Chaouachi Salwa

**RAPPORT**

Analyse de Réseaux de Citations

ligne horizontale

## Introduction

## Ce rapport présente une analyse du réseau de citations cit-DBLP provenant de DBLP, visant à explorer la structure des citations entre articles scientifiques. L’objectif est de comprendre l'influence des articles et d’identifier les communautés thématiques via l’analyse du graphe de citations.

## Partie 1 : Collecte des données

### 1. Identification de la source de données en ligne

Les données utilisées dans ce projet proviennent du site [**NetworkRepository**](https://networkrepository.com/), une plateforme en ligne qui centralise de nombreux jeux de données en lien avec la recherche en analyse de graphes.  
 Le dataset sélectionné est **cit-DBLP**, un réseau de citations d’articles scientifiques issus de la base de données bibliographique **DBLP**.

🔗 **Lien de téléchargement direct** :  
<https://nrvis.com/download/data/cit/cit-DBLP.zip>

### 2. Identification des entités (nœuds) et des relations (liens)

* **Nœuds** : Chaque nœud représente un **article scientifique**.
* **Liens** : Une arête dirigée de u vers v signifie que **l’article u cite l’article v**.
* Le graphe est donc un **graphe orienté**.

### 3. Informations additionnelles

Le fichier contient uniquement les **liens de citation** sous forme de paires d'identifiants numériques. Il ne contient **aucun attribut** tel que :

* le nom de l’auteur,
* le titre ou l’année de publication,
* ni de poids pour les liens.

Le réseau est donc :

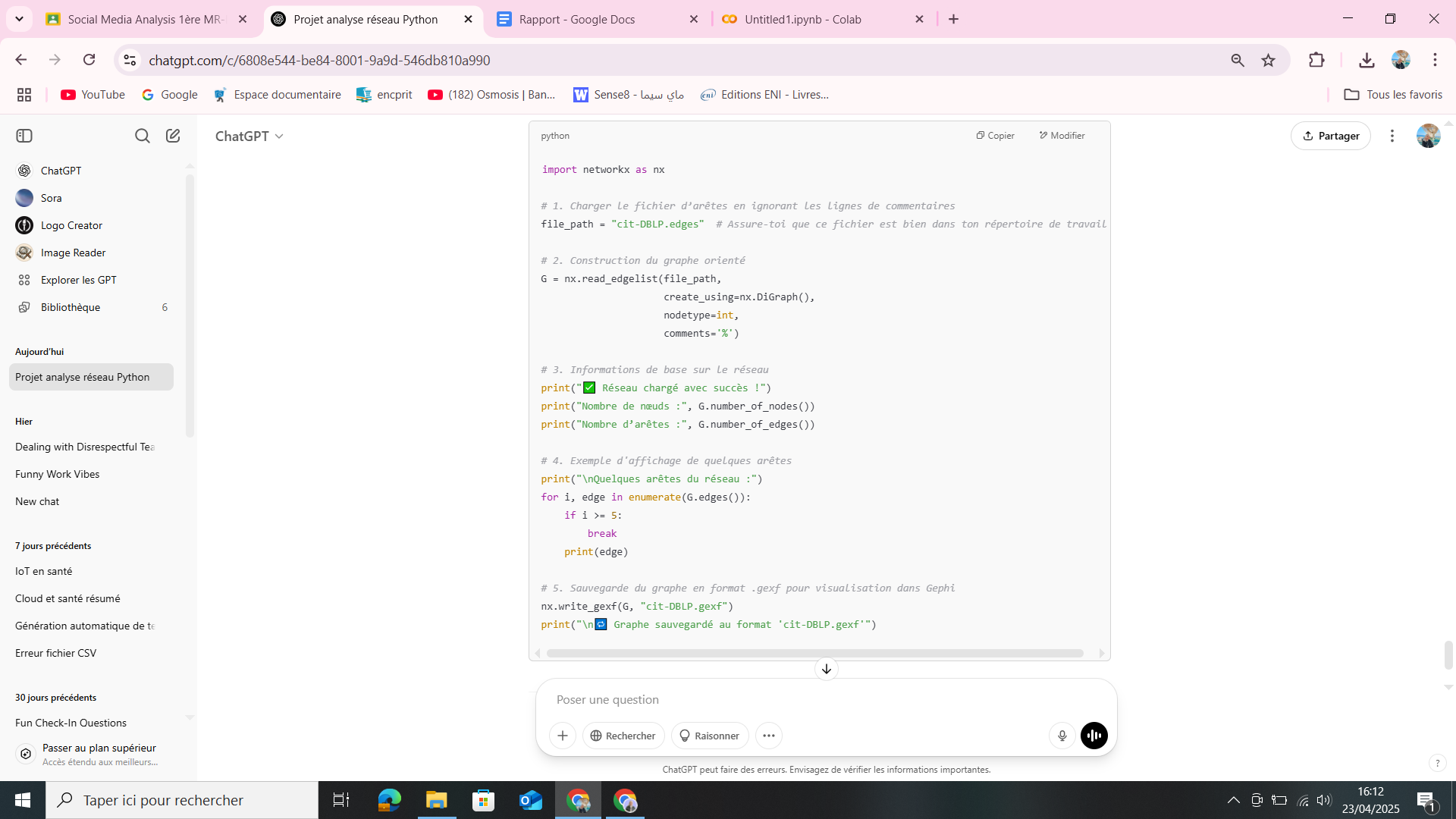
* **orienté**,
* **non pondéré**,
* et **sans attributs supplémentaires**.

### 4. Obtention des données

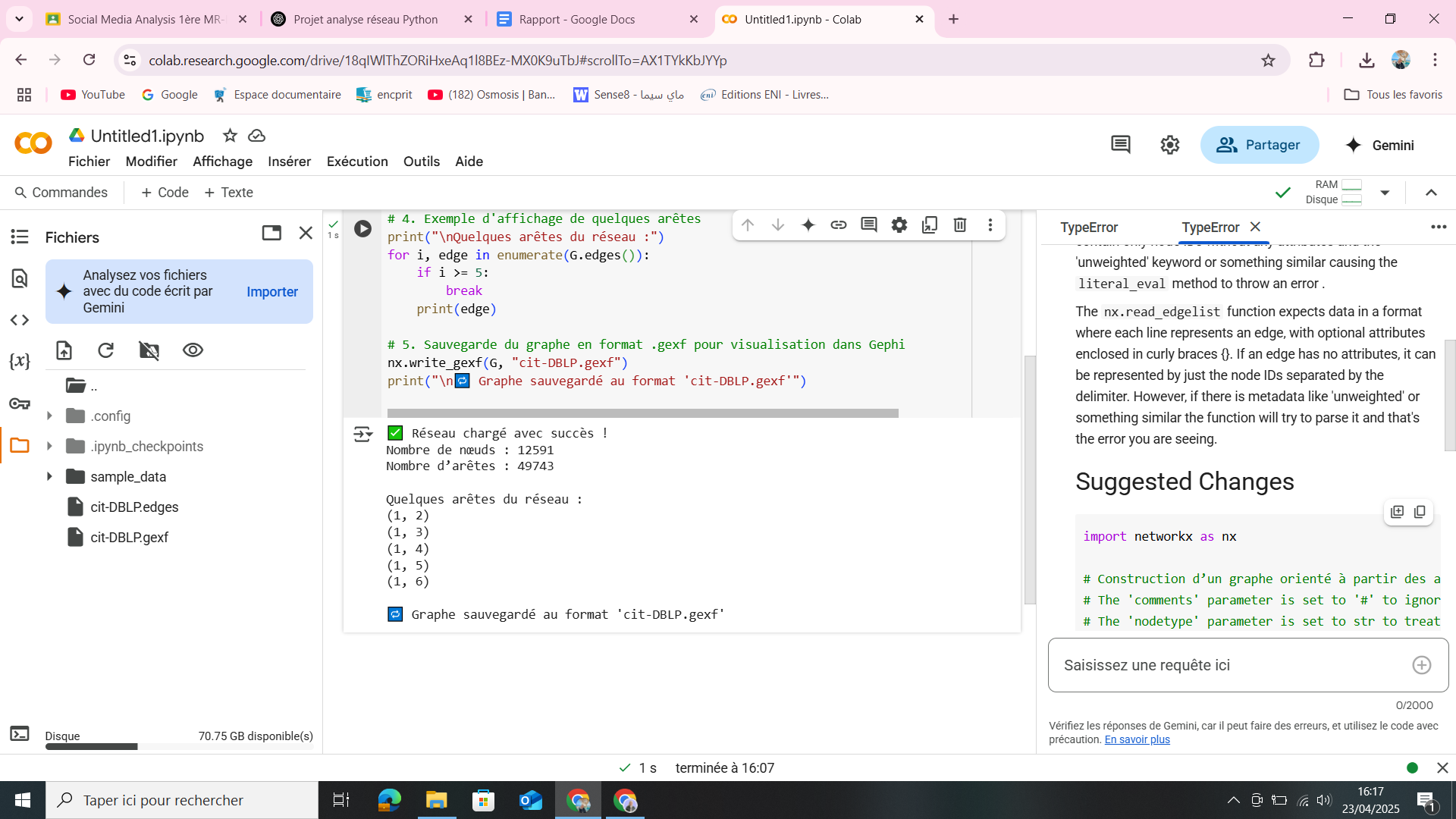
Les données ont été téléchargées sous forme de fichier compressé (cit-DBLP.zip) à partir de la plateforme NetworkRepository. Une fois décompressé, le fichier principal cit-DBLP.edges contient les relations de citation, ligne par ligne.

### 5. Construction du réseau

Le graphe a été construit en Python avec la bibliothèque **NetworkX** :



**Résultat :**

****

## Partie 2 : Analyse du réseau

Le graphe analysé représente un réseau de citations entre articles scientifiques (DBLP). C’est un graphe orienté, constitué de 12 591 nœuds et 49 743 arêtes.

### 1. Distribution des degrés

L’analyse de la distribution des degrés montre que :

* Certains articles sont fortement cités (degré entrant élevé), ce qui révèle leur influence.
* La majorité des nœuds ont un faible degré sortant, ce qui signifie qu’ils citent peu d’autres articles.

Des histogrammes ont été utilisés pour illustrer cette distribution (voir figures).

### 2. Composants connectés

* Composant faiblement connecté (WCC) le plus grand : 12 495 nœuds
* Composant fortement connecté (SCC) le plus grand : 240 nœuds

Cela signifie que la majorité des nœuds peuvent être atteints les uns des autres si l’on ignore l’orientation, mais seuls quelques groupes restreints se citent mutuellement dans les deux sens.

### 3. Analyse des chemins

Sur le plus grand SCC :

* Longueur moyenne des plus courts chemins : 6.27
* Diamètre (plus long plus court chemin) : 16

Ces valeurs montrent que les articles sont relativement proches les uns des autres dans ce sous-réseau, mais que certains chemins peuvent être très longs.

### 4. Clustering et densité

* Coefficient de clustering moyen : 0.1168
* Densité du graphe : 0.00031

Le faible clustering est typique des réseaux de citations : il est rare que deux articles qui en citent un troisième se citent aussi mutuellement. La densité très faible reflète un graphe très creux (sparse).

### 5. Analyse de la centralité

| **Rang** | **ID de l’article** | **Citations** |
| --- | --- | --- |
| 1 | 58 | 227 |
| 2 | 700 | 224 |
| 3 | 61 | 215 |
| 4 | 433 | 214 |
| 5 | 119 | 182 |
| 6 | 62 | 176 |
| 7 | 397 | 174 |
| 8 | 144 | 142 |
| 9 | 249 | 142 |
| 10 | 95 | 137 |

#### a. Top 10 des articles les plus cités (centralité de degré entrant) :**Ces articles sont les plus influents dans le réseau.**

#### **b. Centralité de proximité (closeness)**

Les 10 articles ayant la plus grande proximité sont ceux les plus "accessibles" par les autres dans le SCC. Cela indique une position stratégique dans le réseau.

#### **c. Centralité d’intermédiarité (betweenness)**

Les articles avec la plus grande intermédiarité jouent le rôle de ponts entre différentes communautés. Ils sont cruciaux pour la diffusion de l'information.

## Partie 3 : Détection de communautés

L’analyse de la structure du graphe de citations peut révéler des sous-ensembles d’articles fortement interconnectés, appelés **communautés**. Ces groupes peuvent correspondre à des domaines de recherche, à des écoles de pensée, ou à des thématiques scientifiques. Dans cette section, nous présentons trois méthodes de détection de communautés appliquées au graphe : **Louvain**, **Label Propagation**, et **K-Clique**.

### 3.1 Méthode de Louvain

La méthode de **Louvain** est une approche hiérarchique basée sur l'optimisation de la **modularité**, une mesure de la densité des liens à l'intérieur des communautés par rapport aux liens entre les communautés. Elle est particulièrement adaptée aux grands graphes.

* **Nombre de communautés détectées** : **58**
* Cette méthode produit des communautés bien séparées, souvent utilisées comme base dans des études sur les graphes réels.
* Elle est rapide, efficace, et tend à produire un nombre modéré de communautés.

### 3.2 Méthode de Label Propagation

La méthode de **Label Propagation** repose sur un processus itératif de diffusion d’étiquettes : chaque nœud adopte le label le plus fréquent parmi ses voisins jusqu’à stabilisation.

* **Nombre de communautés détectées** : **308**
* Elle est non déterministe et peut varier selon les exécutions.
* C’est une méthode rapide et sans paramètre, bien adaptée aux très grands graphes, mais elle a tendance à générer des communautés très fragmentées.

### 3.3 Méthode des k-Cliques (k = 4)

La méthode des **k-cliques** repose sur la recherche de sous-graphes complets de taille *k*, ici *k = 4*, qui se chevauchent. Elle permet de détecter des **communautés denses**, avec un haut degré de cohésion interne.

* **Nombre de communautés détectées** : **210**
* La plus grande communauté détectée contient **2623 nœuds**.
* Cette approche permet d’identifier des clusters très fortement connectés, mais elle dépend fortement du paramètre *k* et peut ignorer des structures moins denses.

### 3.4 Comparaison et interprétation

| **Méthode** | **Nombre de communautés** | **Avantages principaux** | **Inconvénients principaux** |
| --- | --- | --- | --- |
| Louvain | 58 | Bonne modularité, stable, hiérarchique | Moins sensible aux petites communautés |
| Label Propagation | 308 | Très rapide, sans paramètre | Résultats variables, sur-segmentation |
| k-Clique (k = 4) | 210 | Détecte des structures denses et précises | Dépend de *k*, exige des graphes denses |

Chaque méthode fournit une vision différente de la structure communautaire du graphe. Pour une analyse approfondie, il peut être intéressant de comparer les communautés détectées à des catégories thématiques (si disponibles) ou à des métadonnées bibliographiques.

**Conclusion**

L’analyse a révélé des informations clés sur la distribution des degrés et la centralité des articles. La méthode Louvain a détecté 58 communautés bien séparées, tandis que Label Propagation et k-Clique ont produit des résultats plus fragmentés. Cette étude fournit des insights sur l’organisation des citations et pourrait aider à mieux comprendre les dynamiques de publication dans la recherche scientifique.