FACULTÉ DES SCIENCES ÉCONOMIQUES ET DE GESTION DE NABEUL

RAPPORT DE PROJET ANALYSE DES RÉSEAUX SOCIAUX

Élaborée par :
Wiem BOUKHARI

Proposé par :
DR. Fériel BEN FRAJ

2023/2024

I. INTRODUCTION :

Ce projet explore le réseau de collaboration académique dans le domaine de la Relativité Générale et de la Cosmologie Quantique (GR-QC), basé sur les articles soumis à la catégorie correspondante sur arXiv.

En examinant les collaborations entre auteurs sur une période de janvier 1993 à avril 2003, nous cherchons à comprendre la structure et l'évolution de cette communauté scientifique.

Notre analyse comprend l'exploration des données, la détection de communautés et la prédiction des liens potentiels entre chercheurs.
Les résultats fourniront des informations précieuses sur la dynamique des collaborations dans ce domaine de recherche fondamental.

II. COLLECTE DES DONNÉES :

1. Identification de la source de données en ligne :

J'ai sélectionné le jeu de données Arxiv GR-QC (General Relativity and Quantum Cosmology) collaboration network.
Le réseau de collaboration Arxiv GR-QC provient des e-preprints de arXiv et couvre les collaborations scientifiques entre les auteurs de documents soumis à la catégorie de la Relativité Générale et de la Cosmologie Quantique.

Si un auteur i a co-écrit un document avec l'auteur j, le graphe contient une arête non dirigée de i à j.
Si le document est co-écrit par k auteurs, cela génère un (sous)-graphe complètement connecté sur k nœuds.

Les données couvrent les documents de la période de janvier 1993 à avril 2003 (124 mois). Elles commencent quelques mois seulement après la création de arXiv, et représentent ainsi essentiellement l'histoire complète de sa section GR-QC.

2. Méthodes de collecte des données :

Les données ont été obtenues directement à partir du site web de Stanford SNAP (Stanford Network Analysis Project).

J'ai téléchargé le fichier de données au format .txt à partir de la page dédiée au jeu de données Arxiv GR-QC.

3. Identification des entités (nœuds) et des relations entre elles (liens) :

Dans notre ensemble de données, les entités sont représentées par les auteurs de documents scientifiques, et les relations entre ces entités sont établies par les collaborations entre ces auteurs.

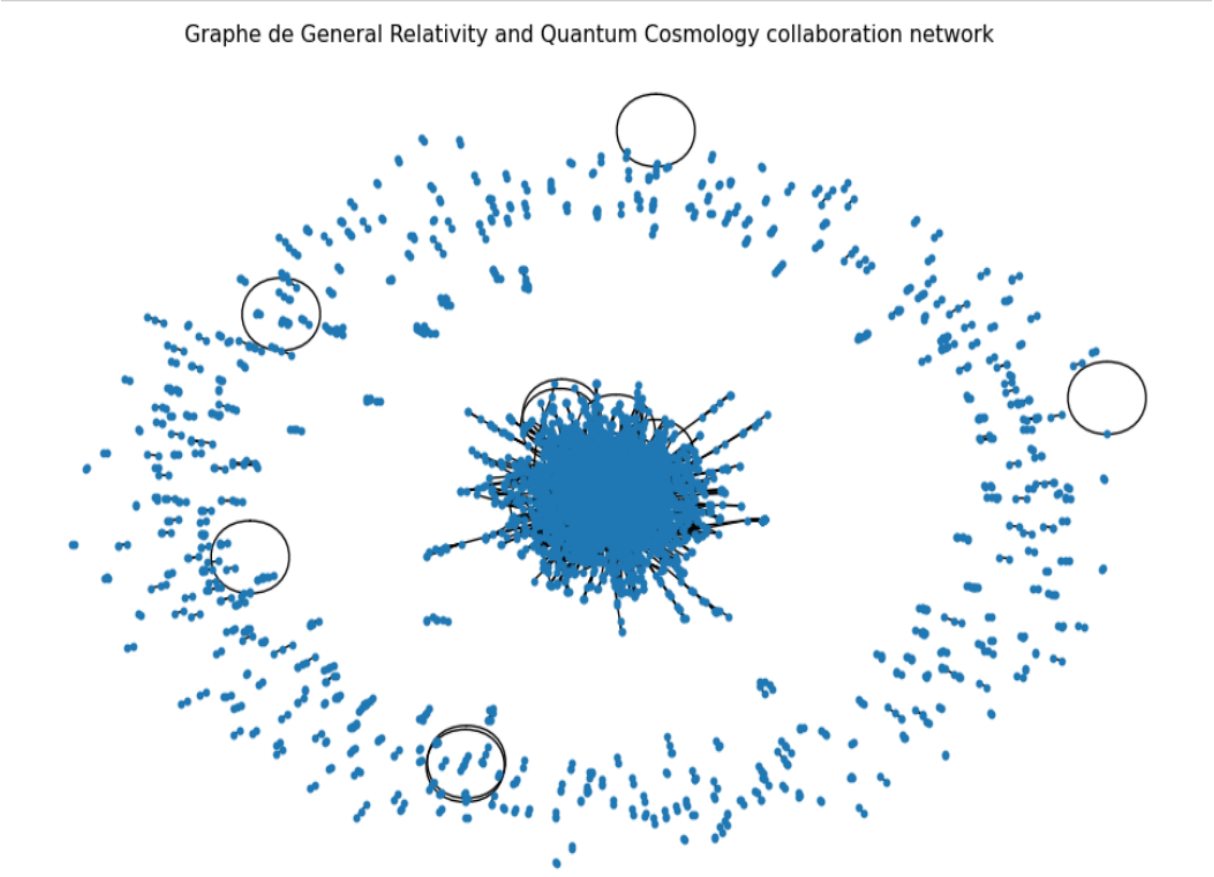
```
print("Nombre total d'entités (nœuds) :", nombre_entites)
print("Nombre total de relations (liens) :", nombre_relations)
```

```
Nombre total d'entités (nœuds) : 5242
Nombre total de relations (liens) : 14496
```

4. Construction du réseau :

Une fois les données téléchargées, j'ai utilisé des bibliothèques Python telles que NetworkX pour construire un réseau à partir des informations de collaboration entre les auteurs. Chaque nœud dans le réseau représente un auteur, et chaque lien représente une collaboration entre deux auteurs.

En construisant ce réseau, j'ai créé une représentation graphique des interactions entre les auteurs dans le domaine de la relativité générale et de la cosmologie quantique.



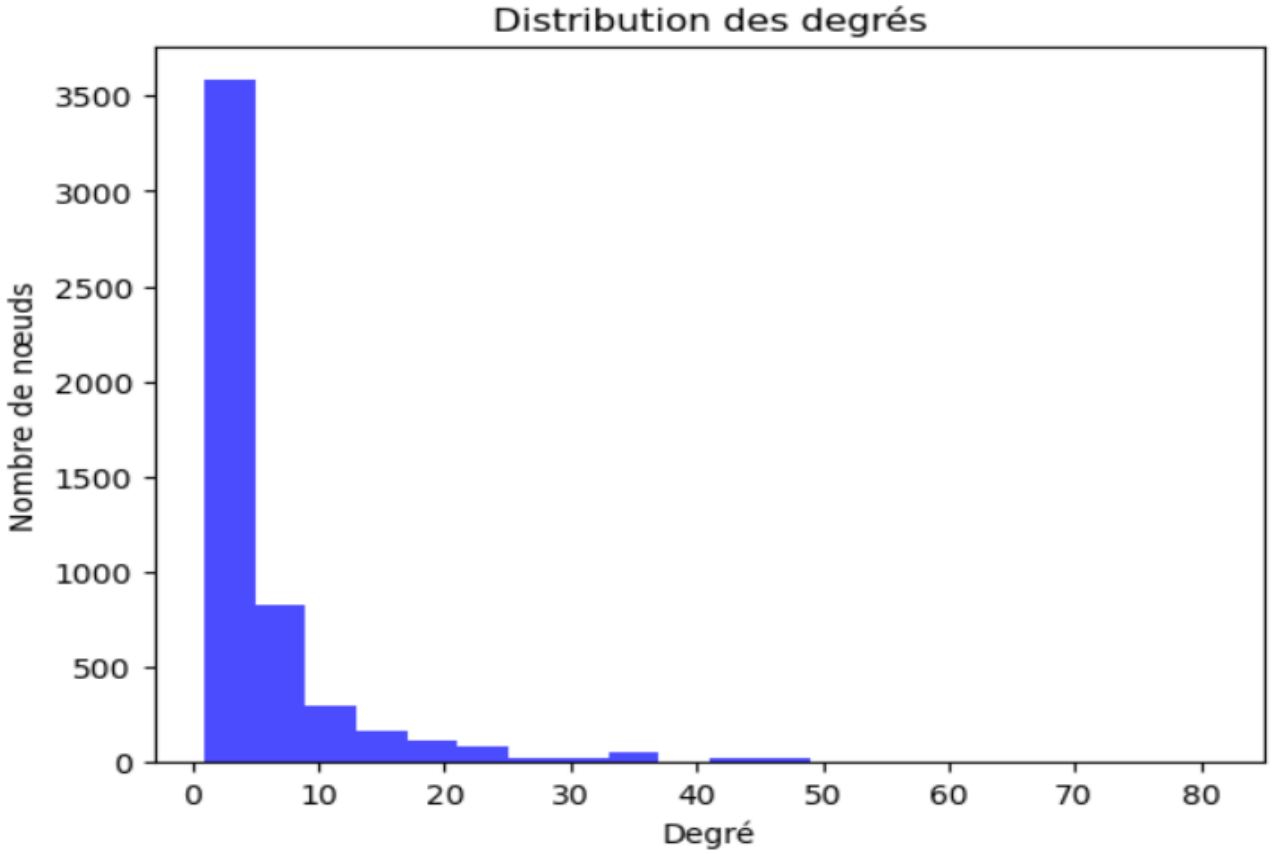
III. ANALYSE DU RÉSEAU :

Dans cette deuxième partie, je me suis concentré sur l'analyse exploratoire des données pour mieux comprendre la structure et les caractéristiques du réseau de collaboration entre les auteurs dans le domaine de la relativité générale et de la cosmologie quantique.

➢ L'analyse de la distribution des degrés :

J'ai calculé la distribution des degrés des nœuds dans le réseau, ce qui m'a permis de comprendre la répartition des collaborations entre les différents auteurs.

Une distribution des degrés étudiée peut révéler des informations sur la structure globale du réseau, telles que la présence de nœuds fortement connectés ou de nœuds isolés.



➢ L'analyse des composants connectés :

J'ai identifié les composants connectés dans le réseau, c'est-à-dire les sous-graphes dans lesquels chaque nœud est relié à tous les autres nœuds par un chemin.

Cette analyse permet de comprendre s'il existe des sous-groupes d'auteurs qui sont fortement connectés entre eux mais moins connectés au reste du réseau.

```
Nombre de composants connectés : 355
Taille du composant connecté 1 : 4158
Taille du composant connecté 2 : 2
Taille du composant connecté 3 : 3
Taille du composant connecté 4 : 2
Taille du composant connecté 5 : 2
Taille du composant connecté 6 : 2
Taille du composant connecté 7 : 2
```

➢ L’analyse des chemins :

J'ai étudié les chemins les plus courts entre les paires de nœuds dans le réseau.

```
Longueur moyenne des chemins les plus courts : 6.049380016182999
```

Cette analyse permet de comprendre la distance moyenne entre les auteurs et la facilité avec laquelle l'information peut se propager à travers le réseau.

```
Distance moyenne entre les nœuds : 3.8061249382218865
```

➢ Le coefficient de clustering et l’analyse de la densité :

J'ai calculé le coefficient de clustering moyen pour évaluer à quel point les nœuds dans le réseau forment des clusters ou des regroupements denses.
La densité du réseau a également été évaluée pour comprendre à quel point le réseau est dense en termes de connexions entre les nœuds.

```
Coefficient de clustering moyen : 0.5294338141753044
Densité du graphe : 0.0010545461581625744
```

➢ L’analyse de la centralité :

J'ai calculé plusieurs mesures de centralité, telles que la centralité d'intermédiarité (betweenness centrality) et la centralité de proximité (closeness centrality), pour identifier les auteurs les plus centraux dans le réseau.
Ces analyses permettent de comprendre qui sont les auteurs les plus influents ou les mieux positionnés dans le réseau de collaboration.

```
Moyenne de la centralité d'intermédiarité : 0.0006062707858585091
```

```
Auteur: 3466
Betweenness Centrality: 0.001715248119663197
Closeness Centrality: 0.15250009440442658
Eigenvector Centrality: 3.489474540927225e-05

Auteur: 937
Betweenness Centrality: 0.0003493412893676944
Closeness Centrality: 0.14789649865964416
Eigenvector Centrality: 1.4596578376254 49e-05

Auteur: 5233
Betweenness Centrality: 0.0
Closeness Centrality: 0.14373793718636851
Eigenvector Centrality: 6.681320201491573e-06

Auteur: 8579
Betweenness Centrality: 0.00035838805830684087
Closeness Centrality: 0.14223737289668725
Eigenvector Centrality: 5.131748375726487e-06

Auteur: 10310
Betweenness Centrality: 0.00875853820067516
Closeness Centrality: 0.17089274080637024
Eigenvector Centrality: 0.0002698843812326264
```

IV. IDENTIFICATION DES COMMUNAUTÉS:

Pour identifier les communautés dans le réseau de collaboration, j'ai utilisé plusieurs algorithmes de détection des communautés, notamment :

Louvain : Cet algorithme est largement utilisé pour détecter les communautés dans les réseaux. Il maximise la modularité du réseau en recherchant des partitions qui optimisent le regroupement des nœuds en communautés.

Propagation des labels : Cette méthode attribue initialement une étiquette à chaque nœud, puis propage ces étiquettes à travers le réseau en fonction des nœuds voisins. Les nœuds finissent par appartenir à la même communauté s'ils partagent des étiquettes similaires.

K-clique : Cet algorithme identifie les sous-graphes complets de taille k (cliques) dans le réseau. Chaque clique est considérée comme une communauté potentielle.

Infomap : Cette méthode utilise la théorie de l'information pour détecter les structures modulaires dans le réseau. Elle cherche à minimiser le code de longueur moyenne nécessaire pour décrire les mouvements d'information sur le réseau.

J'ai implémenté ces algorithmes à l'aide de la bibliothèque NetworkX et évalué leurs performances en termes de qualité des partitions identifiées.
En utilisant les attributs supplémentaires disponibles dans le réseau, tels que les noms des auteurs ou d'autres métadonnées, j'ai cherché à donner un sens aux communautés identifiées.

L'analyse des communautés peut être étendue en explorant d'autres approches disponibles dans la bibliothèque CDlib ou en développant des méthodes personnalisées en fonction des besoins spécifiques du réseau étudié.

```
Communautés détectées avec Louvain :
Communauté 1: ['4583', '9572', '16310', '899', '3096', '16148', '24371', '24559', '7689', '19738', '8916', '22184', '23647',
'23708', '1765', '3032', '7442', '7768', '13276', '17266', '22415', '7264', '19489', '10762', '20647', '3909', '8862', '78',
'4877', '7459', '8254', '12155', '22598', '24932', '888', '1520', '7007', '7712', '10711', '13614', '14102', '18517', '2335
1', '24114', '2465', '2592', '3977', '5055', '5993', '9265', '19890', '20341', '21560', '1375', '5555', '5787', '9721', '1015
8', '10942', '13600', '21316', '22900', '23770', '25143', '25601', '25980', '4870', '5175', '8282', '22046', '4766', '3561',
'4868', '8352', '15365', '16931', '10620', '3679', '13955', '14003', '15235', '18227', '19445', '19495', '23967', '15399', '1
8222', '21287', '23227', '4624', '5355', '6863', '15770', '21322', '22265', '22336', '23099', '23880', '4319', '12758', '2202
3', '16575', '12587', '5840', '6732', '8614', '13847', '15081', '22609', '5143', '9735', '17396', '26138', '2591', '8932', '9
188', '19179', '2127', '11712', '16106', '17207', '17670', '18286', '18612', '19234', '19724', '3412', '894', '4814', '5441',
```

```
Communautés détectées avec Propagation des labels :
Communauté 1: ['14807', '11148', '25251', '950', '9482', '19423', '12981', '9800', '15003', '14512', '25346', '20850', '1661
1', '15659', '19961', '24140', '20774', '8879', '22372', '2560', '3595', '18894', '2952', '13597', '23628', '22887', '23481',
'21281', '18668', '17655', '16594', '11785', '25758', '12365', '4046', '23293', '12553', '21012', '4180', '2803', '4164', '49
75', '4146', '5434', '731', '13164', '18648', '9283', '11175', '25053', '6774', '18867', '7533', '4513', '8922', '12085', '21
847', '9488', '3372', '10243', '20562', '1653', '4382', '231', '46', '345', '15538', '12422', '18866', '13411', '2980', '1555
2', '8053', '25189', '2982', '13995', '20108', '21508', '16726', '6610', '25684', '10041', '17182', '18719', '13060', '2614
1', '19954', '773', '9341', '8517', '8049', '4511', '16722', '1997', '7014', '8503', '4632', '21800', '2741', '11472', '1285
1', '5578', '12496', '23403', '12679', '10549', '7384', '5262', '11053', '7956', '1234', '14376', '6830', '14540', '6179', '2
```

```
Communautés détectées avec Propagation des labels :
Communauté 1: ['22528', '24453', '18580', '7578', '6094', '20148', '22504', '8219', '25251', '12688', '896', '25402', '2577
7', '3291', '12806', '22197', '1694', '3651', '22837', '10623', '6700', '25562', '9800', '18331', '15580', '4466', '2810', '5
510', '7717', '17439', '12710', '17807', '19961', '6963', '15325', '20774', '18790', '22177', '25201', '8134', '20424', '2057
4', '8612', '5346', '15614', '14419', '15846', '25094', '6411', '23628', '4635', '12070', '18286', '23110', '22791', '574',
'13649', '12884', '8374', '25931', '5605', '17655', '19150', '12365', '19870', '21860', '15585', '19923', '21583', '6934', '1
8724', '17979', '6317', '11113', '7386', '10910', '10942', '13636', '11400', '18940', '13624', '13164', '6896', '11801', '876
8', '2287', '16393', '7533', '22188', '9943', '11964', '21379', '24885', '12130', '5136', '10427', '21407', '10519', '24295',
'20562', '23409', '231', '25572', '20879', '3502', '25988', '18866', '13411', '10310', '21321', '2535', '1014', '2980', '1229
```

```
Communautés détectées avec Louvain :
Communauté 1: ['6610', '9785', '19423', '21012', '22691', '15003', '45', '46', '570', '773', '1653', '2212', '2741', '2952',
'3372', '4046', '4164', '4511', '4513', '5262', '6179', '6830', '7956', '8879', '11241', '11472', '12365', '12496', '12781',
'12851', '14540', '14807', '15659', '17655', '17692', '18894', '19961', '20108', '20562', '20635', '21281', '21508', '21847',
'22887', '23293', '24955', '25346', '25758', '12422', '1234', '1997', '5134', '9341', '13060', '23628', '25251', '4632', '110
53', '13411', '14512', '16594', '20774', '23403', '3595', '18174', '12679', '13164', '4382', '17875', '23095', '2055', '256
0', '4146', '10926', '12553', '18668', '12981', '8922', '9488', '17182', '10041', '13995', '7384', '3280']
Communauté 2: ['6512', '2449', '7444', '20574', '88', '1995', '2338', '2530', '2535', '4355', '4755', '5210', '6340', '7197',
'7635', '8727', '9313', '9656', '9869', '10350', '12507', '12710', '14648', '14766', '15010', '15244', '15245', '15457', '159
33', '16654', '17807', '18225', '18453', '18778', '19114', '19446', '19645', '19939', '20346', '20576', '21178', '22504', '23
712', '24453', '24718']
Communauté 3: ['20827', '8978', '9017', '15170', '15455', '16589', '2255', '3056', '6158', '7324', '8365', '9023', '11444',
```

```
Indice de Jaccard entre les communautés détectées par K-clique et Propagation des labels: [0.1380020597322348, 0.0, 0.0, 0.0,
```

```
Indice de Jaccard entre les communautés détectées par K-clique et Louvain: [0.26537013801756587, 0.002325581395348837, 0.0, 0.
0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
```

```
Indice de Jaccard entre les communautés détectées par Propagation des labels et Louvain: [0.0008216926869350862, 0.127314814814
81483, 0.0, 0.0, 0.0, 0.0, 0.0, 0.04678362573099415, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.
```

V. PRÉDICTION DES LIENS :

➢ Répartition des données :

Pour diviser les données du réseau de collaboration en ensembles d'apprentissage et de test, j'ai utilisé une approche de répartition aléatoire en conservant 80% des liens pour l'apprentissage et 20% pour le test.
Cette répartition a été effectuée en veillant à ce qu'elle soit équilibrée et représentative de l'ensemble des liens du réseau. Ainsi, chaque ensemble contient une proportion adéquate de liens entre les différents nœuds du réseau, garantissant ainsi la diversité et la qualité des données utilisées pour l'apprentissage et l'évaluation des modèles.

➢ Application d'algorithmes d'apprentissage automatique :

Pour la prédiction des liens dans le réseau de collaboration, j'ai exploré plusieurs algorithmes d'apprentissage automatique, notamment les méthodes de K-Means, les arbres de décision, Réseaux de neurones, Forêt aléatoire, K-Nearst Neighbors et les méthodes de Multi-Layer Perceptron.

Ces algorithmes ont été entraînés sur l'ensemble d'apprentissage en utilisant des caractéristiques pertinentes extraites du réseau, telles que la similarité entre les nœuds, les

mesures de centralité, etc. Chaque algorithme a été ajusté et évalué en fonction de ses performances prédictives sur l'ensemble de test.

✧ Résultats de l'algorithme de K-Means :

```
Exactitude de prédiction: 0.5139751552795031
```

✧ Résultats de l'algorithme de Forêt aléatoire :

```
Exactitude de prédiction: 1.0
```

✧ Résultats de l'algorithme de Réseaux de neurones :

```
Epoch 10/10
580/580 ━━━━━━━━━━━━━━━━━━━━ 0s 627us/step - accuracy: 1.0000 - loss: 5.3806e-09 - val_accuracy: 1.0000 - val_loss: 9.3524e-06
182/182 ━━━━━━━━━━━━━━━━━━━━ 0s 444us/step - accuracy: 1.0000 - loss: 4.0916e-22
Exactitude de prédiction sur les données de test: 1.0
```

✧ Résultats de l'algorithme de K-Nearest Neighbors :

```
# Évaluer la performance du modèle
accuracy = np.mean(predictions == np.zeros(len(X_test)))
print("Exactitude de prédiction:", accuracy)
```

```
Exactitude de prédiction: 1.0
```

✧ Résultats de l'algorithme de Multi-Layer Perceptron :

```
print("Exactitude de prédiction:", accuracy)
```

```
Exactitude de prédiction: 1.0
```

✧ Résultats de l'algorithme de l'arbre de décision :

```
Exactitude de prédiction: 1.0
```

Il semble que tous les algorithmes utilisés ont une exactitude de prédiction de 1.0 sur ce ensemble de données, cela signifie que ces modèles ont réussi à prédire correctement tous les liens dans l'ensemble de test.

Une exactitude de prédiction de 1.0 peut être inhabituelle et peut indiquer plusieurs choses :

Sur-apprentissage (Overfitting) : Les modèles peuvent avoir sur-appris les données d'entraînement et mémorisé les exemples au lieu d'apprendre des motifs généraux. Cela peut se produire si les modèles sont trop complexes par rapport à la quantité de données d'entraînement disponible.

Données d'entraînement et de test similaires : Il est possible que les données d'entraînement et de test soient très similaires, ce qui permet aux modèles de généraliser efficacement à de nouvelles données.

Données de test trop simples : Les données de test peuvent être trop simples ou peu représentatives, ce qui facilite la prédiction pour les modèles