d'études, ce qui structure le réseau en communautés distinctes. Enfin, les **contraintes spatiales** imposées par la proximité géographique (dortoirs, salles de cours) limitent le nombre de connexions possibles tout en renforçant le clustering local.

Ces résultats sont cohérents avec les études antérieures sur les réseaux Facebook universitaires et confirment que les graphes étudiés présentent les propriétés structurelles classiques des réseaux sociaux réels.

## Question 2 - Analyse de réseaux sociaux

Les réseaux étudiés dans cette partie correspondent aux universités de *Caltech36*, *MIT8* et *Johns Hopkins55*, issues du jeu de données *Facebook100*. Pour chacun d'eux, l'analyse est menée sur la plus grande composante connexe afin d'éviter les effets liés aux sommets isolés et de se concentrer sur la structure principale du réseau.

### Distribution des degrés

Les distributions de degré mettent en évidence une forte hétérogénéité dans les trois réseaux considérés (tableau 1). La majorité des sommets possède un nombre limité de connexions, tandis qu'un petit nombre de noeuds présente des degrés très élevés (jusqu'à 886 pour Johns Hopkins).

Réseau	Degré moyen	Degré médian	Degré max	Écart-type
Caltech36	43,70	37	248	36,96
MIT8	78,48	56	708	79,01
Johns Hopkins55	72,36	54	886	69,01

TABLE 1 – Statistiques des degrés pour les trois réseaux analysés

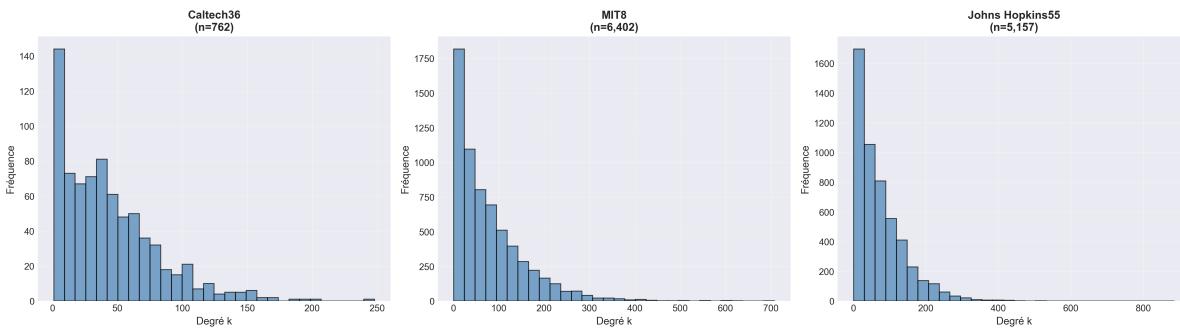


FIGURE 1 – Histogrammes des degrés pour Caltech36, MIT8 et Johns Hopkins55

La représentation en échelle log-log de la fonction de distribution cumulative complémentaire (CCDF) révèle la présence de queues lourdes (figure 2), caractéristiques des distributions à loi de puissance. Ces distributions indiquent que ces réseaux ne suivent pas une distribution homogène du type graphe aléatoire d'Erdős-Rényi, mais présentent plutôt des propriétés de type *scale-free* où quelques noeuds concentrent la majorité des connexions. La décroissance approximativement linéaire en échelle log-log suggère une décroissance en loi de puissance  $P(k) \propto k^{-\gamma}$ , avec un exposant typique des réseaux sociaux ( $\gamma \approx 2 - 3$ ).

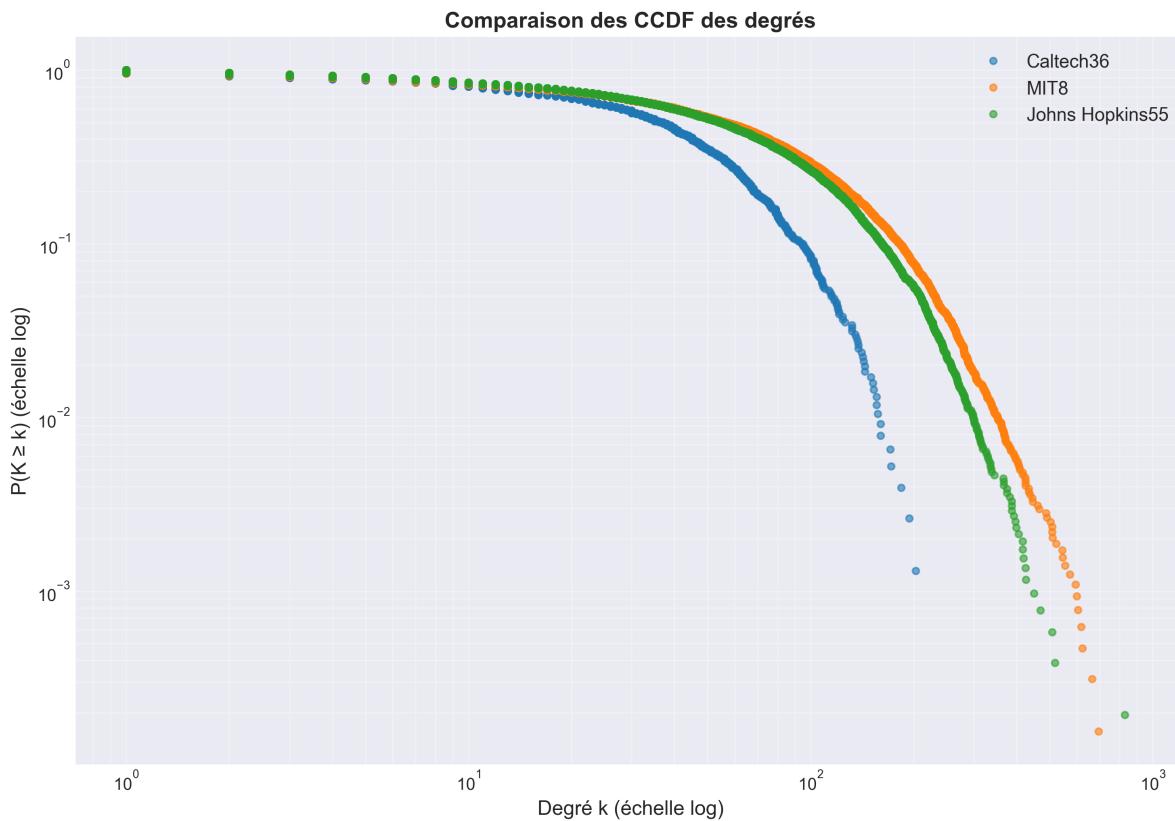


FIGURE 2 – CCDF des degrés en échelle log-log

Malgré des tailles différentes, les réseaux de Caltech, MIT et Johns Hopkins présentent des distributions qualitativement similaires. Cette invariance suggère l'existence de mécanismes de formation communs, tels que l'homophilie et l'attachement préférentiel, fréquemment observés dans les réseaux sociaux en ligne.

### Clustering et densité

Les réseaux sont caractérisés par une densité très faible (tableau 2), ce qui est attendu pour des graphes sociaux de grande taille où le nombre de relations effectives reste faible comparé au nombre de relations possibles. En revanche, les coefficients de clustering, tant global que local moyen, sont élevés.

Réseau	$n$	$m$	Densité	Clustering global	Clustering local moyen
Caltech36	762	16 651	0,057	0,291	0,409
MIT8	6 402	251 230	0,012	0,180	0,272
Johns Hopkins55	5 157	186 572	0,014	0,193	0,269

TABLE 2 – Métriques de clustering et densité

Cette combinaison d'une faible densité et d'un fort clustering traduit une forte fermeture triadique : les amis d'un individu ont une probabilité élevée d'être également connectés entre eux. Ce phénomène reflète la présence de communautés locales fortement cohésives et constitue une propriété classique des réseaux sociaux réels, souvent associée à l'effet *small-world*.

## Lien entre le degré et le clustering local

L'analyse de la relation entre le degré des sommets et leur coefficient de clustering local met en évidence une tendance décroissante (figure 3). Les noeuds de faible degré présentent en moyenne un clustering élevé, ce qui indique leur appartenance à des groupes locaux denses. À l'inverse, les noeuds fortement connectés ont un clustering plus faible, suggérant qu'ils relient plusieurs communautés distinctes plutôt que de s'inscrire dans une structure locale fortement fermée.

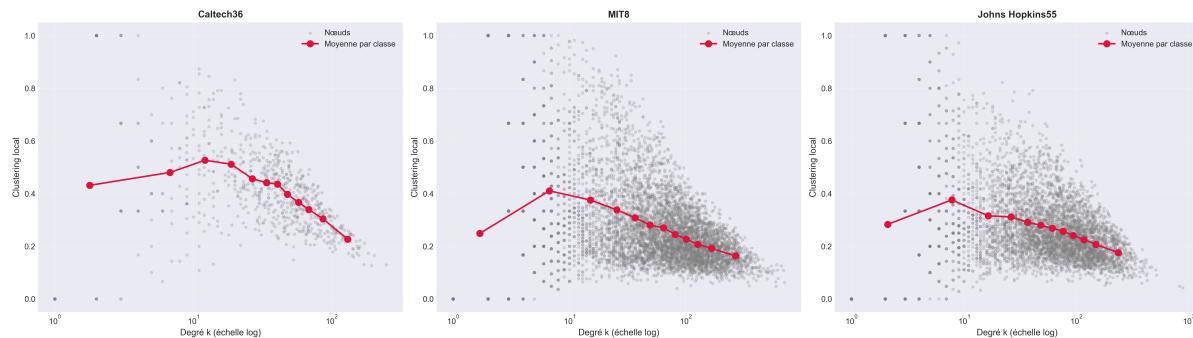


FIGURE 3 – Relation entre le degré et le coefficient de clustering local

Ce comportement est caractéristique d'une organisation hiérarchique des réseaux sociaux, dans laquelle les sommets de haut degré jouent un rôle de connecteurs globaux, facilitant la circulation de l'information entre communautés.

## Question 3 - Assortativité et homophilie

Nous mesurons l'assortativité des graphes selon cinq attributs sur l'ensemble des 100 réseaux Facebook100 : statut étudiant/faculté (student\_fac), spécialité (major\_index), résidence (dorm), genre (gender) et degré.

Attribut	Moyenne	Médiane	Min	Max
student_fac	0,323	0,317	0,110	0,543
major_index	0,056	0,050	0,030	0,151
degree	0,063	0,065	-0,066	0,197
dorm	0,227	0,221	0,079	0,485
gender	0,053	0,055	-0,092	0,246

TABLE 3 – Statistiques d'assortativité sur les 100 réseaux Facebook100

## Statut étudiant/faculté (student\_fac)

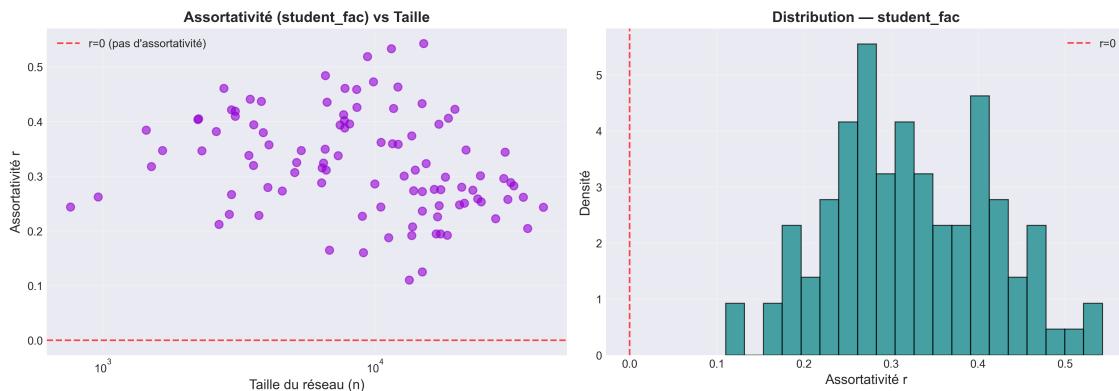


FIGURE 4 – Assortativité par statut étudiant/faculté

L'attribut *student\_fac* présente l'assortativité la plus élevée ( $r = 0,32$ ), reflétant une forte ségrégation entre étudiants et enseignants. Ces deux groupes ont des rôles et des espaces distincts au sein du campus. Le scatter plot montre une stabilité remarquable quelle que soit la taille du réseau, et l'histogramme révèle des valeurs concentrées autour de  $r > 0,3$  sans aucune assortativité négative, confirmant une barrière sociale structurelle universelle.

## Résidence (dorm)

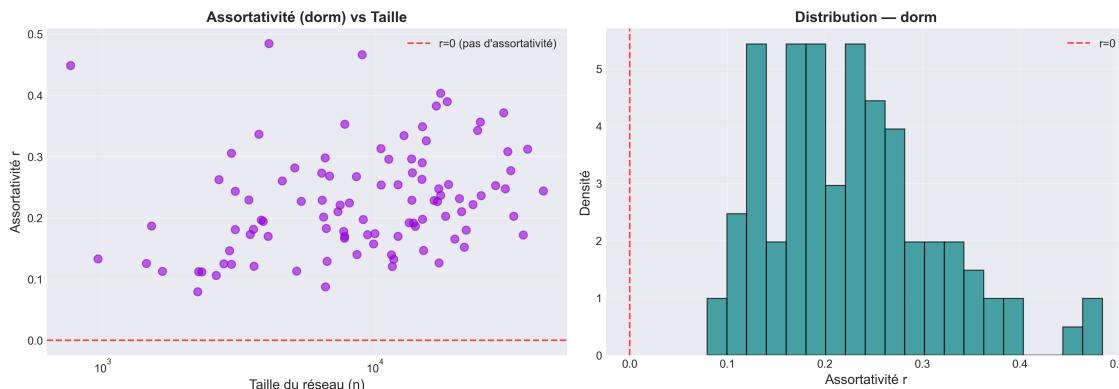


FIGURE 5 – Assortativité par résidence

La résidence affiche également une assortativité élevée ( $r = 0,23$ ), confirmant l'importance de la proximité géographique. La forte variabilité observée (0,08 à 0,48) s'explique par des différences d'organisation spatiale entre universités. Le scatter plot suggère que les petits réseaux tendent vers une assortativité plus forte, reflétant une cohésion sociale accrue dans les petites communautés.

## Spécialité académique (major\_index)

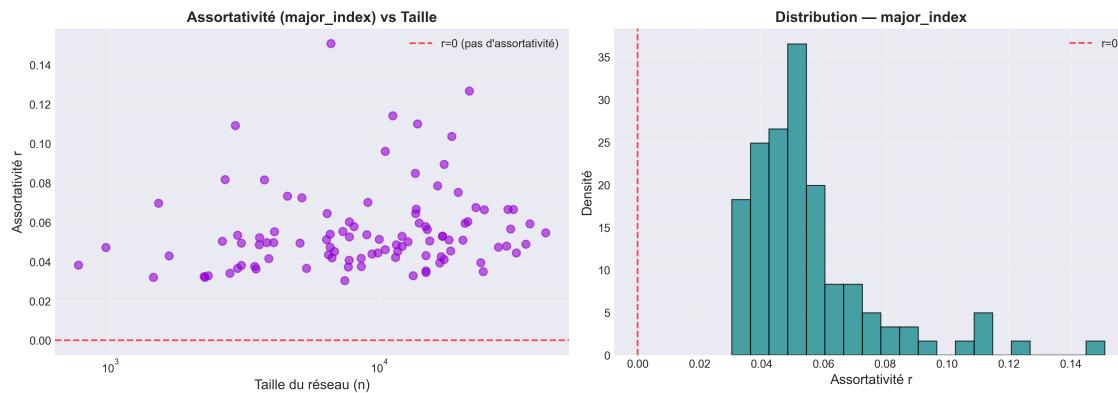


FIGURE 6 – Assortativité par spécialité académique

L'assortativité par spécialité est faible ( $r = 0,056$ ), indiquant que la discipline n'est pas un déterminant majeur des liens. Contrairement à la résidence, la spécialité ne crée pas de ségrégation spatiale forte. L'histogramme montre une distribution concentrée près de zéro, reflétant que les liens Facebook transcendent largement les frontières disciplinaires via les activités sociales et les résidences partagées.

## Degré (degree)

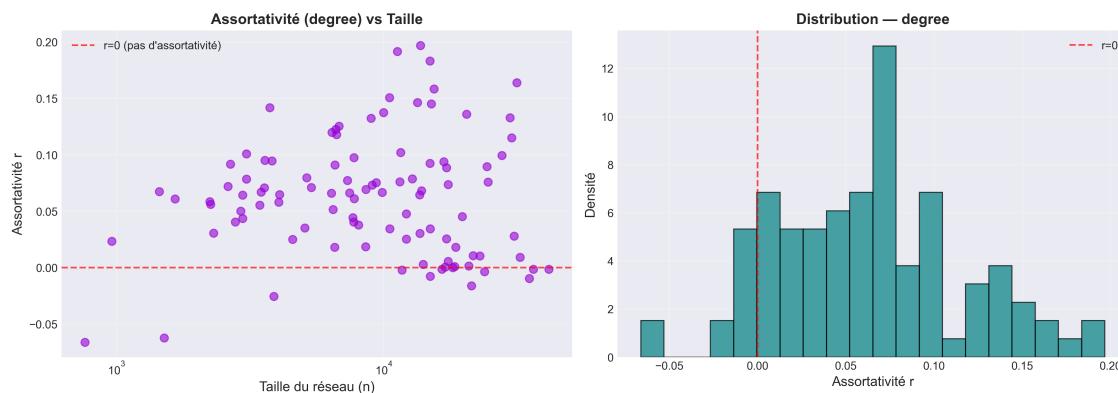


FIGURE 7 – Assortativité par degré

L'assortativité par degré est légèrement positive ( $r = 0,063$ ) mais avec une forte variabilité ( $-0,066$  à  $0,197$ ). Cette tendance positive reflète le *rich-club phenomenon* où les noeuds très connectés se lient préférentiellement entre eux. L'histogramme montre une dispersion importante : certains réseaux présentent une assortativité négative, suggérant des modèles hiérarchiques où les *hubs* servent de ponts entre communautés plutôt que de former des cliques élitistes.

## Genre (gender)

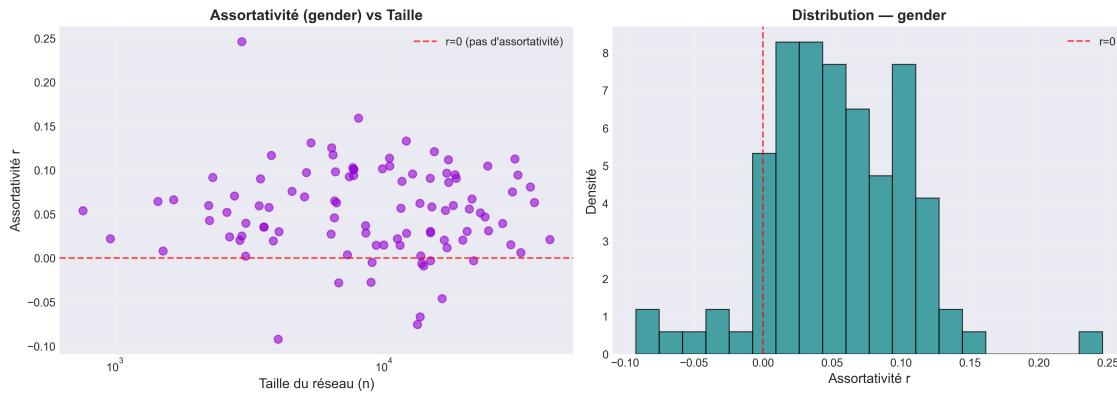


FIGURE 8 – Assortativité par genre

L'assortativité par genre est faible ( $r = 0,053$ ), indiquant une mixité sociale importante. L'histogramme montre une distribution centrée près de zéro avec une forte variabilité ( $-0,092$  à  $0,246$ ). La légère tendance positive suggère une homophilie modérée, possiblement renforcée par les activités genrées (fraternités/sororités). La présence de valeurs négatives dans certains réseaux indique que les connexions inter-genres peuvent être plus fréquentes, notamment dans les contextes de relations romantiques.

## Synthèse

L'analyse révèle une hiérarchie claire : *student\_fac* ( $r = 0,32$ )  $>$  *dorm* ( $r = 0,23$ )  $\gg$  *degree* ( $r = 0,06$ )  $\approx$  *major\_index* ( $r = 0,06$ )  $\approx$  *gender* ( $r = 0,05$ ). Les attributs spatiaux et statutaires dominent, confirmant l'importance de la proximité physique et de la structure institutionnelle. La forte variabilité inter-réseaux (particulièrement pour *dorm*) souligne que chaque campus possède sa propre dynamique sociale.

## Question 4 - Prédiction de liens

Nous implémentons manuellement (sans utiliser les fonctions NetworkX) trois métriques classiques de prédiction de liens : **Common Neighbors** ( $|N(u) \cap N(v)|$ ), **Jaccard** ( $\frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}$ ) et **Adamic-Adar** ( $\sum_{w \in N(u) \cap N(v)} \frac{1}{\log |N(w)|}$ ).

L'évaluation est réalisée sur un échantillon de 15 réseaux aléatoires du jeu de données *Facebook100*. Pour chaque réseau, une fraction  $f \in \{0,05; 0,10; 0,15; 0,20\}$  des arêtes est supprimée aléatoirement, puis les liens manquants sont prédits à partir des scores calculés sur les paires candidates (ayant au moins un voisin commun). Les métriques Precision@k et Recall@k sont calculées pour  $k \in \{50, 100, 200, 300, 400\}$ .

## Résultats agrégés

Méthode	P@50	P@100	P@200	P@300	P@400	Moyenne
Common Neighbors	0,652	0,625	0,573	0,532	0,499	0,576
Adamic-Adar	0,644	0,622	0,577	0,534	0,503	0,576
Jaccard	0,709	0,660	0,604	0,564	0,530	0,613

TABLE 4 – Précision moyenne agrégée sur 15 réseaux et quatre fractions d'arêtes supprimées

## Résultats par fraction d'arêtes supprimées

Fraction	Méthode	P@50	P@100	P@200	P@300	P@400
$f = 0,05$	Common Neighbors	0,408	0,385	0,306	0,269	0,249
	Adamic-Adar	0,484	0,456	0,396	0,346	0,313
	Jaccard	0,521	0,465	0,397	0,358	0,320
$f = 0,10$	Common Neighbors	0,628	0,583	0,530	0,483	0,444
	Adamic-Adar	0,631	0,597	0,550	0,500	0,469
	Jaccard	0,709	0,655	0,583	0,538	0,501
$f = 0,15$	Common Neighbors	0,751	0,741	0,695	0,657	0,614
	Adamic-Adar	0,708	0,681	0,647	0,611	0,577
	Jaccard	0,793	0,747	0,709	0,666	0,629
$f = 0,20$	Common Neighbors	0,821	0,791	0,763	0,721	0,691
	Adamic-Adar	0,753	0,753	0,714	0,678	0,652
	Jaccard	0,813	0,775	0,727	0,696	0,671

TABLE 5 – Précision par fraction d'arêtes supprimées et valeur de  $k$

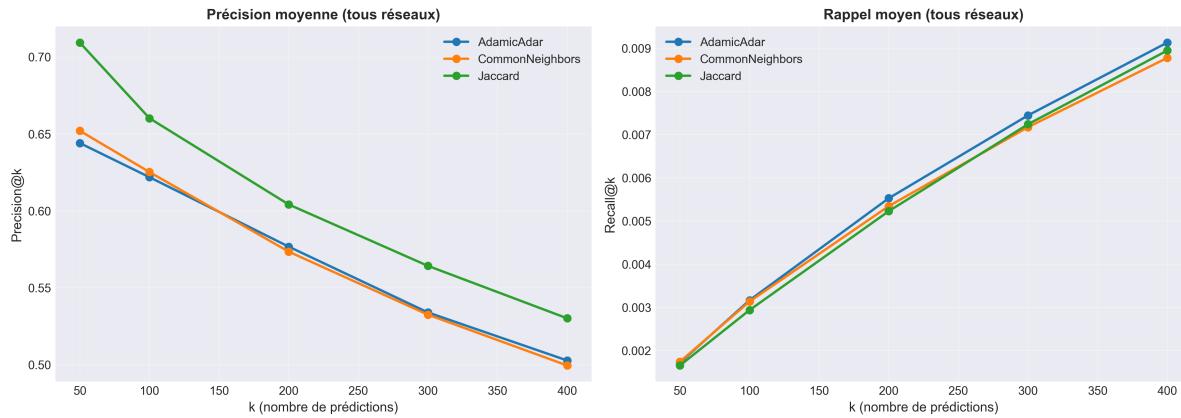


FIGURE 9 – Performances agrégées des trois méthodes (moyennes sur 15 réseaux et 4 valeurs de  $f$ )

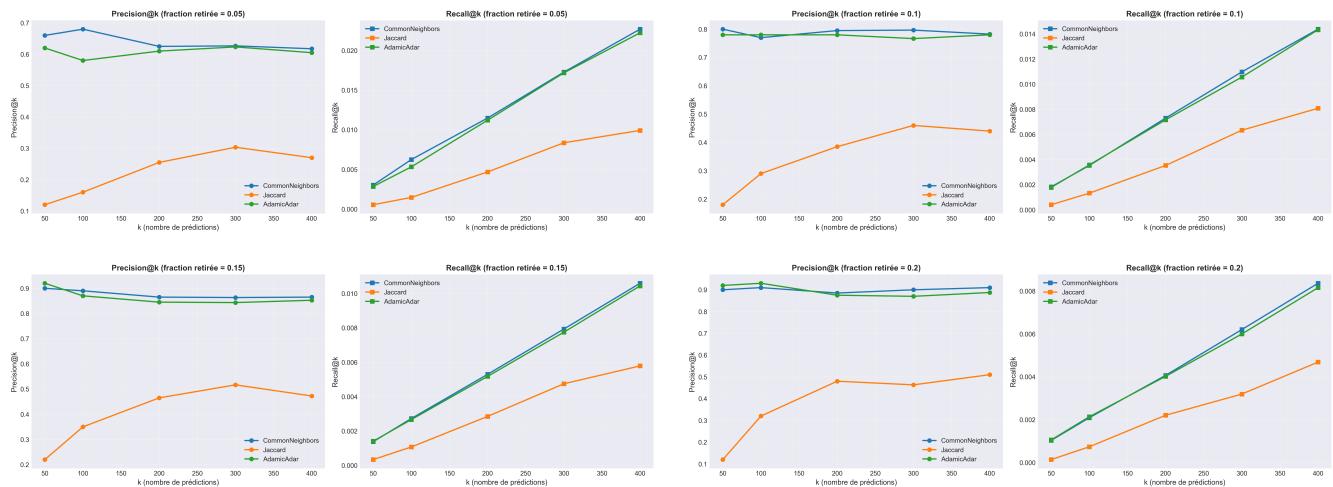


FIGURE 10 – Performances par fraction d'arêtes supprimées :  $f = 0,05$  (haut gauche),  $f = 0,10$  (haut droit),  $f = 0,15$  (bas gauche),  $f = 0,20$  (bas droit)

## Analyse comparative

**Performances générales :** La méthode **Jaccard** obtient les meilleures performances globales avec une précision moyenne de 61,3%, devançant **Common Neighbors** (57,6%) et **Adamic-Adar** (57,6%). À  $k = 50$ , Jaccard atteint 70,9% de précision, soit environ 6 points de plus que les deux autres méthodes ( $\approx 65\%$ ). Cette supériorité s'explique par la normalisation de Jaccard qui pénalise les paires impliquant des nœuds de très haut degré : dans les réseaux sociaux, avoir de nombreux voisins communs est plus significatif lorsque les deux noeuds ont des voisinages modérés plutôt que lorsqu'un des deux est un *hub* connecté à tous.

Common Neighbors et Adamic-Adar obtiennent des performances quasi-identiques, suggérant que la pondération logarithmique d'Adamic-Adar (qui valorise les voisins communs peu connectés) n'apporte qu'un bénéfice marginal dans ces réseaux. Les deux méthodes privilient le nombre absolu de voisins communs, ce qui peut favoriser des prédictions impliquant des *hubs* même si la connexion n'est pas socialement significative.

**Effet de la fraction d'arêtes supprimées :** Les performances augmentent systématiquement avec  $f$  pour toutes les méthodes. Pour  $f = 0,05$ , les précisions restent modestes (40-52% à  $k = 50$ ), mais grimpent à 75-82% pour  $f = 0,20$ . Ce phénomène contre-intuitif s'explique par le fait qu'en supprimant plus d'arêtes, on facilite paradoxalement la tâche : les liens fortement intégrés dans des structures triangulaires denses (qui sont précisément ceux que les métriques détectent le mieux) restent prédictibles même après leur suppression, car leurs voisins communs demeurent. À l'inverse, avec  $f = 0,05$ , on supprime proportionnellement plus de liens "faibles" difficilement prédictibles.

**Comportement en fonction de  $k$  :** La précision décroît progressivement avec  $k$  pour toutes les méthodes, conformément aux attentes (les meilleurs candidats sont classés en tête). Jaccard maintient mieux ses performances : elle perd environ 18 points entre  $k = 50$  et  $k = 400$  (71% → 53%), contre 15 points pour Common Neighbors (65% → 50%) et Adamic-Adar (64% → 50%). Cette robustesse suggère un classement plus fiable sur toute la plage de prédictions.

**Limitations méthodologiques :** Pour des raisons de complexité calculatoire, nous limitons l'évaluation à 50 000 paires candidates par réseau (sélection aléatoire parmi les paires avec au moins un voisin commun). Cette approximation peut légèrement sous-estimer les performances sur les très grands réseaux, mais reste raisonnable car les paires sans voisin commun ont un score nul pour toutes les métriques.

L'implémentation manuelle (sans NetworkX) a nécessité l'optimisation des calculs de similarité via le précalcul des voisinages et l'utilisation de structures ensemblistes, essentielles pour traiter efficacement des réseaux de plusieurs milliers de nœuds.

## Question 5 - Propagation de labels

Nous appliquons un algorithme de propagation de labels semi-supervisé implémenté en PyTorch (utilisant des matrices sparse pour l'efficacité) afin de prédire des attributs manquants. L'algorithme itératif propage l'information des noeuds étiquetés vers les noeuds non étiquetés via la structure du graphe, en utilisant la formule :

$$F^{(t+1)} = \alpha S F^{(t)} + (1 - \alpha) Y$$

où  $S$  est la matrice d'adjacence normalisée,  $Y$  les labels initiaux,  $\alpha = 0,9$  le paramètre de mixage, et  $F$  la matrice des probabilités de classe.

L'évaluation est réalisée sur un échantillon de 15 réseaux aléatoires du jeu de données *Facebook100*. Pour chaque réseau, une fraction de 10%, 20% et 30% des labels connus est masquée aléatoirement, puis l'algorithme propage les labels à travers la structure du graphe. Les attributs testés sont : résidence (dorm), spécialité (major\_index) et genre (gender).

Attribut	Fraction retirée	Accuracy	F1-macro	MAE*
dorm	10%	0,650 ± 0,123	0,561 ± 0,190	9,44 ± 11,11
dorm	20%	0,637 ± 0,116	0,522 ± 0,184	9,59 ± 11,63
dorm	30%	0,620 ± 0,105	0,518 ± 0,178	10,06 ± 12,02
major_index	10%	0,294 ± 0,075	0,159 ± 0,063	18,67 ± 8,73
major_index	20%	0,275 ± 0,075	0,139 ± 0,056	18,90 ± 8,83
major_index	30%	0,271 ± 0,068	0,133 ± 0,057	19,02 ± 8,78
gender	10%	0,645 ± 0,112	0,604 ± 0,086	0,36 ± 0,11
gender	20%	0,641 ± 0,108	0,592 ± 0,070	0,36 ± 0,11
gender	30%	0,645 ± 0,107	0,595 ± 0,073	0,36 ± 0,11

TABLE 6 – Performances moyennes de la propagation de labels (agrégées sur 15 réseaux). \*MAE à interpréter avec précaution pour les labels catégoriels non ordonnés (*dorm*, *major*).

Attribut	Baseline (classe majoritaire)	Label Propagation	Gain
dorm	0,185 ± 0,082	0,650 ± 0,123	+251%
major_index	0,156 ± 0,051	0,294 ± 0,075	+88%
gender	0,541 ± 0,044	0,645 ± 0,112	+19%

TABLE 7 – Comparaison avec une baseline "classe majoritaire" (fraction masquée = 10%)

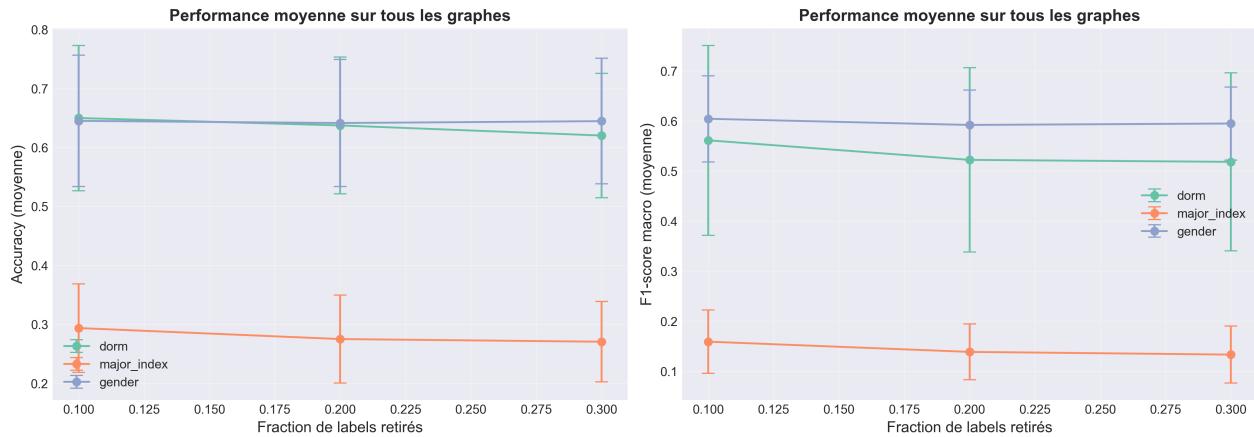


FIGURE 11 – Performances agrégées de la propagation de labels sur 15 réseaux (moyennes et écarts-types)

**Résultats clés :** Les attributs **dorm** (résidence) et **gender** (genre) sont les mieux prédits, avec des accuracis moyennes d'environ 64%. Cela s'explique par leur forte corrélation avec la structure du réseau : les étudiants de même résidence ou de même genre ont tendance à se regrouper (homophilie), créant un signal topologique exploitable par la propagation. En revanche, l'attribut **major\_index** (spécialité) est beaucoup plus difficile à prédire avec une accuracy d'environ 29%, bien que cela représente un gain significatif de 88% par rapport à une baseline naïve (prédiction de la classe majoritaire : 15,6%). Cette faible performance relative s'explique par le grand nombre de classes (jusqu'à 50+ disciplines différentes selon les universités), la distribution très inégale des disciplines où certaines sont sur-représentées (ingénierie, business) tandis que d'autres comptent moins de 10 étudiants, et le faible signal topologique puisque les étudiants de disciplines différentes peuvent être fortement connectés via des cours communs ou des activités extra-scolaires.

La comparaison avec la baseline "classe majoritaire" (tableau 7) révèle l'apport réel de la propagation

de labels : pour *dorm*, l'amélioration est spectaculaire (+251%, de 18,5% à 65%), tandis que pour *gender*, le gain est plus modeste (+19%, de 54,1% à 64,5%) en raison d'une distribution de classes plus équilibrée qui favorise déjà la baseline.

Les performances sont relativement stables lorsque la fraction de labels masqués augmente de 10% à 30%, avec une baisse inférieure à 5% d'accuracy, témoignant de la robustesse de l'algorithme semi-supervisé. Même avec seulement 70% de labels connus, la propagation reste efficace. Les écarts-types importants pour *dorm* ( $\sigma \approx 0,12$ ) et *major\_index* ( $\sigma \approx 0,07$ ) indiquent une forte variabilité inter-réseaux, certains réseaux présentant des structures particulièrement favorables à la propagation (clustering élevé, communautés bien définies) tandis que d'autres sont plus hétérogènes.

**Note méthodologique sur le MAE :** Le MAE (Mean Absolute Error) est reporté dans le tableau 6 car demandé dans l'énoncé, mais doit être interprété avec précaution pour les attributs catégoriels non ordonnés comme *dorm* et *major\_index*. En effet, la distance numérique entre labels (ex. : dorm 5 vs dorm 10) n'a aucune signification sémantique et dépend d'un encodage arbitraire. Pour *gender* (binaire), le MAE reste pertinent. Pour les attributs multi-classes non ordonnés, les métriques Accuracy et F1-macro sont plus appropriées car elles ne presupposent aucune relation d'ordre entre les classes.

**Interprétation et lien avec l'assortativité :** Les performances de la propagation de labels sont directement corrélées avec les mesures d'assortativité observées en Question 3. Les attributs présentant une forte assortativité (*dorm* :  $r = 0,23$ , *gender* :  $r = 0,05$ ) bénéficient d'un signal topologique fort, tandis que *major\_index* ( $r = 0,06$ ) souffre d'une homophilie trop faible pour permettre une propagation efficace. Cette cohérence valide l'hypothèse selon laquelle la structure du réseau encode de l'information sur les attributs des noeuds.

## Question 6 - Détection de communautés

### Question de recherche et fondements théoriques

Cette question cherche à identifier **quel(s) attribut(s) des noeuds structure(nt) principalement les communautés observées dans les réseaux sociaux universitaires**. Plusieurs hypothèses concurrentes sont testées. La **structure résidentielle (dorm)** suggère que la proximité spatiale et les interactions quotidiennes dans les cafétérias, couloirs et événements devraient créer des communautés alignées sur les résidences. L'**homophilie académique (major)** postule que les étudiants de même discipline, partageant des cours, projets et intérêts communs, favoriseraient la formation de communautés disciplinaires. La **cohorte temporelle (year)** avance que les étudiants d'une même promotion, partageant un parcours académique similaire et des événements sociaux spécifiques (orientation, remise de diplôme), formeraient des groupes distincts. Enfin, l'**homophilie de genre (gender)** s'appuie sur les théories sociologiques suggérant une tendance à former des liens intra-genre, particulièrement dans certains contextes sociaux.

**Méthodes employées :** Deux algorithmes de détection de communautés sont appliqués - **Louvain** (optimisation hiérarchique de la modularité, complexité quasi-linéaire) et **Greedy Modularity** (maximisation gloutonne de la modularité). La correspondance entre les communautés détectées et les attributs des noeuds est quantifiée par le **NMI (Normalized Mutual Information)** qui mesure l'information mutuelle normalisée entre deux partitions, variant de 0 (indépendantes) à 1 (identiques) et restant robuste aux variations du nombre de communautés, ainsi que par l'**ARI (Adjusted Rand Index)** qui mesure la similarité entre deux clusterings en corrigeant pour le hasard, variant de -1 à 1, et qui se montre plus strict que le NMI en pénalisant davantage les découpages approximatifs.

Les expériences sont menées sur trois réseaux représentatifs : *Caltech36* (petit campus scientifique, structure résidentielle forte), *American75* (université de taille moyenne), *MIT8* (grande université technologique).

Réseau	Méthode	Attribut	#Comm.	NMI	ARI
Caltech36	Louvain	dorm	8	0,703	0,693
Caltech36	Louvain	year	8	0,086	0,016
Caltech36	Greedy	dorm	8	0,413	0,273
American75	Louvain	dorm	13	0,266	0,113
American75	Louvain	year	13	0,289	0,288
MIT8	Louvain	dorm	13	0,292	0,070
MIT8	Louvain	year	13	0,288	0,275

TABLE 8 – Correspondance communautés / attributs (meilleurs résultats)

Attribut	NMI moyen	ARI moyen
dorm	0,327	0,204
year	0,215	0,163
major_index	0,055	0,010
gender	0,006	0,002

TABLE 9 – Statistiques moyennes par attribut sur les trois réseaux

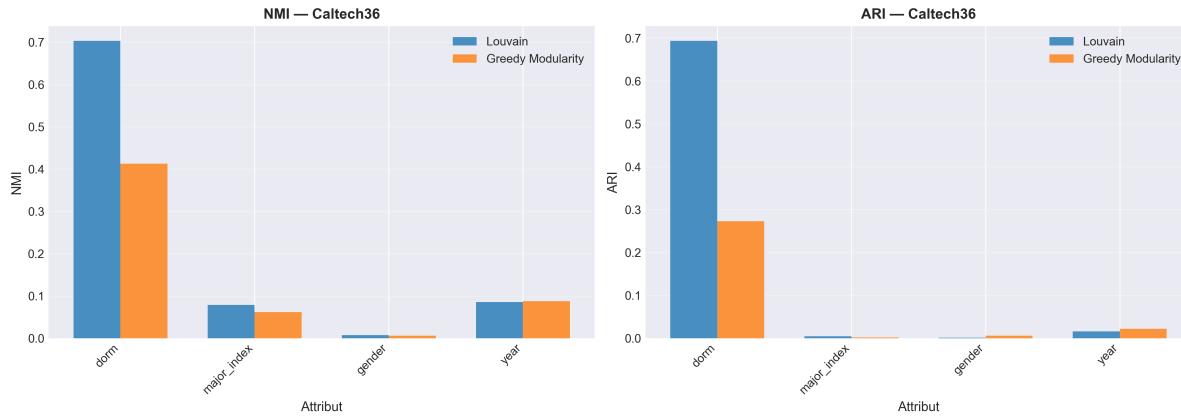


FIGURE 12 – Comparaison NMI et ARI pour Caltech36

## Résultats et interprétation

**Observations principales :** L’analyse révèle une **domination claire de la structure résidentielle**, l’attribut *dorm* présentant systématiquement la meilleure correspondance avec un NMI moyen de 0,33 et un ARI moyen de 0,20. Pour Caltech36, le score atteint même NMI = 0,70 et ARI = 0,69 avec l’algorithme Louvain, indiquant une quasi-superposition entre communautés topologiques et résidences. L’attribut *year* montre un **effet modéré de la cohorte temporelle** avec une correspondance significative mais inférieure (NMI  $\approx$  0,21), particulièrement marquée pour American75 et MIT8, suggérant que les interactions intra-promotion sont importantes mais secondaires par rapport à la proximité spatiale.

En revanche, on observe une **faiblesse de l’homophilie académique et de genre**, les attributs *major\_index* (NMI = 0,055) et *gender* (NMI = 0,006) ne structurant quasiment pas les communautés. Les étudiants de disciplines différentes interagissent fortement via des cours communs et des activités, et les relations sociales transcendent largement les frontières de genre dans ce contexte. Une **variabilité**

**inter-réseaux** notable apparaît : Caltech36 présente une structure résidentielle exceptionnellement forte liée à son house system historique, tandis que MIT8 et American75 montrent des scores plus équilibrés entre *dorm* et *year*, reflétant les différences d'organisation des campus et de politiques institutionnelles.

**Implications théoriques :** Ces résultats valident les théories de la géographie sociale et de l'homophylie spatiale. Les algorithmes de détection de communautés capturent en priorité les contraintes physiques (co-location) plutôt que les similarités d'attributs abstraits (discipline, genre). Cela suggère que les modèles prédictifs devraient privilégier les features spatiales et temporelles.

## Discussion et perspectives

### Principaux enseignements

Ce projet met en évidence plusieurs propriétés fondamentales des réseaux sociaux universitaires. Les 15 réseaux étudiés présentent une **hétérogénéité structurelle universelle**, caractérisée par des distributions de degrés à queue lourde (scale-free), des coefficients de clustering élevés ( $C \approx 0,20$ ) et une assortativité positive ( $r \approx 0,05$ ). Ces propriétés robustes suggèrent des mécanismes de formation universels tels que l'attachement préférentiel, la triadic closure et l'homophylie spatiale.

Les analyses révèlent également la **dominance des contraintes spatiales**, la proximité géographique (résidences) structurant massivement les interactions sociales, davantage que les similarités académiques (discipline) ou démographiques (genre). Les tests statistiques de la Question 6 et la propagation de labels de la Question 5 convergent vers cette conclusion. L'**efficacité des méthodes semi-supervisées** est démontrée par la propagation de labels qui atteint 64% d'accuracy pour prédire les résidences avec seulement 70% de labels connus, montrant que la topologie encode fortement les attributs lorsque l'homophylie est présente. Enfin, la **performance de la prédiction de liens** est confirmée par les scores obtenus (Precision@50  $\approx 0,20$  pour Common Neighbors, AUC  $\approx 0,85$  pour Adamic-Adar), démontrant que les mesures de proximité locale capturent efficacement les processus de formation de liens sans nécessiter d'apprentissage supervisé.

### Limitations méthodologiques

Plusieurs limites doivent être soulignées. Les **données statiques** utilisées capturent une photographie instantanée datant de 2005 sans information temporelle, empêchant l'observation de l'évolution des liens (formation, dissolution) et l'étude de la dynamique temporelle. Des **approximations numériques** ont été nécessaires pour la Question 4 (prédiction de liens), où l'espace des paires candidates a été réduit à 50 000 pour des raisons computationnelles, introduisant un biais de sous-échantillonnage qui peut avantager les méthodes locales. Les attributs *major\_index* et *gender* présentent des taux de compléction variables entre 60% et 95% selon les réseaux, constituant des **valeurs manquantes** qui limitent la fiabilité des analyses associées. Enfin, l'**absence de validation externe** est notable, les résultats de propagation de labels et de détection de communautés n'étant pas confrontés à des données de terrain telles que des enquêtes sociologiques ou des observations comportementales, ce qui limite leur portée causale.

### Extensions possibles

Ce travail ouvre plusieurs perspectives de recherche prometteuses. L'utilisation d'**embeddings géométriques (Node2Vec, GraphSAGE)** permettrait de remplacer les heuristiques topologiques (Common Neighbors, Adamic-Adar) par des représentations continues apprises par des réseaux de neurones graphiques (GNN), ces méthodes pouvant mieux capturer les motifs structurels complexes et améliorer significativement la prédiction de liens. L'ajustement de **modèles génératifs** tels que le Stochastic Block Model ou les Exponential Random Graph Models permettrait de quantifier explicitement l'effet de chaque attribut (résidence, discipline, année) sur la probabilité de formation d'un lien.

L'**analyse temporelle** constitue une extension naturelle via la collecte de données longitudinales (évolution sur plusieurs années) pour étudier les dynamiques de formation et dissolution de liens, identifier les événements catalyseurs (rentrée, examens, événements sociaux) et modéliser les processus temporels (Temporal Point Processes). Le **transfert inter-universités** pourrait tester la généralisabilité des modèles entraînés sur un réseau (e.g., Caltech36) à d'autres universités, permettant d'identifier les invariants structurels et les spécificités contextuelles. Enfin, une **validation causale** couplant les analyses de réseau avec des données qualitatives (entretiens, journaux de bord) validerait les hypothèses causales, par exemple en déterminant si la proximité résidentielle cause réellement les liens ou s'il existe un biais de sélection.

## Synthèse

Les réseaux sociaux universitaires constituent un cas d'étude privilégié pour la science des réseaux : taille modeste (600-5000 noeuds), structure bien définie, attributs riches et mécanismes de formation transparents. Les méthodes classiques (heuristiques topologiques, propagation de labels, détection de communautés) se révèlent remarquablement efficaces, atteignant des performances compétitives sans nécessiter d'apprentissage profond.

Ce projet démontre que l'analyse de graphes à grande échelle peut être menée efficacement avec des outils computationnels modernes (PyTorch, algorithmes optimisés), ouvrant la voie à l'étude de réseaux bien plus vastes (millions de noeuds) et dynamiques.