Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Mohamed El-Bachir El-Ibrahimi Bordj Bou Arréridj
Faculté des Mathématiques et de l'Informatique
Département de Recherche Opérationnelle

Mémoire de Master

Mémoire présenté en vue de l'obtention du diplôme de Master en :
Domaine : Mathématiques et Informatique.
Filière : Mathématiques.
Spécialité : Méthodes et Outils pour la Recherche Opérationnelle.

Similarité globale pour la prédiction de liens dans les réseaux complexes

Présenté par :

- Ouali Aya
- Zitouni Rayane

Sous la direction de :
M. Abdelhamid Saifi

Soutenu publiquement le 14/juin/2025 devant le jury composé de :

M.	Saifi Abdelhamid	Professeur	Université de BBA	Encadreur
M.	Naili,M	Maître de Conférences	Université de BBA	Président
M.	Maza,S	Maître de Conférences	Université de BBA	Examinateur

Année Universitaire 2024/2025

Remerciement

Nous remercions notre Dieu tout puissant pour nous avoir donné la santé, le courage, la force et un excellent encadrement pour finir ce travail.
Tout d'abord nous remercions très sincèrement ce qui nous a encadré durant nos recherches, spécialement, le Docteur Abdelhamid Saifi qui malgré ses nombreuses occupations, a été pour nous autant une source d'inspiration, qu'un guide chaleureux et accueillant, chaque fois que nous frappions à sa porte.
Nous adressons nos sincères remerciements à tous les Membres du jury ici présent, pour avoir fait l'honneur d'évaluer notre travail.
Nous adressons nos vifs remerciements à tous nos enseignants du département de la Recherche Opérationnelle pour les connaissances qu'ils nous ont inculquées durant ces cinq années.
Nous tenons à présenter nos remerciements les plus sincères à nos parents, pour leurs soutiens et encouragements de tous les instants qui ont toujours été là pour nous.
Enfin, Nous disons sincèrement merci à nos camarades pour leurs multiples conseils et encouragements, ainsi qu'à nos nombreux parents et amis.

Dédicace

Je dédie ce mémoire à :

A ma très chère mère Tu présentes pour moi le symbole de la bonté par excellence, la source de tendresse et l'exemple du dévouement qui n'a pas cessé de m'encourager et de prier pour moi.

A mon père Aucune dédicace ne saurait exprimer l'amour, l'estime, le dévouement et le respect que j'ai toujours eu pour vous.

A mes très chérés sœurs et frères Hamza, Mouhamed, Walid, Hanane, Ryma, Iman, Sarah, Meriam et Walid, Aymen, Ahlem, Asma, Marwa .
En témoignage de l'amour, et de l'affectation que je porte pour vous.

A mes chèr(e)s ami(e)s Vous êtes pour moi des frères, sœurs et des amis sur qui je peux compter.
En témoignage de l'amitié qui nous unit et des souvenirs de tous les moments que nous avons passé ensemble, je vous dédie ce travail et je vous souhaite une vie pleine de santé et de bonheur surtout Hiba.

*a*ya et rayane

Résumé

Dans ce mémoire, nous abordons la problématique de la prédiction de liens dans les réseaux complexes, une tâche cruciale pour anticiper l'apparition de relations entre entités.

Nous nous concentrons plus particulièrement sur les méthodes de similarité globale, qui exploitent la structure entière du réseau pour estimer la probabilité d'existence d'un lien entre deux nœuds.

Cinq méthodes ont été étudiées et comparées : Shortest Path, SimRank, Newton's Gravitational Law Index (NGLI) , Katz Index et Common Neighbor Distance (CND).

Après avoir présenté les fondements théoriques des graphes et des réseaux complexes, nous avons implémenté ces méthodes dans un environnement Python et les avons testées sur plusieurs réseaux réels issus de différents domaines (biologie, transport, réseaux sociaux...).

Les performances ont été évaluées à l'aide de métriques classiques telles que la précision, le rappel, la F-mesure, l'exactitude (accuracy).

Les résultats ont montré que chaque méthode présente des avantages spécifiques selon la structure du réseau, et qu'aucune approche n'est universellement supérieure.

Ce travail permet ainsi de mieux guider le choix de la méthode en fonction du contexte d'application.

Mots-clés : Prédiction de liens, Réseaux complexes, Similarité globale, théorie des graphes, Évaluation des performances.

Abstract

This thesis addresses the problem of link prediction in complex networks, a critical task for anticipating the emergence of connections between entities.

We specifically focus on global similarity methods, which leverage the entire network structure to estimate the likelihood of a link between two nodes.

Five methods are studied and compared : Shortest Path, SimRank, Newton's Gravitational Law Index (NGLI), Katz Index, and Common Neighbor Distance (CND).

After presenting the theoretical foundations of graph theory and complex networks, we implemented these methods using Python and applied them to several real-world networks from different domains (biology, transportation, social networks, etc.).

The performance of each method was evaluated using standard metrics such as precision, recall, F-measure, and accuracy.

The results show that each method has its strengths depending on the network structure, and no single method consistently outperforms the others.

This study thus provides valuable insights to guide the choice of link prediction techniques based on specific application contexts.

Keywords : Link prediction, complex networks, global similarity, graph theory, performance evaluation.

ملخص

يتناول هذا البحث مشكلة تنبؤ الروابط في الشبكات المعقدة، وهي مهمة أساسية تهدف إلى توقع ظهور علاقات بين الكيانات داخل الشبكة.
نركز بشكل خاص على طرق التشابه العالمية، والتي تعتمد على البنية الكاملة للشبكة لتقدير احتمالية وجود علاقة بين عقدتين.
لقد قمنا بدراسة ومقارنة خمس طرق تنبؤ وهي: أقصر مسار، ومؤشر قانون الجاذبية لنيوتن، ومؤشر كاتز، ومسافة الجيران المشتركة، وسيم رانك.
بعد استعراض الأسس النظرية لنظرية الرسوم البيانية والشبكات المعقدة، تم تنفيذ هذه الطرق وتطبيقها على شبكات حقيقية من مجالات متنوعة (بيولوجيا، نقل، شبكات اجتماعية...)
تم تقييم أداء هذه الطرق باستخدام مؤشرات شائعة مثل الدقة، والاسترجاع، ومقياس، ومعدل الصحة.
أظهرت النتائج أن لكل طريقة مزاياها الخاصة حسب بنية الشبكة، ولا توجد طريقة تتفوق دائمًا على غيرها.
يساهم هذا العمل في توجيه اختيار الطريقة المناسبة بحسب السياق التطبيق.

الكلمات المفتاحية:
تنبؤ الروابط، الشبكات المعقدة، التشابه العالمي، الرسوم البيانية ، تقييم الأداء.

Table des matières

Table des figures

Liste des tableaux

Introduction Générale

Avec l'essor des systèmes interconnectés dans des domaines variés comme les réseaux sociaux, les systèmes biologiques, les infrastructures technologiques ou encore les réseaux d'information, la compréhension et l'analyse des structures relationnelles deviennent primordiales.

La théorie des graphes s'impose ainsi comme un outil incontournable pour modéliser ces relations complexes à travers des représentations de réseaux.

Parmi les nombreuses tâches d'analyse de réseaux, la prédiction de liens occupe une place centrale.

Elle vise à anticiper l'apparition de nouvelles connexions ou à identifier des liens manquants entre les entités d'un réseau.

Cette capacité d'anticipation est d'une grande utilité pratique, notamment dans la recommandation d'amis, la détection d'interactions biologiques ou encore l'amélioration de la diffusion d'informations.

Dans ce mémoire, nous nous intéressons à une famille spécifique d'approches : les méthodes de similarité globale, qui exploitent la structure entière du réseau pour estimer la probabilité d'existence d'un lien entre deux nœuds.

L'objectif principal est de comparer l'efficacité de cinq mesures globales : Shortest Path, SimRank, Newton's Gravitational Law Index (NGLI), Katz Index et Common Neighbor Distance (CND), appliquées à divers réseaux complexes.

Motivation

La richesse des réseaux réels, caractérisée par leur structure hétérogène, leurs dynamiques évolutives et leur grande taille, rend la prédiction de liens particulièrement difficile.

Si de nombreuses méthodes locales existent, elles peuvent s'avérer limitées dans les cas où les relations ne se manifestent pas uniquement à travers le voisinage immédiat.

Les méthodes de similarité globale, quant à elles, permettent de prendre en compte l'ensemble de la topologie du réseau, offrant une vision plus étendue et potentiellement plus précise de la connectivité future.

Cependant, ces méthodes présentent des caractéristiques computationnelles variées et leurs performances peuvent fortement dépendre du type de réseau.

Dans ce contexte, il est essentiel de comparer et d'évaluer rigoureusement ces approches afin de mieux comprendre leurs avantages respectifs et d'identifier celles qui sont les plus adaptées à des contextes spécifiques.

Contributions

Ce mémoire apporte plusieurs contributions majeures :

Une présentation détaillée de cinq méthodes de similarité globale utilisées pour la prédiction de liens : Shortest Path, SimRank, Newton's Gravitational Law Index (NGLI), Katz Index et Common Neighbor Distance (CND).

L'implémentation de ces méthodes dans un environnement Python, appliquée à des réseaux réels issus de fichiers Pajek.

Une évaluation comparative basée sur des métriques standards : précision, rappel, F-mesure, exactitude (accuracy) .

Organisation du Mémoire

Le mémoire est structuré en 3 chapitres :

Chapitre 1 : Présentation des fondements des graphes et des réseaux complexes, incluant définitions, types de graphes et indicateurs structurels.

Chapitre 2 : Revue des méthodes de prédiction de liens, avec un focus sur les approches de similarité globale.

Chapitre 3 : Description de l'implémentation, des jeux de données utilisés, et du processus d'expérimentation et analyse et comparaison des résultats obtenus avec les cinq méthodes sur différents réseaux.

Enfin, une conclusion synthétise les principales contributions du mémoire, discute des limites des approches étudiées et propose des perspectives de recherche future.

Chapitre 1

Fondements des réseaux complexes et des graphes

1.1 Introduction

La prédiction de liens consiste à anticiper l'apparition de connexions ou d'interactions entre des entités au sein d'un graphe.

Cette problématique, à la croisée de plusieurs disciplines, suscite un intérêt croissant dans des domaines aussi variés que les réseaux sociaux, la biologie ou encore les systèmes d'information et technologiques.

De nombreuses approches ont émergé ces dernières années pour répondre à cette question.

Ce chapitre présente les concepts fondamentaux de la théorie des graphes et introduit les différentes catégories de réseaux complexes, constituant ainsi une base théorique essentielle à la compréhension des mécanismes de prédiction de liens.

1.2 Réseaux complexes

Les graphes sont largement utilisés pour représenter des réseaux complexes issus du monde réel.

Ils modélisent des systèmes tels que les réseaux biologiques, sociaux, de communication, cibles ou le Web (WWW).

Un réseau complexe est un graphe composé de nœuds et de liens.

Les nœuds représentent des entités (individus, organisations, objets).

Les liens traduisent des interactions ou des relations entre ces entités.

Cette représentation permet de modéliser des phénomènes observés.

Par exemple : interactions protéine-protéine, réseaux de neurones, réseaux de gènes.

Elle aide à analyser les structures et dynamiques des systèmes réels.

Les sections suivantes présentent les principaux types de réseaux complexes.

1.2.1 Réseaux sociaux

Un graphe de réseau social sert à modéliser les interactions spécifiques entre individus ou groupes, où les nœuds représentent les acteurs (personnes, organisations, etc.) et les arêtes traduisent les relations qui les unissent.

Ces relations peuvent être variées, allant des liens d'amitié ou de parenté à des collaborations professionnelles ou à des intérêts communs [1].

Les plateformes de réseaux sociaux en ligne offrent une illustration concrète de ce concept : Facebook peut être modélisé par un graphe non orienté, les relations d'« amitié » étant réciproques, tandis que Twitter correspond à un graphe orienté, car les abonnements entre utilisateurs ne sont pas nécessairement mutuels [2].

La figure 1.1, ci-dessous, présente une structure d'un réseau social.

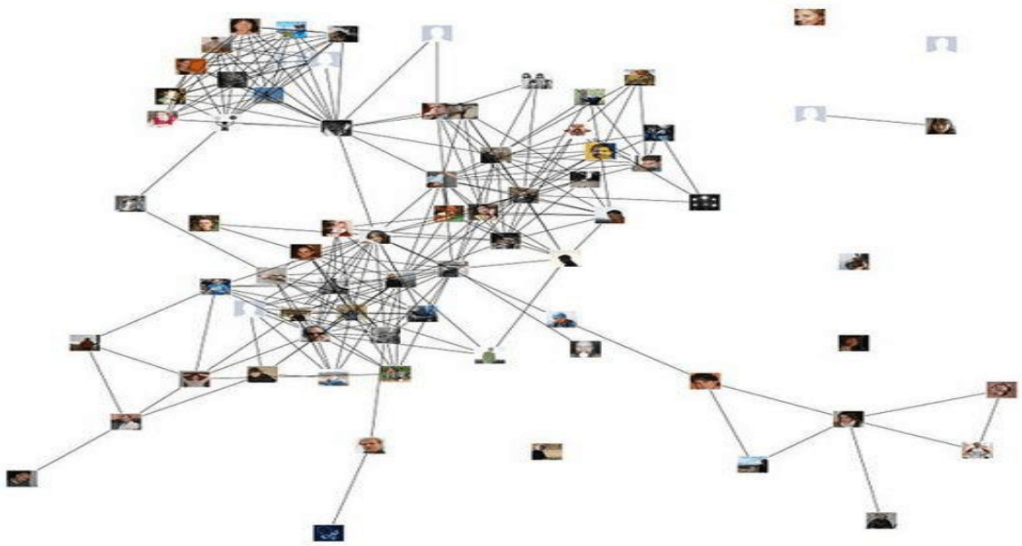

FIG. 1.1 : Structure d'un réseau social [3].

1.2.2 Réseaux biologiques

Les biologistes travaillent avec différents types de réseaux, notamment les réseaux métaboliques, qui modélisent les processus de transformation de matière et de production d'énergie au sein des organismes vivants [4],

les réseaux d'interactions protéiques [5], ainsi que les réseaux de régulation génétique [6].

On peut également mentionner les réseaux trophiques (ou réseaux alimentaires), qui décrivent les relations de prédation entre espèces.

L'ensemble de ces réseaux présente des caractéristiques typiques des réseaux complexes, telles qu'une distribution en loi de puissance, également appelée structure sans échelle, mise en évidence par des analyses topologiques approfondies [7], [8].

La Figure 1.2 illustre un exemple de réseau de régulation génétique, représentant les interactions de contrôle entre gènes, protéines et petites molécules.

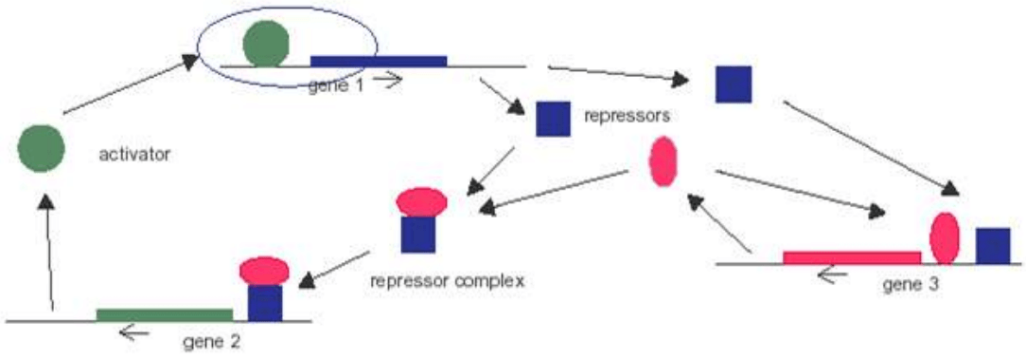

FIG. 1.2 : structure d'un réseau biologique [9].

1.2.3 Réseaux d'information

Les réseaux d'information constituent une autre catégorie importante de réseaux.

Le World Wide Web en est un exemple emblématique : les pages web y sont représentées par des nœuds, et les hyperliens permettant de passer d'une page à une autre forment les arêtes du graphe [10].

Ce réseau, qui compte plusieurs milliards de nœuds, est un graphe orienté, mais il ne comporte généralement pas de cycles fermés, en raison de l'absence de structure hiérarchique stricte dans l'organisation des sites web.

Un autre exemple de réseau d'information est constitué par les réseaux de citations scientifiques, où les publications sont les nœuds et les citations les liens [11].

Ces réseaux sont par nature acycliques, puisque les articles ne peuvent citer que des travaux antérieurs déjà publiés.

La figure 1.3, ci-dessous, présente un exemple de réseaux d'informations
.

FIG. 1.3 : Exemple de réseaux d'informations [12].

1.2.4 Réseaux technologiques

Les réseaux technologiques désignent des infrastructures construites par l'être humain.

Ils incluent notamment les réseaux électriques [13] , et les réseaux de télécommunication , ainsi que les réseaux de transport qu'ils soient routiers [14], ferroviaires [15] , ou aériens.

Parmi ces réseaux, Internet est sans doute le plus étudié.

Il s'agit d'un réseau informatique mondial où les ordinateurs, serveurs et routeurs représentent les nœuds, tandis que les connexions physiques, telles que les câbles en fibre optique, constituent les liens assurant l'interconnexion du système [16].

La Figure 1.4 illustre un exemple du réseau aérien mondial en 2009 .

FIG. 1.4 : réseau de transport aérien [17].

1.3 représentation graphique

Issue des mathématiques et de la recherche opérationnelle, la théorie des graphes s'est imposée comme un outil incontournable dans de nombreux domaines, qu'ils soient civils ou militaires.

Elle permet d'optimiser des itinéraires, par exemple dans les systèmes GPS, et de réduire les coûts dans diverses applications.

On la retrouve largement dans la gestion des réseaux de transport (ferroviaire, métropolitain ou aérien) ainsi que dans les réseaux d'énergie, qu'il s'agisse d'électricité, de gaz ou d'oléoducs.

Les réseaux de communication, tels qu'Internet, reposent également sur ces principes.

La théorie des graphes intervient dans l'organisation logistique des ports et des aéroports, ainsi que dans l'ordonnancement des tâches et la planification des activités.

Bien qu'elle ne constitue pas une branche indépendante des mathématiques, elle entretient des liens étroits avec d'autres disciplines telles que la topologie, la programmation linéaire et les probabilités.

Elle offre des outils puissants pour modéliser et résoudre des problèmes complexes, contribuant ainsi à une meilleure compréhension et à une maîtrise accrue des systèmes structurés.

Depuis 2002, son enseignement est intégré aux programmes scolaires, ce qui témoigne de son importance dans la formation scientifique.

La théorie des graphes connaît un essor constant dans les applications pratiques.

Dans cette section, nous présenterons ces notions fondamentales et les principales catégories de graphes utilisées en pratique.

1.3.1 Définition d'un graphe

Un graphe est une structure composée d'un ensemble de sommets (ou nœuds), noté V, et d'un ensemble d'arêtes (ou arcs, selon qu'ils soient orientés ou non), noté E, reliant certains couples de sommets.

On utilise généralement la notation suivante : $|V| = N$ (nombre de sommets), $|E| = M$ (nombre d'arêtes).

Les arêtes peuvent être pondérées à l'aide d'une fonction de poids $\omega : E \to \mathbb{R}^+$, permettant de représenter plus précisément l'intensité des interactions entre les sommets.

Le graphe pondéré est alors défini par le triplet $G = (V, E, \omega)$.

Le poids de l'arête reliant les sommets i et j est noté ω_{ij}.

Par convention, si une arête n'existe pas entre i et j, on lui attribue un poids nul, soit $\omega_{ij} = 0$ si $\{i, j\} \notin E$.

Dans un graphe non pondéré, les poids sont fixés à 1 pour toutes les arêtes existantes. Ainsi, dans ce cas, $\forall i, j \in V$, $\omega_{ij} \in \{0, 1\}$.

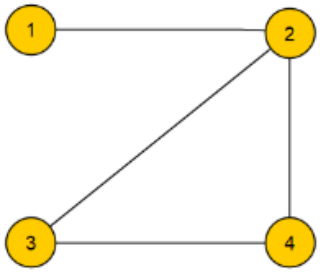

FIG. 1.5 : Un exemple de graphe

1.3.2 Propriétés des graphes

Nous présentons ci-dessous quelques propriétés fondamentales des graphes :

1. Le **poids** $\omega(i)$ d'un sommet i est défini comme la somme des poids des arêtes qui lui sont incidentes :

$$\omega(i) = \sum_{j \in V} \omega_{ij} \tag{1.1}$$

2. Soit un graphe $G = (V, E)$, La **distance** entre deux sommets v_i et v_j est la longueur du plus court chemin les reliant, notée $d(v_i, v_j)$.

   Si aucun chemin n'existe, on peut conventionnellement poser $d(v_i, v_j) = \infty$, ou bien $d(v_i, v_j) = 1$ selon le contexte.

Par exemple, la distance entre les nœuds 1 et 3 selon la figure 1.5 est : $d(1,3) = 2$

3. Dans un graphe $G = (V, E)$, le **voisinage** d'un sommet v_i, noté $\Gamma(v_i)$, est l'ensemble de ses voisins :

$$\Gamma(v_i) = \{v_j \in V \mid (v_i, v_j) \in E\} \tag{1.2}$$

Par exemple, dans la figure 1.5, les voisins du sommet 2 sont les sommets 1, 3 et 4.

4. Dans un graphe non orienté, le **degré** d'un sommet v, noté $d(v)$, correspond au nombre d'arêtes qui lui sont incidentes.

Il s'agit donc du nombre de sommets directement connectés à v, autrement dit, de ses voisins. Formellement :

$$d(v) = |\{u \in V \mid (v, u) \in E\}| \tag{1.3}$$

Dans le cas d'un graphe orienté, on distingue deux notions complémentaires :

- le *degré sortant* $d^+(v)$, qui représente le nombre d'arêtes partant du sommet v.
- le *degré entrant* $d^-(v)$, correspondant au nombre d'arêtes se dirigeant vers ce sommet.

Le degré constitue un indicateur de connectivité locale et joue un rôle clé dans la caractérisation des réseaux, en particulier pour identifier les nœuds centraux ou fortement connectés [18].

Par exemple, dans la figure 1.5, le degré du nœud 1 est $d(1) = 1$.

Le **degré moyen** $\langle k \rangle$ représente le nombre moyen de connexions (ou liens) par nœud dans un réseau. Il se calcule selon la formule suivante :

$$\langle k \rangle = \frac{2 \times L}{N} \tag{1.4}$$

où :

- L est le nombre total de liens dans le réseau.
- N est le nombre total de nœuds.

Chaque lien connectant deux nœuds, on multiplie par 2 le nombre total de liens pour obtenir la somme des degrés de tous les nœuds.

5. ASPL (Average Shortest Path Length) est La longueur moyenne des plus courts chemins entre tous les couples de nœuds.

1.3.3 Représentation matricielle

Représentation des graphes par des matrices :

1.2.3.1. Matrice d'adjacence

La matrice d'adjacence A est une matrice carrée de taille $n \times n$.

Pour un graphe non orienté $G = (V, E)$, cette matrice est symétrique et vérifie :

$$A_{ij} = \begin{cases} 1, & \text{si les sommets } i \text{ et } j \text{ sont adjacents} \\ 0, & \text{sinon} \end{cases} \tag{1.5}$$

Par exemple, la matrice d'adjacence du graphe illustré à la figure 1.5 est :

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

1.2.3.2. Matrice des degrés

La **matrice des degrés** d'un graphe est une matrice diagonale notée D, où chaque élément diagonal D_{ii} représente le degré du sommet i, c'est-à-dire le nombre d'arêtes incidentes à ce sommet.

Tous les autres éléments hors diagonale sont nuls.

Elle s'écrit formellement :

$$D_{ij} = \begin{cases} \deg(v_i), & \text{si } i = j \\ 0, & \text{sinon} \end{cases} \tag{1.6}$$

où $\deg(v_i)$ désigne le degré du sommet v_i.

Dans le cas des graphes pondérés, on remplace le degré par la somme des poids des arêtes connectées au sommet.

Dans les graphes orientés, on distingue la matrice des degrés *entrants* et celle des *sortants* selon la direction des arêtes.

Cette matrice joue un rôle central dans diverses applications, notamment dans la formulation du *Laplacien du graphe* [19].

1.2.3.3. Matrice d'incidence

La **matrice d'incidence** d'un graphe est une représentation matricielle qui décrit la relation entre les sommets et les arêtes du graphe.

Soit $G = (V, E)$ un graphe non orienté, avec $|V| = n$ sommets et $|E| = m$ arêtes.

La matrice d'incidence associée, notée B, est une matrice de taille $n \times m$ telle que :

$$B_{ij} = \begin{cases} 1 & \text{si le sommet } v_i \text{ est incident à l'arête } e_j \\ 0 & \text{sinon} \end{cases} \tag{1.7}$$

Dans le cas d'un graphe orienté, la matrice d'incidence est construite de la façon suivante :

$$B_{ij} = \begin{cases} 1 & \text{si } v_i \text{ est l'origine de l'arête } e_j \\ -1 & \text{si } v_i \text{ est l'extrémité de l'arête } e_j \\ 0 & \text{sinon} \end{cases} \tag{1.8}$$

Cette matrice est particulièrement utile en algèbre des graphes, notamment pour les calculs de flux, les cycles et les couplages [20].

1.3.4 Types de graphes

Dans cette section, nous présenterons quelques types de graphes couramment rencontrés :

1. **Un sous-graphe** est obtenu à partir d'un graphe initial en supprimant certains sommets ainsi que toutes les arêtes qui leur sont associées.

   On dit qu'un graphe $G' = (V', E')$ est un sous-graphe de $G = (V, E)$ si $V' \subseteq V$ et $E' \subseteq E$.

2. **Un sous-graphe partiel** conserve tous les sommets du graphe original, mais seulement une partie de ses arêtes.

   Autrement dit, $G' = (V, E')$ est un graphe partiel de $G = (V, E)$ si $E' \subseteq E$.

3. **Un sous-graphe induit** d'un graphe G est un graphe G' dont les sommets forment un sous-ensemble $V' \subseteq V$, et dont les arêtes sont uniquement celles reliant les sommets de V' dans G.

   Ainsi, G' préserve la structure de G restreinte à V'.

4. Un graphe G est dit **connexe** s'il existe un chemin entre n'importe quelle paire de sommets de G.

   Cela signifie que tous les sommets sont accessibles les uns aux autres.

   - **Composantes connexes** : Les composantes connexes d'un graphe sont des sous-graphes maximaux dans lesquels chaque paire de sommets est reliée par un chemin.

     Un graphe non connexe peut ainsi être décomposé en plusieurs composantes connexes distinctes.
   - **Nombre de composantes connexes** : Le nombre de composantes connexes, noté C, représente le nombre total de sous-graphes connexes présents dans un graphe non connexe.

     Il est défini par la formule :

     $$C = |\mathcal{C}| \tag{1.9}$$

     où $\mathcal{C}$ est l'ensemble des composantes connexes du graphe.

5. Un graphe est dit **complet** lorsque chaque paire de sommets distincts est connectée par une arête.

   Ainsi, dans un graphe complet $G = (V, E)$, tous les sommets sont mutuellement adjacents.

   Le graphe dans la figure 1.6 ci-dessus représente un graphe complet d'ordre 5 car on a chaque nœud à une interaction avec tous les autres nœuds.

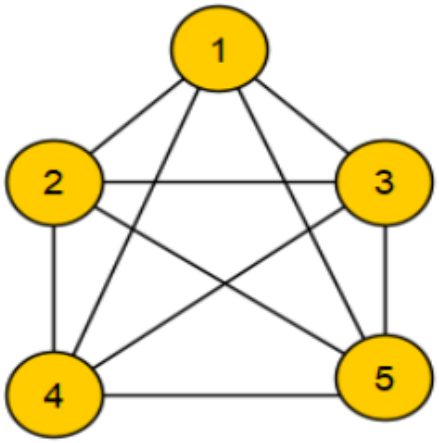

FIG. 1.6 : Exemple d'un graphe complet.

6. **Un graphe aléatoire** est généré selon un processus probabiliste.

   Il se caractérise généralement par une distribution des degrés suivant une loi puissance, une forte densité de triangles, et une densité globale relativement faible lorsque les degrés des sommets sont faibles par rapport à la taille totale du graphe.

   Ces graphes servent de modèles pour analyser de grands réseaux réels tels que les réseaux sociaux, biologiques, d'information ou technologiques.

   La figure 1.7 présente la structure d'un graphe aléatoire.

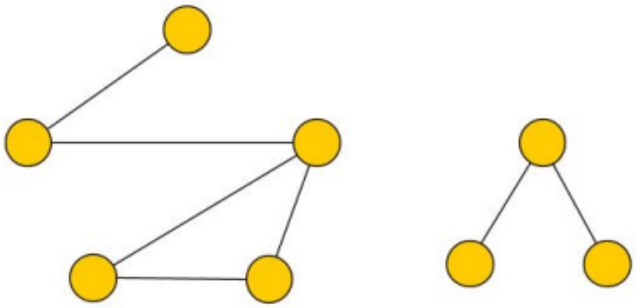

FIG. 1.7 : structure d'un graphe aléatoire .

1.3.5 Indicateurs d'un graphe

Dans cette section, nous présenterons quelques indicateurs d'un graphe couramment rencontrés :

1.3.5.1.diamètre (D)

Le **diamètre** d'un graphe $G = (V, E)$ est la plus grande distance entre deux sommets quelconques du graphe.

Plus précisément, il s'agit du maximum des longueurs des plus courts chemins entre toutes les paires de sommets.

On le note :

$$\text{diam}(G) = \max_{u,v \in V} d(u, v) \tag{1.10}$$

où $d(u, v)$ désigne la distance géodésique (longueur du plus court chemin) entre les sommets u et v.

Par exemple, pour le graphe illustré à la figure 1.5, on a :

$$\text{diam}(G) = 3$$

1.3.5.2.densité

La **densité** d'un graphe mesure la quantité de liens présents dans le réseau et permet d'évaluer sa cohésion.

cette mesure peut être abordée sous deux angles : une perspective *socio-centrée* ou *égo-centrée.*

- Dans une **analyse égo-centrée**, on s'intéresse à la densité des liens autour d'un nœud donné.

  Cela permet de mesurer l'influence de ce nœud sur la structure locale du réseau (sous-graphe induit par ce nœud et ses voisins).

- Dans une **analyse socio-centrée**, la densité est calculée à l'échelle du graphe entier.

  Elle reflète le niveau global de connectivité et de contrainte exercée par le réseau sur ses membres.

Toutefois, la comparaison directe de densité entre différents graphes peut être peu pertinente, notamment à cause de l'influence de la taille et de la structure du réseau sur cette mesure [21].

$$\text{den}(G) = \frac{2|E|}{|V|(|V|-1)} \tag{1.11}$$

1.3.5.3.centralité

La **centralité** permet d'évaluer l'importance structurelle d'un nœud dans un réseau.

on distingue principalement trois types :

- **Centralité de degré** : le nombre de connexions directes d'un nœud. Plus ce nombre est élevé, plus le nœud est central.
- **Centralité d'intermédiarité** : mesure le nombre de fois qu'un nœud apparaît sur les plus courts chemins reliant deux autres nœuds.

  Elle indique le pouvoir d'un nœud à contrôler la circulation d'information dans le réseau.
- **Centralité de proximité** : inverse de la somme des distances les plus courtes reliant un nœud aux autres.

  Elle représente la rapidité avec laquelle un nœud peut atteindre tous les autres[22].

1.3.6 caractéristiques d'un graphe

Dans cette section, nous présenterons quelques caractéristiques d'un graphe couramment rencontrées :

1.3.6.1.Coefficient de clustering élevé

Une caractéristique importante des réseaux complexes est leur tendance marquée au clustering, due à la propension naturelle de l'être humain à se regrouper en communautés.

Cette propriété se traduit par un **coefficient de clustering élevé**, révélant une forte densité locale de connexions.

Une question pertinente que soulève cette propriété est la suivante : **les amis de mes amis deviennent-ils aussi mes amis ?**

Autrement dit, un réseau présente un clustering important si, lorsqu'un nœud X est connecté à un nœud Y, et que Y est lui-même connecté à un nœud Z, alors il est probable que X et Z soient également connectés. Ce phénomène est également appelé transitivité [23].

la figure 1.8 ci-dessus représente un Coefficient de clustering élevé .

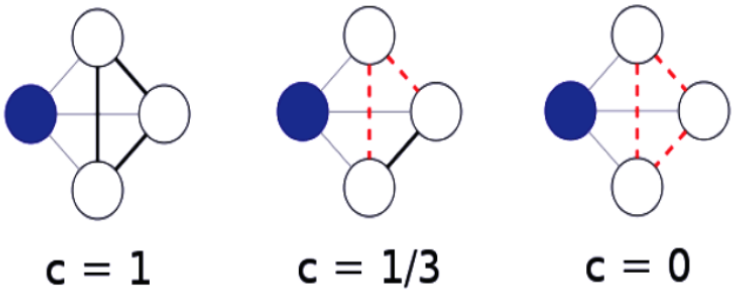

FIG. 1.8 : Coefficient de clustering élevée [24].

1.3.6.2.Distribution des degrés en loi de puissance

Une autre propriété remarquable des réseaux complexes, notamment les réseaux sociaux, est que la **distribution des degrés suit une loi de puissance** [23].

Cela signifie que plus le degré est élevé, plus le nombre de nœuds possédant ce degré est faible.

Cette distribution particulière implique que **la majorité des nœuds ont un faible degré**, tandis qu'un petit nombre de nœuds très connectés, appelés hubs, possèdent un degré élevé.

Les réseaux présentant ce type de distribution sont qualifiés de **réseaux invariants d'échelle** (scale-free).

Mathématiquement, cette distribution s'exprime par la relation suivante :

$$P(k) = k^{-a} \tag{1.12}$$

où k le nombre de nœuds qui ont le degré a .

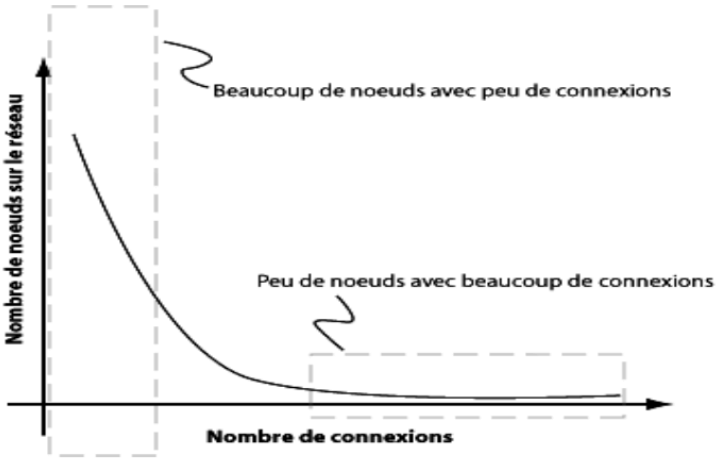

FIG. 1.9 : Distribution de degrés en loi de puissance [25].

1.3.6.3.Structure en communautés

Une autre caractéristique essentielle des réseaux sociaux est leur **structure en communautés** [23].

Cela signifie que les nœuds du réseau se regroupent en sous-ensembles densément connectés, appelés communautés, au sein desquels les liens sont nombreux, tandis que les connexions entre communautés sont plus rares et appelées ponts.

Ce phénomène reflète souvent des regroupements d'individus partageant des intérêts communs, des profils similaires ou entretenant des relations fortes.

Ainsi, la socialisation dans ces réseaux s'accompagne d'une **tendance à l'affiliation**, où les nœuds ayant des caractéristiques similaires ont davantage de chances d'être connectés.

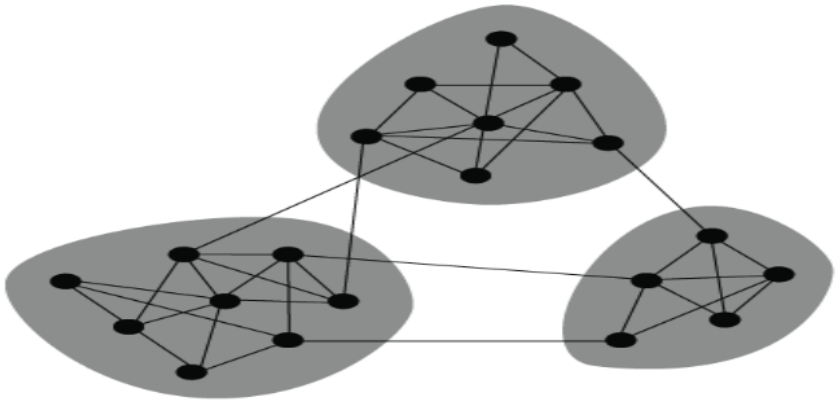

FIG. 1.10 : Structure en communautés [26].

1.4 Conclusion

Ce premier chapitre a posé les bases théoriques indispensables à la compréhension de la problématique de prédiction de liens.

En introduisant les notions fondamentales de la théorie des graphes, telles que les sommets, les arêtes, les chemins ,etc.

il a permis de structurer le cadre conceptuel dans lequel s'inscrit notre étude.

Par ailleurs, une attention particulière a été portée à la typologie des réseaux complexes, qu'il s'agisse de réseaux sociaux, biologiques, technologiques ou informationnels, chacun présentant des propriétés spécifiques influençant les mécanismes de prédiction.

La maîtrise de ces notions est essentielle pour appréhender les modèles et les méthodes qui seront présentés dans les chapitres suivants.

En effet, les algorithmes de prédiction de liens s'appuient étroitement sur les caractéristiques structurelles des graphes et sur les dynamiques propres à chaque type de réseau.

Dans le chapitre suivant, nous proposerons un état de l'art approfondi des approches développées pour répondre à la problématique de prédiction de liens.

Nous explorerons à la fois les méthodes heuristiques classiques et les approches plus récentes reposant sur l'apprentissage automatique et les représentations vectorielles des graphes.

Cette analyse nous permettra de mieux situer les travaux existants et d'identifier les principaux enjeux encore ouverts dans ce champ de recherche.

Chapitre 2

Etat de l'art ; prédiction de liens

2.1 Introduction

La prédiction de liens est une tâche qui vise à anticiper les relations et interactions futures au sein d'un réseau.

Les techniques d'apprentissage automatique sont largement utilisées pour prédire des liens manquants ou encore inexistants entre les nœuds d'un graphe.

L'objectif principal est donc de prévoir l'apparition de connexions qui ne sont pas encore observables dans l'état actuel du réseau.

En raison de son importance, cette problématique a suscité un vif intérêt de la part des chercheurs issus de divers domaines, ce qui a conduit au développement d'un grand nombre de méthodes au fil des dernières années.

Ces approches varient en fonction de plusieurs facteurs, tels que l'évolution du réseau, le type de données exploitées, ou encore la quantité d'informations disponibles.

Dans ce chapitre, nous proposons une étude approfondie des différentes approches de prédiction de liens.

Pour cela, nous apportons plusieurs contributions à cette revue.

Tout d'abord, nous réalisons une analyse détaillée de l'état de l'art des méthodes existantes.

Ensuite, nous présentons une taxonomie permettant de classer ces techniques en fonction de leur méthodologie et du volume d'information qu'elles mobilisent.

Enfin, nous menons une étude empirique en appliquant les principales méthodes à différents réseaux présentant des caractéristiques variées, afin d'évaluer et de comparer leurs performances.

2.2 Prédiction de liens

La prédiction de liens consiste à anticiper l'apparition de relations dans un graphe ou un réseau.

Elle vise soit à identifier des liens manquants ou cachés dans des graphes statiques, soit à estimer la probabilité d'apparition de liens futurs dans des graphes dynamiques.

En raison de ses nombreuses applications, notamment en biologie, en physique, en informatique et dans bien d'autres disciplines, ce domaine a suscité un fort intérêt scientifique.

De nombreuses études y ont été consacrées, menant au développement et à l'application de diverses méthodes de prédiction adaptées à différents types de réseaux.

La majorité des travaux existants se focalisent sur la prédiction de l'existence de liens potentiels. Toutefois, la prédiction de liens englobe plusieurs tâches complémentaires, telles que l'estimation du poids des liens, la prédiction de leur type, ou encore celle de leur cardinalité [27].

2.2.1 Définition formelle de la prédiction de liens

Étant donné un graphe $G_t = (V, E_t)$ à un instant t, la prédiction de liens vise à prévoir l'ensemble des nouveaux liens E' susceptibles d'apparaître dans l'intervalle de temps $[t, t_2]$, où $t_2 > t$.

Le réseau au temps t_2 est alors représenté par :

$$G_{t_2} = (V, E_{t_2}) \quad \text{avec} \quad E_{t_2} = E_t \cup E' \tag{2.1}$$

Il est important de noter que, dans le problème de prédiction de liens, l'ensemble des nœuds V reste généralement fixe au cours du temps, tandis que l'ensemble des liens E évolue au fur et à mesure de l'apparition de nouvelles connexions [28].

De nombreux jeux de données réels peuvent être naturellement modélisés sous forme de graphes, où les nœuds V représentent des instances et les arêtes E correspondent aux relations entre ces instances [29].

Les liens à prédire peuvent être soit des liens manquants entre des nœuds liés, soit des liens potentiels destinés à apparaître dans le futur [29].

Ainsi, l'objectif de la prédiction de liens est d'anticiper l'existence de ces liens futurs ou absents dans le graphe actuel.

La Figure 2.1 illustre ce processus au fil du temps.

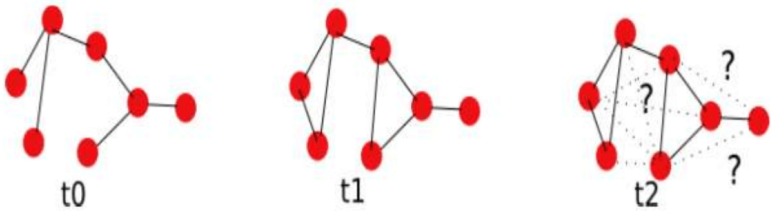

FIG. 2.1 : la prédiction de lien dans les instants t1, t2 [30].

2.2.2 Domaine d'application de la prédiction de liens

Les techniques de prédiction de liens trouvent des applications dans de nombreux domaines variés.

Tout système où les entités interagissent de manière structurée peut potentiellement bénéficier de ces techniques.

Par exemple, elles sont utilisées pour améliorer la sélection d'utilisateurs similaires dans les systèmes de recommandation basés sur le filtrage collaboratif, permettant ainsi d'obtenir des recommandations plus pertinentes [31].

Une application similaire concerne les réseaux sociaux, devenus omniprésents dans la société moderne : les utilisateurs s'attendent à des mécanismes simples et efficaces pour naviguer parmi l'immense nombre d'utilisateurs enregistrés.

La plupart des plateformes sociales utilisent ainsi des techniques de prédiction de liens pour suggérer automatiquement de nouvelles connaissances avec un haut degré de précision.

Dans le domaine de la biologie, la prédiction de liens est utilisée pour identifier des interactions potentielles entre des paires de protéines dans les réseaux d'interaction protéine-protéine (PPI) [32].

Une autre application concerne la prédiction de collaborations futures dans les réseaux de co-auteurs scientifiques.

Grâce aux données publiques issues des bases de journaux scientifiques, les méthodes de prédiction de liens permettent non seulement d'anticiper de nouvelles collaborations, mais aussi de mieux comprendre les structures des domaines de recherche [33].

2.2.3 les approches de prédiction de liens

la figure montre les approches de prédiction de liens basées sur la similarité :

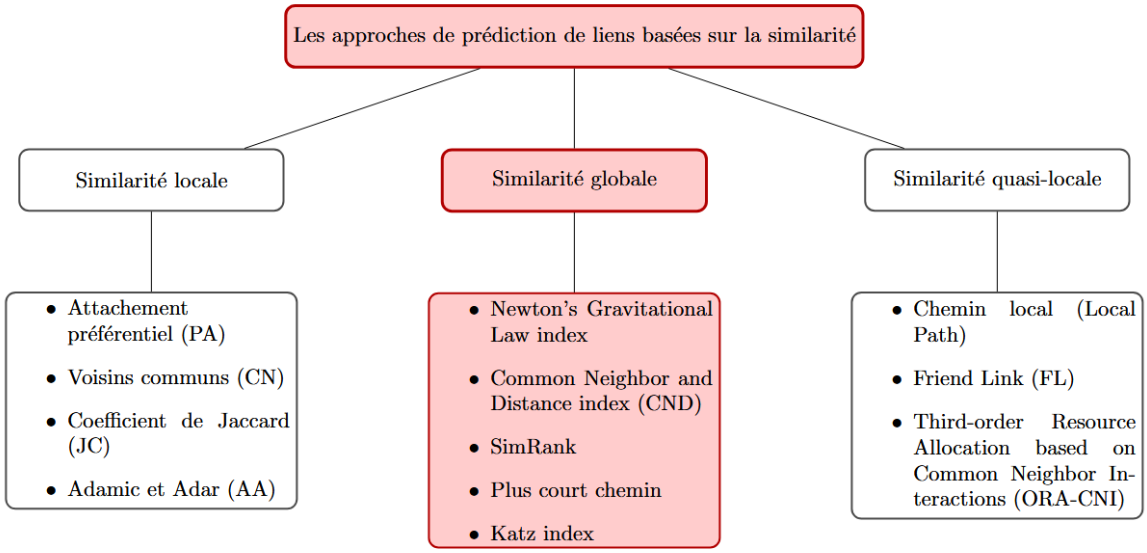

FIG. 2.2 : les approches de prédiction de liens Basées sur la similarité

2.2.4 Les méthodes basées sur la similarité (motif topologique)

Considérons un réseau social simple, ne contenant aucun attribut ni sur les nœuds, ni sur les liens.

Il existe de nombreuses mesures permettant de calculer les similarités entre les paires de nœuds, la majorité se concentrant sur les informations issues de la structure ou de la topologie du réseau.

Les travaux antérieurs ont montré que la structure d'un réseau joue un rôle important dans l'évolution des liens au fil du temps.

En conséquence, de nombreuses mesures de similarité basées uniquement sur la topologie d'un réseau social ont été proposées.

Dans la section suivante, nous présentons une vue systématique des mesures les plus couramment utilisées dans la prédiction de liens.

Il est important de noter que ces mesures se répartissent généralement en trois grandes catégories :

- Les mesures locales, basées sur le voisinage immédiat des nœuds .
- Les mesures globales, basées sur les distances entre les nœuds .
- Les mesures quasi-locales, basées sur les marches aléatoires.

2.2.4.1.Mesures de similarité locales

Dans les réseaux sociaux, les individus ont tendance à établir des relations avec des personnes proches d'eux.

Le voisinage immédiat joue donc un rôle crucial dans la formation de nouveaux liens.

Pour cette raison, de nombreuses mesures ont été définies par les chercheurs, basées uniquement sur les voisins directs d'un nœud.

Ces mesures attribuent un score de similarité aux paires de nœuds non connectés en se fondant uniquement sur leur voisinage respectif, sans prendre en compte l'ensemble de la structure du réseau.

Une valeur élevée du score de similarité entre deux nœuds non connectés indique une forte probabilité que ces nœuds soient reliés dans le futur.

- Attachement préférentiel (PA)

  il existe une forte probabilité que deux nœuds se connectent si ces nœuds, appelés également hubs , sont déjà reliés à un grand nombre d'autres nœuds.

  Cette idée est liée au principe du rich-get-richer (les riches s'enrichissent).

  Le score associé à la probabilité d'existence d'un lien entre deux nœuds x et y est défini comme suit :

$$PA(x, y) = |\Gamma(x)| \times |\Gamma(y)| \tag{2.2}$$

  où $\Gamma(x)$ désigne l'ensemble des voisins du nœud x.

  Cependant, l'attachement préférentiel présente l'inconvénient de favoriser des valeurs de similarité élevées entre des utilisateurs déjà très connectés, au détriment de ceux qui le sont peu.

  Ce biais est dû au fait que les relations dépendent uniquement du degré de connectivité des nœuds.

  dans certaines applications, l'objectif est plutôt de favoriser la connexion de nœuds peu connectés pour améliorer la couverture du réseau.

De plus, une autre limite de cette méthode est qu'elle peut encourager la création de multiples liens superflus entre hubs, au détriment d'une maximisation efficace de la connectivité globale du réseau Newman [34] et Barabási et al. [35] .

- Voisins communs (CN)

  une mesure parmi les plus utilisées dans le problème de prédiction de liens, principalement en raison de sa simplicité.

  Pour deux nœuds x et y, le score CN est défini comme le nombre de nœuds avec lesquels x et y partagent une connexion directe, autrement dit leurs voisins communs.

  Un nombre élevé de voisins communs accroît la probabilité d'apparition d'un lien futur entre x et y. La mesure est définie par la formule suivante :

$$CN(x, y) = |\Gamma(x) \cap \Gamma(y)| \tag{2.3}$$

  Comme mentionné précédemment, cette méthode repose sur l'idée que plus deux utilisateurs partagent de voisins, plus ils sont susceptibles d'être connectés.

  Cependant, à l'instar de l'attachement préférentiel, elle présente l'inconvénient de favoriser les utilisateurs possédant un grand nombre de voisins, attribuant ainsi des similarités faibles, voire nulles, entre les utilisateurs peu connectésNewman [34] .

- Coefficient de Jaccard (JC)

  Le coefficient de Jaccard constitue une amélioration de la méthode des voisins communs.

  Il mesure la similarité entre deux nœuds en divisant le nombre de voisins communs par le nombre total de voisins distincts de ces deux nœuds.

  Cette approche attribue des scores plus élevés aux paires de nœuds présentant une proportion importante de voisins communs par rapport à l'ensemble de leurs voisins.

  La mesure est définie par la formule suivante :

$$JC(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \tag{2.4}$$

Contrairement aux deux méthodes précédentes, le coefficient de Jaccard présente l'avantage de limiter l'influence des utilisateurs disposant d'un grand nombre de voisins.

- Adamic et Adar (AA)

  À l'origine, la méthode d'Adamic et Adar a été proposée pour évaluer la similarité entre deux pages web, en se basant sur les items qu'elles ont en commun.

  La particularité de cette approche est de donner plus de poids aux items rares, c'est-à-dire ceux partagés par peu de pages, contrairement aux items très fréquents qui sont moins informatifs.

  Par la suite, cette mesure a été largement adoptée dans le contexte des réseaux sociaux.

  La mesure Adamic-Adar (AA) entre deux nœuds x et y s'exprime de la manière suivante :

$$AA(x,y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log |\Gamma(z)|} \tag{2.5}$$

  Dans cette formule, au lieu de considérer des items, on considère les voisins communs des nœuds x et y.

  L'idée principale est d'accorder un poids plus important aux voisins communs ayant peu de connexions : plus un voisin commun est rare (peu connecté), plus sa contribution à la similarité est élevée.

TAB. 2.1 : Quelques caractéristiques des mesures de similarité locale [36]

Méthodes	**Normalisée**	**Complexité**	**Caractéristiques**
PA	Non	$O(n^2)$	Simple ; favorise les nœuds ayant un degré élevé
CN	Non	$O(n^2)$	Simple et intuitive
JC	Oui	$O(n^2)$	Compare la proportion de voisins communs par rapport au nombre total de voisins
AA	Non	$O(2n^2)$	Donne plus de poids aux voisins communs ayant peu de connexions

Dans le tableau 2.1, nous comparons plusieurs mesures populaires de similarité basées sur le voisinage, selon trois critères principaux : la normalisation, la complexité temporelle et leurs caractéristiques.

Il existe trois mesures non normalisées, c'est-à-dire que la similarité qu'elles calculent n'a de signification que lorsqu'on établit un classement relatif des scores ; elles ne donnent pas d'information précise sur la topologie du graphe, comme le degré des nœuds ou la proportion de voisins communs.

La complexité temporelle est également un facteur déterminant dans le choix d'une mesure, notamment pour les réseaux sociaux de grande taille.

En supposant que le nombre moyen de voisins par nœud est n, trouver tous les voisins d'un nœud est en $O(n)$, et l'intersection ou l'union de deux ensembles de voisins est en $O(n^2)$.

Les mesures CN (Voisins Communs), PA (Attachement Préférentiel) et JC (Coefficient de Jaccard) ont toutes une complexité temporelle de $O(n^2)$, car elles nécessitent de calculer l'intersection ou l'union de deux ensembles.

La mesure AA (Adamic-Adar) nécessite, en plus de l'intersection, de retrouver les degrés des voisins communs, ce qui porte sa complexité à $O(2n^2)$.

2.2.4.2.Mesures de similarité globales

Contrairement aux mesures de similarité locales, les mesures globales nécessitent la connaissance de l'ensemble de la topologie du réseau social.

Ces approches reposent généralement sur l'hypothèse que l'existence de multiples chemins de différentes longueurs (troisième degré ou plus) entre deux nœuds peut favoriser l'établissement d'une relation directe entre eux dans le futur.

- Newton's Gravitational Law Index

  Méthode inspirée de la loi gravitationnelle, où la force entre deux nœuds est : proportionnelle au produit des centralités, et inversement proportionnelle à leur distance.

  L'indice est défini par :

$$S_{xy}^{\text{NGLI}} = \frac{C_D(x) \cdot C_D(y)}{SP(x,y)} \tag{2.6}$$

  où C_D représente le degré de centralité du nœud, et $SP(x,y)$ la longueur du plus court chemin entre x et y [37] .

- Common Neighbor and Distance index (CND)

  combine deux propriétés classiques utilisées en prédiction de liens : le nombre de voisins communs entre deux nœuds et la distance qui les sépare dans le graphe.

  Il est défini comme suit :

$$S_{CND}^{xy} = \begin{cases} \frac{CN^{xy}+1}{2} & \text{si } \Gamma(x) \cap \Gamma(y) \neq \emptyset \\ \frac{1}{d^{xy}} & \text{sinon} \end{cases} \tag{2.7}$$

  où :

  CN^{xy} est le nombre de voisins communs entre les nœuds x et y ;

  $\Gamma(x)$ désigne l'ensemble des voisins de x ;

  d^{xy} est la distance (plus court chemin) entre x et y.

  Cette mesure vise à améliorer la robustesse des scores de similarité en considérant à la fois la topologie locale et globale.

  Les valeurs obtenues ne sont pas normalisées.

  La complexité est en $O(nk)$, où k est le degré moyen du graphe Yang et Zhang [38] .

- SimRank

  repose sur l'hypothèse que deux nœuds sont similaires s'ils sont connectés à des nœuds similaires [39].

  La mesure est définie récursivement par :

$$score_{SR}(x,y) = \begin{cases} 1 & \text{si } x = y \\ \gamma \cdot \frac{\sum_{a\in\Gamma(x)} \sum_{b\in\Gamma(y)} simRank(a,b)}{|\Gamma(x)|\cdot|\Gamma(y)|} & \text{sinon} \end{cases} \tag{2.8}$$

  où $\Gamma(x)$ désigne l'ensemble des voisins du nœud x et $\gamma \in [0,1]$ est un facteur d'atténuation.

  SimRank peut également être interprété à travers le concept de marche aléatoire : $score_{SR}(x,y)$ représente le moment où deux marcheurs aléatoires partant respectivement des nœuds x et y se rencontrent [40].

  Le calcul de SimRank implique un processus d'expansion récursif de complexité $\mathcal{O}(k^{2l})$, où k est le degré moyen et l la profondeur de la récursion.

La complexité pour chaque paire de nœuds est alors de $\mathcal{O}(k^{2l+2})$, menant à une complexité globale de $\mathcal{O}(v^2k^{2l+2})$ pour un graphe de v nœuds [41].

- Plus court chemin (Shortest Path)

  La mesure la plus directe pour évaluer la similarité entre deux nœuds est la distance la plus courte qui les sépare.

  Plus précisément, on initialise deux ensembles : $S = \{x\}$ et $D = \{y\}$.

  À chaque itération, on élargit S et D en y ajoutant leurs voisins directs respectifs.

  Le processus s'arrête dès que l'intersection $S \cap D$ devient non vide.

  La distance correspond alors au nombre d'itérations nécessaires.

  La similarité est définie comme l'opposé de la distance :

$$S(x, y) = -|d(x, y)| \tag{2.9}$$

  où $d(x, y)$ est la longueur du plus court chemin entre x et y. Ainsi, une distance courte conduit à un score élevé, et une distance longue à un score faible [42] .

- Indice de Katz (Katz Index)

  L'indice de Katz prend en compte tous les chemins possibles entre deux nœuds, tout en pondérant chaque chemin par une valeur décroissante en fonction de sa longueur : les chemins plus courts ont un poids plus élevé.

$$\text{Katz}(x, y) = \sum_{l=1}^{\infty} \beta^l \times \left(A^l\right)_{xy} \tag{2.10}$$

  où A est la matrice d'adjacence du graphe, β est un paramètre de décroissance, et l est la longueur du chemin[43].

TAB. 2.2 : Caractéristiques de quelques mesures de similarité globale

Méthodes	Normalisée	Complexité	Caractéristiques
NGLI	non	$O(n^2)$	Inspirée de la loi de gravitation ; pondère la similarité selon les degrés des nœuds et la distance ; favorise les nœuds proches ayant un fort degré
CND	non	$O(n^2)$	Combine les voisins communs avec la distance entre les nœuds ; affine la mesure CN en intégrant une pénalité selon l'éloignement
SimRank	non	$O(n^2)$ à $O(n^4)$	Basée sur le principe que deux nœuds sont similaires si leurs voisins le sont ; nécessite une itération jusqu'à convergence
Plus court chemin	non	$O(n \log n)$ à $O(n^2)$	Mesure de distance globale ; suppose que plus deux nœuds sont proches, plus ils sont susceptibles d'être liés
Katz Index	non	$O(n^3)$	Prend en compte tous les chemins entre deux nœuds, avec une pénalisation exponentielle selon leur longueur ; capte des relations indirectes

Dans le tableau 2.2, nous comparons plusieurs autres mesures de similarité globales, selon trois critères principaux : la normalisation, la complexité temporelle et leurs caractéristiques distinctives.

Certaines de ces mesures prennent en compte non seulement les voisins directs, mais aussi les chemins plus longs ou la structure globale du graphe.

Cela se traduit généralement par une complexité algorithmique plus élevée, mais aussi par une meilleure capacité à capturer des relations latentes.

2.2.4.3.Approches quasi-locales

Les méthodes quasi-locales ont récemment émergé pour offrir un compromis entre les approches locales et globales.

Elles sont presque aussi efficaces à calculer que les méthodes locales, tout en intégrant davantage d'informations topologiques, comme le font les méthodes globales.

Contrairement aux mesures strictement locales, les approches quasi-locales ne se limitent pas uniquement aux voisins immédiats des nœuds, sans pour autant traiter toutes les paires de nœuds arbitraires du réseau.

Certaines de ces méthodes exploitent l'information globale du graphe, mais leur complexité temporelle reste inférieure à celle des approches globales.

Leur coût algorithmique dépend souvent de paramètres spécifiques, tels que le nombre d'itérations ou la longueur maximale des chemins considérés.

- Chemin local (Local Path)

  La mesure Local Path exploite non seulement les chemins de longueur 2 (voisins communs) mais aussi ceux de longueur 3 pour estimer la similarité entre deux nœuds.

  Les chemins plus courts sont jugés plus importants, et un facteur d'ajustement ϵ est introduit pour pondérer l'impact des chemins de longueur 3.

  $$\mathrm{LP}(x, y) = A^2_{xy} + \epsilon \times A^3_{xy} \tag{2.11}$$

  où A est la matrice d'adjacence du graphe, et ϵ est un paramètre de pondération ($\epsilon \ll 1$) [44].

- FriendLink (FL)

  FriendLink (FL) est une mesure quasi-locale basée sur le comptage des chemins entre deux nœuds, avec une pondération selon leur longueur, similaire à l'indice de chemin local.

  Elle introduit un facteur d'atténuation $\frac{1}{i-1}$ et une normalisation par le nombre de chemins de même longueur dans un graphe complet.

  La similarité entre deux nœuds x et y est donnée par :

  $$s(x, y) = \sum_{i=2}^{l} \frac{1}{i-1} \cdot (A^i)_{x,y} \cdot \prod_{j=2}^{i} \frac{1}{|V| - j} \tag{2.12}$$

  Cette méthode présente une complexité algorithmique de $\mathcal{O}(lv^2k)$ [45].

- Third-order Resource Allocation based on Common Neighbor Interactions (ORA-CNI)

  ORA-CNI est une extension de l'allocation de ressources fondée sur les interactions entre voisins communs, prenant également en compte les chemins de longueur trois.

  Elle redéfinit la distribution des ressources entre deux nœuds séparés par une distance de trois.

La similarité entre deux nœuds x et y est calculée comme suit :

$$s(x,y) = \sum_{z \in \Gamma_x \cap \Gamma_y} \frac{1}{|\Gamma_z|} + \sum_{\substack{e_{i,j} \in E \\ |\Gamma_i| < |\Gamma_j| \\ i \in \Gamma_x, j \in \Gamma_y}} \left(\frac{1}{|\Gamma_i|} - \frac{1}{|\Gamma_j|} \right) + \beta \sum_{[x,p,q,y] \in \text{paths}_3^{x,y}} \frac{1}{|\Gamma_p||\Gamma_q|} \tag{2.13}$$

où β est un facteur d'atténuation ajustant l'influence des chemins à trois sauts.

L'ajout de ce troisième terme augmente la complexité algorithmique à $\mathcal{O}(vk^3(k^2+k^3)) = \mathcal{O}(vk^6)$ [46].

Finalement, il est important de souligner qu'il existe un grand nombre d'algorithmes de prédiction de liens.

Toutefois, plusieurs études expérimentales ont montré qu'aucune mesure de similarité n'offre systématiquement de bonnes performances pour tous les types de réseaux complexes.

2.3 Mesures de performance

Nous avons choisi d'évaluer les performances à l'aide de l'Accuracy, du rappel, de la précision et de la F-mesure [47].

Nous allons détailler ci-dessous le principe de chacune de ces mesures dans le contexte de la prédiction de liens, dans le but de comparer les résultats obtenus par les algorithmes indice de Katz, le plus court chemin, SimRank et Newton's Gravitational Law Index, Common Neighbor and Distance index (CND).

Il est important de noter que nous pouvons évaluer la performance des algorithmes uniquement si nous disposons d'une nouvelle capture du réseau complexe.

Celle-ci permet de comparer l'état réel du réseau à l'état prédit.

Dans ce contexte, nous distinguons quatre types de prédictions, représentés dans la **matrice de confusion** suivante :

	G-test : +	G-test : -
Prédit : +	TP (Vrai positif)	FP (Faux positif)
Prédit : -	FN (Faux négatif)	TN (Vrai négatif)

Tableau 3.8 — Matrice de confusion

Les définitions sont les suivantes :

- **TP (True Positive)** : liens prédits (liens dans G-pred) et effectivement apparus dans (G-test).
- **FP (False Positive)** : liens prédits (liens dans G-pred) mais qui n'apparaissent pas dans (G-test).
- **FN (False Negative)** : liens non prédits (non liens dans G-pred) mais qui apparaissent dans (G-test).
- **TN (True Negative)** : liens non prédits (non liens dans G-pred) et absents également de (G-test).

la figure 2.2 illustre les différents types de liens pour calculer les performances d'un algorithme de prédiction des liens :

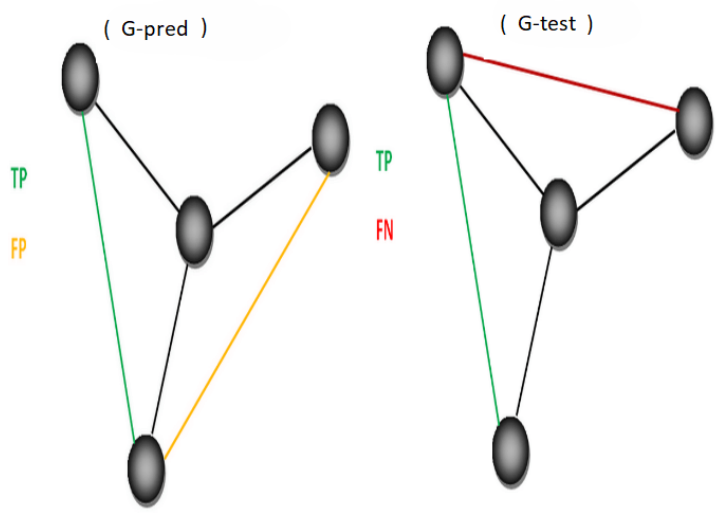

FIG. 2.3 : Les différents types des liens : TP, TN , FP, FN [48].

2.3.1 Rappel

Dans le contexte de la prédiction de liens, le **rappel** (ou sensibilité) correspond au rapport entre :

- le nombre de liens correctement prédits (prévus et réellement observés dans la nouvelle capture du réseau social).
- et le nombre total de liens réellement présents dans cette nouvelle capture.

Il est calculé à l'aide de la formule suivante :

$$\text{Rappel} = \frac{TP}{TP + FN} \tag{2.14}$$

2.3.2 Précision

La **précision** mesure la proportion de liens prédits qui sont effectivement apparus dans la nouvelle capture du réseau social, par rapport à l'ensemble des liens prédits.

Elle est donnée par :

$$\text{Précision} = \frac{TP}{TP + FP} \tag{2.15}$$

2.3.3 F-mesure

La **F-mesure** combine le rappel et la précision en calculant leur moyenne harmonique.

La formule utilisée est :

$$\text{F-mesure} = \frac{2 \times \text{Rappel} \times \text{Précision}}{\text{Rappel} + \text{Précision}} \tag{2.16}$$

2.3.4 Exactitude (Accuracy)

Alors que la précision et le rappel se concentrent uniquement sur les vrais liens positifs, et que la spécificité cible les vrais négatifs, l'**exactitude** est une mesure globale qui prend en compte les deux types de prédictions correctes.

Elle est définie comme :

$$\text{Exactitude} = \frac{TP + TN}{TP + FP + FN + TN} \tag{2.17}$$

2.4 conclusion

Dans ce chapitre, nous avons procédé à une revue approfondie des principales méthodes de prédiction de liens, en mettant en lumière la richesse

et la diversité des approches développées dans ce domaine.

Cette exploration a permis de souligner les fondements théoriques et pratiques de chaque méthode, tout en illustrant leurs domaines d'application respectifs.

Afin de structurer cette diversité, nous avons proposé une taxonomie permettant de classer les techniques selon leur nature méthodologique et le type d'information qu'elles exploitent.

Cette classification contribue à une meilleure compréhension comparative des approches et constitue un outil précieux pour guider le choix des méthodes en fonction des spécificités des réseaux étudiés.

Par ailleurs, une étude empirique a été menée sur plusieurs types de réseaux, afin d'évaluer les performances relatives des méthodes analysées.

Les résultats obtenus ont mis en évidence les points forts et les limites de chaque approche, révélant les contextes dans lesquels elles se montrent les plus efficaces ou, au contraire, les moins adaptées.

Cette analyse critique de l'état de l'art constitue une étape essentielle dans la construction de notre démarche scientifique.

Elle nous permet d'identifier les méthodes les plus pertinentes à mettre en œuvre dans le cadre de notre étude expérimentale.

Dans le chapitre suivant, nous appliquerons concrètement certaines de ces techniques à des jeux de données réels, dans le but de valider empiriquement leur efficacité et de tirer des enseignements pratiques pour la tâche de prédiction de liens.

Chapitre 3

Implémentation et expérimentations

3.1 Introduction

Dans ce chapitre, nous présentons notre projet ainsi que les différentes expérimentations réalisées pour évaluer et comparer plusieurs mesures de similarité globale, telles que le Plus Court Chemin, l'indice de Katz, Newton's Gravitational Law Index, Common Neighbor and Distance Index (CND) , SimRank .

dans le contexte de la prédiction de liens dans les réseaux complexes.

3.2 Environnement matériel

Les expérimentations ont été effectuées sur un PC avec 8 Go de RAM , un Processeur Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz , fonctionnant sous le système d'exploitation Windows 64 bits.

3.3 Environnement logiciel

Pour réaliser l'implémentation de Chapitre 3, nous avons utilisé les environnements de développement suivants :

3.3.1 Python 3.8 (Anaconda3)

Python est un langage de programmation puissant et facile à apprendre.

Il propose des structures de données de haut niveau et permet une approche simple, mais efficace de la programmation orientée objet.

Grâce à sa syntaxe élégante, son typage dynamique et son interprétation directe, Python est particulièrement adapté à l'écriture de scripts ainsi qu'au développement rapide d'applications.

Aujourd'hui, il s'agit du langage de programmation le plus utilisé au monde.

L'interpréteur Python et sa large bibliothèque standard sont disponibles librement pour toutes les plateformes majeures à l'adresse suivante : `https://www.python.org/`.

3.3.2 Spyder

Spyder est un environnement de développement scientifique puissant, écrit en Python pour Python, conçu pour les scientifiques, ingénieurs et analystes de données.

Il offre une combinaison unique de fonctionnalités avancées pour l'édition de code, l'analyse, le débogage et le profilage, ainsi qu'un environnement interactif d'exploration de données avec de fortes capacités de visualisation.

Spyder est un IDE libre intégré dans la distribution Anaconda.

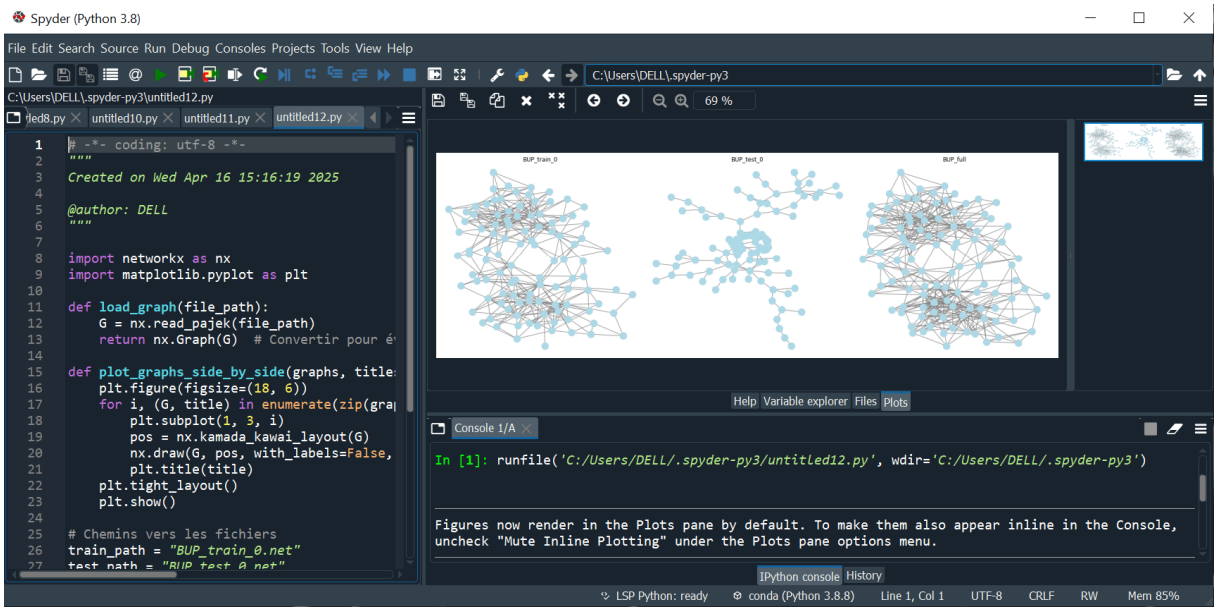

FIG. 3.1 : Interface Spyder .

3.4 Bibliothèques utilisées

Dans cet environnement, plusieurs bibliothèques ont été installées pour faciliter le développement :

- **os** :

  permet d'interagir avec le système d'exploitation pour la gestion des fichiers, des processus et des variables systèmes.

- **NetworkX** :

  bibliothèque pour la création, la manipulation et l'analyse de réseaux complexes.

- **numpy** :

  Bibliothèque pour le calcul scientifique, offrant des structures de données efficaces et des fonctions mathématiques avancées.

3.5 Datasets

Le fichier `datasets.zip` rassemble 22 réseaux provenant de différentes sources et domaines d'application. Nous en avons choisi quatre (BUP , CEG , INF , UAL).

Ces réseaux ont été soigneusement sélectionnés pour couvrir un large éventail de propriétés telles que la taille, le degré moyen,le diamètre et la longueur moyenne des plus courts chemins .

Un résumé des propriétés structurelles des réseaux que nous avons utilisées dans nos expériences se trouve dans le tableau ci-dessous

TAB. 3.1 : Résumé des propriétés structurelles des réseaux

Nom	Nœuds	Liens	⟨k⟩	ASPL	D
CEG	297	2148	14.46	2.46	5
INF	410	2765	13.49	3.63	9
UAL	332	2126	12.81	2.74	6
BUP	105	441	8.4	3.08	7

Obtenus à partir du site

`https://noesis.ikor.org/datasets/link-prediction/datasets.zip`.

Nous travaillons sur 4 fichiers des réseaux suivants :

- **CEG** (Cancer Expression Graph) est un réseau de biologie : Il s'agit d'un réseau issu de données de biologie, généralement utilisé pour représenter les interactions entre gènes ou protéines impliqués dans le développement ou la progression du cancer.

  Les nœuds représentent des gènes ou des protéines, tandis que les arêtes symbolisent des interactions biologiques telles que des co-expressions, des régulations ou des corrélations d'expression.

Ce type de réseau permet d'identifier des modules fonctionnels ou des gènes clés associés à certaines pathologies.

- **INF** (Infectious Exhibition) est un réseau de contacts en face à face dans une exposition : Ce réseau modélise les contacts physiques en face à face entre individus lors d'une exposition, souvent dans le cadre d'une étude sur la propagation de maladies infectieuses.

  Les nœuds représentent des visiteurs ou des participants, et les arêtes indiquent un contact direct ayant eu lieu dans un certain intervalle de temps.

  Ces données sont généralement récoltées à l'aide de capteurs RFID, et servent à analyser les dynamiques de transmission dans des environnements publics.

- **UAL** (U.S. Airport Network) est un réseau de trafic aéroportuaire : Ce réseau représente le trafic aérien entre les aéroports aux États-Unis.

  Les nœuds correspondent aux aéroports, et les arêtes représentent des liaisons aériennes entre eux, pondérées par le nombre de vols ou de passagers.

  Ce type de réseau est crucial pour l'étude de la résilience des infrastructures de transport, la diffusion spatiale de maladies, ou encore l'optimisation des routes aériennes.

- **BUP** (Political Blog Network) est un réseau de blogs politiques : Ce réseau est constitué de blogs politiques américains recensés autour de l'élection présidentielle de 2004.

  Les nœuds représentent des blogs, et les arêtes représentent des hyperliens entre eux.

  Il est souvent utilisé pour étudier la polarisation politique, la structure de l'opinion en ligne et la dynamique de l'information dans des environnements numériques.

  Ce réseau est aussi un exemple classique d'analyse de communautés et de détection de clusters idéologiques.

3.6 Processus de prédiction de liens

Cette section décrit le processus général suivi pour effectuer la prédiction de liens sur les réseaux étudiés :

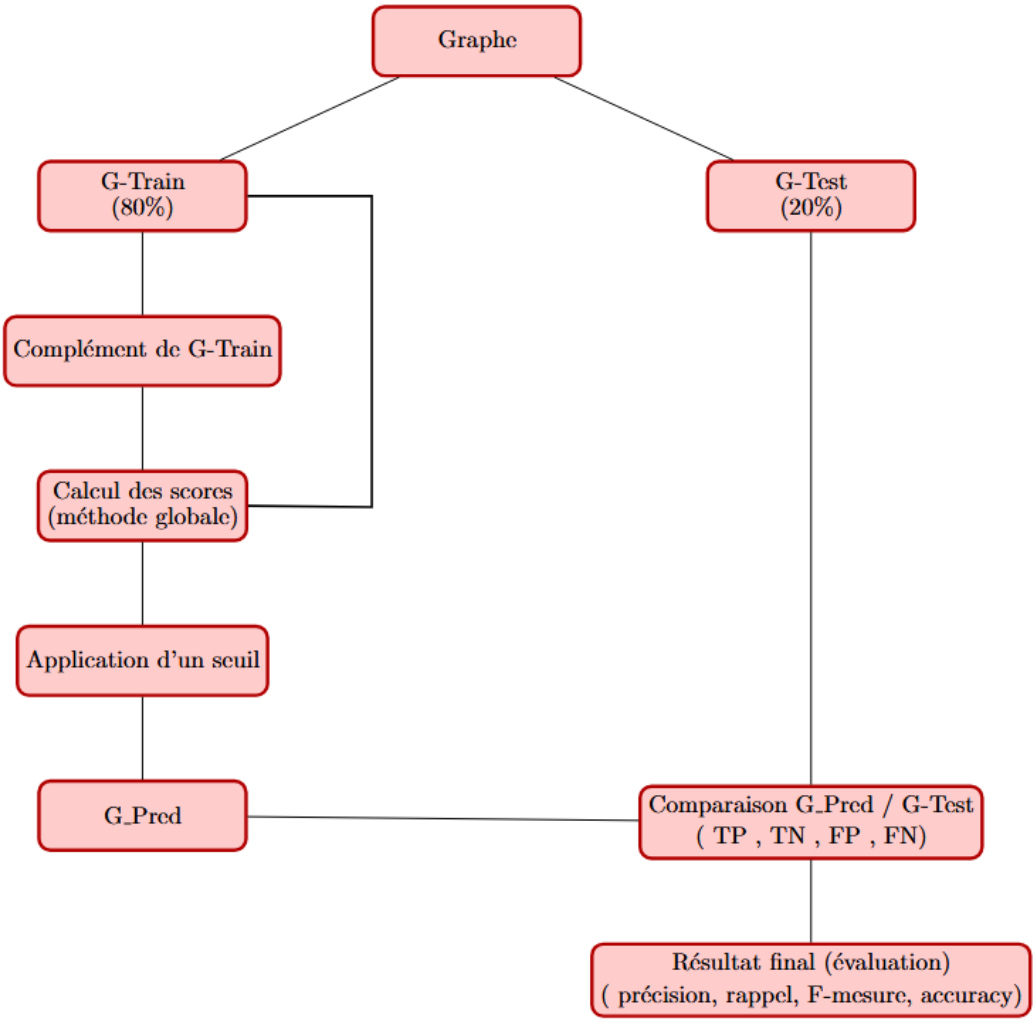

FIG. 3.2 : Processus de prédiction des liens

les étapes du processus de prédiction de liens :

1. **La collecte de données**

   Le point de départ est un graphe initial (nommé ici *Graphe*), représentant un réseau de nœuds et de liens donnés.

2. **Séparation en G-Train (80 %) et G-Test (20 %)**

   Le graphe est séparé en deux sous-ensembles :

   - **G-Train (80 %)** :
     utilisé pour l'apprentissage ou la prédiction .
   - **G-Test (20 %)** :
     utilisé pour évaluer les performances du modèle, les liens qu'il contient étant cachés pendant la phase de prédiction.

3. **Complément de G-Train**

   On identifie les paires de nœuds qui ne sont pas connectées dans *G-Train*, c'est-à-dire les liens absents dans ce sous-graphe. Ces paires constituent les candidats à la prédiction.

4. **Calcul des scores (procédure globale)**

   Pour chaque paire non connectée, un score représentant la probabilité d'existence d'un lien est calculé à l'aide d'une méthode globale (comme *Katz*, *SimRank*, etc.).

   Ces scores quantifient la probabilité qu'un lien existe entre deux nœuds.

5. **Application d'un seuil**

   Un seuil est appliqué aux scores obtenus : seules les paires dont le score dépasse ce seuil sont conservées comme liens prédits.

   On construit alors un graphe prédictif, noté *G_Pred*.

6. **Comparaison entre G_Pred et G-Test**

   Les liens prédits dans *G_Pred* sont comparés aux liens réels présents dans *G-Test*, afin d'évaluer la qualité des prédictions.

7. **Résultat final (évaluation)**

   La performance du modèle est mesurée à l'aide de métriques d'évaluation telles que la précision, le rappel, la F-mesure (F1-score), ou l'accuracy.

   Il s'agit du résultat final du processus de prédiction.

3.7 Expérimentations

Nous avons travaillé sur le thème de la prédiction de liens dans les réseaux complexes.

Nous présentons ici les expérimentations réalisées pour évaluer la mesure de performance des différentes méthodes de similarité globale dans la prédiction de liens .

basée sur la comparaison entre plusieurs mesures de similarité globale, telles que le Plus Court Chemin, l'indice de Katz, Newton's Gravitational Law Index, Common Neighbor and Distance index (CND) , SimRank sur les fichiers de DataSets (BUP, INF, UAL, CEG) en utilisant des approches basées sur la similarité globale visent à explorer comment utiliser la comparaison pour améliorer les prédictions de liens.

Nous avons utilisé un algorithme sur l'environnement (python 3.8) pour calculer les indicateurs de performance. Les résultats obtenus sont mentionnés dans les tableaux ci-dessous.

3.7.1 présentation des résultats de la base (fichier) BUP

1. précision de (Shortest Path , SimRank , CND , NGLI , Katz Index)

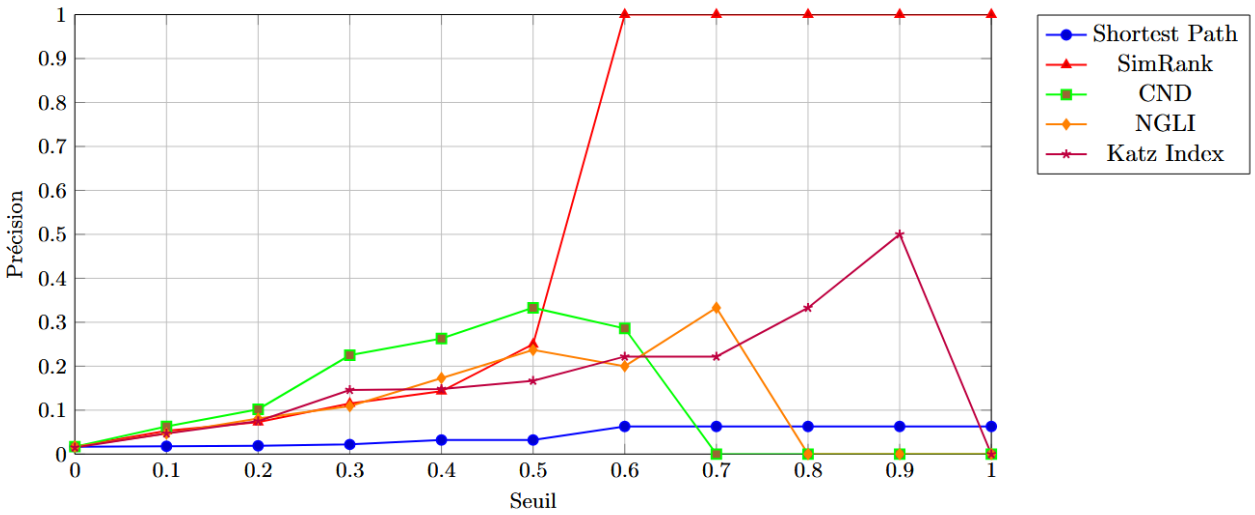

FIG. 3.3 : Comparaison du précision pour différentes méthodes de prédiction de liens de la base BUP

- discussion de résultat :

  Nous constatons que le choix de la méthode et du seuil influence fortement la précision des prédictions.

  La précision évalue la proportion de prédictions correctes parmi toutes les prédictions positives, indiquant ainsi la fiabilité d'une méthode à éviter les faux positifs.

  Nos résultats montrent que le choix du seuil influence fortement cette mesure.

  Parmi les méthodes testées, SimRank se distingue par sa précision nettement supérieure à partir du seuil 0.5.

  Elle atteint même une précision parfaite (1.0) entre 0.6 et 1.0, ce qui signifie que tous les liens prédits dans cette plage sont corrects.

  Cela reflète une capacité remarquable à discriminer les vraies connexions.

2. rappel de (Shortest Path , SimRank , CND , NGLI , Katz Index)

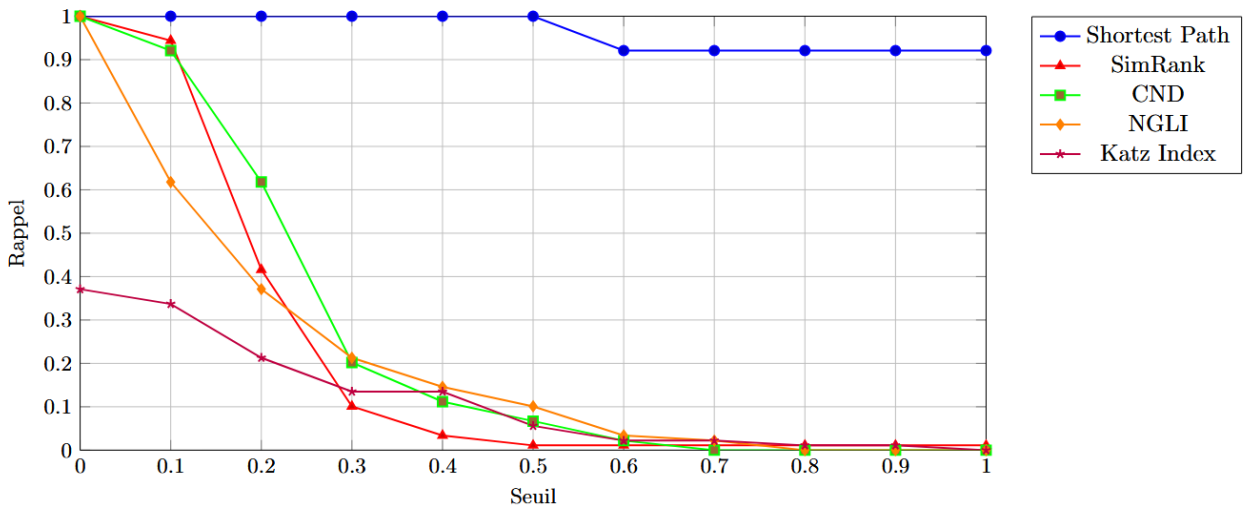

FIG. 3.4 : Comparaison du rappel pour différentes méthodes de prédiction de liens de la base BUP

- discussion de résultat :

  Le rappel mesure la capacité d'une méthode à retrouver tous les liens réellement existants dans le réseau, c'est-à-dire la proportion de vrais liens correctement prédits parmi l'ensemble des liens existants.

  En général, nos résultats montrent que le rappel diminue à mesure que le seuil augmente, car moins de paires sont considérées comme des liens potentiels.

  Toutefois, la méthode Shortest Path fait exception : elle maintient un rappel parfait ou quasi-parfait sur toute la plage de seuils.

  Cela signifie qu'elle parvient à identifier presque tous les liens réels, même à des seuils élevés.

  Cette performance souligne sa capacité à ne pas manquer de vraies relations, ce qui est crucial dans des contextes où l'exhaustivité est prioritaire.

3. f-mesure de (Shortest Path , SimRank , CND , NGLI , Katz Index)

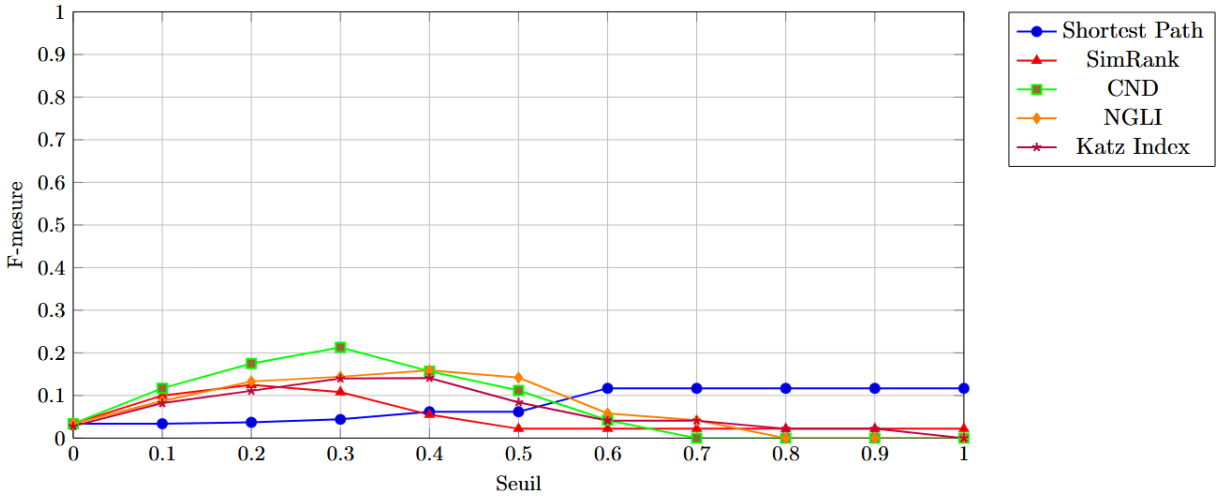

FIG. 3.5 : Comparaison du f-mesure pour différentes méthodes de prédiction de liens de la base BUP

- discussion de résultat :

  Les résultats de la figure montrent que la F-mesure varie fortement selon le seuil et la méthode utilisée.

  La F-mesure (ou F1-score) est l'harmonique entre la précision et le rappel, et permet d'évaluer une méthode selon un compromis équilibré entre ces deux critères.

  Les résultats montrent que la F-mesure varie fortement en fonction du seuil choisi et de la méthode utilisée.

  La méthode Common Neighbors Distance (CND) obtient le meilleur score, atteignant un pic de 0.213 pour un seuil de 0.3. Cela indique qu'à ce seuil, CND parvient à maintenir une précision et un rappel relativement bons simultanément.

  Ce bon équilibre est essentiel dans les situations où il est aussi important d'éviter les faux positifs que de retrouver un maximum de vrais liens.

  Ainsi, CND apparaît comme la méthode la plus robuste et équilibrée dans notre cadre expérimental.

4. accuracy de (Shortest Path , SimRank , CND , NGLI , Katz Index)

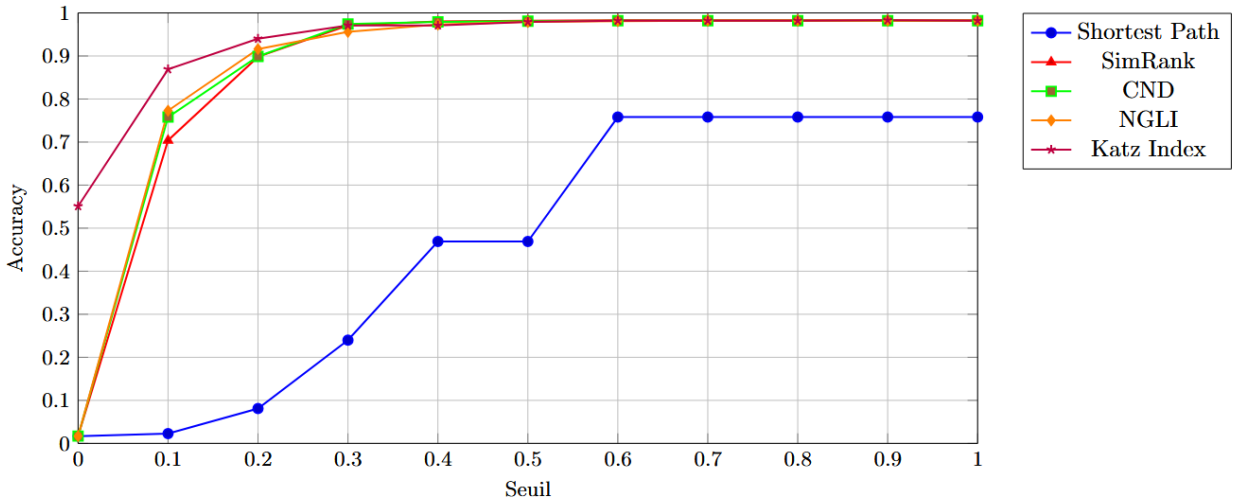

FIG. 3.6 : Comparaison du accuracy pour différentes méthodes de prédiction de liens de la base BUP

- discussion de résultat :

  SimRank (SR) et Katz Index (KI) atteignent tous deux une accuracy maximale de 0.983, ce qui les place en tête.

  Toutefois, SimRank atteint ce niveau plus tôt (dès le seuil 0.6), alors que Katz Index n'y parvient qu'à partir de 0.9.

  Cela fait de SimRank la méthode la plus efficace et stable en termes d'exactitude sur une large plage de seuils.

5. performances obtenues pour la base BUP sur les seuils de 0.0 à 1.0 (Moyenne)

Méthodes et mesures	Précision	Rappel	F-mesure	Accuracy
Shortest Path Inverse	0.041	0.964	0.078	0.463
SimRank	0.514	0.233	0.050	0.861
CN and Distance	0.117	0.267	0.077	0.865
NGLI	0.109	0.228	0.073	0.866
Katz Index	0.134	0.149	0.085	0.931

TAB. 3.2 : performances obtenues pour la base BUP sur les seuils de 0.0 à 1.0 (Moyenne)

3.7.2 présentation des résultats de la base (fichier) CEG

1. précision de (Shortest Path , SimRank , CND , NGLI , Katz Index)

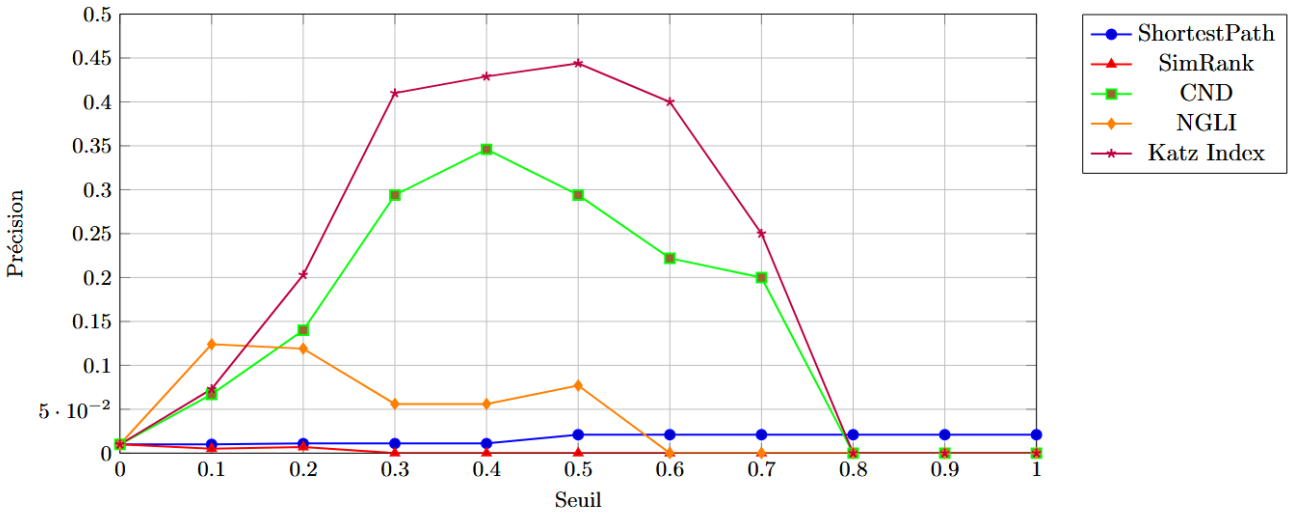

FIG. 3.7 : Comparaison du précision pour différentes méthodes de prédiction de liens de la base CEG

- discussion de résultat :

  La précision mesure la proportion de liens correctement prédits parmi l'ensemble des liens prédits comme existants, reflétant ainsi la capacité d'une méthode à limiter les faux positifs.

  D'après le graphe, la méthode Katz Index (KI) se distingue nettement en termes de précision.

  Elle atteint un maximum d'environ 0.444 à un seuil de 0.5, ce qui signifie que près de 0.45 des liens qu'elle prédit sont effectivement corrects.

  Cette performance dépasse celles des autres méthodes sur une large plage de seuils, ce qui indique une grande fiabilité dans ses prédictions.

  Katz Index favorise donc des prédictions de haute qualité, en privilégiant les liens les plus probables, même si cela peut se faire au détriment du rappel.

2. rappel de (Shortest Path , SimRank , CND , NGLI , Katz Index)

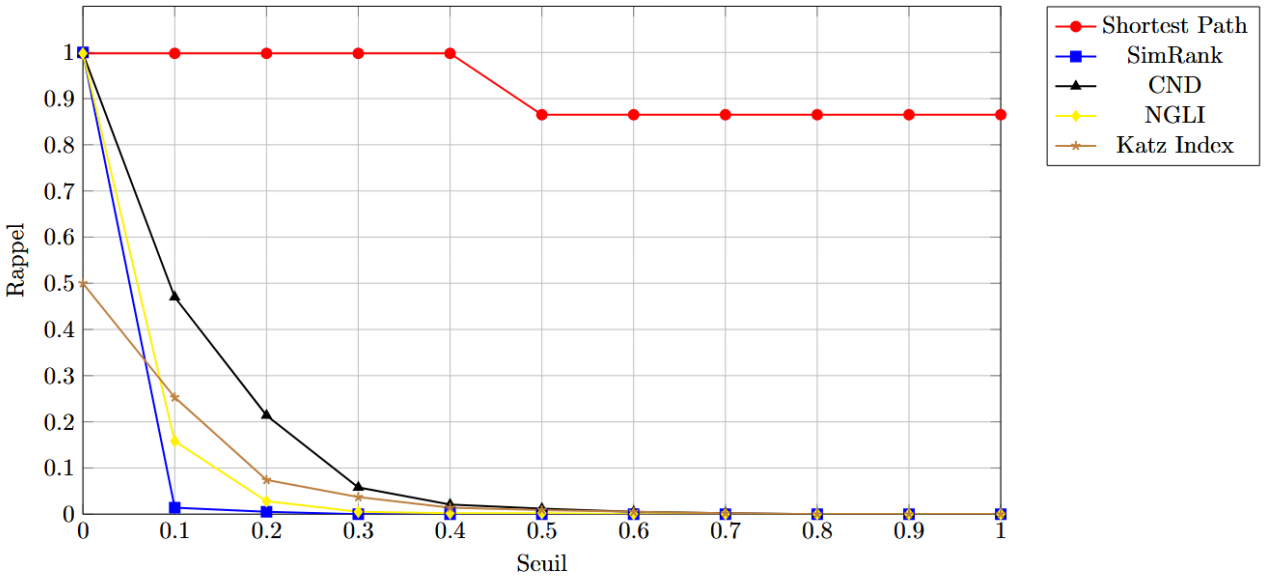

FIG. 3.8 : Comparaison du rappel pour différentes méthodes de prédiction de liens de la base CEG

- discussion de résultat :

  Le rappel mesure la capacité d'une méthode à retrouver tous les liens réels du réseau, c'est-à-dire le taux de vrais positifs par rapport à l'ensemble des liens existants.

  La méthode Shortest Path est celle qui obtient les meilleurs résultats en rappel.

  Elle atteint un score proche de 1 (0.998) sur une large plage de seuils allant de 0.0 à 0.4, ce qui signifie qu'elle identifie presque tous les liens réels dans cette zone.

  Même au-delà, son rappel reste élevé (0.865), montrant une robustesse remarquable.

  Elle surpasse ainsi toutes les autres méthodes de façon constante, ce qui en fait un excellent choix lorsque l'exhaustivité des liens détectés est prioritaire, par exemple dans des applications où rater un lien peut avoir un coût important.

3. f-mesure de (Shortest Path , SimRank , CND , NGLI , Katz Index)

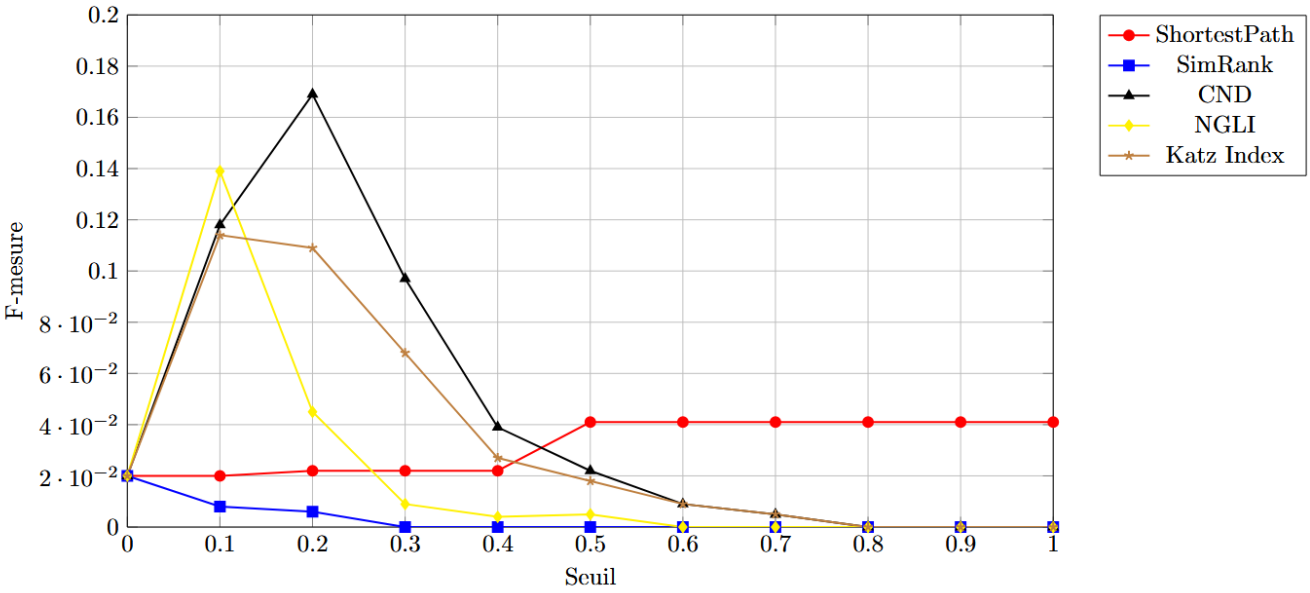

FIG. 3.9 : Comparaison du f-mesure pour différentes méthodes de prédiction de liens de la base CEG

- discussion de résultat :

  La F-mesure (ou F1-score) est une mesure synthétique qui combine la précision et le rappel en une seule valeur, en prenant leur moyenne harmonique.

  Elle permet d'évaluer le compromis entre ces deux indicateurs, ce qui est essentiel lorsque l'on souhaite à la fois éviter les faux positifs et ne pas manquer de vrais liens.

  Dans notre expérimentation, la méthode Common Neighbors Distance (CND) se distingue comme la meilleure selon la F-mesure.

  Elle atteint un pic à environ 0.169 au seuil 0.2, surpassant toutes les autres méthodes. Sa performance reste stable et relativement élevée entre les seuils 0.1 et 0.3, ce qui montre sa capacité à maintenir un bon équilibre.

  Cela fait de CND la méthode la plus équilibrée sur cette base de données, idéale lorsque les objectifs de précision et de rappel sont également importants.

4. accuracy de (Shortest Path , SimRank , CND , NGLI , Katz Index)

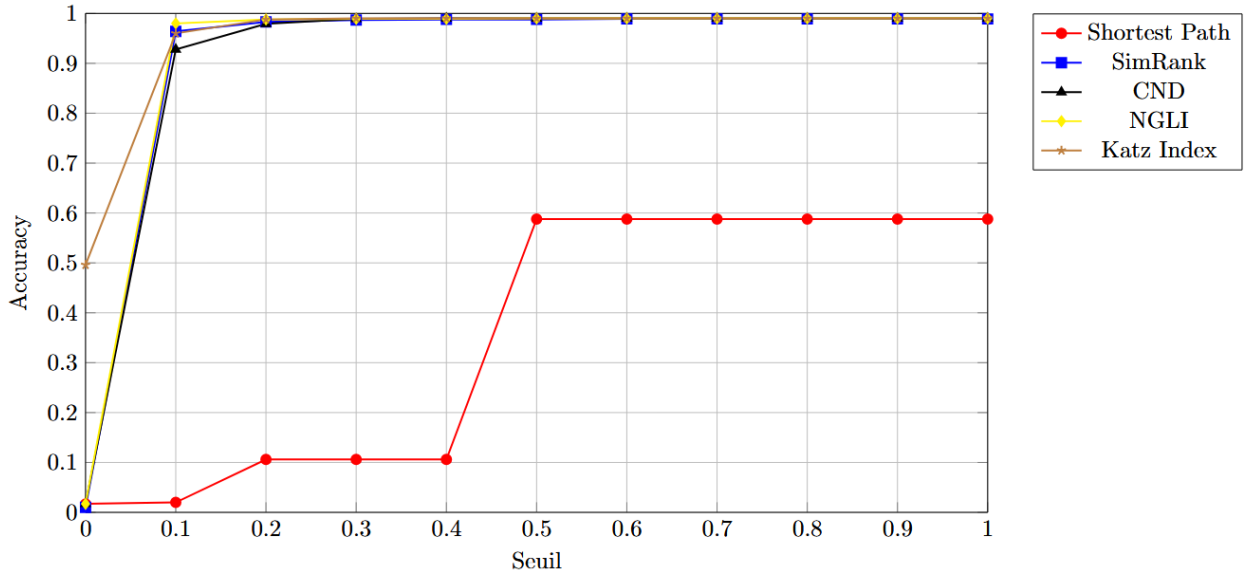

FIG. 3.10 : Comparaison du accuracy pour différentes méthodes de prédiction de liens de la base CEG

- discussion de résultat :

  Meilleure méthode en termes de seuil d'apparition de l'accuracy maximale : Katz Index, qui atteint 0.990 dès le seuil 0.3.

  Si le seuil optimal n'est pas encore fixé, Katz Index est légèrement supérieur car il atteint l'accuracy maximale plus tôt.

  Donc, selon l'accuracy, la meilleure méthode est le Katz Index, suivi de très près par NGLI, CND, et SimRank.

5. performances obtenues pour la base GEG sur les seuils de 0.0 à 1.0 (Moyenne)

Méthodes et mesures	Précision	Rappel	F-mesure	Accuracy
Shortest Path Inverse	0.016	0.922	0.035	0.399
SimRank	0.002	0.092	0.003	0.975
CN and Distance	0.096	0.201	0.045	0.955
NGLI	0.040	0.109	0.022	0.968
Katz Index	0.179	0.093	0.032	0.971

TAB. 3.3 : performances obtenues pour la base GEG sur les seuils de 0.0 à 1.0 (Moyenne)

3.7.3 présentation des résultats de la base (fichier) UAL

1. précision de (Shortest Path , SimRank , CND , NGLI , Katz Index)

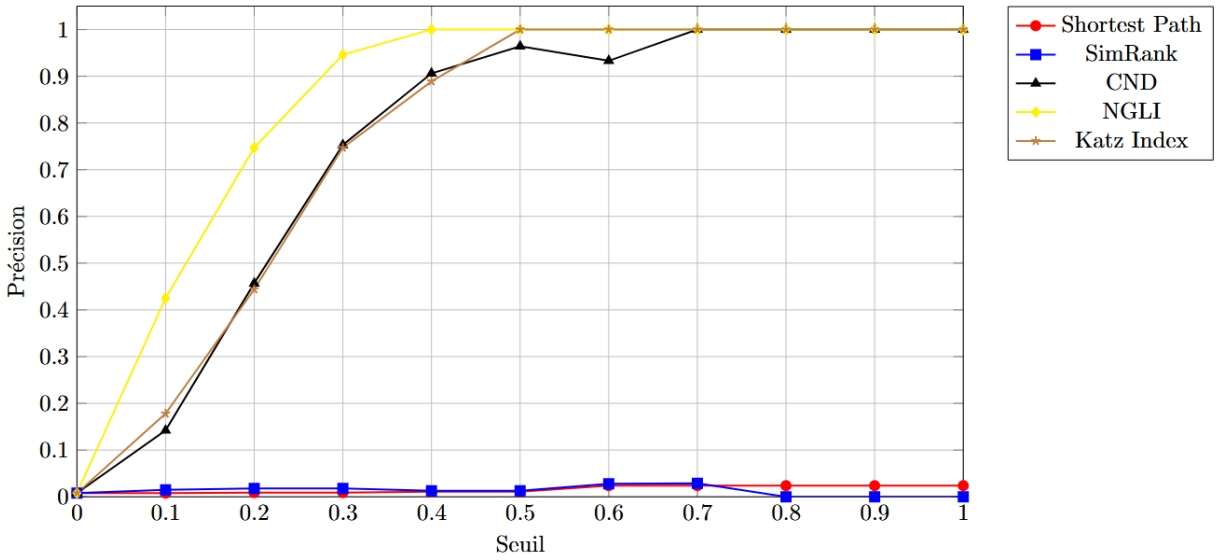

FIG. 3.11 : Comparaison du précision pour différentes méthodes de prédiction de liens de la base UAL

- discussion de résultat :

  La précision reflète la capacité d'un algorithme à faire des prédictions exactes, en mesurant la proportion de liens correctement prédits parmi tous les liens prédits comme existants.

  Dans cette expérimentation, la méthode NGLI se distingue comme la plus performante sur cet indicateur.

  Elle atteint une précision maximale (1.000) dès le seuil de 0.4, soit plus tôt que toutes les autres méthodes.

  Ce score parfait signifie que toutes ses prédictions à partir de ce seuil sont correctes, sans aucun faux positif.

  De plus, cette précision reste stable à 1.000 pour tous les seuils suivants, témoignant d'une fiabilité exceptionnelle.

  NGLI dépasse ainsi les autres méthodes, y compris CND et Katz Index, non seulement en précision maximale mais aussi en rapidité à l'atteindre, ce qui en fait une méthode idéale lorsque l'on recherche des prédictions sûres dès les premiers seuils significatifs.

2. rappel de (Shortest Path , SimRank , CND , NGLI , Katz Index)

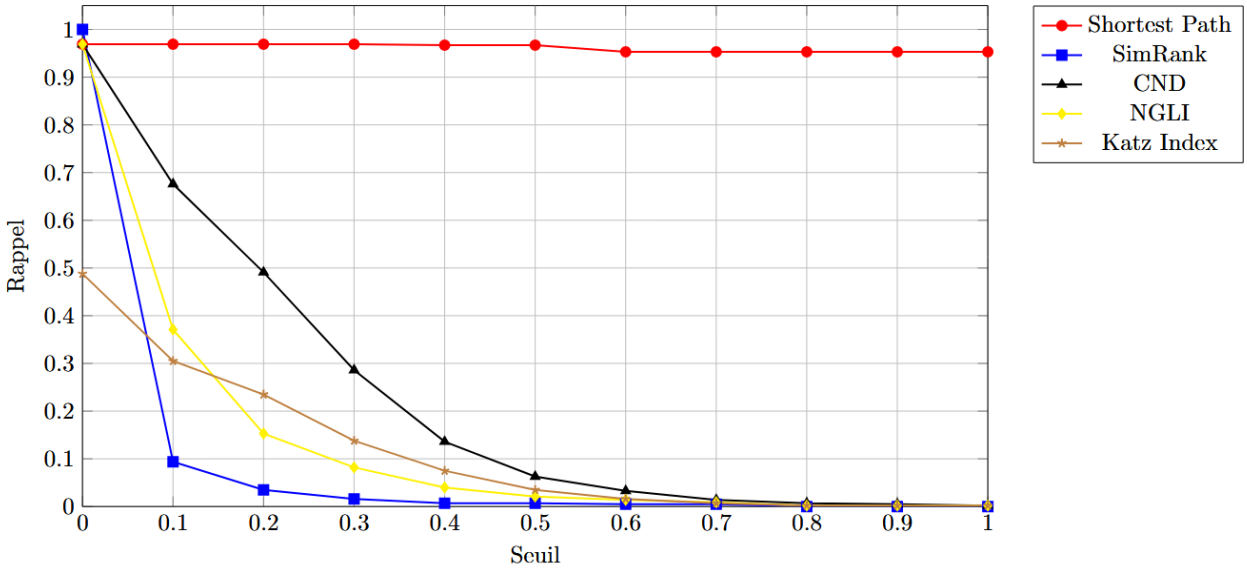

FIG. 3.12 : Comparaison du rappel pour différentes méthodes de prédiction de liens de la base UAL

- discussion de résultat :

  Le rappel évalue la capacité d'une méthode à retrouver l'ensemble des liens réels dans le réseau, c'est-à-dire le rapport entre les liens correctement identifiés et le nombre total de liens existants.

  Une valeur élevée de rappel indique que peu de liens sont manqués.

  La méthode Shortest Path se distingue nettement par un rappel très élevé et remarquablement stable, quelle que soit la valeur du seuil.

  Dès que le seuil dépasse 0.1, elle surpasse toutes les autres méthodes, confirmant sa robustesse.

  Cette stabilité signifie qu'elle conserve une excellente capacité de détection des liens vrais, même à mesure que les critères de sélection deviennent plus stricts.

  Shortest Path est donc idéale dans des contextes où l'exhaustivité prime, comme dans la recherche d'informations ou la détection de relations cachées.

3. f-mesure de (Shortest Path , SimRank , CND , NGLI , Katz Index)

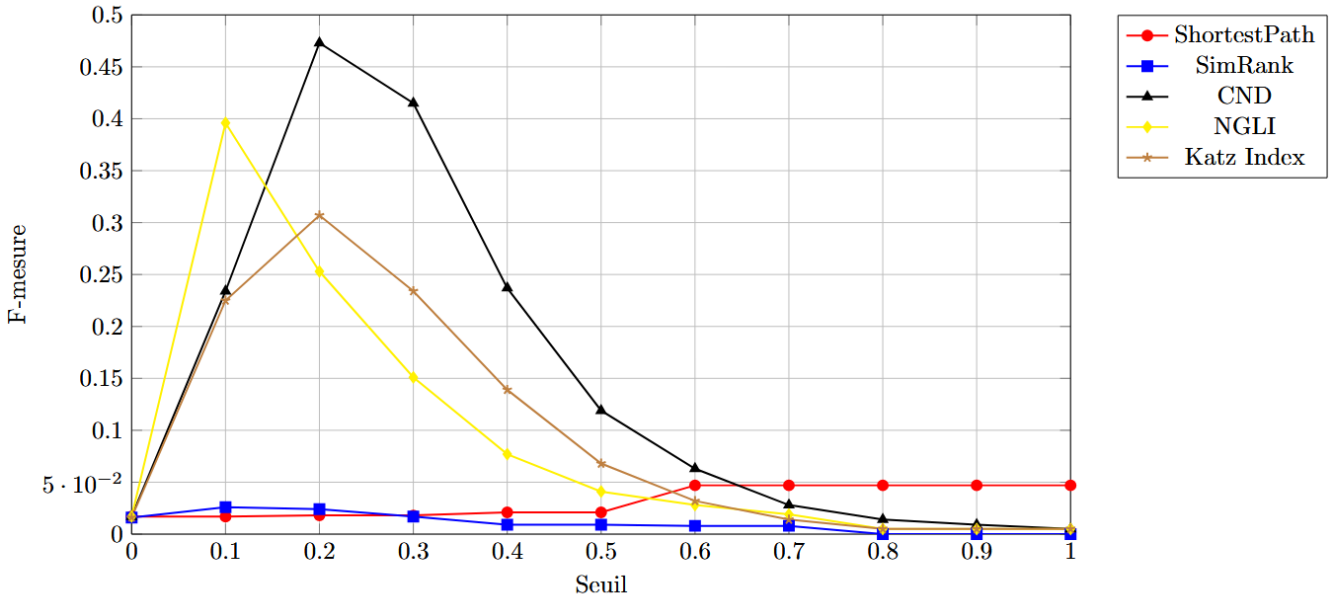

FIG. 3.13 : Comparaison du f-mesure pour différentes méthodes de prédiction de liens de la base UAL

- discussion de résultat :

  La F-mesure combine la précision et le rappel en une seule métrique, représentant leur moyenne harmonique, ce qui permet d'évaluer le compromis global entre la qualité et la complétude des prédictions.

  La méthode Common Neighbors Distance (CND) se distingue comme la meilleure selon la F-mesure, atteignant un score maximal de 0.473.

  Cette valeur élevée reflète une performance équilibrée, où la précision reste forte, assurant que les prédictions sont majoritairement correctes, tandis que le rappel reste acceptable, notamment jusqu'au seuil 0.2, garantissant que beaucoup de liens vrais sont détectés.

  De plus, CND maintient cette performance sur une plage de seuils assez large (de 0.1 à 0.4), démontrant sa robustesse face au choix du seuil.

  Ainsi, CND offre un excellent compromis entre éviter les faux positifs et ne pas manquer trop de liens réels.

4. accuracy de (Shortest Path , SimRank , CND , NGLI , Katz Index)

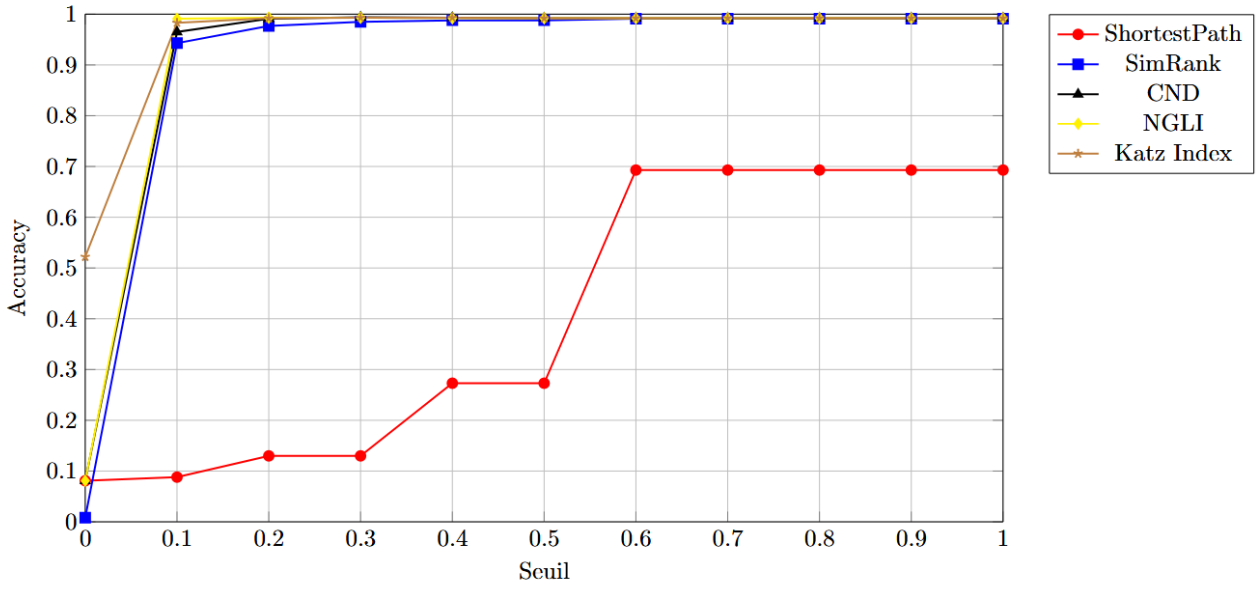

FIG. 3.14 : Comparaison du accuracy pour différentes méthodes de prédiction de liens de la base UAL

- discussion de résultat :

  Common Neighbors Distance est la meilleure méthode , car : Elle atteint l'exactitude maximale de 0.994 (au seuil 0.3).

  Elle maintient une performance très stable (0.992) sur une large plage de seuils.

5. performances obtenues pour la base UAL sur les seuils de 0.0 à 1.0 (Moyenne)

Méthodes et mesures	Précision	Rappel	F-mesure	Accuracy
Shortest Path Inverse	0.016	0.956	0.033	0.435
SimRank	0.012	0.105	0.012	0.967
CN and Distance	0.683	0.263	0.108	0.968
NGLI	0.671	0.143	0.089	0.968
Katz Index	0.595	0.140	0.096	0.974

TAB. 3.4 : performances obtenues pour la base UAL sur les seuils de 0.0 à 1.0 (Moyenne)

3.7.4 présentation des résultats de la base (fichier) INF

1. précision de (Shortest Path , SimRank , CND , NGLI , Katz Index)

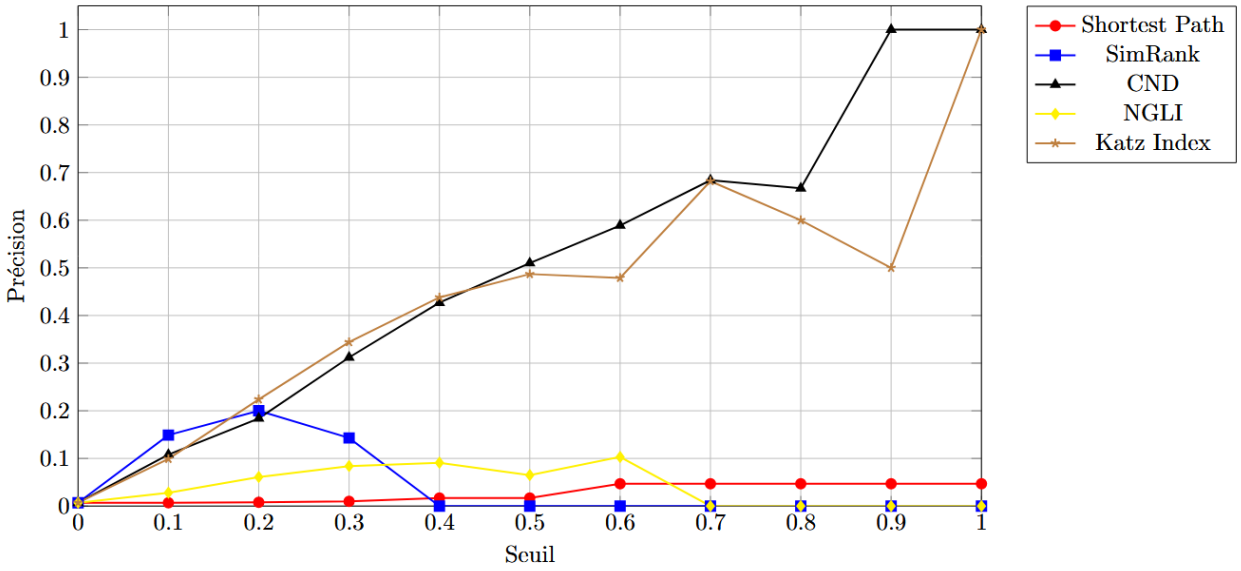

FIG. 3.15 : Comparaison du précision pour différentes méthodes de prédiction de liens de la base INF

- discussion de résultat :

  La précision mesure la proportion de liens correctement prédits parmi toutes les prédictions positives, indiquant la capacité d'une méthode à éviter les faux positifs.

  Parmi les méthodes étudiées, Common Neighbors Distance (CND) se distingue par une précision élevée et une grande stabilité sur la majorité des seuils, garantissant des prédictions fiables sur une large plage.

  La méthode Katz Index montre également de bonnes performances en précision, mais sa stabilité est un peu moindre, notamment entre les seuils 0.6 et 0.9, où la précision fluctue davantage.

  Ainsi, bien que Katz Index soit très performant, CND offre une meilleure constance dans la qualité des prédictions, ce qui est un atout important pour des applications nécessitant une fiabilité constante.

2. rappel de (Shortest Path , SimRank , CND , NGLI , Katz Index)

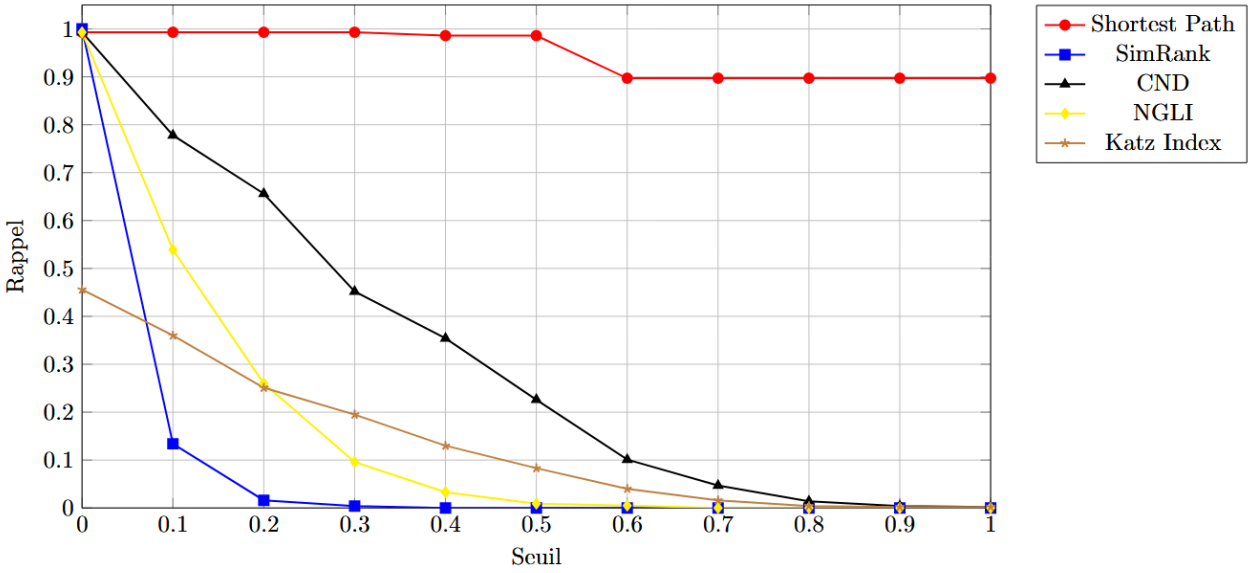

FIG. 3.16 : Comparaison du rappel pour différentes méthodes de prédiction de liens de la base INF

- discussion de résultat :

  Le rappel mesure la capacité d'une méthode à identifier la majorité des liens réellement existants dans le réseau, c'est-à-dire la proportion de vrais liens correctement détectés parmi tous les liens présents.

  La méthode Shortest Path se distingue par un rappel élevé, supérieur à 0.9, et ce, jusqu'au seuil 0.6, ce qui signifie qu'elle retrouve plus de 0.9 des liens réels même en augmentant le seuil.

  Cette performance fait de Shortest Path un choix privilégié lorsque l'exhaustivité est essentielle, notamment dans des contextes où il est crucial de ne pas omettre des relations importantes.

  Sa capacité à maintenir un rappel élevé sur une large plage de seuils démontre une grande robustesse dans la détection des liens vrais.

3. f-mesure de (Shortest Path , SimRank , CND , NGLI , Katz Index)

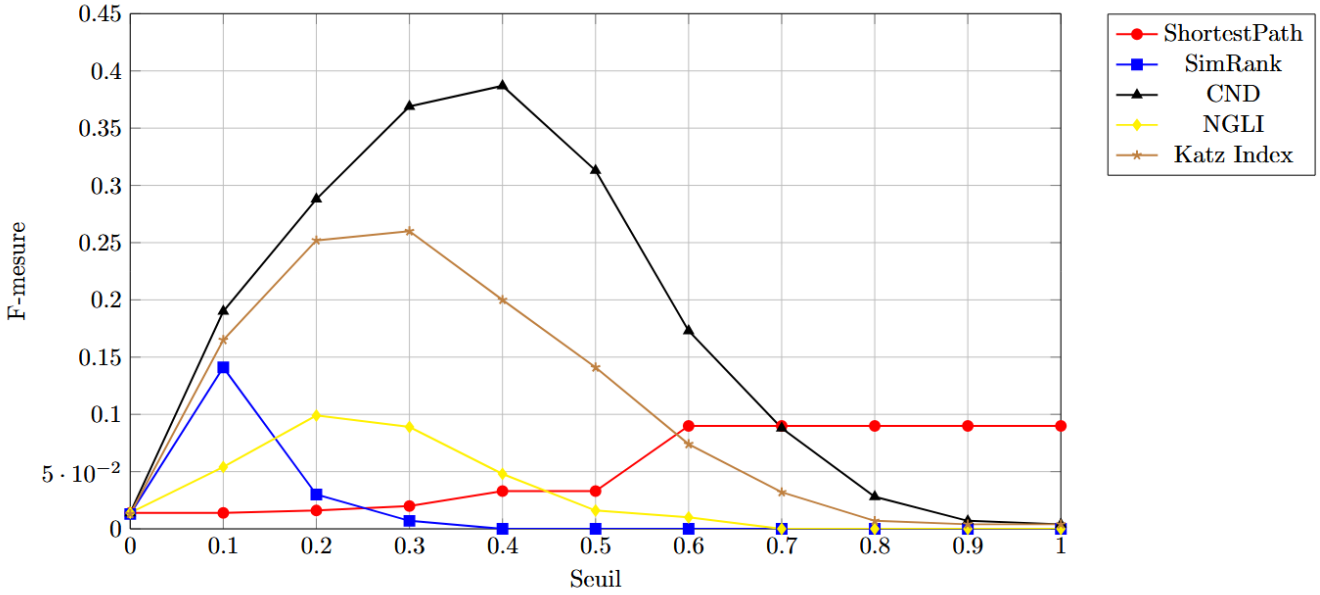

FIG. 3.17 : Comparaison du f-mesure pour différentes méthodes de prédiction de liens de la base INF

- discussion de résultat :

  La F-mesure combine la précision et le rappel en une métrique unique, représentant leur moyenne harmonique, ce qui permet d'évaluer l'équilibre entre la qualité et la quantité des prédictions.

  La méthode Common Neighbors Distance (CND) se distingue comme la meilleure selon cette mesure, offrant la meilleure combinaison entre précision et rappel.

  Autour du seuil 0.4, elle atteint un score maximal de 0.387, indiquant un compromis optimal où les prédictions sont à la fois fiables (précision élevée) et complètes (rappel satisfaisant).

  Cette performance souligne la capacité de CND à équilibrer efficacement le risque de faux positifs et celui de faux négatifs, ce qui est essentiel pour une prédiction de liens robuste.

4. accuracy de (Shortest Path , SimRank , CND , NGLI , Katz Index)

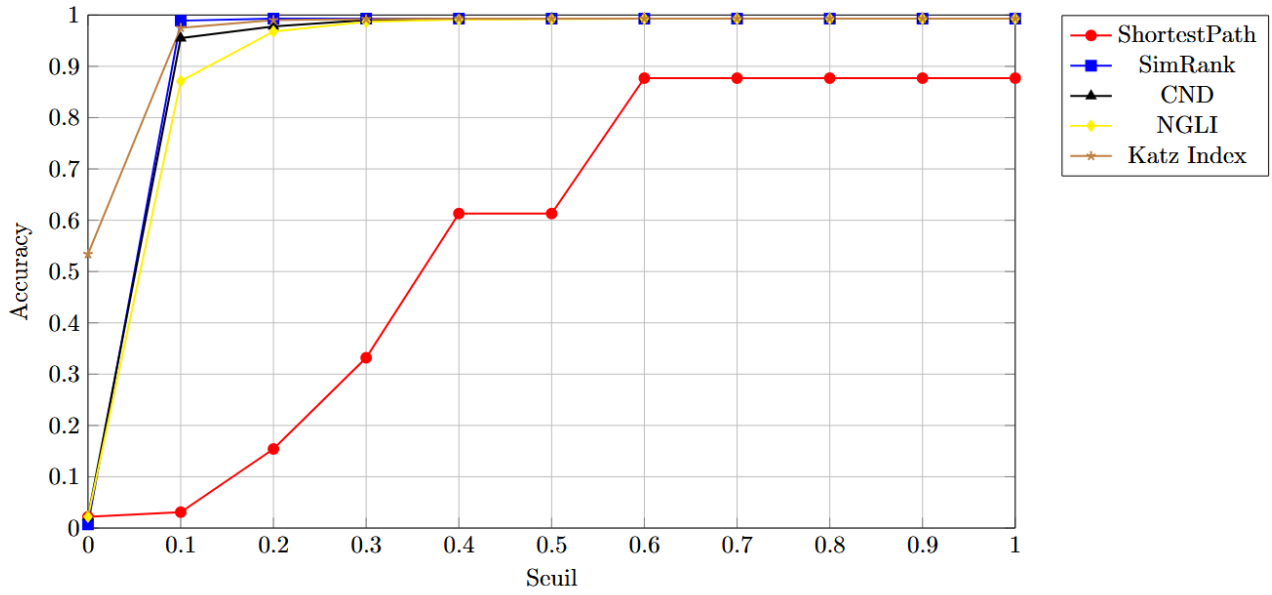

FIG. 3.18 : Comparaison du accuracy pour différentes méthodes de prédiction de liens de la base INF

- discussion de résultat :

  Meilleure méthode selon accuracy est SimRank .

  Atteint accuracy = 0.993 à partir du seuil 0.2, et reste stable jusqu'à 1.0.

5. performances obtenues pour la base INF sur les seuils de 0.0 à 1.0 (Moyenne)

Méthodes et mesures	Précision	Rappel	F-mesure	Accuracy
Shortest Path Inverse	0.027	0.942	0.057	0.598
SimRank	0.041	0.106	0.016	0.991
CN and Distance	0.365	0.391	0.161	0.972
NGLI	0.046	0.183	0.031	0.945
Katz Index	0.379	0.154	0.102	0.970

TAB. 3.5 : performances obtenues pour la base INF sur les seuils de 0.0 à 1.0 (Moyenne)

3.8 Conclusion

Dans ce chapitre, nous avons présenté notre projet de recherche appliqué à la prédiction de liens dans les réseaux complexes, en mettant l'accent sur l'évaluation comparative de plusieurs mesures de similarité globale.

Les expérimentations ont porté sur des indices bien établis tels que le Plus Court Chemin, l'indice de Katz, le Newton's Gravitational Law Index, le Common Neighbor and Distance Index (CND), ainsi que le SimRank.

Les tests ont été menés sur un réseau réel de collaboration scientifique.

Ce cadre expérimental a permis d'évaluer les performances des différentes méthodes en fonction de la structure du réseau et de la nature des interactions observées.

Les résultats obtenus ont mis en évidence les avantages et les limites spécifiques à chaque mesure de similarité, et ont permis d'identifier les approches les plus adaptées à ce type de réseau.

Cette analyse constitue une base solide pour la poursuite de nos travaux, notamment en vue d'améliorer les performances de prédiction en combinant ou en adaptant ces mesures aux caractéristiques particulières du graphe étudié.

Conclusion générale

Ce travail a permis d'explorer en profondeur la problématique de la prédiction de liens dans les réseaux complexes, à travers une analyse comparative de cinq méthodes de similarité globale.

Les expérimentations menées ont mis en évidence la pertinence spécifique de chaque approche, révélant que le choix de la méthode dépend étroitement de la topologie du réseau étudié ainsi que des objectifs visés par l'analyse.

Ainsi, les résultats ont montré que certaines méthodes, comme le Katz Index ou le SimRank, présentent une forte capacité de généralisation, les rendant particulièrement adaptées à des contextes variés.

D'autres, telles que le Shortest Path ou le Newton's Gravitational Law Index (NGLI), se révèlent plus efficaces dans des réseaux à structure bien définie, où les liens sont fortement influencés par la distance ou la hiérarchie.

Par ailleurs, la méthode Common Neighbor Distance (CND) se distingue par un équilibre intéressant entre performance et coût computationnel, ce qui en fait une approche pragmatique dans un cadre applicatif.

Ce mémoire ouvre plusieurs perspectives de recherche. Parmi les prolongements envisageables, on peut citer :

- l'intégration d'algorithmes d'apprentissage automatique pour améliorer la précision de la prédiction ;
- l'exploitation d'attributs de nœuds afin d'enrichir les mesures de similarité ;
- l'extension aux graphes dynamiques, dont l'évolution temporelle introduit de nouveaux défis.

En somme, cette étude pose les fondations d'une exploration plus large des méthodes hybrides et adaptatives pour la prédiction de liens, dans des réseaux de plus en plus complexes et évolutifs.

Bibliographie

[1] J. Scott et P. J. Carrington, *The SAGE Handbook of Social Network Analysis*. SAGE Publications, 2014. doi : 10.4135/9781446294413.

[2] J. Ugander, B. Karrer, L. Backstrom et C. Marlow, *The Anatomy of the Facebook Social Graph*, 2011. arXiv : 1111.4503 [physics.soc-ph]. adresse : https://arxiv.org/abs/1111.4503.

[3] Anonyme, *Exemple illustratif d'un graphe de réseau social*, Image illustrative utilisée à des fins académiques dans ce mémoire, 2025.

[4] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai et A.-L. Barabási, « The large-scale organization of metabolic networks », *Nature*, t. 407, n° 6804, p. 651-654, 2000.

[5] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori et Y. Sakaki, « A comprehensive two-hybrid analysis to explore the yeast protein interactome », *Proceedings of the National Academy of Sciences*, t. 98, n° 8, p. 4569-4574, 2001.

[6] N. Guelzim, S. Bottani, P. Bourgine et F. Képès, « Topological and causal structure of the yeast transcriptional regulatory network », *Nature genetics*, t. 31, n° 1, p. 60-63, 2002.

[7] R. Albert, « Scale-free networks in cell biology », *Journal of cell science*, t. 118, n° 21, p. 4947-4957, 2005.

[8] X. Zhu, M. Gerstein et M. Snyder, « Getting connected : analysis and principles of biological networks », *Genes & development*, t. 21, n° 9, p. 1010-1024, 2007.

[9] Anonyme, *Structure d'un réseau biologique*, Figure 1.2 : Image illustrative d'un réseau biologique utilisée à des fins académiques sans source précisée, 2025.

[10] B. A. Huberman, *The Laws of the Web : Patterns in the Ecology of Information*. MIT Press, 2001, isbn : 978-0-262-27583-5.

[11] L. Egghe et R. Rousseau, « Introduction to Informetrics : Quantitative Methods in Library, Documentation and Information Science », *The Library Quarterly*, t. 61, n° 2, p. 220-221, 1991. doi : 10.1086/602337.

[12] Anonyme, *Exemple de réseaux d'information*, Figure 1.3 : Image illustrative d'un réseau d'information utilisée à des fins académiques sans source précisée, 2025.

[13] L. Amaral, A. Scala, M. Barthélemy et H. Stanley, « Classes of small-world networks », *Proceedings of the National Academy of Sciences*, t. 97, n° 21, p. 11 149-11 152, 2000. doi : 10.1073/pnas.200327197.

[14] V. Kalapala, V. Sanwalani, A. Clauset et C. Moore, « Scale invariance in road networks », *Physical Review E*, t. 73, n° 2, p. 026 130, 2006. doi : 10.1103/PhysRevE.73.026130.

[15] V. Latora et M. Marchiori, « Is the Boston subway a small-world network ? », *Physica A : Statistical Mechanics and its Applications*, t. 314, n° 1, p. 109-113, 2002. doi : 10.1016/S0378-4371(02)01089-0.

[16] Q. Chen, H. Chang, R. Govindan et S. Jamin, « The origin of power laws in Internet topologies revisited », in *Proceedings. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies*, t. 2, IEEE, 2002, p. 608-617. doi : 10.1109/INFCOM.2002.1019306.

[17] Wikipédia, *Réseau de transport aérien*, https://fr.wikipedia.org/wiki/Réseau_de_transport_aérien, Figure 1.4 : Image illustrative issue de Wikipédia, consultée en 2025, 2025.

[18] M. Newman, *Networks : An Introduction*. Oxford : Oxford University Press, 2010.

[19] D. B. West, *Introduction to Graph Theory*, 2nd. Upper Saddle River, NJ : Prentice Hall, 2001.

[20] J. Bondy et U. Murty, *Graph Theory* (Graduate Texts in Mathematics). London : Springer, 2008, t. 244.

[21] J. Scott, *Social Network Analysis : A Handbook*, 2nd. London : SAGE Publications, 2000.

[22] L. C. Freeman, « Centrality in social networks : Conceptual clarification », *Social Networks*, t. 1, n° 3, p. 215-239, 1979.

[23] R. Kanawati, « Prédiction de liens dans les réseaux sociaux », thèse de doct., Université Paris 13, 2010.

[24] Anonyme, *Coefficient de clustering élevé*, Figure 1.8 : Illustration d'un coefficient de clustering élevé utilisée à des fins académiques sans source précisée, 2025.

[25] Anonyme, *Distribution de degrés en loi de puissance*, Figure 1.9 : Illustration d'une distribution de degrés suivant une loi de puissance, utilisée à des fins académiques sans source précisée, 2025.

[26] Anonyme, *Structure en communautés*, Figure 1.10 : Illustration d'une structure en communautés utilisée à des fins académiques sans source précisée, 2025.

[27] M. M. Hasan et S. Zaki, « Survey of link prediction in social networks », *Social network analysis and mining*, t. 3, n° 4, p. 1-25, 2014.

[28] V. Srinivas et P. Mitra, *Link prediction in social networks : Role of power law distribution*. Springer International Publishing, 2016.

[29] R. Guns, « Link prediction », in *Measuring scholarly impact*, Springer, Cham, 2014, p. 35-55.

[30] Anonyme, *La prédiction de lien dans les instants t1, t2*, Figure 2.1 : Illustration du processus de prédiction de lien entre les instants t1 et t2, utilisée à des fins académiques sans source précisée, 2025.

[31] J. B. Schafer, D. Frankowski, J. Herlocker et S. Sen, « Collaborative filtering recommender systems », in *The adaptive web*, Springer, Berlin, Heidelberg, 2007, p. 291-324.

[32] S. Leininger, T. Urich, M. Schloter et al., « Archaea predominate among ammonia-oxidizing prokaryotes in soils », *Nature*, t. 442, n° 7104, p. 806-809, 2006.

[33] M. Pavlov et R. Ichise, « Finding experts by link prediction in co-authorship networks », in *FEWS*, t. 290, 2007, p. 42-55.

[34] M. E. J. Newman, « Clustering and preferential attachment in growing networks », *Physical Review E*, t. 64, p. 025 102, 2001.

[35] A.-L. Barabási, H. Jeong, Z. Neda, E. Ravasz, A. Schubert et T. Vicsek, « Evolution of the social network of scientific collaborations », *Physica A : Statistical Mechanics and its Applications*, t. 311, n° 3-4, p. 590-614, 2002.

[36] X. Wang et Y. Peng, « Link prediction in social networks : the state-of-the-art », *Science China Information Sciences*, t. 58, n° 1, p. 1-38, 2015.

[37] U. Ashraf, A. Mahmood, M. Bennamoun et F. Boussaid, « Link prediction in complex networks using node similarity metrics based on gravitational law », in *2018 International Conference on Information Networking (ICOIN)*, IEEE, 2018, p. 234-239.

[38] Y. Yang et Z.-K. Zhang, « Link prediction based on common neighbors and distance in complex networks », *Physica A : Statistical Mechanics and its Applications*, t. 443, p. 129-135, 2016.

[39] S. Spiegel, J. Clausen, S. Albayrak et J. Kunegis, « Link prediction on evolving data using tensor factorization », in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, 2011, p. 100-110.

[40] L. Lü et T. Zhou, « Link prediction in complex networks : A survey », *Physica A : Statistical Mechanics and its Applications*, t. 390, n° 6, p. 1150-1170, 2011.

[41] M. E. J. Newman, « Clustering and preferential attachment in growing networks », *Physical Review E*, t. 64, n° 2, p. 025 102, 2001.

[42] A. Winslow, *Link Prediction Algorithms*, Online resource, 2014. adresse : `https://web.eecs.umich.edu/~winlo/notes/link-prediction.pdf`.

[43] L. Katz, « A new status index derived from sociometric analysis », *Psychometrika*, t. 18, n° 1, p. 39-43, 1953. doi : `10.1007/BF02289026`.

[44] T. Zhou, L. Lü et C. Jin, « Similarity index based on local paths for link prediction of complex networks », *Physical Review E*, t. 80, n° 4, p. 046 122, 2009. doi : `10.1103/PhysRevE.80.046122`.

[45] A. Papadimitriou, P. Symeonidis et Y. Manolopoulos, « Fast and accurate link prediction in social networking systems », *Journal of Systems and Software*, t. 85, n° 9, p. 2119-2132, 2012.

[46] J. Zhang, Y. Zhang, H. Yang et J. Yang, « A link prediction algorithm based on socialized semi-local information », *Journal of Computational Information Systems*, t. 10, n° 10, p. 4459-4466, 2014.

[47] R. Lichtenwalter et N. V. Chawla, « Link prediction : fair and effective evaluation », in *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, IEEE, 2012, p. 376-383. doi : `10.1109/ASONAM.2012.58`.

[48] Anonyme, *Les différents types de liens : TP, TN, FP, FN*, Figure 2.3 : Illustration des types de liens dans une matrice de confusion (TP, TN, FP, FN), utilisée à des fins académiques sans source précisée, 2025.