

# Projet NSGL - Network Science and Graph Learning

Tanguy CESAR

lien github : <https://github.com/tanguycesar/NSGL-Projet>

## 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.

## 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.

### 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.

**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 tendance à la fermeture des triades (les amis de mes amis sont mes amis). L'analyse du clustering local en fonction du degré montre une relation décroissante : les noeuds faiblement connectés appartiennent souvent à des groupes locaux denses, tandis que les noeuds de fort degré ont un coefficient de clustering plus faible.

**Assortativité positive :** L'assortativité par degré est globalement positive (moyenne  $\approx 0,06$ ). Cela suggère une légère tendance des noeuds très connectés à être reliés entre eux.

### Synthèse

Les analyses confirment que les graphes étudiés présentent les propriétés statistiques classiques attendues pour des réseaux sociaux de cette taille : une densité faible, un clustering local fort et une distribution de degrés à queue lourde.

## 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.

### 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

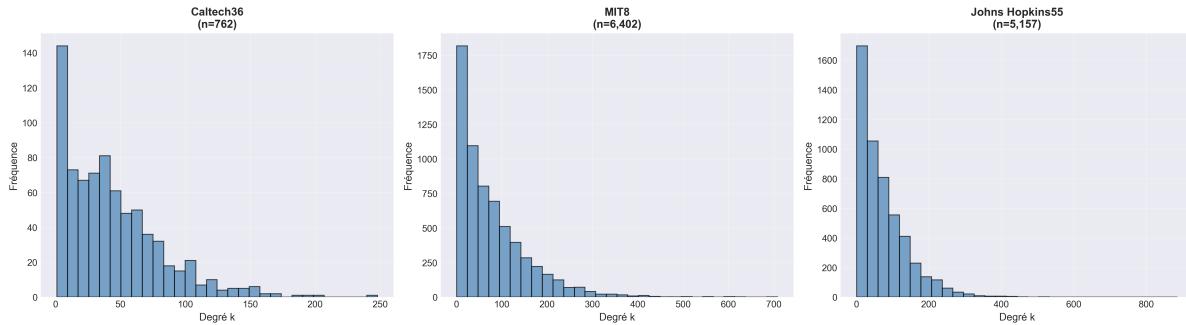


FIGURE 1 – Histogrammes des degrés pour Caltech36, MIT8 et Johns Hopkins55

La représentation en échelle log-log de la fonction de distribution cumulative complémentaire (CCDF) révèle la présence de queues lourdes (figure 2), se rapprochant d'une loi de puissance. La décroissance approximativement linéaire en échelle log-log est typique des réseaux où quelques noeuds concentrent une part importante des connexions.

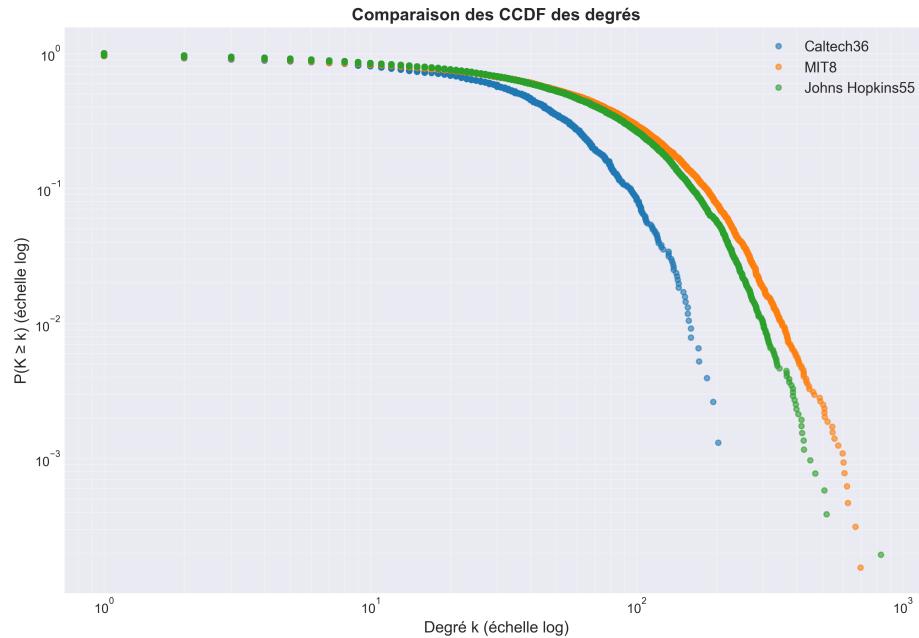


FIGURE 2 – CCDF des degrés en échelle log-log

On constate que malgré des différences de taille (de 762 à 6402 noeuds), les trois réseaux présentent des distributions qualitativement similaires, indiquant une organisation structurelle comparable.

## Clustering et densité

Les réseaux sont caractérisés par une densité très faible (tableau 2), ce qui est cohérent pour des graphes sociaux de cette dimension. En revanche, les coefficients de clustering sont élevés comparativement à ce que l'on observerait dans un graphe aléatoire de même densité.

| 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

## 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 structures locales denses. À l'inverse, les noeuds fortement connectés ont un clustering plus faible.

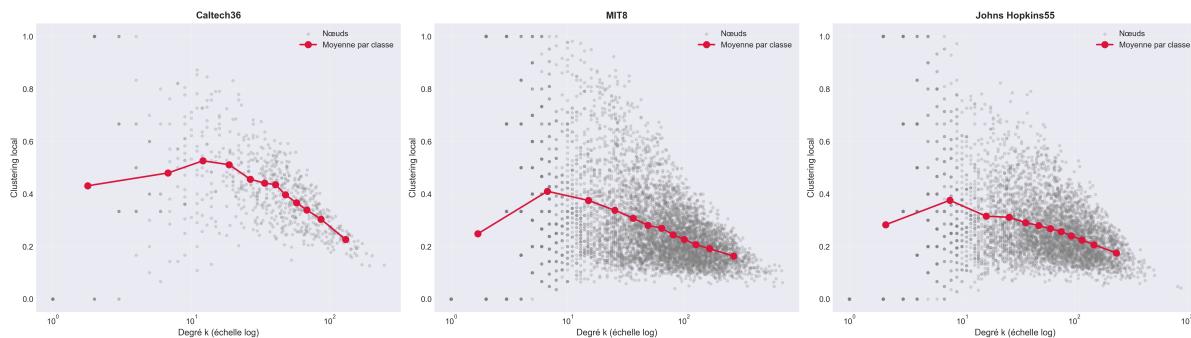


FIGURE 3 – Relation entre le degré et le coefficient de clustering local

## 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)

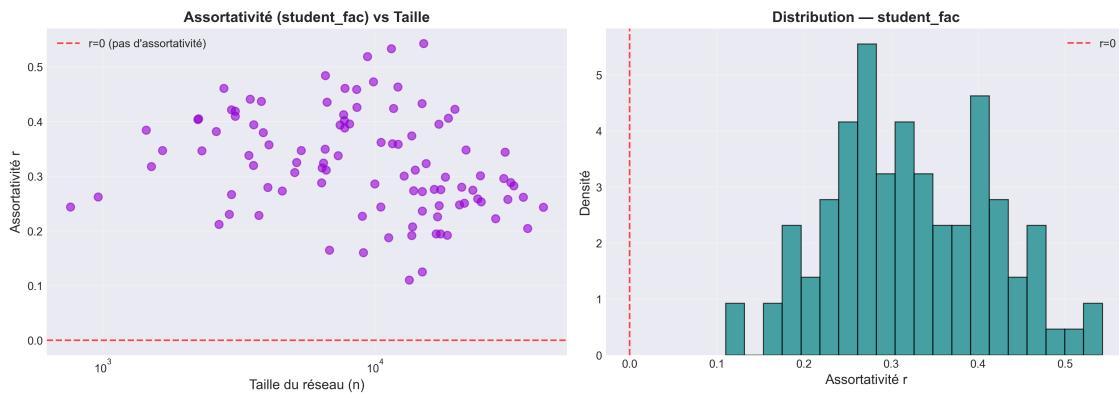


FIGURE 4 – Assortativité par statut étudiant/faculté

L’attribut *student\_fac* présente l’assortativité la plus élevée ( $r = 0,32$ ), montrant une tendance marquée des étudiants à se connecter entre eux, et des membres de la faculté à faire de même. Les valeurs sont systématiquement positives ( $r > 0,1$ ) sur l’ensemble des réseaux.

## Résidence (dorm)

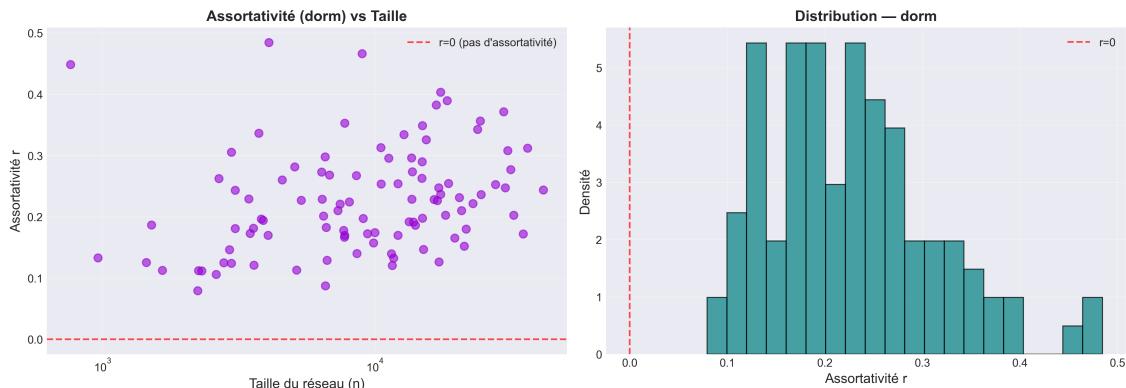


FIGURE 5 – Assortativité par résidence

La résidence affiche également une assortativité significative ( $r = 0,23$ ), bien que l’on observe une forte variabilité (0,08 à 0,48). Cela indique que le lieu de vie est un facteur corrélé aux connexions, mais son importance varie fortement selon l’organisation spécifique de chaque campus.

## Spécialité académique (major\_index)

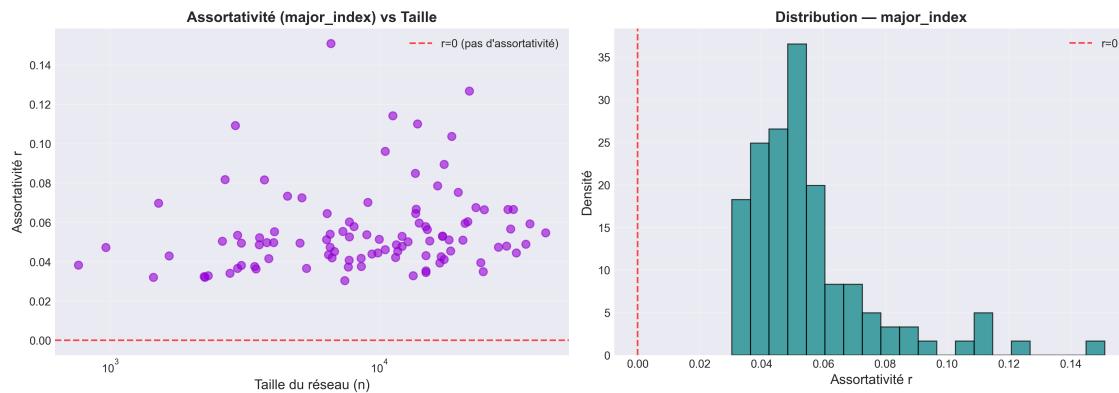


FIGURE 6 – Assortativité par spécialité académique

L'assortativité par spécialité est faible ( $r = 0,056$ ). Contrairement à la résidence ou au statut, la discipline académique ne semble pas être un facteur discriminant fort dans la structure des connexions sur Facebook pour ces données.

## Degré (degree)

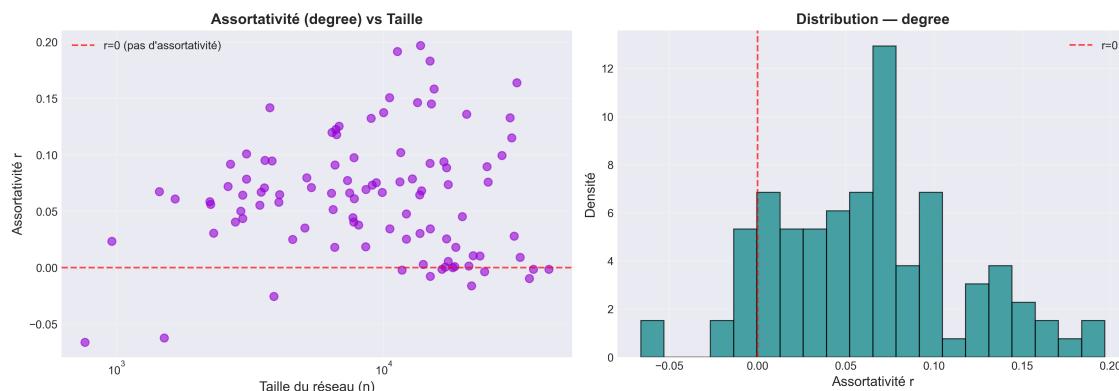


FIGURE 7 – Assortativité par degré

L'assortativité par degré est légèrement positive ( $r = 0,063$ ) mais très dispersée. Certains réseaux présentent une assortativité négative, ce qui signifie que les noeuds très connectés y sont reliés à des noeuds peu connectés, tandis que d'autres montrent une tendance inverse.

## Genre (gender)

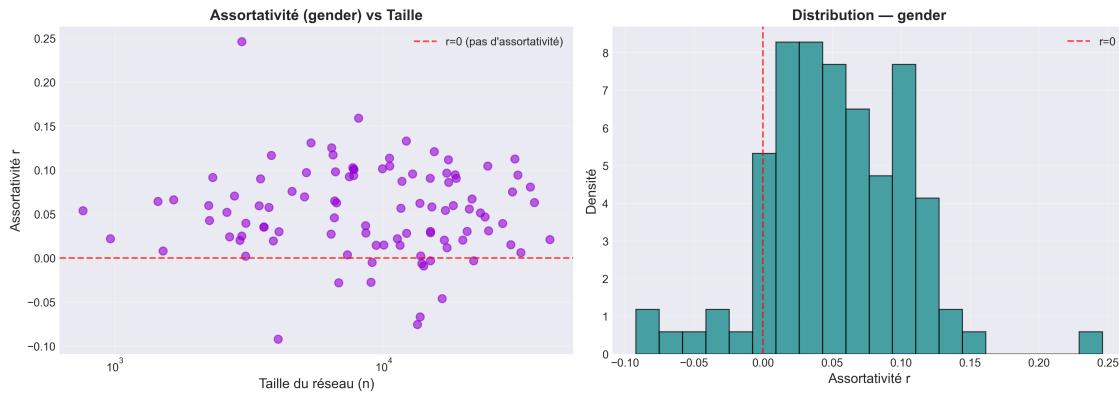


FIGURE 8 – Assortativité par genre

L'assortativité par genre est faible ( $r = 0,053$ ) et centrée proche de zéro. Cela indique une mixité importante dans les connexions : le genre n'apparaît pas comme un facteur structurant majeur des interactions dans ces graphes.

## Synthèse

L'analyse des corrélations met en évidence que les attributs liés à l'organisation spatiale et institutionnelle (*student\_fac*, *dorm*) sont plus fortement corrélés à la structure du graphe que les attributs individuels ou académiques (*major*, *gender*).

## Question 4 - Prédiction de liens

Nous implémentons manuellement 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. Pour chaque réseau, une fraction  $f \in \{0,05; 0,10; 0,15; 0,20\}$  des arêtes est supprimée aléatoirement, puis les scores sont calculés pour tenter de retrouver ces liens manquants.

## 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  $k$

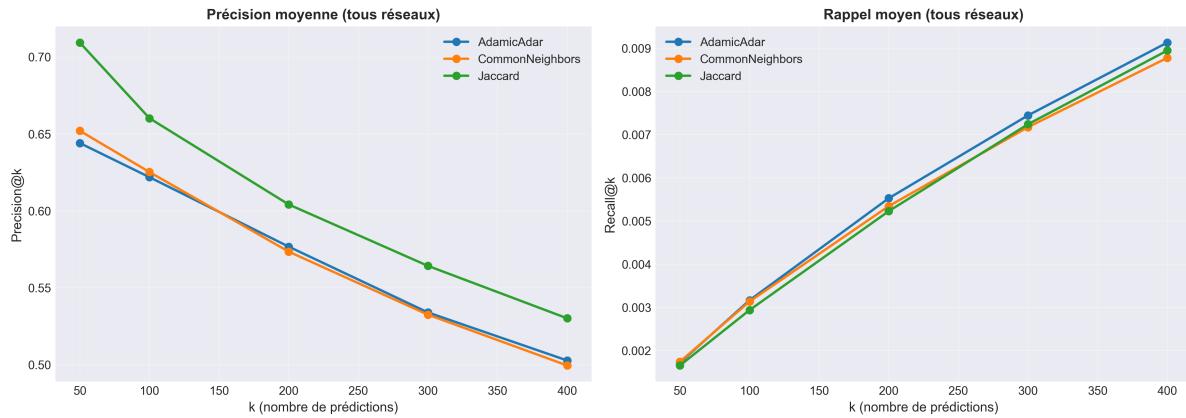


FIGURE 9 – Performances agrégées des trois méthodes (moyennes sur 15 réseaux et 4 valeurs de  $f$ )

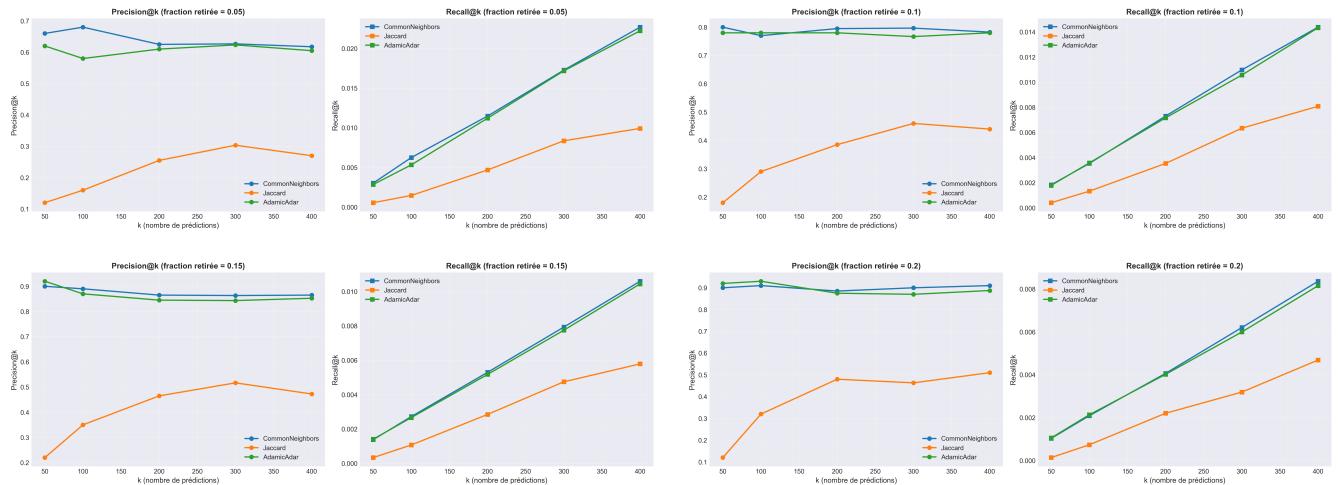


FIGURE 10 – Performances par fraction d'arêtes supprimées

## Analyse comparative

**Performances générales :** Dans nos expériences, la métrique **Jaccard** obtient les meilleurs résultats globaux (précision moyenne de 61,3%), devant **Common Neighbors** et **Adamic-Adar** (57,6%). Cette différence est notable pour les petites valeurs de  $k$  ( $k = 50$ ). Jaccard semble avantageée ici car elle normalise par la taille de l'union des voisinages, pénalisant ainsi les prédictions impliquant simplement des nœuds de très fort degré (hubs).

**Impact de la fraction supprimée :** On observe que la précision mesurée augmente avec la fraction d'arêtes supprimées  $f$ . Ce résultat peut sembler contre-intuitif, mais s'explique souvent par la nature des liens supprimés : plus on supprime d'arêtes, plus on a de chances de supprimer des liens faisant partie de triangles (qui sont faciles à prédire par ces méthodes locales), alors qu'avec un faible  $f$ , la proportion de liens "faciles" à retrouver dans l'ensemble de test peut être moindre.

**Conclusion :** Les méthodes basées sur le voisinage local fonctionnent bien sur ces réseaux, confirmant la forte transitivité (clustering) observée durant l'analyse descriptive. Jaccard offre le meilleur compromis dans ce contexte spécifique.

## Question 5 - Propagation de labels

Nous appliquons un algorithme de propagation de labels semi-supervisé ( $F^{(t+1)} = \alpha SF^{(t)} + (1 - \alpha)Y$ ) pour prédire les attributs manquants. L'évaluation est faite sur 15 réseaux avec masquage aléatoire de 10%, 20% et 30% des labels.

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

TABLE 6 – Performances moyennes de la propagation de labels

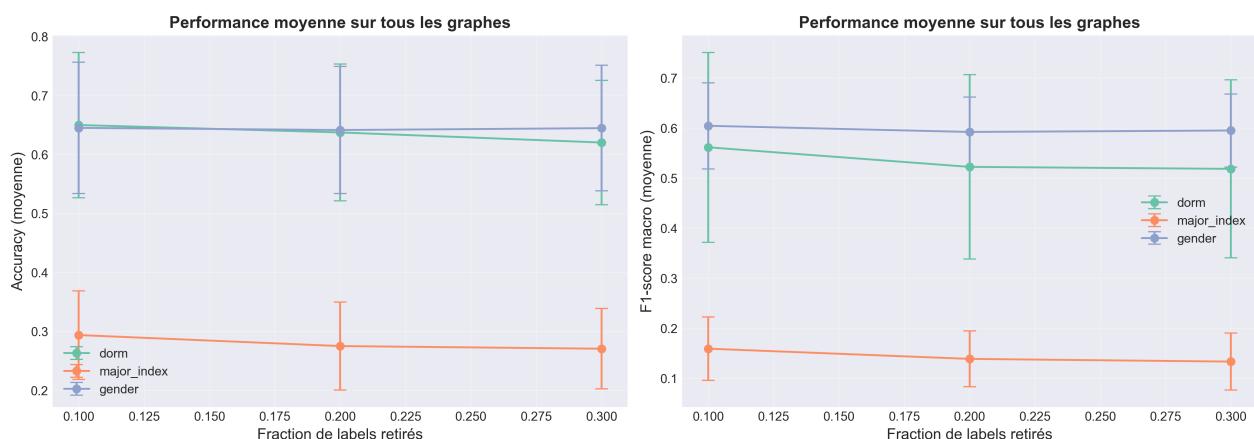


FIGURE 11 – Performances agrégées de la propagation de labels sur 15 réseaux

Les attributs **dorm** et **gender** sont les mieux prédits (Accuracy  $\approx$  64-65%), tandis que **major\_index** est difficile à prédire (Accuracy  $\approx$  29%), bien que nettement supérieur au hasard compte tenu du grand nombre de disciplines. Ces résultats sont cohérents avec l'assortativité mesurée en Question 3 : les attributs les plus assortatifs (comme le dortoir) se propagent mieux dans le graphe car les voisins partagent souvent la même valeur.

## Question 6 - Détection de communautés

### Objectif et Méthodologie

L'objectif est de déterminer si les communautés structurelles détectées uniquement à partir de la topologie du graphe correspondent à des attributs réels des étudiants (dortoir, année, discipline, genre). Nous utilisons deux algorithmes de détection de communautés (**Louvain** et **Greedy Modularity**) et comparons les partitions obtenues avec les attributs réels via deux métriques : le **NMI** (Information Mutuelle Normalisée) et l'**ARI** (Adjusted Rand Index).

Les tests sont effectués sur trois réseaux aux profils différents : *Caltech36*, *American75* et *MIT8*.

### Analyse des résultats

| 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 7 – Comparaison des communautés détectées avec les attributs réels (meilleurs scores)

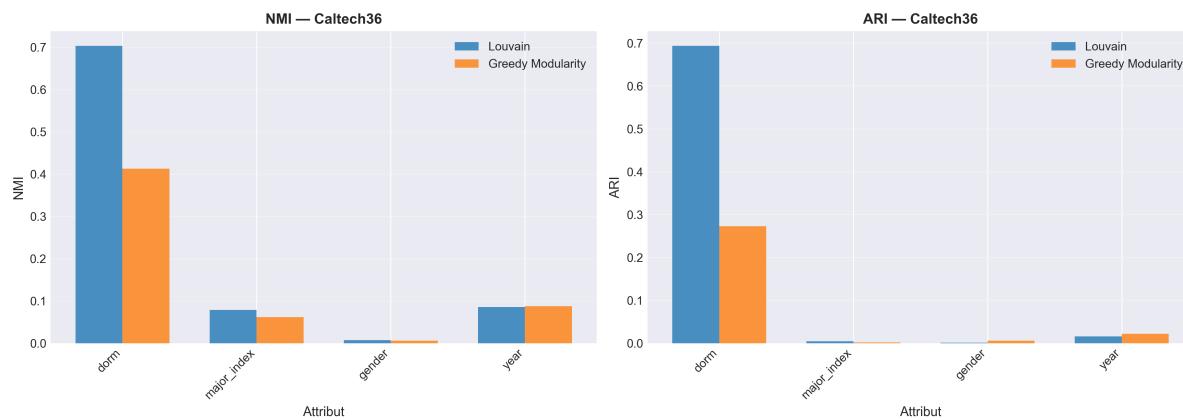


FIGURE 12 – Scores NMI et ARI pour Caltech36

L'analyse des scores NMI et ARI permet de dégager plusieurs constats réalistes sur la structure de ces réseaux :

**Le cas particulier de Caltech36 :** Sur ce réseau, l'attribut **dorm** (résidence) obtient des scores très élevés avec l'algorithme Louvain ( $NMI = 0,70$ ;  $ARI = 0,69$ ). Cela indique une correspondance presque directe entre les communautés topologiques et les résidences. C'est une spécificité connue de Caltech, organisée autour d'un "House System" très structurant pour la vie sociale.

**Résultats mitigés pour les autres universités :** Pour *American75* et *MIT8*, les résultats sont différents. Bien que les attributs **dorm** et **year** (année de promotion) obtiennent les meilleurs scores comparés aux autres attributs, les valeurs absolues restent modérées ( $NMI$  autour de  $0,27$ - $0,29$ ). Cela signifie que si la résidence et l'année influencent la formation des groupes, elles ne suffisent pas à expliquer la totalité de la structure communautaire. Les communautés détectées par les algorithmes sont probablement le résultat d'une combinaison complexe de plusieurs facteurs (cours communs, activités extra-scolaires, résidences) qu'un seul attribut ne peut capturer entièrement.

**Attributs non structurants :** Les attributs **major** (discipline) et **gender** (genre) obtiennent systématiquement des scores très faibles ( $NMI$  proche de 0). Topologiquement, les graphes ne se divisent pas selon ces lignes : les étudiants de différentes disciplines et genres sont fortement mélangés.

**Conclusion :** L'algorithme de détection de communautés retrouve efficacement la structure résidentielle là où elle est l'organisateur principal de la vie sociale (Caltech). Pour les campus plus grands et moins cloisonnés, les communautés détectées sont plus floues et ne correspondent pas uniquement à un critère administratif simple.

## Conclusion Générale

Ce projet a permis d'explorer les propriétés des réseaux sociaux universitaires via le jeu de données Facebook100. L'analyse descriptive a confirmé des caractéristiques structurelles communes : forte hétérogénéité des degrés, clustering élevé et présence de phénomènes "small-world". L'analyse des attributs a montré que les facteurs spatiaux (résidence) et temporels (année d'étude) sont davantage corrélés aux connexions que les facteurs académiques ou de genre, un résultat confirmé à la fois par l'étude de l'assortativité, la propagation de labels et la détection de communautés. Enfin, les méthodes classiques de prédiction de liens basées sur le voisinage local se sont montrées performantes, validant l'importance des structures locales (triangles) dans ces réseaux.

Les limites de cette étude résident principalement dans l'aspect statique des données et l'absence d'informations contextuelles supplémentaires qui permettraient d'affiner l'interprétation des communautés détectées sur les grands réseaux.