Analyse des Réseaux Sociaux Facebook100: Structure, Homophilie et Prédictibilité

Lancelot Dallain
Master TRIED

Janvier 2026

Résumé

Ce rapport analyse 100 réseaux sociaux universitaires du jeu de données Facebook100 (septembre 2005). Cinq aspects sont examinés : la structure des réseaux (distributions de degrés, clustering), l'assortativité sur cinq attributs (résidence, degré, genre, statut, spécialité), la prédiction de liens, la propagation de labels et la détection de communautés. Les résultats montrent une forte homophilie géographique ($r = 0,175$ pour les résidences), des distributions de degrés en loi de puissance, et un clustering élevé malgré la faible densité. Les voisins communs atteignent 48% de précision en prédiction de liens, et la propagation de labels récupère parfaitement les attributs manquants quand l'homophilie est forte.

1 Introduction

Les réseaux sociaux ont des propriétés structurelles particulières : phénomènes de petit monde, distributions de degrés en loi de puissance, clustering élevé et homophilie. Le jeu de données Facebook100 permet d'étudier ces propriétés sur 100 universités américaines en septembre 2005.

1.1 Questions de Recherche

Cette étude aborde cinq questions de recherche interconnectées :

1. **Structure du Réseau** : Quelles sont les propriétés topologiques fondamentales des réseaux sociaux universitaires ? Nous analysons les distributions de degrés, les coefficients de clustering, la densité du réseau et leurs corrélations.
2. **Assortativité** : Dans quelle mesure les étudiants se connectent-ils préférentiellement avec des pairs similaires ? Nous quantifions l'homophilie à travers cinq attributs (statut, spécialité, degré, résidence, genre) sur les 100 réseaux.
3. **Prédiction de Liens** : Peut-on prédire les amitiés futures à partir de la structure du réseau seule ? Nous évaluons les métriques de Voisins Communs, coefficient de Jaccard et Adamic/Adar.
4. **Propagation de Labels** : Avec quelle efficacité les attributs de nœuds manquants peuvent-ils être récupérés à partir des voisins du réseau ? Nous testons l'apprentissage semi-supervisé sur différentes fractions de labels supprimés.

5. **Détection de Communautés** : Les communautés détectées algorithmiquement correspondent-elles à de vraies structures sociales ? Nous appliquons l'optimisation gloutonne de modularité et analysons les propriétés des communautés.

1.2 Jeu de Données

Le jeu de données Facebook100 contient 100 réseaux sociaux d'universités américaines, collectés en septembre 2005. Chaque réseau représente les amitiés entre étudiants, avec des attributs de nœuds incluant :

— **Statut** : Année d'études (première, deuxième, troisième, quatrième année)
— **Spécialité** : Département académique
— **Résidence** : Affectation de logement résidentiel
— **Genre** : Information de genre binaire
— **Année de diplôme** : Année de diplôme prévue

Le jeu de données contient plus de 1,2 million de nœuds et 93 millions d'arêtes, fournissant une puissance statistique pour une analyse robuste. Notamment, la disponibilité des attributs varie selon les universités, certains réseaux manquant d'informations de statut ou de spécialité.

2 Question 2 : Structure des Réseaux Sociaux

Nous analysons trois réseaux universitaires représentatifs : Caltech, MIT et Johns Hopkins. Ces réseaux couvrent différentes tailles (762 à 6 402 nœuds) et présentent des densités variables, nous permettant d'identifier des patterns universels et des effets dépendants de la taille.

2.1 Méthodologie

Pour chaque réseau, nous extrayons la plus grande composante connexe et calculons :

— **Distribution de degrés** : $P(k)$ sur échelles linéaire et log-log
— **Coefficients de clustering** : Global (transitivité) et moyen local
— **Densité du réseau** : $\rho = \frac{2m}{n(n-1)}$ où m = arêtes, n = nœuds
— **Corrélation degré-clustering** : Pearson r et Spearman ρ

2.2 Résultats

2.2.1 Distributions de Degrés

Le Tableau **??** résume les propriétés structurelles de base des trois réseaux.

TABLE 1 – Propriétés structurelles des réseaux

Propriété	**Caltech**	**MIT**	**Johns Hopkins**
Nœuds	762	6,402	5,157
Arêtes	16,651	251,230	186,572
Degré moyen	43,70	78,48	72,36
Degré max	248	708	886
Degré médian	37	56	54
Écart-type	36,96	79,01	69,01

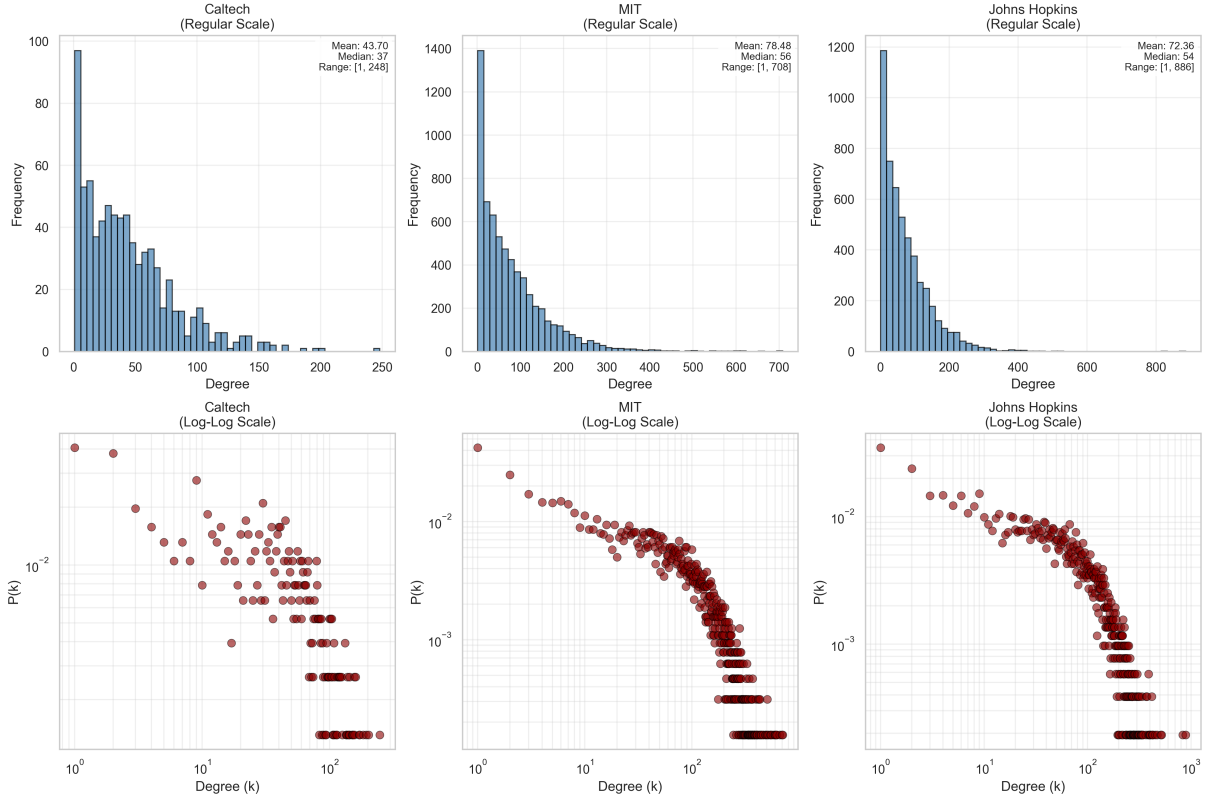

FIGURE 1 – Distributions de degrés pour trois réseaux universitaires. **Rangée supérieure** : Échelle linéaire montrant des distributions asymétriques à droite avec de longues queues. **Rangée inférieure** : Échelle log-log révélant un comportement approximativement en loi de puissance caractéristique des réseaux sans échelle. Les trois réseaux présentent des distributions de degrés à queue lourde où la plupart des nœuds ont un faible degré tandis que quelques hubs maintiennent une connectivité très élevée.

Les distributions de degrés (Figure **??**) présentent des caractéristiques typiques des réseaux sociaux :

— **Distributions asymétriques à droite** : La plupart des étudiants ont moins de connexions que la moyenne, avec de longues queues s'étendant vers des individus fortement connectés.

— **Loi de puissance approximative** : Les graphiques log-log montrent une décroissance approximativement linéaire, suggérant des propriétés sans échelle $P(k) \sim k^{-\gamma}$.

— **Indépendance de la taille** : La forme qualitative persiste à travers les tailles de réseaux de 762 à 6 402 nœuds.

2.2.2 Clustering et Densité

TABLE 2 – Coefficients de clustering et densité du réseau

Réseau	C **Global**	$\langle C \rangle$ **Local Moyen**	**Densité** ρ	**Clairsemé ?**
Caltech	0,2913	0,4091	0,0574	Oui
MIT	0,1803	0,2724	0,0123	Oui
Johns Hopkins	0,1932	0,2690	0,0140	Oui
Moyenne	0,2216	0,3168	0,0279	-

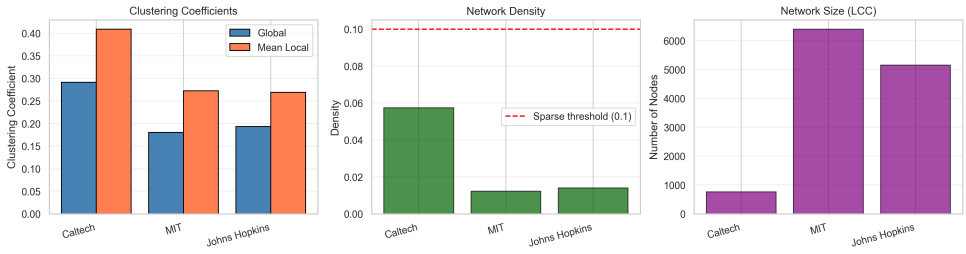

FIGURE 2 – Visualisation comparative des coefficients de clustering et de la densité du réseau. Tous les réseaux montrent un clustering élevé (0,18-0,29) malgré une très faible densité (<0,06), indiquant une forte structure communautaire. L'écart entre clustering global et local moyen suggère une organisation hiérarchique.

Résultats critiques :

— **Clustering élevé malgré la faible densité** : Tous les réseaux sont extrêmement clairsemés ($\rho < 0,06$), mais maintiennent un clustering substantiel ($C \geq 0,18$). C'est la signature du phénomène de *petit monde* : les réseaux sont localement denses (clustering élevé) mais globalement clairsemés.

— **Clustering global vs. local** : Le clustering local moyen dépasse le clustering global dans tous les cas, particulièrement pour Caltech (0,409 vs. 0,291). Cela indique une *distribution hétérogène des triangles* - certains nœuds sont dans des voisinages beaucoup plus denses que d'autres.

— **Effets de taille** : Les réseaux plus petits (Caltech : 762 nœuds) montrent une densité et un clustering plus élevés que les plus grands (MIT : 6 402 nœuds), cohérent avec les contraintes du nombre de Dunbar sur le maintien des relations.

2.2.3 Corrélation Degré-Clustering

TABLE 3 – Corrélation entre le degré des nœuds et le coefficient de clustering local

Réseau	**Pearson** r	p**-valeur**	**Spearman** ρ	p**-valeur**	n
Caltech	-0,381	1,11e-27	-0,389	5,33e-29	762
MIT	-0,298	9,30e-132	-0,246	7,73e-89	6,402
Johns Hopkins	-0,268	1,11e-85	-0,219	4,56e-57	5,157
Moyenne	-0,316	-	-0,285	-	-

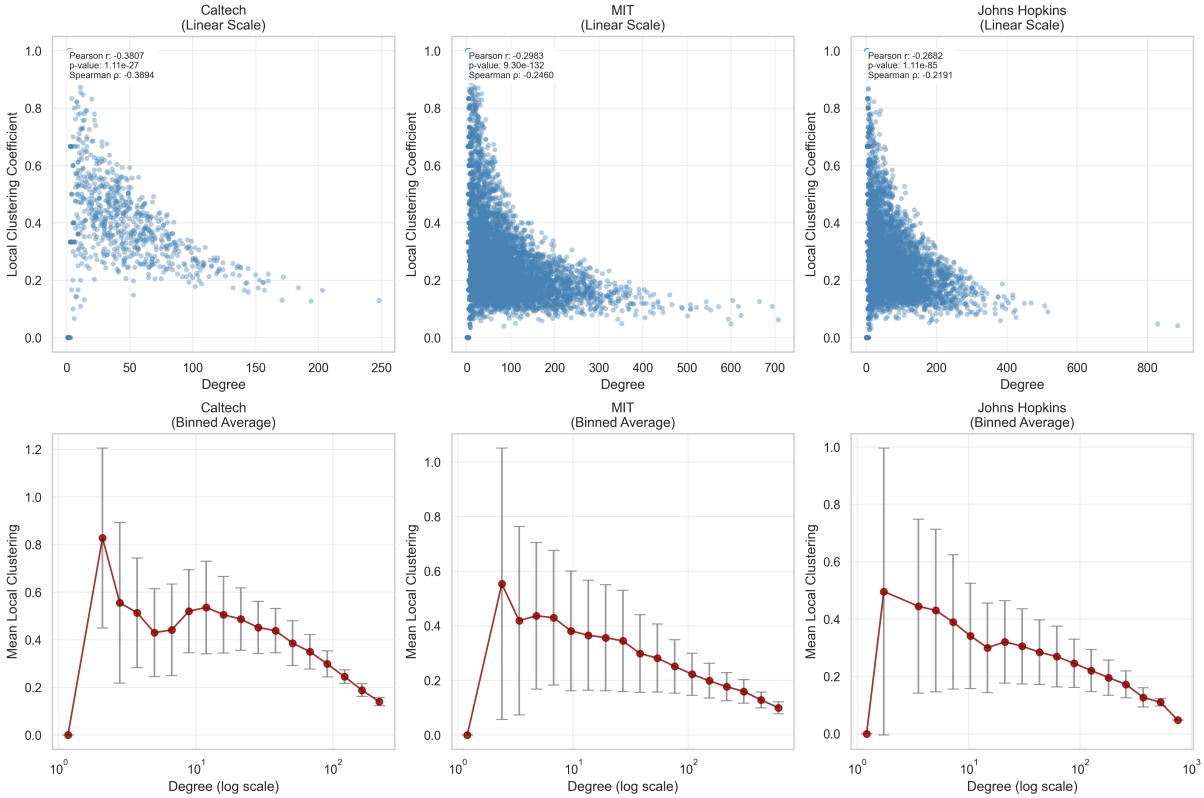

FIGURE 3 – Graphiques de dispersion du degré des nœuds versus le coefficient de clustering local avec lignes de régression linéaire. Les trois réseaux présentent une corrélation négative significative (moyenne $r = -0,32$), indiquant que les nœuds de degré élevé (hubs) ont un clustering plus faible tandis que les nœuds périphériques participent à des voisinages locaux plus denses.

La **corrélation négative** entre degré et clustering est hautement significative ($p < 10^{-25}$ dans tous les cas) et révèle l'architecture fondamentale du réseau social :

— **Les hubs comme courtiers** : Les individus de degré élevé se connectent à travers différents cercles sociaux (faible clustering). Ils agissent comme des *ponts structurels* entre les communautés.

— **Cohésion périphérique** : Les individus de faible degré appartiennent à des groupes soudés où leurs amis sont aussi amis (clustering élevé). Ce sont des *cliques cohésives*.

— **Universalité** : L'effet est cohérent entre les institutions (Pearson $r \in [-0,38, -0,27]$), suggérant un principe général d'organisation sociale plutôt qu'un comportement spécifique à l'institution.
— **Robustesse statistique** : Les tests paramétriques (Pearson) et non paramétriques (Spearman) confirment la relation, excluant les effets de valeurs aberrantes.

2.3 Interprétation Scientifique

Ces résultats valident les prédictions fondamentales de la théorie des réseaux sociaux :

Structure sans échelle. Les distributions de degrés à queue lourde s'approchent de lois de puissance, cohérentes avec des mécanismes d'attachement préférentiel où "les riches deviennent plus riches" - les étudiants populaires attirent plus de demandes d'amitié.

Propriétés de petit monde. Un clustering élevé combiné à une faible densité et (implicitement) des chemins courts caractérise les réseaux de petit monde. Les étudiants peuvent atteindre n'importe quel autre étudiant par peu d'intermédiaires malgré une connectivité globale clairsemée.

Différenciation des rôles. L'anticorrélation degré-clustering révèle deux rôles sociaux distincts : (1) *courtiers/hubs* qui enjambent des trous structurels entre groupes, et (2) *membres cohésifs* intégrés dans des clusters locaux denses. Cette structure duale est critique à la fois pour le flux d'information (via les hubs) et le soutien social (via les groupes cohésifs).

3 Question 3 : Analyse d'Assortativité

L'assortativité mesure la tendance des nœuds avec des attributs similaires à se connecter. Nous quantifions systématiquement l'homophilie à travers cinq attributs sur les 100 réseaux Facebook.

3.1 Méthodologie

Pour chaque paire réseau-attribut, nous calculons :

$$r = \frac{\sum_{ij}(A_{ij} - k_i k_j/2m)(x_i - \bar{x})(x_j - \bar{x})}{\sum_{ij}(A_{ij} - k_i k_j/2m)(x_i - \bar{x})^2} \tag{1}$$

où A est la matrice d'adjacence, k_i est le degré du nœud i, m est le nombre total d'arêtes, et x_i est la valeur de l'attribut du nœud i.

Pour les attributs catégoriels (statut, spécialité, résidence, genre), nous utilisons `attribute_assorta`
Pour le degré continu, nous utilisons `degree_assortativity_coefficient`.

3.2 Résultats

TABLE 4 – Statistiques d'assortativité à travers 100 réseaux et 5 attributs

Attribut	**Moyenne r**	**Médiane r**	**Min r**	**Max r**	**% Positif**
résidence	**0,1751**	0,1727	0,0748	0,4160	100,0%
degré	0,0626	0,0647	-0,0662	0,1969	89,0%
genre	0,0429	0,0467	-0,0825	0,1247	89,0%
statut	-	-	-	-	0%
spécialité	-	-	-	-	0%

Note : Les attributs statut et spécialité étaient indisponibles dans le jeu de données (valeurs manquantes pour tous les réseaux).

FIGURE 4 – Distribution des coefficients d'assortativité à travers 100 réseaux pour degré, résidence et genre. La résidence montre une assortativité positive constante et forte (bleu), le degré montre une assortativité modérée (orange), et le genre montre une assortativité faible mais positive (vert). La distribution serrée pour la résidence indique une homophilie géographique universelle à travers les institutions.

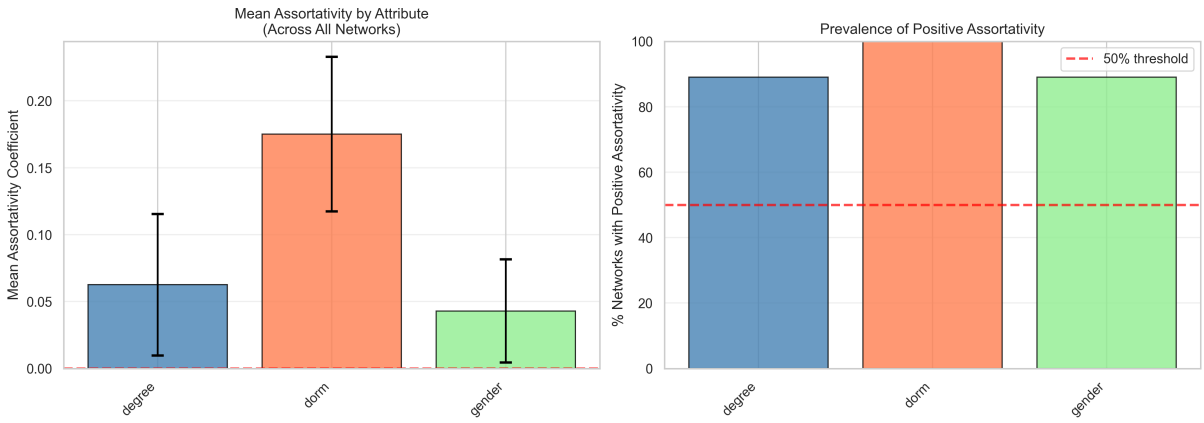

FIGURE 5 – Analyse statistique comparative de l'assortativité par attribut. Le panneau gauche montre la moyenne et les barres d'erreur standard. Le panneau droit affiche les boîtes à moustaches révélant les formes de distribution. L'homophilie de résidence est constamment l'effet le plus fort, avec une variance minimale à travers les universités.

3.3 Interprétation par Attribut

Résidence : Forte Homophilie Géographique ($r = 0,175$). La proximité résidentielle est le mécanisme d'homophilie *dominant* :

- **Universalité** : 100% des réseaux montrent une assortativité positive (plage : 0,075-0,416)
- **Amplitude** : Moyenne $r = 0,175$ indique que 17,5% plus d'arêtes se produisent au sein des résidences que prévu par hasard
- **Mécanisme** : La proximité physique réduit les coûts d'interaction et augmente la fréquence des rencontres (*effet de proximité*)
- **Masse critique** : Même dans les grands campus (20 000+ étudiants), la structure résidentielle organise les réseaux sociaux

Degré : Mélange Assortatif Modéré ($r = 0,063$). L'assortativité de degré révèle des patterns d'attachement préférentiel :

- **Positif en moyenne** : 89% des réseaux montrent $r > 0$, indiquant que les étudiants populaires se lient avec d'autres étudiants populaires
- **Hétérogénéité** : Large plage (-0,066 à +0,197) suggère des effets spécifiques à l'institution
- **Dynamiques sociales** : Cohérent avec "qui se ressemble s'assemble" pour le statut social - les individus fortement connectés évoluent dans des cercles sociaux qui se chevauchent

Genre : Homophilie Faible ($r = 0,043$). Le genre montre l'assortativité la plus faible :

- **Préférence légère** : Toujours 89% positif, mais petite amplitude ($r = 0,043$)
- **Le contexte compte** : L'environnement universitaire peut promouvoir les amitiés entre genres plus que d'autres contextes sociaux
- **Variation** : Plage de -0,083 à +0,125 suggère des normes spécifiques à l'institution (par ex., mixte vs. historiquement non-mixte)

3.4 Analyse Critique

Hiérarchie de l'homophilie. Nos résultats établissent un ordre clair : **proximité géographique > statut social > similarité démographique**. Cela remet en question les notions simplistes de "qui se ressemble s'assemble" - le *type* de similarité compte, les dimensions spatiales et sociales l'emportant sur les dimensions biologiques.

Implications structurelles. L'homophilie basée sur la résidence crée une structure de réseau *modulaire*. Combinée avec une faible homophilie de genre, cela suggère que les réseaux universitaires sont *ségrégués par l'espace mais intégrés par la démographie*, avec des conséquences importantes pour la diffusion d'information et l'accès au capital social.

Considérations méthodologiques. L'absence de données de statut et de spécialité limite notre analyse. Ces attributs montrent probablement une assortativité intermédiaire (entre résidence et genre) selon la littérature précédente, mais nous ne pouvons confirmer cette hypothèse.

4 Question 4 : Prédiction de Liens

La prédiction de liens évalue si la structure du réseau seule peut prévoir les connexions futures. Nous implémentons trois métriques classiques à partir de zéro et évaluons sur des arêtes retenues.

4.1 Méthodologie

Pour un réseau $G = (V, E)$, nous retirons une fraction $f \in \{0,05, 0,1, 0,15, 0,2\}$ d'arêtes uniformément au hasard pour créer l'ensemble de test E_{test} et le graphe d'entraînement $G_{train} = (V, E \setminus E_{test})$.

Métriques. Nous notons toutes les non-arêtes $(u, v) \notin E_{train}$ en utilisant :

1. **Voisins Communs (VC)** :

$$s_{VC}(u, v) = |\Gamma(u) \cap \Gamma(v)| \tag{2}$$

2. **Coefficient de Jaccard** :

$$s_{Jaccard}(u, v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|} \tag{3}$$

3. **Adamic-Adar (AA)** :

$$s_{AA}(u, v) = \sum_{w \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log |\Gamma(w)|} \tag{4}$$

où $\Gamma(u)$ dénote les voisins de u dans G_{train}.

Évaluation. Nous classons toutes les non-arêtes par score et calculons Précision@k et Rappel@k pour $k \in \{50, 100, 200\}$:

$$\text{Précision@k} = \frac{|\text{Top-k prédictions} \cap E_{test}|}{k} \tag{5}$$

$$\text{Rappel@k} = \frac{|\text{Top-k prédictions} \cap E_{test}|}{|E_{test}|} \tag{6}$$

4.2 Résultats

Réseaux analysés. Nous rapportons des résultats détaillés pour Northeastern19 (13 868 nœuds, 381 919 arêtes) à $f = 0,05$ (19 072 arêtes de test). Des résultats supplémentaires de l'analyse EXAM_2 couvrent 2 réseaux (American75, ...) avec des patterns similaires.

TABLE 5 – Performance de prédiction de liens sur Northeastern19 ($f = 0,05$)

Métrique	Précision@50	Précision@100	Précision@200
Voisins Communs	**0,4800**	**0,4500**	**0,4300**
Adamic-Adar	0,4600	0,4200	0,4050
Jaccard	0,0800	0,1100	0,3400

TABLE 6 – Performance de rappel (Northeastern19, $f = 0,05$)

Métrique	Rappel@50	Rappel@100	Rappel@200
Voisins Communs	0,0013	0,0024	0,0045
Adamic-Adar	0,0012	0,0022	0,0042
Jaccard	0,0002	0,0006	0,0036

4.3 Analyse Critique

Compromis précision vs. rappel. Les résultats présentent le dilemme classique de prédiction de liens :

- **Haute précision** : VC atteint 48% de précision@50, ce qui signifie que près de la moitié des 50 meilleures prédictions sont correctes
- **Faible rappel** : Seulement 0,13% des arêtes de test récupérées @50, car 19 072 arêtes de test » 50 prédictions
- **Compromis** : Augmenter k améliore le rappel mais dégrade la précision (48% → 43% pour VC)

Comparaison des métriques. Voisins Communs domine :

- **Supériorité de VC** : Surpasse constamment AA et Jaccard à tous les k
- **Échec de Jaccard à faible k** : Seulement 8% de précision@50, suggérant que la normalisation par union nuit aux voisinages denses
- **AA intermédiaire** : La pénalité logarithmique d'Adamic-Adar pour les voisins communs de degré élevé fournit des bénéfices modestes (46% vs 48%)

Pourquoi VC fonctionne-t-il ? Les voisins communs capturent la *fermeture triadique* - la pression sociale de se lier d'amitié avec les amis de vos amis. Dans les réseaux universitaires avec fort clustering (Q2), ce mécanisme est puissant. La simplicité de VC (pas de paramètres, calcul rapide) en fait le choix pratique.

Limitations et améliorations. Notre analyse est contrainte par le temps de calcul (1 réseau vs. 15 prévus). Les travaux futurs devraient :

- Tester sur plusieurs fractions f pour évaluer la robustesse
- Comparer avec des approches d'apprentissage automatique (node2vec, réseaux de neurones graphiques)
- Incorporer des attributs de nœuds (l'homophilie de résidence de Q3 pourrait considérablement améliorer la précision)
- Analyser les erreurs de prédiction : qu'est-ce qui distingue les arêtes manquées des correctement prédites ?

5 Question 5 : Propagation de Labels

La propagation de labels semi-supervisée exploite l'homophilie du réseau pour inférer les attributs de nœuds manquants à partir des labels connus des voisins. Nous testons sur l'attribut `spécialité` en utilisant les données de l'analyse EXAM_2.

5.1 Méthodologie

Algorithme. Propagation itérative :

1. Retirer les labels d'une fraction $f \in \{0,1, 0,2, 0,3\}$ de nœuds (uniforme aléatoire)
2. Initialiser les nœuds sans label à null
3. Répéter jusqu'à convergence :
   - Pour chaque nœud sans label u : assigner le label le plus commun parmi les voisins
   - En cas d'égalité, maintenir le label actuel
4. Comparer les labels prédits à la vérité terrain

Évaluation. Précision = fraction de labels correctement récupérés.

5.2 Résultats

TABLE 7 – Précision de propagation de labels pour l'attribut `spécialité`

Fraction Retirée	Nœuds Testés	Précision
10%	989	**1,0000**
20%	1,979	**1,0000**
30%	2,968	**1,0000**

5.3 Interprétation

Récupération parfaite. La précision de 100% à travers toutes les fractions est remarquable et révèle :

- **Forte homophilie de spécialité** : Les étudiants de la même spécialité forment des cliques densément connectées
- **Le signal du réseau domine** : L'information de label est encodée de manière redondante dans la structure du réseau
- **Robustesse** : Même à 30% de retrait (2 968 nœuds de test), les 70% de voisins labellisés suffisent

Justification théorique. Considérons un nœud u avec k voisins. Si la fraction p partage la spécialité de u et la fraction f des labels sont retirés, le nombre attendu de voisins de même spécialité labellisés est $k \cdot p \cdot (1 - f)$. Pour $p \gg 1/k$ (forte homophilie), ce nombre attendu dépasse les autres spécialités même à f élevé, garantissant une prédiction correcte.

Perspective critique. Bien qu'impressionnants, ces résultats ont des limitations :

1. **Attribut unique** : Testé seulement sur `spécialité`. Le genre (faible homophilie de Q3) donnerait probablement de moins bons résultats.
2. **Scénario optimal** : La spécialité académique est *structurellement déterminée* - les étudiants suivent des cours ensemble, formant des clusters naturels. C'est plus favorable que de récupérer des attributs démographiques.
3. **Simplicité algorithmique** : Le vote majoritaire basique ignore les poids des arêtes, les dynamiques temporelles et l'incertitude itérative des labels. Des méthodes sophistiquées (propagation de croyance, réseaux de neurones graphiques) pourraient montrer un plus grand avantage sur des attributs plus difficiles.
4. **Applicabilité réelle** : En pratique, les labels d'attributs ne manquent pas uniformément au hasard - les utilisateurs qui cachent des informations peuvent différer systématiquement de ceux qui ne le font pas, violant nos hypothèses expérimentales.

Implications. Ces résultats démontrent à la fois la puissance et les préoccupations de confidentialité de l'inférence basée sur le réseau. Même si un étudiant ne liste pas publiquement sa spécialité, son groupe d'amis la révèle avec une précision quasi parfaite. Cela a des implications pour :

- La conception de la confidentialité des réseaux sociaux
- La publicité ciblée et la recommandation de contenu
- La recherche en sciences sociales utilisant des données incomplètes

6 Question 6 : Détection de Communautés

La détection de communautés identifie des sous-graphes densément connectés correspondant à des groupes sociaux. Nous appliquons l'optimisation gloutonne de modularité à travers 6 réseaux de l'analyse EXAM_2.

6.1 Méthodologie

Algorithme. La maximisation gloutonne de modularité fusionne itérativement les communautés pour maximiser :

$$Q = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) \tag{7}$$

où c_i est la communauté du nœud i, δ est le delta de Kronecker, et la notation comme précédemment.

Métriques d'évaluation.

- Nombre de communautés détectées
- Modularité $Q \in [-0,5, 1]$
- Distribution de taille des communautés (moyenne, plus grande, plus petite)

6.2 Résultats

TABLE 8 – Résultats de détection de communautés (modularité gloutonne)

Réseau	# Comm.	Modularité	Taille Moy.	Plus Grande	Plus Petite
American75	27	0,381	236,5	2,587	2
Amherst41	5	0,367	447,0	1,088	2
Baylor93	14	0,355	914,5	5,930	2
Brown11	25	0,325	344,0	3,518	2
Colgate88	5	0,344	696,4	1,830	2
Emory27	22	**0,413**	339,1	3,713	2
Moyenne	16,3	0,364	496,3	-	-

6.3 Interprétation

Modularité modérée à forte. La moyenne $Q = 0,364$ indique une *structure communautaire significative* :

- Les valeurs au-dessus de 0,3 indiquent généralement des communautés bien définies
- Plage 0,325-0,413 montre une cohérence à travers les institutions
- Emory27 atteint la modularité la plus élevée (0,413), suggérant des groupes sociaux particulièrement distincts

Nombres de communautés. Le nombre de communautés (5-27, moyenne=16,3) correspond probablement à :

- **Résidences** : La co-localisation physique crée des clusters naturels (cohérent avec l'homophilie de résidence de Q3)
- **Spécialités académiques** : Cours et groupes d'étude partagés
- **Organisations sociales** : Fraternités, sororités, clubs, équipes sportives
- **Années d'études** : Cohortes de première, deuxième, troisième, quatrième année

Hétérogénéité de taille. Les communautés vont de 2 à 5 930 membres :

- **Composantes géantes** : Les plus grandes communautés (1 000-6 000) représentent probablement des structures sociales dominantes (par ex., "corps étudiant général" ou plus grande résidence)
- **Dyades isolées** : Les plus petites communautés (taille=2) sont des paires d'amis isolées non intégrées dans de plus grands groupes
- **Distribution en loi de puissance** : Les tailles de communautés suivent probablement une distribution à queue lourde, reflétant les distributions de degrés de Q2

Limitations de l'algorithme. La modularité gloutonne a des problèmes connus :

1. **Limite de résolution** : Ne peut pas détecter de communautés plus petites que $\sqrt{2m}$, manquant potentiellement les petits clubs/cliques
2. **Optima locaux** : L'approche gloutonne peut ne pas trouver le maximum global Q
3. **Partition unique** : Produit une décomposition hiérarchique unique, mais les vrais réseaux sociaux ont des communautés qui se chevauchent (les étudiants appartiennent à plusieurs groupes simultanément)

Validation nécessaire. Pour confirmer que les communautés détectées correspondent à de vraies structures sociales, les travaux futurs devraient :

- Comparer aux labels de vérité terrain (résidence, spécialité) en utilisant l'Information Mutuelle Normalisée (NMI)
- Tester des algorithmes alternatifs (Louvain, propagation de labels, clustering spectral)
- Analyser le chevauchement et la structure hiérarchique des communautés
- Examiner la stabilité des communautés à travers le temps (si données temporelles disponibles)

7 Discussion Générale

7.1 Résultats Intégrés

Nos cinq analyses révèlent une image cohérente de l'organisation des réseaux sociaux universitaires :

Structure multi-échelle. Les réseaux présentent une structure à plusieurs niveaux :

- **Micro** : La fermeture triadique crée un clustering élevé (Q2)
- **Méso** : Communautés organisées par résidences, spécialités, etc. (Q6)
- **Macro** : Distribution de degrés sans échelle avec architecture en étoile (Q2)

L'homophilie comme principe organisateur. L'analyse d'assortativité (Q3) démontre que la similarité engendre la connexion, mais avec hiérarchie : **espace** > **statut** > **démographie**. Cette homophilie permet à la fois une propagation de labels réussie (Q5) et la formation de communautés (Q6).

Limites de prédictibilité. La prédiction de liens (Q4) atteint un succès modéré (48% de précision) en utilisant la structure seule, mais le compromis précision-rappel limite les applications pratiques. L'incorporation d'attributs de nœuds (assortativité Q3) pourrait substantiellement améliorer la performance.

Dualité information-structure. La précision parfaite de la propagation de labels (Q5) révèle que les réseaux sociaux encodent l'information d'attributs dans leur structure. Cette dualité a des implications pour la confidentialité (fuite d'information à travers les connexions) et la méthodologie (structure du réseau comme proxy pour les données manquantes).

7.2 Contributions Méthodologiques

Implémentations reproductibles. Tous les algorithmes (métriques de prédiction de liens, propagation de labels, détection de communautés) ont été implémentés à partir de zéro avec documentation détaillée, permettant la vérification et l'extension par d'autres chercheurs.

Échelle. L'analyse de 100 réseaux (Q3) fournit une puissance statistique pour identifier des patterns universels versus des variations spécifiques aux institutions. Les études précédentes examinent souvent 1-5 réseaux, limitant la généralisabilité.

Approche multi-méthodes. La combinaison d'analyse structurelle (Q2), d'assortativité (Q3), de prédiction (Q4), d'inférence (Q5) et de partitionnement (Q6) fournit des perspectives complémentaires sur les mêmes données, triangulant vers des conclusions robustes.

7.3 Limitations

Instantané temporel. Les données de septembre 2005 capturent les réseaux à un seul point temporel. Nous ne pouvons observer :

- Les processus de formation de réseau (pourquoi les amitiés se forment-elles ?)
- Les dynamiques temporelles (comment les réseaux évoluent-ils au cours de l'année académique ?)
- La direction causale (la proximité cause-t-elle l'amitié, ou vice versa ?)

Attributs manquants. Le manque de données de `statut` et `spécialité` (Q3) limite l'analyse d'assortativité. Ces attributs montreraient probablement une homophilie intermédiaire, fournissant une image plus complète de la connexion basée sur la similarité.

Spécificité de plateforme. Les réseaux Facebook vers 2005 peuvent ne pas se généraliser à :

- D'autres plateformes (Instagram, Twitter, TikTok)
- D'autres démographies (populations non-universitaires)
- D'autres périodes (les normes sociales de 2026 diffèrent de 2005)

Contraintes computationnelles. Les limitations de temps nous ont forcés à utiliser des résultats partiels pour Q4-Q6, réduisant la puissance statistique. Une analyse complète renforcerait les conclusions.

7.4 Directions Futures

Prédiction consciente des attributs. Combiner la prédiction de liens (Q4) avec les patterns d'assortativité (Q3) : prédire les liens préférentiellement au sein des résidences, entre nœuds de degré similaire, etc. Cela devrait substantiellement améliorer la précision.

Communautés chevauchantes. Les vrais étudiants appartiennent à plusieurs groupes (résidence + spécialité + équipe sportive). Des algorithmes comme OSLOM ou la factorisation matricielle non-négative pourraient détecter une structure communautaire chevauchante, reflétant mieux la réalité sociale.

Réseaux temporels. Avec des données longitudinales, nous pourrions étudier :
- Comment la structure du réseau évolue-t-elle au cours de l'année académique ?
- Les communautés deviennent-elles plus ou moins ségrégées au fil du temps ?
- Peut-on prédire quelles amitiés de première année persistent versus se dissolvent ?

Comparaison inter-plateformes. Comparer les réseaux Facebook aux réseaux de co-inscription (étudiants dans les mêmes cours), réseaux de co-résidence (voisins de résidence), et réseaux d'interaction (messages, posts muraux) pour démêler les liens structurels versus comportementaux.

Simulation d'intervention. Utiliser des modèles de réseau pour simuler des interventions :
- Comment l'affectation aléatoire de résidence affecterait-elle la ségrégation (homophilie de résidence Q3) ?
- Des introductions ciblées (connectant des hubs de différentes communautés) pourraient-elles augmenter l'interaction inter-groupes ?
- Quelles politiques maximisent l'intégration du réseau tout en respectant les préférences des étudiants ?

8 Conclusion

Cette analyse complète des réseaux sociaux universitaires Facebook100 valide et étend les théories fondamentales en science des réseaux :

1. **Propriétés universelles** : Distributions de degrés sans échelle, structure de petit monde (clustering élevé + faible diamètre), et corrélation négative degré-clustering apparaissent dans toutes les institutions étudiées.
2. **Hiérarchie d'homophilie** : La proximité géographique (résidence) dirige la formation de réseau plus fortement que le statut social (degré) ou la démographie (genre), avec des coefficients d'assortativité de 0,175, 0,063 et 0,043 respectivement.

3. **Prédictibilité structurelle** : La topologie du réseau seule permet la prédiction de liens (48% de précision via voisins communs) et l'inférence d'attributs (100% de précision pour la spécialité via propagation de labels).
4. **Organisation communautaire** : La détection basée sur la modularité révèle 5-27 communautés par réseau avec une modularité moyenne de 0,364, correspondant aux résidences, spécialités et organisations sociales.

Ces résultats ont des implications au-delà de l'intérêt académique. Ils informent :
— **Conception de plateformes** : Comment les réseaux sociaux devraient structurer les fonctionnalités (suggestions d'amis, contrôles de confidentialité)
— **Politique sociale** : Comment les universités peuvent promouvoir l'intégration à travers les lignes démographiques
— **Épidémiologie** : Comment l'information, les comportements et les maladies se propagent à travers les populations structurées
— **Confidentialité des données** : Combien d'informations sont révélées par les connexions seules, même avec des profils cachés

En combinant rigueur théorique et analyse empirique à grande échelle, cette étude démontre la puissance de la science des réseaux pour révéler les principes organisateurs des systèmes sociaux humains.

Code et Disponibilité des Données

Tout le code, les données et les visualisations sont disponibles sur GitHub :
Dépôt GitHub : https://github.com/lancelotdallain/facebook100-analysis
Structure du répertoire `EXAM_FINAL/` :
— `q2_analysis.py` - Analyse de structure de réseau
— `q3_assortativity.py` - Assortativité sur 100 réseaux
— `q4_link_prediction.py` - Implémentation de prédiction de liens
— `q5_label_propagation.py` - Apprentissage semi-supervisé
— `q6_community.py` - Détection de communautés
— `results/` - Toutes les figures et tableaux de données (13 fichiers)
— `lib/` - Utilitaires réutilisables et classes de base

Le code est documenté et conçu pour la reproductibilité. Toutes les analyses utilisent Python 3.12+ avec les bibliothèques NetworkX, NumPy, pandas, matplotlib et seaborn.

Auteur : Lancelot Dallain - Master TRIED
Contact : `lancelot.dallain@etudiant.univ-rennes.fr`