

Projet

Analyse des Réseaux Sociaux

Coauthorships in network science



Hassouna Hiba

1ère année Mastère de recherche Business computing(BC)

Partie 1 : Collecte des données.....	1
Partie 2 : Analyse du réseau.....	4
❖ Distribution des poids des arrêts.....	5
❖ Densité.....	6
❖ Coefficient de clustering.....	6
❖ les voisins d'un noeud.....	6
❖ Trouver les chercheurs que je peux collaborer avec.....	6
❖ Centralité.....	6
→ Diamètre et centre de chaque composante connectée.....	8
→ plus court chemin.....	8
→ Liste des voisinages de tous les noeuds.....	9
partie 3: Identification des communautés.....	9
Partie 4 : Prédiction des liens.....	11

Partie 1 : Collecte des données

1. Identifier une source de données en ligne

réseau de coauteurs des scientifiques travaillant sur la théorie et l'expérimentation des réseaux, compilé par M. Newman en mai 2006. Ce réseau a été compilé à partir des bibliographies de deux articles de revue sur les réseaux : M. E. J. Newman, SIAM Review 45, 167-256 (2003) et S. Boccaletti et al., Physics Reports 424, 175-308 (2006), avec quelques références supplémentaires ajoutées manuellement.

lien : [Network data \(umich.edu\)](http://networkdata.umich.edu)

2. Identifier les entités (nœuds) et les relations entre elles (liens)

le réseau non orienté avec 1589 sommets et 2742 arêtes. Chaque nœud représente un auteur ; les arêtes représentent les articles écrits conjointement par les auteurs. La valeur est le poids MEJ Newman. Il n'y a pas de boucles ni d'arêtes multiples incluses.

3. Identifier les informations additionnelles valables

Chaque sommet est représentée par (id:numéro de sommet,Label:Nom de l'auteur)

Le réseau est pondéré, avec des poids variables attribués comme décrit dans M.E.J. Newman, Phys. Rev. E 64, 016132 (2001). Les poids représentent la force de la collaboration scientifique.

calcul des poids

Dans l'article de M.E.J. Newman, Phys. Rev. E 64, 016132 (2001), les poids sont calculés en utilisant une méthode basée sur la théorie des graphes. Plus précisément, Newman propose une métrique appelée poids

MEJ (MEJ weight) qui est déterminée par la fréquence à laquelle deux auteurs apparaissent ensemble sur des articles scientifiques.

Le poids entre deux auteurs est calculé en utilisant la formule suivante :

$$w_{ij} = \sum_k \frac{n_{ik} \cdot n_{jk}}{(\sum_{k'} n_{ik'}) \cdot (\sum_{k'} n_{jk'})}$$

où w_{ij} est le poids entre les auteurs i et j , n_{ik} est le nombre d'articles auxquels l'auteur i a contribué avec l'auteur k , et n_{jk} est le nombre d'articles auxquels l'auteur j a contribué avec l'auteur k . Les termes $\sum_{k'} n_{ik'}$ et $\sum_{k'} n_{jk'}$ représentent respectivement le nombre total d'articles auxquels l'auteur i et l'auteur j ont contribué.

4. Obtenir les données à partir de la source de données

Le fichier netscience.gml est téléchargé à partir de lien : [Network data \(umich.edu\)](https://networkdata.umich.edu)

Le fichier netscience.gml contient un réseau de coauteurs de scientifiques travaillant sur la théorie et l'expérimentation des réseaux

5. Construire un réseau à partir des données

★ Chargement des données :

```
#Charger le fichier dans un graphe
G = nx.read_gml('/content/netscience.gml', label="id", destringizer=int)
```

Informations sur le graphe :
Nombre de nœuds : 1589
Nombre d'arêtes : 2742

Le graphe est dirigé : False
Le graphe est vide : False
Le graphe est bipartite : False
Le graphe est biconnecté : False
Le graphe est Connecté : False

Liste des nœuds avec leurs étiquettes		
ID: 0	Label: ABRAMSON, G	Liaison entre SALWINSKI, L et BARON, M avec un poids de 0.2
ID: 1	Label: KUPERMAN, M	Liaison entre YANG, K et HUANG, L avec un poids de 0.5
ID: 2	Label: ACEBRON, J	Liaison entre YANG, K et YANG, L avec un poids de 0.5
ID: 3	Label: BONILLA, L	Liaison entre HUANG, L et YANG, L avec un poids de 0.5
ID: 4	Label: PEREZVICENTE, C	Liaison entre YAN, G et ZHOU, T avec un poids de 0.25
ID: 5	Label: RITORT, F	Liaison entre YAN, G et FU, Z avec un poids de 0.25
ID: 6	Label: SPIGLER, R	Liaison entre ZHOU, T et FU, Z avec un poids de 0.25
ID: 7	Label: ADAMIC, L	Liaison entre YAOUM, Y et LAUMANN, E avec un poids de 1
ID: 8	Label: ADAR, E	Liaison entre YEHA, A et JEANDUPREUX, D avec un poids de 0.333333
ID: 9	Label: HUBERMAN, B	Liaison entre YEHA, A et ALONSO, F avec un poids de 0.333333
ID: 10	Label: LUKOSE, R	Liaison entre YEHA, A et GUEVARA, M avec un poids de 0.333333
ID: 11	Label: PUNTYANI, A	Liaison entre JEANDUPREUX, D et ALONSO, F avec un poids de 0.333333
ID: 12	Label: AERTSEN, A	Liaison entre JEANDUPREUX, D et GUEVARA, M avec un poids de 0.333333
ID: 13	Label: GERSTEIN, G	Liaison entre ALONSO, F et GUEVARA, M avec un poids de 0.333333
ID: 14	Label: HABIB, M	Liaison entre YOOK, S et TU, Y avec un poids de 0.333333
ID: 15	Label: PALM, G	Liaison entre SAGER, J et CSARDI, G avec un poids de 0.333333
ID: 16	Label: AFRAIMOVICH, V	Liaison entre SAGER, J et HAGA, P avec un poids de 0.333333
ID: 17	Label: VERICHEV, N	Liaison entre CSARDI, G et HAGA, P avec un poids de 0.333333
ID: 18	Label: RABINOVICH, M	Liaison entre YUSONG, T et LINGJIANG, K avec un poids de 0.333333
ID: 19	Label: AGRAWAL, H	Liaison entre YUSONG, T et MUREN, L avec un poids de 0.333333
ID: 20	Label: AHUJA, R	Liaison entre LINGJIANG, K et MUREN, L avec un poids de 0.333333
ID: 21	Label: MAGNANTI, T	Liaison entre ZAKS, M et PARK, E avec un poids de 0.333333
ID: 22	Label: ORLIN, J	Liaison entre ZASLAVER, A et MAYO, A avec un poids de 0.142857
		Liaison entre ZASLAVER, A et ROSENFELD, B avec un poids de 0.142857

★ Visualisation de réseau

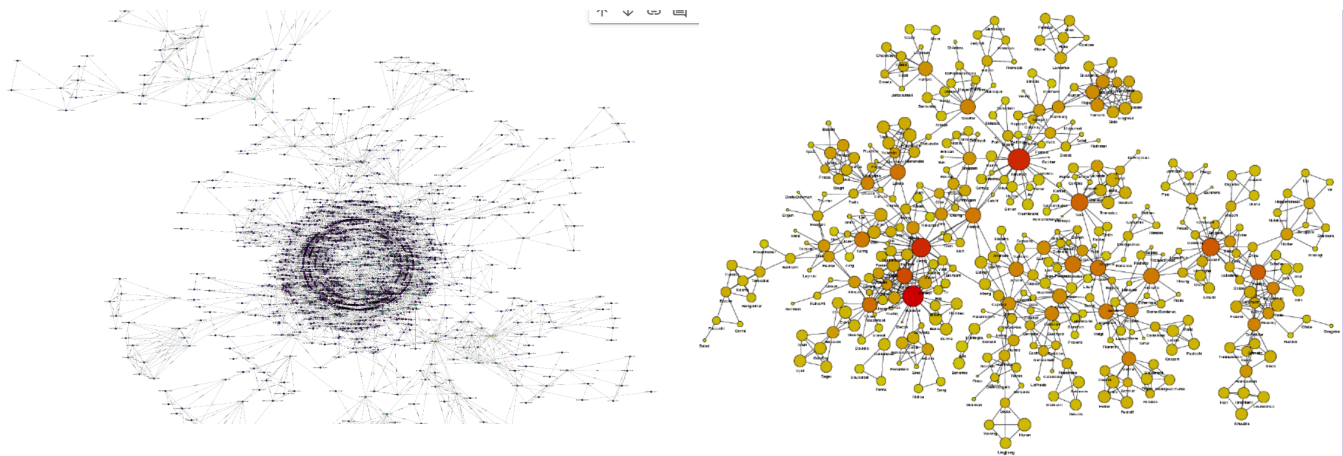


figure 1 le graphique complet du réseau NetScience

Partie 2 : Analyse du réseau

- ❖ Distribution des degrés: Le degré d'un sommet est le nombre d'arêtes incidentes sur le sommet. La figure 2 montre la distribution du degré de collaboration entre auteurs.

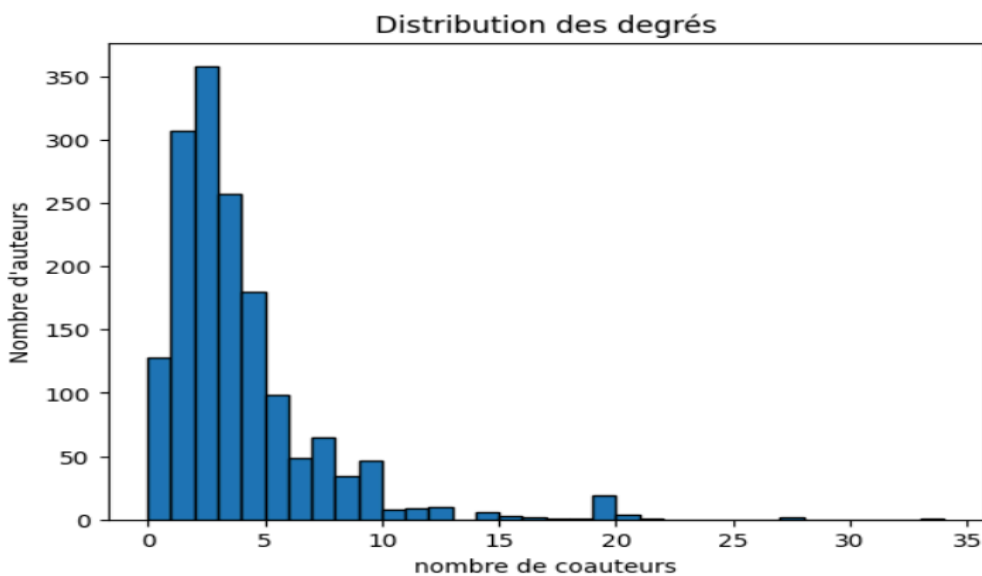


figure 2 Distribution des degrés

degrés minimale est égale à 0 et degrés maximale est égale à 34

Cela montre la présence des nœuds isolés.

❖ Distribution des poids des arrêts

Le Tableau 1 montre la distribution des poids des arêtes dans le réseau NetScience, la valeur la plus basse est 0.0526 et la valeur la plus élevée est 4.7500.

Table 1: Distribution des poids des arêtes dans le réseau NetScience
Le poids minimale 0.0526316 et le poids maximale est 4.75.

Poids	Frequence
0.0526316	187
0.111111	135
3.5	2
3.58333	1
3.83333	1
4.225	1
4.75	1
Poids minimale	0.0526316

Table 1 Distribution de poids

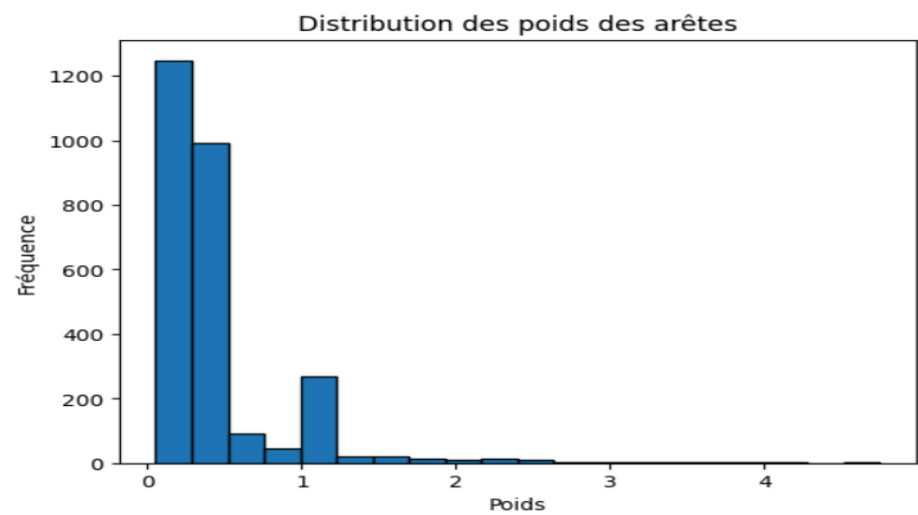


Figure 3 la distribution des poids des arêtes

❖ Densité

La densité est le quotient entre le nombre de connexions d'un réseau donné et le nombre maximal possible de connexions dans le même réseau. Il est évident qu'un réseau complet a une densité maximale. Cependant, la densité n'est pas la meilleure mesure car elle dépend de la taille du réseau, c'est-à-dire du nombre de connexions possibles. La densité du réseau NetScience est de 0.0021 \Rightarrow environ 0.21% des paires de nœuds possibles sont reliées par des arêtes dans le graphe

❖ Coefficient de clustering

Le coefficient de clustering est un moyen de mesurer à quel point les nœuds d'un graphe sont regroupés en clusters ou en communautés.

Coefficient de clustering moyen: 0.6377905695067805

❖ les voisins d'un nœud

Les voisins du DURAN, O (911) sont : [(909, 'KLEVECZ, R'), (910, 'BOLEN, J')]

❖ Trouver les chercheurs que je peux collaborer avec

Voisin: KLEVECZ, R - Voisins des voisins avec poids maximal: ['BOLEN, J', 'DURAN, O']
Voisin: BOLEN, J - Voisins des voisins avec poids maximal: ['KLEVECZ, R', 'DURAN, O']

les auteurs que DURAN, O peut collaborer avec sont:

Le Voisin : KLEVECZ, R - Voisins des voisins avec poids maximal: ['BOLEN, J', 'DURAN, O']

Le Voisin : BOLEN, J - Voisins des voisins avec poids maximal: ['KLEVECZ, R', 'DURAN, O']

❖ Centralité

	Statistique	Valeur
0	degree centrality	{0: 0.0012594458438287153, 1: 0.001889168765743073, 2: 0.0025188916876574307, 3: 0.0031489926876574307, 4: 0.0037890936876574307, 5: 0.0044291946876574307, 6: 0.0050692956876574307, 7: 0.0057093966876574307, 8: 0.0063494976876574307, 9: 0.0069895986876574307}
1	betweenness centrality	{0: 0.0, 1: 1.5872033318572343e-06, 2: 0.0, 3: 0.0, 4: 0.0, 5: 0.0, 6: 0.0, 7: 0.0, 8: 0.0, 9: 0.0}
2	closeness centrality	{0: 0.0014168765743073047, 1: 0.001889168765743073, 2: 0.0025188916876574307, 3: 0.0031489926876574307, 4: 0.0037890936876574307, 5: 0.0044291946876574307, 6: 0.0050692956876574307, 7: 0.0057093966876574307, 8: 0.0063494976876574307, 9: 0.0069895986876574307}

+ La centralité de proximité(Closeness Centrality) peut être interprétée dans les réseaux sociaux comme la facilité pour un nœud d'atteindre les autres nœuds dans l'ensemble du réseau.

+Le concept d'intermédierité (betweenness) est la mesure dans laquelle un nœud se trouve entre d'autres nœuds dans le réseau social. Cette mesure prend en compte la connectivité des voisins du nœud. Elle reflète le nombre de nœuds auxquels un nœud est connecté indirectement via leurs liens directs

+ Les chercheurs en réseaux sociaux mesurent l'activité du réseau pour un nœud en utilisant le concept de degrés - le nombre de connexions directes qu'un nœud possède. Il peut être défini comme la quantité de liens qui se produisent sur un nœud (c'est-à-dire le nombre de liens qu'un nœud possède).

❖ L'analyse des composants connectés

Nombre de composantes connectées: 396

La plus grande composant connectée = {30, 31, 32, 33, 34, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 69, 70, 71, 72, 77

Nombre de nœuds de la plus grande composante connectée: 379

Nombre d'arêtes dans la plus grande composante connexe : 914

Nœuds isolés: [19, 26, 41, 89, 101, 110, 115, 125, 159, 168, 178, 204, 232, 236, 253, 257, 272, 295, 407, 420, 451, 504, 510, 536, 543,

Nombre des noeuds isolés 128

Le centre du sous graphe la plus large est : [78, 131, 203, 756, 757, 758, 759, 1123]

Le diamètre du sous graphe la plus large est : 17

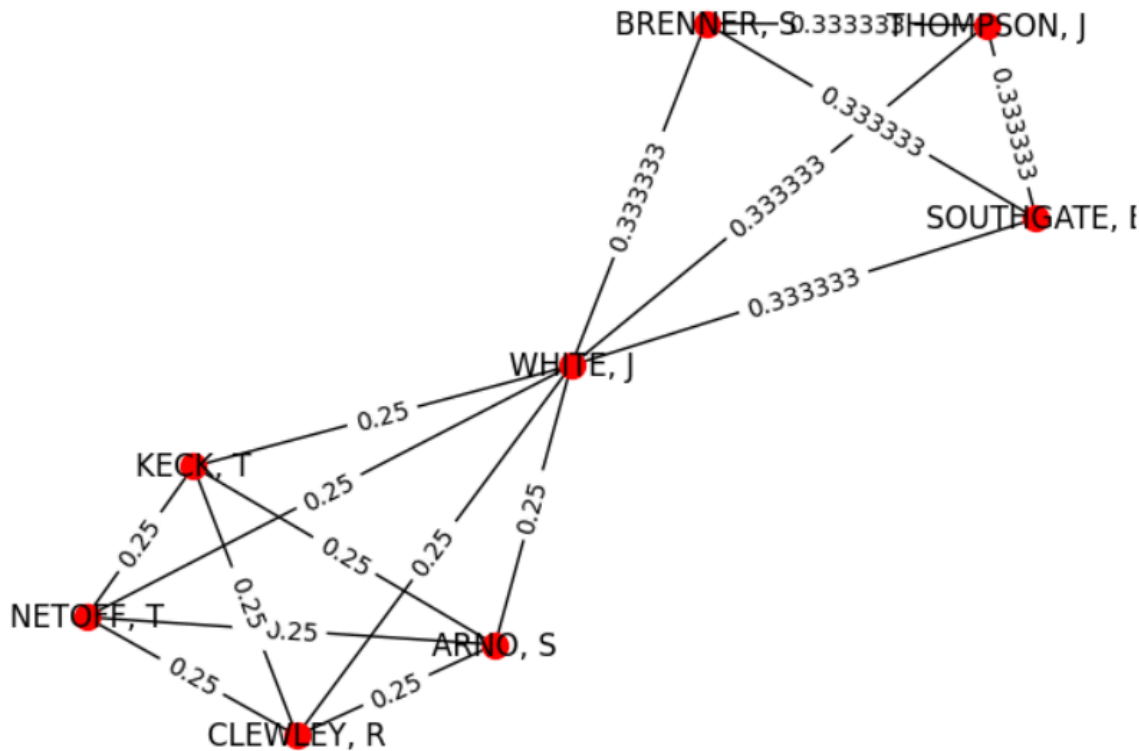
Le rayon du sous graphe la plus large est : 9



figure 4 graphe de plus grand composant

→ Diamètre et centre de chaque composante connectée

Composante 1 : 4 noeuds - Diamètre : 2, Centre : [1], nombre de noeud : 4, nombre des arrêts : 4
 Composante 2 : 5 noeuds - Diamètre : 1, Centre : [2, 3, 4, 5, 6], nombre de noeud : 5, nombre des arrêts : 10
 Composante 3 : 8 noeuds - Diamètre : 3, Centre : [7, 9, 10, 11], nombre de noeud : 8, nombre des arrêts : 11
 Composante 4 : 8 noeuds - Diamètre : 2, Centre : [12], nombre de noeud : 8, nombre des arrêts : 16
 Composante 5 : 3 noeuds - Diamètre : 1, Centre : [16, 17, 18], nombre de noeud : 3, nombre des arrêts : 3
 Composante 6 : 1 noeuds - Diamètre : 0, Centre : [19], nombre de noeud : 1, nombre des arrêts : 0
 Composante 7 : 3 noeuds - Diamètre : 1, Centre : [20, 21, 22], nombre de noeud : 3, nombre des arrêts : 3
 Composante 8 : 7 noeuds - Diamètre : 3, Centre : [201, 202, 24, 25], nombre de noeud : 7, nombre des arrêts : 12
 Composante 9 : 1 noeuds - Diamètre : 0, Centre : [26], nombre de noeud : 1, nombre des arrêts : 0
 Composante 10 : 3 noeuds - Diamètre : 1, Centre : [27, 28, 29], nombre de noeud : 3, nombre des arrêts : 3
 Composante 11 : 379 noeuds - Diamètre : 17, Centre : [78, 131, 203, 756, 757, 758, 759, 1123], nombre de noeud : 379, nombre des arrêts : 379
 Composante 12 : 6 noeuds - Diamètre : 1, Centre : [35, 36, 37, 38, 39, 40], nombre de noeud : 6, nombre des arrêts : 15
 Composante 13 : 1 noeuds - Diamètre : 0, Centre : [41], nombre de noeud : 1, nombre des arrêts : 0
 Composante 14 : 2 noeuds - Diamètre : 1, Centre : [42, 43], nombre de noeud : 2, nombre des arrêts : 1
 Composante 15 : 3 noeuds - Diamètre : 1, Centre : [59, 60, 61], nombre de noeud : 3, nombre des arrêts : 3
 Composante 16 : 31 noeuds - Diamètre : 4, Centre : [62], nombre de noeud : 31, nombre des arrêts : 97
 Composante 17 : 3 noeuds - Diamètre : 1, Centre : [66, 67, 68], nombre de noeud : 3, nombre des arrêts : 3
 Composante 18 : 57 noeuds - Diamètre : 7, Centre : [523, 1356, 742, 746], nombre de noeud : 57, nombre des arrêts : 149



Nombre de noeuds dans le composant 281: 8
 Nombre d'arêtes dans le composant 281: 16

→ plus court chemin

Plus court chemin entre NETOFF, T et THOMPSON, J : ['NETOFF, T', 'WHITE, J', 'THOMPSON, J']

→ Liste des voisinages de tous les noeuds

```
The neighbors of node 2 are [3, 4, 5, 6]
The neighbors of node 0 are [1, 1084]
The neighbors of node 1 are [0, 946, 1084]
The neighbors of node 2 are [3, 4, 5, 6]
The neighbors of node 3 are [2, 4, 5, 6]
The neighbors of node 4 are [2, 3, 5, 6]
The neighbors of node 5 are [2, 3, 4, 6]
The neighbors of node 6 are [2, 3, 4, 5]
The neighbors of node 7 are [8, 9, 10, 11]
The neighbors of node 8 are [7]
The neighbors of node 9 are [7, 10, 11, 1424, 1425, 1532]
The neighbors of node 10 are [7, 9, 11]
The neighbors of node 11 are [7, 10, 9]
The neighbors of node 12 are [13, 14, 15, 1047, 1048, 1049, 1050]
The neighbors of node 13 are [12, 14, 15]
The neighbors of node 14 are [12, 13, 15]
```

partie 3: Identification des communautés

Une communauté est un organisme vivant partageant un environnement commun et interagissent entre eux .

bibliothèque CDlib : est une bibliothèque Python dédiée à la détection de communautés dans les réseaux complexes. Elle fournit une gamme d'algorithmes de détection de communautés et d'évaluation de partitions communautaires.

Dans cette partie, On identifie, évalue et valide l'échantillon du réseau choisi par l'implémentation des algorithmes de détections de la communautés , tel que :

- ☒ Propagation des labels
- ☒ Louvain
- ☒ Infomap

1. Définition des algorithmes

- Propagation des labels

La propagation de labels est l'une des méthodes les plus rapides pour la détection de communautés, de complexité quasi-linéaire en termes d'arêtes. Il s'agit d'une méthode locale où chaque nœud possède son propre label qui change par interaction avec son voisinage.

- Louvain

Est un algorithme hiérarchique d'extraction de communautés applicable à de grands réseaux. La méthode a été proposée par Vincent Blondel et al. de l'Université de Louvain en 2008.

- Infomap

introduite pour la première fois par Rosvall et Bergstrom (2008) :

Minimise la longueur de la description d'un random matching ;La méthode infomap utilise une équation cartographique . Cette équation exploite la recherche aléatoire de structure de cluster et leur connexion en minimisant la longueur du marcheur .

plus la connexion d'un objet est pondérée, plus il est possible que le marcheur aléatoire utilisera cette connexion.

2. Résultat

Algorithme	Nombre de communautés
Louvain	405
Infomap	442
Label Propagation	467

Indice de similarité de Jaccard entre Louvain et Infomap: 0.8574561403508771

Indice de similarité de Jaccard entre Louvain et Propagation des labels: 0.7942386831275721

Indice de similarité de Jaccard entre Infomap et Propagation des labels: 0.8107569721115537

- Pour l'indice de similarité de Jaccard entre Louvain et Infomap, la valeur est d'environ 0.86. Cela signifie que 86% des nœuds sont affectés aux mêmes communautés par ces deux algorithmes.
- Pour l'indice de similarité de Jaccard entre Louvain et la Propagation des labels, la valeur est d'environ 0.80. Cela signifie que 80% des nœuds sont affectés aux mêmes communautés par ces deux algorithmes.
- Pour l'indice de similarité de Jaccard entre Infomap et la Propagation des labels, la valeur est d'environ 0.82. Cela signifie que 82% des nœuds sont affectés aux mêmes communautés par ces deux algorithmes.

Partie 4 : Prédiction des liens

precision du modele 0.4918032786885246
 rappel du modele 0.4918032786885246
 f1-score du modele 0.6593406593406593

