République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université de Mohamed El Bachir El Ibrahimi de Bordj Bou Arreridj
Présenté à la Faculté des Mathématiques et d'Informatique

Faculté des Mathématiques et d'Informatique

Département d'Informatique

MEMOIRE

Présenté en vue de l'obtention du diplôme
Master en informatique
Spécialité : **Réseau et multimédia**

Thème :

Étude expérimentale de la prédiction des liens dans les réseaux complexes

---

Soutenu publiquement le : 22/06/2024

Devant le jury composé de :

Président : **Dr**. Saifi Abdelhamid
Examinateur : **Dr**. Belazoug Mouhoub
Encadreur : **Dr**. Charikhi Mourad

Présenté par :

Bourezg Dounyazed
Chala Khaoula

2023/2024

DEDICACE

((وَ آخِرُ دَعْوَاهُمْ أَنِ الْحَمْدُ لِلّٰهِ رَبِّ الْعَالَمِين))

Louange à Allah au début et à la fin, car aucun chemin n'est terminé, aucun effort n'est accompli, aucune quête n'est achevée sans Sa grâce. Le rêve n'était pas court et la route n'était pas semée de facilités, mais je l'ai fait et je l'ai atteint. Je dédie avec tout mon amour mon mémoire de fin d'études :

*À celui qui m'a soutenue sans limites, et m'a encouragée à atteindre mes ambitions, "**mon père**".*

*À celle dont les prières ont été le secret de ma réussite, à **ma chère mère**, je te dédie cette réussite qui n'aurait pas été possible sans tes sacrifices.*

*À mes **sœurs** et **frères** « Hadjer, Hadil, Mohamed, Abdeljalil et Abdelaziz » mon soutien dans la vie.*

*À mon cher **grand-père**, qui nous a quittés en corps mais reste vivant dans nos cœurs et nos souvenirs. Ta petite-fille va obtenir son diplôme aujourd'hui. J'ai réussi et je te dédie ma réussite et sache que tu es fier de moi comme tu l'as toujours été. Que Dieu ait ton âme.*

*À chaque membre de ma **famille** qui a prié pour moi et souhaité ma réussite.*

*À celles qui m'ont entourée d'amour, m'ont toujours apporté de la force et ont été mon soutien à chaque épreuve, **mes amies**.*

*À mes proches, **Dounyazed** et **Maroua**.*

Khaoula.

DEDICACE

Je dédie ce mémoire à mes parents ***Mohammed*** *et* ***Samira*** *pour leur soutien constant et leur amour inconditionnel, Merci de toujours croire en moi.*

Je dédie ce mémoire à tous les membres de ma famille décédés. Votre amour, votre soutien et vos enseignements continuent de me guider chaque jour, Vous restez à jamais dans mon cœur et votre souvenir m'inspire à poursuivre mes rêves.

Je dédie ce mémoire à mon cher frère ***Amar****. Ta présence et ton inspiration ont été une source constante de force pour moi, Merci d'avoir toujours été là pour moi, même dans les moments les plus difficiles.*

Je dédie ce mémoire à mes Famille ***Bourezg*** *et* ***Boutankik****.*

Je dédie ce mémoire à tous mes amis.

Je dédie ce mémoire à mon binôme ***Khaoula****, pour son travail acharné, sa collaboration et son soutien tout au long de ce projet. Merci pour cette belle aventure partagée.*

Dounyazed.

REMERCIEMENT

Chaque page que nous avons soumis portait une part de notre santé. Nous y avons mis tout notre effort pour atteindre le meilleur de nos capacités. Quelle que soit le résultat et quelles que soient les notes, nous serons satisfaits et dirons ***Alhamdulillah****.*

Cette expérience nous a enseigné la patience et la persévérance, et nous a confirmé que le travail acharné n'est jamais vain, même si nous n'atteignons pas tous nos objectifs.

Merci à tous ceux qui nous ont soutenus et aidés tout au long de ce parcours, que ce soit par une parole gentille ou un conseil sincère. Nous remercions Dieu pour sa guidance et Sa protection tout au long de ce travail.

Nous tenons à exprimer notre profonde gratitude à notre encadreur ***Dr. Charikhi Mourad****, pour son encadrement précieux, sa patience et ses conseils avisés tout au long de ce projet, ainsi qu'à tous les professeurs du jury.*

RÉSUMÉ

Dans un contexte où les réseaux complexes jouent un rôle crucial dans de nombreux domaines tels que la suggestion d'amis dans les réseaux sociaux, la biologie et les systèmes de recommandation, la prédiction de liens devient un enjeu majeur pour comprendre et analyser ces structures interconnectées. Le thème de la prédiction des liens pour les réseaux complexes explore les méthodes visant à anticiper les connexions potentielles entre les entités au sein de ces réseaux, en se concentrant sur l'anticipation des connexions potentielles entre les nœuds d'un réseau. L'objectif principal de l'étude est d'évaluer et de comparer les performances des méthodes de prédiction de liens basées sur la similarité qui sont largement utilisées, en utilisant des ensembles de données variés. Une méthodologie rigoureuse incluant un processus de validation croisée à cinq volets et l'utilisation de mesures de performance est mise en œuvre pour comparer les différentes approches. Les résultats de l'étude mettent en évidence les forces et les faiblesses des différentes méthodes de prédiction de liens dans les réseaux complexes, ouvrant la voie à de futures recherches, notamment le développement d'une nouvelle mesure de prédiction de liens pour améliorer l'efficacité des algorithmes dans les réseaux complexes.

Mots Clés : la prédiction de liens, les réseaux complexes, similarité.

ABSTRACT

In a context where complex networks play a crucial role in many fields such as friend suggestion in social networks, biology, and recommendation systems, link prediction becomes a major issue for understanding and analyzing these interconnected structures. The topic of link prediction for complex networks explores methods aimed at anticipating potential connections between entities within these networks, focusing on anticipating potential connections between nodes. The main objective of this study is to evaluate and compare the performance of widely used similarity-based link prediction methods using various datasets. A rigorous methodology, including a five-fold cross-validation process and the use of performance measures, is implemented to compare different approaches. The study's results highlight the strengths and weaknesses of different link prediction methods in complex networks, paving the way for future research. This includes the development of a new link prediction measure to improve the efficiency of algorithms in complex networks.

Keywords: link prediction, complex networks, similarity.

ملخص

في إطار تلعب فيه الشبكات المعقدة دورًا حيويًا في العديد من المجالات مثل اقتراح الأصدقاء في الشبكات الاجتماعية، علم الأحياء وأنظمة التوصية، يصبح تنبؤ الروابط مسألة رئيسية لفهم وتحليل هذه الهياكل المترابطة. يستكشف موضوع تنبؤ الروابط للشبكات المعقدة الأساليب التي تهدف إلى توقع الروابط المحتملة بين الكيانات داخل هذه الشبكات، مع التركيز على توقع الروابط المحتملة بين العقد في الشبكة. الهدف الرئيسي من هذه الدراسة هو تقييم ومقارنة أداء أساليب تنبؤ الروابط القائمة على التشابه التي تستخدم على نطاق واسع، باستخدام مجموعات بيانات متنوعة. يتم تنفيذ منهجية صارمة تتضمن عملية تحقق متقاطع من خمسة أجزاء واستخدام مقاييس الأداء لمقارنة الأساليب المختلفة. تبرز نتائج الدراسة نقاط القوة والضعف في أساليب تنبؤ الروابط المختلفة في الشبكات المعقدة، مما يمهد الطريق لأبحاث مستقبلية. يشمل ذلك تطوير مقياس جديد لتنبؤ الروابط لتحسين كفاءة الخوارزميات في الشبكات المعقدة

الكلمات الرئيسية: تنبؤ بالروابط، الشبكات المعقدة، التشابه.

Table des matières

LISTE D'ABRÉVIATIONS

AA	Adamic Adar index
ACT	Average Commute Time
AP	Average Precision
AUC	Area Under the Curve
AUROC	Area Under the ROC Curve
CN	Common Neighbors
CNC	Common neighbor centrality
DNGR	Deep Neural Networks for Graph Representation
GAT	Graph Attention Network
GCPN	Graph Convolution Policy Network
GLHN	The Global LHN Index
GNN	Graph Neural Network
GTPN	Graph Transformation Policy Network
HMT	Hollywood Movie-Tie-In
HPI	Hub Promoted Index
HTC	High-Energy Theory Collaboration
JI	Jaccard index
KI	The Katz Index
LPI	The Local Path Index.
LRW	Local Random Walks

MF	Matrix Factorization
NSP	Negated Shortest Path
ORA-CNI	N Third-Order Resource Allocation Based on CN
PA	Preferential Attachment Index
PRM	Probabilistic Relational Method
RA	Resource allocation index
RA-CNI	Resource Allocation Based on Common Neighbor Interactions
ROC	Receiver Operating Characteristic
SDNE	Structural Deep Network Embedding
SR	SimRank
SRW	Superposed Random Walks
SVM	Support Vector Machine
UAL	United Airlines Network

LISTE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION GÉNÉRALE

INTRODUCTION GÉNÉRALE

Les réseaux complexes sont devenus des outils essentiels pour modéliser et analyser une grande variété de systèmes du monde réel, tels que les réseaux sociaux, les réseaux biologiques et les réseaux de transport. Ces réseaux sont caractérisés par un grand nombre de nœuds (entités) et d'arêtes (interactions) et présentent souvent des structures et des dynamiques complexes. Leur étude permet de mieux comprendre les interactions et comportements des systèmes complexes. L'importance croissante de ces réseaux dans la représentation des relations entre entités a conduit à une problématique intrigante : la prédiction des liens.

La prédiction de liens, est d'une grande importance dans de nombreux domaines d'application. Par exemple, la prédiction de liens dans les réseaux sociaux peut aider à identifier des communautés potentielles ou à recommander des amis à des utilisateurs. Dans les réseaux biologiques, la prédiction de liens peut aider à identifier des interactions protéine-protéine et à déceler des maladies. Les réseaux de transport et de télécommunications bénéficient également de cette capacité prédictive, améliorant la gestion des flux et des connexions.

La complexité de ces réseaux pose des défis uniques en termes d'analyse et de prédiction de liens, notamment en raison de la connectivité élevée des réseaux complexes et de la distribution inégale des degrés de connexion, nécessitant des approches spécifiques pour anticiper les connexions futures et modéliser les dynamiques du réseau.

L'utilisation de méthodes basées sur la similarité est une approche prometteuse pour la prédiction de liens dans les réseaux complexes. Ces méthodes exploitent les caractéristiques des nœuds et des liens existants pour prédire la probabilité de création de nouveaux liens. Plusieurs types de similarités peuvent être utilisés, tels que la similarité des attributs, la similarité des chemins et la similarité des voisins. Elles peuvent être basées sur des mesures de similarité locales, qui se concentre sur les informations structurelles de voisinage, ou sur des mesures de similarité globales, telles que la structure du réseau ou la dynamique des liens ainsi que les méthodes de similarité quasi-locales qui offrent un compromis entre les indices locaux (qui se concentrent sur les nœuds voisins) et les indices globaux (qui considèrent tout le réseau).

La question à laquelle nous répondrons dans cette thèse est : Quelles sont les méthodes efficaces basées sur la similarité pour prédire les liens dans les réseaux complexes ?

L'objectif de ce mémoire est d'explorer l'utilisation de méthodes basées sur la similarité pour la prédiction de liens dans les réseaux complexes et d'étudier, présenter et évaluer l'efficacité de ces approches dans les déférents data sets à travers les déférents métriques d'évaluation. Ces méthodes exploitent les caractéristiques des nœuds et des liens existants pour prédire la probabilité de création de nouveaux liens.

Notre travail consiste à implémenter et comparer les performances des méthodes de prédiction de liens dans les réseaux complexes.

Ce mémoire composé de quatre chapitres, d'une introduction et d'une conclusion générale.

Dans le premier chapitre, intitulé « Réseaux Complexes », nous présentons les concepts fondamentaux des réseaux complexes. Ce chapitre explore les différentes caractéristiques des réseaux complexes, les types de réseaux existants, et souligne l'efficacité de la représentation graphique des réseaux complexes pour en faciliter l'analyse.

Dans le deuxième chapitre, notre objectif est de fournir une compréhension globale du concept de « prédiction de lien » et de ses diverses applications. Nous explorerons les principales approches utilisées pour prédire les liens dans des réseaux complexes, avec un accent particulier sur les méthodes basées sur la similarité qui offrent des solutions efficaces à ce problème.

Le troisième chapitre fournit une introduction détaillée sur le processus de prédiction de liens dans les réseaux complexes. Ensuite, il présente la description des data sets étudiés, les métriques d'évaluation de la performance. Il se poursuit avec des informations sur la mise en œuvre et l'expérimentation, détaillant l'environnement de développement, les outils utilisés.

Le quatrième chapitre présente les résultats détaillés de nos expérimentations, ainsi que les tests et analyses effectués.

Enfin nous terminions notre travail par une conclusion générale.

CHAPITRE 01

LES RÉSEAUX COMPLEXES

I. CHAPITRE 01 : LES RÉSEAUX COMPLEXES

I.1. Introduction

Le monde qui nous entoure est constitué d'une multitude de systèmes complexes composés d'un grand nombre d'entités en interaction, qu'il s'agisse de réseaux biologiques, de réseaux sociaux, de réseaux de transport ou de réseaux technologiques.

Les réseaux complexes, présents dans une grande diversité de systèmes naturels et artificiels, peuvent être représentés et étudiés efficacement à l'aide de graphes. La représentation graphique des réseaux complexes constitue une approche fondamentale et puissante pour comprendre la structure et le fonctionnement des systèmes complexes qui nous entourent.

Dans ce chapitre, nous explorerons les concepts fondamentaux d'un réseau complexe. Nous examinerons ces caractéristiques. Nous discuterons également les différents types de réseau complexe et sa représentation graphique et matricielle.

I.2. Définition d'un réseau complexe

Un réseau complexe est un système composé d'un grand nombre d'entités interconnectées, appelées nœuds ou sommets, par des liens ou des arêtes. Ces entités peuvent représenter divers éléments, tels que des individus, des ordinateurs, des protéines, des entreprises, etc. Ce qui distingue un réseau complexe des réseaux simples ou réguliers est la présence de structures non triviales et de propriétés émergentes qui émergent des interactions entre les entités. Ces propriétés incluent, entre autres, la petite taille du monde, la distribution des degrés en loi de puissance, la modularité, la résilience aux pannes en cascade, la présence de communautés ou de clusters, et la navigation efficace [1].

Les réseaux complexes sont souvent étudiés à l'aide de modèles mathématiques, de techniques d'analyse de réseaux et de simulations informatiques pour comprendre leur structure globale, leur fonctionnement dynamique, leur évolution dans le temps, et pour prédire leur évolution. Ces réseaux se retrouvent dans de nombreux domaines tels que les réseaux sociaux, les réseaux biologiques, les réseaux informatiques, les réseaux de transport, les réseaux d'information, et bien d'autres, et constituent un domaine de recherche interdisciplinaire en constante évolution [2].

I.3. Caractéristiques des réseaux complexes

Les réseaux complexes présentent plusieurs caractéristiques qui ne se retrouvent pas dans les réseaux simples ou réguliers.

I.3.1. Connectivité élevée

La connectivité élevée est l'une des caractéristiques clés des réseaux complexes. Elle indique que la plupart des nœuds dans ces réseaux sont reliés entre eux par des liens ou des connexions. Cette propriété permet une transmission rapide et efficace de l'information à travers le réseau. Elle favorise également l'échange d'idées, de ressources et de connaissances entre les différents nœuds, ce qui peut stimuler l'innovation et la collaboration [3] [4].

I.3.2. Petits mondes

Les réseaux complexes ont souvent une structure de "petit-monde", ce qui signifie que la distance topologique moyenne entre deux nœuds est relativement courte par rapport à la taille totale du réseau. Cela se traduit par un faible nombre d'étapes nécessaires pour relier deux nœuds quelconques dans le réseau [1].

La propriété "Petits mondes" est une caractéristique fascinante des réseaux complexes qui a des implications importantes pour leur fonctionnement et leur dynamique. Cette propriété permet une propagation rapide de l'information, une synchronisation des événements, et une meilleure résilience aux pannes. La compréhension de la propriété "Petits mondes" est essentielle pour l'étude et la modélisation des réseaux complexes dans divers domaines.

I.3.3. Hétérogénéité : Distribution des degrés en loi de puissance

L'hétérogénéité des réseaux complexes se manifeste par la répartition inégale des connexions entre les nœuds. Certains nœuds ont un grand nombre de connexions (hubs), tandis que la majorité des nœuds ont un plus petit nombre de connexions. Cette propriété est souvent décrite comme une loi de puissance dans la distribution des degrés de nœuds, ce qui signifie qu'il existe un petit nombre de nœuds très fortement connectés, alors que la plupart des nœuds ont un nombre limité de connexions [5].

Cette particularité, observée dans de nombreux réseaux réels tels que le réseau Internet, remet en question les modèles traditionnels en montrant que la propagation d'informations ou

de maladies peut se produire même avec un petit nombre de nœuds fortement connectés. Comprendre cette distribution spécifique est essentiel pour analyser la dynamique, la résilience et la propagation dans les réseaux complexes, ouvrant ainsi de nouvelles perspectives pour la modélisation et la gestion efficace de ces réseaux [6].

I.3.4. Clustering « Transitivité ou clustering »

Le clustering, également appelé coefficient de regroupement ou coefficient de clustering, est une mesure de la densité locale d'un réseau complexe. Cette mesure permet de quantifier la tendance des nœuds d'un réseau à se regrouper en communautés ou en clusters.

Les réseaux complexes présentent une forte densité locale, mesurée par le coefficient de clustering, qui contraste avec une faible densité globale dans le graphe. Ce coefficient de clustering est défini comme la moyenne du ratio du nombre de voisins d'un nœud qui sont reliés entre eux sur le nombre total de liens possibles entre ces voisins [6].

I.3.5. Structure communautaire

La caractéristique de structure communautaire dans les réseaux complexes fait référence à la présence de groupes de nœuds fortement interconnectés entre eux, mais moins connectés avec le reste du réseau. Ces groupes de nœuds forment ce que l'on appelle des communautés, qui sont des sous-ensembles de nœuds qui partagent des propriétés ou des rôles similaires au sein du réseau complexe.

La détection de la structure communautaire dans un réseau complexe est essentielle pour comprendre la relation entre les nœuds individuels à l'échelle microscopique et les groupes ou communautés à l'échelle macroscopique. En identifiant ces communautés, on peut mieux appréhender l'organisation et les interactions au sein du réseau [6]**.**

Les communautés peuvent correspondre à diverses entités dans différents types de réseaux complexes. Par exemple, dans un réseau de pages web, les communautés pourraient représenter des groupes de pages traitant de sujets similaires. Dans un réseau social, les communautés pourraient correspondre des groupes d'individus ayant des interactions fréquentes entre eux [6]**.**

I.4. Les types de réseaux complexes

Les réseaux complexes se retrouvent dans une grande variété de domaines tels que l'informatique, la sociologie, la biologie, et la psychologie. Ces domaines incluent des réseaux aussi variés que l'Internet**,** les réseaux de transport, les réseaux sociaux humains, et les réseaux de protéines.

En conséquence, de nombreuses études ont été menées pour les comprendre. On peut les classer en quatre catégories distinctes :

I.4.1. Les réseaux sociaux

Un réseau social est défini comme un ensemble d'acteurs (individus, groupes ou organisations) qui sont reliés par des interactions sociales de nature familiale, amicales, sentimentales (liens forts) ou par des interactions plus distantes, les relations d'affaires (liens faibles) [7]. Ces connexions forment un tissu complexe de relations qui peuvent être représentées graphiquement à l'aide d'un graphe de réseau social. Dans un graphe de réseau social, les individus ou les groupes sont représentés par des nœuds, et les liens entre eux par des arêtes [8].

La structure du graphe, c'est-à-dire la manière dont les nœuds sont connectés, offre des informations précieuses sur les dynamiques sociales et les structures relationnelles au sein du réseau. En analysant ces interactions à travers les graphes de réseau social, il est possible de décrypter les schémas de contacts, les influences, et les comportements qui régissent les échanges au sein de ces réseaux interconnectés, permettant ainsi une meilleure compréhension des dynamiques sociales et des interactions humaines. Tous les réseaux qui impliquent des interactions sociales peuvent être considérés comme des réseaux sociaux indirects, tels que :

- Les réseaux sociaux en ligne (Facebook, Twitter, MySpace, etc.),
- Les réseaux en ligne de partage du contenu média (YouTube, Flickr, etc.),
- Les réseaux de télécommunications, les réseaux de communication par email,
- Les réseaux de discussion instantanée (Skype, Google Talk, Messenger, etc.).

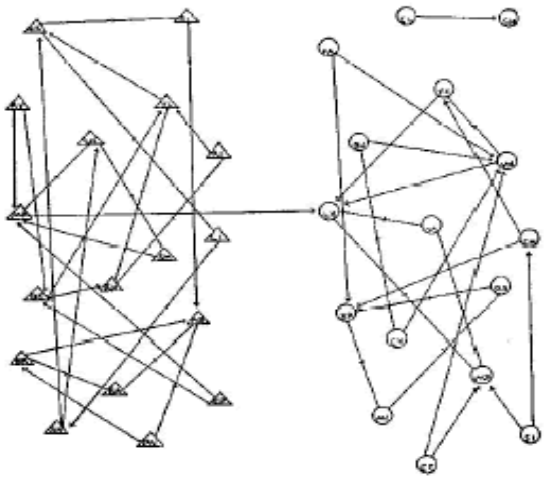


Figure I.1. Un des premiers réseaux sociaux dessinés à la main, datant de 1934, représentant les amitiés entre écoles enfants d'après Moreno [9]

I.4.2. Les réseaux d'information

Un réseau d'information peut être défini comme un ensemble de sommets interconnectés où chaque sommet est associé à des données, qu'elles soient structurées ou non structurées. Ces données peuvent prendre diverses formes, telles que des données numériques présentées sous forme d'un ensemble ou d'un vecteur, des données textuelles, ou tout autre type de données pertinentes [10].

Ces réseaux permettent de modéliser et d'analyser les flux d'informations, les relations entre les données et les schémas de connectivité dans divers contextes, tels que la science, l'Internet ou d'autres domaines où les données et leurs relations sont importantes.

Le World Wide Web peut être considéré comme un réseau d'information [11], avec ses pages web comme nœuds contenant des informations et ses hyperliens comme liens entre les pages. Les réseaux de citations universitaires et juridiques, les réseaux de brevets, les réseaux peer-to-peer, ainsi que les réseaux de recommandation sont d'autres exemples de réseaux d'information importants.

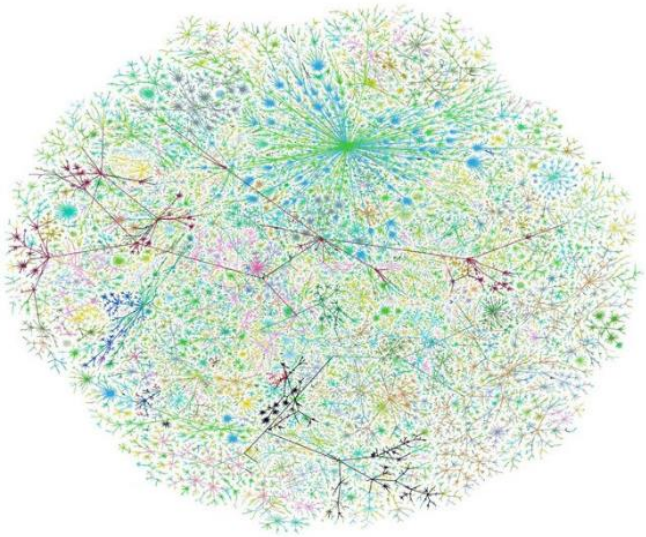


Figure I.2. Carte Internet complète du 29 juin 1999 [12]

I.4.3. Les réseaux biologiques

Les réseaux biologiques font partie des types de réseaux complexes servent à représenter et à analyser les interactions entre diverses entités biologiques comme les protéines [13], les gènes et les cellules [14]. Ces réseaux permettent de mieux comprendre les mécanismes physiologiques, les maladies, et les processus évolutifs [15].

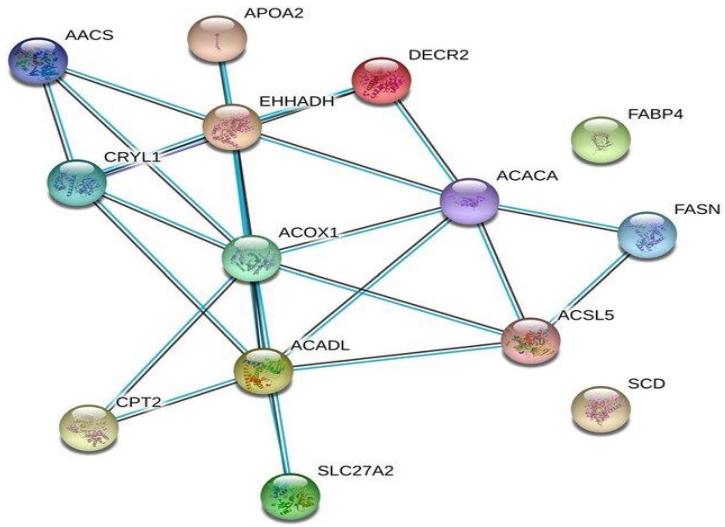


Figure I.3. Un réseau d'interactions entre protéines [16]

I.4.4. Les réseaux technologiques

Les réseaux technologiques également appelés réseaux artificiels, conçues pour la distribution de services ou de ressources spécifiques, tels que l'électricité, l'information, le transport, etc. Souvent créés par l'homme. Parmi les exemples courants de réseaux technologiques figurent les réseaux électriques, les réseaux de télécommunications, les réseaux informatiques, et les réseaux de distribution de données telles qu'Internet. Ils sont conçus pour permettre le transport efficace et la distribution de ces ressources sur de vastes étendues géographiques. Les études sur les réseaux technologiques incluent des analyses statistiques pour comprendre leur structure, leur fonctionnement, et leur évolution.

Les réseaux routiers et ferrés sont également des exemples de réseaux technologiques destinés à faciliter le déplacement des personnes et des marchandises. Ces réseaux sont composés de routes, autoroutes, voies ferrées et autres infrastructures connexes, conçues et construites pour répondre aux besoins de transport.

Internet, le réseau informatique mondial, est l'un des réseaux technologiques les plus étudiés et utilisés à ce jour. Il relie des millions d'ordinateurs à travers le monde, permettant le partage d'informations, la communication en temps réel, et l'accès à une multitude de services et de ressources en ligne [17].

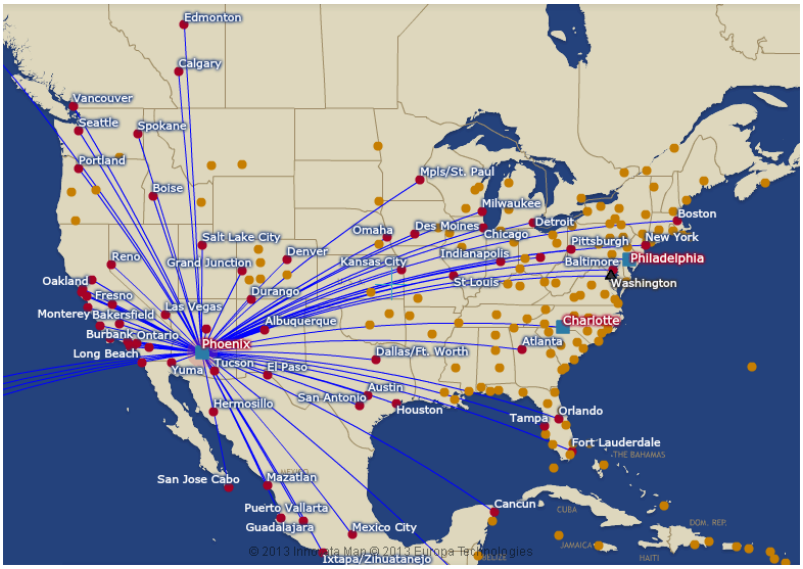


Figure I.4. Cartes routières des vols d'American Airlines en Amérique du Nord depuis Phoenix [18]

I.5. Représentation graphique

La représentation graphique des réseaux complexes est un domaine d'étude essentielle dans divers domaines tels que l'informatique, la biologie, la technologie, et bien d'autres. Ces réseaux, qui représentent des systèmes interconnectés d'entités et de leurs interactions, nécessitent souvent une visualisation claire et concise. Les graphes sont utilisés pour représenter les relations entre les individus dans les réseaux sociaux tels que Facebook, LinkedIn, et Twitter, pour modéliser les systèmes de transport tels que les réseaux routiers.

La représentation graphique, utilisant des structures mathématiques qui étudie les relations entre des objets discrets, composées :

- Un ensemble fini de sommets (nœuds), noté ***V***. Ce sont les entités individuelles du graphe.
- Un ensemble des liens (arêtes) noté ***E*** tel que $E \subseteq V \times V$. Ce sont les connexions entre les sommets du graphe.

Donc, un graphe ***G*** est défini par un couple ***G = (V, E)***, Où ***V*** est l'ensemble des sommets du graphe et ***E*** l'ensemble de liens.

Exemple :

Le graphe de la figure 5 représente un graphe non orienté ***G = (V, E)*** avec ***V*** = *{1, 2, 3, 4, 5}* et ***E*** = *{{1, 2}, {2, 3}, {2, 4}, {3, 4}, {3, 5}, {4, 5}}*.

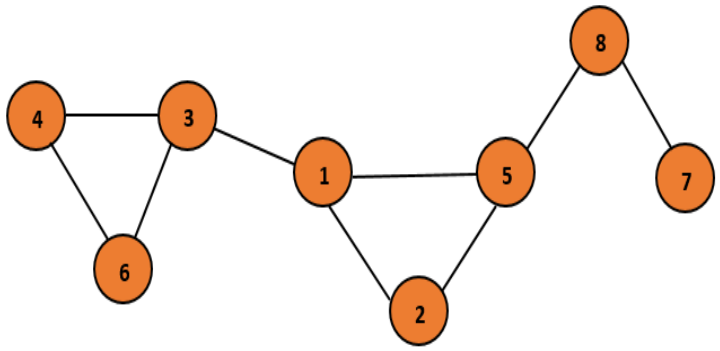


Figure I.5. Exemple d'un graphe non orienté

Pour représenter les différentes relations possibles entre les nœuds, il existe plusieurs types de graphes. Les plus courants sont les suivants :

- **Un graphe non orienté :** est un ensemble de nœuds connectés par des arêtes où chaque arête représente une relation non directionnelle entre deux sommets, ce qui signifie que la relation entre deux nœuds est symétrique.

- **Un graphe orienté :** À la différence d'un graphe non orienté, dans un graphe orienté, les arêtes ont une direction spécifique. Les arêtes sont appelées arcs et sont représentés graphiquement par $\boldsymbol{vi} \rightarrow \boldsymbol{vj}$. $\boldsymbol{vi}$ est le sommet initial, et $\boldsymbol{vj}$ le sommet terminal, Cela signifie que la relation entre deux nœuds peut être asymétrique.

- **Un graphe complet :** D'autre part, un graphe complet est un graphe dans lequel chaque paire de sommets distincts est reliée par une seule arête. La figure I.6(b) montre un exemple de graphe complet**.**

- **Un graphe connexe** : est un graphe dans lequel il existe un chemin reliant chaque paire de sommets distincts. Chaque nœud est relié à au moins un autre nœud par une séquence d'arêtes, ce qui signifie qu'il n'y a pas de "composantes" isolées dans le graphe. Un exemple de graphe connexe est représenté à la figure I.6(c).

- Une **composante connexe** ***C*** d'un réseau ***G*** est définie comme un sous-réseau connecté de ***G***. Deux composantes connexes ***C1 = (V1, E1)*** *et* ***C2 = (V2, E2)*** de ***G*** sont dites déconnectées s'il n'existe aucun chemin reliant un nœud $\boldsymbol{vi}$ de ***V1*** à un nœud $\boldsymbol{vj}$ de ***V2*** [19].

- **Un graphe pondéré :** Un graphe pondéré attribue un poids qui correspond à une valeur numérique affectée au lien**.** Ce poids pourrait représenter une distance, un coût, ou toute autre quantité pertinente associée à la connexion entre les nœuds. Un graphe pondéré peut être observé sur la figure 6(d).

- **Un graphe biparti :** dans un graphe l'ensemble de ses nœuds peut être divisé en deux sous-ensembles distincts, de telle manière que chaque lien (ou arête) du graphe relie un sommet d'un ensemble à un sommet de l'autre ensemble. La figure 6(e) explique cette définition.

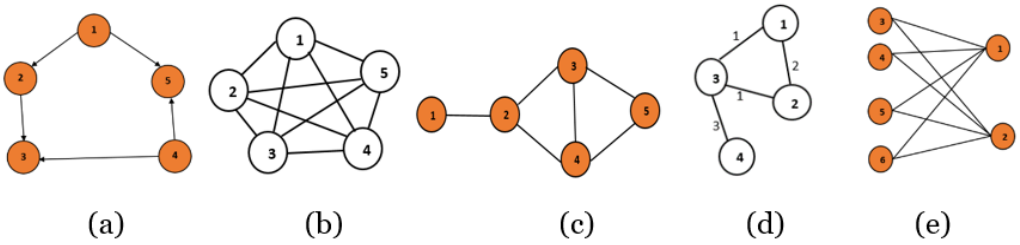
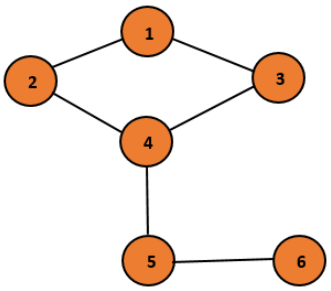


Figure I.6. Représentation graphique des différents types de graphes (a) graphe orienté, (b) graphe complet, (c) graphes connexes, (d) graphe pondéré, (e) graphe biparti

I.6. Représentation par matrices

I.6.1. Matrice d'adjacence

Une matrice d'adjacence ***M*** est une matrice carrée utilisée pour représenter un graphe ***G(V, E)*** de dimension ***n*** x ***n***, où n est le nombre de nœuds du graphe (chaque ligne et chaque colonne correspondent à un nœud). L'entrée **(*i, j*)** indique s'il existe une connexion entre les nœuds **i** et ***j***, telle que M_{ij} = {***1*** *si* ***(i, j)*** $\in$ ***E***, *et* M_{ij} = ***0*** *sinon*}.



	1	**2**	**3**	**4**	**5**	**6**
1	0	1	1	0	0	0
2	1	0	0	1	0	0
3	1	0	0	1	0	0
4	0	0	0	0	1	0
5	0	0	0	1	0	1
6	0	0	0	0	1	0

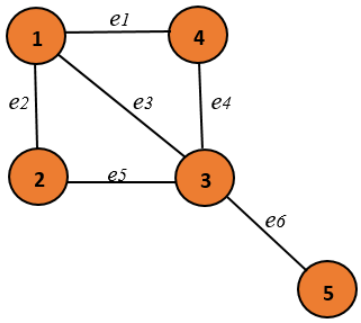
Figure I.7. Représentation par matrice d'adjacence du graphe non orienté

I.6.2. Matrice d'incidence

Une matrice d'incidence est une matrice ***n*** x ***p***, où n est le nombre de nœuds du graphe et *p* le nombre de d'arrêts (liens). Chaque élément de la matrice indique si un nœud est relié à une arête spécifique. Un graphe non orienté peut être représenté par une matrice d'incidence défini par :

M_{ij}= {**1** si le nœud v_i est une extrémité de l'arête e_j, **0** sinon}.

Exemple :



	e_1	e_2	e_3	e_4	e_5	e_6
1	1	1	1	0	0	0
2	0	1	0	0	1	0
3	0	0	1	1	1	1
4	1	0	0	1	0	0
5	0	0	0	0	0	1

Figure I.8. Représentation par matrice d'incidence du graphe non orienté

I.7. Mesures et propriétés

Les mesures des propriétés structurelles des graphes permettent de les comparer et de les comprendre et analyser leur fonctionnement.

I.7.1. Le voisinage d'un nœud

Le voisinage d'un nœud, souvent désigné par ***Γ(v),*** est constitué de tous les nœuds reliés directement à ce nœud par des arêtes.

Dans un graphe non orienté ***G = (V, E),*** l'ensemble ***{e, v} ∈ E***, où **e** et **v** sont des nœuds de ***V***, représente une arête reliant ces deux nœuds. Le voisinage ***Γ(v)*** d'un nœud ***v ∈ V*** est constitué de tous les nœuds ***e ∈ V*** qui sont reliés à **v** par de telles arêtes**.**

I.7.2. Degré d'un nœud

Dans un graphe non orienté, le degré d'un nœud couramment appelé ***k(x),*** est la mesure du nombre d'arêtes incidentes à ce nœud, en d'autres termes le nombre de voisins que possède ce nœud.

I.7.3. Le chemin

Un chemin de longueur ***k*** entre un nœud ***x*** et un nœud ***y*** d'un graphe, appelée $\boldsymbol{paths}_{x,y}^{k}$ est un ensemble d'arêtes connectée qui assurent la liaison d'une suite de nœuds du graphe.

I.7.4. La densité d'un graphe

La densité d'un graphe **D** est une échelle qui évalue le rapport entre le nombre réel d'arêtes du graphe et le nombre maximal de ses arêtes possibles.

Soit ***G = (V, E)*** un graphe non orienté simple, la densité du graphe **G** est :

$$\boldsymbol{D = \frac{2|E|}{|V|.([V|-1)}}$$

I.8. Conclusion

En conclusion, ce chapitre a présenté les concepts fondamentaux de réseau complexe. Nous avons exploré les différentes caractéristiques qui les définissent, les types de réseau complexe qui existent.

La représentation des réseaux complexes sous forme de graphe a grandement facilité leur analyse, permettant le calcul et l'extraction d'informations et de propriétés essentielles caractérisant les nœuds, les liens et l'ensemble du réseau.

Le but de ce chapitre est de mettre en lumière l'importance cruciale des réseaux complexes en tant que technologie permettant de modéliser une multitude de systèmes naturels ou artificiels, englobant divers domaines de notre vie.

Le chapitre suivant approfondira la question de la prédiction de liens en explorant les différentes méthodes utilisées dans les réseaux complexes.

CHAPITRE 02

MÉTHODES DE PRÉDICTION DE LIENS DANS LES RESÉAUX COMPLEXES

II. CHAPITRE 02 : MÉTHODES DE PRÉDICTION DE LIENS DANS LES RÉSEAUX COMPLEXES

II.1. Introduction

La prédiction de liens dans les réseaux complexes est un problème fondamental dans le domaine de l'analyse des réseaux. Cette discipline vise à anticiper ou à prédire les connexions potentielles entre les entités au sein d'un réseau, telles que les liens entre des utilisateurs dans un réseau social, des protéines dans un réseau biologique ou des sites Web dans un réseau d'information ou d'autres types de réseaux complexes.

Différentes approches ont été explorées pour la prédiction des liens, notamment les méthodes d'apprentissage supervisées et non supervisées et les mesures de similarité et de probabilité. Ces méthodes sont classées sur plusieurs différentes classifications en fonction des règles spécifiques qu'elles suivent et de la quantité et du type d'informations qu'elles utilisent [20].

L'objectif de ce chapitre est d'exposer le terme « prédiction de liens » ainsi que ses domaines d'application. Il présente les principales approches de prédiction de liens dans les réseaux complexes qui peuvent être employés pour prédire ces liens, en se concentrant sur les méthodes basées sur la similarité pour résoudre ce problème.

II.2. Le problème de prédiction de liens

Le problème de prédiction de lien dans les réseaux complexes consiste à inférer l'existence de liens manquants (c'est-à-dire des liens existants mais non observés) dans un réseau donné à un instant donné, ou à prédire les relations futures qui se formeront dans ce réseau. En d'autres termes, il s'agit de prédire les interactions ou relations potentielles entre les entités ou acteurs du réseau sur la base des connexions déjà observées. Ce problème est souvent formalisé comme un problème de classement, où des paires de nœuds non connectés se voient attribuer un score reflétant la probabilité d'existence d'un lien entre eux [20]. Le problème de classement de nouveaux liens qui apparaissent dans le réseau en deux classes : **{normal, anormal}**, en fonction de l’évolution du réseau [21]. Cela soulève des défis uniques en termes

de prédiction, nécessitant des approches innovantes pour anticiper efficacement ces connexions.

II.3. Applications de la prédiction de lien

Les applications de prédiction de liens permettent de prédire les connexions entre différents éléments d'un réseau. Ces applications sont largement utilisées dans des domaines variés tels que les réseaux sociaux, les systèmes de recommandation, la biologie, la prédiction collaborative, les réseaux routiers, etc. Ces prédictions sont précieuses pour prendre des décisions éclairées, formuler des recommandations d'optimisation et améliorer les performances globales du système. En général, la prédiction des liens peut être utilisée dans n'importe quel réseau où la prévision des connexions ou des relations futures est précieuse :

II.3.1. Les Réseaux sociaux

Dans le domaine des réseaux sociaux, la prédiction de liens permet de recommander des amis ou des contacts potentiels à un utilisateur en se basant sur ses interactions passées. Cela améliore l'expérience utilisateur en leur permettant de retrouver plus facilement des connaissances parmi un grand nombre d'utilisateurs enregistrés et favorise l'expansion de son réseau. De plus, cela aide les plateformes à personnaliser le contenu affiché, augmentant ainsi l'engagement et la fidélité des utilisateurs [20].

II.3.2. Les systèmes de recommandation

L'utilisation de la prédiction de liens dans la recommandation de produits permet aux entreprises de cibler plus efficacement les besoins et les préférences des consommateurs. En analysant les habitudes d'achat et les interactions passées, les systèmes de recommandation peuvent proposer des produits pertinents, augmentant ainsi les ventes et la satisfaction client. Cela marque un tournant majeur dans l’e-commerce [22].

II.3.3. Biologie

En biologie, les techniques de prédiction de liens sont appliquées pour identifier les interactions potentielles entre des paires de protéines dans les réseaux d'interaction protéine-protéine ou des réseaux de régulation des gènes. Cela permet de réduire les coûts et le temps associés aux expériences in vitro en ciblant les interactions les plus prometteuses [23].

II.3.4. La Santé

Les applications de prédictions de liens dans le domaine de la santé ont de vastes implications à travers divers aspects. Elles permettent aux médecins d'améliorer le diagnostic médical en identifiant les risques de maladies, en diagnostiquant plus rapidement les affections et en recommandant des traitements personnalisés. De plus, elles facilitent la prévention des maladies en identifiant les facteurs de risque et en permettant des mesures préventives précoces. Enfin, dans la gestion des données médicales, elles optimisent le suivi des patients, la planification des traitements et la prise de décisions cliniques en détectant des corrélations cachées entre les données médicales, en prévoyant les besoins des patients et en anticipant des complications potentielles, permettant ainsi d'améliorer l'efficacité opérationnelle et la qualité des services dans les établissements de soins de santé [24] [25] [26].

II.3.5. Collaboration scientifique

Dans le domaine de la recherche scientifique, la prédiction de liens est utilisée pour anticiper les collaborations potentielles entre auteurs ou groupes de recherche. Cela aide à mieux comprendre l'évolution des domaines de recherche en identifiant les collaborations futures [27].

II.4. Description du problème de prédiction de liens

Le problème de la prédiction de liens dans les réseaux complexes peut être représenté par un graphe non orienté $G = (\boldsymbol{V}, \boldsymbol{E})$ où $\boldsymbol{V}$ est l'ensemble de nœuds et $\boldsymbol{E}$ est l'ensemble des arêtes. Soit $\boldsymbol{n} = |\boldsymbol{V}|$ le nombre de nœuds dans le réseau, également appelé taille du réseau et $\boldsymbol{e} = |\boldsymbol{E}|$ le nombre de liens.

S'il y a un lien entre chaque paire de nœuds, donc le graphe **G** est un graphe complet et contient un total de $\boldsymbol{n}.\frac{\boldsymbol{n-1}}{\boldsymbol{2}}$ liens. Au sein de ce graphe, il existe des liens manquants qui ne sont pas encore présents mais qui pourraient se former dans le futur entre certains nœuds. On note $\mathbf{E'}$ l'ensemble de ces liens absents : $\mathbf{E'} = (\boldsymbol{x}, \boldsymbol{y})\ \{\in \mathbf{V}\ \mathbf{2}\ |\ (\boldsymbol{x}, \boldsymbol{y})\ \notin \mathbf{E}\}$. En général, la taille de. $\boldsymbol{E'}$est $\boldsymbol{n}.\frac{\boldsymbol{n-1}}{\boldsymbol{2}} - \boldsymbol{e}.$

Identifier et anticiper ces liens manquants ou futurs dans l'ensemble $\mathbf{E'}$ est l'objectif de la prédiction de liens. Pour résoudre le problème, chaque lien $\boldsymbol{e}(\boldsymbol{x}, \boldsymbol{y})$ tel que $\boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{V}$ et

$(x, y) \in E'$, est attribué à un score $s(x, y)$ qui représente la mesure de la probabilité d'existence de liens entre $x\ et\ y$. Une liste ordonnée décroissante de tous les liens non observés est fournie en fonction de leurs scores et les liens en haut sont les plus possibles d'exister [21].

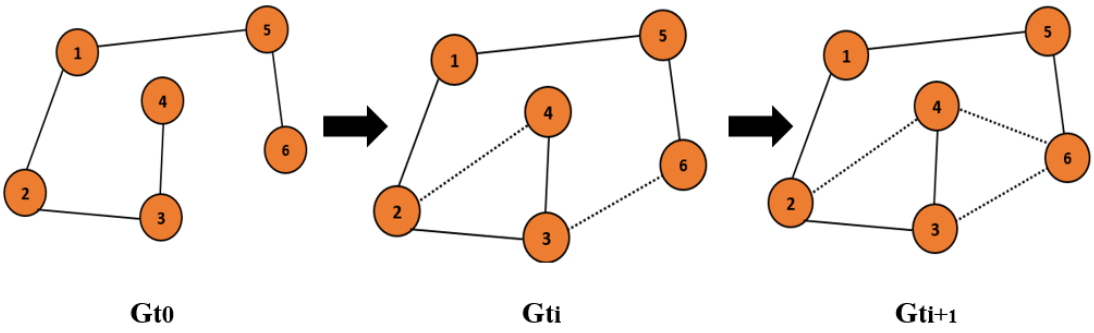


Figure II.1. La prédiction de liens dans un graphe

II.5. Classification des approches de prédiction de liens

La prédiction des liens est devenue un domaine de recherche largement étudié dans la littérature, avec de nombreuses études et revues analysées les techniques de prédiction et qui ont proposé différentes classifications de ces techniques.

On peut classer les approches de prédiction des liens en deux grandes catégories : les approches basées sur l'apprentissage et les approches basées sur les caractéristiques.

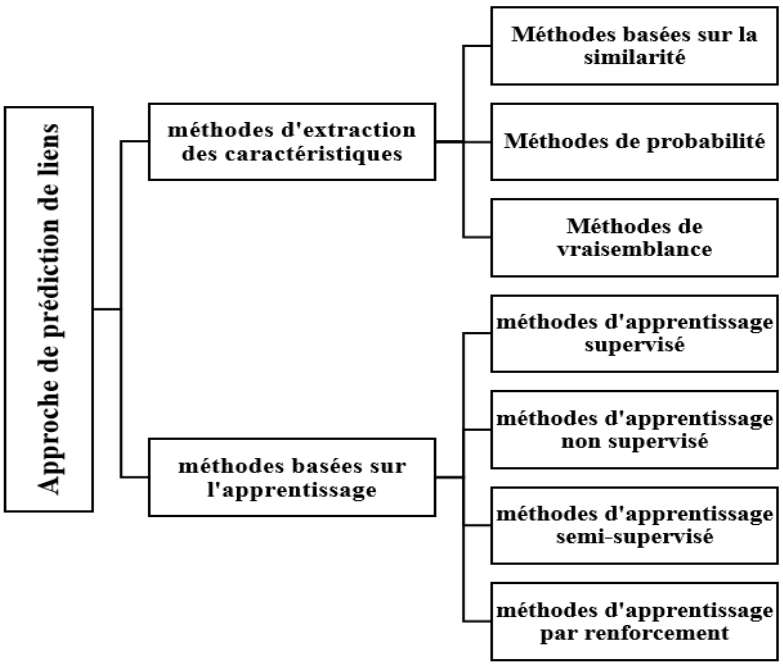


Figure II.2. Classification des approches de prédiction des liens [28]

II.5.1. Les méthodes sur l'extraction de caractéristiques

L'analyse de caractéristiques permet d'identifier des informations essentielles à partir de données de réseau. Ces informations peuvent ensuite servir à l'entraînement de modèles prédiction de liens, englobant des approches fondées sur la similarité, les probabilités et la vraisemblance.

II.5.1.1. Les méthodes basées sur la similarité

Les méthodes basées sur la similarité sont des approches utilisées pour prédire les liens dans les réseaux complexes. Elles exploitent les similarités existantes entre les nœuds du réseau pour prédire les liens manquants, en se concentrant sur l'idée que des nœuds similaires ont tendance à être connectés entre eux. Plusieurs types de similarités peuvent être utilisés, tels que la similarité des attributs, la similarité des chemins et la similarité des voisins. Ces approches calculent des scores de probabilité ou de similarité pour anticiper quels nœuds non connectés pourraient éventuellement former des liens dans le futur [20].

Les méthodes de similarité dans les réseaux complexes peuvent être classées en trois catégories principales [29]:

- ✓ Méthodes de similarité locale,
- ✓ Méthodes de similarité globale,
- ✓ Méthodes de similarité Quasi-locale.

- **Similarité locale**

La similarité locale dans le contexte de la prédiction de liens dans les réseaux complexes se réfère à la mesure de la similitude entre les nœuds d'un réseau en se basant sur leurs informations structurelles de voisinage. Ces approches sont plus rapides que les techniques non locales, efficaces et hautement parallélisables et plus nombreux, par contre les attributs des nœuds ne sont pas généralement disponibles ou sont cachés [20] [29]. Dans la section suivante nous allons lister les mesures les plus connus cités dans la littérature.

Dans ce qui suit : $\boldsymbol{\Gamma(x)}$ représente le voisinage de $\boldsymbol{x}$ qui est l'ensemble des nœuds connectés à un nœud $\boldsymbol{x}$ par une arête. Le degré d'un nœud $\boldsymbol{x}$ est représenté par le symbole $\boldsymbol{|\Gamma(x)|}$ ou $\boldsymbol{k(x)}$ et est défini comme le degré ou le nombre d'arêtes qui sont reliées au nœud.

- **Voisins communs (Common Neighbors)**

Est une mesure de similarité locale qui compte le nombre de voisins communs entre deux nœuds $\boldsymbol{x}$ et $\boldsymbol{y}$, plus ce nombre est élevé, plus la similarité entre les deux nœuds est considérée comme élevée [30]. Peut être mesurée comme suite :

$$\boldsymbol{S_{xy}^{CN} = |\Gamma(x) \cap \Gamma(y)|} \qquad \text{(II.1)}$$

- **Indice de Jaccard (JI)**

C'est une Mesure de similarité locale qui calcule le rapport entre les voisins communs et le nombre total de voisins pour deux nœuds [31], donnée par :

$$\boldsymbol{S_{xy}^{JI} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}} \qquad \text{(II.2)}$$

- **Indice d'Adamic-Adar (AA)**

Cette méthode permet d'ajouter la somme des poids des nœuds qui sont connectés aux deux nœuds $\boldsymbol{x}$ et $\boldsymbol{y}$, le point est dépendu de degré des nœuds [32], la formule de l'indice Adamic-Adar est la suivante :

$$\boldsymbol{S_{xy}^{AA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log\ k(z)}} \qquad \text{(II.3)}$$

- **Indice d'allocation des ressources (RA)**

L'indice d'allocation des ressources est une mesure de similarité locale entre deux nœuds et ne sont pas directement connectés, le nœud x peut envoyer des ressources au nœud y via leurs voisins communs. Chaque voisin commun agit comme un transmetteur de ressources et distribue équitablement une unité de ressource à tous ses voisins [33]. Cette méthode peut être calculée comme :

$$\boldsymbol{S_{xy}^{RA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k(z)}} \qquad \text{(II.4)}$$

- **Allocation de ressources basée sur les interactions de voisins communs (RA-CNI)**

Cette approche est motivée par le processus d'allocation de ressources où chaque nœud distribue une unité de ressource à ses voisins. Cependant, cette méthode considère également le retour des ressources dans la direction opposée [34]. Cela est défini comme :

$$S_{xy}^{\mathbf{RA-CNI}} = \sum_{Z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k(z)} + \sum_{e_{i,j} \in E, \Gamma(i) < \Gamma(j), i \in \Gamma(x), j \in \Gamma(y)} \left(\frac{1}{k(i)} - \frac{1}{k(j)}\right) \quad \text{(II.5)}$$

- **L'indice de Salton (similarité cosinus)**

Est une mesure de similarité entre deux nœuds dans un réseau. Cette mesure est étroitement liée à l'indice de Jaccard et est souvent utilisée dans des contextes pratiques pour évaluer la similarité entre deux ensembles de voisins. L'indice de Salton donne une valeur environ deux fois supérieure à l'indice de Jaccard [35]. Cette mesure évalue la similarité entre deux ensembles en mesurant le cosinus de l'angle entre eux dans un espace vectoriel [36]. La formule générale de l'indice de Salton est la suivante :

$$S_{xy}^{SI} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{\Gamma(x) * \Gamma(y)}} \quad \text{(II.6)}$$

- **Indice de Sørensen**

Cet indice est principalement utilisé pour comparer la similarité entre différentes données de communautés écologiques [37]. Malgré sa similarité avec l'indice Jaccard, il est moins sensible aux valeurs aberrantes [38]. La similitude de Sørensen est définie comme :

$$S_{xy}^{Sorensen} = \frac{2|\Gamma(x) \cap \Gamma(y)|}{k(x) + k(y)} \quad \text{(II.7)}$$

- **L'index promu par le hub (Hub Promoted Index (HPI)**

Est une mesure de similarité introduite pour capturer la structure hiérarchique des réseaux, en favorisant la formation de liens entre les nœuds de faible degré et les hubs, tout en évitant les liens entre les nœuds centraux. La formule de similarité de l'indice HPI compare le nombre de voisins communs entre deux nœuds par rapport au degré minimum des deux nœuds. Cette mesure vise à mettre en évidence les relations entre les nœuds centraux (hubs) et les nœuds périphériques de faible degré dans un réseau [39]. Donnée par :

$$S_{xy}^{HPI} = \frac{|\Gamma(x) \cap \Gamma(y)|}{min\{k(x), k(y)\}} \quad \text{(II.8)}$$

- **L'indice d'attachement préférentiel (PA)**

Mesure la probabilité qu'un lien se forme entre deux nœuds en se basant sur les degrés de ces nœuds. En partant du principe que les nœuds présentant des degrés plus élevés sont plus susceptibles de former de nouvelles connexions [40]. Peut être estimée comme suite :

$$S_{xy}^{PA} = k(x) * k(y) \qquad \text{(II.9)}$$

- **Allocation de ressources de troisième ordre basée sur des interactions de voisinage commun (ORA-CNI)**

Est une méthode avancée de prédiction de liens qui étend le concept d'allocation de ressources basée sur les interactions de voisins communs en prenant en compte la distance entre trois chemins. Dans cette méthode, l'allocation des ressources est redéfinie pour les nœuds distants de trois chemins en considérant les interactions entre ces nœuds [34]. La formule de calcul de l'indice ORA-CNI est donnée par :

$$S_{xy}^{\mathrm{ORA-CNI}} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k(z)} + \sum_{e_{i,j} \in E, \Gamma(i) < \Gamma(j), i \in \Gamma(x), j \in \Gamma(y)} \left(\frac{1}{k(i)} - \frac{1}{k(j)}\right) + \beta \sum_{[x,p,q,y] \in path_{x,y}^{3}} \frac{1}{k(p)k(q)} \qquad \text{(II.10)}$$

D'où le coefficient d'amortissement β permet de moduler l'effet du terme d'allocation de ressources sur une période de trois sauts.

- **Méthodes de similarité globale**

Ces méthodes de similarité globale évaluent la similarité entre les nœuds en prenant en compte des informations plus étendues sur la topologie du réseau, par opposition aux méthodes de similarité locale qui se concentrent sur les voisins immédiats des nœuds. Cette mesure est basée sur la comparaison des propriétés globales des nœuds telles que la distance des chemins les reliant. Cependant, en raison de leur complexité computationnelle, ces approches peuvent être difficilement applicables aux grands réseaux et leur parallélisation peut s'avérer très complexe, surtout dans des environnements distribués où la topologie complète du réseau n'est pas forcément connue de tous les agents informatiques [20].

Voici quelques-unes des méthodes les plus répandues de cette catégorie :

- **Chemin le plus court annulé (NSP)**

Est une mesure de similarité de base basée sur le graphe. Cette méthode nécessite le calcul des chemins les plus courts entre toutes les paires de nœuds dans le réseau, ce qui peut être réalisé efficacement avec l'algorithme de Dijkstra. Cependant, malgré sa simplicité, la méthode NSP a une précision de prédiction relativement faible par rapport à d'autres méthodes plus complexes prenant en compte des chemins multiples [41]. La similarité peut être calculée comme suit :

$$S_{xy}^{NSP} = -\left|\boldsymbol{chemin\ le\ plus\ court}_{x,y}\right| \qquad \text{(II.11)}$$

- **L'indice de Katz (KI)**

Cette mesure calcule le nombre de chemins différents qui relient des paires de nœuds, le poids de chaque chemin est basé sur sa longueur, ainsi, les chemins courts ont un poids plus élevé que les chemins longs qu'en lui affectant des poids faible [42]. Donnée par :

$$S_{xy}^{\mathbf{KI}} = \sum_{l=1}^{\infty} \boldsymbol{\beta}^{\mathbf{l}} \cdot \left|\boldsymbol{paths}_{\mathbf{x,y}}^{\mathbf{l}}\right| \qquad \text{(II.12)}$$

D'où $\boldsymbol{paths}_{\mathbf{x,y}}^{\mathbf{l}}$ est l'ensemble de tous les chemins de longueur **l** reliant **x** et **y**, et **β** est un paramètre libre contrôlant les poids des chemins **(0 < β < 1).**

- **L'indice global de Leicht-Holme-Newman (GLHN)**

Est un indice qui combine des aspects des indices de Katz et des chemins locaux pour évaluer la similarité entre les nœuds d'un réseau complexe, en se basant sur la similarité de leurs voisins immédiats. La similarité entre toutes les paires de nœuds est définie comme :

$$S_{xy}^{\mathbf{GLHNI}} = \boldsymbol{2m\lambda_1 D^{-1}(I - \frac{\varphi A}{\lambda_1})^{-1} D^{-1}} \qquad \text{(II.13)}$$

D'où $\boldsymbol{\lambda_1}$ est la plus grande valeur propre de $\boldsymbol{A}$ et $\boldsymbol{m}$ est le nombre total d'arêtes dans le réseau. $\boldsymbol{D}$ est une matrice diagonale de degré de la forme $\boldsymbol{diag\{k1,..,kn\}}$, $\boldsymbol{\varphi}$ est un paramètre libre ($0 < \boldsymbol{\varphi} < 1$). Le choix de $\boldsymbol{\varphi}$ dépend du réseau étudié, et un $\boldsymbol{\varphi}$ plus petit attribue plus de poids aux chemins plus courts [43].

- **Temps de trajet moyen (ACT)**

Cette mesure se base sur le concept du nombre moyen d'étapes qu'un marcheur aléatoire doit effectuer pour se déplacer d'un nœud x à un autre y dans le réseau. En considérant que deux nœuds sont plus similaires s'ils ont un temps de trajet moyen plus court entre eux [44]. L'**ACT** peut être calculé comme :

$$\boldsymbol{S_{xy}^{ACT}} = \frac{\mathbf{1}}{\boldsymbol{L_{x,x}^{+}+L_{y,y}^{+}-2L_{x,y}^{+}}} \qquad \text{(II.14)}$$

$\boldsymbol{L = D - A}$, où D est une matrice diagonale de taille |V| dont chaque élément $\boldsymbol{Di,i = |\Gamma i|}$ correspond au degré entrant du nœud i et A est la matrice d'adjacence du graphe

$$\boldsymbol{L^{+}} = \left(\boldsymbol{L} - \frac{\boldsymbol{ee^{T}}}{\boldsymbol{n}}\right)^{\mathbf{-1}} + \frac{\boldsymbol{ee^{T}}}{\boldsymbol{n}} \qquad \text{(II.15)}$$

$\boldsymbol{n}$ c'est le nombre des nœuds, $\boldsymbol{e}$ c'est un vecteur composé de 1.

- **Les méthodes de similarité Quasi-Locales**

Les indices quasi-locaux sont des mesures de similarité qui offrent un compromis entre les indices locaux (qui se concentrent sur les nœuds voisins) et les indices globaux (qui considèrent tout le réseau). Ces méthodes prennent en compte des informations topologiques plus étendues que les méthodes purement locales, mais évitent la complexité computationnelle des méthodes globales en ne considérant pas l'ensemble du réseau pour calculer la similarité entre les nœuds. Diverses mesures de similarité quasi-locales peuvent être utilisées à des fins de prédiction de lien, notamment :

- **L'indice du chemin local (LPI)**

Dans le contexte de l'indice de chemin local, une matrice de similarité est calculée en prenant en compte des chemins avec un nombre fini de longueurs, généralement utilisée avec longueur l= 3 en raison de sa complexité algorithmique [33], et donné par :

$$\boldsymbol{S_{xy}^{LPI}} = (\boldsymbol{A})^{\mathbf{2}} + \boldsymbol{\beta}(\boldsymbol{A})^{\mathbf{3}} \qquad \text{(II.16)}$$

D'où : $\boldsymbol{A}$ est la matrice d'adjacence du graphe (représente les connexions entre les nœuds du graphe), $\beta < 1$ est défini sur une valeur minimale pour que les chemins augmentent le poids des chemins plus.

- **Marches aléatoires locales (LRW)**

Cette approche est une variante des méthodes de marche aléatoire classiques, mais elle limite le nombre d'itérations à un petit nombre fixe a priori **t**, ce qui la classe parmi les indices quasi-locaux [44]. La métrique est formalisée par :

$$S_{xy}^{\mathbf{LRW}}(t) = \frac{k_x}{2|E|} . \pi_{xy}(t) + \frac{k_y}{2|E|} . \pi_{yx}(t) \qquad \text{(II.17)}$$

D'où : $\pi_{xy}(t)$ dénote la probabilité obtenue par le processus marche aléatoire lors de l'itération ***t***.

- **Marches aléatoires superposées (SRW)**

Est une méthode de prédiction de liens dans les réseaux complexes qui vise à améliorer la sensibilité des méthodes basées sur la marche aléatoire à la topologie du réseau dans les zones distantes. Cette approche est basée sur la méthode de marche aléatoire locale et a été proposée pour surmonter ce problème en relâchant continuellement le marcheur au nœud de départ.

Dans cette méthode, chaque contribution du marcheur est superposée pour calculer la similarité entre deux nœuds [44]. La formule de calcul de la similarité est donnée par :

$$S_{xy}^{\mathbf{SRW}}(t) = \sum_{l=1}^{t} \left(\frac{k_x}{2|E|} . \pi_{xy}(l) + \frac{k_y}{2|E|} . \pi_{yx}(l) \right) \qquad \text{(II.18)}$$

$\pi_{xy}(l)$ Représente la probabilité de transition à l'étape l du processus de marche aléatoire, indépendamment des nœuds de départ et d'arrivée.

- **Common neighbor centrality index (CNC)**

Cette méthode combine deux aspects importants des nœuds dans un réseau : le nombre de voisins communs entre deux nœuds (Common Neighbors) et la centralité des nœuds (Centrality). Le voisin commun fait référence aux nœuds communs entre deux nœuds distincts dans un réseau. Plus le nombre de voisins communs est élevé, plus il est probable qu'il existe une relation entre ces deux nœuds. La Centralité, quant à elle, mesure l'importance d'un nœud dans un réseau. Dans le contexte du Common Neighbors Centrality Index, la centralité est souvent basée sur des mesures telles que la proximité (closeness) ou l'intermédiarité (betweenness) d'un nœud [45]. Voici la formule pour calculer le score de similarité entre les nœuds ***x*** et ***y*** :

$$S_{xy}^{CNC} = \alpha(|\Gamma(x) \cap \Gamma(y)|) + (1 - \alpha).\frac{n}{d_{xy}} \qquad \text{(II.19)}$$

α est un paramètre qui varie entre [0,1], La valeur d_{xy} représente la distance la plus courte entre les nœuds x et y dans le réseau.

II.5.1.2. Les méthodes basées sur la probabilité

Les méthodes basées sur les probabilités utilisent la nature probabiliste de la connectivité réseau pour établir la probabilité qu'un lien existe entre deux nœuds. Par la suite, ces méthodes utilisent cette probabilité comme moyen d'anticiper la présence ou l'absence d'un lien.

Nous allons vous montrer un exemple de méthode basée sur la probabilité. Une méthode relationnelle probabiliste (PRM) décrit une distribution de probabilité sur des instanciations qui sont en accord avec un graphe d'instanciation donné en spécifiant un modèle probabiliste complexes tels que les réseaux bayésiens [46]. Voici quelques autres exemples de méthodes basées sur la probabilité :

- Les méthodes relationnelles d'entités.
- Les méthodes relationnelles stochastiques.

II.5.1.3. Les méthodes basées sur la vraisemblance

On peut classée les méthodes de vraisemblance en deux catégories principales d'algorithmes les algorithmes conçus pour les modèles de structure hiérarchique et les algorithmes conçus pour les modèles de blocs stochastiques.

Les réseaux possédant une structure hiérarchique spécifique utilisent des algorithmes de modèle de structure hiérarchique pour prédire les connexions manquantes, Cette approche est applicable aux réseaux qui possèdent une structure hiérarchique perceptible, tels que les réseaux impliqués dans des attaques terroristes et les chaînes alimentaires [47], Lorsque certains réseaux ne respectent pas un schéma hiérarchique, on utilise les algorithmes des modèles de blocs stochastiques. Ils reposent sur l'estimation de la vraisemblance [48].

II.5.2. Les méthodes basées sur l'apprentissage

L'apprentissage est une approche basée sur la recherche visant à extraire la structure du graphe.

L'objectif de cette approche est d'analyser les modèles présents dans les réseaux complexes. L'utilisation des modèles tels que l'apprentissage supervisé, l'apprentissage non supervisé, l'apprentissage semi-supervisé et l'apprentissage par renforcement lui permet d'extraire des informations utiles à partir des données de réseau. On utilise ces méthodes pour résoudre des problèmes comme la classification de nœuds, la prédiction de liens et la détection de communautés dans les réseaux complexes [49].

II.5.2.1. Les méthodes d'apprentissage supervisées

Un modèle supervisé est une technique d'apprentissage automatique qui consiste à entraîner un modèle en utilisant un ensemble de données étiquette où chaque exemple comprend à la fois des caractéristiques d'entrée (extraites des nœuds ou des arêtes du réseau) et des caractéristiques sorties (des étiquettes ou des valeurs associées à ces nœuds ou arêtes). La principale tâche de cette méthode est d'acquérir des compétences pour prédire avec précision les étiquettes pour de nouvelles données, En se basant sur les exemples d'entraînement existants [50].

On peut classer les modèles supervisés en deux catégories : la classification cherche à prédire des étiquettes discrètes, et la régression prédit des valeurs continues.

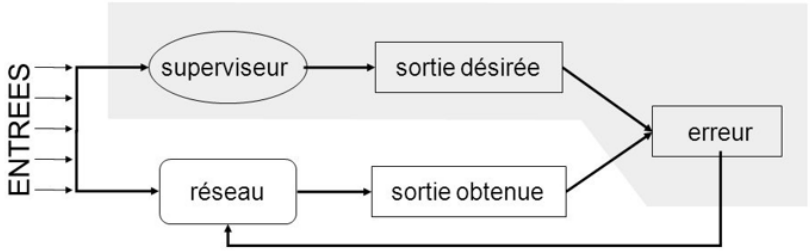


Figure II.3. Schéma d'un modèle supervisé [51]

Les méthodes d'apprentissage supervisé utilisées pour prédire les liens sont :

- **Support Vector Machine (SVM) :** est un algorithme d'apprentissage supervisé qui cherche le meilleur équilibre entre deux catégories de données, Grâce à son efficacité, il est adapté aux ensembles de données de petite et moyenne taille, ce qui le rend utilisable dans des applications en temps réel) [52].
- **Matrix Factorization (MF) :** est une méthode de diminution de taille qui sépare une matrice dense en deux matrices de petite taille, en saisissant la structure sous-jacente et

les connexions dans les données, il détecte les liens inexistants ou potentiels entre les nœuds d'un réseau [53] **.**

II.5.2.2. Les méthodes d'apprentissage non supervisées

C'est une méthode d'apprentissage utilisée dans les réseaux complexes, où le modèle est formé en se basant sur des données non étiquetées ou classées, contrairement à l'apprentissage supervisé Il n'y a pas de sorties attendues pour les données, nous avons seulement des données entrantes. Le principal objectif de cette méthode est d'analyser les données non étiquetées d'entraînement afin de déduire une fonction pour représenter un sous-jacent en se basant sur ces données [54].

On peut aussi classifier l'apprentissage non supervisé en deux grandes catégories : le regroupement (Clustering) et la réduction de la dimensionnalité (dimensionality reduction).

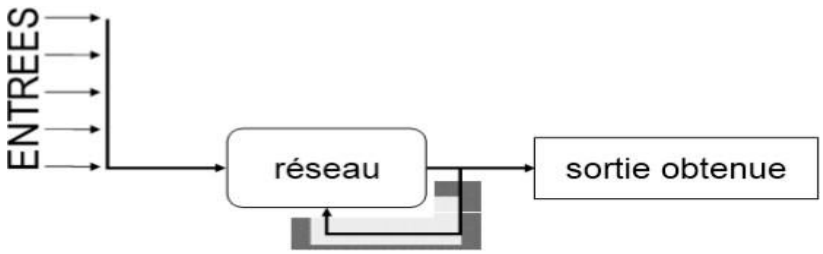


Figure II.4. Apprentissage non supervisé [51]

Les méthodes de prédiction de lien d'apprentissage non supervisé sont :

- **Deep Neural Networks for Graph Representation (DNGR)** : est une technique basée sur des réseaux de neurones profonds pour acquérir des représentations vectorielles des nœuds dans un graphe, il peut détecter des liens et des structures complexes dans le graphe. [55]**.**
- **Structural Deep Network Embedding (SDNE)** : est une méthode d'apprentissage non supervisé utilisée pour apprendre des représentations vectorielles de nœuds dans un graphe, Ces représentations, connues sous le nom **embeddings,** il se montre plus performant contre le bruit et les nœuds isolés dans le graphe par rapport à d'autres techniques d'embeddings. [56]**.**

II.5.2.3. Les méthodes d'apprentissage semi-supervisé

La méthode semi-supervisée est une approche qui intègre à la fois l'apprentissage supervisé et non supervisé. Cela implique l'utilisation simultanée de données étiquetées et de données non étiquetées [57].

L'objectif de cette méthode est d'utiliser les données non étiquetées pour améliorer les performances des modèles d'apprentissage. Ainsi, l'utilisation de données non étiquetées, qui sont en général plus abondantes que les données étiquetées, est rendue efficace par la méthode semi-supervisée [58].

Parmi les méthodes d'apprentissage semi-supervisé pour la prédiction de liens, on peut citer :

- **Graph Attention Network (GAT) :** est un type de réseau neuronal qui permet de modéliser des données structurées sous la forme de graphes, il peut servir à évaluer la présence ou l'absence d'un lien entre deux nœuds du graphe [59]**.**
- **Graph Neural Network (GNN) :** est un modèle de réseau neuronal conçus pour traiter des données graphiques, ils utilisent des données non étiquetées pour transmettre l'information à travers les nœuds et les arêtes du graphe [60].

II.5.2.4. L’apprentissage par renforcement

Est une méthode d'apprentissage qui nous permet de développé des algorithmes qui amélioré la capacité de prendre des décisions et d'interagir avec l'environnement basé sur un réseau complexe. La méthode est constituée d'un ensemble prédéterminé de données à utiliser pour formuler une approche efficace de la prise de décision. L'objectif de l'apprentissage par renforcement est visé la meilleure stratégie d'action pour optimiser la récompense [61].

Les difficultés examinées par l'apprentissage par renforcement tournent fréquemment autour de séquences d'actions, ce qui indique qu'un état spécifique peut sembler défavorable dans la période immédiate, mais que les actions ultérieures peuvent apporter des avantages substantiels [62].

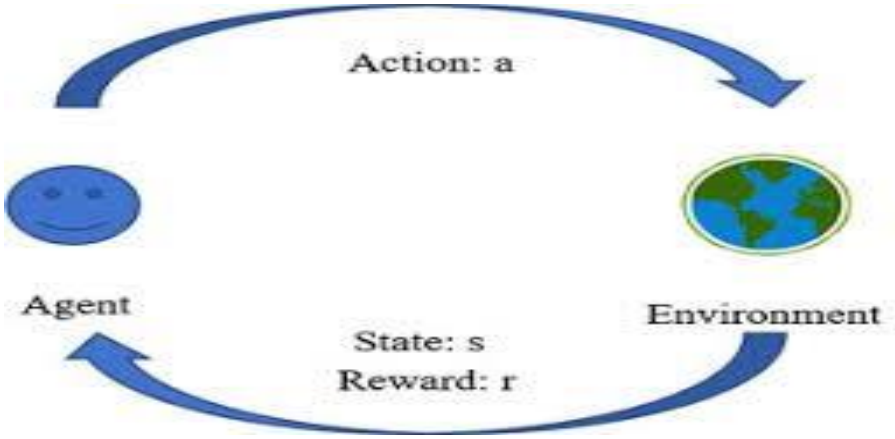


Figure II.5. Apprentissage par renforcement [63]

Les méthodes d'apprentissage par renforcement pour la prédiction de liens, incluent :

- **Graph Convolution Policy Network (GCPN)** : est un type de modèle d'apprentissage par renforcement spécialement conçu pour créer des graphes moléculaires axés sur un objectif [64].
- **Graph Transformation Policy Network (GTPN) :** est type de modèle d'apprentissage par renforcement profond qui a été spécialement créé dans le but de résoudre le problème complexe de prédiction des réactions chimiques [65].

II.6. Conclusion

Ce chapitre sur la prédiction de liens dans les réseaux complexes a mis en lumière l'importance de cette discipline dans divers domaines tels que la biologie, la santé, la collaboration scientifique, les réseaux sociaux et les systèmes de recommandation.

Les différentes méthodes de prédiction, qu'elles soient basées sur la similarité, les modèles probabilistes ou l'apprentissage, offrent des outils puissants pour anticiper les connexions potentielles entre les entités d'un réseau.

Nous avons exploré en profondeur les différentes méthodes de similarité locales, globales et quasi-locales utilisées pour la prédiction de liens dans les réseaux complexes. Chaque approche présente ses propres avantages et limites, et leur performance peut varier en fonction des caractéristiques spécifiques du réseau étudié.

Les méthodes de similarité locales se concentrent sur les caractéristiques locales des nœuds, offrant une approche simple et rapide. Les méthodes de similarité globales prennent en

compte l'ensemble du réseau pour identifier des motifs globaux, mais peuvent être plus complexes à mettre en œuvre. Les méthodes de similarité quasi-locales combinent des aspects des deux approches pour obtenir des prédictions plus précises, bien que leur mise en œuvre puisse être plus complexe.

En résumé, comprendre et choisir judicieusement entre les méthodes de similarité locale, globale et quasi-locale est essentiel pour anticiper efficacement les liens dans les réseaux complexes.

CHAPITRE 03

EXPÉRIMENTATION

III. CHAPITRE 03 : EXPÉRIMENTATION

III.1. Introduction

Ce chapitre constitue une entré à la réalisation et à la mise en œuvre du processus de prédiction de liens dans les réseaux complexes. Nous avons étudié quatre ensembles de données représentant des réseaux issus de domaines variés. Chacun de ces ensembles de données a été utilisé pour évaluer et comparer les performances des méthodes de prédiction de liens dans des contextes réels. Pour notre processus de prédiction en utilisant un processus de validation croisée à cinq volets et des mesures de performance telles que l'AUC et l'AP. Nous terminerons ce chapitre par une présentation de l'environnement et les outils de développement utilisés pour nos expérimentations ainsi que les méthodes de prédiction de lien utilisés

III.2. Description des data sets étudiés

Les expériences menées ont été réalisées sur plusieurs ensembles de données, tous provenant de https://noesis.ikor.org/datasets/link-prediction.

Nous nous sommes concentrés sur quatre ensembles de données de prédiction de liens distincts, chacun représentant des réseaux complexes qui se trouvent dans différents domaines et peuvent être classés en plusieurs types, notamment la biologie, les interactions sociales, le transport et la collaboration. Ces ensembles de données ont été utilisés pour évaluer et comparer les performances des méthodes de prédiction de liens dans divers contextes réels.

- **Le réseau HMT** est un exemple de réseau social qui représente les relations entre acteurs, films et studios de production hollywoodiens [66].

- **Le réseau YST** fait référence au réseau protéine-protéine de la levure en bourgeonnement (Saccharomyces cerevisiae). Ce réseau représente les interactions entre les protéines de la levure en, fournissant ainsi un aperçu des relations et des connexions entre ces protéines au niveau moléculaire. L'analyse de ce réseau permet de comprendre la structure et l'organisation des interactions protéiques dans la levure en bourgeonnement [67].

- **Le réseau HTC (High-Energy Theory Collaboration)** est un réseau de collaboration scientifique qui représente les liens entre chercheurs travaillant dans le domaine de la

haute énergie. Les chercheurs sont considérés comme connectés dans ce réseau s'ils ont co-écrit des articles ensemble [68].

- **UAL** est un réseau de trafic aéroportuaire représentant les connexions aériennes entre les aéroports desservis par United Airlines aux États-Unis, utilisé pour étudier différentes facettes du trafic aérien [69].

Ce tableau présente les caractéristiques structurelles des réseaux que nous avons utilisés dans nos expériences ou **|v|** le nombre de nœuds, **|E|** le nombre d'arêtes

Tableau III.1. Synthèse des caractéristiques des réseaux expérimentaux

	Type de réseau	**|v|**	**|E|**
HMT	**Social**	**2426**	**16630**
YST	**biologique**	**2284**	**6646**
HTC	**co-auteurs**	**7610**	**15751**
UAL	**transport**	**332**	**2126**

III.3. Métriques d'évaluation de la performance

Les métriques d'évaluation sont utilisées pour mesurer la performance des méthodes de prédiction des liens dans les réseaux complexes. Elles permettent d'évaluer l'exactitude et l'efficacité de ces méthodes. Nous pouvons définir une matrice de confusion ou un tableau de contingence comparant la classe prédite avec la classe réelle. Elle est généralement organisée comme suit pour un problème de classification binaire.

Considérons un graphe réel noté $G^r(V^r,E^r)$ et un graphe prédit noté $G^p(V^p,E^p)$, les quatre scénarios possibles seraient les suivants :

- **True Positive (TP) :** Les valeurs réelles et prédites sont identiques, et où le modèle prédit correctement la présence d'un lien dans le réseau Cela signifie que La valeur réelle est positive et la valeur prédite est positive. Si $e(x,y) \in E^p$ *et* $e(x,y) \in E^r$

- **True Negative (TN) :** Dans ce cas, les valeurs réelles et prédites sont identiques, et le modèle prédit correctement l'absence de lien dans le réseau, la valeur prédite du modèle est négative, avec une valeur négative réelle. Si le lien $e(x,y) \notin E^p$ et $e(x,y) \notin E^r$

- **False Positive (FP) :** Lorsqu'il y a une discordance entre les valeurs réelles et prédites, et que le modèle prédit à tort la présence d'un lien qui n'existe pas réellement dans le réseau, la valeur réelle est négative (pas de lien), mais le modèle prédit une valeur positive (existe un lien) et la prédiction est **fausse**. Si le lien $e(x,y) \in E^p$ et $e(x,y) \notin E^r$

- **False Negative (FN) :** C'est lorsque le modèle prédit à tort l'absence d'un lien qui existe réellement dans le réseau, créant ainsi une discordance entre les valeurs réelles et prédites.La valeur prédite du modèle est négative (pas de lien) et la prédiction est fause. Cependant, la valeur réelle est positive (avec lien). Si le lien $e(x,y) \notin E^p$ *et* $e(x,y) \in E^r$

Tableau III.2. Matrice de confusion

		Classe Prédite	
		Positive (Lien)	Negative (Pas de Lien)
Classe Réelle	Positive(Lien)	**TP**	**FN**
	Negative (Pas de Lien)	**FP**	**TN**

Différentes métriques d'évaluation peuvent être utilisées pour évaluer les prédictions effectuées par n'importe quelle technique de prédiction de lien notamment :

Précision

La précision est le taux de vrais positifs par rapport au nombre total de prédictions positives faites par le modèle. Elle est calculée en utilisant la formule :

$$\textbf{Précision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \qquad \text{(III.1)}$$

AUROC (Area Under the ROC Curve)

L'AUC est une mesure numérique de la performance globale du modèle représentée par la courbe ROC. Elle varie de 0 à 1, où une valeur de 1 indique un modèle parfait et une valeur

de 0,5 correspond à un modèle aléatoire. Plus l'AUC est proche de 1, meilleure est la capacité de discrimination du modèle. L'AUC est défini comme :

$$\mathbf{AUC} = \frac{n' + 0.5n''}{(n' + n'')} \qquad \text{(III.2)}$$

n' : le nombre de paires pour lesquelles le lien manquant a été mieux classé que le lien inexistant.

n'' : correspond en fait au nombre de paires où les liens existants et non-existants ont été classés de manière équivalente lors de la prédiction.

III.4. Processus de prédiction de liens

Afin d'évaluer les performances de nos méthodes de prédiction de liens nous avons implémenté un processus de validation croisée à cinq volets. La validation croisée est une technique essentielle en apprentissage automatique pour garantir la robustesse de notre modèle. Cette technique consiste à diviser notre ensemble de données en k sous-ensembles (dans notre cas, $k = 5$), appelés "volets" ou "folds".

Avant d'appliquer la validation croisée, nous avons effectué un prétraitement sur le réseau. Cela comprend tout d'abord l'élimination des nœuds isolés, qui sont des nœuds n'ayant aucun lien avec d'autres nœuds dans le réseau.

Dans notre processus, nous avons divisé notre réseau en cinq sous-graphes de taille égale. À chaque itération de la validation croisée, l'un de ces sous-graphes est désigné comme le sous-graphe de test (G_{test}), tandis que les quatre autres sont combinés pour former le sous-graphe d'entraînement (G_{train}). Cette procédure est répétée pour chaque sous-graphe, de sorte que chaque sous-graphe est utilisé une fois comme ensemble de test. Cela garantit une évaluation complète de nos méthodes sur l'ensemble du réseau.

Après avoir séparé le réseau en sous-graphes d'entraînement et de test, nous avons appliqué une étape supplémentaire pour limiter l'analyse aux nœuds communs aux sous-réseaux d'apprentissage et de test. Cela consiste à éliminer les nœuds qui ne sont pas présents dans les deux sous-réseaux, ce qui permet de s'assurer que les données utilisées pour l'évaluation sont

cohérentes. Nous avons appliqué nos méthodes de prédiction de liens sur le sous-graphe d'entraînement pour obtenir des prédictions de liens.

Ensuite, nous avons évalué la performance de chaque méthode en comparant ses prédictions avec les véritables liens présents dans le sous-graphe de test. Les performances de chaque méthode sont évaluées à l'aide de mesures telles que l'**AUC** (aire sous la courbe ROC) et l'**AP** (précision moyenne).

Enfin, la moyenne des performances obtenues sur les cinq itérations est calculée pour chaque méthode, fournissant ainsi une évaluation globale de leur efficacité dans la prédiction des liens dans le réseau. Ce processus nous permet de tester nos méthodes de prédiction de liens de manière rigoureuse et de nous assurer qu'elles sont robustes et généralisables.

III.5. Mise en œuvre des expérimentations

III.5.1. L'environnement de développement

Nous présenterons et fournirons une description détaillée de notre mise en œuvre. Nous décrirons l'environnement de développement utilisés dans ce processus pour développer et tester les différentes méthodes de prédiction de liens.

III.5.1.1. Outils de développement

- **Visual Studio Code**

C'est un éditeur de code source populaire développé par Microsoft. Il est disponible sur plusieurs plateformes, y compris Windows, macOS et Linux. Il offre un environnement optimal et personnalisable pour la rédaction, le débogage et la gestion de notre code. Tout en étant compatible avec nombreux langages de programmation et frameworks.

- **Jupyter Notebook**

Est une application web open source qui permet aux utilisateurs de créer, partager et exécuter des documents interactifs qui contenant du code en direct, du texte, des équations. Il est largement utilisé pour l'analyse de données, les tâches de science des données et de machine learning, d'apprentissage automatique. Le Jupyter Notebook prend également en charge plusieurs langages de programmation, notamment Python, JavaScript [70].

III.5.1.2. Langages de programmation

- **Python**

Est désormais l'un des langages de programmation les plus populaires et largement utilisés de l'ère actuelle, basé sur des principes orientés objet. Il offre une syntaxe propre et simple. L'une des principales caractéristiques de Python est qu'il dispose d'une vaste collection de packages. Python favorise la lisibilité du code et une structure permettant au programmeur d'exprimer des concepts en moins de lignes de code.

La version majeure la plus récente de Python est Python 3, que nous utiliserons. L'interpréteur Python et les bibliothèques standard sont disponibles sous forme source ou binaire pour toutes les plateformes majeures et peuvent être distribués librement et gratuitement [71].

III.5.1.3. Bibliothèques et frameworks

- **NetworkX**

Est un package Python open source, qui offre des fonctionnalités pour la création, la manipulation et l'analyse de graphes complexes. Il propose une variété de structures de données, d'algorithmes et d'outils de dessin pour les graphes. Grâce à son intégration avec d'autres bibliothèques Python telles que NumPy, SciPy et Matplotlib, NetworkX facilite la manipulation et la visualisation des données. Son objectif principal est de fournir une plateforme flexible et facile à utiliser pour explorer la science des réseaux et il est largement utilisé dans divers domaines disciplinaires.

- **NumPy**

Est une bibliothèque open-source populaire pour le langage de programmation Python, principalement utilisée pour les calculs scientifiques et numériques. Elle offre des fonctionnalités avancées pour la manipulation de tableaux multidimensionnels. Cette bibliothèque est précieuse dans divers domaines tels que le traitement d'images, l'analyse de données et la simulation numérique [72].

- **Scikit-learn**

Est une bibliothèque open source d'apprentissage automatique pour Python. Elle fournit des outils pour la manipulation, le prétraitement, la sélection des caractéristiques et la construction de modèles de classification et de régression. Elle est compatible avec d'autres bibliothèques populaires telles que NumPy et Pandas. Sklearn est largement utilisé dans la communauté de l'apprentissage automatique pour son efficacité, sa facilité d'utilisation et sa flexibilité [73].

💻 **Math**

La bibliothèque **math** est un module standard de Python qui fournit un ensemble de fonctions mathématiques couramment utilisées, ce qui la rend très pratique pour effectuer des calculs mathématiques dans nos programmes Python. Elle est incluse par défaut dans l'installation de Python et ne nécessite donc aucune installation supplémentaire.

💻 **Matplotlib**

Matplotlib est une bibliothèque de visualisation en 2D pour Python largement utilisée dans le domaine scientifique et analytique. Elle offre aux développeurs la possibilité de créer facilement des graphiques de qualité professionnelle et de tracer des fonctions et d'afficher leurs courbes dans des graphiques.

III.5.2. Choix et paramétrage des méthodes

Dans notre expérimentation portant sur la prédiction de liens dans les réseaux complexes, nous avons travaillé sur dix méthodes parmi les trois catégories : locales, globales et quasi- locales. Les méthodes programmées et testées incluent **CN**, **JI**, **AA**, **RA**, **HPI**, **NSP**, **SR**, **KI**, **LPI** et **CNC**.

Tableau III.3. Les Méthodes de similarité utilisées pour notre expérimentation

Méthodes	Descriptions	Formules	Références
CN	Common Neighbors	$S_{xy}^{CN} = \lvert\Gamma(x) \cap \Gamma(y)\rvert$	[30]
JI	Jaccard index	$S_{xy}^{JI} = \frac{\lvert\Gamma(x) \cap \Gamma(y)\rvert}{\lvert\Gamma(x) \cup \Gamma(y)\rvert}$	[31]
AA	Adamic-Adar	$S_{xy}^{AA} = \sum_{Z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log k(z)}$	[32]
RA	Resource allocation	$S_{xy}^{RA} = \sum_{Z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k(z)}$	[33]

HPI	Hub Promoted Index	$S_{xy}^{HPI} = \frac{\lvert\Gamma(x) \cap \Gamma(y)\rvert}{min\{k(x), k(y)\}}$	[39]
NSP	Negated Shortest Path	$S_{xy}^{NSP} = -\lvert chemin\ le\ plus\ court_{x,y}\rvert$	[41]
SR	SimRank	$S_{xy}^{SR} = \begin{cases} 1 & if \quad x = y \\ \gamma . \frac{\sum_{a \in \Gamma(x)} \sum_{b \in \Gamma(y)} S_{ab}^{SR}}{\lvert\Gamma(x)\rvert . \lvert\Gamma(y)\rvert} & sinon \end{cases}$	[74]
KI	Katz Index	$S_{xy}^{\mathbf{KI}} = \sum_{l=1}^{\infty} \beta^{l} \cdot \lvert paths_{x,y}^{l}\rvert$	[42]
LPI	Local Path IndeX	$S_{xy}^{LPI} = (A)^2 + \beta(A)^3$	[33]
CNC	Common neighbor centrality	$S_{xy}^{CNC} = \alpha(\lvert\Gamma(x) \cap \Gamma(y)\rvert) + (1 - \alpha) . \frac{n}{d_{xy}}$	[45]

Parmi les méthodes de prédiction de liens énumérées dans notre programme, plusieurs sont paramétrables. Ces paramètres permettent d'ajuster la performance de ces méthodes en fonction des caractéristiques spécifiques du réseau ou des objectifs de prédiction.

Pour la méthode Local Path Index (**LPI**), nous avons fixé le paramètre **beta** à **0.01**. Quant à la méthode Common Neighbor Centrality (**CNC**), nous l'avons exécutée en réglant le paramètre **alpha** sur **0.5**. En ce qui concerne la méthode Katz (**KI**), nous avons testé la méthode en utilisant $\mathbf{beta = 0.001}$ et $\mathbf{l = 5}$.

III.6. Conclusion

Ce chapitre présente une introduction détaillée à la réalisation et à la mise en œuvre du processus de prédiction de liens dans les réseaux complexes. L'évaluation des méthodes de prédiction de liens a été réalisée en utilisant un processus de validation croisée à cinq volets.

Les résultats détaillés de nos expérimentations, ainsi que les tests et analyses effectués, seront présentés dans le chapitre suivant.

CHAPITRE 04

ANALYSE DES RESULTATS

IV. CHAPITRE 04 : ANALYSE DES RESULTATS

IV.1. Introduction

Dans ce chapitre, nous examinerons et évaluerons l'efficacité des dix méthodes de prédiction de liens dans les différents data sets étudiés (HMT, YST, HTC, UAL). Cette section vise à comparer les performances des méthodes en termes d'**AUC** moyen et de précision moyenne (**AP**) afin d'évaluer leur capacité à identifier correctement les liens existants et inexistants. À travers des tableaux et des figures, nous explorerons les résultats obtenus sur différents data sets.

IV.2. Analyse et évaluation des résultats

Les tableaux suivants présentent les résultats de test comparant les différentes méthodes de prédiction de liens en termes d'AUC moyen et de précision moyenne (AP). Ces mesures d'évaluation permettent d'évaluer la capacité des différentes méthodes à identifier correctement les liens existants et inexistants dans les réseaux complexes étudiés.

IV.2.1. Comparaison globale des différentes méthodes

Tableau IV.1. Les résultats d'AUC moyen de chaque méthode sur les quatre data sets

Méthode	HMT	YST	HTC	UAL
CN	0.9461	0.6987	0.8658	0.9011
JI	0.9423	0.6978	0.8662	0.8655
AA	0.9520	0.6998	0.8663	0.9171
RA	0.9543	0.6995	0.8663	**0.9295**
HPI	0.9424	0.6977	0.8661	0.8499
NSP	0.8132	0.7845	0.6816	0.7151
SR	0.9346	0.7755	0.9173	0.7235
KI	**0.9591**	**0.8238**	**0.9210**	0.8992
LPI	0.9545	0.8167	0.9084	0.8638
CNC	0.9569	0.8036	0.9182	0.9014

D’après les résultats affichés sur le **tableau 4** qui montre la comparaison des scores d’**AUC** moyen entre les méthodes sur les quatre réseaux étudiés, on remarque que la méthode

KI a des résultats d'**AUC** moyen nettement meilleurs que les autres méthodes dans les réseaux **HMT**, **YST**, **HTC**. Toutefois, dans le cas du réseau **UAL**, c'est l'**AUC** moyen de la méthode **RA** qui est le plus élevé.

D'autre part, Sur les réseaux **HTC**, **HMT** et **UAL**, la méthode **NSP** obtient des scores d'**AUC** inférieurs à la moyenne. Cependant, sur le réseau **YST**, ses performances sont comparables à celles des autres méthodes.

Par conséquent, nous pouvons conclure que **KI** présente une performance globale plus élevée parmi les dix méthodes étudiées pour la prédiction de liens.

Tableau IV.2. Le top 5 et les positions de chaque méthode en fonction de l'AUC moyen

Méthodes	Position 1	Position 2	Position 3	Position 4	Position 5	TOP 5
CN				1		1
AA		1			2	3
RA	1			1		2
NSP				1		1
SR			1		1	2
KI	3				1	4
LPI		1	1	1		3
CNC		2	2			4

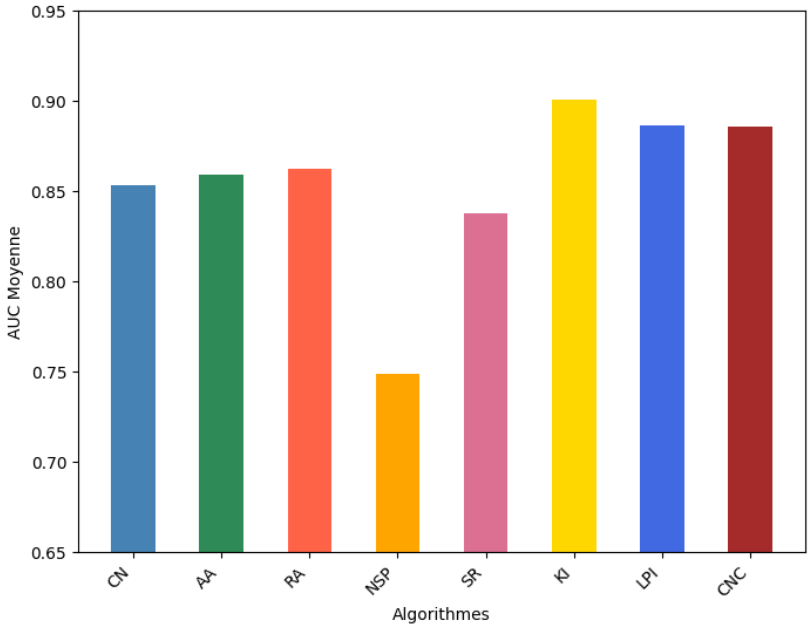


Figure III.1. AUC moyen des algorithmes

Le **tableau 5** présente les résultats des performances moyennes de différentes méthodes de prédiction. Chaque méthode est classée selon sa performance moyenne, avec les positions

allant de 1 à 5. Les positions sont déterminées en comparant l'**AUC** moyen de chaque méthode par rapport aux autres. Les cinq meilleures méthodes sont également identifiées dans le TOP 5.

Les méthodes les plus efficaces sont celles qui se classent le plus souvent dans le top cinq pour l'**AUC** et celles qui obtiennent les positions les plus élevées de manière constante.

D'après le tableau et la **figure 17** qui représente l'**AUC** moyen des algorithmes considérés sur tous les ensembles de données, **KI** se démarque comme la méthode la plus performante avec le plus grand nombre d'apparitions dans le top 5 (4 fois) et la position la plus fréquente en première position (3 fois).

CNC et **LPI** affichent également des scores **AUC** moyens élevés et figurent régulièrement dans le top 4. **CNC** se classe fréquemment parmi les trois premières positions (deuxième et troisième) et tandis que **LPI** occupe la deuxième, troisième et quatrième position.

RA, **AA** et **SR** apparaissent dans le top 5 mais moins souvent que les trois premiers. **RA** apparaissant une fois en première position. Enfin, **CN** et **NSP** sont ceux qui apparaissent le moins souvent dans le top 5.

Tableau IV.3. Les résultats Précision moyenne pour chaque méthode sur les quatre data sets

Méthode	HMT	YST	HTC	UAL
CN	0.1722	0.0435	0.1926	0.4237
JI	0.2228	0.0233	0.2235	0.1865
AA	0.2503	0.0547	**0.3392**	0.4677
RA	**0.3476**	0.0459	0.3321	**0.5311**
HPI	0.0729	0.0165	0.1307	0.0749
NSP	0.0105	0.0090	0.0014	0.0422
SR	0.0493	0.0163	0.0808	0.0379
KI	0.1815	**0.0606**	0.1977	0.4388
LPI	0.0812	0.0358	0.0862	0.1304
CNC	0.1723	0.0451	0.1946	0.4238

En se basant sur les scores de précision moyenne (**AP**) pour chaque réseau, on peut voir que la méthode **NSP** atteint la précision moyenne la plus basse par rapport aux autres approches sur les quatre réseaux étudiés, tandis que la méthode **RA** se distingue en affichant la précision moyenne la plus élevée sur les réseaux **HMT** et **UAL** c'est-à-dire une bonne prédiction de liens pour ces types de réseaux. Sur le réseau **UAL**, la méthode **AA** se donne une meilleure précision moyenne. La méthode **KI** obtient une précision moyenne plus élevée sur le réseau **HTC**.

Tableau IV.4. Le top 5 et les positions de chaque méthode en fonction de la précision moyenne

Méthodes	Position 1	Position 2	Position 3	Position 4	Position 5	TOP 5
CN					2	2
JI			2			2
AA	1	3				4
RA	2	1	1			4
KI	1		1	2		4
CNC				2	2	4

L'analyse des résultats de précision moyenne (**AP**) montre plusieurs méthodes efficaces. Du **tableau 7**, on peut déduire que les méthodes les plus efficaces en moyenne sont les méthodes **RA**, **AA**, **KI** et **CNC**. Ces méthodes sont plus courantes dans le top 5. Parmi les méthodes, la méthode **RA** se distingue par le plus grand nombre de premières positions (2 fois), ce qui la positionne comme la méthode la plus efficace et la plus cohérente. De plus, **KI** et **AA** figurent fréquemment dans le top 5 et occupent chacun une première position. Bien que **CNC** n'ait jamais atteint le sommet, il se classe régulièrement dans le top 5 (quatrième et cinquième). A l'inverse, **CN** et **JI** sont les moins performantes et apparaissant rarement dans les dans le top 5 (2 fois).

IV.2.2. Comparaison selon la classification

De l'autre côté, nous allons comparer ces méthodes selon la catégorie à laquelle chacune d'elles appartient, les méthodes de similarité locale, les méthodes de similarité globale, les méthodes de similarité Quasi-Locales. Nous avons tracé la courbe **AUC** pour chaque méthode implantée.

IV.2.2.1. Comparaison des méthodes de similarité locale

Nous comparons les cinq algorithmes de similarité locale **CN**, **JI**, **AA**, **RA** et **HPI** sur les quatre data sets (**HMT**, **YST**, **HTC**, **UAL**).

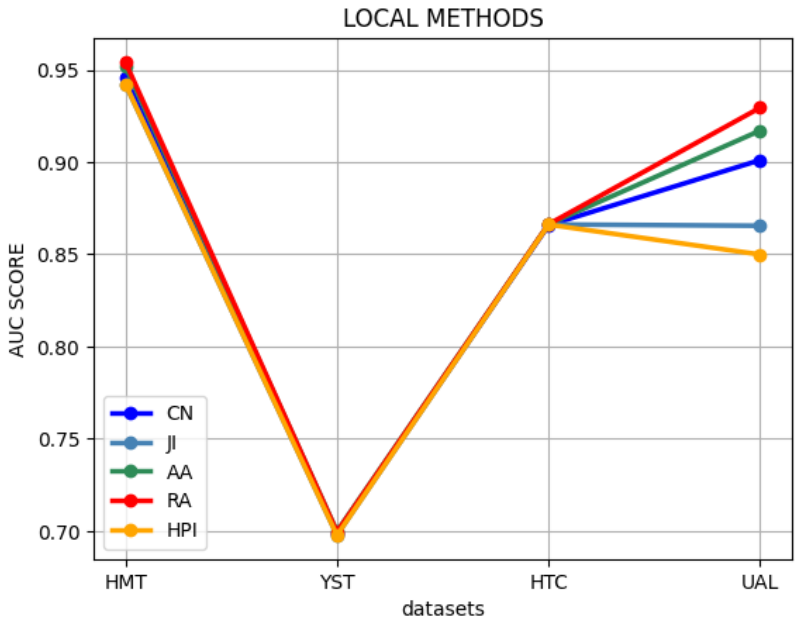


Figure III.2. Comparaison d'AUC moyenne des algorithmes locales sur chaque data sets

Sur l'ensemble de données **HMT**, tous les algorithmes se comportent bien, avec des scores supérieurs à **0,94**. **RA** obtient le score le plus élevé, indiquant qu'il est l'algorithme le plus performant pour cet ensemble de données.

En revanche, l'ensemble de données **YST** est le plus difficile pour tous les algorithmes, avec des scores AUC moyens autour de **0,7**. Cela suggère que la tâche de classification pour cet ensemble de données est difficile.

L'ensemble de données **HTC** présente des scores supérieurs à **0,86** avec tous les algorithmes obtenant. **CN**, **JI**, **AA**, **RA** et **HPI** ont des scores similaires, ce qui indique qu'ils se comportent de manière similaire sur cet ensemble de données.

Enfin, l'ensemble de données **UAL** présente une large gamme de scores, avec **RA** et **AA** obtenant les meilleurs résultats et **HPI** obtenant les moins bons résultats. Cela suggère que certains algorithmes sont mieux adaptés à cet ensemble de données que d'autres.

Les résultats d'AUC moyens des cinq méthodes sont très proches sur les data sets HMT, YST, HTC et un peu divergents sur l'UAL et la méthode **RA** (Resource Allocation) se distingue comme la meilleure méthode locale pour la prédiction de lien sur ces ensembles de données étudiés.

IV.2.2.2. Comparaison des méthodes de similarité globale

Ensuite, nous comparons les trois méthodes de similarité globale (**NSP**, **SR** et **KI**) :

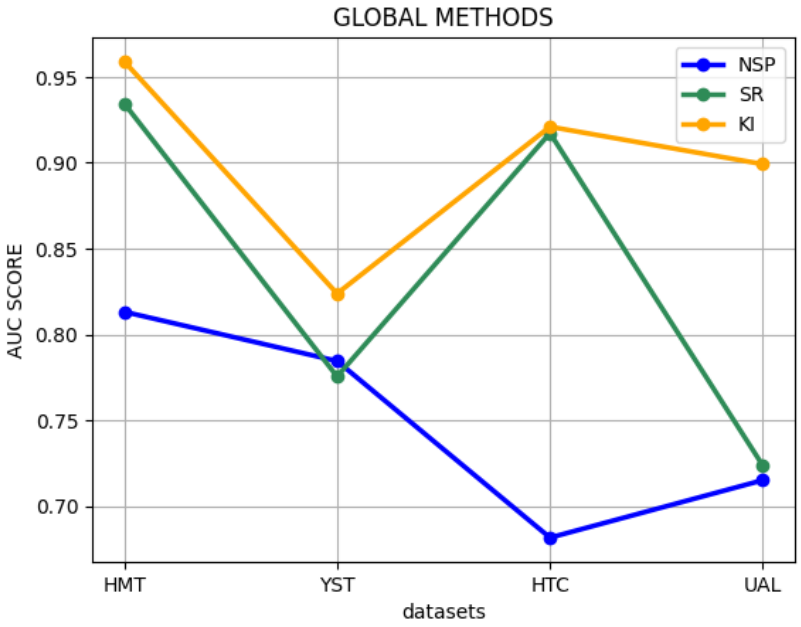


Figure III.3. Comparaison d'AUC moyenne des algorithmes globales sur chaque data sets

Sur l'ensemble de données **HMT**, **KI** a le score AUC le plus élevé supérieurs à **0,95**, indiquant qu'il est l'algorithme le plus performant pour cet ensemble de données. **SR** obtient également de bons résultats, avec un score AUC plus de **0,90**, tandis que **NSP** a le score le plus bas.

L'ensemble de données **YST** est le plus difficile pour tous les algorithmes, **KI** ayant le score le plus élevé de plus de **0,80**, suivi de **NSP** et **SR** avec des scores similaires autour de **0,78**.

L'ensemble de données **HTC** est relativement facile pour **KI** et **SR**, la méthode **KI** a obtenue de meilleurs. La méthode **SR** a obtenu des résultats moins bons que la métrique **KI**, mais elle est considérée comme tout aussi bonne. Cependant, **NSP** a du mal avec cet ensemble de données, avec un score AUC inférieurs à de **0,70**.

Enfin, l'ensemble de données **UAL** présente une large gamme de scores, **KI** obtenant les meilleurs résultats et **SR** les moins bons. **NSP** a un score modéré, indiquant qu'il fonctionne mieux sur cet ensemble de données que sur l'ensemble de données **HTC**, mais reste derrière **KI**.

Dans l'ensemble, ces résultats montrent que **KI** est l'algorithme le plus cohérent sur tous les ensembles de données, avec les scores AUC les plus élevés. **SR** performe bien sur certains ensembles de données mais a des difficultés sur d'autres, tandis que **NSP** a les scores les plus bas dans tous les ensembles de données, indiquant qu'il n'est peut-être pas le meilleur choix pour ces tâches de classification.

IV.2.2.3. Comparaison des méthodes de similarité Quasi-Locales

La figure montre les scores moyens d'AUC pour deux algorithmes de similarité quasi-locale (**LPI** et **CNC**) sur quatre data sets.

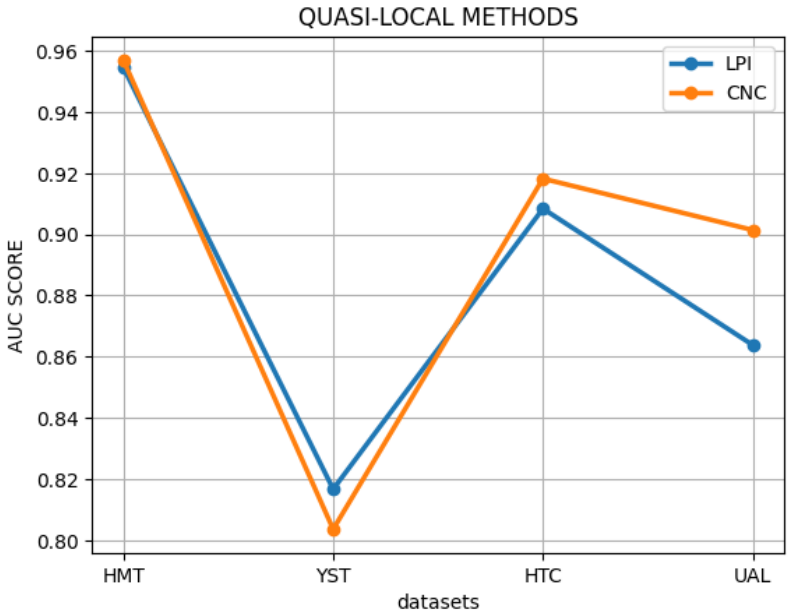


Figure III.4. Comparaison d'AUC moyenne des algorithmes quasi-locales sur chaque data sets

Sur l'ensemble de données **HMT**, les deux algorithmes, **LPI** et **CNC**, obtiennent de bons résultats avec des scores AUC supérieurs à **0,95**. **CNC** a un score légèrement plus élevé, indiquant qu'il est l'algorithme le plus performant pour cet ensemble de données.

L'ensemble de données **YST** est le plus difficile pour les deux algorithmes, avec des scores AUC autour de **0,81**. **LPI** a un score légèrement, indiquant qu'il performe légèrement mieux que **CNC** sur cet ensemble de données.

L'ensemble de données **HTC** est relativement facile pour les deux algorithmes, avec des scores AUC supérieurs à **0,90**. **CNC** a un score légèrement plus élevé, indiquant qu'il performe légèrement mieux que **LPI** sur cet ensemble de données.

Enfin, l'ensemble de données **UAL** présente une large gamme de scores, **CNC** obtenant les meilleurs résultats et **LPI** les moins bons.

En résumé, **LPI** et **CNC** performent bien sur **HMT** et **HTC**, avec **CNC** légèrement supérieur. Tous deux ont des difficultés avec **YST**. **CNC** est nettement meilleur que **LPI** sur **UAL**, indiquant qu'il est préférable pour cette tâche.

Dans la section suivante, nous allons analyser les performances globales des méthodes de prédiction de liens sur les ensembles de données étudiés.

IV.2.3. Comparaison des Performances des algorithmes sur les ensembles de Données

Nous présentons les performances globales des dix algorithmes sur les quatre data sets. Notre objectif est d'identifier quels ensembles de données sont difficiles à prédire par rapport aux autres. Pour y parvenir, nous avons calculé l'AUC moyen pour chaque algorithme sélectionné sur tous les ensembles de données. Les résultats sont résumés dans la figure 21 :

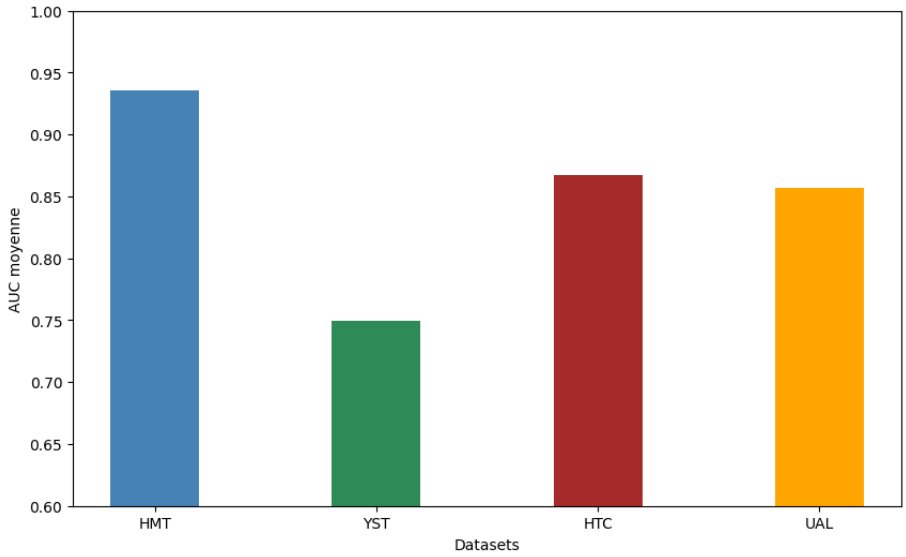


Figure III.5. AUC moyenne des dix algorithmes sur chaque ensemble de données

Il a été noté que l'AUC moyen la plus élevée (0,93) a été obtenue sur l'ensemble de données (**HMT**), tandis que l'AUC moyen la plus faible (0,74) a été observée sur l'ensemble de données (**YST**).

En comparant ces résultats, nous pouvons conclure que l'ensemble de données **YST** est le plus difficile à prédire, car les algorithmes y ont les AUC moyens les plus faibles. En

revanche, les ensembles de données **HMT**, **HTC** et **UAL** sont globalement faciles à prédire, car les algorithmes y obtiennent des AUC moyens élevés.

En somme, les résultats de comparaison des différentes méthodes montrent que les performances des algorithmes de prédictions peuvent varier considérablement en fonction de l'ensemble de données. Alors que certains algorithmes se comportent bien sur tous les ensembles de données, d'autres peuvent exceller sur des ensembles de données spécifiques. Certaines méthodes peuvent offrir des performances plus élevées, mais elles sont souvent plus complexes, ce qui indique que le choix de l'algorithme doit être adapté à la tâche et au réseau spécifique à traiter.

IV.3. Conclusion

Ce chapitre présente une analyse et une évaluation des résultats de différentes méthodes de prédiction de liens dans les réseaux complexes étudiés. Les scores mesurés par l'AUC moyen et la précision moyenne pour chaque méthode sur les quatre data sets ont été comparés afin de mettre en évidence les forces et les faiblesses de chaque méthode. Cette analyse approfondie permettra de mieux comprendre l'efficacité des algorithmes de prédiction de liens dans les réseaux complexes.

CONCLUSION GÉNÉRALE

CONCLUSION GÉNÉRALE

Ce mémoire a exploré en profondeur le domaine de la prédiction de liens dans les réseaux complexes, en se concentrant sur l'anticipation des connexions potentielles entre les entités d'un réseau. À travers l'évaluation de différentes méthodes de prédiction de liens sur des ensembles de données, nous avons pu analyser et comparer les performances de ces approches dans divers contextes.

Nous avons appris que la prédiction de liens dans les réseaux complexes est un défi complexe nécessitant des approches innovantes et des méthodes adaptées à la structure spécifique de chaque réseau. En évaluant les performances des méthodes basées sur la similarité (**CN**, **JI**, **AA**, **RA**, **HPI**, **NSP**, **SR**, **KI**, **LPI** et **CNC**).

Les principaux résultats de cette étude ont mis en lumière l'efficacité de certaines méthodes de prédiction de liens par rapport aux d'autres, en se basant sur un processus de validation croisée rigoureux et des mesures de performance telles que l'**AUC** et l'**AP**. Ces résultats ont permis de mieux comprendre les forces et les faiblesses de chaque méthode évaluée en offrant ainsi des pistes pour améliorer la prédiction de liens dans les réseaux complexes.

Nous avons également acquis des compétences essentielles en programmation, notamment en utilisant Python pour implémenter et évaluer les méthodes de prédiction de liens. Python s'est révélé être un outil puissant et polyvalent pour le traitement des données, l'analyse statistique et la mise en œuvre d'algorithmes, ce qui a enrichi notre expérience dans le domaine de la science des données.

En somme, ce mémoire a permis d'approfondir notre expertise dans le domaine de la prédiction des liens pour les réseaux complexes, en mettant en avant l'importance de ces méthodes pour anticiper les interactions au sein de ces structures complexes et en proposant des perspectives prometteuses pour de futures études dans ce domaine en constante évolution.

Nous souhaitons développer une nouvelle mesure de prédiction de liens, basée sur des méthodes existantes, afin d'apporter des contributions supplémentaires à ce domaine et de surmonter certaines des limitations identifiées dans les méthodes actuelles.

RÉFÉRENCES

[1] M. E. J. Newman, «The structure and function of complex networks,» *SIAM Review,* vol. 45, pp. 167-256, 25 3 2003.

[2] V. Levorato, «Contributions à la Modélisation des Réseaux : Prétopologie et Applications,» 2008.

[3] D. Rodighiero, «Conspirations des réseaux plats: dessiner des visualisations sur une surface sphérique continue,» *OSF.io,* 2022.

[4] D. Rodighiero, «Flat-network conspiracies-Drawing visualizations on a continuous spherical surface,» *Études digitales,* 2022.

[5] A.-L. Barabási, R. Albert et A. Vespignani, Réseaux complexes et physique statistique, P. Publishing, Éd., 2002.

[6] S. Yacine et D. Ahlem, «Découverte de communautés dans les réseaux complexes,» 2016.

[7] W. Mbarek, «Les réseaux sociaux des femmes entrepreneures tunisiennes et leurs accès aux ressources informationnelles et financières externes,» *International Journal of Scientific Research and Management,* pp. 30-47, 2024.

[8] J. Scott et P. Carrington, The SAGE Handbook of Social Network Analysis, 2021.

[9] J. L. Moreno, Who Shall Survive?, Beacon House, 1934.

[10] D. Combe, «Détection de communautés dans les réseaux d'information utilisant liens et attributs,» 2013.

[11] B. Huberman, The Laws of the Web: Patterns in the Ecology of Information, M. Press, Éd., 2001.

[12] W. Cheswick, «Internet Mapping Project: Map Gallery,» [En ligne]. Available: http://www.cheswick.com/ches/map/gallery/index.html.

[13] . A. Mashaghi, . A. Ramezanpour et V. Karimipour , «Investigation of a protein complex network,» *Eur. Phys. J. B,* vol. 41, p. 113–121, 2004.

[14] S. S. Shen-Orr, R. Milo, S. Mangan et U. Alon, «Network motifs in the transcriptional regulation network of Escherichia coli,» *Nature Genetics,* 2002.

[15] A. L. Barabási, N. Gulbahce et J. Loscalzo, «Network medicine: a network-based approach to human disease,» *Nature Reviews. Genetics*, p. 56–68, 2011.

[16] Z. Hamid , M. Summa et A. Armirotti, «Swath Label-Free Proteomics insight into the FAAH −/− Mouse Liver,» *Scientific Reports*, vol. 8, 2018.

[17] L. A. Amaral, A. Scala, M. Barthelemy et H. E. Stanley, «Classes of small-world networks,» *Proceedings of the National Academy of Sciences*, vol. 97, pp. 11149-11152, 2000.

[18] «US Airways North America Phoenix Route Map,» Airline Route Maps.com, [En ligne]. Available: https://www.airlineroutemaps.com/maps/US_Airways/North_America/Phoenix.

[19] E. Stattner, «Contributions à l'étude des réseaux sociaux : propagation, fouille, collecte de données,» 2012.

[20] V. Martínez, F. Berzal et J. C. Cubero, «A Survey of Link Prediction in Complex Networks,» *ACM Computing Surveys*, vol. 1, 2015.

[21] A. Saifi, «Fouille d'interaction dans les réseaux complexes,» 2024.

[22] J. Ben Schafer, D. Frankowski, J. Herlocker et S. Sen, «Collaborative Filtering Recommender Systems,» chez *The Adaptive Web*, vol. 4321, P. Brusilovsky, A. Kobsa et W. Nejdl, Éds., 2007.

[23] S. Leininger, T. Urich, M. Schloter et al, «Archaea predominate among ammonia-oxidizing prokaryotes in soils,» *Nature*, vol. 442, pp. 806-809, 2006.

[24] A. Godmer, Y. Kherabi et G. Pasquie, «Intelligence artificielle et autres outils digitaux : apport à la microbiologie et aux maladies infectieuses,» chez *Médecine et Maladies Infectieuses Formation*, vol. 2, 2023, pp. 117-129.

[25] I. BESTANI et N. BENADJROUDA, «Développement des Systèmes de Recommandation de Santé à l'aide des Réseaux Convolutifs Graphiques,» 2023.

[26] L. Fettah et H. Mehloul, «Application de Deep Learning pour un système de santé Intelligent dans un environnement cloud,» 2023.

[27] M. Pavlov et R. Ichise, «Finding Experts by Link Prediction in Co-authorship Networks,» 2007.

[28] H. Wang et Z. Le, «Seven-layer model in complex networks link prediction: A survey.,» 2020.

[29] L. Lü et T. Zhou, «Link Prediction in Complex Networks: A Survey,» *Physica A : Statistical Mechanics and its Applications*, vol. 390, pp. 1150-1170, 2011.

[30] D. Liben-Nowell et J. Kleinberg, «The link-prediction problem for social networks. J.,» *J. Am. Soc. Inf. Sci. Technol,* vol. 58, p. 1019–1031, 2007a.

[31] M. Pujari et R. Kanawati, « Link Prediction in Complex Networks,» pp. 58-97, 2016.

[32] L. A. Adamic et E. Adar, «Friends and neighbors on the web,» *Social networks*, vol. 25, pp. 211-230, 2003.

[33] L. Lü et T. Zhou, «Role of weak ties in link prediction of complex networks,» chez *Proceedings of the 1st ACM international workshop on Complex networks meet information & knowledge management*, 2009.

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

[35] L. Hamers, «Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula.,» *Information Processing and Management,* vol. 25, pp. 315-318, 1989.

[36] N. J. V. Eck et L. Waltman, «How to normalize cooccurrence data? An analysis of some well-known similarity measures,» *Journal of the American society for information science and technology,* vol. 60, n° %18, pp. 1635-1651, 2009.

[37] T. J. Sørensen, «A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons,» *I kommission hos E. Munksgaard,* 1948.

[38] B. McCune et J. B. Grace, «Analysis of ecological communities,» 2002.

[39] E. Ravasz, A. L. Somera, D. A. Mongru, Z. Oltvai et A. L. Barabasi, «Hierarchical organization of modularity in metabolic networks. Science,» vol. 297, p. 1551–1555, 2002.

[40] A.-L. Barabási et R. Albert , «Emergence of Scaling in Random Networks,» *Science 286*, pp. 509-512, 1999.

[41] D. Liben-Nowell, «An algorithmic approach to social networks (PhD Thesis),» 2005.

[42] L. Katz, «A new status index derived from sociometric analysis,» *Psychometrika,* vol. 18, p. 39–43, 1953.

[43] E. A. Leicht, P. Holme et M. E. Newman, «Vertex similarity in networks.,» *Physical Review,* vol. 73, 2006.

[44] W. Liu et L. Lü, «Link prediction based on local random walk,» *Europhysics Letters,* vol. 89, p. 58007, 2010.

[45] I. Ahmad, M. U. Akhtar, S. Noor et A. Shahnaz, «Missing Link Prediction using Common Neighbor and Centrality based Parameterized Algorithm,» *Sci Rep 10,* p. 1–9, 2020.

[46] B. Taskar, P. Abbeel et D. Koller, «Discriminative Probabilistic Models for Relational Data,» 2012.

[47] A. Clauset, C. Moore et M. E. J. Newman, ,, «Hierarchical structure and the prediction of missing links in networks,» 2008.

[48] P. Marsden, «Generalized Blockmodeling, P. Doreian, V. Batagelj, A. Ferligoj.,» *Cambridge University Press, New York* , p. 275–282, 2006.

[49] P. Wang et J. Hu, «Machine learning for complex networks: A survey,» 2021.

[50] L. Backstrom et J. Leskovec, «Supervised random walks: predicting and recommending links in social networks,» *In Processings of the fourth ACM international conference on Web search and data mining,* pp. 635-644, February 2011.

[51] M. ZABOUA, M. YAO, M. KOUAKOU et M. LASME, «Bulletin de veille technologique,» *ARTCI,* p. 8, 05 Décembre 2019.

[52] Steinwart, Ingo et A. Christmann, «Support vector machines,» *Wiley interdiscriplinary Reviews:Computation Statistic,* vol. 1, 12 August 2008.

[53] K. M. Aditya et E. Charles , «Link Prediction via Matrix Factorization,» vol. 6912, p. pp 437–452, September 2011.

[54] B. Equipe, *Data Scientist,* 12 avril 2022.

[55] S. Cao, W. Lu et Q. Xu, «Deep Neural Networks for Learning Graph,» vol. 30, 21 fevrier 2016.

[56] D. Wang, P. Cui et W. Zhu, «Structural Deep Network Embedding,» *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 1225–1234, 13 August 2016.

[57] O. CHAPELLE et A. ZIEN , «Semi-supervised classification by low density separation,» *In Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics (AI STATS'05).,* 2005.

[58] h. boulanger, t. Lavergne, s. Rosse et Rechercheurs en informatique, «Tri-apprentissage génératif: génération de données pour de la reconnaissance d'entitées nommées semi-supervisé.,» *hal.science*, 2023.

[59] W. Gu, F. Gao , X. Lou et J. Zhang , «Link Prediction via Graph Attention Network.,» 2019.

[60] F. Scarselli, M. Gori, C. A. Tsoi, M. Hagenbuchner et G. Monfardini, «The Graph Neural Network Model,» *IEEE Transactions on Neural Networks,* vol. 20, pp. 61 - 80, January 2009.

[61] D. D. Omar , «Exploration in Reinforcement Learning Beyond Finite State-Spaces Artificial Intelligence Université de Lille,» 2022.

[62] R. S. Sutton et A. G. Barto, «Reinforcement learning: An introduction,» *MIT press,* 2018.

[63] T. M. Shawan , B. Andreas , A. Gerd et D. Guido , «Big Data Analytics for Cyber-Physical Systems,» pp. 187-213, 2019.

[64] H. Peng, B. Du, M. Liu, M. Liu, S. Ji, S. Wang, X. Zhang et L. He, «Dynamic graph convolutional network for long-term traffic flow prediction with reinforcement learning,» *Information Sciences,* vol. 578, pp. 401-416, November 2021.

[65] K. Do, T. Tran et S. Venkatesh, «Graph Transformation Policy Network for Chemical Reaction Prediction,» *KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 750–760, 25 July 2019.

[66] J. Kunegis, «KONECT - The Koblenz Network Collection,» *Proc. Int. Conf. on World Wide Web Companion*, p. 1343–1350, 2013.

[67] D. BU, Y. Zhao, L. Cai, H. Xue, X. Zhu, H. Lu et R. Chen, «Topological structure analysis of the protein–protein interaction network in budding yeast,» *Nucleic acids research,* vol. 31, pp. 2443-2450 , 2003.

[68] M. E. Newman, «The structure of scientific collaboration networks,» *Proceedings of the national academy of sciences,* vol. 98, pp. 404-409, 2001.

[69] P. Massa, M. Salvetti et D. Tomasoni, «Bowling alone and trust decline in social network sites,» *chez 2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure,* 2009.

[70] «Qu'est-ce qu'un notebook Jupyter ?,» [En ligne]. Available: www.databricks.com/fr/glossary/jupyter-notebook. [Accès le 2 Mai 2024].

[71] «Le tutoriel Python — Documentation Python 3.12.3,» [En ligne]. Available: docs.python.org/fr/3/tutorial/. [Accès le 3 Mai 2024].

[72] «NumPy,» [En ligne]. Available: https://fr.wikipedia.org/wiki/NumPy. [Accès le 3 Mai 2024].

[73] «20 meilleures bibliothèques Python pour l'apprentissage automatique,» 6 novembre 2023. [En ligne]. Available: https://www.carmatec.com/fr/blog/20-meilleures-bibliotheques-python-pour-lapprentissage-automatique/.

[74] G. Jeh et J. Widom, «SimRank:a measure of structural-context similarity,» *Proceedings,* 2002.