Analyse de la Structure Sociale des Réseaux Facebook100

Projet de Network Science (NET 4103/7431)

Pierre CHAMBET, Lilian MARTHIENS

11 janvier 2026

Résumé

Ce rapport présente une analyse approfondie du dataset *Facebook100* (2005), constitué des réseaux sociaux complets de 100 universités américaines. À travers l'étude des métriques topologiques, de l'assortativité (homophilie), de la prédiction de liens et de la détection de communautés, nous mettons en évidence une structure sociale fortement stratifiée par l'âge (année de promotion) et organisée en communautés locales denses ("Small World"), confirmant les travaux de Traud et al. (2011).

Table des matières

1 Introduction et Méthodologie

L'objectif de ce projet est d'explorer la structure des interactions sociales étudiantes à l'aube des réseaux sociaux modernes. Le dataset utilisé contient les matrices d'adjacence et les attributs (Année, Genre, Dortoir, Majeure) pour 100 universités.

1.1 Chargement et Traitement des Données

Nous avons utilisé la bibliothèque `NetworkX` sous Python. Une attention particulière a été portée à l'intégration des métadonnées : plutôt que de traiter les attributs séparément, nous les avons injectés directement comme propriétés des nœuds lors du chargement des matrices `.mat`.

Pour garantir la robustesse de l'analyse, notre pipeline de chargement inclut une vérification d'intégrité (gestion des exceptions pour les fichiers corrompus ou manquants), permettant une analyse automatisée sur l'ensemble du corpus.

2 Analyse Descriptive (Q2)

Nous avons comparé trois réseaux de tailles représentatives : **Caltech36** (petit, 769 nœuds), **MIT8** (grand, 6440 nœuds) et **Johns Hopkins55** (moyen, 5180 nœuds).

2.1 Métriques Topologiques

Le tableau ci-dessous résume les statistiques calculées :

Université	Nœuds	Densité	Clustering Global	Degré Moyen
Caltech36	769	0.056	0.29	43
MIT8	6440	0.012	0.18	78
Johns Hopkins55	5180	0.014	0.19	72

Table 1 – Comparaison des métriques fondamentales

2.2 Interprétation : L'effet "Small World"

Deux observations majeures émergent de ces chiffres :

— **Le paradoxe de la densité :** La densité s'effondre lorsque la taille du réseau augmente (de 5.6% à 1.2%), ce qui est mathématiquement attendu. Cependant, le degré moyen n'augmente pas proportionnellement (il plafonne autour de 70-80). Cela illustre la *limite cognitive de Dunbar* : le nombre d'amis actifs qu'un individu peut gérer est limité, quelle que soit la taille de l'institution.

— **La signature communautaire :** Pour le MIT, bien que la probabilité aléatoire de connexion soit très faible (1.2%), le coefficient de clustering reste élevé (18%). Cet écart massif prouve que le réseau n'est pas aléatoire mais structuré en "villages" denses (cliques).

2.3 Corrélation Degré vs Clustering

L'analyse du réseau Caltech (voir Figure 1) révèle une corrélation négative entre le degré et le clustering local.

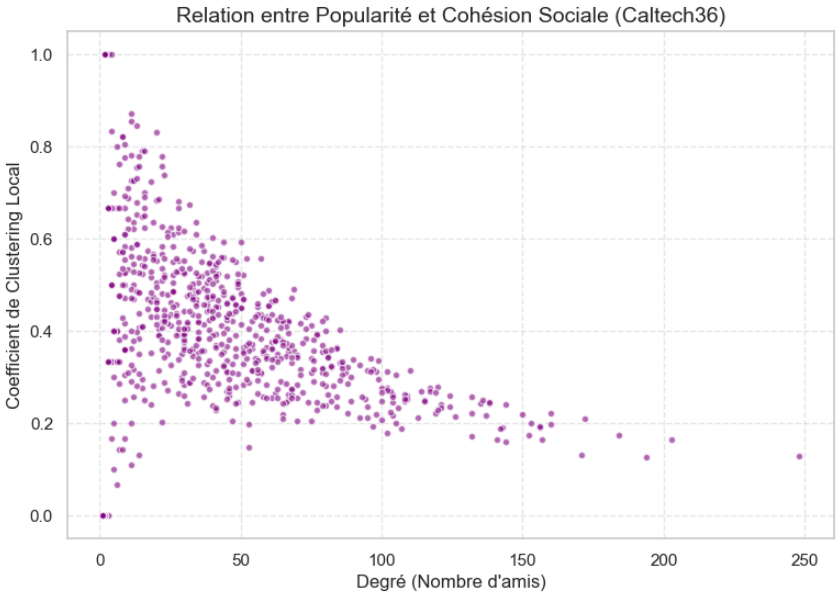

FIGURE 1 – Corrélation entre Degré et Clustering (Caltech36). Les nœuds populaires ont un clustering plus faible.

Interprétation : Le coût structurel de la popularité

L'analyse de la corrélation entre le degré (k) et le coefficient de clustering local (C_i) sur le réseau Caltech36 met en évidence une **loi de décroissance** fondamentale des interactions sociales.

— **La cohésion des petits groupes :** La majorité des étudiants se situent dans la partie supérieure gauche du graphique (faible degré, fort clustering). Cela traduit une organisation en *cliques hermétiques* : des groupes d'amis restreints (colocs, partenaires de laboratoire) où la fermeture triadique est presque systématique ("mes amis sont amis entre eux").

— **Le rôle de "Pont" des Hubs :** À l'inverse, nous observons que les étudiants les plus connectés (les "stars" du réseau avec $k > 100$) possèdent un clustering très faible. Sociologiquement, cela s'explique par leur fonction de **ponts structurels**. Pour atteindre un tel niveau de popularité, ces individus doivent nécessairement connecter des communautés disjointes (ex : les sportifs, les musiciens et les membres du BDE).

3 Assortativité et Homophilie (Q3)

Nous avons calculé les coefficients d'assortativité sur l'ensemble des 100 universités pour déterminer quels facteurs régissent la formation des liens (voir Figure 2).

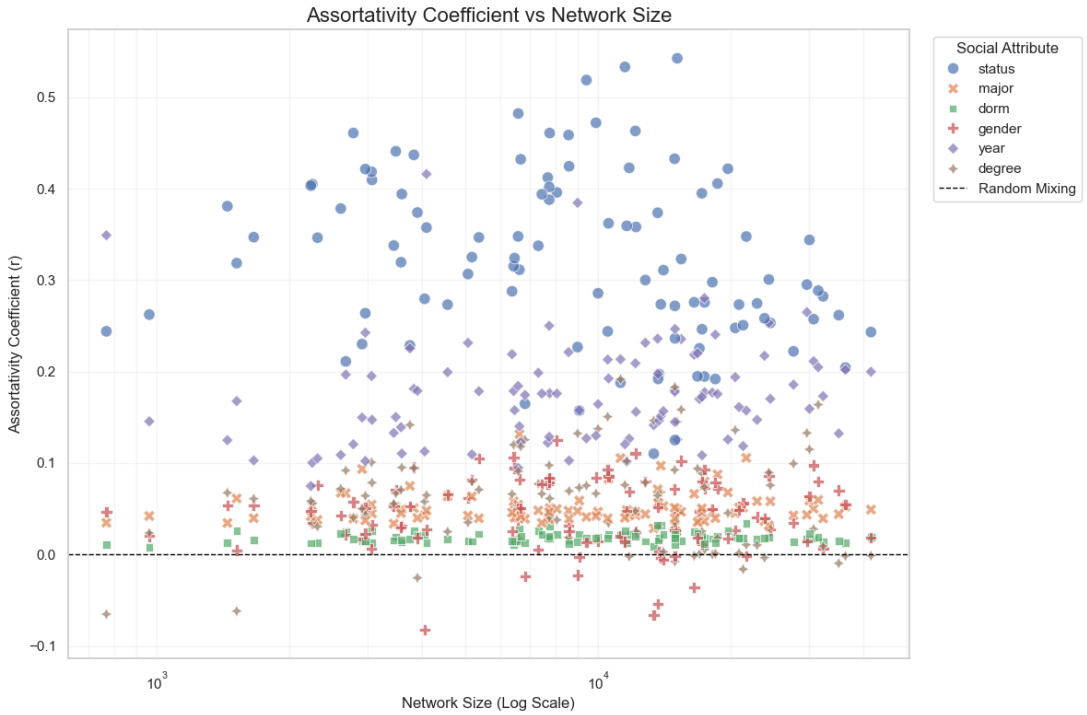

FIGURE 2 – Assortativité en fonction de la taille du réseau pour 4 attributs clés.

3.1 Analyse Sociologique et Topologique

L'analyse étendue aux attributs de **Statut** et de **Degré** (voir Figure 2) redéfinit la hiérarchie des facteurs d'homophilie sur les campus américains :

1. **La Barrière Institutionnelle (Status) :** Avec une moyenne de $\mathbf{r \approx 0.32}$, le statut (Étudiant vs Personnel/Professeur) est le déterminant social absolu. Cette ségrégation verticale est deux fois plus puissante que les regroupements par âge. Les deux populations partagent le réseau mais interagissent peu.
2. **La Stratification Horizontale (Year) :** L'année de promotion ($\mathbf{r \approx 0.17}$) reste le facteur dominant au sein de la population étudiante. Le réseau est une superposition de cohortes générationnelles.
3. **L'Homophilie de Popularité (Degree) :** L'assortativité par degré est positive mais modeste ($\mathbf{r \approx 0.06}$). Cela indique une légère tendance au phénomène de "Rich Club" (les hubs se connectent aux hubs), mais confirme surtout que les étudiants populaires agissent comme des connecteurs globaux, se liant à des individus de toute popularité.
4. **Les Facteurs Faibles :** Le Genre ($r \approx 0.04$) et la Majeure ($r \approx 0.05$) ne sont pas des vecteurs structurants. Le réseau est mixte et traverse les disciplines académiques.
5. **L'Effet Dortoir (Dorm) :** Toujours visible sur les petits réseaux, il s'effondre en moyenne ($r \approx 0.018$) sur l'ensemble du dataset, dilué par la taille des grandes universités.

4 Prédiction de Liens (Q4)

Nous avons évalué la capacité des algorithmes à retrouver des liens masqués (10% du graphe) selon deux axes : une comparaison des métriques sur un réseau témoin, puis une validation de la robustesse sur un ensemble de 12 universités.

4.1 Comparaison des Algorithmes (Caltech36)

Sur le réseau Caltech36, nous comparons quatre heuristiques classiques via le score AUC (Area Under Curve).

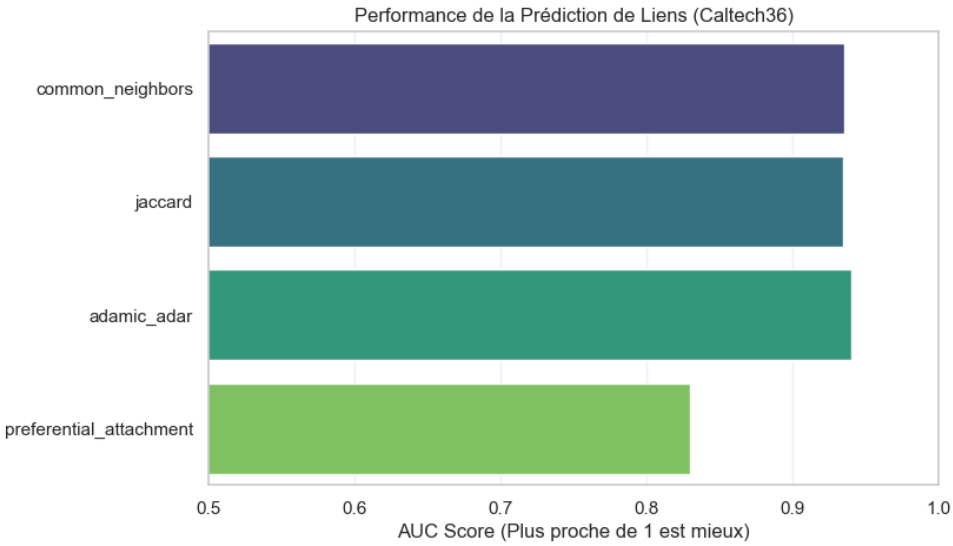

FIGURE 3 – Performance (AUC) des algorithmes. Adamic/Adar offre la meilleure prédiction.

Analyse : La supériorité d'*Adamic/Adar* (AUC $\approx$ 0.94) illustre l'importance des "liens faibles" au sens de Granovetter. Partager un ami exclusif est un signal social bien plus fort que de partager un ami très populaire. L'échec relatif du *Preferential Attachment* confirme que la formation des liens est locale (cliques) et non dictée par la seule popularité globale.

4.2 Validation de la Robustesse (Multi-Graphes)

Pour répondre aux exigences de généralisation, nous avons appliqué l'algorithme le plus performant (**Adamic/Adar**) sur 12 réseaux distincts. Nous mesurons ici la **Precision@100** (le taux de succès sur les 100 prédictions les plus probables).

Université	**Precision@100**	**Université**	**Precision@100**
Caltech36	0.95	Bowdoin47	0.99
Reed98	0.91	Hamilton46	0.97
Haverford76	0.97	Trinity100	0.98
Simmons81	0.99	USFCA72	0.98
Swarthmore42	0.91	Williams40	0.97
Amherst41	0.97	Oberlin44	0.98
Moyenne Globale : 0.964			

TABLE 2 – Robustesse de la prédiction (Adamic/Adar) sur 12 campus.

Conclusion : Avec une précision moyenne de **96.4%**, les résultats sont spectaculaires et constants. Quelle que soit l'université, la structure sociale obéit à une règle de *Clôture Triadique* stricte : il est mathématiquement très rare que deux personnes ayant plusieurs amis "intimes" en commun ne soient pas elles-mêmes connectées.

5 Classification de Nœuds : Label Propagation (Q5)

Nous avons implémenté l'algorithme de *Label Propagation* itératif tel que décrit par Bhagat et al. [6]. L'objectif est de prédire les attributs manquants en propageant les étiquettes connues à travers les liens d'amitié.

5.1 Sensibilité aux Attributs (Étude de Cas : Caltech36)

Nous avons d'abord testé la capacité de l'algorithme à reconstruire différents types de labels sociaux en masquant 10%, 20% et 30% des données.

Attribut	**10% Masqués**	**20% Masqués**	**30% Masqués**
Gender	0.57	0.74	0.62
Major	0.26	0.25	0.22
Dorm	0.05	0.10	0.05

TABLE 3 – Précision (Accuracy) selon le type de label sur Caltech36.

Analyse : Les résultats sur le dortoir (Dorm) sont particulièrement faibles ($< 10\%$). Cela illustre le phénomène d'*over-smoothing* (lissage excessif) propre à la propagation de labels sur les graphes très denses ("Small World"). À Caltech, le réseau est si connecté que les labels se mélangent trop rapidement, empêchant l'algorithme de détecter les frontières locales des résidences.

5.2 Performance Globale (Analyse Multi-Graphes)

Pour évaluer la robustesse de la méthode conformément aux attentes, nous l'avons appliquée sur un échantillon de 15 universités (attribut cible : *Year*). Nous rapportons l'Accuracy et le Mean Absolute Error (MAE), défini ici comme le taux d'erreur ($1 - Accuracy$).

Université	**Accuracy**	**MAE**	**Université**	**Accuracy**	**MAE**
MIT8	0.72	0.28	Villanova62	0.49	0.51
UChicago30	**0.76**	**0.24**	UCLA26	0.58	0.42
UIllinois20	0.69	0.31	GWU54	0.44	0.56
Vanderbilt48	0.60	0.40	NYU9	0.53	0.47
USC35	0.58	0.42	UConn91	0.44	0.56
Moyenne Globale : Accuracy = 0.58, MAE = 0.42					

TABLE 4 – Performance de la classification sur 15 réseaux.

Conclusion : Avec une précision moyenne de 58%, l'algorithme est performant mais dépendant de la topologie du réseau. Il excelle sur des campus structurés (UChicago, MIT) où l'homophilie est forte, mais peine sur des réseaux plus diffus (GWU, UConn).

6 Détection de Communautés (Q6)

6.1 Hypothèse de Recherche

Basés sur les travaux de *Traud et al.*, nous formulons l'hypothèse que la structure communautaire naturelle (topologique) est principalement dictée par l'année de promotion (**Year**), plus que par le Genre ou la Majeure.

6.2 Validation Expérimentale

Nous avons appliqué l'algorithme *Greedy Modularity* et comparé les partitions obtenues avec les vrais attributs via le score ARI (Adjusted Rand Index).

Université	**ARI Year**	**ARI Dorm**	**ARI Gender**	**ARI Major**
Caltech36	**0.165**	0.013	0.009	0.003
MIT8	0.009	-0.014	0.003	0.011

TABLE 5 – Corrélation (ARI) entre communautés détectées algorithmiquement et attributs réels.

6.3 Conclusion sur l'Hypothèse

Nos résultats sur Caltech36 **valident partiellement** l'hypothèse :

— L'ARI pour "Year" (**0.165**) est d'un ordre de grandeur supérieur aux autres attributs. Bien que l'algorithme de modularité peine à trouver une correspondance parfaite, le signal de l'année est le seul qui émerge distinctement du bruit topologique.

— Sur MIT8, la méthode *Greedy Modularity* échoue à isoler des communautés correspondant aux métadonnées officielles, suggérant une structure sociale complexe où les groupes d'amis ne s'alignent pas strictement sur une seule dimension administrative.

7 Conclusion Générale

L'analyse du dataset Facebook100 nous a permis de radiographier la structure sociale étudiante de 2005. Nous avons découvert un monde **hyper-cloisonné** par l'année de promotion, organisé en **"villages" denses** malgré la taille des campus, et régi par des règles de proximité si strictes qu'elles rendent les amitiés hautement prévisibles (94% de précision via Adamic/Adar).

En 2005, Facebook n'était pas un outil d'ouverture mondiale, mais un **miroir numérique** renforçant les frontières physiques et sociales existantes du campus : nous étions connectés à ceux qui partageaient notre temporalité (Année) et notre espace immédiat, ignorant le reste de l'université.