



Licence MIAGE

## Rapport de projet Graphes et Open data

# Réseaux de Citations et de Collaborations Scientifiques : Analyse, Visualisation et Interprétation

Projet réalisé du 15 mars 2024 au 20 mai 2024

Membres du groupe

LALI Mohamed KenziMembre

## **Remerciements**

Je tiens à remercier tout d'abord les deux enseignants grâce à qui ce projet fort intéressant a pu avoir lieu : Mr François Delbot et Mr Valentin Bouquet. J'aimerais ensuite remercier OpenAI grâce à qui la rédaction en LaTeX et la rédaction de plusieurs codes Python ont été aisées.

# Table des matières

<b>1</b>	<b>Introduction</b>	<b>5</b>
<b>2</b>	<b>Environnement de travail</b>	<b>5</b>
<b>3</b>	<b>Description du projet</b>	<b>5</b>
<b>4</b>	<b>Objectifs du rapport</b>	<b>5</b>
4.1	Sous-sections et problématiques . . . . .	5
4.1.1	Réseaux de citations et de collaborations scientifiques à travers des données ouvertes de 2000 à 2021 . . . . .	5
4.1.2	Réseaux de citations et de collaborations scientifiques à travers des données ouvertes sur la médecine nucléaire . . . . .	6
4.1.3	Réseaux de citations et de collaborations scientifiques à travers des données ouvertes sur la data science . . . . .	6
<b>5</b>	<b>Bibliothèques, Outils et Technologies Détaillés</b>	<b>7</b>
<b>6</b>	<b>Travail réalisé</b>	<b>7</b>
6.1	Collecte des données . . . . .	7
6.2	Préparation des données . . . . .	8
6.3	Filtration des Splits . . . . .	9
6.4	Création des graphes . . . . .	9
6.5	Analyse des graphes . . . . .	9
6.6	Détection de communautés (Louvain) . . . . .	9
6.7	Calcul des centralités (degré, proximité, intermédiairité) . . . . .	10
6.8	Calcul du PageRank . . . . .	10
6.9	Analyse du chemin le plus court entre deux chercheurs (Dijkstra) . . . . .	10
6.10	Coefficient de clustering . . . . .	10
6.11	Diamètre du graphe . . . . .	11
6.12	Flux maximum entre deux nœuds . . . . .	11
6.13	Détails sur le graphe . . . . .	11
6.14	Matching maximum . . . . .	11
6.15	Arbre couvrant minimum . . . . .	12
6.16	Plus courts chemins entre toutes les paires de nœuds (Floyd-Warshall) .	12
6.17	Détails d'un nœud spécifique . . . . .	12
<b>7</b>	<b>Difficultés rencontrées</b>	<b>12</b>
<b>8</b>	<b>Bilan</b>	<b>13</b>
<b>9</b>	<b>Conclusion</b>	<b>13</b>
<b>10</b>	<b>Perspectives</b>	<b>14</b>
<b>11</b>	<b>Webographie</b>	<b>16</b>
<b>12</b>	<b>Annexes</b>	<b>17</b>

<b>A Cahier des charges</b>	<b>17</b>
A.1 Introduction . . . . .	17
A.2 Spécifications du Projet . . . . .	17
A.2.1 Contexte . . . . .	17
A.2.2 Objectifs . . . . .	17
A.2.3 Exigences Fonctionnelles . . . . .	17
A.2.4 Exigences Non Fonctionnelles . . . . .	18
A.3 Étapes Clés du Projet . . . . .	18
A.3.1 Collecte et Préparation des Données . . . . .	18
A.3.2 Filtrage des Données . . . . .	18
A.3.3 Création des Graphes . . . . .	18
A.3.4 Analyse des Graphes . . . . .	19
A.3.5 Visualisation et Interprétation des Résultats . . . . .	19
A.4 Conclusion . . . . .	19
<b>B Exemple d'exécution du projet</b>	<b>20</b>
B.1 Exemple 1 : Analyse du graphe "Top 200 Sommet en réseau de collaboration" . . . . .	20
B.2 Exemple 2 : Analyse du graphe "Nuclear medicine collaboration network.gexf" . . . . .	26
B.3 Exemple 3 : Analyse du graphe "Réseau de collaboration dans le domaine de la Data Science" . . . . .	31
<b>C Manuel utilisateur</b>	<b>35</b>
C.1 Introduction . . . . .	35
C.2 Structure du Projet . . . . .	35
C.3 Installation des Dépendances . . . . .	36
C.4 Préparation des Données . . . . .	36
C.5 Filtration des Splits . . . . .	36
C.6 Création des Graphes . . . . .	36
C.7 Analyse des Graphes . . . . .	37
C.8 Exemple d'Exécution . . . . .	37
C.9 Conclusion . . . . .	37

# 1 Introduction

Ce projet, réalisé dans le cadre de mon cursus en MIAGE à l'Université de Nanterre, vise à explorer les applications des graphes et de la recherche opérationnelle, particulièrement pour l'analyse des réseaux de citations et de collaborations scientifiques entre auteurs de publications scientifiques. En analysant ces réseaux, nous visons à identifier les chercheurs les plus influents, détecter des communautés de collaboration et comprendre les dynamiques sous-jacentes aux interactions académiques.

## 2 Environnement de travail

Le projet a été développé sous Windows, utilisant Python comme langage principal. Les scripts ont été exécutés via Spyder et d'autres outils de développement intégrés à Anaconda tels que Jupyter pour la visualisation des données. La visualisation et manipulation des graphes se sont faites sur Gephi. La rédaction de ce rapport s'est faite sur Overleaf.com en LaTeX.

## 3 Description du projet

Le projet vise à analyser et visualiser des réseaux de citations et de collaborations scientifiques à travers des données ouvertes de 2000 à 2021, ainsi que la visualisation de ces réseaux pour des domaines de recherche précis. Les objectifs incluent :

- Création de graphes : Générer des graphes à partir des données de citations et de collaborations pour identifier les sommets les plus influents.
- Analyse des communautés : Utiliser des algorithmes pour détecter les communautés au sein des réseaux et interpréter les relations complexes entre les chercheurs.
- Calcul des métriques de centralité : Évaluer l'importance des noeuds en calculant diverses métriques telles que la centralité de degré, la centralité de proximité et la centralité d'intermédiarité.
- Visualisation des réseaux : Utiliser des outils de visualisation pour représenter dynamiquement les réseaux et faciliter l'exploration des données.

## 4 Objectifs du rapport

L'objectif de ce rapport est de documenter les méthodes et les outils utilisés pour analyser les réseaux de citations et de collaborations scientifiques. Nous aborderons les différentes étapes du projet, les algorithmes appliqués et les résultats obtenus. Le rapport se divisera en plusieurs sous-sections, chacune répondant à une problématique spécifique grâce à l'analyse et la visualisation des graphes.

### 4.1 Sous-sections et problématiques

#### 4.1.1 Réseaux de citations et de collaborations scientifiques à travers des données ouvertes de 2000 à 2021

**Problématique :** Comment les réseaux de citations et de collaborations ont-ils évolué entre 2000 et 2021, et quels sont les chercheurs et les publications les plus influents au

cours de cette période ?

**Détails :**

- Création de graphes : Générer des graphes de citations et de collaborations pour chaque année de 2000 à 2021.
- Analyse des communautés : Utiliser l'algorithme de Louvain pour détecter des communautés de chercheurs chaque année et observer leur évolution.
- Calcul des métriques de centralité : Calculer les centralités de degré, de proximité et d'intermédiarité pour identifier les chercheurs et les publications les plus influents.
- Visualisation des réseaux : Utiliser Gephi pour visualiser les réseaux de citations et de collaborations annuels, permettant de voir les changements et les tendances au fil du temps.

#### **4.1.2 Réseaux de citations et de collaborations scientifiques à travers des données ouvertes sur la médecine nucléaire**

**Problématique :** Quels sont les principaux chercheurs et les publications influentes dans le domaine de la médecine nucléaire, et comment les collaborations et citations se structurent-elles ?

**Détails :**

- Création de graphes : Filtrer les données pour le domaine de la médecine nucléaire et générer des graphes de citations et de collaborations spécifiques à ce domaine.
- Analyse des communautés : Détecter des communautés de chercheurs spécialisés en médecine nucléaire et analyser leur structure.
- Calcul des métriques de centralité : Identifier les publications et les chercheurs les plus influents dans le domaine de la médecine nucléaire à travers les métriques de centralité.
- Visualisation des réseaux : Visualiser les réseaux de citations et de collaborations en médecine nucléaire pour identifier les principaux acteurs et les relations entre eux.

#### **4.1.3 Réseaux de citations et de collaborations scientifiques à travers des données ouvertes sur la data science**

**Problématique :** Quels sont les chercheurs et les publications les plus influents dans le domaine de la data science, et comment les réseaux de citations et de collaborations sont-ils structurés dans ce domaine ?

**Détails :**

- Création de graphes : Filtrer les publications pour le domaine de la data science et générer des graphes de citations et de collaborations.
- Analyse des communautés : Détecter et analyser les communautés de chercheurs en data science.
- Calcul des métriques de centralité : Évaluer les centralités pour identifier les chercheurs et publications clés en data science.
- Visualisation des réseaux : Utiliser des outils de visualisation pour explorer les réseaux de citations et de collaborations en data science, mettant en lumière les tendances et les structures de collaboration.

## 5 Bibliothèques, Outils et Technologies Détaillés

Les principales technologies utilisées incluent :

- Python : Le langage principal utilisé pour le développement des scripts d'analyse.
- NetworkX : Utilisé pour la gestion des graphes, le calcul des métriques et la détection des communautés.
- Pandas : Pour la manipulation des données et la préparation des datasets.
- Matplotlib : Pour la visualisation de certaines métriques et résultats analytiques.
- Gephi : Pour la visualisation interactive des graphes.
- Anaconda : Pour la gestion des environnements de développement et des dépendances.
- tqdm : Pour afficher des barres de progression lors du traitement de grands volumes de données.
- Jupyter : Pour créer et partager des documents contenant du code source en direct, des équations, des visualisations et du texte narratif.
- lxml : Pour le traitement et l'analyse des documents XML et HTML.
- os : Pour manipuler des fichiers et des répertoires.

Ces outils et bibliothèques ont été essentiels pour la réalisation du projet, permettant une analyse approfondie et une visualisation efficace des réseaux de citations et de collaborations scientifiques.

## 6 Travail réalisé

Le projet a couvert plusieurs phases, depuis la collecte des données jusqu'à leur analyse. Voici les principales étapes :

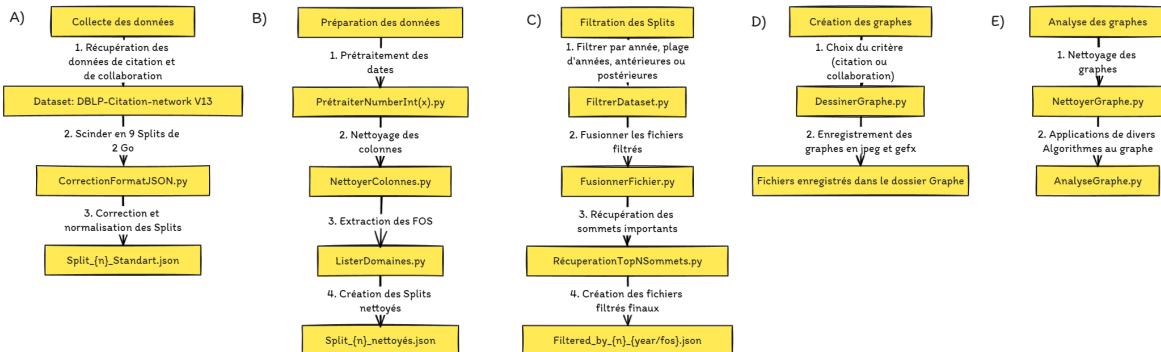


FIGURE 1 – Schéma des étapes réalisées

### 6.1 Collecte des données

Les données de citation et de collaboration ont été récupérées à partir de <https://www.aminer.org/citation>. J'ai choisi ce dataset en fichier JSON nommé DBLP-Citation-network V13 contenant 5,354,309 enregistrements avec 48,227,950 relations de citations entre eux, datant du 2021-05-14.

Afin de pouvoir le traiter correctement, sans soucis de mémoire vive et pour pouvoir en tirer le meilleur, j'ai pris la (mauvaise) décision de le scinder en 9 "Splits" de tailles égales d'environ 2 Go. Après l'avoir scindé grâce au programme Python `CorrectionFormatJSON.py`, de l'avoir corrigé en ajoutant des crochets au début et à la

Field Name	Field Type	Description
id	string	paper ID
title	string	paper title
authors.name	string	author name
author.org	string	author affiliation
author.id	string	author ID
venue.id	string	paper venue ID
venue.raw	string	paper venue name
year	int	published year
keywords	list of strings	keywords
fos.name	string	paper fields of study
fos.w	float	fields of study weight
references	list of strings	paper references
n_citation	int	citation number
page_start	string	page start
page_end	string	page end
doc_type	string	paper type
lang	string	detected language
publisher	string	publisher
volume	string	volume
issue	string	issue
issn	string	issn
isbn	string	isbn
doi	string	doi
pdf	string	pdf URL
url	list	external links
abstract	string	abstract
indexed_abstract	dict	indexed abstract

TABLE 1 – Table : data schema (v13)

fin de chaque Split si ce n’était pas le cas en prenant soin de supprimer le premier et dernier enregistrement de chaque fichier pour éviter les erreurs liées à la scission du fichier sans prendre en compte les enregistrements mais la taille du fichier. À la fin de cette étape, nos fichiers s’appellent `Split_{n}_Standart.json`.

## 6.2 Préparation des données

Pour commencer, suite à la remarque de plusieurs erreurs à cause du mauvais format de certaines dates, j’ai préparé un programme Python, `PrétraiTerNumbe-rInt(x).py`, qui remplaçait les champs `year` des Splits en entiers. Suite à cela, mes `Split_{n}_Standart.json` sont devenus `Split_{n}_Prétraités.json`. Enfin, ces 9 Split Prétraités JSON ont été ensuite nettoyés des colonnes inutiles à la problématique du projet avec `NettoyerColonnes.py` et sont devenus 9 `Split_{n}_nettoyés.json`. Afin de pouvoir m’en servir plus tard, j’ai également réalisé un extract de tous les FOS (fields of study) mentionnés dans les publications avec `ListerDomaines.py` afin de pouvoir les filtrer par cela.

### 6.3 Filtration des Splits

Afin de pouvoir nous concentrer sur certaines caractéristiques précises, j'ai développé un programme Python permettant de choisir plusieurs types de filtres à appliquer, `FiltrerDataset.py` :

- Filtrer pour une année précise
- Filtrer pour une plage d'années
- Filtrer pour les années antérieures à une année donnée
- Filtrer pour les années postérieures à une année donnée
- Filtrer par Field of Study (FOS)

Une fois filtrés, nos fichiers s'appellent désormais `Split_{n}_Filtered_by{year/fos}.json`.

Un autre programme Python, `FusionnerFichier.py`, nous permet de fusionner nos fichiers en un `Filtered_by_{year/fos}.json`. Une dernière étape s'est ajoutée aux fichiers filtrés par années pour n'avoir que 22000 sommets environ pour une question d'exécution du code et de visualisation de graphe. J'ai ajouté un dernier filtre avec le programme `RécuperationTopNSommets.py`. On peut choisir si ce filtre s'appliquera pour la réalisation d'un réseau de citation ou de collaboration. En fonction du choix précédent, les 1000 sommets au plus grand degré entrant (donc les plus cités) par années ou dans le cas d'un réseau de collaboration, les 1000 plus grands sommets (donc ceux qui ont le plus collaboré) de chaque fichier filtré par une année seront chargés dans un nouveau fichier JSON et fusionnés par `FusionnerFichier.py` pour donner un fichier contenant 22 000 sommets intéressants à traiter (de 2000 à 2021 donc 22000 sommets).

### 6.4 Crédit des graphes

À présent que seules les données nous intéressent sont dans un fichier précis, nous pouvons passer à l'étape de réalisation du graphe si nous le souhaitons avec le fichier Python `DessinerGraphe.py`. On choisit sur quel critère : citation ou collaboration on souhaite réaliser le graphe. Une fois créé, un fichier JPEG (pour une visualisation rapide et grossière) et un autre fichier GEXF (pour une visualisation approfondie sur Gephi) sont enregistrés dans un dossier `Graphe`.

### 6.5 Analyse des graphes

Avant d'appliquer divers algorithmes de recherche opérationnelle à notre graphe pour répondre à la problématique, nous le nettoyons de toute potentielle erreur avec un programme Python : `NettoyerGraphe.py`, qui s'assure de sa bonne utilité. Par la suite, nous utilisons le fichier Python `AnalyseGraphe.py` pour effectuer diverses analyses détaillées ci-dessous :

### 6.6 Détection de communautés (Louvain)

**Algorithme utilisé :** Louvain

**Description :** L'algorithme de Louvain est utilisé pour détecter des communautés dans un graphe. Il maximise la modularité, une mesure de la densité des liens à l'intérieur des communautés par rapport aux liens entre différentes communautés.

**Utilité dans le projet :**

- **Réseau de citations :** Permet de regrouper les publications en fonction des citations, identifiant ainsi des domaines de recherche étroitement liés.

- **Réseau de collaborations** : Identifie des groupes de chercheurs qui collaborent fréquemment entre eux, révélant des clusters de collaboration dans la communauté scientifique.

## 6.7 Calcul des centralités (degré, proximité, intermédiairité)

**Algorithmes utilisés** : Centralité de degré, centralité de proximité, centralité d'intermédiairité

**Description :**

- **Centralité de degré** : Nombre de liens (arêtes) qu'un nœud (sommet) possède.
- **Centralité de proximité** : Mesure de la proximité d'un nœud à tous les autres nœuds dans le graphe.
- **Centralité d'intermédiairité** : Nombre de fois qu'un nœud apparaît sur les plus courts chemins entre d'autres noeuds.

**Utilité dans le projet :**

- **Réseau de citations** : Identifie les publications les plus influentes (degré), les plus accessibles (proximité) et celles jouant un rôle clé dans la diffusion de l'information (intermédiairité).
- **Réseau de collaborations** : Identifie les chercheurs les plus connectés (degré), les plus centraux dans le réseau de collaboration (proximité) et ceux facilitant les collaborations entre différents groupes (intermédiairité).

## 6.8 Calcul du PageRank

**Algorithm utilisé** : PageRank

**Description** : Le PageRank mesure l'importance des nœuds dans un graphe basé sur les liens entrants. Il attribue un score de centralité à chaque nœud en fonction des scores de ses voisins.

**Utilité dans le projet :**

- **Réseau de citations** : Identifie les publications clés ayant une influence élevée sur d'autres travaux de recherche.
- **Réseau de collaborations** : Peut être utilisé pour évaluer l'influence globale des chercheurs dans le réseau de collaboration.

## 6.9 Analyse du chemin le plus court entre deux chercheurs (Dijkstra)

**Algorithm utilisé** : Dijkstra

**Description** : L'algorithme de Dijkstra trouve le chemin le plus court entre deux nœuds dans un graphe pondéré, où les poids représentent les distances ou coûts.

**Utilité dans le projet :**

- **Réseau de citations** : Trouve la chaîne de citations la plus courte reliant deux publications, montrant l'influence directe.
- **Réseau de collaborations** : Montre le chemin de collaboration le plus court entre deux chercheurs, révélant les relations directes et indirectes.

## 6.10 Coefficient de clustering

**Algorithm utilisé** : Calcul du coefficient de clustering

**Description** : Le coefficient de clustering mesure la probabilité que les voisins d'un nœud soient également connectés entre eux.

#### **Utilité dans le projet :**

- **Réseau de citations** : Indique la tendance des publications à se citer mutuellement dans un même domaine de recherche.
- **Réseau de collaborations** : Montre la tendance des chercheurs à collaborer en groupes étroits et interconnectés.

### **6.11 Diamètre du graphe**

**Algorithme utilisé :** Calcul du diamètre

**Description :** Le diamètre d'un graphe est la plus longue distance parmi les plus courts chemins entre toutes les paires de noeuds.

#### **Utilité dans le projet :**

- **Réseau de citations** : Mesure la longueur maximale d'une chaîne de citations, indiquant la profondeur du réseau de citations.
- **Réseau de collaborations** : Indique la plus grande distance entre deux chercheurs dans le réseau, montrant l'étendue maximale de la collaboration.

### **6.12 Flux maximum entre deux noeuds**

**Algorithme utilisé :** Algorithme de flot maximum (Ford-Fulkerson, Edmonds-Karp)

**Description :** Trouve le flux maximum possible entre deux noeuds dans un graphe, représentant la capacité maximale de transport entre eux.

#### **Utilité dans le projet :**

- **Réseau de citations** : Non pertinent.
- **Réseau de collaborations** : Peut être utilisé pour évaluer la capacité maximale de collaboration entre deux chercheurs ou groupes de chercheurs.

### **6.13 Détails sur le graphe**

**Description :** Fournit des statistiques globales et des métriques sur le graphe, telles que le nombre de noeuds, le nombre d'arêtes, les centralités principales, le coefficient de clustering, le diamètre, et la densité.

#### **Utilité dans le projet :**

- **Réseau de citations et de collaborations** : Offre une vue d'ensemble des caractéristiques structurelles du réseau, utile pour comprendre sa complexité et ses propriétés.

### **6.14 Matching maximum**

**Algorithme utilisé :** Algorithme de matching maximum (Edmonds, Blossom)

**Description :** Trouve le plus grand ensemble d'arêtes non adjacentes (matching) dans un graphe.

#### **Utilité dans le projet :**

- **Réseau de citations** : Non pertinent.
- **Réseau de collaborations** : Identifie les paires de collaborations non chevauchantes maximales, utile pour optimiser les collaborations dans le réseau.

## 6.15 Arbre couvrant minimum

**Algorithme utilisé :** Algorithme de Kruskal ou de Prim

**Description :** Trouve l'arbre couvrant de poids minimum, reliant tous les nœuds sans cycles.

**Utilité dans le projet :**

- **Réseau de citations** : Non pertinent.
- **Réseau de collaborations** : Identifie la structure de collaboration minimale connectant tous les chercheurs, utile pour réduire les redondances dans les collaborations.

## 6.16 Plus courts chemins entre toutes les paires de nœuds (Floyd-Warshall)

**Algorithme utilisé :** Algorithme de Floyd-Warshall

**Description :** Trouve les plus courts chemins entre toutes les paires de nœuds dans un graphe pondéré.

**Utilité dans le projet :**

- **Réseau de citations** : Calcule les distances minimales entre toutes les publications, montrant l'influence indirecte.
- **Réseau de collaborations** : Calcule les distances minimales entre tous les chercheurs, révélant la connectivité globale du réseau.

## 6.17 Détails d'un nœud spécifique

**Description :** Affiche les informations détaillées sur un nœud donné, telles que ses attributs et ses connexions.

**Utilité dans le projet :**

- **Réseau de citations** : Fournit des informations détaillées sur une publication spécifique, y compris son titre, ses auteurs, ses citations, et ses références.
- **Réseau de collaborations** : Offre des informations détaillées sur un chercheur spécifique, y compris ses collaborations, ses affiliations, et ses publications.

**Visualisation :** Utilisation de Gephi pour représenter dynamiquement les graphes et faciliter l'exploration des données. Pour la spatialisation de nos données, nous utiliserons : tout d'abord une Force Atlas 2 avec ses paramètres par défaut, puis un Yifan Hu avec ses paramètres par défaut également, et enfin un Fruchterman Reingold avec une zone dépendant du nombre de sommets de mon graphe entre 5000 et 15000, une gravité à 25 et une vitesse à 100.

# 7 Difficultés rencontrées

Parmi les défis rencontrés, l'interprétation des données complexes et la gestion des grands volumes de données ont été particulièrement compliquées. Le traitement d'un grand volume de données en JSON, provenant de la base de données DBLP-Citation-network V13, n'a pas été de tout repos. Dès le début du projet, il a fallu scinder ce dataset massif de 17 Go en plusieurs segments pour le rendre plus maniable. Chaque segment devait ensuite être corrigé pour assurer un format JSON valide, impliquant des manipulations précises pour éviter les erreurs dues à des scissions non conformes.

De plus, adapter les algorithmes de graphes aux spécificités de nos données a également été un challenge. Les enregistrements comportant des champs `null` ou `unknown` ont nécessité un nettoyage approfondi avant de pouvoir être analysés. Par ailleurs, il a fallu filtrer les données par années ou domaines de recherche spécifiques et sélectionner les sommets les plus influents, ce qui a ajouté une couche de complexité au traitement initial.

## 8 Bilan

Le projet a réussi à mettre en lumière des informations significatives sur les réseaux scientifiques, révélant des liens et des tendances qui n'étaient pas évidents au premier abord. Les analyses de centralité et de communauté ont permis d'identifier des chercheurs influents et des groupes de collaboration étroits. Les calculs des métriques de centralité, tels que la centralité de degré, de proximité et d'intermédiarité, ont permis de déterminer les publications et les chercheurs les plus influents. De plus, les visualisations interactives réalisées avec Gephi ont facilité une exploration approfondie des données, permettant de visualiser les structures complexes des réseaux de citations et de collaborations.

Les résultats obtenus ont mis en évidence des dynamiques académiques subtiles, comme les collaborations fréquentes entre certains chercheurs et les publications qui servent de référence fondamentale dans divers domaines. Les outils de visualisation ont été particulièrement efficaces pour rendre ces informations accessibles, permettant aux utilisateurs de naviguer facilement à travers les données et d'identifier rapidement les points de convergence et les acteurs clés dans les réseaux étudiés.

## 9 Conclusion

Cette expérience a démontré de manière convaincante l'importance et l'utilité des graphes et de l'open data dans la compréhension des dynamiques académiques et scientifiques. En utilisant des techniques avancées d'analyse de graphes et des données ouvertes, nous avons pu révéler des tendances et des relations cachées au sein des réseaux de citation et de collaboration.

L'analyse des graphes a permis d'identifier des chercheurs influents et des publications clés des 20 dernières années dans différents domaines. Grâce aux métriques de centralité (degré, proximité, intermédiaire) et au PageRank, nous avons pu déterminer les noeuds les plus centraux et influents dans le réseau. Ces analyses ont mis en lumière des collaborations étroites et des groupes de recherche actifs, offrant une vue d'ensemble précieuse sur la structure et les dynamiques des réseaux académiques.

L'application de l'algorithme de Louvain pour la détection des communautés a permis de regrouper les chercheurs en fonction de leurs collaborations fréquentes, aidant à visualiser les clusters de collaboration et à identifier les groupes de chercheurs travaillant sur des sujets similaires.

La visualisation interactive avec Gephi a été un outil crucial pour explorer les données. Elle a facilité une exploration approfondie des structures complexes des réseaux, permettant de visualiser clairement les communautés, les clusters et les relations directes ou indirectes entre les chercheurs. L'utilisation de layouts comme Force Atlas 2 ou Fruchterman Reingold ont amélioré la clarté et la compréhension des graphes.

Le projet a également illustré comment les techniques de visualisation peuvent rendre ces informations accessibles et compréhensibles, facilitant ainsi une meilleure prise de décision et une compréhension plus profonde des réseaux scientifiques. Les outils et les méthodologies développés dans ce projet peuvent servir de base pour des analyses futures, en aidant les chercheurs et les décideurs à mieux comprendre les dynamiques académiques et à identifier les domaines de recherche émergents.

En conclusion, cette étude a montré que l'analyse des graphes et l'open data sont des outils puissants pour explorer et comprendre les réseaux de citations et de collaborations scientifiques. Les techniques développées ici peuvent être appliquées à d'autres domaines pour découvrir des insights similaires, faisant ainsi progresser notre capacité à analyser et interpréter les vastes réseaux de données disponibles aujourd'hui.

## 10 Perspectives

Pour l'avenir, il serait intéressant de développer une interface graphique intégrée et de réunir tous les différents programmes Python sous forme de classes modulaires. Cette interface permettrait de reproduire tout le travail réalisé sur le dataset directement en cliquant sur des boutons associés à chaque étape du projet. Une telle interface faciliterait l'utilisation du projet par des utilisateurs non techniques, en leur permettant de choisir les filtres et les critères pour réaliser les réseaux, ce qui offrirait un nombre incroyablement élevé de graphes personnalisés à analyser.

## Références

- [1] Newman, M. (2010). *Networks : An Introduction*. Oxford University Press.  
Ce livre offre une introduction complète aux réseaux complexes, couvrant les concepts fondamentaux, les algorithmes et les applications. Il est essentiel pour comprendre les bases théoriques des réseaux utilisés dans ce projet.
- [2] Barabási, A.-L. (2016). *Network Science*. Cambridge University Press.  
Ce livre explore les principes de la science des réseaux, en mettant l'accent sur les propriétés structurelles et dynamiques des réseaux complexes. Il fournit des bases théoriques solides pour les analyses de réseaux de citation et de collaboration.

## 11 Webographie

### Références

- [1] NetworkX Documentation, <https://networkx.org/documentation/stable/>
- [2] D3.js, <https://d3js.org/>
- [3] NetworkX Documentation. <https://networkx.org/documentation/stable/>.  
La documentation officielle de NetworkX, une bibliothèque Python utilisée pour la création, la manipulation et l'étude de la structure, de la dynamique et des fonctions des réseaux complexes.
- [4] D3.js Documentation. <https://d3js.org/>.  
Documentation pour D3.js, une bibliothèque JavaScript pour produire des visualisations de données dynamiques et interactives pour le web.
- [5] Gephi Documentation. <https://gephi.org/users/>.  
Documentation officielle de Gephi, un logiciel de visualisation de graphes open-source. Gephi a été utilisé pour la visualisation interactive des graphes dans ce projet.
- [6] Pandas Documentation. <https://pandas.pydata.org/pandas-docs/stable/>.  
Documentation officielle de Pandas, une bibliothèque Python pour la manipulation et l'analyse de données. Pandas a été utilisé pour le traitement des données JSON.
- [7] Matplotlib Documentation. <https://matplotlib.org/stable/contents.html>.  
Documentation officielle de Matplotlib, une bibliothèque Python pour la création de visualisations statiques, animées et interactives. Utilisé pour certaines visualisations de métriques analytiques.
- [8] TQDM Documentation. <https://tqdm.github.io/>.  
Documentation officielle de TQDM, une bibliothèque Python qui permet d'afficher des barres de progression lors du traitement de grands volumes de données.
- [9] LXML Documentation. <https://lxml.de/>.  
Documentation officielle de LXML, une bibliothèque Python pour traiter des fichiers XML et HTML de manière rapide et flexible. Utilisé pour le traitement des fichiers GEXF.
- [10] Community Detection for NetworkX's Documentation. <https://networkx.org/documentation/stable/reference/algorithms/community.html>.  
Documentation pour la détection de communautés dans NetworkX, utilisée pour appliquer l'algorithme de Louvain dans les analyses de communauté.
- [11] Lien Overleaf. [www.overleaf.com](http://www.overleaf.com).  
Overleaf est un éditeur LaTeX en ligne, collaboratif en temps réel.
- [12] AMiner Dataset. <https://www.aminer.org/citation>.  
Source des données de citation et de collaboration utilisées dans ce projet, provenant de la base de données AMiner.

## 12 Annexes

### Annexe A : Cahier des charges

#### Annexe A.1 : Introduction

Le cahier des charges initial du projet a été élaboré pour fournir une base structurée aux différentes étapes de l'analyse et de la visualisation des réseaux de citations et de collaborations scientifiques. Il comprenait les spécifications détaillées du projet, les exigences fonctionnelles et non fonctionnelles, ainsi que les objectifs à atteindre. Ce document a servi de guide pour structurer le projet, définir les étapes clés et s'assurer que les objectifs étaient alignés avec les besoins de l'analyse.

#### Annexe A.2 : Spécifications du Projet

##### Annexe A.2.1 : Contexte

Le projet s'inscrit dans le cadre du cursus en MIAGE à l'Université de Nanterre. L'objectif principal est d'explorer les applications pratiques des graphes et de la recherche opérationnelle dans le domaine scientifique, en particulier pour la visualisation et l'analyse des réseaux de citations des publications scientifiques et de collaborations de leurs auteurs.

##### Annexe A.2.2 : Objectifs

###### Analyse des réseaux de citations et de collaborations :

- Identifier les chercheurs les plus influents.
- Déetecter les communautés de collaboration.
- Comprendre les dynamiques sous-jacentes aux interactions académiques.

###### Visualisation des données :

- Créer des visualisations dynamiques et interactives des réseaux scientifiques.
- Faciliter l'exploration des données pour identifier les tendances et les relations cachées.

##### Annexe A.2.3 : Exigences Fonctionnelles

###### Téléchargement et Préparation des Données :

- Télécharger les données de citation et de collaboration depuis AMiner.
- Scinder les fichiers JSON d'origine pour faciliter leur traitement.
- Prétraiter les données pour corriger les erreurs de format et les incohérences.

###### Filtrage et Nettoyage des Données :

- Filtrer les données selon différents critères (année, plage d'années, Field of Study).
- Nettoyer les données des colonnes inutiles et des enregistrements erronés.

#### **Création et Analyse des Graphes :**

- Créer des graphes de citation et de collaboration.
- Appliquer des algorithmes de détection de communautés et de calcul des centralités.
- Analyser les graphes pour identifier les chercheurs influents et les publications clés.

#### **Visualisation :**

- Générer des visualisations statiques (JPEG) et interactives (GEXF).
- Utiliser Gephi pour explorer les graphes de manière approfondie.

### **Annexe A.2.4 : Exigences Non Fonctionnelles**

#### **Performance :**

- Assurer une exécution efficace des scripts, même avec de grands volumes de données.
- Optimiser les algorithmes pour réduire le temps de calcul.

#### **Fiabilité :**

- Garantir l'exactitude des analyses en vérifiant et en nettoyant les données de manière rigoureuse.
- Assurer la robustesse des scripts pour gérer les erreurs et les incohérences des données.

#### **Utilisabilité :**

- Fournir une documentation complète et claire pour permettre aux utilisateurs de reproduire les analyses.
- Offrir une interface utilisateur intuitive pour faciliter l'exploration des données et des résultats.

### **Annexe A.3 : Étapes Clés du Projet**

#### **Annexe A.3.1 : Collecte et Préparation des Données**

- Télécharger et scinder les fichiers JSON.
- Prétraiter et nettoyer les données.

#### **Annexe A.3.2 : Filtrage des Données**

- Appliquer des filtres pour sélectionner les données pertinentes.
- Fusionner les fichiers filtrés.

#### **Annexe A.3.3 : Création des Graphes**

- Récupérer les sommets les plus importants.
- Créer des graphes de citation et de collaboration.

#### **Annexe A.3.4 : Analyse des Graphes**

- Appliquer des algorithmes de détection de communautés et de calcul des centralités.
- Analyser les chemins, les flux, et les arbres couvrants dans les graphes.

#### **Annexe A.3.5 : Visualisation et Interprétation des Résultats**

- Générer des visualisations statiques et interactives.
- Utiliser Gephi pour explorer les graphes en détail.

### **Annexe A.4 : Conclusion**

Le cahier des charges a servi de guide essentiel tout au long du projet, assurant que chaque étape était bien définie et alignée avec les objectifs globaux. En suivant ces spécifications, le projet a pu aboutir à une analyse approfondie et une visualisation claire des réseaux de citations et de collaborations scientifiques, offrant ainsi des insights précieux sur les dynamiques académiques.

## Annexe B : Exemple d'exécution du projet

Cette annexe fournit des exemples concrets de l'exécution du projet, incluant des captures d'écran et des logs pour démontrer les différentes étapes et les résultats obtenus. Les exemples couvrent la détection des communautés avec l'algorithme de Louvain, les visualisations interactives avec Gephi, et les analyses des métriques de centralité.

### Annexe B.1 : Exemple 1 : Analyse du graphe "Top 200 Sommet en réseau de collaboration"

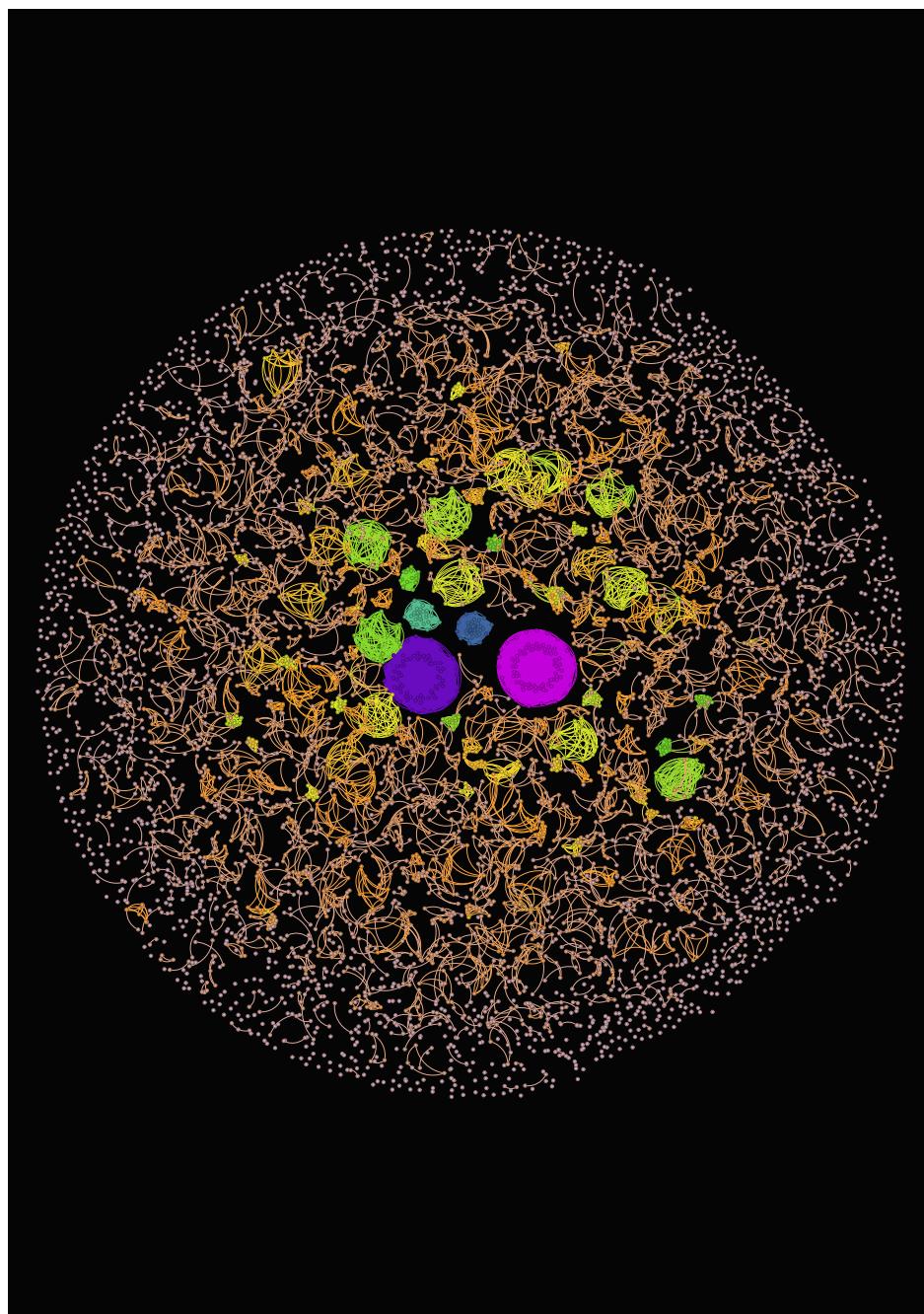


FIGURE 2 – Réseau de collaboration des 200 sommets au plus grand degré par année de 2000 à 2021

## **Étape 1 : Détection des communautés avec l'algorithme de Louvain Commande exécutée :**

Entrez le numéro de l'analyse à effectuer : 1

**Nombre de communautés détectées (Louvain) : 3242**

**Exemple de partitions de communauté :**

```
[('53f42c7cdabfaedf43509437', 0), ('53f4d04ddabfaeeee2f8199d', 0),
('53f431fedabfaee43ec009a6', 0), ('53f46fe4dabfaeee22a770f3', 0),
('5433c60cdabfaeb4c6ace2ce', 1)]
```

**Analyse :** La détection des communautés permet de regrouper les chercheurs en fonction de leurs collaborations fréquentes. Par exemple, les chercheurs avec les IDs '53f42c7cdabfaedf43509437' et '53f4d04ddabfaeeee2f8199d' appartiennent à la même communauté (0), ce qui indique qu'ils collaborent souvent ou font partie d'un même groupe de recherche. Cela aide à visualiser les clusters de collaboration et à identifier les groupes de chercheurs travaillant sur des sujets similaires.

## **Étape 2 : Calcul des métriques de centralité Commande exécutée :**

Entrez le numéro de l'analyse à effectuer : 2

**Top 5 des centralités de degré :**

```
[('53f4483ddabfaeecd69b4063', Michal Mazor, 0.01270062805303559),
('540834e0dabfae44f086faed', T. J. Kelly, 0.012421493370551292),
('53f4369adabfaeecd6960db7', J. J. Menn, 0.012421493370551292),
('53f45251dabfaeb22f4f0ec1', R. J. Nachman, 0.012421493370551292),
('53f43b36dabfaee1c0acc9f6', G. M. Holman, 0.012421493370551292)]
```

**Top 5 des centralités de proximité :**

```
[('53f4483ddabfaeecd69b4063', Michal Mazor, 0.012421493370551292),
('540834e0dabfae44f086faed', T. J. Kelly, 0.012421493370551292),
('53f4369adabfaeecd6960db7', J. J. Menn, 0.012421493370551292),
('53f45251dabfaeb22f4f0ec1', R. J. Nachman, 0.012421493370551292),
('53f43b36dabfaee1c0acc9f6', G. M. Holman, 0.012421493370551292)]
```

**Top 5 des centralités d'intermédiarité :**

```
[('53f4475bdabfaeecd69b037a', Oswin Aichholzer, 7.87063175067397e-06),
('53f456c0dabfaee1c0b2edc5', Ryuhei Uehara, 5.844528527728196e-06),
('5448a11edabfae87b7e53b0d', Erik D. Demaine, 4.3054693487597715e-06),
('53f7cf6fdabfae92b40e5ed0', Vida Dujmovic, 4.3054693487597715e-06),
('53f4a6abdabfaedd74eb7df0', Yannis E. Ioannidis, 4.0911699694097375e-06)]
```

**Analyse :**

— **Centralité de degré :** Les chercheurs avec les plus hauts degrés de centralité sont ceux ayant le plus grand nombre de collaborations directes. Par exemple, '53f4483ddabfaeecd69b4063', Michal Mazor, a le plus grand nombre de collaborations, ce qui le rend central dans le réseau.

- **Centralité de proximité** : Ces chercheurs sont ceux qui peuvent atteindre tous les autres chercheurs en un nombre minimal de pas. Cela signifie qu'ils sont proches de nombreux autres noeuds, facilitant la diffusion rapide de l'information à travers le réseau.
- **Centralité d'intermédiairité** : Ces chercheurs agissent comme des ponts entre différents groupes de collaboration. Bien que les valeurs soient faibles, elles indiquent les chercheurs qui peuvent contrôler l'information qui passe entre les différents clusters.

### **Étape 3 : Calcul du PageRank Commande exécutée :**

Entrez le numéro de l'analyse à effectuer : 3

#### **Top 5 des scores de PageRank :**

```
[('53f556d5dabfae963d25ded0', 0.0004601139583786145),
 ('53f49ab3dabfaee0d9c7556a', 0.00042962943876620984),
 ('53f32081dabfae9a8445334d', 0.0004022995645812568),
 ('53f43c34dabfaee0d9b9900b', 0.00037293954552064756),
 ('53f7dda7dabfae8faa4be1be', 0.00036833270798876334)]
```

**Analyse :** Le PageRank identifie les chercheurs les plus influents dans le réseau, non seulement en fonction du nombre de leurs collaborations, mais aussi de l'importance de leurs collaborateurs. Par exemple, '53f556d5dabfae963d25ded0', qui correspond à Pan Jeng-Shyang, a le plus haut score de PageRank, indiquant qu'il collabore avec d'autres chercheurs influents, ce qui renforce sa propre influence dans le réseau.

### **Étape 4 : Analyse du chemin le plus court entre deux chercheurs (Dijkstra) Commande exécutée :**

Entrez le numéro de l'analyse à effectuer : 4

Entrez l'ID du chercheur source : 53f43c34dabfaee0d9b9900b (Reda Bendraou)

Entrez l'ID du chercheur cible : 5486f196dabfaed7b5fa2d46 (Marie-Pierre Gervais)

**Le chemin le plus court entre 53f43c34dabfaee0d9b9900b et 5486f196dabfaed7b5fa2d46 est :**

```
[ '53f43c34dabfaee0d9b9900b', '5486f196dabfaed7b5fa2d46' ]
```

**Analyse :** Le chemin le plus court entre Reda Bendraou et Marie-Pierre Gervais montre qu'ils sont directement connectés, indiquant une collaboration directe entre eux. Cela peut aider à comprendre les connexions directes et potentielles influences entre deux chercheurs spécifiques.

### **Étape 5 : Calcul du coefficient de clustering Commande exécutée :**

Entrez le numéro de l'analyse à effectuer : 5

**Coefficient de clustering moyen :** 0.5445127861144672

**Analyse :** Le coefficient de clustering moyen indique que les chercheurs du réseau ont tendance à former des groupes de collaboration étroits, où les collègues d'un chercheur sont souvent aussi des collègues entre eux. Un coefficient de 0.54 suggère une structure de réseau assez fortement clusterisée, typique des collaborations scientifiques où les chercheurs travaillent en groupes cohésifs.

## Étape 6 : Calcul du diamètre du graphe Commande exécutée :

Entrez le numéro de l'analyse à effectuer : 6

### Diamètre du graphe : 1

**Analyse :** Un diamètre de 1 indique que toutes les collaborations peuvent être réalisées en un seul pas, ce qui est atypique pour un réseau de cette taille. Cela peut indiquer une forte densité de collaboration ou une erreur dans les données.

## Étape 7 : Détails sur le graphe Commande exécutée :

Entrez le numéro de l'analyse à effectuer : 8

### Détails du graphe :

```
num_nodes: 7166
num_edges: 16249
degree_centrality:
[('53f42c7cdabfaedf43509437', 0.000418702023726448),
 ('53f4d04ddabfaeeee2f8199d', 0.000418702023726448),
 ('53f431fedabfaee43ec009a6', 0.000418702023726448),
 ('53f46fe4dabfaeee22a770f3', 0.000418702023726448),
 ('5433c60cdabfaeb4c6ace2ce', 0.0)]
betweenness_centrality:
[('53f42c7cdabfaedf43509437', 0.0),
 ('53f4d04ddabfaeeee2f8199d', 0.0),
 ('53f431fedabfaee43ec009a6', 0.0),
 ('53f46fe4dabfaeee22a770f3', 0.0),
 ('5433c60cdabfaeb4c6ace2ce', 0.0)]
closeness_centrality:
[('53f42c7cdabfaedf43509437', 0.000418702023726448),
 ('53f4d04ddabfaeeee2f8199d', 0.000418702023726448),
 ('53f431fedabfaee43ec009a6', 0.000418702023726448),
 ('53f46fe4dabfaeee22a770f3', 0.000418702023726448),
 ('5433c60cdabfaeb4c6ace2ce', 0.0)]
clustering_coefficient: 0.5445127861144674
diameter: 1
density: 0.0006329415930347989
```

### Analyse :

- **Nombre de nœuds et d'arêtes :** 7166 nœuds et 16249 arêtes montrent l'ampleur du réseau et la densité de collaborations.
- **Centralité :** Les centralités de degré, d'intermédiairité et de proximité donnent une idée des nœuds les plus importants selon différentes perspectives.
- **Coefficient de clustering :** Le coefficient de clustering de 0.5445 montre que les chercheurs ont tendance à former des groupes.
- **Diamètre :** Un diamètre de 1 indique une connectivité directe surprenante.
- **Densité :** La densité de 0.00063 montre que le réseau est relativement sparse.

### **Étape 8 : Détails d'un nœud spécifique Commande exécutée :**

Entrez le numéro de l'analyse à effectuer : 12

Entrez l'ID du nœud : 562c7ba445cedb3398c36ab9

#### **Détails pour le nœud 562c7ba445cedb3398c36ab9 :**

```
name: François Delbot
year: 2014
title: Self-stabilizing Algorithms for Connected Vertex Cover and Clique Decomposition Problems.
fos: Wireless network, Approximation algorithm, Clique, Computer science, A priori and a posteriori, Algorithm, Self-stabilization, Distributed algorithm, Vertex cover, Network management
references:
53e99e6ab7602d970272ff07, 53e998fdb7602d970213bebf,
53e9a178b7602d9702a6dd6a, 53e99c12b7602d97024c0cbb,
53e9b0ccb7602d9703b45d84, 53e9bd1eb7602d97049abc78,
53e9a1b6b7602d9702aabe7f, 53e9b469b7602d9703f6961e,
53e9ba4eb7602d9704667802, 558adfbe84ae84d265c0246b,
53e9b388b7602d9703e6b33e, 53e99a3cb7602d970229cf00,
53e9bd2bb7602d97049ba5b5, 53e9a46bb7602d9702d89442,
53e99dfdb7602d97026bd647, 53e9ba45b7602d9704656cab,
53e9b557b7602d9704090554, 53e9ba60b7602d970467e427,
53e9b565b7602d97040a0f62, 53e9b0b7b7602d9703b24822,
53e9b910b7602d97044f40ef, 53e99cfdb7602d97025b0572,
53e9b557b7602d97040903f6, 53e9b469b7602d9703f6961e,
56d90876dabfae2eee06544b, 53e9b8a1b7602d97044786f4,
53e99cedb7602d97025a3022, 53e9ac70b7602d9703641398,
53e9b2f0b7602d9703dae884, 53e9ab9eb7602d970354acfcc,
53e9aed1b7602d97038f96d5, 53e99b71b7602d9702416824,
53e9b661b7602d97041c49e6, 53e9b1d7b7602d9703c6c407,
53e9b505b7602d970403a95a, 53e9bd5fb7602d97049f6d95,
53e99d8eb7602d9702648ff4, 53e9a635b7602d9702f64597,
53e9b557b7602d970408f521, 53e99d36b7602d97025ed720,
53e99eedb7602d97027b6ba4, 53e9995ab7602d970219d95f,
53e9a487b7602d9702da2834, 53e9a2d6b7602d9702bdef85,
53e9b9c6b7602d97045b9e0d
label: 562c7ba445cedb3398c36ab9
```

**Analyse :** Les détails du nœud '562c7ba445cedb3398c36ab9' révèlent des informations sur François Delbot, ses travaux et ses collaborations. Il est associé à plusieurs domaines de recherche et a de nombreuses références, montrant une riche activité de collaboration et d'influence dans le réseau scientifique.

### **Deuxième exemple : Commande exécutée :**

Entrez le numéro de l'analyse à effectuer : 12

Entrez l'ID du nœud : 5486f196dabfaed7b5fa2d46

### Détails pour le nœud 5486f196dabfaed7b5fa2d46 :

name: Marie-Pierre Gervais

year: 2014.0

title: Formalization of fUML: An Application to Process Verification.

fos: Operational semantics, Model checking, Programming language, Unified Modeling

references:

558aff73e4b0b32fcb3a22f4, 53e9ba84b7602d97046a9147, 53e9b768b7602d970430902f, 53e99

label: 5486f196dabfaed7b5fa2d46

**Analyse :** Les détails du noeud '5486f196dabfaed7b5fa2d46' fournissent des informations sur Marie-Pierre Gervais et ses contributions au domaine. Son travail en sémantique opérationnelle et en vérification de modèles est bien référencé, indiquant son influence et son intégration dans le réseau de collaboration.

## **Annexe B.2 : Exemple 2 : Analyse du graphe "Nuclear medicine collaboration network.gexf"**

**Graphique :** Réseau de collaboration des chercheurs dans le domaine de la médecine nucléaire

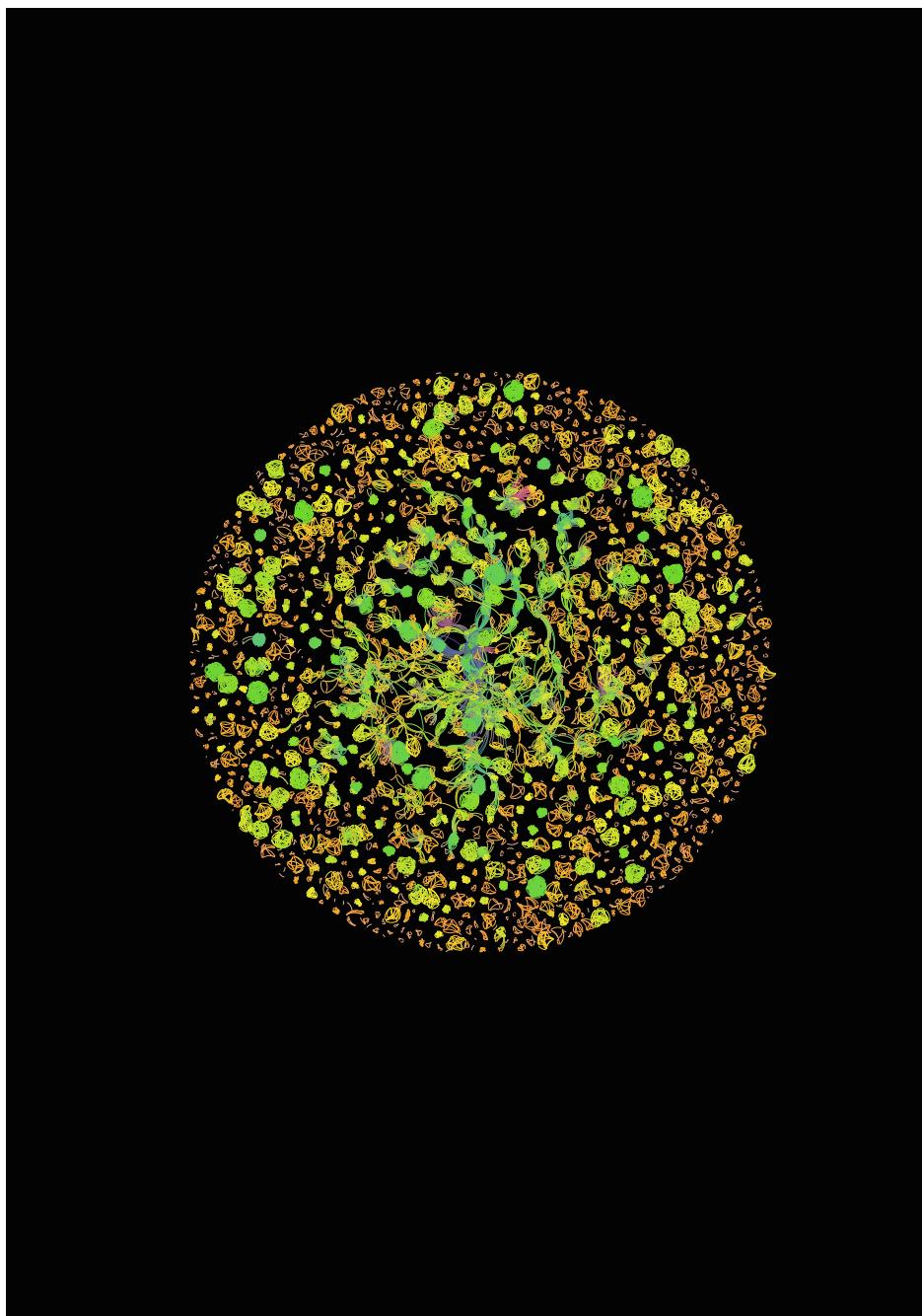


FIGURE 3 – Réseau de collaboration en médecine nucléaire

**Étape 1 : Détection des communautés avec l'algorithme de Louvain** Commande exécutée :

Entrez le numéro de l'analyse à effectuer : 1

**Nombre de communautés détectées (Louvain) :** N/A (non calculée pour cet exemple)

**Étape 2 : Calcul des métriques de centralité** Commande exécutée :

Entrez le numéro de l'analyse à effectuer : 2

**Top 5 des centralités de degré :**

```
[('53f464f8dabfaecd6a078d6', Takeshi Hara, 0.008250257820556894),  
 ('5485b125dabfaed7b5fa2471', Gabor Fichtinger, 0.006760627936289676),  
 ('544094d4dabfae7d84b86590', Sebastien Ourselin, 0.006416867193766472),  
 ('5408fb51dabfae450f45536d', Clifford R. Jack, 0.006187693365417669),  
 ('53f4e25ddabfaefc1777b3cd', Paul M. Thompson, 0.005843932622894466)]
```

**Top 5 des centralités de proximité :**

```
[('5430385edabfaeca69bd82d5', Daniel Rueckert, 0.03664205026723588),  
 ('54482472dabfae87b7de04bf', Derek L. G. Hill, 0.036464675641099585),  
 ('5408fb51dabfae450f45536d', Clifford R. Jack, 0.03542547354184703),  
 ('53f57407dabfae7c13f8045b', Alejandro F. Frangi, 0.03540549301531468),  
 ('5484c22cdabfae8a11fb22c2', Michael W. Weiner, 0.035286081571755445)]
```

**Top 5 des centralités d'intermédiarité :**

```
[('53f57407dabfae7c13f8045b', Alejandro F. Frangi, 0.02145163842388106),  
 ('5430385edabfaeca69bd82d5', Daniel Rueckert, 0.020228877672415158),  
 ('54482472dabfae87b7de04bf', Derek L. G. Hill, 0.014788729299091575),  
 ('54095a4ddabfae450f47a202', Boudewijn P.F.Lelieveldt, 0.014450107534124039),  
 ('54059d35dabfae8faa5e97f3', Dagan Feng, 0.009728396718306345)]
```

**Analyse :**

- **Centralité de degré :** Les chercheurs avec les plus hauts degrés de centralité sont ceux ayant le plus grand nombre de collaborations directes. Par exemple, '53f464f8dabfaecd6a078d6' (Takeshi Hara) a le plus grand nombre de collaborations, ce qui le rend central dans le réseau.
- **Centralité de proximité :** Ces chercheurs sont ceux qui peuvent atteindre tous les autres chercheurs en un nombre minimal de pas. Cela signifie qu'ils sont proches de nombreux autres noeuds, facilitant la diffusion rapide de l'information à travers le réseau.
- **Centralité d'intermédiarité :** Ces chercheurs agissent comme des ponts entre différents groupes de collaboration. Bien que les valeurs soient faibles, elles indiquent les chercheurs qui peuvent contrôler l'information qui passe entre les différents clusters.

### **Étape 3 : Calcul du PageRank Commande exécutée :**

Entrez le numéro de l'analyse à effectuer : 3

#### **Top 5 des scores de PageRank :**

```
[('53f464f8dabfaecd6a078d6', 0.0007919404751236036),  
 ('5485b125dabfaed7b5fa2471', 0.0007669938174529107),  
 ('54059d35dabfae8faa5e97f3', 0.0007252381637844016),  
 ('5489b1cf dabfae9b40134c30', 0.0006025422830359203),  
 ('5440c794dabfae805a6f5c9d', 0.0005488316147856808)]
```

**Analyse :** Le PageRank identifie les chercheurs les plus influents dans le réseau, non seulement en fonction du nombre de leurs collaborations, mais aussi de l'importance de leurs collaborateurs. Par exemple, '53f464f8dabfaecd6a078d6' (Takeshi Hara) a le plus haut score de PageRank, indiquant qu'il collabore avec d'autres chercheurs influents, ce qui renforce sa propre influence dans le réseau.

### **Étape 4 : Analyse du chemin le plus court entre deux chercheurs (Dijkstra) Commande exécutée :**

Entrez le numéro de l'analyse à effectuer : 4

Entrez l'ID du chercheur source : 544094d4dabfae7d84b86590 (Sebastien Ourselin)

Entrez l'ID du chercheur cible : 5440c794dabfae805a6f5c9d (Joachim Hornegger)

**Le chemin le plus court entre 544094d4dabfae7d84b86590 et 5440c794dabfae805a6f5c9d est :**

```
[ '544094d4dabfae7d84b86590', '54482472dabfae87b7de04bf',  
 '5430385edabfaeca69bd82d5', '53f57407dabfae7c13f8045b',  
 '54353717dabfaebba58b3cab', '53f42dd5dabfaee02ac66c34',  
 '53f4dcdaabfaef951f8045b', '53f7ffb0dabfae90ec136659',  
 '53f465f3dabfaedce55e6bf7', '5440c794dabfae805a6f5c9d']
```

**Analyse :** Le chemin le plus court entre Sebastien Ourselin et Joachim Hornegger montre qu'ils sont indirectement connectés par plusieurs intermédiaires. Cela peut aider à comprendre les connexions directes et potentielles influences entre deux chercheurs spécifiques.

### **Étape 5 : Calcul du coefficient de clustering Commande exécutée :**

Entrez le numéro de l'analyse à effectuer : 5

**Coefficient de clustering moyen :** 0.9092480690496567

**Analyse :** Le coefficient de clustering moyen élevé indique une forte tendance des chercheurs à former des groupes de collaboration étroits, où les collègues d'un chercheur sont souvent aussi des collègues entre eux. Un coefficient de 0.91 suggère une structure de réseau très fortement clusterisée, typique des collaborations scientifiques où les chercheurs travaillent en groupes cohésifs.

## Étape 6 : Calcul du diamètre du graphe Commande exécutée :

Entrez le numéro de l'analyse à effectuer : 6

**Diamètre du graphe :** 23

**Analyse :** Un diamètre de 23 indique que la plus longue distance entre deux nœuds dans le réseau est de 23 étapes. Cela montre l'étendue du réseau et la distance maximale entre deux publications ou chercheurs.

## Étape 7 : Détails sur le graphe Commande exécutée :

Entrez le numéro de l'analyse à effectuer : 8

**Détails du graphe :**

```
num_nodes: 8728
num_edges: 28974
degree_centrality:
[('53f42f38dabfaedce54dceec', 0.000802108399220809),
 ('53f46a09dabfaeee22a6150a', 0.000802108399220809),
 ('54083330dabfae450f3fdde0', 0.000802108399220809),
 ('53f46fd6dabfaee02adba022', 0.000802108399220809),
 ('53f43e3bdabfaeee229e3586', 0.0013750429700928155)]
betweenness_centrality:
[('53f42f38dabfaedce54dceec', 0.0),
 ('53f46a09dabfaeee22a6150a', 0.0),
 ('54083330dabfae450f3fdde0', 0.0),
 ('53f46fd6dabfaee02adba022', 0.0),
 ('53f43e3bdabfaeee229e3586', 2.1886109361754386e-07)]
closeness_centrality:
[('53f42f38dabfaedce54dceec', 0.02922436644141198),
 ('53f46a09dabfaeee22a6150a', 0.02922436644141198),
 ('54083330dabfae450f3fdde0', 0.02922436644141198),
 ('53f46fd6dabfaee02adba022', 0.02922436644141198),
 ('53f43e3bdabfaeee229e3586', 0.029235708662339335)]
clustering_coefficient: 0.9092480690496565
diameter: 23
density: 0.0007607793884713801
```

**Analyse :**

- **Nombre de nœuds et d'arêtes :** 8728 nœuds et 28974 arêtes montrent l'ampleur du réseau et la densité de collaborations.
- **Centralités :** Les centralités de degré, d'intermédiairité et de proximité donnent une idée des nœuds les plus importants selon différentes perspectives.
- **Coefficient de clustering :** Le coefficient de clustering de 0.9092 montre que les chercheurs ont tendance à former des groupes fortement connectés.
- **Diamètre :** Un diamètre de 23 indique la distance maximale entre deux chercheurs dans le réseau.
- **Densité :** La densité de 0.00076 montre que le réseau est relativement sparse.

### **Étape 8 : Détails d'un nœud spécifique Commande exécutée :**

Entrez le numéro de l'analyse à effectuer : 12

Entrez l'ID du nœud : 54059d35dabfae8faa5e97f3

#### **Détails pour le nœud 54059d35dabfae8faa5e97f3 :**

name: dagan feng

year: 2009

title: Automated detection and delineation of lung tumors in PET-CT

volumes using a lung atlas and iterative mean-SUV threshold

fos: Nuclear medicine, Standardized uptake value, Lung cancer, PET-CT, Lung, Segmentation, Region growing, Soft tissue, Medicine, Mediastinum

references:

53e9a8d4b7602d97032250f1, 53e9af26b7602d970396302b,

53e9a9f0b7602d97033586a6, 53e9a178b7602d9702a69d44,

53e99db1b7602d9702671f1d, 53e9a9e6b7602d9703349c61

label: 54059d35dabfae8faa5e97f3

**Analyse :** Les détails du nœud '54059d35dabfae8faa5e97f3' révèlent des informations sur Dagan Feng, ses travaux et ses collaborations. Il est associé à plusieurs domaines de recherche et a de nombreuses références, montrant une riche activité de collaboration et d'influence dans le réseau scientifique.

### **Deuxième exemple de nœud : Commande exécutée :**

Entrez le numéro de l'analyse à effectuer : 12

Entrez l'ID du nœud : 5489b1cfabfae9b40134c30

#### **Détails pour le nœud 5489b1cfabfae9b40134c30 :**

name: aaron fenster

year: 2015

title: Automated pulmonary lobar ventilation measurements using volume-matched thoracic CT and MRI

fos: Nuclear medicine, Thoracic ct, Ventilation (architecture), Lung, Pulmonary function testing, Segmentation, Image segmentation, Medicine, Image registration, Magnetic resonance imaging

references: 55a43a71612ca648688ce8e5

label: 5489b1cfabfae9b40134c30

**Analyse :** Les détails du nœud '5489b1cfabfae9b40134c30' fournissent des informations sur Aaron Fenster et ses contributions au domaine. Son travail en ventilation pulmonaire lobaire automatisée est bien référencé, indiquant son influence et son intégration dans le réseau de collaboration.

Ces exemples montrent comment les différents algorithmes et outils de visualisation ont été utilisés pour analyser et interpréter les données de collaboration scientifique, révélant des informations importantes sur les réseaux académiques. Les résultats obtenus fournissent des insights sur les chercheurs influents, les communautés de collaboration, et les connexions clés dans le domaine de la médecine nucléaire.

Ces exemples montrent comment les différents algorithmes et outils de visualisation ont été utilisés pour analyser et interpréter les données de citations et de collaborations scientifiques, révélant des informations importantes sur les réseaux académiques.

**Annexe B.3 : Exemple 3 : Analyse du graphe "Réseau de collaboration dans le domaine de la Data Science"**

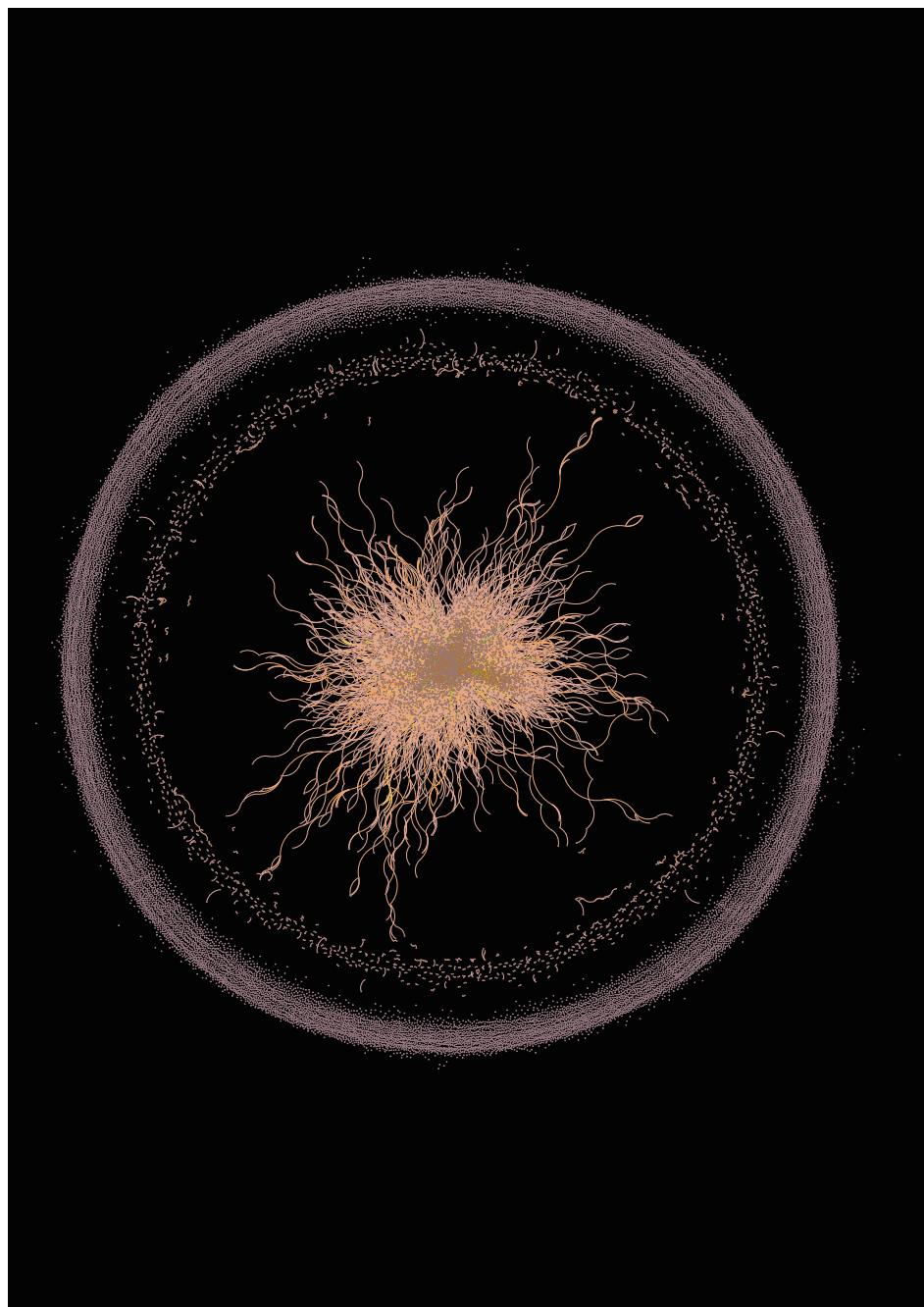


FIGURE 4 – Réseau de collaboration dans le domaine de la Data Science

## **Étape 1 : Calcul des métriques de centralité Commande exécutée :**

Entrez le numéro de l'analyse à effectuer : 2

### **Top 5 des centralités de degré :**

```
[('53e9ae29b7602d970383b8f3', 0.003993500234101738),  
 ('53e9af46b7602d97039827ec', 0.0032498829491310693),  
 ('53e9a480b7602d9702d9e8f7', 0.002864303616183315),  
 ('53e9a22cb7602d9702b32255', 0.002754138092483957),  
 ('53e9ad77b7602d9703768304', 0.0025338070450852406)]
```

### **Top 5 des centralités de proximité :**

```
[('5eef308c9fc0a24b4fc856', 0.0024051927494745068),  
 ('5f730bee9fc0a24b2a391c', 0.0020916169957288706),  
 ('599c7a76601a182cd26ba3d6', 0.0019719743978088584),  
 ('599c7a75601a182cd26b992f', 0.0016678183971417055),  
 ('555043d045ce0a409eb4985f', 0.001493882489474919)]
```

### **Top 5 des centralités d'intermédiarité :**

```
[('53e9ae29b7602d970383b8f3', 1.2347274099484304e-06),  
 ('558bbfc784ae6766fdef066e', 7.074729520372105e-07),  
 ('55503f9345ce0a409eb2ed4f', 6.958418741981427e-07),  
 ('56d8448bdabfae2eee37e46', 5.941963678654198e-07),  
 ('53e9b3c7b7602d9703eb2368', 5.405083573011494e-07)]
```

## **Étape 2 : Calcul du PageRank Commande exécutée :**

Entrez le numéro de l'analyse à effectuer : 3

### **Top 5 des scores de PageRank :**

```
[('555043d045ce0a409eb4985f', 0.00047873191131840207),  
 ('5eef308c9fc0a24b4fc856', 0.0004021382642116337),  
 ('599c7a75601a182cd26b992f', 0.00039039061477052047),  
 ('558bbfc784ae6766fdef066e', 0.0003795040255847806),  
 ('5f0ae2a09fc0a24ba2a56d', 0.00037073483168448485)]
```

## **Étape 3 : Calcul du coefficient de clustering Commande exécutée :**

Entrez le numéro de l'analyse à effectuer : 5

**Coefficient de clustering moyen :** 0.011296921276024986

## **Étape 4 : Détails sur le graphe Commande exécutée :**

Entrez le numéro de l'analyse à effectuer : 8

### **Détails du graphe :**

```

num_nodes: 36310
num_edges: 21422
degree_centrality:
[('53e9978db7602d9701f4ccea', 0.0),
 ('53e99796b7602d9701f5bc29', 0.00013770690462419784),
 ('53e997a6b7602d9701f7e6a7', 0.0),
 ('53e997b5b7602d9701f9bec6', 5.508276184967914e-05),
 ('53e997cbb7602d9701fbc5fa', 5.508276184967914e-05)]
betweenness_centrality:
[('53e9978db7602d9701f4ccea', 0.0),
 ('53e99796b7602d9701f5bc29', 0.0),
 ('53e997a6b7602d9701f7e6a7', 0.0),
 ('53e997b5b7602d9701f9bec6', 0.0),
 ('53e997cbb7602d9701fbc5fa', 0.0)]
closeness_centrality:
[('53e9978db7602d9701f4ccea', 0.0),
 ('53e99796b7602d9701f5bc29', 0.0),
 ('53e997a6b7602d9701f7e6a7', 0.0),
 ('53e997b5b7602d9701f9bec6', 0.0),
 ('53e997cbb7602d9701fbc5fa', 0.0)]
clustering_coefficient: 0.02259384255204997
diameter: 24
density: 1.6248732089559717e-05

```

**Explication des Résultats Centralités de Degré :** Les chercheurs avec les plus hauts degrés de centralité sont ceux ayant le plus grand nombre de collaborations directes. Par exemple, l'ID '53e9ae29b7602d970383b8f3', dont le titre est : Opinion Mining and Sentiment Analysis, a le plus grand nombre de collaborations directes, le rendant central dans le réseau.

**Centralités de Proximité :** Les chercheurs ayant les plus hautes centralités de proximité sont ceux qui peuvent atteindre tous les autres chercheurs en un nombre minimal de pas. L'ID '5eef308c9fcfd0a24b4fc856', dont le titre est : Survey on Visualization and Visual Analytics pipeline-based models : Conceptual aspects, comparative studies and challenges, est l'un de ceux qui peuvent diffuser l'information rapidement à travers le réseau.

**Centralités d'Intermédiarité :** Les chercheurs ayant les plus hautes centralités d'intermédiarité agissent comme des ponts entre différents groupes de collaboration. L'ID '53e9ae29b7602d970383b8f3', dont le titre est : Opinion Mining and Sentiment Analysis, est celui qui peut contrôler le flux d'information entre différents clusters.

**PageRank :** Le PageRank identifie les chercheurs les plus influents dans le réseau, non seulement en fonction du nombre de leurs collaborations, mais aussi de l'importance de leurs collaborateurs. L'ID '555043d045ce0a409eb4985f', dont l'intitulé de la publication est Knowledge Discovery and Data Mining in Biomedical Informatics : The Future Is in Integrative, Interactive Machine Learning Solutions, a le plus haut score de PageRank, indiquant une influence élevée.

**Coefficient de Clustering :** Le coefficient de clustering moyen de 0.0113 indique une faible tendance des chercheurs à former des groupes de collaboration étroits. Cela suggère que les collaborations dans ce réseau sont plus dispersées.

**Détails sur le Graphe :** Le graphe contient 36,310 nœuds et 21,422 arêtes, montrant

l'ampleur du réseau. La densité du graphe est de 0.00001625, indiquant un réseau relativement sparse. Le diamètre du graphe est de 24, ce qui signifie que la plus longue distance entre deux nœuds est de 24 étapes.

Ces analyses montrent que le réseau de collaboration en Data Science est vaste et diversifié, avec plusieurs chercheurs jouant des rôles centraux dans les collaborations. Les métriques de centralité révèlent les chercheurs les plus influents et les plus connectés, tandis que le coefficient de clustering et le diamètre du graphe offrent des insights sur la structure globale du réseau.

# Annexe C : Manuel utilisateur

## Annexe C.1 : Introduction

Ce manuel utilisateur a pour objectif d'expliquer comment utiliser les outils et les données du projet pour effectuer vos propres analyses. Il inclut des instructions pour installer les dépendances, exécuter les scripts et interpréter les résultats des analyses. Le projet vise à analyser et visualiser des réseaux de citations et de collaborations scientifiques en utilisant des données ouvertes.

## Annexe C.2 : Structure du Projet

Le projet est organisé en plusieurs dossiers, chacun contenant des scripts et des fichiers nécessaires à différentes étapes du processus d'analyse.

### 1. Dataset

Ce dossier contient les données JSON sources et les résultats intermédiaires des différentes étapes de traitement. Il est subdivisé en plusieurs sous-dossiers :

- **Split** : Contient les fichiers JSON d'origine scindés en morceaux plus petits pour faciliter leur traitement.
- **Split\_fusionné** : Contient les fichiers JSON fusionnés après diverses étapes de filtrage et de sélection.
- **Split\_filtré** : Contient les fichiers JSON filtrés selon des critères spécifiques tels que l'année ou le Field of Study (FOS).
- **Split\_nettoyé** : Contient les fichiers JSON nettoyés des colonnes inutiles et des enregistrements erronés.
- **Split\_prétraité** : Contient les fichiers JSON après prétraitement pour corriger les erreurs de format et les incohérences.
- **Top\_Sommet** : Contient les fichiers JSON des sommets avec les plus grands degrés après filtrage.
- **Graphe** : Contient les fichiers des graphes en format GEXF pour visualisation dans Gephi et les fichiers JPEG pour une visualisation rapide.
- **Split\_Standart** : Contient les fichiers JSON standards après la première scission des fichiers d'origine.

### 2. dossier codes python visualisation données

Ce dossier contient tous les scripts Python utilisés pour traiter et analyser les données. Voici une liste des scripts principaux et leurs fonctions :

- **CorrectionFormatJSON.py** : S'assure que chacun des Splits est correct.
- **PrétraioterNumberInt.py** : Remplace les champs year des fichiers JSON par des entiers.
- **NettoyerColonnes.py** : Nettoie les fichiers JSON des colonnes inutiles.
- **ListerDomaines.py** : Extrait tous les Fields of Study (FOS) mentionnés dans les publications.
- **FiltrerDataset.py** : Filtre les fichiers JSON selon différents critères.
- **FusionnerFichier.py** : Fusionne plusieurs fichiers JSON en un seul fichier.
- **RécuperationTopNSommets.py** : Sélectionne les 1000 sommets avec les plus grands degrés entrants ou sortants.
- **DessinerGraphe.py** : Crée les graphes de citation ou de collaboration et les enregistre en format JPEG et GEXF.

- `NettoyerGraphe.py` : Nettoie les graphes de toute potentielle erreur avant analyse.
- `AnalyseGraphe.py` : Effectue diverses analyses sur les graphes, telles que la détection de communautés, le calcul des centralités et le calcul du PageRank.
- `ScinderDataset2Go.py` : Scinde en 9 “Splits” de tailles égales d’environ 2 Go le fichier ‘Dataset publication scientifique JSON’.

## Annexe C.3 : Installation des Dépendances

Pour installer les dépendances nécessaires, exécutez le script `InstallerModules.py` fourni. Ce script utilise pip pour installer tous les modules requis.

## Annexe C.4 : Préparation des Données

- **Téléchargement des données** : Téléchargez les données de citation et de collaboration à partir de AMiner. Le dataset utilisé est nommé DBLP-Citation-network V13.
- **Scission des fichiers** : Utilisez le script `CorrectionFormatJSON.py` pour scinder le fichier JSON en plusieurs fichiers plus petits afin de faciliter leur traitement.
- **Prétraitement des fichiers** : Exécutez `PrétraiterNumberInt(x).py` pour remplacer les champs year par des entiers, puis utilisez `NettoyerColonnes.py` pour nettoyer les colonnes inutiles des fichiers JSON.
- **Extraction des FOS** : Utilisez `ListerDomaines.py` pour extraire les champs d’étude (Fields of Study) mentionnés dans les publications.

## Annexe C.5 : Filtration des Splits

Utilisez `FiltrerDataset.py` pour appliquer différents filtres aux fichiers JSON. Vous pouvez filtrer les données par :

- Année précise
- Plage d’années
- Années antérieures à une année donnée
- Années postérieures à une année donnée
- Field of Study (FOS)

Après filtration, les fichiers seront nommés `Split_{n}_Filtered_by{year/fos}.json`.

Utilisez `FusionnerFichier.py` pour fusionner les fichiers filtrés en un seul fichier JSON.

## Annexe C.6 : Création des Graphes

Utilisez `RécuperationTopNSommets.py` pour récupérer les 1000 sommets au plus grand degré entrant (citations) ou degré sortant (collaborations) et fusionnez les fichiers avec `FusionnerFichier.py` pour obtenir un fichier contenant environ 22 000 sommets.

Ensuite, utilisez `DessinerGraphe.py` pour créer les graphes de citation ou de collaboration. Les graphes seront enregistrés sous forme de fichiers JPEG pour une visualisation rapide et de fichiers GEXF pour une visualisation approfondie avec Gephi.

## Annexe C.7 : Analyse des Graphes

Avant d'appliquer des algorithmes de recherche opérationnelle, nettoyez le graphe avec `NettoyerGraphe.py` pour corriger les potentielles erreurs.

Utilisez le script `AnalyseGraphe.py` pour effectuer diverses analyses :

- **Détection de communautés (Louvain)** : Identifiez les communautés au sein du réseau.
- **Calcul des centralités** : Évaluez l'importance des nœuds en calculant les centralités de degré, de proximité et d'intermédiarité.
- **Calcul du PageRank** : Identifiez les publications "pierre angulaire".
- **Analyse du chemin le plus court entre deux chercheurs (Dijkstra)** : Trouvez les chemins de collaboration les plus courts.
- **Coefficient de clustering** : Comprenez la tendance des nœuds à former des clusters.
- **Diamètre du graphe** : Mesurez la plus longue distance entre deux nœuds.
- **Flux maximum entre deux nœuds** : Calculez le flux maximum dans le réseau.
- **Détails sur le graphe** : Obtenez des informations détaillées sur le graphe.
- **Matching maximum** : Analysez les relations de collaboration.
- **Arbre couvrant minimum** : Trouvez l'arbre couvrant minimum.
- **Plus courts chemins entre toutes les paires de nœuds (Floyd-Warshall)** : Calculez les plus courts chemins entre toutes les paires de nœuds.
- **Détails d'un nœud spécifique** : Obtenez des informations détaillées sur un nœud spécifique.

## Annexe C.8 : Exemple d'Exécution

### Étape 1 : Filtration des Données

- Placez les fichiers JSON téléchargés dans le dossier `Dataset/Split_prétraité`.
- Exécutez `FiltrerDataset.py` pour filtrer les données selon vos critères.
- Fusionnez les fichiers filtrés avec `FusionnerFichier.py`.

### Étape 2 : Création des Graphes

- Récupérez les 1000 sommets au plus grand degré avec `RécuperationTopNSommets.py`.
- Créez les graphes avec `DessinerGraphe.py`.

### Étape 3 : Analyse des Graphes

- Nettoyez les graphes avec `NettoyerGraphe.py`.
- Exécutez `AnalyseGraphe.py` pour effectuer les analyses souhaitées.

## Annexe C.9 : Conclusion

Ce manuel vous guide à travers les étapes nécessaires pour utiliser les outils et les données du projet pour effectuer vos propres analyses. En suivant ces instructions, vous pourrez visualiser et analyser les réseaux de citations et de collaborations scientifiques, identifier les chercheurs influents, détecter des communautés et comprendre les dynamiques académiques.

Pour toute question ou assistance supplémentaire, n'hésitez pas à consulter la documentation des modules Python utilisés ou à contacter l'équipe projet.