

Projet de fin d’année

**Data extraction and analysis of Co-author network analysis**

Fait par: Encodé par:

Abdessamd akharaz Ikram el asri

Houdayfa Housny

# Contents

Table des matières

[Chapitre 1 1](#_Toc167919600)

[1.1 Introduction](#_Toc167919604) 2

[1.2 les technologies](#_Toc167919605) 3

[Chapiter 2 M´ethodologies 5](#_Toc167919601)

[2.1 Collecte des donn´ees 5](#_Toc167919602)

[2.1.1 Semantic Scholar 5](#_Toc167919603)

[2.2 Analyse des donn´ees 6](#_Toc167919604)

[2.3 Visualisation 9](#_Toc167919605)

[Chapter 3 R´esultats 12](#_Toc167919606)

[Chapter 4 Discussion 13](#_Toc167919607)

[Chapter 5 13](#_Toc167919609)

[5.1 Conclusion 6](#_Toc167919604)

[5.2 Référence 9](#_Toc167919605)

**Chapitre 1 Introduction**

La recherche scientifique repose sur la collaboration entre chercheurs, facilitée par la publication d'articles en co-auteur. Analyser ces collaborations peut révéler des informations précieuses sur la structure et la dynamique des réseaux de recherche. Ce projet vise à extraire des données sur les auteurs et leurs co-auteurs à partir de la plateforme Semantic Scholar, puis à construire et analyser un réseau de collaboration.

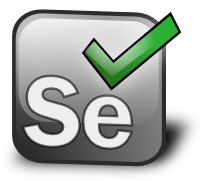
Semantic Scholar est une base de données académique qui indexe des millions d'articles de recherche et fournit des informations détaillées sur les auteurs, les publications, et les relations de co-auteur. En utilisant des techniques de web scraping et d'analyse de données, nous avons collecté des informations sur les auteurs et leurs collaborations. Ces données ont ensuite été utilisées pour construire un réseau de co-auteur, où les nœuds représentent des auteurs et les liens représentent des collaborations entre eux.

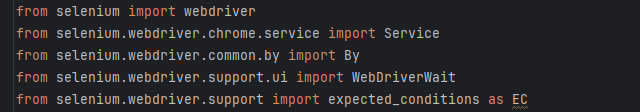
L'analyse de ce réseau de co-auteur permet d'identifier des tendances et des motifs dans les collaborations scientifiques. Par exemple, il est possible de repérer des clusters de chercheurs travaillant sur des sujets similaires, de mesurer la centralité des auteurs influents, et d'analyser l'évolution des réseaux de collaboration au fil du temps. Ces informations peuvent être utilisées pour comprendre comment les idées se propagent dans la communauté scientifique, identifier des opportunités de collaboration, et évaluer l'impact de certains chercheurs ou institutions.

Ce rapport présente les méthodes utilisées pour l'extraction des données et la construction du réseau, les résultats de l'analyse, et les implications de ces résultats pour la compréhension des réseaux de recherche scientifique. En fournissant une vue d'ensemble des collaborations académiques, ce projet contribue à la littérature existante sur l'analyse des réseaux sociaux et offre des outils pour une gestion plus efficace des collaborations scientifiques.

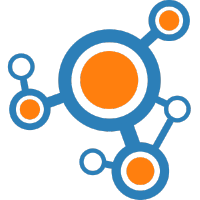
**Les technologies :**

**Selenium :**





Selenium est un outil puissant en Python qui permet d'automatiser les navigateurs web. Il est couramment utilisé pour le test d'applications web, la navigation automatisée, ainsi que pour le scraping de données à partir de pages web. Grâce à Selenium, les scripts peuvent interagir avec les éléments d'une page web de la même manière qu'un utilisateur humain, ce qui inclut la soumission de formulaires, la navigation entre les pages, et l'extraction de contenu dynamique.

**Networkx :**

NetworkX est une bibliothèque Python destinée à la création, la manipulation et l'étude des structures, dynamiques et fonctions complexes des réseaux. Elle permet de travailler avec différents types de graphes (non orientés, orientés, multigraphes), de calculer des métriques de réseau (comme le degré, la centralité, les chemins les plus courts), et de visualiser les réseaux de manière efficace. NetworkX est largement utilisée dans les domaines de la science des données, de l'analyse des réseaux sociaux, et de la recherche opérationnelle, entre autres.

**Pyvis :**

Pyvis est une bibliothèque Python qui facilite la visualisation interactive des réseaux et des graphes. Elle s'appuie sur la bibliothèque JavaScript vis.js pour générer des graphes dynamiques et interactifs pouvant être visualisés dans les navigateurs web. Pyvis permet aux utilisateurs de créer des visualisations attrayantes et intuitives des réseaux, où l'on peut zoomer, faire glisser les nœuds et explorer les connexions. Cette bibliothèque est particulièrement utile pour ceux qui souhaitent présenter des réseaux complexes de manière visuellement engageante et facile à comprendre.

**Matplotlib :**

Matplotlib est une bibliothèque Python largement utilisée pour la création de visualisations statiques, animées et interactives. Elle offre une grande flexibilité pour générer divers types de graphiques, tels que des courbes, des histogrammes, des diagrammes en barres, des scatter plots, et bien d'autres. Matplotlib permet de personnaliser presque tous les aspects des visualisations, y compris les axes, les légendes, les annotations et les couleurs. C'est un outil essentiel pour les scientifiques des données, les chercheurs et les ingénieurs, facilitant l'exploration et la communication des données de manière claire et informative.

# Chapitre 2 M´ethodologies

## Collecte des donn´ees

### Semantic Scholar



Semantic Scholar est une plateforme de recherche académique avancée, développée par l'Allen Institute for AI (AI2) en 2015. Son objectif principal est de faciliter l'accès et la découverte d'articles scientifiques pour les chercheurs, étudiants, et professionnels du monde entier. Contrairement aux moteurs de recherche académiques traditionnels, Semantic Scholar utilise des techniques avancées d'intelligence artificielle et de traitement du langage naturel pour améliorer la pertinence et la qualité des résultats de recherche.

**Caractéristiques Principales de Semantic Scholar**

***1. Indexation Étendue:***

- Semantic Scholar indexe des millions d'articles de recherche provenant de diverses disciplines scientifiques, incluant les sciences de la vie, la médecine, les sciences sociales, l'informatique, l'ingénierie, et bien d'autres.

***2. Recherche Sémantique:***

- Utilisant des algorithmes de traitement du langage naturel (NLP), Semantic Scholar comprend le contexte des requêtes de recherche, permettant ainsi de fournir des résultats plus pertinents. Par exemple, au lieu de se baser uniquement sur les mots-clés exacts, le moteur de recherche peut interpréter l'intention derrière une recherche et proposer des articles liés thématiquement

***3. Graphiques de Citations:***

- L'une des fonctionnalités distinctives de Semantic Scholar est sa capacité à visualiser les réseaux de citations. Les utilisateurs peuvent voir comment les articles sont interconnectés par les citations, ce qui aide à identifier les travaux les plus influents dans un domaine spécifique.

***4. Résumé Automatique:***

- Semantic Scholar génère des résumés automatiques des articles, appelés "highly influential citations", permettant aux utilisateurs de comprendre rapidement l'importance et le contenu des travaux de recherche sans avoir à lire chaque article en entier.

***5. Support de Données:***

- Semantic Scholar soutient les initiatives de science ouverte en offrant des liens vers les ensembles de données associés aux publications, lorsqu'ils sont disponibles, ce qui encourage la transparence et la reproductibilité dans la recherche scientifique.

Semantic Scholar révolutionne la manière dont les chercheurs accèdent à la littérature scientifique en fournissant des outils plus intelligents et des analyses plus profondes. Les avantages principaux incluent :

**- Découverte de Connaissances:** La visualisation des réseaux de citations et les filtres avancés aident à découvrir des articles et des relations qui pourraient ne pas être immédiatement apparents avec des moteurs de recherche traditionnels.

**- Amélioration de la Recherche Collaborative:** Les fonctionnalités de profil d'auteur et de co-auteur encouragent la collaboration et la mise en réseau, ce qui est essentiel dans un paysage de recherche de plus en plus interdisciplinaire.

En somme, Semantic Scholar représente une avancée significative dans la recherche académique, offrant des outils puissants pour améliorer la découverte, la compréhension et l'application des connaissances scientifiques.

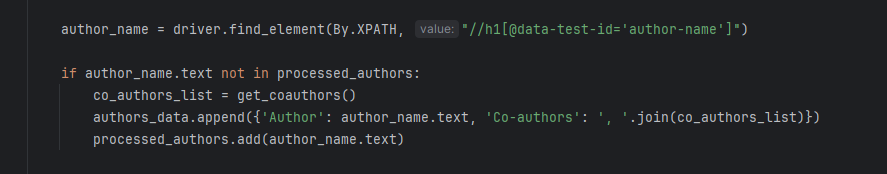
## Analyse des donn´ees

Une fois les donn´ees collect´ees, plusieurs ´etapes d’analyse sont r´ealis´ees :

* **Nettoyage des donn´ees pour ´eliminer les doublons et corriger les erreurs de formatage.**

Le nettoyage des données est une étape cruciale dans toute analyse de données, particulièrement lorsqu'il s'agit de données collectées via des techniques de web scraping. Cette étape assure que les données utilisées pour la construction et l'analyse du réseau sont précises et fiables. Voici un détail des processus impliqués dans le nettoyage des données pour éliminer les doublons et corriger les erreurs de formatage.

Les doublons peuvent se produire lorsque la même information est collectée plusieurs fois, soit à partir de différentes sources soit à partir de la même source mais sous différentes formes. L'élimination des doublons est essentielle pour garantir que chaque auteur et chaque co-auteur ne soient comptés qu'une seule fois dans le réseau.



* **Construction d’un graphe ou` les nœuds repr´esentent des auteurs et les arˆetes repr´esentent des collaborations.**

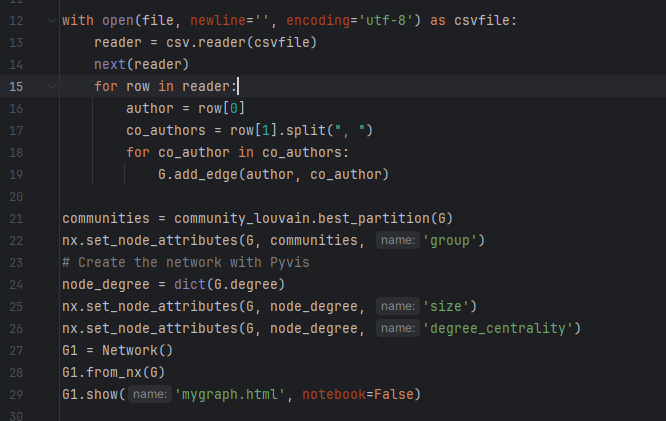
La construction d’un graphe est une étape clé dans l’analyse de réseau, permettant de visualiser et de comprendre les relations entre les auteurs. Voici un guide détaillé pour construire un graphe où les nœuds représentent des auteurs et les arêtes représentent des collaborations.

1. Étapes de Construction du Graphe
2. Chargement des Données Nettoyées
3. Initialisation du Graphe
4. Ajout des Nœuds
5. Ajout des Arêtes
6. Calcul des Attributs des Nœuds
7. Sauvegarde et Visualisation du Graphe



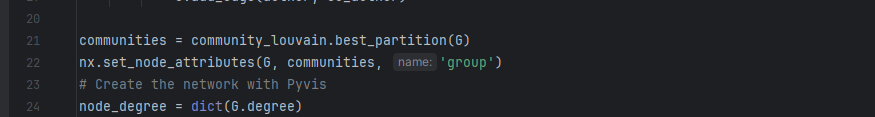
* **Calcul des m´etriques de r´eseau telles que la centralit´e, le degr´e, et les communaut´es.**

Dans l'analyse des réseaux, plusieurs métriques sont utilisées pour comprendre la structure et les dynamiques du réseau. Les métriques couramment calculées incluent la centralité, le degré, et l'identification des communautés. Voici un guide détaillé pour calculer ces métriques.



* **Utilisation de l’algorithme de Louvain pour la d´etection des commu- naut´es.**

L'algorithme de Louvain est une méthode populaire pour détecter les communautés dans les réseaux complexes. Il est basé sur l'optimisation de la modularité, une mesure de la densité des connexions à l'intérieur des communautés par rapport aux connexions entre les communautés. Voici un guide détaillé pour utiliser l'algorithme de Louvain pour détecter les communautés dans un réseau de collaborations entre auteurs.



## Visualisation

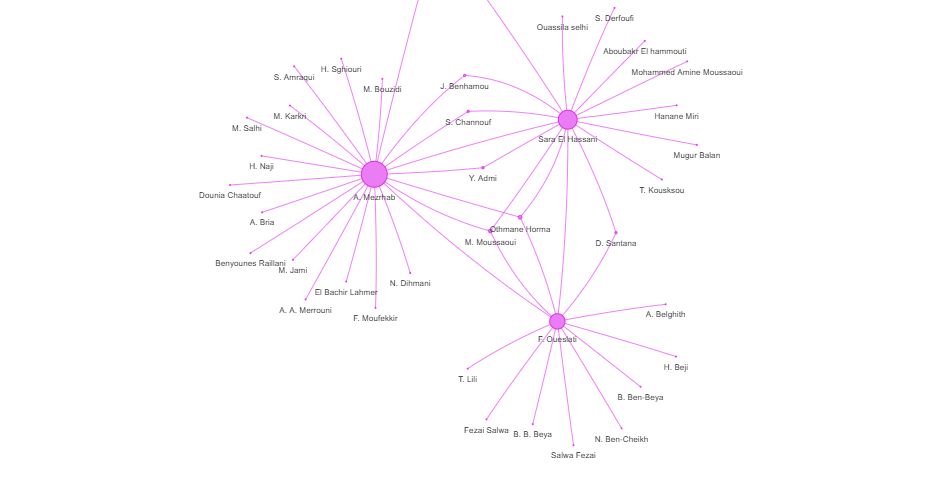
La visualisation des données joue un rôle crucial dans l'interprétation des résultats de l'analyse de réseau. Nous avons utilisé la bibliothèque Pyvis pour générer des graphes interactifs qui permettent une exploration détaillée du réseau de collaboration entre auteurs. De plus, nous pouvons également utiliser la bibliothèque Matplotlib pour créer des visualisations statiques et des tracés qui complètent notre analyse. Voici comment nous avons utilisé ces deux bibliothèques pour visualiser les données et interpréter les résultats.

#### Visualisation avec Pyvis

Pyvis est une bibliothèque Python qui permet de créer des graphes interactifs basés sur les données de NetworkX. Nous avons utilisé Pyvis pour créer des visualisations interactives des réseaux de collaboration entre auteurs. Dans ces visualisations :

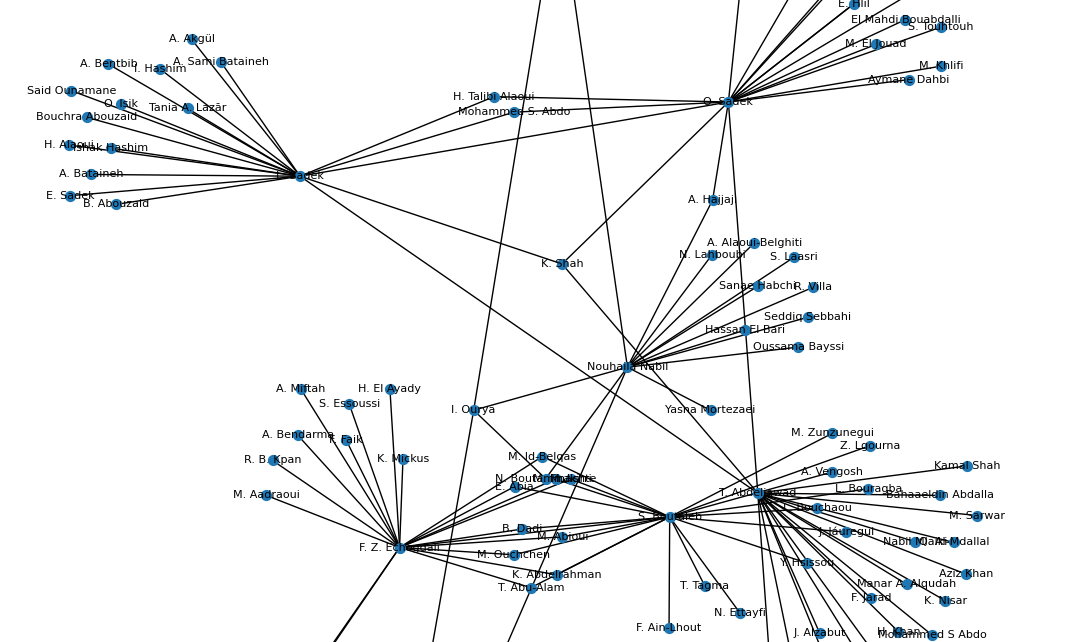
* Les nœuds représentent les auteurs, et les arêtes représentent les collaborations.
* Les nœuds sont colorés et dimensionnés en fonction des différentes métriques calculées, telles que le degré et la centralité.
* En survolant un nœud, des informations détaillées sur l'auteur sont affichées, y compris ses métriques calculées.

Ces visualisations interactives nous permettent d'explorer le réseau de collaboration de manière intuitive, en identifiant les auteurs les plus influents, les communautés de collaboration, et les tendances générales dans le réseau.



#### Interprétation avec Matplotlib

En plus des visualisations interactives, nous pouvons utiliser la bibliothèque Matplotlib pour créer des tracés statiques qui complètent notre analyseUn histogramme des degrés des nœuds peut nous donner un aperçu de la distribution de la connectivité des auteurs dans le réseau.

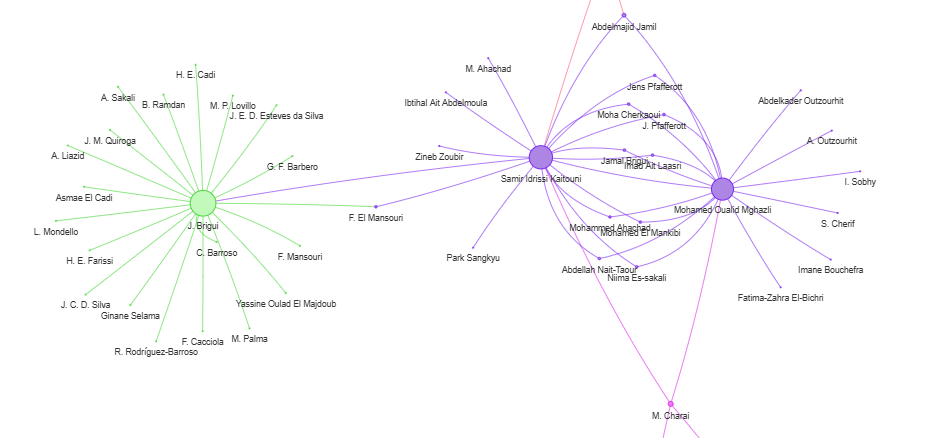


# Chapitre 3 R´esultats

Les r´esultats de cette ´etude r´ev`elent plusieurs tendances int´eressantes dans les r´eseaux de co-auteurs :

* Identification des auteurs les plus influents par leur centralit´e.
* D´etection des communaut´es de chercheurs travaillant sur des th`emes similaires.
* Analyse des dynamiques de collaboration a` travers diff´erentes p´eriodes.

Des visualisations interactives permettent de naviguer dans le r´eseau et d’explorer les relations individuelles entre les chercheurs.



# Chapitre 4

# Discussion

Les r´esultats obtenus soul`event plusieurs points de discussion :

* Les structures de collaboration peuvent varier significativement selon les disciplines.
* L’impact des collaborations internationales sur la diffusion des connais- sances.
* Les limitations de l’approche utilis´ee, notamment en termes de couver- ture des donn´ees et de biais possibles.

La discussion aborde également des résultats qu’on n’arrive pas à simuler, notamment en ce qui concerne la visualisation du réseau :

* Dans la liste des co-auteurs, il peut y avoir des auteurs travaillant sur d'autres recherches. Il est important d'établir des liens entre eux.
* De plus, dans la partie extraction des données, il est nécessaire d'extraire les citations pour chaque auteur afin de les utiliser dans le calcul du degré.
* Aussi dans la partie d’identification des communistes dans veut le faire selon le domaine ou bien la spécialité de chaque auteur

# Chapitre 5

# Conclusion

Cette ´étude sur les r´eseaux de co-auteurs d´emontre l’utilit´e des analyses de r´eseau pour comprendre les dynamiques de collaboration scientifique. Les techniques utilis´ees, incluant le scraping de donn´ees et l’analyse de r´eseau, offrent des outils puissants pour explorer et visualiser les interactions entre les chercheurs. Des pistes de recherche futures incluent l’int´egration de donn´ees suppl´ementaires et l’am´elioration des algorithmes de d´etection de communaut´es.

**Reference :**

<https://pyvis.readthedocs.io/en/latest/>

<https://networkx.org/>

<https://matplotlib.org/>

<https://selenium-python.readthedocs.io/waits.html>