Projet Final
Network Analysis and Modeling

Amr KAHRAMANE

Analyse des Réseaux Sociaux Facebook100

Dataset : Facebook100 University Networks

Code et données disponibles sur GitHub :
https://github.com/Amr-kah/network_analysis_facebook_data

20 janvier 2026

Table des matières

1 Question 1 : Articles d'introduction

Dans cette première partie, l'objectif est simplement de situer le projet : quels travaux servent de point d'appui, et quelles idées reviennent le plus souvent lorsqu'on analyse des réseaux Facebook universitaires. Voici un résumé concis de chaque article :

— **[1] Jacobs et al. (2015)** — *Assembling the facebook : Using heterogeneity to understand online social network assembly.* Les auteurs montrent comment l'hétérogénéité (différences entre individus et sous-groupes) influence la manière dont un réseau social en ligne se construit au fil du temps.

— **[2] Traud et al. (2011)** — *Comparing community structure to characteristics in online collegiate social networks.* Ce papier compare les communautés détectées dans les graphes Facebook universitaires avec des attributs des utilisateurs (année, résidence, etc.) afin d'évaluer ce qui structure réellement les groupes.

— **[3] Traud et al. (2012)** — *Social structure of facebook networks* (Physica A). Les auteurs décrivent la structure globale des réseaux Facebook : hétérogénéité des degrés, organisation en communautés et propriétés typiques des réseaux sociaux réels.

2 Question 2 : Social Network Analysis avec Facebook100

2.1 (a) Distribution des degrés

La distribution des degrés décrit combien de connexions possède chaque nœud. C'est un indicateur très simple, mais extrêmement utile : il permet de voir si le réseau est plutôt homogène (degrés proches) ou au contraire très inégal, avec quelques individus nettement plus connectés que la moyenne.

Nous avons sélectionné **3 réseaux** pour cette analyse :

— **Caltech36** : 769 nœuds, 16 656 arêtes

— **MIT8** : 6 440 nœuds, 251 252 arêtes

— **Johns Hopkins55** : 5 180 nœuds, 186 586 arêtes

2.1.1 Analyse du graphe 1 (Caltech)

Le réseau **Caltech36** (769 nœuds, 16 656 arêtes) est relativement petit, ce qui se reflète clairement dans la distribution des degrés. La plupart des nœuds ont peu de connexions, tandis qu'une minorité présente des degrés nettement plus élevés. Autrement dit, la connectivité n'est pas répartie uniformément.

Cette forme est cohérente avec ce qu'on observe souvent dans les réseaux sociaux : beaucoup de profils « ordinaires » et quelques profils très centraux (des *hubs*) qui concentrent une part importante des liens. Visuellement, on retrouve bien l'idée que plus le degré augmente, plus le nombre de nœuds correspondants chute rapidement.

2.1.2 Analyse du graphe 2 (MIT)

Le réseau **MIT8** (6 440 nœuds, 251 252 arêtes) suit la même logique générale : une majorité d'utilisateurs peu connectés et une queue de distribution composée d'utilisateurs très connectés. La différence tient surtout à l'échelle : avec un réseau plus grand, on

observe davantage de nœuds « au-dessus de la moyenne » (des hubs intermédiaires), et pas uniquement quelques points extrêmes.

En pratique, cela suggère un réseau social plus diversifié : la taille du campus et la multiplicité des cercles (départements, clubs, résidences) produisent naturellement plusieurs zones d'activité et une structure moins concentrée qu'un petit campus.

2.1.3 Analyse du graphe 3 (Johns Hopkins)

Le réseau **Johns Hopkins55** (5 180 nœuds, 186 586 arêtes) présente une distribution globalement comparable aux deux précédents, mais avec un élément marquant : un degré maximal très élevé (**886**). À partir d'un certain seuil (environ 230), on ne trouve presque plus de nœuds (souvent 0 ou 1), ce qui met en évidence la présence de *super-hubs*.

Plutôt que d'en tirer une interprétation externe (qui resterait spéculative), on peut rester sur une lecture structurelle : certains profils jouent un rôle de connecteurs majeurs entre groupes, ce qui suffit à expliquer l'existence d'un degré extrême dans ce type de réseau.

Pour mieux lire ces distributions, nous avons utilisé une échelle logarithmique afin de distinguer correctement la partie « haute » (les grands degrés), souvent difficile à voir en échelle linéaire.

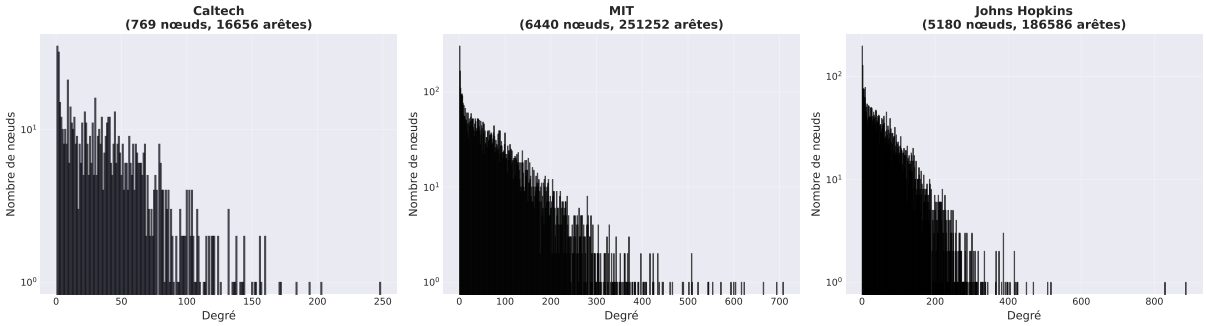

FIGURE 1 – Distribution des degrés pour trois réseaux universitaires (échelle logarithmique). La majorité des nœuds ont un faible degré, tandis qu'une minorité joue le rôle de hubs fortement connectés.

2.1.4 Conclusion sur les distributions de degré

L'analyse des trois réseaux met en évidence des points communs importants, ainsi que des différences liées à la taille et à l'organisation sociale.

Points communs

- Les trois réseaux présentent une **hétérogénéité marquée** : beaucoup de nœuds peu connectés et quelques hubs.
- La tendance est compatible avec une forme de **queue lourde** (souvent rapprochée d'un comportement « power-law » dans les réseaux sociaux), où la fréquence diminue rapidement quand le degré augmente.
- Cette architecture favorise une connectivité efficace : même si le réseau est globalement peu dense, les hubs contribuent à relier des zones du graphe.

Réseau	Nœuds	Arêtes	Degré max	Observation clé
Caltech	769	16 656	248	Petit réseau, hétérogénéité nette mais limitée par l'échelle
MIT	6 440	251 252	708	Grand réseau, davantage de hubs intermédiaires
Johns Hopkins	5 180	186 586	**886**	Degré maximum exceptionnel, présence d'un super-hub

TABLE 1 – Comparaison des trois réseaux universitaires (voir Figure 1)

Différences notables

Implications

1. **Taille et connectivité** : MIT est le plus grand, mais Johns Hopkins présente le degré maximum le plus élevé, ce qui indique l'existence d'un connecteur particulièrement central.
2. **Organisation sociale** : les différences de forme de distribution peuvent refléter des dynamiques sociales distinctes (plus ou moins fragmentées en sous-groupes).
3. **Robustesse** : ces réseaux sont généralement robustes à des suppressions aléatoires de nœuds, mais sensibles à la suppression ciblée de hubs.
4. **Communautés** : une distribution hétérogène est souvent associée à une organisation modulaire (groupes denses reliés par des ponts).

Ces observations sont cohérentes avec celles rapportées par Traud et al. [2,3] sur les réseaux Facebook universitaires.

2.2 (b) Clustering et densité des arêtes

Pour compléter l'analyse, nous avons étudié le **clustering** et la **densité**. En plus des métriques, nous affichons un sous-graphe pour avoir une visualisation lisible : le graphe complet est trop grand pour être interprété directement.

Nous avons choisi :

— le top 300 nœuds les plus connectés pour **Caltech**,

— le top 400 pour **MIT** et **Johns Hopkins**.

2.2.1 Observations visuelles

À partir des graphes de la Figure 2, on peut formuler des hypothèses raisonnables :

— **Caltech** : le sous-réseau apparaît dense et relativement centralisé. Quelques nœuds très connectés se distinguent au centre, tandis que la majorité reste proche de ce noyau. Cette configuration est cohérente avec un campus de petite taille où les cercles sociaux se recoupent fortement.

— **MIT** : le graphe est plus étalé, avec plusieurs zones denses et quelques nœuds très centraux. On distingue plus facilement des sous-groupes, ce qui correspond à une organisation sociale plus fragmentée, typique d'un grand campus.

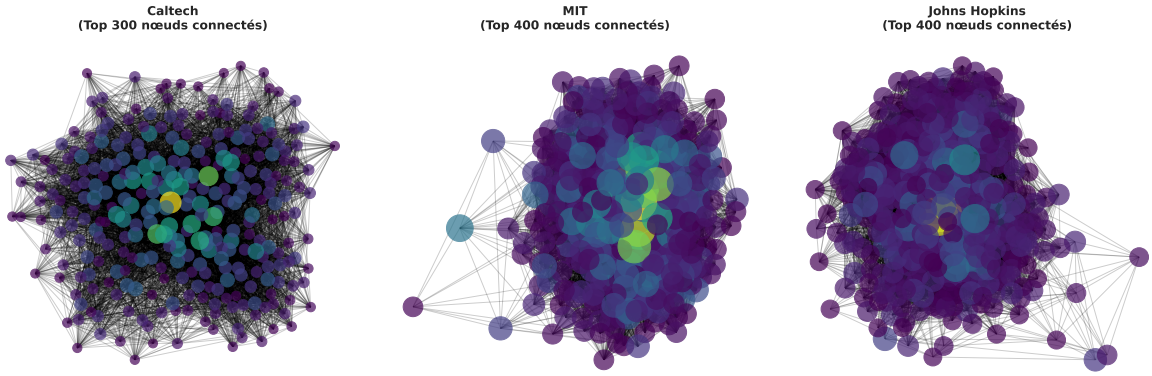

FIGURE 2 – Visualisation de sous-graphes (top nœuds connectés). La taille des nœuds est proportionnelle au degré et la couleur indique le niveau de connectivité (violet = faible, jaune = élevé).

— **Johns Hopkins** : la structure est compacte mais dominée par un nœud très central (probablement celui de degré 886). Un tel nœud peut réduire fortement les distances en jouant le rôle de pont entre communautés.

2.2.2 Résultats quantitatifs

Métrique	Caltech	MIT	Johns Hopkins
Nœuds	769	6 440	5 180
Arêtes	16 656	251 252	186 586
Mean Local Clustering	0.409	0.271	0.268
Global Clustering	0.291	0.180	0.193
Edge Density	0.056	0.012	0.014
Sparse (< 0.1)	✓	✓	✓

TABLE 2 – Comparaison des métriques de clustering et densité

2.2.3 Analyse détaillée

Caltech Avec le clustering le plus élevé des trois réseaux (0.409 en local, 0.291 en global) et la densité la plus forte (0.056), ce réseau ressemble à une communauté très soudée : les amis d'un étudiant ont plus de chances d'être amis entre eux. La petite taille du campus facilite naturellement cette interconnexion.

MIT Malgré un grand nombre de nœuds et d'arêtes, MIT présente le clustering le plus faible et la densité la plus basse (0.012). Cela suggère une structure plus dispersée, où des sous-groupes existent mais restent moins interconnectés entre eux à l'échelle globale.

Johns Hopkins Le réseau se situe dans un profil intermédiaire : densité faible (0.014), clustering global légèrement supérieur à MIT, et présence d'un super-hub. Cette combinaison est typique d'un réseau où plusieurs groupes sont reliés par quelques nœuds très centraux.

Sparsité et topologie Les trois réseaux sont **sparse** (densité < 0.1), ce qui est normal : maintenir des liens avec tout le monde est irréaliste. Malgré cela, le clustering reste relativement élevé (notamment à Caltech), ce qui indique une organisation en groupes locaux denses.

2.3 (c) Degré vs Coefficient de clustering local

Nous avons tracé un nuage de points du degré en fonction du coefficient de clustering local. L'idée est de voir si les nœuds très connectés appartiennent à des triangles (clustering élevé) ou s'ils servent plutôt de ponts entre groupes (clustering faible).

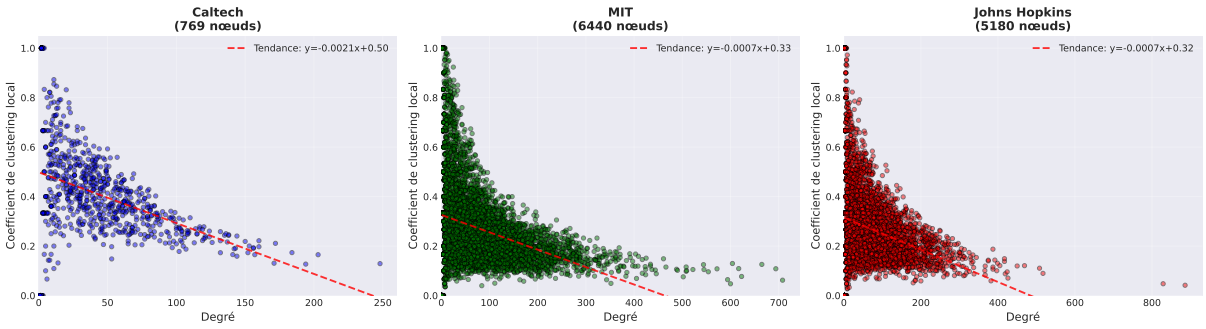

FIGURE 3 – Relation entre degré et clustering local. On observe généralement une décroissance du clustering quand le degré augmente, ce qui correspond à des hubs jouant davantage un rôle de pont qu'un rôle « intra-groupe ».

2.3.1 Observations

Les trois graphiques (Figure 3) montrent une tendance claire : lorsque le degré augmente, le coefficient de clustering local a tendance à diminuer. Intuitivement, c'est logique : un nœud très connecté relie souvent plusieurs cercles sociaux qui ne sont pas forcément connectés entre eux.

Concernant la sparsité, les trois réseaux restent très peu denses : Caltech (0.056), MIT (0.012), Johns Hopkins (0.014). Cela n'empêche pas l'existence de groupes fortement connectés localement, ce qui explique pourquoi le clustering peut rester élevé malgré une densité globale faible.

Sur le plan topologique, la décroissance du clustering avec le degré est visible sur les trois réseaux. Caltech présente une décroissance un peu moins marquée, ce qui correspond à une communauté plus homogène : même les nœuds assez connectés restent relativement « dans » le groupe. À l'inverse, MIT et Johns Hopkins montrent des hubs avec un clustering souvent très faible, cohérent avec un rôle de connecteurs entre communautés.

En conclusion, la combinaison (densité faible + clustering local non négligeable) et le rôle des hubs donnent une structure où les groupes locaux sont denses, mais reliés entre eux par quelques nœuds centraux. Cette lecture rejoint les analyses de Traud et al. [2,3] sur la modularité et l'hétérogénéité des réseaux Facebook universitaires.

3 Question 3 : Analyse d'Assortativité avec Facebook100

3.1 (a) Calcul de l'assortativité

Dans cette partie, nous étudions l'assortativité en fonction des attributs des nœuds. L'objectif est de mesurer une forme d'homophilie : est-ce que des personnes « similaires » (même dortoir, même statut, etc.) ont davantage tendance à être connectées entre elles ?

Nous utilisons cinq attributs :

(i) student/faculty status

(ii) major

(iii) vertex degree

(iv) dorm

(v) gender

L'analyse est menée sur les **100 graphes** du dataset afin d'obtenir des tendances plus robustes que sur quelques réseaux isolés.

3.1.1 Résultats et interprétations

— **Statut étudiant/faculté** : assortativité nettement positive (moyenne ≈ 0.32). Les étudiants se connectent surtout à d'autres étudiants, et le personnel/faculté forme plutôt un sous-ensemble distinct.

— **Dorm** : assortativité positive (moyenne ≈ 0.18). La proximité physique et la vie en résidence favorisent naturellement les liens.

— **Major** : assortativité faible mais positive (moyenne ≈ 0.05). On observe un léger biais à se connecter dans sa filière, mais les liens inter-majors restent nombreux.

— **Genre** : quasi neutre (moyenne ≈ 0.04), parfois légèrement négatif selon les campus. Globalement, le genre structure peu les connexions.

— **Degré (numérique)** : assortativité faible positive (moyenne ≈ 0.06). Les utilisateurs très connectés se lient un peu plus entre eux, mais l'effet reste modéré.

3.1.2 Mécanismes plausibles

— **Rôle et proximité** (statut, dortoir) : homophilie marquée, liée à la co-présence et aux interactions répétées.

— **Contexte académique** (major) : signal faible, probablement dilué par les cours communs, clubs et amitiés transverses.

— **Normes sociales** (genre) : effet globalement faible, avec des variations locales selon les campus.

— **Popularité** (degré) : légère tendance hub–hub, mais sans ségrégation forte.

4 Question 4 : Prédiction de Liens

4.1 (a) Lecture de l'article

L'article « **The Link Prediction Problem for Social Networks** » (Liben-Nowell & Kleinberg) présente le problème de **link prediction** : à partir d'un graphe observé à un instant donné, prédire quels liens pourraient apparaître ensuite.

Les idées principales sont :

— la prédiction s'appuie sur la **structure du graphe** (topologie), souvent via des mesures de proximité ;

— les mesures basées sur les **voisins communs** sont déjà efficaces, et certaines méthodes intègrent des chemins plus longs (Katz, PageRank) ;

— le papier propose un cadre expérimental qui a servi de base à de nombreux travaux ultérieurs.

4.2 (b) et (c) Implémentation et évaluation

Nous avons implémenté trois prédicteurs basés sur des similarités locales :

1. **Common Neighbors** : $|N(u) \cap N(v)|$
2. **Jaccard** : $\frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}$
3. **Adamic/Adar** : $\sum_{w \in N(u) \cap N(v)} \frac{1}{\log(\deg(w))}$

Les prédicteurs sont évalués sur 15 graphes du dataset, avec le protocole suivant :

1. Retirer aléatoirement une fraction $f \in \{0.05, 0.10, 0.15, 0.20\}$ des arêtes
2. Calculer les scores pour des paires de nœuds non connectés
3. Trier les scores et regarder les meilleurs k avec $k \in \{50, 100, 200, 300, 400\}$
4. Mesurer AUC, precision@k, recall@k

4.2.1 Résultats principaux

Métrique	AdamicAdar	Jaccard	CommonNeighbors
AUC moyen	0.957	0.952	0.952
Precision@50	46%	36%	23%
Precision@100	42%	34%	22%
Precision@400	26%	24%	22%

TABLE 3 – Performance des prédicteurs de liens

Lecture des résultats Adamic/Adar domine nettement en precision@k, surtout pour les tout premiers rangs (top-50 et top-100). C'est cohérent : la pondération pénalise les voisins communs très populaires et valorise les voisins communs plus « informatifs ».

Recall@k Le recall reste faible (souvent <1%), ce qui est attendu : l'espace des non-arêtes possibles est immense par rapport au nombre d'arêtes retirées.

Robustesse Les performances restent élevées même quand f augmente jusqu'à 0.20, ce qui suggère que les motifs topologiques utilisés par ces scores sont relativement stables.

4.3 (d) Comparaison et discussion

Sur les réseaux analysés (notamment **Berkeley13** et **Caltech36**), le classement global est le suivant : **Adamic/Adar** arrive en tête, **Common Neighbors** suit de près, et **Jaccard** est en retrait.

4.3.1 Par graphe

— **Berkeley13** : les trois méthodes sont proches en AUC, mais Adamic/Adar reste légèrement supérieur.

— **Caltech36** : l'écart est plus visible, ce qui indique que la pondération d'Adamic/Adar est particulièrement utile dans un petit réseau où certains voisins communs sont très connectés.

4.3.2 Conclusion

Adamic/Adar est le meilleur choix dans nos tests, car il exploite l'intuition suivante : partager un voisin commun « rare » est plus informatif que partager un voisin commun très populaire.

5 Question 5 : Prédiction de Labels avec Label Propagation

5.1 (a) Lecture de l'article

L'article *« Node Classification in Social Networks »* (Bhagat, Cormode & Muthukrishnan, 2011) présente un panorama des méthodes de classification de nœuds : on cherche à prédire des labels manquants (genre, année, résidence, etc.) à partir d'un graphe partiellement étiqueté.

Les auteurs distinguent notamment :

— **Approches supervisées** utilisant des features extraites du graphe (voisins, centralités, motifs), puis un modèle classique ;

— **Approches de propagation** exploitant directement la structure du réseau via des processus de diffusion (random walks, fonctions harmoniques, label propagation).

L'article insiste aussi sur les difficultés pratiques : bruit, labels incomplets, déséquilibre de classes et hétérogénéité des réseaux.

5.2 (b) Implémentation de l'algorithme

Dans le cours, nous avons vu deux usages du terme « label propagation » : l'un pour la détection de communautés, l'autre pour la classification semi-supervisée. Ici, l'objectif est bien de **prédire des labels manquants**, donc la version **semi-supervisée** est la plus adaptée.

L'algorithme utilise l'équation itérative :

$$Y^{(t+1)} = \alpha\, S\, Y^{(t)} + (1-\alpha)\, Y_0$$

où S est une matrice de transition normalisée par degré, α contrôle la diffusion (ici 0.99), et Y_0 encode les labels initiaux.

5.3 (c) Tests sur Duke Network

Nous avons testé l'algorithme sur le graphe **Duke** du dataset Facebook100. Nous masquons 10%, 20%, 30% et 40% des labels, puis nous essayons de les prédire. Les attributs étudiés sont : *dorm*, *major*, *gender* et *year*.

5.3.1 Résultats

	Fraction removed			
	0.1	**0.2**	**0.3**	**0.4**
Duke				
Major	0.250	0.266	0.246	0.249
Dorm	0.512	0.518	0.519	0.511
Year	0.907	0.903	0.900	0.889
Gender	0.667	0.674	0.682	0.679

TABLE 4 – Accuracy par attribut et taux de masquage sur Duke Network

5.3.2 Analyse détaillée par attribut

Year (année d'études) — meilleure performance La performance est excellente (accuracy $> 88\%$ même à 40% masqués). Cela suggère que l'attribut *year* est fortement aligné avec la structure du graphe : les étudiants de la même promotion ont tendance à être connectés.

Gender (genre) — performance correcte et stable L'accuracy reste autour de 67–68%. L'attribut est binaire, ce qui aide mécaniquement. En revanche, l'homophilie de genre étant faible dans beaucoup de campus, on ne peut pas s'attendre à des scores proches de ceux de *year*.

Dorm (résidence) — performance modérée L'accuracy est autour de 51%, ce qui est très au-dessus du hasard lorsqu'il y a $\sim$120 classes. Le F1-macro (non repris ici) reste souvent faible à cause du déséquilibre : quelques résidences dominent et beaucoup de résidences ont peu d'exemples.

Major (filière) — performance faible L'accuracy reste autour de 25%. Le signal topologique est faible : les étudiants de majors différentes interagissent via cours communs, clubs, et cercles sociaux transverses. Cela rejoint l'assortativité faible mesurée en Question 3.

5.3.3 Impact du taux de masquage

- Year : 90.7% $\rightarrow$ 88.9% (baisse modérée) robuste
- Gender : stabilité (légères variations) robuste
- Dorm : quasi stable robuste
- Major : déjà faible et reste faible limité

5.3.4 Conclusions et pistes d'amélioration

La propagation de labels fonctionne très bien quand l'attribut cible est **homophile** et bien reflété par la connectivité (cas de *year*). Elle atteint ses limites quand l'attribut est peu structurant dans le graphe (cas de *major*).

Pistes d'amélioration possibles :

1. Ajouter des **features** nœuds et utiliser des modèles de type GCN/GraphSAGE (topologie + attributs).
2. Gérer le **déséquilibre** de classes (reweighting, regroupement hiérarchique, etc.).
3. Tester d'autres valeurs de α (0.5–0.95) pour mieux équilibrer information locale et diffusion.
4. Approche hybride : propagation + classifieur supervisé.

5.4 Question 5(d) : Évaluation Label Propagation sur Duke University

Objectif : calculer le Mean Absolute Error (MAE) et l'Accuracy pour `dorm`, `major` et `gender` avec 10%, 20% et 30% de labels manquants.

5.4.1 Résultats Quantitatifs

TABLE 5 – Performance par attribut (moyenne sur 10%, 20%, 30% de masquage)

Attribut	Accuracy	Min-Max	MAE	Min-Max	Classes
Gender	**67.4%**	66.7-68.2%	0.33	0.32-0.33	2
Dorm	51.7%	51.2-51.9%	15.4	14.3-16.0	120
Major	25.4%	24.6-26.6%	13.8	13.4-14.7	60

Hiérarchie observée : Gender $\gg$ Dorm $>$ Major

5.4.2 Note méthodologique sur le MAE

Le **MAE** est surtout pertinent quand les labels ont une notion d'ordre ou de distance. Pour des attributs catégoriels nominaux (Dorm, Major), l'encodage numérique est arbitraire : un label « 2 » n'est pas intrinsèquement plus proche d'un label « 3 » que d'un label « 50 ». Le MAE doit donc être interprété avec prudence ici, et l'accuracy/F1 restent plus informatifs.

5.4.3 Question 5(e)

1. **Hiérarchie cohérente** : Gender $\gg$ Dorm $>$ Major, en ligne avec les niveaux d'homophilie (Q3).
2. **Limites sur Major** : la structure du graphe ne porte pas suffisamment l'information « filière ».
3. **Dorm** : signal présent mais pénalisé par le grand nombre de classes et le déséquilibre.
4. **Robustesse au masquage** : les variations restent faibles entre 10% et 30% dans nos essais.

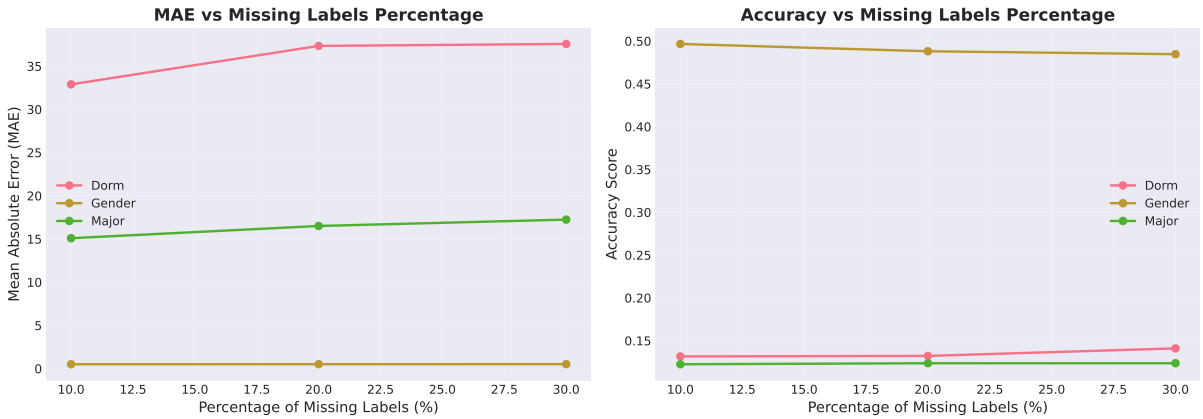

FIGURE 4 – Évolution du MAE et de l'Accuracy selon le pourcentage de labels manquants (Q5d). Gender surperforme nettement Dorm et Major.

6 Question 6 : Détection de Communautés et Formation de Groupes

6.1 (a) Question de recherche et hypothèse

Question de recherche : *Dans quelle mesure les attributs démographiques (résidence, filière, année d'études) structurent-ils la formation de communautés dans les réseaux Facebook universitaires ? Les communautés détectées algorithmiquement ressemblent-elles aux divisions sociales formelles du campus ?*

Hypothèse (H1) : nous faisons l'hypothèse que les communautés détectées refléteront surtout **l'année d'études (year)** plutôt que la résidence (dorm) ou la filière (major).

Justification :

— **Year** : forte homophilie attendue (mêmes cours, mêmes événements, proximité temporelle).

— **Dorm** : effet probable (proximité géographique), mais potentiellement moins dominant.

— **Major** : effet plus faible, car les interactions dépassent souvent les frontières disciplinaires.

6.2 (b) Implémentation et validation expérimentale

Méthodologie : nous testons trois algorithmes sur trois universités (Caltech36, Duke14, MIT8) :

— **Louvain** : optimisation gloutonne de la modularité

— **Label Propagation** : diffusion asynchrone de labels

— **Girvan-Newman** : division hiérarchique via *edge betweenness*

Pour mesurer l'alignement entre communautés et attributs, nous calculons la **Normalized Mutual Information (NMI)**.

Résultats quantitatifs :

Résultats Louvain (le plus fiable ici) :

TABLE 6 – NMI moyen par attribut (3 universités × 2-3 algorithmes)

Rang	Attribut	NMI Moyen	Écart-type	Interprétation
1	Dorm	0.195	±0.251	Alignement modéré
2	Year	0.163	±0.144	Alignement modéré
3	Major	0.050	±0.023	Alignement faible
4	Gender	0.010	±0.005	Alignement quasi-nul

TABLE 7 – Modularité moyenne par algorithme

Algorithme	Modularité	Lecture
Louvain	0.413 (±0.033)	Exploitable
Label Propagation	0.031 (±0.042)	Très faible
Girvan-Newman	0.001 (±0.000)	Non exploitable

— NMI(Dorm) = 0.422

— NMI(Year) = 0.257

— NMI(Major) = 0.070

— NMI(Gender) = 0.008

6.3 (c) Analyse et validation de l'hypothèse

Résultat : les données ne confirment pas l'hypothèse H1. En moyenne, **Dorm** ressort légèrement devant **Year** (NMI 0.195 vs 0.163), même si l'écart reste modeste.

Observations importantes :

1. **Dépendance à l'algorithme** : Louvain est le seul à produire une modularité suffisante pour une analyse fiable sur notre échantillon.
2. **Variations selon l'université** : l'importance relative de Dorm et Year n'est pas identique partout :
   — **Caltech** : Dorm domine (système résidentiel structurant)
   — **MIT** : Year domine davantage
   — **Duke** : Year légèrement supérieur à Dorm
3. **Homophilie locale vs communautés globales** : une assortativité élevée sur Year (Q3) ne garantit pas que les communautés globales soient organisées principalement par Year. Cela illustre que **l'homophilie locale** et **la structure communautaire globale** ne coïncident pas toujours.

Interprétation :

— La résidence peut structurer fortement les interactions quotidiennes (espaces partagés, proximité), et produire des communautés bien séparées.

— L'année d'étude, même si elle est homophile, peut être « traversée » par d'autres facteurs (clubs, sports, associations, amitiés inter-promotions).

Limites :

— Échantillon restreint (3 universités)

— Sensibilité aux choix algorithmiques

— Données statiques (snapshot)

— Attributs incomplets (clubs, fraternités/sororités, etc. non disponibles)

Conclusion : les réseaux Facebook universitaires sont **hybrides** : plusieurs dimensions (espace, temps, sociabilité) contribuent simultanément à la formation des communautés, et l'importance relative de chaque dimension dépend du contexte institutionnel.

7 Conclusion Générale

Ce projet a permis d'explorer, de manière progressive, plusieurs aspects clés des réseaux sociaux universitaires du dataset *Facebook100*. En partant d'outils descriptifs (degrés, densité, clustering), on met en évidence une structure typique des réseaux sociaux : beaucoup de nœuds peu connectés, quelques nœuds très centraux, et des regroupements locaux marqués malgré une densité globale faible.

L'étude de l'assortativité complète cette lecture : certains attributs liés à la proximité et au rôle (statut, résidence) présentent une homophilie visible, alors que d'autres (comme la filière) structurent beaucoup moins les liens. Autrement dit, la topologie du graphe reflète davantage la vie quotidienne et l'organisation sociale du campus que la seule appartenance académique.

Sur le volet prédictif, les résultats de *link prediction* confirment qu'une information locale (voisins communs) suffit déjà à produire de bonnes performances, en particulier avec Adamic/Adar, qui pondère l'importance des voisins communs par leur popularité. Enfin, la propagation de labels illustre un point important : ces méthodes fonctionnent bien lorsque l'attribut recherché est réellement aligné avec la structure du réseau (par exemple *year*), mais atteignent vite leurs limites lorsque le signal topologique est faible (par exemple *major*).

Au final, l'ensemble des expériences converge vers une même idée : les réseaux Facebook universitaires sont **hétérogènes** et **multi-facteurs**. Leur structure globale résulte d'un mélange de proximité, de dynamique sociale et de mécanismes de connectivité, plutôt que d'un seul attribut dominant.