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

MEMOIRE

Présenté en vue de l'obtention du Diplôme

Master en mathématique

Spécialité
Recherche opérationnelle

Thème

Prédiction de liens dans les réseaux complexes basée sur la décomposition en composantes connexes

Par
DAHMANI Mohamed
BELABAS Ferial

Soutenue publiquement le : 15 / 06 / 2023

Devant le jury composé de :

Président	**M.** MAZA Sofiane	Maître Conférence, Université de BBA
Examinateur	**Mme**. BENABID Sonia	Maître assistant, Université de BBA
Encadreur	**M.** SAIFI Abdelhamid	Maître Conférence, Université de BBA

Année : 2022/2023

Remerciement

Nous remercions tout d'abord ALLAH, d'avoir donné la force et la volonté pour réaliser ce modeste travail.

Mes vifs et sincères remerciements, accompagnés de toute ma gratitude vont à mon encadreur Mr. ***SAIFI Abdelhamid****, qui s'est toujours montré à l'écoute ainsi son précieux conseil et son aide durant toute la période du travail, pour le temps qu'il a consacré à m'apporter les outils méthodologiques indispensables à la conduite de cette recherche et dirigé dans mon travail, pour sa disponibilité et son avis éclairés.*

J'exprime mes gratitudes aux mes parent pour leur soutien et patience. Je n'oublie pas mes amis et mes proches qui m'ont toujours soutenue et encouragé.

Merci à toute et à tous

Dédicaces

À mes chers parents pour tous leurs sacrifices, leur amour, leur tendresse, leur soutien et leurs prières tout au long de mes études,

A mon épouse pour ses encouragements constants et son soutien moral,

A mes chers frères pour leur soutien et leurs encouragements,

A toute ma famille pour leur soutien tout au long de mon parcours universitaire,

Merci à tous ceux qui ont toujours été à mes côtés.

Dédicaces

الى اغلى ما املك الى امي الغالية وابي الى الاخوة الاحباء سهام وفؤاد وندى الريحان و ياسين الى احب واعز شخص لي جلولي الخير الى اعز الاصدقاء كفية فريال ياسمين هاجر خديجة خالد والى كل

من ساندني و دعمني في مشواري الدراسي.

RÉSUMÉ :

La prédiction de liens dans les réseaux complexes est un domaine de recherche important qui vise à anticiper les liens manquants ou les futures interactions entre les nœuds d'un réseau donné. Une approche prometteuse pour aborder ce défi consiste à utiliser la décomposition en composantes connexes du réseau. Cette décomposition divise le réseau en sous-ensembles appelés composantes connexes, où chaque composante représente un groupe de nœuds étroitement liés les uns aux autres.

L'utilisation de la décomposition en composantes connexes permet de réduire la complexité du problème de prédiction de liens en se concentrant sur les connexions internes à chaque composante. Cela réduit le nombre de liens à considérer lors des calculs de similarité et de prédiction, ce qui permet d'économiser du temps d'exécution et de réduire l'espace mémoire requis.

Les méthodes basées sur la similarité locale, qui exploitent les motifs de connectivité à petite échelle dans le réseau, sont souvent utilisées pour la prédiction de liens. En utilisant la décomposition en composantes connexes, ces méthodes peuvent être appliquées spécifiquement à chaque composante, améliorant ainsi la précision des prédictions.
Mots-clés : prédiction de liens, réseaux complexes, décomposition en composantes connexes, similarité locale, temps d'exécution, espace mémoire, motifs de connectivité.

ABSTRACT:

Link prediction in complex networks is an important research area that aims to anticipate missing links or future interactions between nodes in a given network. A promising approach to address this challenge is to use the connected component decomposition of the network. This decomposition divides the network into subsets called connected components, where each component represents a group of nodes that are closely related to each other.

The use of the decomposition into connected components makes it possible to reduce the complexity of the link prediction problem by focusing on the internal connections to each component. This reduces the number of links to consider during similarity and prediction calculations, saving execution time and reducing required memory space.

Local similarity-based methods, which exploit small-scale connectivity patterns in the network, are often used for link prediction. By using the decomposition into connected components, these methods can be applied specifically to each component, thus improving the accuracy of the predictions.

Keywords: link prediction, complex networks, decomposition into connected components, local similarity, running time, memory space, connectivity patterns.

الملخص:

يعد التنبؤ بالارتباط في الشبكات المعقدة مجال بحث مهمًا يهدف إلى توقع الروابط المفقودة أو التفاعلات المستقبلية بين العقد في شبكة معينة. تمثل إحدى الطرق الواعدة لمواجهة هذا التحدي في استخدام تحلل المكون المتصل بالشبكة. يقسم هذا التحلل الشبكة إلى مجموعات فرعية تسمى المكونات المتصلة، حيث يمثل كل مكون مجموعة من العقد التي ترتبط ارتباطًا وثيقًا ببعضها البعض.

يتيح استخدام التحلل إلى مكونات متصلة تقليل تعقيد مشكلة توقع الارتباط من خلال التركيز على التوصيلات الداخلية لكل مكون. يؤدي ذلك إلى تقليل عدد الارتباطات التي يجب مراعاتها أثناء حسابات التشابه والتنبؤ ، مما يوفر وقت التنفيذ ويقلل من مساحة الذاكرة المطلوبة.

غالبًا ما تُستخدم الأساليب القائمة على التشابه المحلي، والتي تستغل أنماط الاتصال صغيرة النطاق في الشبكة، للتنبؤ بالارتباط. باستخدام التحلل إلى مكونات متصلة، يمكن تطبيق هذه الطرق على وجه التحديد على كل مكون، وبالتالي تحسين دقة التنبؤات.

الكلمات المفتاحية: توقع الارتباط، الشبكات المعقدة، التحلل إلى مكونات متصلة، التشابه المحلي، وقت التشغيل، مساحة الذاكرة، أنماط الاتصال.

Table des matières

Chapitre III : Implémentation et expérimentations

Liste des figures

Liste des tableaux

INTRODUCTION GENERALE

Introduction

La prédiction de liens dans les réseaux complexes est un domaine de recherche qui vise à anticiper les liens manquants ou les futures interactions entre les nœuds d'un réseau donné. Ces réseaux, tels que les réseaux sociaux, les réseaux biologiques ou les réseaux d'infrastructure, sont caractérisés par leur nature complexe et dynamique.

L'un des défis majeurs de la prédiction de liens dans les réseaux complexes est la gestion de la complexité du calcul des liens potentiels. En effet, le nombre de liens possibles peut être extrêmement élevé, ce qui rend les méthodes de prédiction traditionnelles coûteuses en termes de temps d'exécution et d'utilisation de l'espace mémoire.

C'est dans ce contexte que la décomposition en composantes connexes émerge comme une approche prometteuse pour la prédiction de liens. La décomposition en composantes connexes consiste à partitionner le réseau en sous-ensembles appelés composantes connexes, où chaque composante représente un groupe de nœuds étroitement liés les uns aux autres.

L'idée principale derrière l'utilisation de la décomposition en composantes connexes est de réduire la complexité du problème de prédiction de liens en se concentrant sur les connexions internes à chaque composante. En limitant l'analyse aux liens à l'intérieur des composantes connexes, on peut réduire considérablement le nombre de liens à considérer, ce qui permet d'économiser du temps de calcul et de réduire l'espace mémoire requis.

De plus, la décomposition en composantes connexes offre la possibilité d'appliquer des méthodes de prédiction de liens spécifiques à chaque composante, en prenant en compte les caractéristiques locales et les motifs de connectivité propres à chaque sous-ensemble de nœuds. Cela peut améliorer la précision des prédictions de liens en exploitant les relations étroites entre les nœuds à l'intérieur des composantes connexes.

Dans cette perspective, cette étude vise à explorer l'utilisation de la décomposition en composantes connexes pour la prédiction de liens dans les réseaux complexes. Nous examinerons différentes méthodes de prédiction basées sur la similarité locale et

évaluerons leur efficacité en termes de temps d'exécution et de nombre de liens calculés, en comparant les résultats avec et sans décomposition en composantes connexes.

En utilisant cette approche, nous espérons trouver des méthodes efficaces de prédiction de liens qui permettent de réduire le temps d'exécution nécessaire tout en minimisant le nombre de liens calculés et l'espace mémoire occupé. Ces résultats contribueront à l'avancement de la prédiction de liens dans les réseaux complexes et ouvriront de nouvelles perspectives pour la modélisation et l'analyse de ces systèmes complexes.

Problématique :

Quelles sont les méthodes efficaces basées sur la similarité pour prédire les liens dans les réseaux complexes, en utilisant la décomposition des composants pour réduire le grand nombre de liens calculés et la grande quantité de mémoire occupée, tout en minimisant le long temps d'exécution ?

En répondant à cette question, nous pourrons identifier des méthodes basées sur la similarité et la décomposition des composants associés qui réduisent efficacement le temps d'exécution nécessaire, réduisent le nombre de liens calculés et occupent de l'espace mémoire.

Objectifs du projet :

L'objectif de ce projet est d'améliorer le processus de prédiction de liens dans les réseaux complexes basés sur les méthodes similarité dans le calcule de liens sur le voisinage et les chemins entre les nœuds, tout en réduisant le temps d'exécution nécessaire grâce à l'utilisation de la décomposition en composantes connexes, afin de minimiser le nombre de liens calculés et l'espace mémoire occupé.

Tel que :

$$\forall\, x \in c_i\,, \forall\, y \in c_j\,, i \neq j \begin{cases} |\Gamma x\ \cap \Gamma y| = 0 \\ paths_{x,y} = \infty \end{cases}.$$

Où :

- x, y sont des nœuds du graphe.
- c_i, c_j sont des composantes connexes du graphe.

Organisation de mémoire :

Pour atteindre les objectifs visés de notre travail, on a organisé notre mémoire en 3 chapitres :

Dans le premier chapitre : nous présentons les concepts fondamentaux liés aux réseaux complexes et la description des différentes topologies et les notions fondamentales relatives à la théorie des graphes.

Le deuxième chapitre est consacré aux définitions importantes, ainsi que les différentes méthodes basées sur la similarité utilisées dans réseaux complexes et Nous introduisons le problème de la prédiction des liens et nous décrivons les principales approches pour résoudre problème.

Dans le troisième chapitre, nous introduisons des méthodes efficaces basées sur la similarité locale en utilisant la décomposition pour réduire le nombre de liens calculés et le temps d'exécution et nous appliquons les principales approches locales (Voisins Communs (CN), L'indice Jaccard (JA), L'indice de Sørensen (SO), Le Hub Promoted Index (HPI), Le Hub Depressed Index, L'indice local de Leicht-Holme-Newman (LLHN), L'indice de Salton (SA)) pour résoudre ce problème et on clôturera notre chapitre avec d'implémentation et expérimentations.

Finalement, on termine notre mémoire avec une conclusion générale.

CHAPITRE I:
RÉSEAUX COMPLEXE

I.1 Introduction:

En théorie des graphes, un réseau complexe est un réseau possédant une architecture et une topologie complexe et irrégulière. Comme tous les réseaux, ils sont composés de nœuds (ou sommets ou points) représentant des objets, interconnectés par des liens (ou arêtes ou lignes). Ces réseaux sont des représentations abstraites des relations principalement présentes dans la vie réelle dans une grande diversité de systèmes biologiques et technologiques.

L'étude des réseaux complexes à fait l'objet d'une grande attention de la part de la communauté scientifique depuis le début des années 2000[1], et s'est montrée utile dans de nombreux domaines tels que la physique, la biologie, les télécommunications, l'informatique, la sociologie, l'épidémiologie entre autres.

I.2 Les réseaux complexe dans la vie réelle :

Les réseaux complexes sont omniprésents autour de nous et ont de nombreuses applications dans la vie courante. Pour n'en nommer que quelques-uns, nous pouvons citer le World Wide Web, Internet, les réseaux trophiques (ou chaîne alimentaire) ou encore les réseaux métaboliques. Cette grande diversité de réseaux complexes rend leur classification selon leurs propriétés communes difficiles, mais nous pouvons retenir quatre groupes principaux[2] : les réseaux sociaux, les réseaux d'informations, les réseaux technologiques, et les réseaux biologiques.

I.3 Types des Réseaux complexes :

Ces réseaux peuvent être regroupés en quatre catégories :

1. Les réseaux sociaux :

Un graphe de réseau social est un graphe permettant de représenter les interactions spécifiques entre différents groupes de personnes, représenté respectivement par les liens et les nœuds du graphe. Ces interactions peuvent être très variées, comme des liens d'amitié ou de parenté, des activités professionnelles ou personnelles communes, ou encore partager les mêmes opinions[3]. Les réseaux sociaux en ligne en sont un bon exemple, où Facebook peut être vu comme un graphe non orienté, puisque les "amitiés" sont bidirectionnelles, et Twitter quant à lui est un graphe orienté, puisque les "abonnements" sont à sens unique[4].

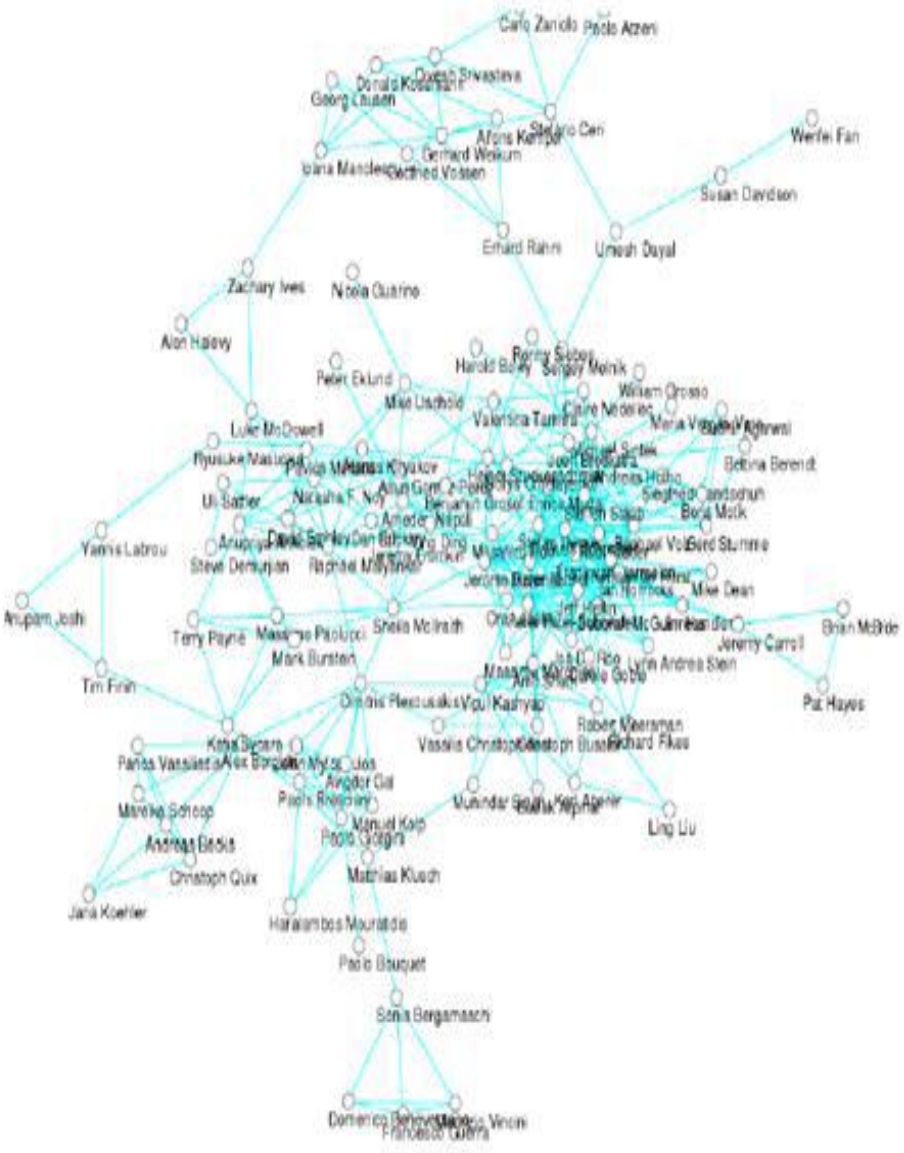


Figure I.1 : *Réseau social d'un site web en communauté.*

2. Les réseaux d'information:

Les réseaux d'information sont une autre catégorie de réseaux. Un exemple typique de ce type de réseau est le World Wide Web[5], où les nœuds correspondent aux pages web contenant de l'information, et les liens sont les hyperliens permettant de naviguer d'une page à l'autre. Ce réseau de plusieurs milliards de nœuds est un graphe dirigé, mais qui ne contient malgré tout pas de boucles fermées, puisqu'il n'y a pas de contraintes dans le classement des sites internet.

Les réseaux de citations des articles académiques sont également un bon exemple de réseau d'information [6]. Ces réseaux sont acycliques, puisque des articles ne peuvent citer que des travaux déjà publiés.

Figure I.2 : *Exemples de réseaux d'informations*

3. Les réseaux technologiques:

Nous pouvons également identifier les réseaux technologiques. Ce sont généralement des réseaux créés par l'Homme, comme les réseaux électriques[7], les réseaux de télécommunications, les réseaux aériens, les réseaux routiers[8] ou ferrés[9]. Mais, le réseau technologique le plus étudié est actuellement Internet [10], le réseau informatique mondial. Dans ce réseau, les ordinateurs et les routeurs sont les nœuds du réseau, et ces derniers sont connectés par des liens physiques comme la fibre optique modélisant les liens de ce réseau complexe.

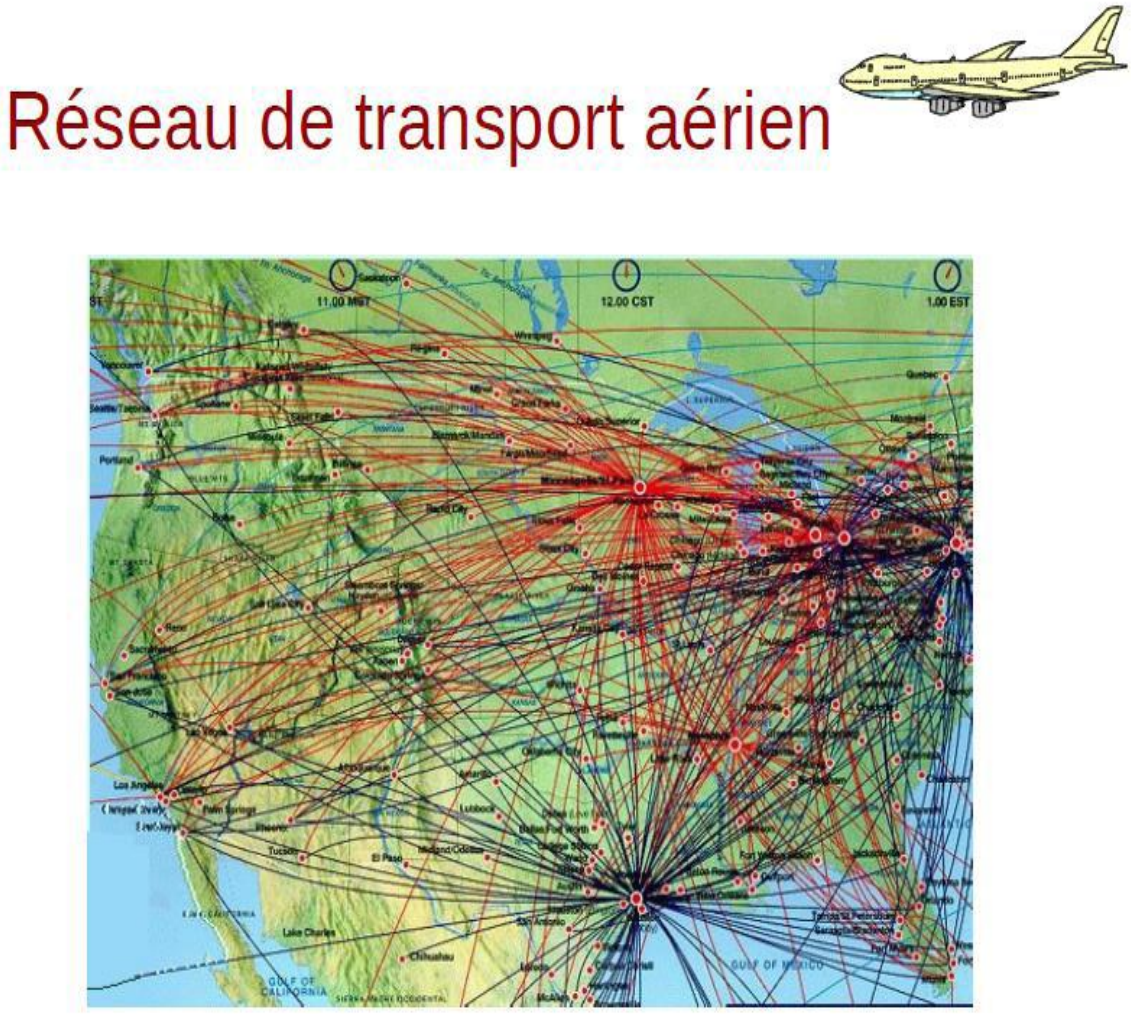


Figure I.3 : *réseau de transport aérien.*

4. Les réseaux biologiques:

Les réseaux complexes permettent également de représenter la majorité des systèmes biologiques. Ils sont de ce fait très étudiés en biologie des réseaux et en bio-informatique.

Les organismes vivants étant très complexes, le nombre de réseaux biologiques présents dans une cellule vivante est énorme. Ces réseaux complexes possèdent des fonctions spécifiques souvent indispensables au bon fonctionnement cellulaire. De plus, ces réseaux sont fortement interconnectés et fonctionnent de façon coordonnée et synchronisée avec une grande précision, puisque le moindre dysfonctionnement peut entraîner une maladie.

Parmi ces nombreux réseaux, nous pouvons citer les réseaux d'interaction protéine-protéine[11], les réseaux de régulation des gènes[12], les réseaux de

signalisation ou encore les réseaux métaboliques[13]. Les réseaux métaboliques (ou voies métaboliques) sont un exemple typique de réseau biologique. Ils représentent l'ensemble des réactions biochimiques permettant de convertir un composé en un autre dans les cellules. Dans un tel réseau, les nœuds seront les molécules biochimiques et les liens les réactions ayant permis de les obtenir.

En plus de permettre de mieux comprendre le fonctionnement cellulaire complexe, l'analyse de ces réseaux permet d'identifier plus précisément les causes de différentes maladies, et ainsi de développer de nouveaux traitements, ce qui a même mené à la création d'une nouvelle discipline : la médecine des réseaux[14].

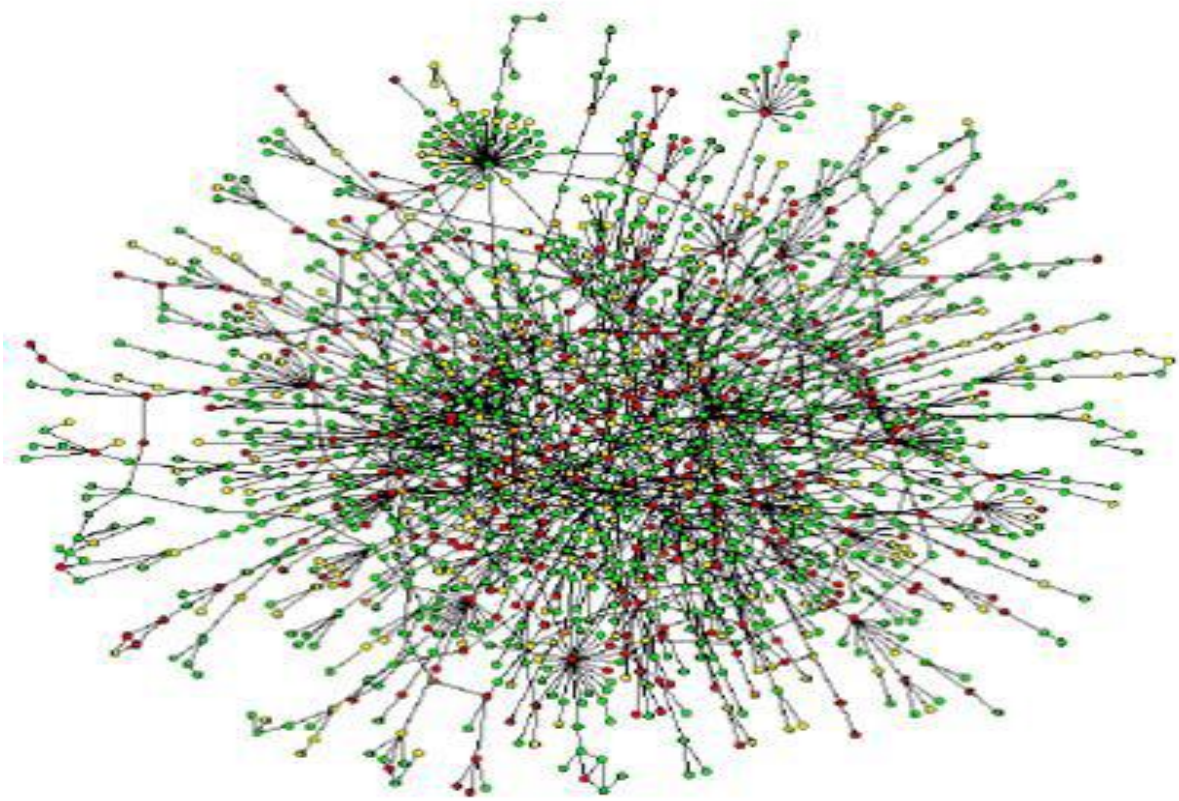


Figure I.4 : *un réseau d'interactions entre protéines.*

I.4 Théorie des graphes :

les applications de la théorie des graphes et de la recherche opérationnelle sont aujourd'hui immenses tant au plan civil que militaire : aide à la décision, stratégie, optimisation (plus court chemin, GPS, coût minimal), réseaux de transports : chemins de fer, métropolitain, lignes aériennes, électricité, gaz, oléoducs (transport de l'énergie), Internet (réseau de l'information), ports et aéroports, ordonnancement des tâches, etc.

La théorie des graphes n'est pas une branche indépendante des mathématiques, elle se rattache à la programmation linéaire, la programmation convexe (où le concept plus général de fonction convexe remplace les fonctions linéaires et affines), la topologie, le calcul des probabilités. Depuis l'année 2002, une initiation à la théorie des graphes est donnée en classe Terminale ES dans son enseignement de spécialité.

1.5. Concepts de base de la théorie des graphes

Nous allons présenter dans cette section les notions fondamentales relatives à la théorie des graphes ainsi que les différents types de ces derniers.

1.5.1. Définition d'un graphe

Un graphe est un ensemble de sommets (ou appelés nœuds) noté V (pour *Vertices*, en anglais) et d'arêtes notés E (pour *Edges*, en anglais) liant certains couples de nœuds. On écrit $G=(V,E)$ où $V=\{v1,v2,...,vn\}$ est l'ensemble de nœuds et $E=\{e1,e2,...,em\}$ est l'ensemble d'arêtes. La figure I.5 : présente un graphe d'ordre 4.

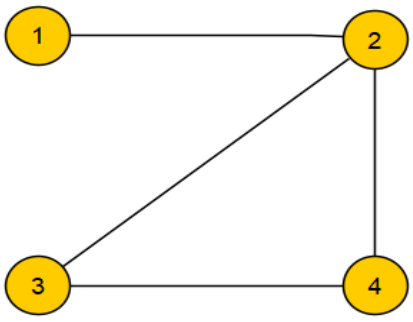


Figure I.5 : *Exemple d'un graphe d'ordre 4.*

1.5.2 Propriétés des graphes :

Les graphes, qui sont des structures composées de nœuds (ou sommets) reliés par des arêtes, peuvent avoir différentes propriétés selon leur structure et leurs caractéristiques. Voici quelques propriétés des graphes avec des formules correspondantes :

1. **Nœud (ou sommet):** Un nœud représente un élément individuel dans un graphe. Il peut être noté par un symbole ou une étiquette.
2. **Nombre de nœuds (ou sommets) (V) :** Il s'agit du nombre total de nœuds dans un graphe.
   Formule : N = |V| (où N est le cardinal de l'ensemble des nœuds).
3. **Lien (ou arête):** Une arête est une connexion entre deux nœuds dans un graphe. Dans un graphe non orienté, une arête est représentée par une paire non ordonnée de nœuds (u, v). Dans un graphe orienté, une arête est représentée par une paire ordonnée de nœuds (u, v), indiquant une direction de u vers v.
4. **Nombre d'arêtes (E) :** Il s'agit du nombre total d'arêtes dans un graphe.
   Formule : A = |E| (où A est le cardinal de l'ensemble des arêtes).
5. **Graphe:** Un graphe est défini comme une paire ordonnée G = (V, E), où V est un ensemble de nœuds (ou sommets) et E est un ensemble d'arêtes.
6. **Graphe complet:** Un graphe complet est un graphe dans lequel chaque paire de nœuds est reliée par une arête. Autrement dit, il y a une arête entre chaque paire de nœuds distincts.
   Formule : n(n-1)/2 arêtes dans un graphe complet ayant n nœuds , ce qui en fait un graphe densément connecté.
7. **Sous-graphe:** Un sous-graphe est un graphe formé à partir d'un graphe initial en supprimant certains de ses nœuds et arêtes tout en conservant les connexions entre les nœuds restants.
8. **Graphe non orienté:** Dans un graphe non orienté, les arêtes ne sont pas directionnelles. Si un graphe non orienté possède une arête entre les nœuds u et v, alors il possède également une arête entre les nœuds v et u.
9. **Connexité:** Un graphe est dit connexe s'il existe un chemin entre chaque paire de nœuds. Sinon, il est considéré comme non connexe, avec plusieurs composantes connexes distinctes.

10. **Composantes connexes:** Les composantes connexes d'un graphe sont les sous-graphes dans lesquels chaque paire de nœuds est connectée par un chemin. Les graphes non connexes ont plusieurs composantes connexes.
11. **Nombre de composantes connexes (C):** Il s'agit du nombre de sous-graphes connexes dans un graphe non connexe.

    Formule : C = |C| (où |C| est le cardinal de l'ensemble des composantes connexes).

Les formules permettent de quantifier et d'évaluer différentes caractéristiques des graphes, ce qui est utile pour l'analyse et la comparaison de différents graphes.

I.6 Conclusion

Cette représentation du problème des réseaux complexes nécessaire à une compréhension globale de ce système. Nous nous efforcerons de définir la notion de « réseau complexe» en nous appuyant dans Le deuxième chapitre sur des outils théoriques et mathématique dont nous justifierons le choix.

CHAPITRE II: PRÉDICTION DE LIENS DANS RÉSEAUX COMPLEXES

II.1 Introduction :

Un lien est une connexion entre deux nœuds dans un réseau. Ce concept simple peut être utilisé pour représenter des systèmes extrêmement complexes où un grand nombre d'éléments interagissent parmi eux. La prolifération des données pouvant être représentées sous forme de réseaux a créé de nouvelles opportunités mais aussi de nouveaux défis dans le domaine du data mining. Un grand nombre de problèmes liés au minage en réseau sont actuellement à l'étude, notamment détection de communautés [15], analyse structurelle de réseaux [16], et visualisation de réseau [17]. L'un des plus intéressants problèmes liés au réseau est la prédiction de lien, qui consiste à déduire l'existence de nouvelles relations ou d'interactions encore inconnues entre paires d'entités basées sur leurs propriétés et les liens actuellement observés [18].

Les approches et techniques conçues pour résoudre ce problème permettent d'extraire les informations implicites présentes dans le réseau et l'identification des liens parasites, comme ainsi que la modélisation et l'évaluation des mécanismes d'évolution du réseau.

Les méthodes de prédiction de liens ont été appliquées avec succès à des réseaux biologiques dans afin de prédire des interactions jusque-là inconnues entre protéines [19], réduisant significativement les coûts des approches empiriques [20]. Ils ont également été utilisés pour modéliser des systèmes hautement dynamiques, tels que le courrier électronique ou les réseaux d'appels téléphoniques [21]. Les techniques de prédiction de liens sont largement présent dans notre vie quotidienne, suggérant des personnes que nous connaissons peut-être mais que nous ne sommes pas encore connectés sur nos réseaux sociaux ou des produits que nous pourraient s'intéresser au commerce électronique [22].Les réseaux ont été largement étudiés depuis la proposition des premiers modèles de base identifier les lois qui régissent la formation des réseaux et conduisent à leurs caractéristiques structurelles [23].

Quelques techniques pouvant être considérées comme des méthodes de prédiction de lien ont alors été proposées. Cependant, ce n'est que lorsqu'un séminal spécifique axé sur la prédiction de liens travaux [24], qui a effectué une analyse complète du problème, que ce domaine est venu sous les projecteurs en raison de son applicabilité et de son utilité dans une grande variété de contextes.

La prédiction de liens est fondée sur la preuve empirique que deux entités sont plus susceptibles d'interagir s'ils sont similaires. La similarité dans les réseaux doit être comprise comme un concept abstrait et peut varier d'un réseau à l'autre. Comprendre le

domaine dans lequel réseau représente est une étape cruciale pour définir la similarité entre deux nœuds. Dans la plupart domaines, il a été observé que les nœuds ont tendance à former des communautés hautement connectées [25]. Cela a conduit à la définition commune de la similarité comme la quantité de chemins directs ou indirects pertinents entre les nœuds. L'une des principales difficultés de la prédiction de liens est d'atteindre un bon équilibre entre la quantité d'informations considérées pour effectuer la prédiction et l'algorithme complexité des techniques nécessaires pour collecter ces informations. Puisque les réseaux réels sont généralement formés de centaines de milliers voire de millions de nœuds, les techniques utilisé pour effectuer la prédiction de lien doit être très efficace. Cependant, en ne considérant que les informations locales pourraient conduire à de mauvaises prévisions, en particulier dans les réseaux très clairsemés.

Différentes revues sur ce sujet ont déjà été publiées et ont influencé ce travail [26]. Cependant, de nouvelles approches ont été développées depuis leur publication et une nouvelle revue de l'état de l'art est souhaitable. Un grand nombre de techniques de prédiction de liens spécifiques à un domaine ont été proposées. Cependant, ces techniques sont exclues de cette revue car la plupart d'entre eux sont des variations accordées de la base méthodes qui seront décrites plus loin. Dans ce travail, nous nous concentrons sur les techniques de prédiction de liens dans des réseaux non orientés utilisant des caractéristiques topologiques dérivées. Ces techniques sont plus polyvalentes que les méthodes basées sur les attributs puisque les techniques basées sur la topologie ne sont pas spécifique à un domaine.

Nous apportons plusieurs contributions à cette enquête. Dans un premier temps, nous effectuons une analyse détaillée et étude approfondie de l'état de l'art des approches et méthodes de prédiction de liens utilisant une notation unifiée. Cette étude comprend l'analyse de la complexité de calcul des techniques les plus importantes, qui ne sont pas toujours fournies par leurs auteurs originaux. Deuxièmement, nous proposons une taxonomie pour classer les techniques de prédiction de liens méthodologie qu'ils emploient et la quantité d'informations qu'ils utilisent. Enfin nous effectuer une étude empirique des techniques en appliquant les méthodes les plus importantes à un ensemble de réseaux avec des propriétés différentes et évaluer leurs résultats.

II.2 Le problème de prédiction de liens :

Le problème de prédiction de liens dans les réseaux non orientés est défini comme suit. Donné un instantané d'un réseau non orienté au temps t où chaque nœud représente une entité ou acteur et chaque lien représente une interaction ou une relation entre la paire de entités connectées par le lien, le problème de prédiction de lien peut être formellement défini comme déduire le sous-ensemble de liens manquants (liens existants mais non observés) dans le courant instantané ou qui sera formé au temps t'.

Ce peut être expliqué en utilisant un simple ensemble de données (datasets) d'un réseau de sept personnes (nœuds) avec des relations d'amitiés représentants les liens (sommets) entre les nœuds du graphe de la figure (Figure II.1).

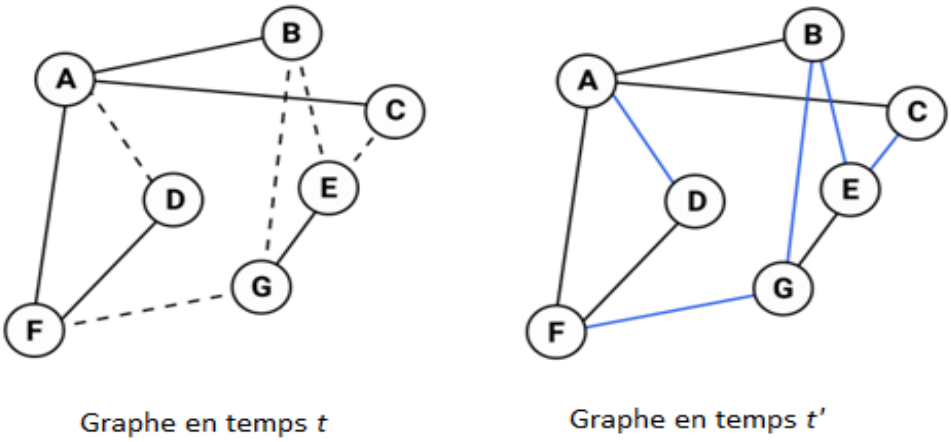


Figure II.1 : *Un exemple pour expliquer le problème de prédiction de liens*

La plupart des techniques de prédiction de liens existantes considèrent qu'il s'agit d'un problème de classement où les paires des nœuds non connectés reçoivent un score proportionnel à la probabilité d'existence d'un lien entre eux. Un seuil est généralement établi : toutes les paires avec un score supérieur le seuil sont considérées comme des instances positives et toutes les paires en dessous du seuil sont considérées comme des instances négatives. Ce seuil peut être spécifié par l'utilisateur, l'application dépendante ou déterminée automatiquement. À notre connaissance, la sélection automatique du seuil dans la prédiction de lien reste un problème inexploré. Le problème de prédiction de lien peut être considéré comme un problème de classification binaire pour les liens du réseau où deux les classes sont considérées : positif ou existence de lien et négatif ou absence de lien.

Un grand nombre de techniques de prédiction de liens ont été proposées dans le domaine spécialisé littérature. Ces techniques diffèrent par différents aspects, y compris les règles d'évolution qu'ils modélisent, leur complexité de calcul ou le type ou la quantité d'informations ils considèrent. Nous proposons une mini version personnalisée (voir Figure II.1) de la taxonomie présenté par [29].

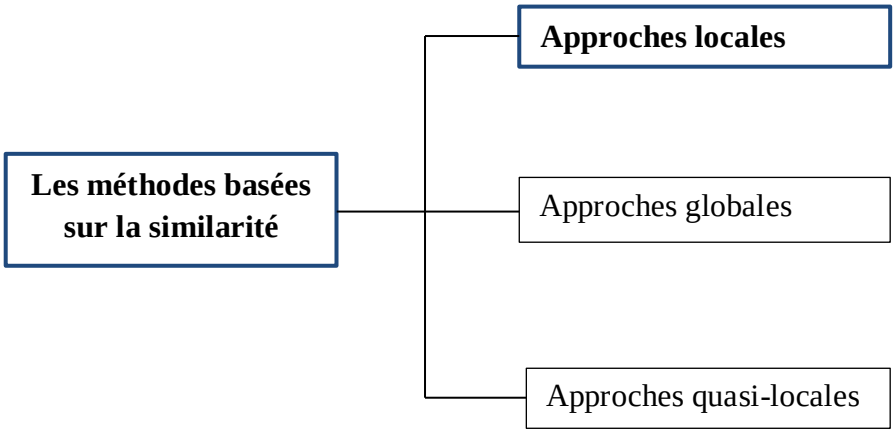


Figure II.2. Taxonomie proposée pour les techniques de prédiction de liens.

Ces taxonomies classent les méthodes en fonction de l'approche qu'ils suivent et de la quantité d'informations qui ils tiennent compte. Chaque méthode est décrite en détail ci-après.

II.2.1 Applications de la prédiction de lien :

Les techniques de prédiction de liens ont trouvé un grand nombre d'applications dans des domaines très différents des champs. Tout domaine où les entités interagissent de manière structurée peut potentiellement bénéficier de la prédiction de lien. Quelques applications intéressantes ou largement utilisées de la prédiction de liens sont décrites ultérieurement.

Les techniques de prédiction de liens sont utilisées pour améliorer la sélection d'utilisateurs similaires dans le système de recommandation des systèmes qui suivent une approche collaborative, conduisant à une meilleure recommandation résultats. Une application similaire est liée aux réseaux sociaux, qui sont devenus extrêmement populaires dans la société moderne. Les utilisateurs de ces systèmes s'attendent à disposer de mécanismes simples et efficaces pour retrouver leurs connaissances parmi les quantités

massives d'utilisateurs enregistrés. La plupart des réseaux sociaux utilisent la prédiction de liens techniques pour suggérer automatiquement des connaissances avec un haut degré de précision.

Dans le domaine de la biologie, des techniques de prédiction de liens sont appliquées pour trouver d'éventuelles interactions entre paires de protéines dans un réseau d'interaction protéine-protéine (PPI réseau). Expériences in vitro pour déterminer quelles protéines interagissent coûtent cher en argent et en temps, les cibles étudiées sont donc soigneusement sélectionnées lorsqu'il y a est une preuve antérieure, qui pourrait être obtenue par calcul.

Une autre application se trouve dans la prédiction de collaboration en co-rédaction scientifique réseaux. Les données de collaboration sont facilement accessibles, puisque certains sites d'indexation de revues rendent publiques leurs collections. Les méthodes de prédiction de liens sont devenues un outil pour mieux comprendre comment certains domaines de recherche vont évoluer en prédisant quels auteurs ou groupes pourraient potentiellement collaborer à l'avenir [30].

La résolution d'entités, également connue sous le nom de couplage d'enregistrements ou déduplication, consiste à trouver références ou enregistrements dupliqués dans un ensemble de données. Traditionnellement, résolution d'entité uniquement repose sur la similarité des attributs entre les entrées. Cependant, récemment, certains auteurs ont montré que la prise en compte des informations de contexte dans les domaines structurés en réseau à l'aide de liens techniques de prédiction pour prendre en compte la similarité entre les instances peut conduire à des améliorations dans la résolution des entités [31].

L'analyse des réseaux sociaux a été largement utilisée pour analyser la structure des réseaux terroristes pour lutter contre le crime organisé [32]. Par exemple, [33] ont montré que la topologie de certains réseaux criminels ne changer si une fraction importante de liens est réinsérée à l'aide de techniques de prédiction de liens.

Ces résultats suggèrent que la prédiction des liens peut révéler des liens réels dans les réseaux criminels, permettant d'anticiper certains actes criminels.

Enfin, les réseaux peuvent être utilisés pour analyser comment les tendances se propagent dans la société. Réseau l'analyse peut être utilisée pour améliorer les études marketing. Certains auteurs ont montré comment la prédiction de liens peut être utilisée dans le marketing viral afin d'obtenir un meilleur marketing plan [34].

II.2.2 Terminologie et notation :

Un graphe ou réseau G est un couple ordonné G = (V, E), où V est un ensemble de sommets ou nœuds étiquetés et E est un ensemble de liens entre des paires d'éléments de l'ensemble V. Un lien entre deux nœuds x et y est noté $e_{x,y}$. Le nombre de nœuds dans le réseau, également connu sous le nom de taille du réseau, est noté |V|. Le nombre de liens est noté |E|. On distingue les liens orientés (notés arcs), qui connecter un nœud source à un nœud destination, et des liens non orientés (notés comme bords), lorsqu'il n'y a pas de concept de source et de destination. Un graphe orienté est composé uniquement d'arcs. De même, un graphe non orienté est composé uniquement d'arêtes. Enfin, un graphique mixte peut contenir les deux types de liens (arcs et arêtes).

L'ensemble des nœuds connectés par une arête à un nœud x ∈ V est appelé le voisinage de x et est noté x. Dans les graphes non orientés, le degré d'un nœud x est défini comme le nombre d'arêtes connectées au nœud et sera noté |x|. Dans les graphes orientés, le degré d'un nœud est la somme du degré sortant et du degré entrant, qui sont le nombre d'arcs sortants et d'arcs entrants, respectivement. Le degré moyen d'un réseau est noté et est égal au degré moyen de tous ses nœuds.

Soit une boucle une arête ou un arc reliant un nœud à lui-même. Un graphique simple est défini sous forme de graphe sans boucles et avec au plus une arête ou un arc entre chaque paire de sommets. Les techniques examinées dans cette enquête supposent qu'il n'y a pas de boucles dans réseau.

Un chemin est une séquence de liens qui relient une séquence de nœuds dans le graphe. En dirigé réseaux, les étapes du chemin sont limitées pour se déplacer du nœud source à la destination nœud du même arc. La longueur du chemin est le nombre de liens dans le chemin Le plus court.

Plusieurs chemins les plus courts pour une paire de sommets peuvent coexister. Un graphe est dit connexe si il existe un chemin entre chaque paire de nœuds x, y ∈ V. Si le graphe n'est pas connexe, il est composé de composants. Un composant est un sous-graphe connexe.

Un graphe connexe a un seul composant. Si l'un des composants a un nombre significativement plus grand de nœuds par rapport aux autres composants, il est généralement appelé composant principal ou composant géant.

II.3. Les méthodes basées sur la similarité :

Les méthodes basées sur la similarité supposent que les nœuds ont tendance à former des liens avec d'autres nœuds. Ces méthodes partent de l'hypothèse que deux nœuds sont similaires s'ils sont connectés à des nœuds similaires ou proches dans le réseau selon une distance donnée fonction. Ces approches définissent une fonction s(x, y) qui attribue un score appelé similitude pour chaque paire de nœuds x et y. Cette mesure est calculée pour chaque intéressante paire de nœuds, généralement ceux avec des liens non observés entre eux. Les paires de nœuds sont classés par ordre décroissant en fonction de leurs scores de similarité, donc les liens en haut de la liste classement sont censés être plus susceptibles d'être présents dans l'ensemble des chaînons manquants.

La définition de la similarité n'est pas une tâche triviale, car elle a une composante heuristique. La fonction de similarité peut varier entre les réseaux même du même domaine. Comme un résultat sans surprise, un grand nombre de méthodes basées sur la similarité avec des définitions de similarité ont été proposées. Il a été démontré empiriquement que le la similarité entre les nœuds peut être définie en termes de propriétés topologiques du réseau.

En tant que contribution supplémentaire de cette enquête, la complexité algorithmique est également calculée pour chaque méthode basée sur la similarité utilisant la notation grand *O*. Trois variables ont été considéré pour mesurer la taille du problème : v comme le nombre de nœuds, e comme le nombre d'arêtes, et k comme degré maximum d'un nœud. Quelques optimisations simples seront prises en compte dans notre analyse de complexité algorithmique.

II.3.1 Approches locales :

Les approches basées sur la similarité locale utilisent des informations structurelles liées au voisinage des nœuds pour calculer la similarité de chaque nœud avec les autres nœuds du réseau. Ces Les approches sont plus rapides que les techniques non locales et hautement parallélisables. En outre, ils nous permettent de traiter efficacement le problème de prédiction de lien dans des environnements très dynamiques et réseaux changeants tels que les réseaux sociaux en ligne. Leur principal inconvénient est que l'utilisation seules les informations locales limitent l'ensemble des nœuds la similarité peut être calculée pour distance-deux nœuds (voisins de voisins). Cela peut être un gros inconvénient car de nombreux liens sont formés à des distances supérieures à deux dans de nombreux réseaux

du monde réel, en particulier dans non-petit monde. Cependant, ces méthodes ont montré une précision de prédiction très compétitive par rapport à des techniques plus complexes.

II.3.1.1 Voisins Communs (CN) :

Les voisins communs sont la technique locale la plus simple. La similarité entre deux nœuds est définie comme le nombre de voisins partagés entre les deux nœuds. Il est logique de supposer que, si deux individus partagent de nombreuses connaissances, ils sont plus susceptibles de se rencontrer que deux individus sans contacts communs. Différentes études ont confirmé cette hypothèse en observant une corrélation entre le nombre de voisins partagés entre paires de nœuds et la probabilité d'être liés. Cette méthode définit la fonction de similarité comme

$$S(x, y) = |\Gamma_x \cap \Gamma_y| \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Malgré sa simplicité, cette mesure fonctionne étonnamment bien dans la plupart des situations réelles. Réseaux et bat des approches très complexes. Cette méthode est la base d'autres approches présentées plus tard. Utilisation de cette méthode pour calculer la similarité pour toutes les paires possibles aboutit à une technique de prédiction de liaison locale.

II.3.1.2 L'indice Jaccard (JA) :

Ce coefficient largement utilisé dans les systèmes de recherche d'informations a été proposé par Paul Jaccard (1868-1944) pour comparer la similitude et la diversité de jeux d'échantillons [35]. Il mesure le ratio de voisins partagés dans l'ensemble ensemble de voisins pour deux nœuds. Cette fonction de similarité est définie comme

$$S(x, y) = \frac{|\Gamma x \cap \Gamma y|}{|\Gamma x \cup \Gamma y|} \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

On peut facilement voir que cette méthode est encore une autre variante des voisines communes méthodes où il y a une pénalisation pour chaque voisin non partagé. L'algorithmique la complexité temporelle de cette méthode.

II.3.1.3 L'indice de Sørensen (SO) :

Cet index a été développé par le botaniste Thorvald Sørensen en 1948 pour comparer la similarité entre différentes communautés écologiques échantillons de données [36]. Malgré sa similitude avec l'indice Jaccard, il est moins sensible aux valeurs aberrantes [37]. La similarité de Sørensen est définie comme

$$S(x, y) = \frac{2\,|\Gamma x \cap \Gamma y|}{|\Gamma x| + |\Gamma y|} \text{.....(3)}$$

II.3.1.4 Le Hub Promoted Index (HPI) :

Cet indice a été proposé à la suite d'une étude modularité dans les réseaux métaboliques [38]. Ces réseaux présentent une hiérarchie structure avec de petits modules hautement connectés en interne qui sont également hautement isolés les uns des autres. L'objectif principal de cette mesure de similarité est d'éviter le lien formation entre les nœuds hub et favoriser la formation de liens entre les nœuds de bas degré et hubs. Cet indice définit la similarité comme

$$S(x, y) = \frac{|\Gamma x \cap \Gamma y|}{\min(|\Gamma x|, |\Gamma y|)} \text{.....(4)}$$

II.3.1.5 Le Hub Depressed Index (HDI) :

Cet index est basé sur l'index promu hub mais a un but opposé [38]. L'index déprimé du moyeu favorise le lien formation entre hubs et entre nœuds de bas degré, mais pas entre hubs et nœuds de bas degré. Cette fonction de similarité peut être définie comme

$$S(x, y) = \frac{|\Gamma x \cap \Gamma y|}{\max(|\Gamma x|, |\Gamma y|)} \text{.....(5)}$$

II.3.1.6 L'indice local de Leicht-Holme-Newman (LLHN) :

Cet indice est défini comme le rapport de chemins réels de longueur deux entre deux nœuds et une valeur proportionnelle à la valeur attendue nombre de chemins de longueur deux entre eux [39]. Ses propres auteurs proclament que cet indice est une mesure plus sensible de l'équivalence structurelle que d'autres comme l'indice Salton ou l'indice Jaccard. La fonction de similarité définie par cet indice peut être calculée comme

$$S(x, y) = \frac{|\Gamma x \cap \Gamma y|}{|\Gamma x|\,|\Gamma y|} \text{.....(6)}$$

II.3.1.7 L'indice de Salton (SA) :

Cet indice est également connu sous le nom de similarité cosinus [39]. Cette mesure est étroitement liée à l'indice de Jaccard, et certains travaux ont montré que, dans la plupart des situations pratiques, l'indice de Salton donne une valeur qui est environ le double de l'indice de Jaccard [40]. Cette fonction de similarité est définie comme

$$S(x, y) = \frac{|\Gamma x \cap \Gamma y|}{\sqrt{|\Gamma x|\,|\Gamma y|}} \ldots\ldots\ldots\ldots\ldots\ldots (7)$$

II.3.2 Approches globales :

Les indices globaux basés sur la similarité utilisent l'ensemble des informations topologiques du réseau pour marquer chaque lien. Ces méthodes ne se limitent pas à mesurer la similarité entre la distance deux nœuds. Cependant, leur complexité de calcul peut les rendre irréalisables pour les grands réseaux et leur parallélisassions peuvent être très complexes, en particulier dans les réseaux distribués. Environnements où la topologie complète du réseau peut ne pas être connue de tous agent de calcul.

II.3.2.1. Chemin le plus court inversé (NSP).

Le chemin le plus court nié [24] est une base mesure de similarité graphique qui nécessite de calculer le chemin le plus court entre un pair de nœuds. Les chemins les plus courts peuvent être efficacement calculés avec l'algorithme de Dijkstra.

Etant donné le chemin le plus court entre un paire de nœuds x et y, leur similarité peut être calculée comme

$$s(x, y) = -|shortest\ path x, y| \ldots\ldots\ldots\ldots\ldots\ldots (8)$$

La précision de la prédiction de chemin inversée est médiocre même par rapport à la plupart des méthodes locales. D'autres méthodes décrites plus loin, basées sur plusieurs chemins, obtenir des résultats nettement meilleurs. Ce fait illustre l'importance de considérer les chemins indirects dans les techniques de prédiction de liens.

II.3.2.2. L'indice de Katz (KI).

Cet indice résume l'influence de tous les chemins possibles entre deux paires de nœuds, pénalisant progressivement les chemins par leur longueur [41]. Ce l'indice est défini comme :

$$\mathbf{s}(\mathbf{x}, \mathbf{y}) = \sum_{l=1}^{\infty} \beta^l |paths_{x,y}^l| = \sum_{l=1}^{\infty} \beta^l (A^l)_{x,y} \ldots\ldots\ldots\ldots\ldots\ldots (9)$$

Où $\boldsymbol{paths}_{x,y}^{l}$ est l'ensemble des chemins de longueur l entre les nœuds *x* et *y*, et *A* est le matrice d'adjacence du réseau. Il convient de noter que la lième puissance de la matrice a chacune de ses entrées égale au nombre de chemins de longueur l entre les paire de nœuds. Le paramètre β est un facteur d'amortissement où $0 < \beta < 1$. Donnant un plus grand La valeur de ce paramètre augmente l'influence des chemins plus longs. Si 1 est ajouté à chaque élément de la diagonale de la matrice de similarité résultante S, cette expression peut être écrit en termes matriciels comme $S = \beta AS + I$. ***……..(10)***

La similitude entre toutes les paires de nœuds peut être directement calculée en utilisant la forme fermée en réarrangeant pour S dans la précédente expression et en soustrayant les 1 précédemment ajoutés aux éléments de la diagonale :

$$\boldsymbol{S = (I - \beta A)^{-1} - I}\text{***……..(11)***}$$

Où *I* est la matrice identité. La similarité pour chaque paire de nœuds *x* et *y* est $s(x, y) = S_{x,y}$, où $S_{x,y}$ est l'élément (x, y) de la matrice *S*. L'indice de Katz a une grande puissance prédictive mais la complexité algorithmique élevée requise pour calculer l'inverse d'une matrice limite son applicabilité aux petits réseaux.

II.3.2.3. Index de noyau de forêt aléatoire (RFK).

En théorie des graphes, un arbre couvrant d'un graphe Gis défini comme un sous-graphe connexe non orienté sans cycles qui inclut tous les sommets et certaines ou toutes les arêtes de G. Le théorème de l'arbre matriciel [42] indique que le nombre d'arbres couvrants dans G est égal à tout cofacteur d'une entrée de sa représentation la placienne. Un cofacteur est le déterminant de la matrice obtenue par supprimer la ligne et la colonne d'un élément donné. Une forêt enracinée est définie comme l'union d'arbres couvrants à racines disjointes. On peut prouver que le cofacteur de $(I + L)_{x,y}$ est égal au nombre de forêts enracinées couvrantes dans lesquelles x et y sont contenus dans le même arbre couvrant à racine x. L'inverse de ce nombre peut être considéré comme une mesure d'accessibilité entre x et y. Par conséquent, une mesure de similarité peut être définie comme

$$\boldsymbol{S = (I + L)^{-1}}\text{***……..(12)***}$$

Étant donné cette matrice de similarité, la similarité entre une paire de nœuds est $s(x, y) = S_{x,y}$.

II.3.3 Approches quasi-locales :

Des méthodes quasi-locales ont récemment émergé pour trouver un équilibre entre le local et le mesures globales. Les approches quasi-locales sont presque aussi efficaces à calculer que les approches locales méthodes globales, mais prennent également en compte des informations topologiques supplémentaires, comme le font les méthodes globales.

Ils ne tiennent pas compte de la similitude entre une paire arbitraire de nœuds dans le réseau, mais ils ne sont pas non plus limités aux voisins des voisins. Certains quasi-locaux les méthodes ont accès à l'ensemble du réseau, mais leur complexité temporelle algorithmique est encore en deçà de la complexité temporelle des méthodes globales. où s dépend de paramètres spécifiques qui définissent le nombre d'itérations ou la longueur des chemins envisagés.

II.3.3.1. L'indice de chemin local (LPI).

Cet indice est fortement basé sur l'indice de Katz mais il ne considère qu'un nombre fini de longueurs de chemin [43]. La matrice de similarité peut être calculé comme

$$S = \sum_{i=2}^{l} \beta^{i-2} A^{i} \text{.....} \textbf{(13)}$$

Où $l > 2$ est la longueur maximale du trajet et β est un facteur d'amortissement. Ça devrait être noté que, lorsque $l = 2$, cela équivaudrait à la méthode des voisins communs. Similarité

pour chaque paire de nœuds x et y est défini comme $s(x, y) = S_{x,y}$.

Cette mesure est généralement utilisée avec $l = 3$ en raison de sa complexité algorithmique. Lorsque le facteur d'amortissement est réglé sur valeur faible, cette mesure obtient des résultats très similaires à l'indice de Katz mais évite le calcul d'inversion de matrice. Si la puissance de chaque matrice d'adjacence du précédent terme de sommation est réutilisée.

II.4 Conclusion :

Le chapitre II sur la prédiction de liens dans les réseaux complexes a exploré l'utilisation de la décomposition en composantes connexes comme une approche efficace pour améliorer la précision et l'efficacité des méthodes de prédiction de liens. En se concentrant sur les connexions internes à chaque composante, la décomposition permet de réduire le nombre de liens à considérer, ce qui se traduit par des gains de temps d'exécution et d'utilisation de l'espace mémoire.

L'utilisation de méthodes basées sur la similarité locale, qui exploitent les motifs de connectivité à petite échelle, a été mise en avant comme une approche pertinente pour la prédiction de liens. En appliquant ces méthodes spécifiquement à chaque composante, il est possible d'améliorer la précision des prédictions en tenant compte des caractéristiques locales et des relations étroites entre les nœuds.

En résumé, la prédiction de liens dans les réseaux complexes basée sur la décomposition en composantes connexes offre une approche prometteuse pour améliorer les performances des méthodes de prédiction. En exploitant les similarités locales et en réduisant le nombre de liens calculés, il est possible de réduire le temps d'exécution et l'espace mémoire requis tout en maintenant une bonne précision de prédiction. Cependant, des recherches supplémentaires sont nécessaires pour optimiser ces approches, prendre en compte les informations globales et étendre leur applicabilité à différents types de réseaux complexes.

CHAPITRE III
IMPLÉMENTATION ET EXPÉRIMENTATION

III.1. Introduction

Ce chapitre est essentiellement consacré aux expérimentations réalisées afin d'atteindre l'objectif de notre thème d'étude. Dans ce but, nous décrivons dans cette partie les outils exploités pour le développement du projet, tels que le choix du langage de programmation, l'environnement de programmation et le matériel utilisé. Le processus de prédiction de lien appliqué dans ce travail est exprimés et les résultats des expérimentations sont communiqués et expliqués. Nous enchainons ensuite avec une discussion des résultats obtenus. A la fin de ce chapitre, nous terminons par une conclusion.

III.2. Environnement matériel

Les expérimentations ont été effectuées sur un PC avec un processeur Intel® Core(TM) i3-2310M CPU @2.10 GHz 2.10 GHz, 4.00G de RAM, fonctionnant sous le système d'exploitation Windows 64 bits.

III.3. Environnement logiciel

Pour réaliser l'implémentation de Chapitre III, nous avons utilisé environnements de développement suivant:

III.3.1. Python-3.8 (Anaconda3)

Est un langage de programmation puissant et facile à apprendre. Il dispose de structures de données de haut niveau et permet une approche simple mais efficace de la programmation orientée objet. Parce que sa syntaxe est élégante, que son typage est dynamique et qu'il est interprété, Python est un langage idéal pour l'écriture de scripts et le développement rapide d'applications. Il est actuellement le langage le plus utilisé au monde.

L'interpréteur Python et sa vaste bibliothèque standard sont disponibles librement, sous forme de sources ou de binaires, pour toutes les plateformes majeures depuis le site Internet https://www.python.org/.

III.3.2. Spyder

Spyder est un puissant environnement scientifique écrit en Python, pour Python, et conçu par et pour des scientifiques, des ingénieurs et des analystes de données. Il présente une combinaison unique des fonctionnalités avancées d'édition, d'analyse, de débogage et de profilage d'un outil de développement complet avec l'exploration de données, l'exécution interactive, l'inspection approfondie et les belles capacités de visualisation d'un package scientifique, est un environnement de développement intégré (IDE) gratuit inclus avec Anaconda. Il comprend des fonctionnalités d'édition, de test interactif, de débogage et d'introspection.

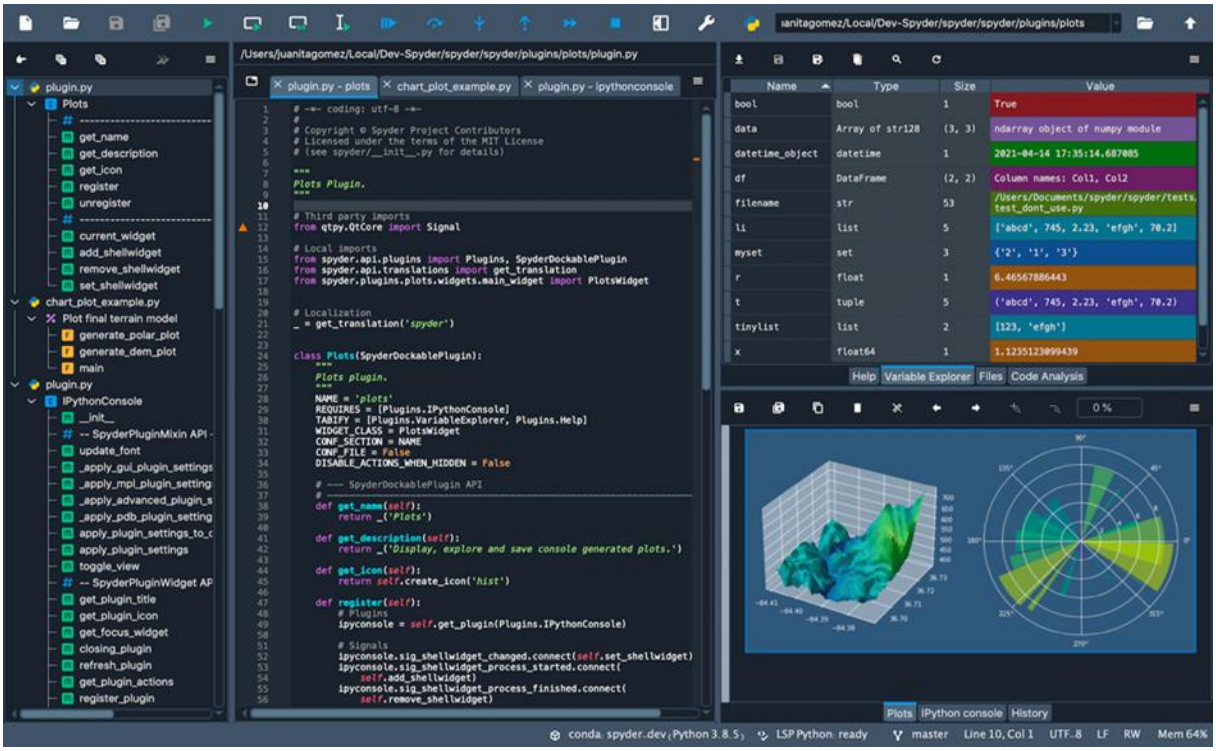


Figure III.1: *Interface Spyder*

III.4. Bibliothèques utilisées

Dans cet environnement on a installé des packages qui nous ont facilité la programmation qui sont :

- **Os**

Le module os est un module fournit par Python dont le but d'interagir avec le système d'exploitation, il permet ainsi de gérer l'arborescence des fichiers, de fournir des informations

sur le système d'exploitation processus, variables systèmes, ainsi que de nombreuses fonctionnalités des systèmes.

- **NetworkX**

NetworkX est un package Python pour la création, la manipulation et l'étude de la structure, de la dynamique et des fonctions pour les réseaux complexes.

- **Math**

Le module math est un ensemble de variables et de fonctions qui sont regroupées en un seul fichier. Les variables et les fonctions d'un module sont généralement liés entre eux d'une manière ou d'une autre. La plupart des fonctions mathématiques ne sont pas intégrées dans le noyau de Python, mais sont disponibles en chargeant le module math.

L'utilisation d'un objet de classe math peut accéder à n'importe quelle fonction math en python.

- **Time**

Ce module fournit différentes fonctions liées au temps. Pour les fonctionnalités associées, voir aussi les modules datetime et calendar.

III.5. Datasets.

Le fichier ZIP (datasets.zip) rassemble 22 réseaux provenant de différentes sources et domaines d'application. Ces réseaux ont été soigneusement sélectionnés pour couvrir un large éventail de propriétés, y compris différentes tailles, degrés moyens, coefficients de regroupement et indices d'hétérogénéité. Un résumé des propriétés structurelles des réseaux que nous avons utilisés dans nos expériences se trouve dans le tableau ci-dessous.

Nom	Nombre des nœuds	Nombre des liens	(k)	C	ASPL	D	H	r
EML	1133	5451	9.62	0,22	3.61	8	1.9421	0,0782
SMG	1024	4916	9.6	0,31	2,98	6	3.9475	-0,1925
INF	410	2765	13h49	0,46	3.63	9	1,3876	0,2258
UAL	332	2126	12.81	0,63	2,74	6	3,4639	-0,2079
NSC	1461	2742	3,75	0,69	2,59	17	1.8486	0,4616

Tableau III.5 : *Résumé des propriétés structurelles des réseaux.*

Obtenus à partir du site (https://noesis.ikor.org/datasets/link-prediction/datasets.zip) .

Nous travaillons sur six fichiers des réseaux suivant :

- SMG et NSC sont des réseaux de coauteurs pour différents domaines d'études.
- CEG est un réseau de biologique.
- EML est un réseau d'individus qui partagent des e-mails.
- INF est un réseau de contacts en face à face dans une exposition.
- UAL est un réseau de trafic aéroportuaire.

III.6 Processus de prédiction de liens :

Dans cette section, nous proposons une architecture pour le problème de prédiction de liens qui est basée sur plusieurs tapes principales, à savoir :

1- **La collecte de données :** Cette étape consiste à préparer les bases de données correspondant à un réseau complexe, en sous-ensembles d'apprentissage et de test, la base de données complète est découpée selon cas suivants

   Nous divisons aléatoirement l'ensemble des liens de la base de données en deux sous-ensembles, à savoir, 80% de liens dans un sous-graphe d'entraînement et 20% des liens dans un sous-ensemble de graphe de test.

2- **Le prétraitement :** Cette étape consiste à Limiter l'analyse aux nœuds communs aux réseaux d'apprentissage et de test et éliminer les autres nœuds.

3- **La décomposition :** Dans cette étape, nous proposons de décomposer le graphe d'apprentissage en ses composantes connexes, Pour étudier :
   - L'impact de la décomposition sur le nombre de liens calculés.
   - Impact de la décomposition sur le temps d'exécution

4- **La Prédiction :** À ce stade, une seule mesure de similarité est utilisée à la fois, pour prédire les nouveaux liens, soit pour les composantes connexes ou pour le graphe d'apprentissage complet.

Notre architecture est illustrée à la Figure III.6.

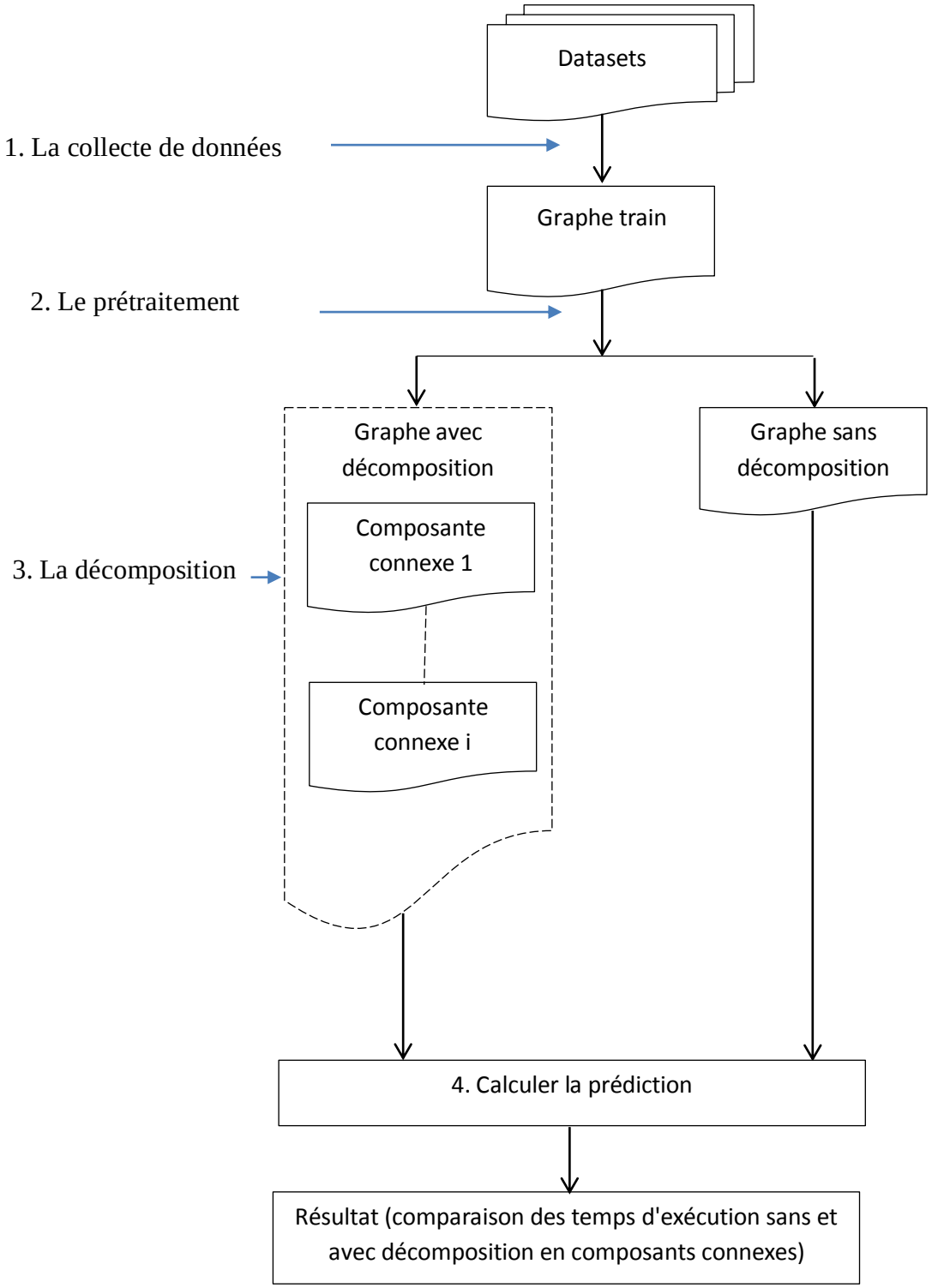


Figure III.6 : Processus de prédiction des liens.

III.7 Expérimentation

Nous avons travaillé sur le thème de la prédiction de liens dans les réseaux complexes basée sur la décomposition en composantes connexes en utiliser les méthodes HDI, HPI, SO, JA, CN, LLHN et SA des approches basées sur la similarité locale vise à explorer comment utiliser la décomposition en composantes connexes pour améliorer les prédictions de liens, tout en prenant en compte les contraintes de temps d'exécution, sur les fichiers de DataSets (SMG, CEG, EML, INF, UAL, NSC) où :

HDI: Le Hub Depressed Index

HPI: Le Hub Promoted Index

SO: L'indice de Sørensen

JA: L'indice Jaccard

(CN): Voisins Communs

LLHN: L'indice local de Leicht-Holme-Newman

SA: L'indice de Salton

III.7.1 Résultat théorique sans et avec décomposition en composants connexes.

D'après les valeurs des fichiers (bases), en calcule le gain en nombre de liens.

- **Tableau III.7.1 :** Calcule le gain en nombre de liens.

Caractéristique / **Datasets**	Nombre des nœuds	Nombre des liens	Nombre de liens calculé sans décomposition	Nombre de composantes connexes	Nombre de liens calculé avec décomposition	Le gain en nombre de liens
SMG	1024	3932	519844	35	483648	**6.96 %**
CEG	297	1718	42238	2	41942	**0.70 %**
EML	1133	4360	636918	48	583711	**8.36 %**
INF	410	2212	81633	4	80409	**1.50 %**
UAL	332	1700	53246	13	49340	**7.34 %**
NSC	1461	2193	1064337	334	63477	**94.04 %**

III.7.2 Discussion des résultats théorique sans et avec décomposition en composants connexes.

D'après le tableau, le gain en nombre de liens représente la différence entre le nombre de liens calculés sans décomposition et avec décomposition en composantes connexes. Un grand gain du nombre de liens indique que nous avons réduit le temps d'exécution (temps de calcule) requis, le nombre de liens calculés et l'espace mémoire occupé, tandis qu'un petit gain signifie que le temps d'exécution (temps de calcule) requis, le nombre de liens calculés et la mémoire occupée sont grand par rapport à avec décomposition en composantes connexes.

III.7.3 Résultat pratique sans et avec décomposition en composants connexes.

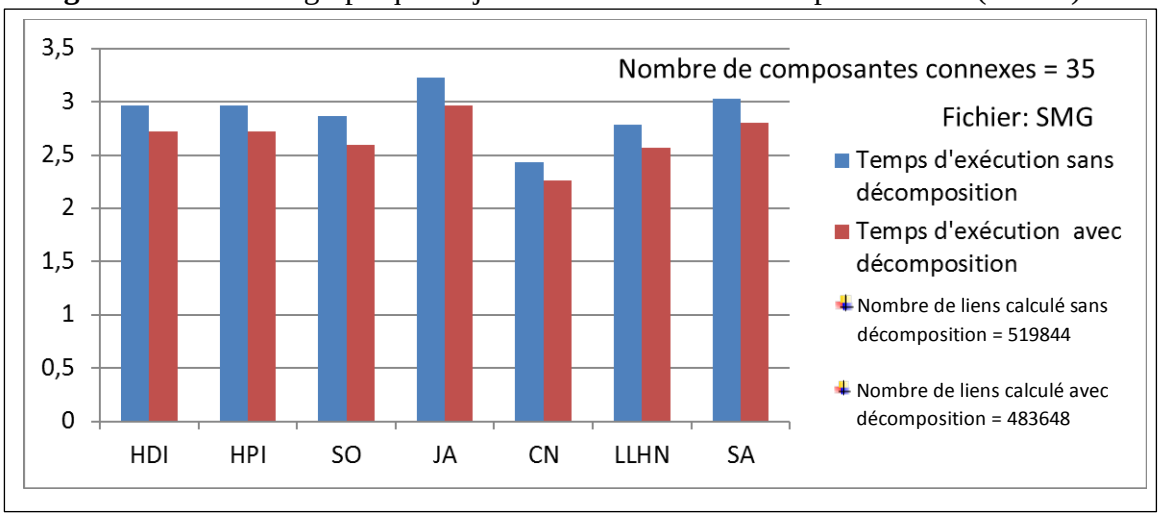
En calcule le temps d'exécutions avec et sans décomposition des fichiers.

III.7.3.1 La base (fichier) SMG :

- **Tableau III.7.3.1:** Temps d'exécutions (temps de calcules) sans et avec décomposition obtenues pour la base (fichier) SMG.

Temps d'exécution / Méthodes	Temps d'exécution sans décomposition	Temps d'exécution avec décomposition
HDI	2,968	2,726
HPI	2,964	2,720
SO	2,865	2,598
JA	3,226	2,968
CN	**2,435**	**2,258**
LLHN	2,784	2,569
SA	3,033	2,804

- **Figure III.7.3.1:** Le graphique ci-joint du tableau 1 obtenue pour la base (fichier) SMG.

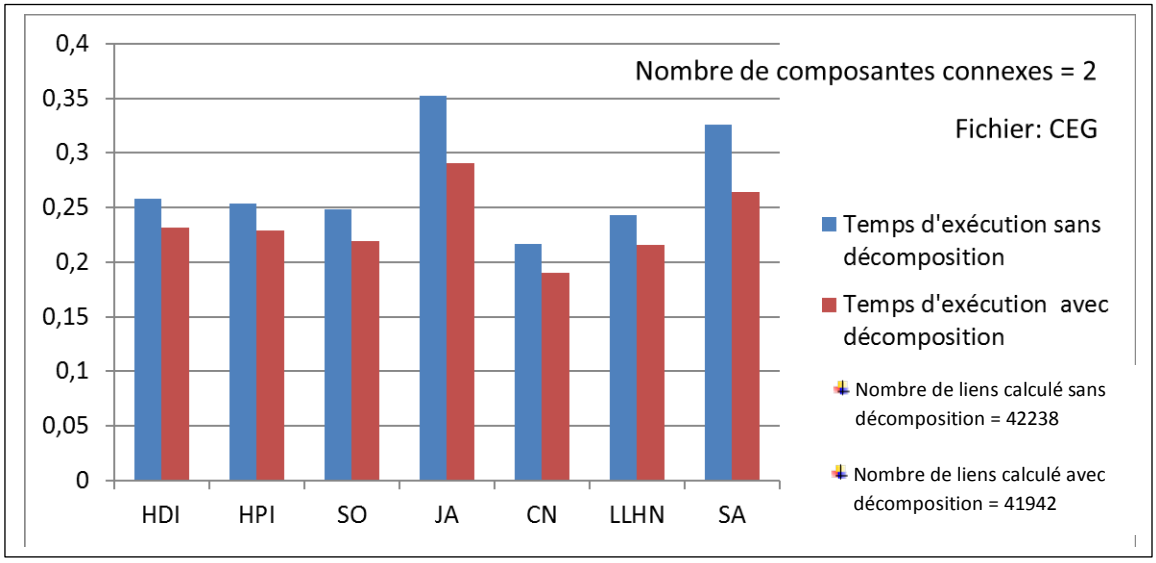


III.7.3.2 La base (fichier) CEG.

- **Tableau III.7.3.2:** Temps d'exécutions (temps de calcules) sans et avec décomposition obtenues pour la base (fichier) CEG

Temps d'exécution / Méthodes	Temps d'exécution sans décomposition	Temps d'exécution avec décomposition
HDI	0,258	0,232
HPI	0,254	0,229
SO	0,248	0,219
JA	0,352	0,291
CN	**0,217**	**0,190**
LLHN	0,243	0,216
SA	0,326	0,264

- **Figure III.7.3.2:** Le graphique ci-joint du tableau 2 obtenue pour la base (fichier) CEG.

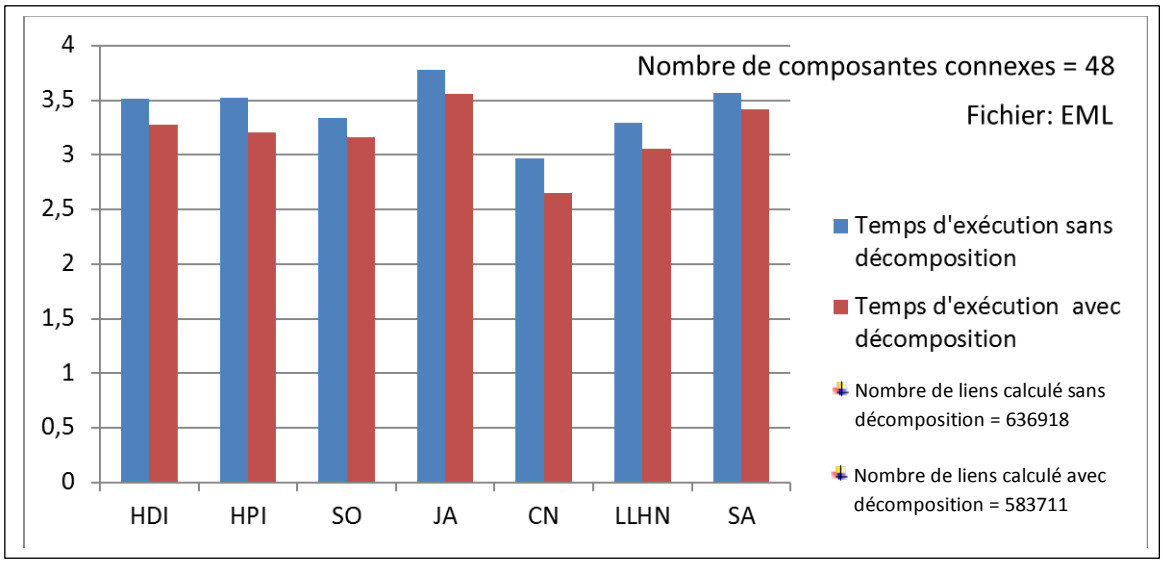


III.7.3 La base (fichier) EML

- **Tableau III.7.3.3:** Temps d'exécutions (temps de calcules) sans et avec décomposition obtenues pour la base (fichier) EML

Temps d'exécution / Méthode	Temps d'exécution sans décomposition	Temps d'exécution avec décomposition
HDI	3,512	3,275
HPI	3,520	3,206
SO	3,337	3,163
JA	3,780	3,558
CN	**2,972**	**2,655**
LLHN	3,296	3,059
SA	3,565	3,418

- **Figure III.7.3.3:** Le graphique ci-joint du tableau 3 obtenue pour la base (fichier) EML.

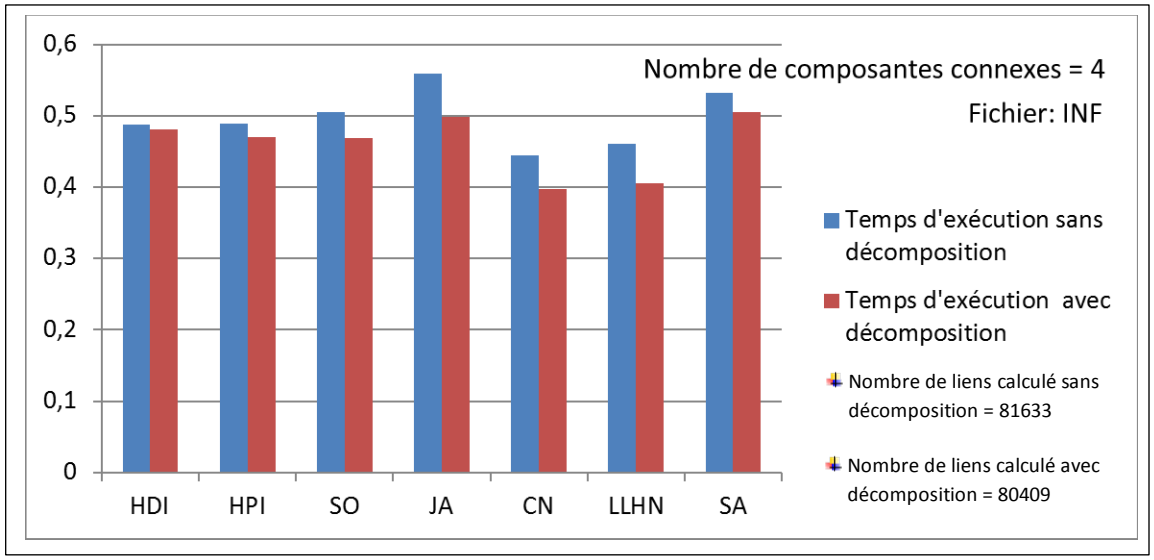


III.7.3.4 La base (fichier) INF

- **Tableau III.7.3.4:** Temps d'exécutions (temps de calcules) sans et avec décomposition obtenues pour la base (fichier) INF

Temps d'exécution / Méthodes	Temps d'exécution sans décomposition	Temps d'exécution avec décomposition
HDI	0,488	0,481
HPI	0,489	0,470
SO	0,505	0,469
JA	0,559	0,498
CN	**0,445**	**0,398**
LLHN	0,461	0,406
SA	0,532	0,506

- **Figure III.7.3.4:** Le graphique ci-joint du tableau 4 obtenue pour la base (fichier) INF.

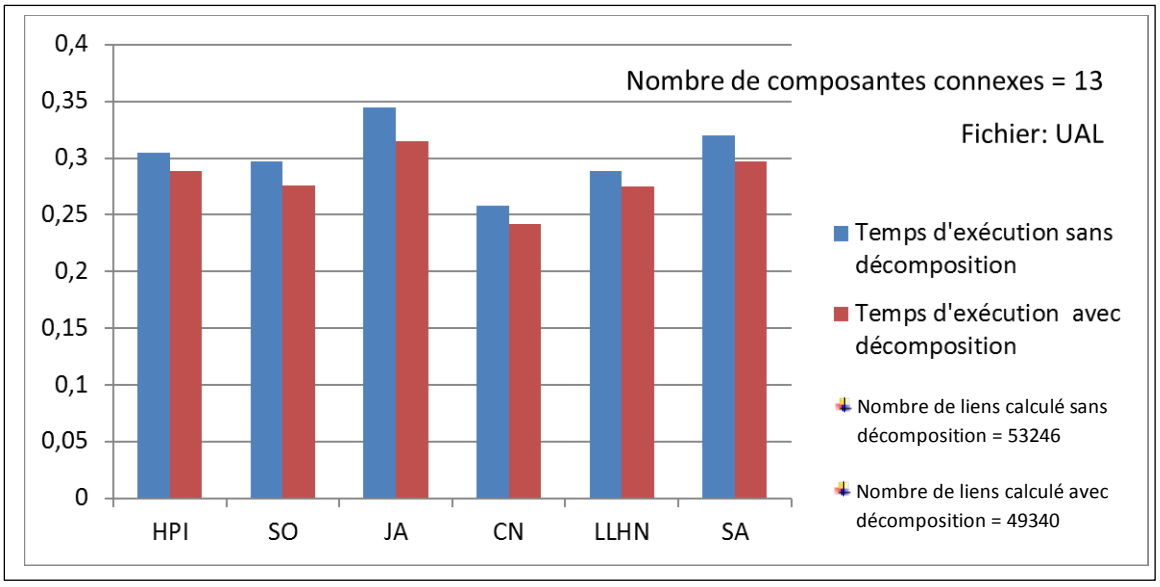


III.7.3 La base (fichier) UAL

- **Tableau III.7.3.5:** Temps d'exécutions (temps de calcules) sans et avec décomposition obtenues pour la base (fichier) UAL

Temps d'exécution / Méthodes	Temps d'exécution sans décomposition	Temps d'exécution avec décomposition
HDI	0,306	0 ,285
HPI	0,305	0,289
SO	0,297	0,276
JA	0,345	0,315
CN	**0,258**	**0,242**
LLHN	0,289	0,275
SA	0,320	0,297

- **Figure III.7.3.5:** Le graphique ci-joint du tableau 5 obtenue pour la base (fichier) UAL.

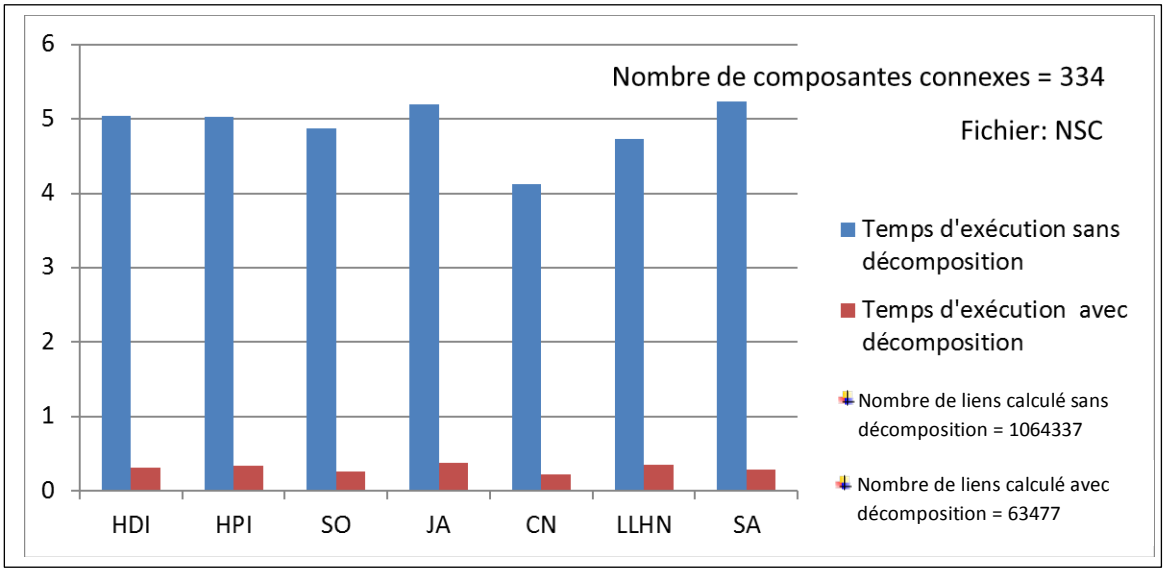


III.7.3.6 La base (fichier) NSC

- **Tableau III.7.3.6:** Temps d'exécutions (temps de calcules) sans et avec décomposition obtenues pour la base (fichier) NSC.

Temps d'exécution / Méthodes	Temps d'exécution sans décomposition	Temps d'exécution avec décomposition
HDI	5,042	0,312
HPI	5,035	0,338
SO	4,876	0,266
JA	5,202	0,383
CN	**4,127**	**0,224**
LLHN	4,732	0,350
SA	5,232	0,292

- **Figure III.7.3.6:** Le graphique ci-joint du tableau 6 obtenue pour la base (fichier) NSC.



III.7.4 Discussion des résultats pratique sans et avec décomposition en composants connexes.

Lorsqu'il s'agit de comparer le temps d'exécution, l'espace mémoire requis et le nombre de liens calculé entre Les approches basées sur la similarité locale avec et sans décomposition en composantes connexes pour la prédiction de liens dans les réseaux complexes, il existe trios aspects importants à prendre en considération.

Voici une discussion sur ces trois aspects :

Temps d'exécution :

- **Avec décomposition :** Avec décomposition en composantes connexes, le temps d'exécution peut être réduit de manière significative. En se concentrant uniquement sur les liens à l'intérieur de chaque composante connexe, on réduit le nombre de liens à considérer et donc le temps de calcul global. Les algorithmes et les techniques de similarité locale peuvent être appliqués spécifiquement à chaque composante, ce qui permet des calculs plus rapides et plus efficaces. De plus, les structures de données optimisées pour les sous-graphes individuels peuvent accélérer le traitement des données.
- **Sans décomposition** : Sans décomposition en composantes connexes, le temps d'exécution peut être plus long. Il est nécessaire de prendre en compte l'ensemble du réseau, ce qui implique un plus grand nombre de liens à considérer et donc des calculs plus complexes et plus lents.

Espace mémoire :

- **Avec décomposition :** La décomposition en composantes connexes peut également avoir un impact sur l'espace mémoire requis pour stocker les données du réseau. En utilisant la décomposition, on peut diviser le réseau en sous-graphes plus petits, correspondant à chaque composante connexe. Cela permet de réduire l'espace mémoire nécessaire pour stocker les informations du réseau global, car seules les données spécifiques à chaque composante sont conservées. Cela peut être particulièrement bénéfique dans les réseaux complexes de grande taille, où l'espace mémoire peut être une contrainte.

- **Sans décomposition :** Sans décomposition en composantes connexes, l'espace mémoire nécessaire peut être plus important, car il faut stocker les informations pour l'ensemble du réseau. Cela peut augmenter la consommation de mémoire, en particulier lorsque le réseau est de grande taille.

Nombre de liens :

- **Avec décomposition :** La décomposition en composantes connexes permet de réduire le nombre de liens à considérer pour les calculs de similarité et de prédiction de liens. En se concentrant sur les liens à l'intérieur de chaque composante connexe, on élimine les liens entre les composantes connexes, qui ne sont pas pertinents pour la prédiction de liens. Cela peut réduire le nombre total de liens à calculé et donc le temps de calcul global.
- **Sans décomposition :** Sans décomposition en composantes connexes, tous les liens du réseau sont pris en compte, ce qui implique un plus grand nombre de liens à considérer pour les calculs de similarité et de prédiction.

En résumé, l'utilisation de la décomposition en composantes connexes peut réduire le temps d'exécution en limitant le nombre de liens à considérer et en permettant des calculs plus efficaces. Elle peut également réduire l'espace mémoire requis en se concentrant sur les sous-ensembles de nœuds. En revanche, l'absence de décomposition en composantes connexes peut entraîner des temps d'exécution plus longs, une consommation de mémoire plus élevée et une augmentation du nombre de liens à calculer.

III.8 Conclusion

En conclusion, les résultats obtenus montrent que l'utilisation de la décomposition en composantes connexes permet de réduire à la fois le temps d'exécution et le nombre de liens calculés dans toutes les méthodes de similarité locale utilisées.

Les résultats obtenus indiquent que cette approche de décomposition en composantes connexes permet d'améliorer l'efficacité globale du processus de prédiction de liens dans les réseaux complexes. Non seulement elle réduit le temps d'exécution, mais elle permet également d'économiser des ressources computationnelles en calculant uniquement les liens pertinents.

Ces observations démontrent l'importance de la décomposition en composantes connexes dans la prédiction de liens, en particulier lorsque les méthodes de similarité locale sont utilisées. Cette approche offre une alternative efficace pour réduire la complexité du calcul de liens dans les réseaux complexes, tout en maintenant la précision des prédictions.

Ces observations démontrent également l'importance de la décomposition en composants connexes sur l'espace mémoire nécessaire pour stocker les données du réseau. En utilisant la décomposition, on peut diviser le réseau en sous-graphes plus petits, correspondant à chaque composante connexe. Cela permet de réduire l'espace mémoire nécessaire pour stocker les informations du réseau global.

En résumé, l'utilisation de la décomposition en composantes connexes permet de réduire le nombre de liens à considérer et d'appliquer des techniques de similarité locale plus efficaces, ce qui se traduit par un temps d'exécution réduit. En revanche, l'approche sans décomposition nécessite de calculer les similarités pour l'ensemble du réseau, ce qui entraîne un temps d'exécution plus long. Ainsi, l'utilisation de la décomposition en composantes connexes peut considérablement améliorer l'efficacité du processus de prédiction de liens dans les réseaux complexes.

Conclusion générale :

Vu la croissance du flux et de la masse d'information disponible, il est nécessaire de livrer aux personnes la compréhension des interactions entre eux, et faciliter la visualisation et la navigation dans les réseaux énormes.

Dans le cadre de notre projet de fin d'études, on a travaillé sur un domaine de recherche récent est très intéressant qui est l'analyse des réseaux complexes. Ce domaine constitue une réponse de mise en œuvre des techniques intelligentes de recherches et de connaissances à partir de vastes ensembles de données.

Ces connaissances étés utilisées pour des applications se rattachant à plusieurs domaines. Parmi celle-ci, La prédiction de lien qui est un domaine en phase d'exploration et pour lequel il faudra encore attendre quelques années avant d'arriver à un stade de maturation. Bien que de nombreux algorithmes de prédiction de liens aient été développés récemment dans divers domaine (prédiction des protéines dans le domaine biologique, prédiction des liens dans les réseaux complexes…etc.), la grande majorité des algorithmes existants ne combine pas la structure générale de graphe avec les caractéristiques des nœuds et des liens. La prédiction de lien dans les réseaux complexes, qui est un domaine de recherche rendu très actif par la multiplication du nombre de documents numérique actuellement disponibles. Nous a permis de décrire et analyser quelques méthodes (Voisins Communs (CN) , L'indice Jaccard (JA) , L'indice de Sørensen (SO) , Le Hub Promoted Index (HPI), Le Hub Depressed Index ,L'indice local de Leicht-Holme-Newman (LLHN) , L'indice de Salton (SA)) existantes pour la prédiction des liens.

La partie pratique de ce travail a été dédiée aux expérimentations. Nous avons implémenté sept méthodes locales de prédiction de lien basées sur des mesures de proximité sur quatre réseaux sociaux.

Dans une autre perspective, Nous avons travaillé actuellement sur l'utilisation de la décomposition en composantes connexes dans la prédiction de liens peut réduire de manière significative le temps d'exécution par rapport à l'approche locales sans décomposition. Cela permet d'améliorer l'efficacité globale du processus de prédiction de liens dans les réseaux complexes en réduisant les calculs inutiles. Plus tard Nous pouvons intéresser sur les performances des méthodes proposées avec d'autres approches existantes en termes de temps d'exécution nécessaire et de précision des prédictions de liens.

Bibliographie

[1]. Réka Albert et Albert-László Barabási, « Statistical mechanics of complex networks », Reviews of Modern Physics, vol. 74, no 1, 30 janvier 2002, p. 47–97 (DOI 10.1103/RevModPhys.74.47, lire en ligne [archive], consulté le 16 février 2021)

[2]. M. E. J. Newman, « The Structure and Function of Complex Networks », SIAM Review, vol. 45, no 2, janvier 2003, p. 167–256 (ISSN 0036-1445 et 1095-7200, DOI 10.1137/S003614450342480, lire en ligne [archive], consulté le 15 février 2021)

[3]. John Scott et Peter Carrington, « The SAGE Handbook of Social Network Analysis », Manuel, 2014 (DOI 10.4135/9781446294413, lire en ligne [archive], consulté le 15 février 2021)

[4]. Johan Ugander, Brian Karrer, Lars Backstrom et Cameron Marlow, « The Anatomy of the Facebook Social Graph », arXiv:1111.4503 [physics], 18 novembre 2011 (lire en ligne [archive], consulté le 15 février 2021)

[5]. B. A. Huberman, The laws of the Web : patterns in the ecology of information, MIT Press, 2001 (ISBN 978-0-262-27583-5, 0-262-27583-X et 0-585-44840-X, OCLC 52289489, lire en ligne [archive])

[6]. « Introduction to Informetrics: Quantitative Methods in Library, Documentation and Information Science. Leo Egghe, Ronald Rousseau », The Library Quarterly, vol. 61, no 2, 1er avril 1991, p. 220–221 (ISSN 0024-2519, DOI 10.1086/602337, lire en ligne [archive], consulté le 15 février 2021)

[7]. Revenir plus haut en :a et b (en) L. A. N. Amaral, A. Scala, M. Barthelemy et H. E. Stanley, « Classes of small-world networks », Proceedings of the National Academy of Sciences, vol. 97, no 21, 10 octobre 2000, p. 11149–11152 (ISSN 0027-8424 et 1091-6490, PMID 11005838, PMCID PMC17168, DOI 10.1073/pnas.200327197, lire en ligne [archive], consulté le 15 février 2021)

[8]. Vamsi Kalapala, Vishal Sanwalani, Aaron Clauset et Cristopher Moore, « Scale invariance in road networks », Physical Review E, vol. 73, no 2, 27 février 2006, p. 026130 (ISSN 1539-3755 et 1550-2376, DOI 10.1103/PhysRevE.73.026130, lire en ligne [archive], consulté le 15 février 2021)

[9]. Vito Latora et Massimo Marchiori, « Is the Boston subway a small-world network? », Physica A: Statistical Mechanics and its Applications, horizons in Complex Systems, vol. 314, no 1, 1er novembre 2002, p. 109–113 (ISSN 0378-4371, DOI 10.1016/S0378-4371(02)01089-0, lire en ligne [archive], consulté le 15 février 2021)

[10]. Qian Chen, Hyunseok Chang, R. Govindan et S. Jamin, « The origin of power laws in Internet topologies revisited », Proceedings.Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies, IEEE, vol. 2, 2002, p. 608–617 (ISBN 978-0-7803-7476-8, DOI 10.1109/INFCOM.2002.1019306, lire en ligne [archive], consulté le 15 février 2021)

[11]. A. R. Mashaghi, A. Ramezanpour et V. Karimipour, « Investigation of a protein complex network », The European Physical Journal B - Condensed Matter and Complex Systems, vol. 41, no 1, 1er septembre 2004, p. 113–121 (ISSN 1434-6036, DOI 10.1140/epjb/e2004-00301-0, lire en ligne [archive], consulté le 15 février 2021)

[12]. Shai S. Shen-Orr, Ron Milo, Shmoolik Mangan et Uri Alon, « Network motifs in the transcriptional regulation network of Escherichia coli », Nature Genetics, vol. 31, no 1, mai 2002, p. 64–68 (ISSN 1546-1718, DOI 10.1038/ng881, lire en ligne [archive], consulté le 15 février 2021)

[13]. Jörg Stelling, Steffen Klamt, Katja Bettenbrock et Stefan Schuster, « Metabolic network structure determines key aspects of functionality and regulation », Nature, vol. 420, no 6912, novembre 2002, p. 190–193 (ISSN 0028-0836 et 1476-4687, DOI 10.1038/nature01166, lire en ligne [archive], consulté le 15 février 2021)

[14]. Albert-László Barabási, Natali Gulbahce et Joseph Loscalzo, « Network medicine: a network-based approach to human disease », Nature Reviews Genetics, vol. 12, no 1, janvier 2011, p. 56–68 (ISSN 1471-0064, PMID 21164525, PMCID PMC3140052, DOI 10.1038/nrg2918, lire en ligne [archive], consulté le 15 février 2021)

[15]. David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology 58, 7 (2007), 1019–1031.

[16]. Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. 2006. Link prediction usingsupervised learning. In Proceedings of the Workshop on Link Analysis, Counter-terrorism and Security(SDM'06).

[17]. Indrajit Bhattacharya and Lise Getoor. 2007. Collective entity resolution in relational data. ACM Transactions on Knowledge Discovery from Data (TKDD) 1, 1 (2007), 5v.

[18]. Santo Fortunato. 2010. Community detection in graphs. Physics Reports 486, 3 (2010), 75–174.

[19]. Lieve Hamers, Yves Hemeryck, Guido Herweyers, Marc Janssen, Hans Keters, Ronald Rousseau, and Andre´Vanhoutte. 1989. Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. Information Processing & Management 25, 3 (1989), 315–318.

[20]. Zan Huang, Xin Li, and Hsinchun Chen. 2005. Link prediction approach to collaborative filtering. In Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'05). ACM, 141–142.

[21]. Paul Jaccard. 1901. Etude comparative de la distribution florale dans une portion des alpes et des jura. Bulletin de la Soci´et´e Vaudoise des Sciences Naturelles 37 (1901), 547579.

[22]. Gueorgi Kossinets and Duncan J. Watts. 2006. Empirical analysis of an evolving social network. Science 311,5757 (2006), 88–90.

23. Valdis E. Krebs. 2002. Mapping networks of terrorist cells. Connections 24, 3 (2002), 43–52

[24]. Elizabeth A. Leicht, Petter Holme, and Mark E. J. Newman. 2006. Vertex similarity in networks. Physical Review E 73, 2 (2006), 026120...

[25]. Zhen Liu, Qian-Ming Zhang, Linyuan Lu, and Tao Zhou. 2011. Link prediction in complex networks: A local ¨Na¨ive Bayes model. EPL (Europhysics Letters) 96, 4 (2011), 48007.

[26]. Linyuan Lu and Tao Zhou. 2011. Link prediction in complex networks: A survey. ¨ Physica A: Statistical Mechanics and Its Applications 390, 6 (2011), 1150–1170.

[27]. Vıctor Mart´ınez, Carlos Cano, and Armando Blanco. 2014. ProphNet: A generic prioritization method through propagation of information. BMC Bioinformatics 15, Suppl 1 (2014), S5.

[28]. Mark E. J. Newman. 2001. Clustering and preferential attachment in growing networks. Physical Review E 64, 2 (2001), 025102.

[29]. Joshua O'Madadhain, Jon Hutchins, and Padhraic Smyth. 2005. Prediction and ranking algorithms for event-based network data. ACM SIGKDD Explorations Newsletter 7, 2 (2005), 23–30.

[30]. Gergely Palla, Imre Derenyi, Ill ´ es Farkas, and Tam ´ as Vicsek. 2005. Uncovering the overlapping community ´structure of complex networks in nature and society. Nature 435, 7043 (2005), 814–818.

[31]. Milen Pavlov and Ryutaro Ichise. 2007. Finding experts by link prediction in co-authorship networks. FEWS 290 (2007), 42–55.

[32]. Matthew Richardson and Pedro Domingos. 2002. Mining knowledge-sharing sites for viral marketing. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02). ACM, 61–70.

[33]. Gerard Salton and Michael J. McGill. 1983. Introduction to Modern Information Retrieval. McGraw-Hill,New York.

[34]. Benno Schwikowski, Peter Uetz, and Stanley Fields. 2000. A network of protein–protein interactions in yeast. Nature Biotechnology 18, 12 (2000), 1257–1261.

[35]. Thorvald Sørensen. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biologiske Skrifter 5 (1948), 1–34.

[36]. Roberto Tamassia. 2013. Handbook of Graph Drawing and Visualization. CRC Press.

[37]. Peng Wang, BaoWen Xu, YuRong Wu, and XiaoYu Zhou. 2015. Link prediction in social networks: The state-of-the-art. Science China Information Sciences 58, 1 (2015), 1–38.

[38]. Jennifer Xu and Hsinchun Chen. 2008. The topology of dark networks. Communications of the ACM 51, 10 (2008), 58–65.

[39]. Gerard Salton and Michael J. McGill. 1983. Introduction to Modern Information Retrieval. McGraw-Hill, New York.

[40]. Lieve Hamers, Yves Hemeryck, Guido Herweyers, Marc Janssen, Hans Keters, Ronald Rousseau, and Andre´Vanhoutte. 1989. Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. Information Processing & Management 25, 3 (1989), 315–318.

[41]. Leo Katz. 1953. A new status index derived from sociometric analysis. Psychometrika 18, 1 (1953), 39–43.

[42]. Pavel Chebotarev and Elena Shamis. 2006. Matrix-forest theorems. arXiv preprint math/0602575 (2006).

[43]. Linyuan Lu, Ci-Hang Jin, and Tao Zhou. 2009. Similarity index based on local paths for link prediction of ¨complex networks. Physical Review E 80, 4 (2009), 046122.

Conclusion générale :

Vu la croissance du flux et de la masse d'information disponible, il est nécessaire de livrer aux personnes la compréhension des interactions entre eux, et faciliter la visualisation et la navigation dans les réseaux énormes.

Dans le cadre de notre projet de fin d'études, on a travaillé sur un domaine de recherche récent est très intéressant qui est l'analyse des réseaux complexes. Ce domaine constitue une réponse de mise en œuvre des techniques intelligentes de recherches et de connaissances à partir de vastes ensembles de données.

Ces connaissances étés utilisées pour des applications se rattachant à plusieurs domaines. Parmi celle-ci, La prédiction de lien qui est un domaine en phase d'exploration et pour lequel il faudra encore attendre quelques années avant d'arriver à un stade de maturation. Bien que de nombreux algorithmes de prédiction de liens aient été développés récemment dans divers domaine (prédiction des protéines dans le domaine biologique, prédiction des liens dans les réseaux complexes…etc.), la grande majorité des algorithmes existants ne combine pas la structure générale de graphe avec les caractéristiques des nœuds et des liens. La prédiction de lien dans les réseaux complexes, qui est un domaine de recherche rendu très actif par la multiplication du nombre de documents numérique actuellement disponibles. Nous a permis de décrire et analyser quelques méthodes (Voisins Communs (CN) , L'indice Jaccard (JA) , L'indice de Sørensen (SO) , Le Hub Promoted Index (HPI), Le Hub Depressed Index ,L'indice local de Leicht-Holme-Newman (LLHN) , L'indice de Salton (SA)) existantes pour la prédiction des liens.

La partie pratique de ce travail a été dédiée aux expérimentations. Nous avons implémenté sept méthodes locales de prédiction de lien basées sur des mesures de proximité sur six réseaux.

Dans une autre perspective, Nous avons travaillé actuellement sur l'utilisation de la décomposition en composantes connexes dans la prédiction de liens peut réduire de manière significative le temps d'exécution par rapport à l'approche locales sans décomposition. Cela permet d'améliorer l'efficacité globale du processus de prédiction de liens dans les réseaux complexes en réduisant les calculs inutiles. Plus tard Nous pouvons intéresser sur les performances des méthodes proposées avec d'autres approches en termes de temps d'exécution nécessaire et de précision des prédictions de liens.

Bibliographie

[1]. Réka Albert et Albert-László Barabási, « Statistical mechanics of complex networks », Reviews of Modern Physics, vol. 74, no 1, 30 janvier 2002, p. 47–97 (DOI 10.1103/RevModPhys.74.47, lire en ligne [archive], consulté le 16 février 2021)

[2]. M. E. J. Newman, « The Structure and Function of Complex Networks », SIAM Review, vol. 45, no 2, janvier 2003, p. 167–256 (ISSN 0036-1445 et 1095-7200, DOI 10.1137/S003614450342480, lire en ligne [archive], consulté le 15 février 2021)

[3]. John Scott et Peter Carrington, « The SAGE Handbook of Social Network Analysis », Manuel, 2014 (DOI 10.4135/9781446294413, lire en ligne [archive], consulté le 15 février 2021)

[4]. Johan Ugander, Brian Karrer, Lars Backstrom et Cameron Marlow, « The Anatomy of the Facebook Social Graph », arXiv:1111.4503 [physics], 18 novembre 2011 (lire en ligne [archive], consulté le 15 février 2021)

[5]. B. A. Huberman, The laws of the Web : patterns in the ecology of information, MIT Press, 2001 (ISBN 978-0-262-27583-5, 0-262-27583-X et 0-585-44840-X, OCLC 52289489, lire en ligne [archive])

[6]. « Introduction to Informetrics: Quantitative Methods in Library, Documentation and Information Science. Leo Egghe, Ronald Rousseau », The Library Quarterly, vol. 61, no 2, 1er avril 1991, p. 220–221 (ISSN 0024-2519, DOI 10.1086/602337, lire en ligne [archive], consulté le 15 février 2021)

[7]. Revenir plus haut en :a et b (en) L. A. N. Amaral, A. Scala, M. Barthelemy et H. E. Stanley, « Classes of small-world networks », Proceedings of the National Academy of Sciences, vol. 97, no 21, 10 octobre 2000, p. 11149–11152 (ISSN 0027-8424 et 1091-6490, PMID 11005838, PMCID PMC17168, DOI 10.1073/pnas.200327197, lire en ligne [archive], consulté le 15 février 2021)

[8]. Vamsi Kalapala, Vishal Sanwalani, Aaron Clauset et Cristopher Moore, « Scale invariance in road networks », Physical Review E, vol. 73, no 2, 27 février 2006, p. 026130 (ISSN 1539-3755 et 1550-2376, DOI 10.1103/PhysRevE.73.026130, lire en ligne [archive], consulté le 15 février 2021)

[9]. Vito Latora et Massimo Marchiori, « Is the Boston subway a small-world network? », Physica A: Statistical Mechanics and its Applications, horizons in Complex Systems, vol. 314, no 1, 1er novembre 2002, p. 109–113 (ISSN 0378-4371, DOI 10.1016/S0378-4371(02)01089-0, lire en ligne [archive], consulté le 15 février 2021)

[10]. Qian Chen, Hyunseok Chang, R. Govindan et S. Jamin, « The origin of power laws in Internet topologies revisited », Proceedings.Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies, IEEE, vol. 2, 2002, p. 608–617 (ISBN 978-0-7803-7476-8, DOI 10.1109/INFCOM.2002.1019306, lire en ligne [archive], consulté le 15 février 2021)

[11]. A. R. Mashaghi, A. Ramezanpour et V. Karimipour, « Investigation of a protein complex network », The European Physical Journal B - Condensed Matter and Complex Systems, vol. 41, no 1, 1er septembre 2004, p. 113–121 (ISSN 1434-6036, DOI 10.1140/epjb/e2004-00301-0, lire en ligne [archive], consulté le 15 février 2021)

[12]. Shai S. Shen-Orr, Ron Milo, Shmoolik Mangan et Uri Alon, « Network motifs in the transcriptional regulation network of Escherichia coli », Nature Genetics, vol. 31, no 1, mai 2002, p. 64–68 (ISSN 1546-1718, DOI 10.1038/ng881, lire en ligne [archive], consulté le 15 février 2021)

[13]. Jörg Stelling, Steffen Klamt, Katja Bettenbrock et Stefan Schuster, « Metabolic network structure determines key aspects of functionality and regulation », Nature, vol. 420, no 6912, novembre 2002, p. 190–193 (ISSN 0028-0836 et 1476-4687, DOI 10.1038/nature01166, lire en ligne [archive], consulté le 15 février 2021)

[14]. Albert-László Barabási, Natali Gulbahce et Joseph Loscalzo, « Network medicine: a network-based approach to human disease », Nature Reviews Genetics, vol. 12, no 1, janvier 2011, p. 56–68 (ISSN 1471-0064, PMID 21164525, PMCID PMC3140052, DOI 10.1038/nrg2918, lire en ligne [archive], consulté le 15 février 2021)

[15]. David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology 58, 7 (2007), 1019–1031.

[16]. Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. 2006. Link prediction usingsupervised learning. In Proceedings of the Workshop on Link Analysis, Counter-terrorism and Security(SDM'06).

[17]. Indrajit Bhattacharya and Lise Getoor. 2007. Collective entity resolution in relational data. ACM Transactions on Knowledge Discovery from Data (TKDD) 1, 1 (2007), 5v.

[18]. Santo Fortunato. 2010. Community detection in graphs. Physics Reports 486, 3 (2010), 75–174.

[19]. Lieve Hamers, Yves Hemeryck, Guido Herweyers, Marc Janssen, Hans Keters, Ronald Rousseau, and Andre´Vanhoutte. 1989. Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. Information Processing & Management 25, 3 (1989), 315–318.

[20]. Zan Huang, Xin Li, and Hsinchun Chen. 2005. Link prediction approach to collaborative filtering. In Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'05). ACM, 141–142.

[21]. Paul Jaccard. 1901. Etude comparative de la distribution florale dans une portion des alpes et des jura. Bulletin de la Soci´et´e Vaudoise des Sciences Naturelles 37 (1901), 547579.

[22]. Gueorgi Kossinets and Duncan J. Watts. 2006. Empirical analysis of an evolving social network. Science 311,5757 (2006), 88–90.

23. Valdis E. Krebs. 2002. Mapping networks of terrorist cells. Connections 24, 3 (2002), 43–52

[24]. Elizabeth A. Leicht, Petter Holme, and Mark E. J. Newman. 2006. Vertex similarity in networks. Physical Review E 73, 2 (2006), 026120...

[25]. Zhen Liu, Qian-Ming Zhang, Linyuan Lu, and Tao Zhou. 2011. Link prediction in complex networks: A local ¨Na¨ive Bayes model. EPL (Europhysics Letters) 96, 4 (2011), 48007.

[26]. Linyuan Lu and Tao Zhou. 2011. Link prediction in complex networks: A survey. ¨ Physica A: Statistical Mechanics and Its Applications 390, 6 (2011), 1150–1170.

[27]. Vıctor Mart´ınez, Carlos Cano, and Armando Blanco. 2014. ProphNet: A generic prioritization method through propagation of information. BMC Bioinformatics 15, Suppl 1 (2014), S5.

[28]. Mark E. J. Newman. 2001. Clustering and preferential attachment in growing networks. Physical Review E 64, 2 (2001), 025102.

[29]. Joshua O'Madadhain, Jon Hutchins, and Padhraic Smyth. 2005. Prediction and ranking algorithms for event-based network data. ACM SIGKDD Explorations Newsletter 7, 2 (2005), 23–30.

[30]. Gergely Palla, Imre Derenyi, Ill´ es Farkas, and Tam´ as Vicsek. 2005. Uncovering the overlapping community ´structure of complex networks in nature and society. Nature 435, 7043 (2005), 814–818.

[31]. Milen Pavlov and Ryutaro Ichise. 2007. Finding experts by link prediction in co-authorship networks. FEWS 290 (2007), 42–55.

[32]. Matthew Richardson and Pedro Domingos. 2002. Mining knowledge-sharing sites for viral marketing. In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02). ACM, 61–70.

[33]. Gerard Salton and Michael J. McGill. 1983. Introduction to Modern Information Retrieval. McGraw-Hill,New York.

[34]. Benno Schwikowski, Peter Uetz, and Stanley Fields. 2000. A network of protein–protein interactions in yeast. Nature Biotechnology 18, 12 (2000), 1257–1261.

[35]. Thorvald Sørensen. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. Biologiske Skrifter 5 (1948), 1–34.

[36]. Roberto Tamassia. 2013. Handbook of Graph Drawing and Visualization. CRC Press.

[37]. Peng Wang, BaoWen Xu, YuRong Wu, and XiaoYu Zhou. 2015. Link prediction in social networks: The state-of-the-art. Science China Information Sciences 58, 1 (2015), 1–38.

[38]. Jennifer Xu and Hsinchun Chen. 2008. The topology of dark networks. Communications of the ACM 51, 10 (2008), 58–65.

[39]. Gerard Salton and Michael J. McGill. 1983. Introduction to Modern Information Retrieval. McGraw-Hill, New York.

[40]. Lieve Hamers, Yves Hemeryck, Guido Herweyers, Marc Janssen, Hans Keters, Ronald Rousseau, and Andre´ Vanhoutte. 1989. Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula. Information Processing & Management 25, 3 (1989), 315–318.

[41]. Leo Katz. 1953. A new status index derived from sociometric analysis. Psychometrika 18, 1 (1953), 39–43.

[42]. Pavel Chebotarev and Elena Shamis. 2006. Matrix-forest theorems. arXiv preprint math/0602575 (2006).

[43]. Linyuan Lu, Ci-Hang Jin, and Tao Zhou. 2009. Similarity index based on local paths for link prediction of ¨complex networks. Physical Review E 80, 4 (2009), 046122.