

Détection de communautés dans un réseau social

Présenté à

Mme Dalila Tayachi

**Épreuve de :
Projets tutorés**

Réalisé par :

**Arfaoui Rihab
Belaidi Khazri Farah
Boussabat Nizar**

**Analyse et algorithmes avancés pour comprendre les
interactions sociales**

Année Universitaire 2024-2025



TABLE DES MATIÈRES

- 1. Introduction**
- 2. Problématique**
- 3. Objectifs**
- 4. Méthodologie**
- 5. Analyse des Données**
- 6. Résultats des Algorithmes**
- 7. Interprétations**
- 8. Applications Pratiques**
- 9. Limites et Enjeux Éthiques**
- 10. Conclusion**

1. Introduction

À l'ère du numérique, les réseaux sociaux sont devenus des plateformes incontournables d'échange, de communication et de diffusion d'informations. Chaque jour, des millions d'utilisateurs interagissent, créent du contenu, partagent des opinions et forment des liens, générant ainsi une quantité massive de données relationnelles. Ces interactions, bien que souvent spontanées, suivent des dynamiques sociales complexes qui peuvent être modélisées et analysées à l'aide de graphes.

Dans ce contexte, la détection de communautés émerge comme une discipline clé de l'analyse des réseaux sociaux. Elle vise à identifier des groupes d'utilisateurs fortement connectés entre eux, mais faiblement liés au reste du réseau. Ces communautés peuvent refléter des intérêts communs, des comportements similaires ou des affinités sociales, et leur identification permet de mieux comprendre la structure sous-jacente du réseau.

L'intérêt pour cette thématique est croissant, tant dans le domaine académique que dans les secteurs industriels. En effet, la capacité à détecter des communautés permet d'optimiser les stratégies de marketing, de personnaliser les recommandations de contenu, de détecter des comportements anormaux (comme les bots ou les campagnes de désinformation), ou encore de renforcer la sécurité des plateformes.

Ce rapport s'inscrit dans cette dynamique et propose une étude approfondie de la détection de communautés dans un réseau social réel. Il s'appuie sur des données issues de la plateforme YouTube et met en œuvre plusieurs algorithmes de détection, notamment Louvain, Girvan-Newman et K-Means. L'objectif est d'analyser, comparer et interpréter les résultats obtenus afin de dégager des insights pertinents sur les comportements des utilisateurs et la structure du réseau.

2. Problématique

Les réseaux sociaux modernes, tels que YouTube, Facebook ou Twitter, génèrent chaque jour des millions d'interactions entre utilisateurs. Ces interactions, bien qu'individuellement simples (commentaires, likes, partages), forment collectivement des structures complexes et dynamiques. Comprendre ces structures est devenu un enjeu majeur pour les chercheurs, les entreprises et les institutions.

Cependant, cette compréhension se heurte à plusieurs défis. D'abord, la taille massive des réseaux rend difficile l'analyse manuelle ou visuelle des relations entre utilisateurs. Ensuite, les interactions sont souvent bruitées, incomplètes ou influencées par des facteurs externes (algorithmes de recommandation, tendances virales, bots, etc.). Enfin, les groupes d'utilisateurs ne sont pas toujours explicitement définis : ils émergent de manière implicite à travers les comportements et les connexions.

La problématique centrale de ce projet est donc la suivante :

Comment identifier automatiquement des communautés d'utilisateurs dans un réseau social à partir de leurs interactions, et quelles méthodes permettent d'obtenir des résultats fiables, interprétables et exploitables ?

Cette question soulève plusieurs sous-problèmes :

- Quels algorithmes sont les plus adaptés à la structure des réseaux sociaux ?
- Comment évaluer la qualité des communautés détectées ?
- Comment visualiser les résultats pour en faciliter l'interprétation ?
- Quelles sont les limites de ces approches, notamment en termes de scalabilité, de précision et d'éthique ?

Répondre à cette problématique permet non seulement de mieux comprendre les dynamiques sociales en ligne, mais aussi de proposer des solutions concrètes pour la personnalisation des services, la détection d'anomalies ou la modération des contenus.

3. Objectifs

L'objectif principal de ce projet est de concevoir, implémenter et évaluer des méthodes de détection de communautés dans un réseau social, en s'appuyant sur des données réelles issues de la plateforme YouTube. Cette démarche vise à mieux comprendre les dynamiques sociales à travers les interactions entre utilisateurs, et à proposer des outils d'analyse exploitables dans divers contextes applicatifs.

Les objectifs spécifiques du projet sont les suivants :

1. Étudier la structure des réseaux sociaux

Comprendre comment les utilisateurs interagissent entre eux, comment les liens se forment, et comment ces interactions peuvent être représentées sous forme de graphes. Cette étape implique une modélisation rigoureuse du réseau, en tenant compte des types de nœuds (utilisateurs, vidéos) et des types d'arêtes (commentaires, réponses, likes).

2. Identifier des communautés cohérentes

Mettre en œuvre des algorithmes capables de détecter des groupes d'utilisateurs fortement connectés. Ces communautés doivent refléter des comportements ou des intérêts communs, et permettre une segmentation pertinente du réseau.

3. Comparer plusieurs algorithmes de détection

Évaluer les performances de différents algorithmes (Louvain, Girvan-Newman, K-Means) selon des critères objectifs tels que la modularité, la couverture et la performance. L'objectif est de déterminer les forces et les limites de chaque méthode dans le contexte des réseaux sociaux.

4. Visualiser les résultats

Produire des représentations graphiques claires et interactives des communautés détectées, afin de faciliter l'interprétation des résultats. Ces visualisations doivent permettre d'identifier les structures dominantes, les nœuds centraux (influenceurs), et les zones de forte interaction.

5. Explorer les applications pratiques

Analyser comment les résultats obtenus peuvent être utilisés dans des domaines concrets :

- Marketing ciblé : segmentation d'audience.
- Recommandation de contenu : personnalisation des suggestions.
- Cybersécurité : détection de comportements anormaux.
- Modération : surveillance des groupes à risque.

6. Réfléchir aux enjeux éthiques

Identifier les risques liés à l'exploitation des données sociales : respect de la vie privée, biais algorithmiques, polarisation, etc. L'objectif est de proposer une approche responsable et éthique de l'analyse des réseaux

4. Méthodologie

L'approche méthodologique adoptée dans ce projet repose sur une combinaison rigoureuse de modélisation, d'implémentation algorithmique et de visualisation des résultats. Elle se décline en plusieurs étapes complémentaires, chacune jouant un rôle essentiel dans la détection et l'analyse des communautés au sein d'un réseau social.

1. Modélisation du réseau social

Le réseau social est représenté sous forme de graphe non orienté pondéré, où :

- Les nœuds correspondent aux utilisateurs ou aux entités (ex. : vidéos).
- Les arêtes représentent les interactions entre ces entités (commentaires, réponses, likes).
- Le poids des arêtes reflète l'intensité ou la fréquence des interactions (nombre de messages échangés, nombre de likes, etc.).

Cette modélisation permet de capturer la structure relationnelle du réseau et de préparer les données pour l'analyse communautaire.

2. Prétraitement et analyse exploratoire

Avant l'application des algorithmes, un travail de nettoyage et de transformation des données est réalisé :

- Suppression des doublons.
- Imputation des valeurs manquantes (moyenne ou médiane selon la distribution).
- Conversion des timestamps en variables temporelles (jour, heure, week-end, etc.).
- Création de nouvelles variables comme la fréquence d'interaction ou la force de la relation.

Une analyse exploratoire des données (EDA) est ensuite menée pour comprendre la distribution des interactions, identifier les utilisateurs les plus actifs et détecter d'éventuelles anomalies.

3. Application des algorithmes de détection de communautés

Trois algorithmes principaux sont utilisés :

Louvain

- Basé sur la maximisation de la modularité.
- Très efficace pour les grands graphes.
- Fonctionne en deux phases : attribution locale des communautés, puis regroupement hiérarchique.

Girvan-Newman

- Supprime les arêtes les plus centrales (selon la centralité d'intermédiarité) pour faire émerger les communautés.
- Plus précis mais coûteux en temps de calcul.
- Adapté aux graphes de taille moyenne.

K-Means

- Appliqué sur les coordonnées spatiales des nœuds (issues d'un algorithme de layout).
- Permet de regrouper les nœuds selon leur position dans l'espace du graphe.
- Nécessite de fixer le nombre de clusters à l'avance.

4. Visualisation des communautés

Les résultats sont visualisés à l'aide de bibliothèques comme NetworkX, Gephi ou D3.js. Deux types de visualisations sont privilégiés :

- Graphes interactifs : pour observer la structure globale et les connexions entre communautés.
- Cartes de chaleur : pour analyser l'intensité des interactions entre groupes.

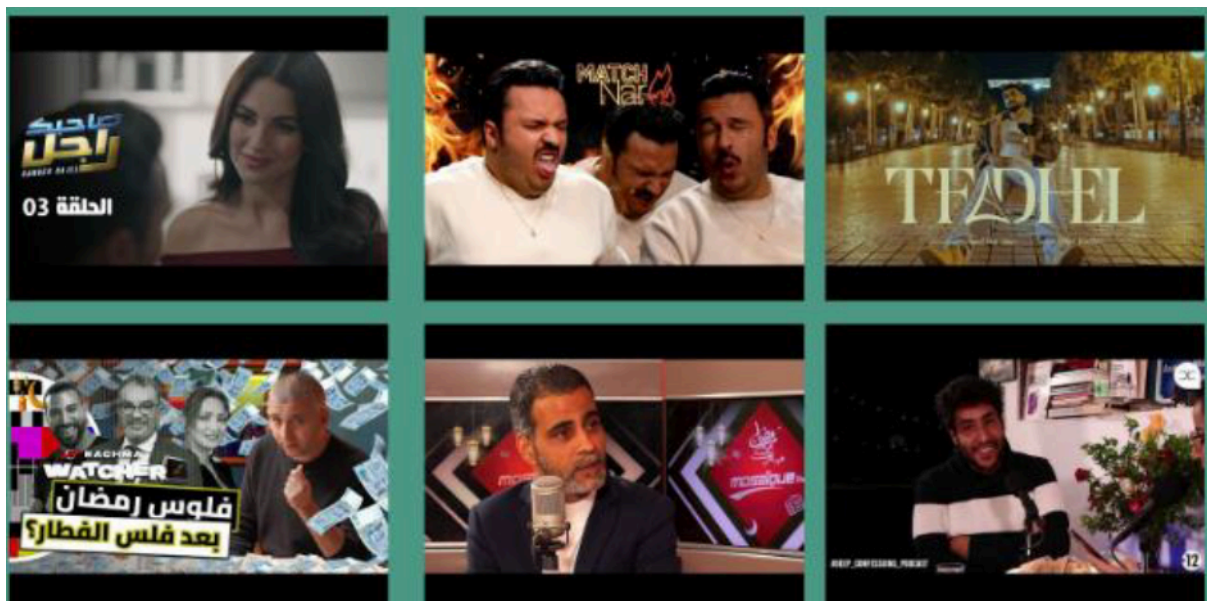
Ces visualisations facilitent l'interprétation des résultats et permettent d'identifier visuellement les communautés dominantes, les hubs d'interaction et les zones de faible connectivité.

5. Analyse des Données

L'analyse des données constitue une étape fondamentale dans le processus de détection de communautés. Elle permet de comprendre la nature des interactions entre les utilisateurs, d'identifier les tendances comportementales, et de préparer les données pour une exploitation optimale par les algorithmes de détection.

1. Source des données

Les données utilisées dans ce projet proviennent de la plateforme YouTube, collectées via l'API Google Cloud. Elles concernent les interactions entre utilisateurs autour de plusieurs vidéos populaires, notamment des commentaires, des réponses à des commentaires, et des métadonnées associées (timestamp, nombre de likes, etc.).



2. Structure du dataset

```
✓ [12] # Afficher les informations générales sur le DataFrame  
print(df.info())
```

```
↩ <class 'pandas.core.frame.DataFrame'>  
RangeIndex: 18810 entries, 0 to 18809  
Data columns (total 17 columns):  
#   Column                                Non-Null Count  Dtype  
---  -  
0   user_id                               18810 non-null  object  
1   interacted_with_id                    18801 non-null  object  
2   video_id                               18810 non-null  object  
3   interaction_type                       18810 non-null  object  
4   interaction_timestamp                  18801 non-null  object  
5   interaction_content                    18795 non-null  object  
6   like_count                            18810 non-null  int64  
7   interaction_frequency                  18810 non-null  int64  
8   relationship_strength                  18810 non-null  int64  
9   total_interactions                    18810 non-null  int64  
10  average_like_count                     18810 non-null  float64  
11  author_display_name                    18796 non-null  object  
12  author_profile_image_url               18801 non-null  object  
13  author_channel_url                     18801 non-null  object  
14  node_type                              18810 non-null  object  
15  edge_type                              18801 non-null  object  
16  edge_weight                            18810 non-null  int64  
dtypes: float64(1), int64(5), object(11)  
memory usage: 2.4+ MB  
None
```

```
✓ [13] # Afficher des statistiques descriptives sur les données numériques  
print(df.describe())
```

```
↩
```

	like_count	interaction_frequency	relationship_strength	\
count	18810.000000	18810.000000	18810.000000	
mean	3.794949	8.441893	16.061882	
std	40.406083	20.159422	109.498539	
min	0.000000	1.000000	0.000000	
25%	0.000000	1.000000	0.000000	
50%	0.000000	1.000000	0.000000	
75%	1.000000	4.000000	2.000000	
max	1658.000000	109.000000	3316.000000	

	total_interactions	average_like_count	edge_weight
count	18810.000000	18810.000000	18810.000000
mean	8.441893	3.794949	0.999522
std	20.159422	37.520906	0.021869
min	1.000000	0.000000	0.000000
25%	1.000000	0.000000	1.000000
50%	1.000000	0.000000	1.000000
75%	4.000000	1.000000	1.000000
max	109.000000	1658.000000	1.000000

3. Prétraitement des données

Avant toute analyse, un important travail de nettoyage et de structuration a été réalisé :

- Suppression des doublons : 13 lignes dupliquées ont été éliminées.

```
➡ Number of duplicate rows: 13  
Shape after removing duplicates: (3748, 17)
```

- Imputation des valeurs manquantes : les colonnes numériques ont été complétées par la moyenne ou la médiane selon la distribution ; les colonnes textuelles ont été remplies par des valeurs par défaut.

```
[ ] 1 # Impute missing values based on skewness  
2 for col in missing_df[missing_df['Missing Values'] > 0].index:  
3     if df[col].dtype in ['int64', 'float64']:  
4         if abs(skew(df[col].dropna())) > 1:  
5             df[col].fillna(df[col].median(), inplace=True) # Skewed data -> use median  
6         else:  
7             df[col].fillna(df[col].mean(), inplace=True) # Normal data -> use mean  
8
```

```
[ ] 1 # For object columns, fill with appropriate placeholders  
2 for col in df.select_dtypes(include=['object']).columns:  
3     df[col] = df[col].fillna('unknown')
```

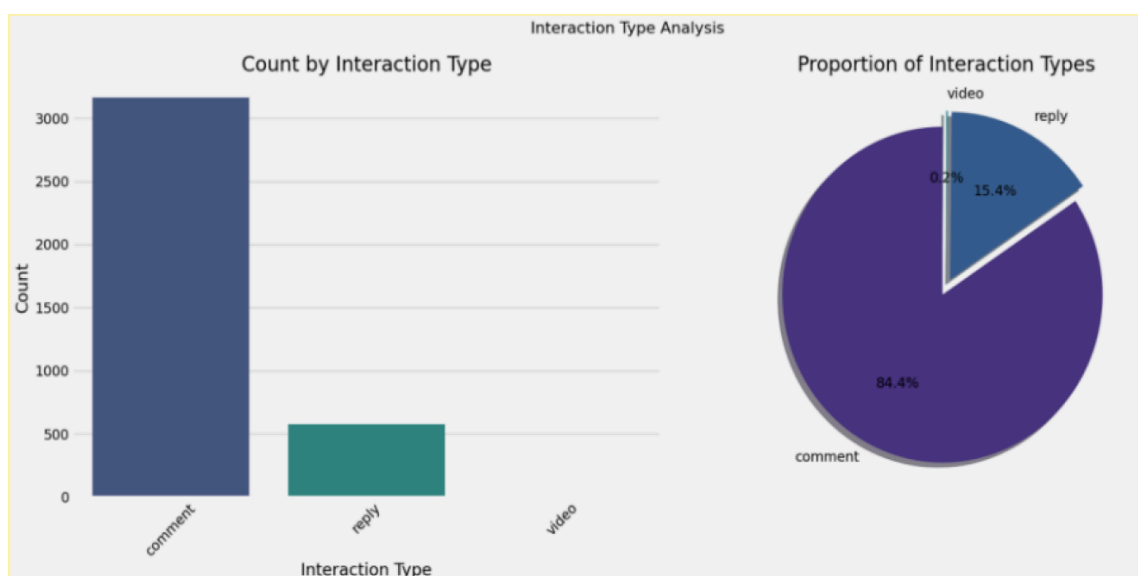
- Conversion des timestamps : transformation en variables temporelles (année, mois, jour, heure, jour de la semaine, période de la journée).

```
[ ] 1 # Advanced timestamp handling and feature engineering  
2 df['interaction_timestamp'] = pd.to_datetime(df['interaction_timestamp'], errors='coerce')  
  
[ ] 1 # Create time-based features  
2 df['year'] = df['interaction_timestamp'].dt.year  
3 df['month'] = df['interaction_timestamp'].dt.month  
4 df['day'] = df['interaction_timestamp'].dt.day  
5 df['hour'] = df['interaction_timestamp'].dt.hour  
6 df['minute'] = df['interaction_timestamp'].dt.minute  
7 df['dayofweek'] = df['interaction_timestamp'].dt.dayofweek  
8 df['is_weekend'] = df['dayofweek'].apply(lambda x: 1 if x >= 5 else 0)  
9 df['time_of_day'] = pd.cut(  
10     df['hour'],  
11     bins=[0, 6, 12, 18, 24],  
12     labels=['Night', 'Morning', 'Afternoon', 'Evening']  
13 )
```

4. Statistiques descriptives

L'ensemble de données contient 18 810 interactions réparties comme suit :

- Commentaires : 84,4 %
- Réponses : 15,4 %
- Autres types : 0,2 %



Les utilisateurs les plus actifs ont été identifiés, avec des profils ayant jusqu'à 45 interactions. Les like counts varient fortement, allant de 0 à plus de 1600, ce qui indique une forte hétérogénéité dans l'engagement des utilisateurs.

```

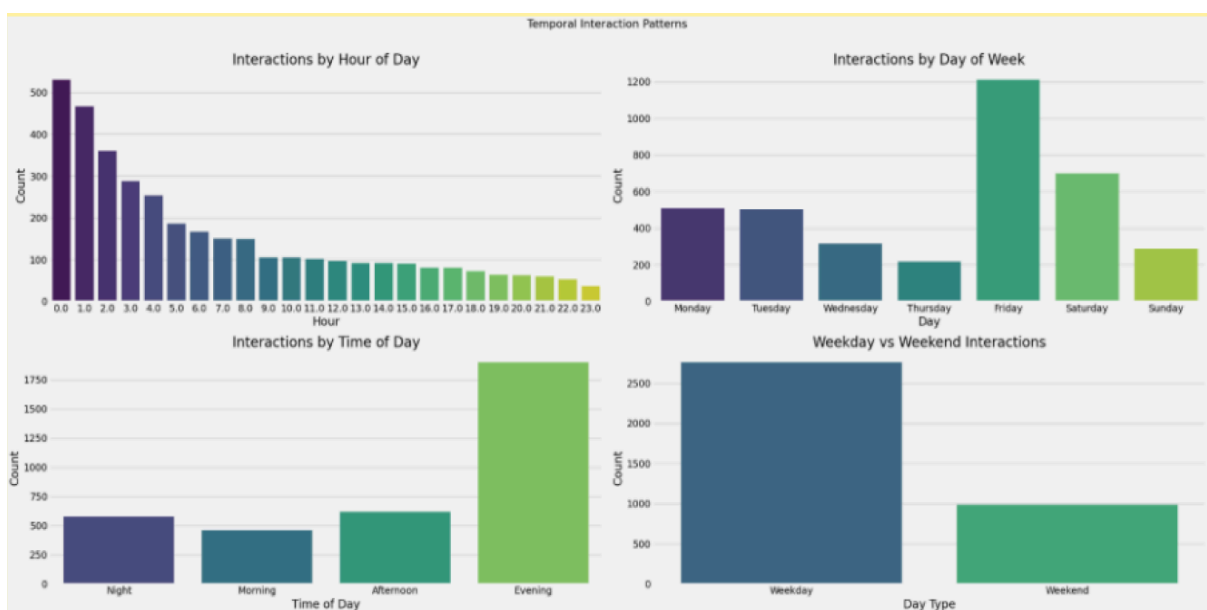
Top 10 most active users:
author_display_name
@AnasBemri          45
@Alaadinalhaj       25
@ChaymaKhadrani     17
@BLIZOOO            17
@dn-devil-2528      14
@TamerAlkhayat      10
@zandyeniyt5221     10
@MohamedKhalifaoui-g9f 10
@RouiaRouia-g1r     9
@Moebensaleh        8
Name: count, dtype: int64

```

5. Analyse temporelle

L'analyse des timestamps a révélé :

- Des pics d'activité en soirée et en fin de semaine.
- Une répartition relativement homogène des interactions sur les jours de la semaine, avec une légère hausse le week-end.
- Une activité concentrée entre 18h et 23h, correspondant aux heures de forte audience.



6. Structure du graphe

Le graphe construit à partir des données comprend :

- 18 801 nœuds (principalement des utilisateurs).
- Arêtes pondérées représentant les interactions, avec un poids moyen proche de 1.
- Une densité relativement faible, typique des réseaux sociaux réels, mais avec des sous-structures denses correspondant aux communautés.

7. Observations clés

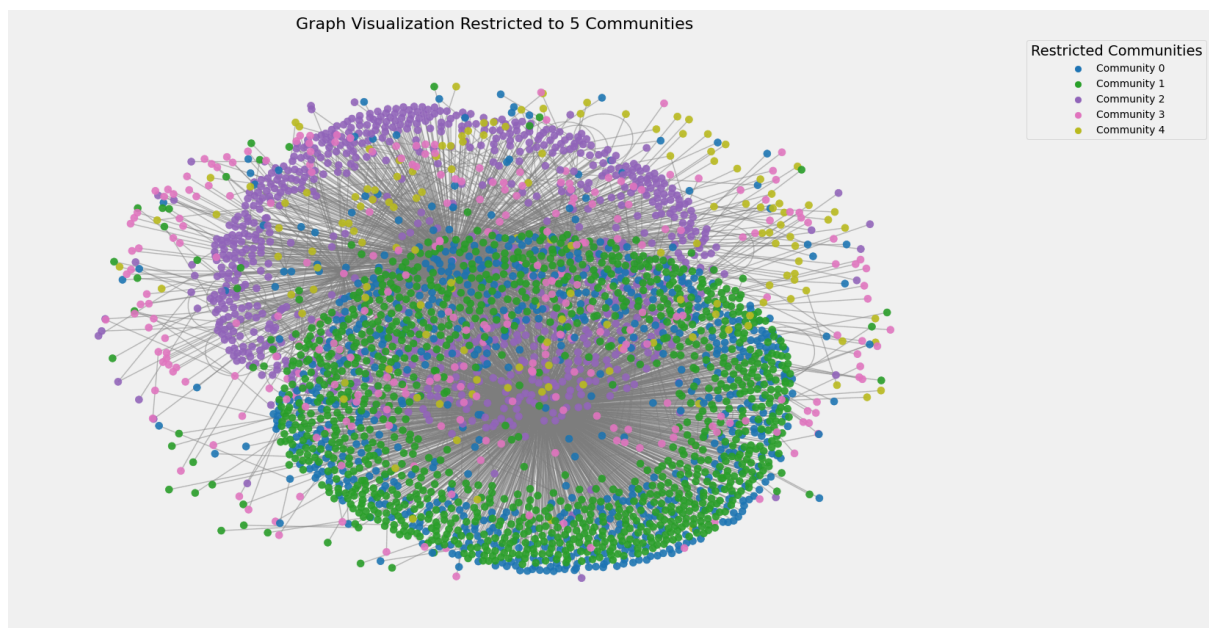
- La majorité des interactions sont concentrées autour de quelques vidéos très populaires.
- Certains utilisateurs jouent un rôle central dans le réseau, avec un grand nombre de connexions (potentiels influenceurs).
- Des sous-groupes d'utilisateurs interagissent de manière répétée, suggérant l'existence de communautés naturelles.

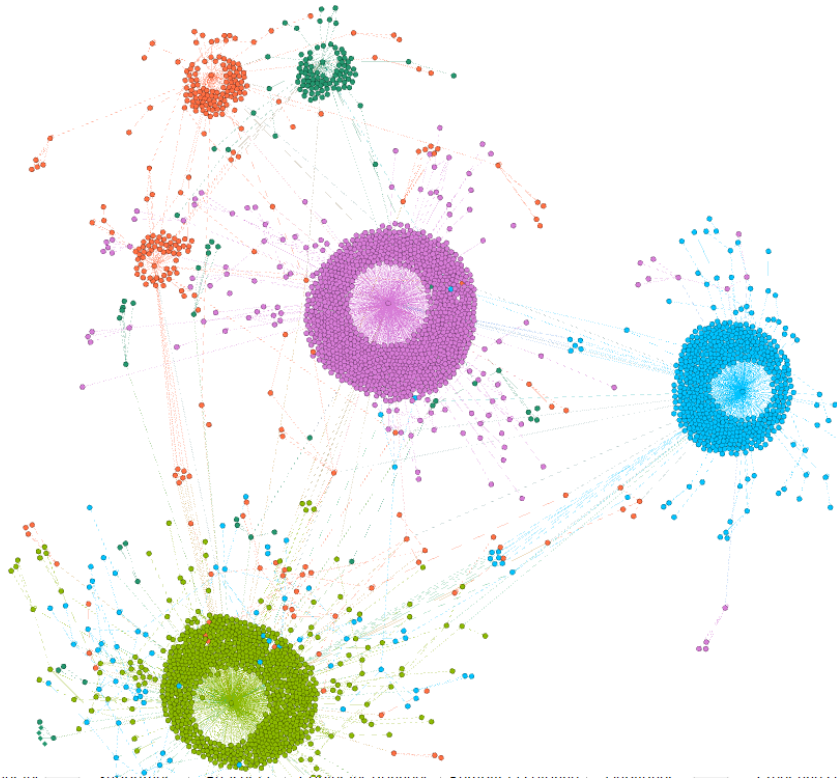
6. Résultats des Algorithmes

L'analyse des communautés a été réalisée à l'aide de trois algorithmes complémentaires : Louvain, Girvan-Newman et K-Means. Chacun d'eux repose sur des principes différents et offre des perspectives variées sur la structure du réseau social étudié. Cette section présente les résultats obtenus, les performances mesurées et les observations clés pour chaque méthode.

6.1 Algorithme de Louvain

L'algorithme de Louvain repose sur la maximisation de la modularité, une mesure qui évalue la densité des liens à l'intérieur des communautés par rapport à un graphe aléatoire.





- Nombre de communautés détectées : 5 (après regroupement via K-Means).
- Modularité obtenue : 0.6785
- Taille des communautés :
 - Communauté 4 : 1186 nœuds
 - Communauté 0 : 899 nœuds
 - Communauté 2 : 819 nœuds
 - Communauté 1 : 259 nœuds
 - Communauté 3 : 149 nœuds

Avantages :

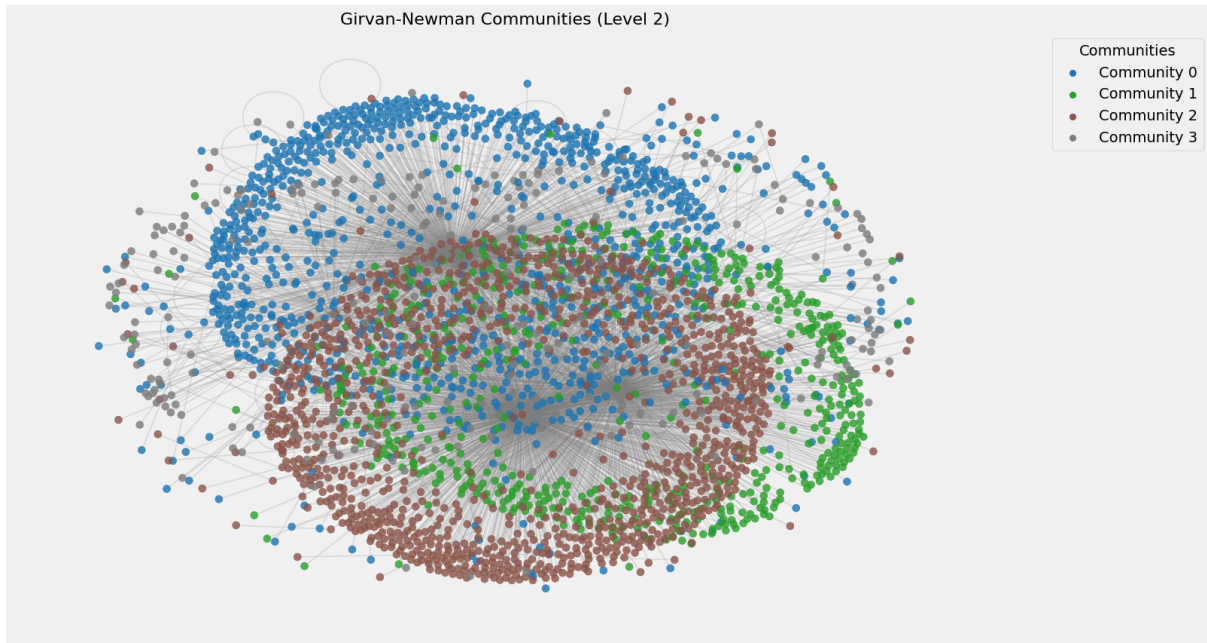
- Très rapide, même sur de grands graphes.
- Capacité à détecter des structures complexes.

Limites :

- Résolution limitée : difficulté à identifier de petites communautés.
- Ne prend pas en compte les attributs des nœuds.

6.2 Algorithme de Girvan-Newman

Cet algorithme identifie les communautés en supprimant progressivement les arêtes les plus centrales (selon la centralité d'intermédiation), ce qui fragmente le graphe en sous-groupes.



- Nombre de communautés détectées : 4 (niveau 2 du dendrogramme).
- Modularité : 0.5302
- Coverage : 0.9926
- Performance : 1.0000
- Répartition :
 - Communauté 2 : 1194 nœuds
 - Communauté 0 : 1107 nœuds
 - Communauté 1 : 624 nœuds
 - Communauté 3 : 250 nœuds

Avantages :

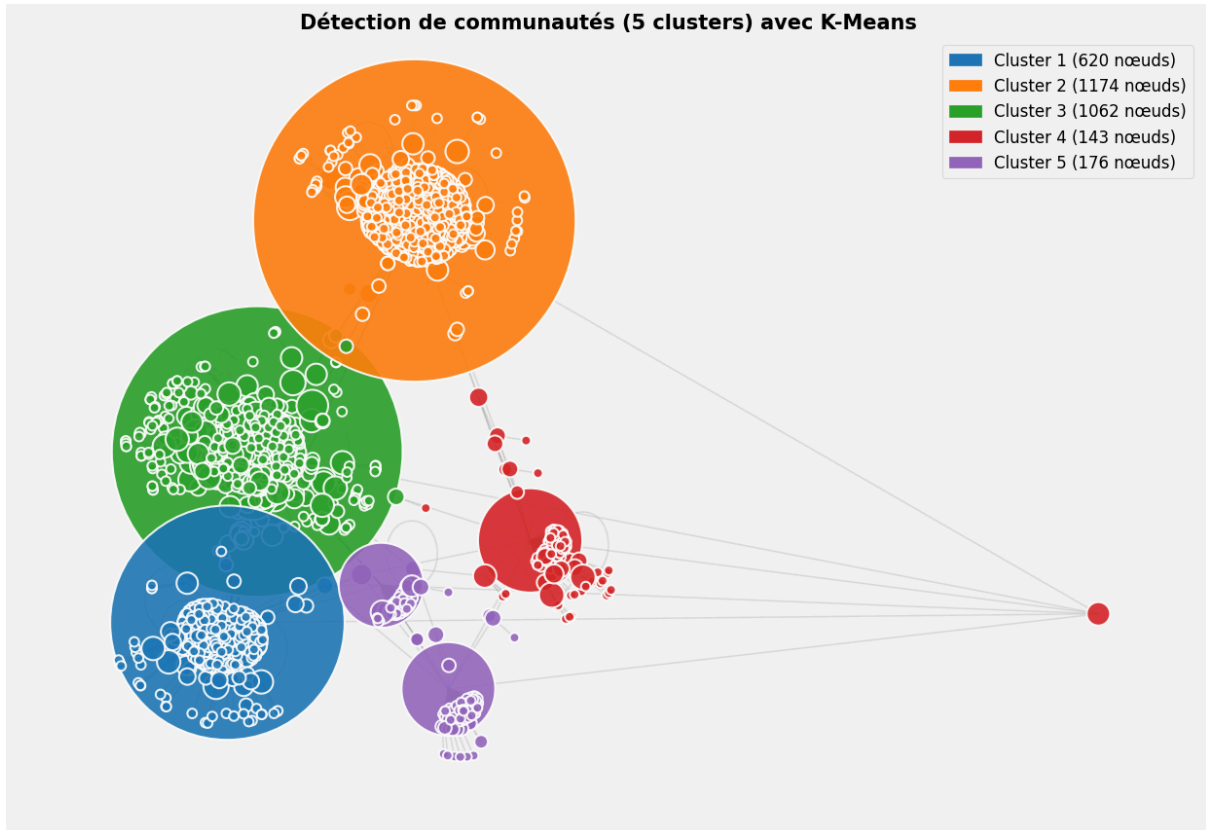
- Très bonne séparation des groupes.
- Visualisation intuitive des communautés.

Limites :

- Coût computationnel élevé.
- Moins adapté aux très grands réseaux.

6.3 Algorithme K-Means

K-Means a été appliqué sur les coordonnées spatiales des nœuds (issues d'un layout de type spring). Il permet de regrouper les nœuds selon leur position dans l'espace du graphe.



- **Nombre de clusters** : 5
- **Modularité** : 0.6750
- **Coverage** : 0.9642
- **Performance** : 0.7179
- **Répartition** :
 - Cluster 2 : 1179 nœuds
 - Cluster 1 : 1080 nœuds
 - Cluster 3 : 680 nœuds
 - Cluster 5 : 252 nœuds
 - Cluster 4 : 117 nœuds

Avantages :

- Rapide et scalable.
- Permet de forcer un nombre fixe de communautés.

Limites :

- Nécessite de fixer le nombre de clusters à l'avance.
- Sensible à la distribution spatiale des nœuds.

Comparaison des performances

Algorithme	Modularité	Coverage	Performance
Louvain	0.6785	0.9593	0.7200
Girvan-Newman	0.5302	0.9926	1.0000
K-Means	0.6750	0.9642	0.7179

Ces résultats montrent que Girvan-Newman offre la meilleure performance en termes de classification des paires de nœuds, mais au prix d'un coût computationnel élevé. Louvain et K-Means sont plus efficaces pour les grands graphes, avec des scores de modularité et de couverture très satisfaisants.

7. Interprétations

L'analyse des résultats issus des algorithmes de détection de communautés permet de tirer des enseignements précieux sur la structure du réseau social étudié, les comportements des utilisateurs, et les dynamiques d'interaction. Cette section propose une lecture qualitative des regroupements détectés, en s'appuyant sur les visualisations, les métriques et les caractéristiques des clusters.

7.1 Structure des communautés

Les résultats montrent une organisation hiérarchique et inégale du réseau :

- Les communautés 0 et 4 (Louvain et K-Means) ou 0 et 2 (Girvan-Newman) regroupent une majorité d'utilisateurs. Elles représentent des espaces d'interaction ouverts, probablement autour de contenus viraux ou de créateurs très populaires.
- Les petites communautés (ex. : cluster 3 ou 5) sont plus compactes et homogènes, suggérant des groupes d'intérêt spécifiques ou des micro-communautés centrées sur un thème ou un influenceur.

7.2 Rôle des utilisateurs

L'analyse des nœuds montre la présence de hubs : des utilisateurs très connectés, jouant un rôle central dans la diffusion de l'information. Ces profils peuvent être :

- Des créateurs de contenu influents.
- Des modérateurs ou figures communautaires.
- Des utilisateurs très engagés, interagissant avec plusieurs groupes.

À l'inverse, certains nœuds périphériques n'interagissent qu'avec un petit nombre d'autres utilisateurs, ce qui peut refléter une participation occasionnelle ou ciblée.

7.3 Cohérence des regroupements

Les scores de modularité et de couverture indiquent une bonne séparation entre les communautés. Cela signifie que :

- Les utilisateurs interagissent principalement à l'intérieur de leur groupe.
- Les liens entre communautés sont plus rares, ce qui renforce la cohésion interne.

Cette structure est typique des réseaux sociaux, où les utilisateurs se regroupent autour de centres d'intérêt communs, de vidéos spécifiques ou de créateurs particuliers.

7.4 Hypothèses thématiques

Même si les algorithmes utilisés ne prennent pas en compte le contenu des interactions, certaines hypothèses peuvent être formulées :

- Les grandes communautés sont probablement liées à des contenus populaires (humour, musique, actualité).
- Les petites communautés peuvent refléter des thématiques de niche (éducation, politique, activisme).
- Les groupes très denses peuvent indiquer une forte affinité sociale ou culturelle entre les membres.

Ces hypothèses pourraient être validées par une analyse sémantique des commentaires ou des métadonnées des vidéos.

7.5 Visualisation et lisibilité

Les visualisations générées (graphes interactifs, cartes de chaleur) confirment la clarté des regroupements :

- Les communautés sont bien séparées visuellement.
- La taille des nœuds permet d'identifier les utilisateurs les plus actifs.
- Les couleurs attribuées aux clusters facilitent la lecture et l'interprétation.

8. Applications Pratiques

La détection de communautés dans les réseaux sociaux ne se limite pas à une analyse descriptive : elle ouvre la voie à de nombreuses applications concrètes dans des domaines variés. Les résultats obtenus dans ce projet peuvent être exploités pour améliorer les services numériques, renforcer la sécurité des plateformes, et optimiser les stratégies de communication.

8.1 Marketing ciblé

L'un des usages les plus directs de la segmentation communautaire est le marketing personnalisé. En identifiant des groupes d'utilisateurs partageant des intérêts ou des comportements similaires, les entreprises peuvent :

- Adapter leurs campagnes publicitaires à chaque segment.
- Diffuser des messages plus pertinents et engageants.
- Maximiser le retour sur investissement publicitaire.

Par exemple, une communauté centrée sur des vidéos humoristiques pourrait être ciblée par des marques de divertissement, tandis qu'un groupe engagé autour de contenus éducatifs pourrait intéresser des institutions de formation.

8.2 Recommandation de contenu

Les plateformes sociales peuvent utiliser les communautés détectées pour améliorer leurs algorithmes de recommandation :

- Suggérer des vidéos ou des créateurs populaires au sein de la communauté.
- Proposer du contenu apprécié par des utilisateurs similaires.
- Éviter la redondance en diversifiant les recommandations selon les clusters.

Cette approche permet de renforcer l'engagement des utilisateurs tout en leur offrant une expérience plus personnalisée.

8.3 Détection d'anomalies et cybersécurité

La structure communautaire peut également servir à identifier des comportements suspects :

- Des groupes très denses et isolés peuvent signaler des campagnes de désinformation ou des réseaux de bots.
- Des utilisateurs interagissant de manière anormale avec plusieurs communautés peuvent être surveillés pour prévenir les abus.

Ces analyses sont particulièrement utiles pour les équipes de modération et de sécurité informatique.

8.4 Détection d'influenceurs

Les nœuds centraux dans chaque communauté (ayant un degré élevé ou un poids d'interaction important) peuvent être considérés comme des influenceurs :

- Ils jouent un rôle clé dans la diffusion de l'information.
- Ils peuvent être sollicités pour des campagnes de sensibilisation ou de partenariat.
- Leur comportement peut être analysé pour comprendre les dynamiques virales.

8.5 Analyse comportementale et sociologique

Au-delà des usages commerciaux, la détection de communautés permet de mieux comprendre les dynamiques sociales :

- Étudier la formation de groupes d'intérêt.
- Observer les effets de polarisation ou de fragmentation.
- Analyser les trajectoires d'engagement des utilisateurs.

Ces analyses peuvent être utiles dans des contextes de recherche académique, de veille stratégique ou de politiques publiques.

9. Limites et Enjeux Éthiques

Bien que la détection de communautés offre de nombreuses opportunités analytiques et applicatives, elle soulève également des limitations techniques et des questions éthiques majeures. Il est essentiel de les prendre en compte pour garantir une utilisation responsable et pertinente des résultats.

9.1 Limites techniques

a. Sensibilité à la structure du graphe

Les algorithmes comme Louvain ou Girvan-Newman sont fortement influencés par la topologie du réseau. Une mauvaise qualité des données (bruit, liens manquants, biais de collecte) peut fausser les résultats.

b. Scalabilité

Certains algorithmes, notamment Girvan-Newman, sont peu adaptés aux grands graphes en raison de leur complexité algorithmique. Cela limite leur usage dans des réseaux sociaux à très grande échelle.

c. Paramétrage

K-Means nécessite de fixer à l'avance le nombre de clusters, ce qui peut introduire un biais si ce paramètre n'est pas bien choisi. De plus, les résultats peuvent varier selon l'initialisation.

d. Absence de sémantique

Les algorithmes utilisés se basent uniquement sur la structure du graphe, sans prendre en compte le contenu des interactions (texte des commentaires, tonalité, etc.). Cela limite la profondeur de l'analyse.

9.2 Enjeux éthiques

a. Respect de la vie privée

L'analyse des interactions sociales, même anonymisées, peut révéler des comportements sensibles ou des affiliations implicites. Il est crucial de garantir le consentement éclairé des utilisateurs et de respecter les réglementations (ex. : RGPD).

b. Risque de discrimination algorithmique

Les regroupements automatiques peuvent renforcer des stéréotypes ou exclure certains profils si les données sont biaisées. Une vigilance particulière est nécessaire pour éviter des décisions injustes basées sur des clusters mal interprétés.

c. Bulles informationnelles

La segmentation en communautés peut accentuer la polarisation en enfermant les utilisateurs dans des groupes homogènes, réduisant leur exposition à des opinions divergentes. Cela peut nuire à la diversité de l'information.

d. Surveillance et manipulation

Les résultats de la détection de communautés peuvent être utilisés à des fins de profilage, de ciblage politique ou de manipulation sociale. Il est essentiel de poser des limites claires à l'usage de ces technologies.

9.3 Recommandations éthiques

- Intégrer des audits algorithmiques pour évaluer les biais.
- Associer des experts en éthique et en sciences sociales aux projets d'analyse.
- Favoriser la transparence des méthodes utilisées.
- Informer les utilisateurs sur l'usage de leurs données et leur offrir des mécanismes de contrôle.

10. Conclusion

La détection de communautés dans les réseaux sociaux représente un enjeu stratégique majeur à l'ère du numérique. Elle permet de révéler des structures sociales implicites, de mieux comprendre les dynamiques d'interaction entre utilisateurs, et d'orienter des décisions dans des domaines aussi variés que le marketing, la cybersécurité, la recommandation de contenu ou la modération.

Dans ce projet, nous avons exploré plusieurs approches algorithmiques — Louvain, Girvan-Newman et K-Means — appliquées à un réseau social construit à partir de données issues de YouTube. Chaque méthode a permis de mettre en évidence des regroupements cohérents d'utilisateurs, avec des performances variables selon les critères de modularité, de couverture et de précision.

L'analyse des résultats a montré que :

- Les grandes communautés sont souvent liées à des contenus viraux ou à des influenceurs.
- Les petites communautés peuvent représenter des niches thématiques ou des sous-cultures.
- Les visualisations permettent une lecture intuitive des structures détectées.

Au-delà des résultats techniques, ce travail a mis en lumière les opportunités offertes par la détection de communautés, mais aussi les risques associés à une mauvaise utilisation des données sociales. Il est donc essentiel d'adopter une démarche éthique, transparente et respectueuse de la vie privée.

En conclusion, la détection de communautés constitue un outil puissant pour explorer, comprendre et agir sur les réseaux sociaux. Elle ouvre la voie à des analyses plus fines, à des services plus personnalisés, et à une meilleure gouvernance des espaces numériques.