Réseaux complexes : exploration, fouille et analyse multicouche

Maria Malek

Ecole doctorale : Économie, Management, Mathématiques, Physique et Sciences Informatiques (EM2PSI)

Habilitation à diriger des recherches

PRESENTÉE À

CY Cergy Paris université

Spécialité INFORMATIQUE

par **Maria Malek**

Réseaux complexes : exploration, fouille et analyse multicouche

Soutenue le 6 janvier 2022 devant le jury composé de :

M. Hocine CHERIFI,	Pr., Université de Bourgogne	Rapporteur
Mme Florence SÈDES,	Pr., Université Paul Sabatier - Toulouse III	Rapporteure
Mme Rosanna VERDE,	Pr., University of the Campania "Luigi Vanvitelli"	Rapporteure
M. Younès BENNANI,	Pr., Université Sorbonne Paris Nord	Examinateur
Mme Christine LARGERON,	Pr., Université Jean Monnet - Saint-Etienne	Examinatrice
M. Dominique LAURENT,	Pr. émérite, CY Cergy Paris Université	Garant
M. Dan VODISLAV,	Pr., CY Cergy Paris Université	Examinateur

Remerciements

Je tiens tout d'abord à exprimer ma profonde gratitude aux membres du jury qui ont accepté de m'accompagner pendant cette période importante de ma vie professionnelle. Je remercie chaleureusement le président du jury Younès Bennani d'avoir assuré un déroulement du jury qui a suscité des discussions très intéressantes. Je remercie vivement tous les membres du jury de leurs retours, remarques et conseils, ils me seront très utiles pour enrichir mes perspectives de recherche. Je remercie vivement Florence Sèdes, Rosanna Verde et Hocine Cherifi qui m'ont fait l'honneur en acceptant d'être rapporteurs de mon manuscrit malgré leurs diverses responsabilités respectives et leurs emplois du temps chargé. Mes remerciements les plus sincères à Christine Largeron d'avoir accepté d'examiner avec grande attention et intérêt mon manuscrit. Je remercie très vivement Dan Vodislav pour ses conseils, son aide et sa réactivité ainsi que son implication dans la responsabilité de l'équipe MIDI au sein du laboratoire ETIS. Mes chaleureux remerciements à mon garant Dominique Laurent pour son soutien, sa disponibilité, ses précieux conseils, je salue notre collaboration fructueuse courant ces dernières années.

J'adresse également mes vifs remerciements à Olivier Romain, directeur du laboratoire ETIS qui m'a fait confiance en me nommant chargée de mission communication et rayonnement, cette nouvelle expérience enrichissante m'a permis d'interagir avec les chercheurs du laboratoire et d'avoir des échanges fructueux avec les diverses instances des co-tutelles. Je remercie également tous les chercheurs et le personnel du laboratoire ETIS, je remercie particulièrement Dimitris, Claudia, Boris, Hajer, Nistor, Katerina, Inbar, Veronica, Sylvain, Aude, Astrid et Annick.

Mes pensées chaleureuses vont aussi à mes anciens collègues du laboratoire Quartz, j'ai beaucoup apprécié notre collaboration et notre travail soutenu jusqu'à la création du laboratoire Quartz et au-delà, je pense particulièrement à eJan-Pierre Barbot, Marc Zolghadri, Jean-Yves Choley, Achour Ouslimani et Imad Tawfiq.

Je pense aussi à mes (ex) doctorants Dalia Sulieman, Jean-Philippe Attal, Sarra Djemili et Kossi Folly et je les remercie du fond de cœur de m'avoir fait confiance et d'avoir écouté mes conseils, je leur dédie ce manuscrit qui est l'aboutissement de nos travaux communs.

Mes pensées chaleureuses vont à mes collègues de CY Tech (ex-EISTI) pour notre histoire commune que nous avons écrite ensemble, pour les moments de solidarité et d'entraide que nous avons partagés tout au long de notre parcours laborieux courant ces années. Je pense à Chris Baskiotis, à ses interventions pertinentes et

nos discussions très intéressantes. Merci aux collègues qui m'ont soutenu et m'ont exprimé leur sympathie et plus particulièrement pendant cette dernière période de préparation, je pense à Sonia et Marietta. Je pense également à tous mes collègues du département informatique et je n'oublie pas notre investissement collectif qui s'est achevé par la création de notre première entité de recherche à l'EISTI que nous avons baptisée LARIS. Je tiens à remercier plus particulièrement mes deux collègues Sébastien Rufiange et Stephan Bornhofen pour notre petite équipe soudée que nous avons formée lors de notre implication dans le projet Blizaar. Et finalement, je remercie évidemment de tout mon cœur, Nesim Fintz fondateur de l'EISTI de m'avoir toujours exprimé sa confiance, son encouragement et son enthousiasme dans toutes les étapes de ma vie professionnelle.

Résumé

L'étude des réseaux complexes nommés aussi *graphes de terrain* est un domaine de recherche scientifique jeune et actif, largement inspiré des découvertes empiriques sur les réseaux réels tels que les réseaux informatiques, les réseaux biologiques, les réseaux technologiques et les réseaux sociaux.

Un réseau complexe est un graphe (réseau) avec des caractéristiques topologiques non triviales, des caractéristiques qui n'apparaissent pas dans des réseaux simples tels que des treillis ou des graphes aléatoires, mais qui se produisent souvent dans des réseaux représentant des systèmes réels.

Les études ont montré que les réseaux complexes réels présentent des propriétés intéressantes, nous citons : la distribution des degrés en loi de puissance, la propriété du petit monde, l'existence de nœuds jouant des rôles centraux ainsi que l'existence de structures modulaires.

Cependant, les systèmes du monde réel seront mieux représentés par un ensemble de réseaux ou de couches en interaction. Ces réseaux multicouches ont été récemment étudiés par les chercheurs du domaine des réseaux complexes.

Nous nous intéressons dans ce mémoire à l'analyse des réseaux complexes modélisés par des graphes non orientés et plus particulièrement à : a) l'exploration des réseaux complexes pour la recommandation b) la détection des communautés chevauchantes et c) l'analyse des réseaux multicouches.

Exploration des réseaux complexes pour la recommandation : nous proposons une approche fondée sur la combinaison des informations topologiques et nodales extraites à partir du réseau complexe. Cette approche a été appliquée aux systèmes de recommandation explorant les réseaux bibliographiques et les graphes des co-achats en e-commerce.

Détection des communautés chevauchantes : une méthode de détection des communautés chevauchante à partir de communautés disjointes pré-calculées est proposée. L'algorithme sélectionne les nœuds candidats pour le chevauchement et utilise *des fonctions d'appartenance* pour décider de l'affectation ou non d'un nœud candidat à chacune de ses communautés voisines. Ces fonctions d'appartenance sont fondées sur des mesures topologiques extraites du réseau.

Analyse des réseaux multicouches : le but de cette étude est de proposer aux experts des mesures et des méthodes leur permettant de comprendre, d'évaluer ou de compléter leurs données privées en les confrontant avec des données ouvertes lorsque les deux sont modélisées par des réseaux multicouches ; nous proposons une extension des algorithmes et des métriques des

réseaux complexes aux réseaux multicouches. L'étude réalisée sur des réseaux biologiques composés de trois couches (protéines, gènes et métabolites), a mené à une proposition d'un cadre générique pour l'exploration et la fouille de graphes multicouches.

Table des matières

Chapitre 1

Introduction

L'étude des réseaux complexes nommés aussi *graphes de terrain* est un domaine de recherche scientifique jeune et actif depuis 2000 (Barabási et Bonabeau [13], Newman [103]), largement inspiré des découvertes empiriques sur les réseaux réels tels que les réseaux informatiques, les réseaux biologiques, les réseaux technologiques et les réseaux sociaux.

Un réseau complexe est un graphe (réseau) avec des caractéristiques topologiques non triviales, des caractéristiques qui n'apparaissent pas dans des réseaux simples tels que des treillis ou des graphes aléatoires, mais qui se produisent souvent dans des réseaux représentant des systèmes réels.

Les études ont montré que les réseaux complexes réels présentent des propriétés intéressantes, nous citons : la distribution des degrés en loi de puissance, la propriété du petit monde, l'existence de nœuds jouant des rôles centraux ainsi que l'existence de structures modulaires (Newman [105]).

Cependant, les systèmes du monde réel seront mieux représentés par un ensemble de réseaux ou de couches en interaction. Ces réseaux multicouches ont été récemment étudiés par les chercheurs du domaine des réseaux complexes.

Nous nous intéressons dans ce mémoire à l'analyse des réseaux complexes modélisés par des graphes non orientés et plus particulièrement à : a) l'exploration des réseaux complexes pour la recommandation b) la détection des communautés chevauchantes et c) l'analyse des réseaux multicouches.

Exploration des réseaux complexes pour la recommandation : nous proposons une approche fondée sur la combinaison des informations topologiques et nodales extraites à partir du réseau complexe. Cette approche a été appliquée aux systèmes de recommandation explorant les réseaux bibliographiques et les graphes des co-achats en e-commerce.

Détection des communautés chevauchantes : une méthode de détection des communautés chevauchante à partir de communautés disjointes pré-calculées est proposée. L'algorithme sélectionne les nœuds candidats pour le chevauchement et utilise *des fonctions d'appartenance* pour décider de l'affectation ou non d'un nœud candidat à chacune de ses communautés voisines. Ces fonctions d'appartenance sont fondées sur des mesures topologiques extraites du réseau.

Analyse des réseaux multicouches : le but de cette étude est de proposer aux experts des mesures et des méthodes leur permettant de comprendre, d'évaluer ou de compléter leurs données privées en les confrontant avec des données ouvertes lorsque les deux sont modélisées par des réseaux multi-couches ; nous proposons une extension des algorithmes et des métriques des réseaux complexes aux réseaux multicouches. L'étude réalisée sur des réseaux biologiques composés de trois couches (protéines, gènes et métabolites), a mené à une proposition d'un cadre générique pour l'exploration et la fouille de graphes multicouches.

La suite de ce chapitre est organisé comme suit.

Nous présentons dans la suite les notions et notations relatives aux graphes non orientés et plus particulièrement les notions de centralités qui jouent un rôle principal dans l'analyse des réseaux complexes. Les caractéristiques des réseaux complexes ainsi que les différents types de réseaux complexes sont présentés ensuite, nous décrivons plus particulièrement les types de réseaux traités dans ce mémoire à savoir les réseaux d'affiliation, les réseaux projetés ainsi que les réseaux multi-couches.

1.1 Graphes : notions, notations et mesures de centralité

Nous présentons dans cette section les notions et les notations usuelles utilisées pour modéliser les graphes de terrain. Nous nous intéressons plus particulièrement aux graphes non orientés.

- Un graphe G est défini par un couple (V, E), avec $V = \{v_1, ..., v_n\}$ étant l'ensemble des sommets (ou nœuds) et E est l'ensemble des arêtes (ou liens) $E = \{e_1, ..., e_m\}$. Une arête $e_k \in E$ reliant les sommets v_i à v_j peut être représenté par un couple de sommets (v_i, v_j). La notation usuelle pour définir un graphe G est $G = (V, E)$. Nous notons $|V| = n$ le nombre de sommets du graphe et $|E| = m$ le nombre d'arêtes.
- Deux sommets $u, v \in V$ sont dits voisins (ou adjacents) s'ils sont reliés par une arête.
- Un graphe peut également être pondéré ou valué, lorsqu'il existe une fonction $W : E \to \mathbb{R}$ qui associe à chaque lien une valeur réelle. On note le graphe pondéré par $G = (V, E, W)$.
- On note $N(u)$ le voisinage du nœud u, où $N(u) = \{u \in V, (u, v) \in E\}$. Le cardinal du voisinage d'un sommet est son *degré* (ou *degré d'incidence*) que l'on note $d(u)$, on note également k_u le nombre de liens qui est égal au nombre de liens qui lui sont incidents (Freeman [46]) .
- Le degré moyen d'un graphe noté λ_G est la moyenne des degrés des nœuds du graphe : $\lambda_G = \frac{1}{n} \sum_{u \in V} d(u)$.
- Etant donné un ensemble $V' \subseteq V$, le sous-graphe de G engendré (ou induit) par V' est le graphe $G' = (V', E')$, où $E' = \{(u, v) \in E : u, v \in V'\}$
- Un *graphe partiel* de G est un graphe $G' = (V, E')$ qui a le même ensemble de sommets mais pour lequel certaines arêtes ont été éliminées.
- On appelle *graphe complet*, un graphe dans lequel tous les sommets sont adjacents, c'est-à-dire si tout couple de sommets distincts est lié par une

arête. Pour tout entier naturel n, on note K_n le graphe complet d'ordre n. Le nombre d'arêtes du graphe complet K_n est égal à $\frac{n(n-1)}{2}$. On appelle *clique* un sous-graphe complet de G.

— Une chaîne [1] reliant le sommet s au sommet t dans un graphe G est une suite $v_0, v_1, ..., v_k$ de sommets telle que $v_0 = s$, $v_k = t$, $(v_{i-1}, v_i) \in E$, pour tout $1 \leq i \leq k$. k est la longueur de la chaîne. Un graphe non orienté est dit *connexe* ssi deux sommets quelconques sont reliés par une chaîne.

— La *distance géodésique* entre deux sommets dans un graphe est définie par la longueur d'un plus court chemin entre ces deux sommets.

— La densité d'un graphe non orienté, noté ρ_G (ou d_G) est donnée par : $\rho_G = \frac{2m}{n(n-1)}$. La densité d'un graphe complet est égale à 1.

Mesures de centralité Il existe plusieurs mesures pour caractériser la topologie d'un graphe et révéler l'importance d'un nœud. Nous rappelons les mesures les plus utilisées en analyse des réseaux complexes.

La centralité de degré pour un nœud u est définie par $CD_u = \frac{d(u)}{n-1}$, $d(u)$ étant le degré du nœud u.

Elle permet d'exprimer l'importance qu'a le nœud u vis-à-vis de son degré au sein du graphe.

Cependant, il est intéressant d'examiner la variation de la centralité de degré des nœuds du graphe pour en étudier la distribution. En 1978, Freeman [46] a ainsi proposé le *degré de centralisation du graphe G* comme étant $C_D = \frac{\sum_{i=1}^{n}[d(u*)-d(i)]}{[(n-1)(n-2)]}$ où $d(u*)$ est le degré du nœud maximal du réseau. Ce nombre varie entre 0 et 1. La valeur 1 est atteinte pour le graphe en forme d'étoile, c'est-à-dire un nœud connecté à tous les autres, tous de degré 1. La valeur 0 est atteinte pour une clique.

Le cœfficient de clustering (CC) est une mesure de regroupement des nœuds dans un réseau. Il mesure à quel point le voisinage d'un sommet est connecté, et calcule plus exactement la probabilité que deux nœuds liés à un autre nœud soient également liés. Le CC a une forme locale qui concerne un nœud donné et une forme globale qui concerne le graphe entier. Soit $G = (V, E)$ un graphe quelconque, pour un nœud $u \in V$, le CC est défini par :

$$CC_u = \frac{\text{nombre de triangles contenant le nœud } u}{\text{nombre de triplets contenant le nœud } u} \tag{1.1}$$

Les triplets contenant le nœud u correspondent aux paires de voisins du sommet u. Le cœfficient de clustering global pour le graphe G est calculé en utilisant la valeur locale $CC_u, \forall u \in V$

$$CC(G) = \frac{1}{n} \sum_{u \in G} CC_u \tag{1.2}$$

Par définition, $\forall u \in V$, $0 \leq CC_u \leq 1$ et $0 \leq CC(G) \leq 1$. Pour un nœud u, plus grand est son coefficient de clustering, plus la probabilité que ses voisins

1. Le terme chemin est utilisé pour les graphes orientés

soient liés est élevée.

La centralité d'intermédiarité (node betweenness centrality) correspond au ratio des plus courts chemins du graphe passant par chaque sommet pour toutes paires de nœuds s et t d'un graphe G, elle est définie par :

$$CI(v) = \sum_{s \neq v, t \neq v, s \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{1.3}$$

avec σ_{st} le nombre de plus courts chemins entre s et t et $\sigma_{st}(v)$ le nombre de plus courts chemins entre s et t passant par v.

La centralité d'intermédiarité des arêtes : étant donné une arête $e \in E$, la centralité d'intermédiarité de e correspond au ratio de plus courts chemins qui passent par e, et définie par :

$$CI(e) = \sum_{s,t \in V, s \neq t} \frac{\sigma_{st}(e)}{\sigma_{st}} \tag{1.4}$$

$\sigma_{st}(e)$ représentant le nombre de plus courts chemins entre les sommets s à et t et passant par e et σ_{st} étant le nombre de plus courts chemins entre les sommets s et t.

La complexité du calcul de la centralité d'intermédiarité est en $\mathcal{O}(n^3)$ (Brandes [26]). Cependant, les travaux de Brandes [26], Bader *et al.* [11] et Geisberger *et al.* [47] portant sur l'approximation de cette mesure ont pu réduire la complexité en $\mathcal{O}(nm)$.

1.2 Analyse des réseaux complexes

Les réseaux complexes constituent aujourd'hui un outil important permettant de décrire et d'analyser des systèmes complexes qui sont représentés par des graphes. Il existe de nombreuses applications qui illustrent cette utilisation comme les sciences sociales, biologiques, physiques, de l'information et de l'ingénierie (Gosak *et al.* [51], Kivelä *et al.* [72], Newman [105]).

De nombreuses études, notamment celles de Barabási et Albert [12], de Newman [105] et de Clauset *et al.* [33] se sont intéressées à trouver les caractéristiques liées aux réseaux complexes. Elles ont montré des caractéristiques communes concernant la loi de la distribution des degrés des nœuds, un faible nombre de nœuds ayant une forte centralité de degré (les *hubs*), un nombre de triangles important, une distance moyenne assez faible entre chaque paire de sommets ainsi que l'existence de groupes de nœuds fortement connectés ensemble et faiblement avec le reste du graphe, appelés *communautés*.

Distribution des degrés De nombreuses études telles que Albert *et al.* [5], Barabási et Albert [12] et Clauset *et al.* [34] ont constaté que la majorité des graphes de terrain possèdent une distribution des degrés non uniforme, qui est approximée par

une distribution en loi de puissance de type $P(k) = Ck^{-\gamma}$, $P(k)$ étant la proportion de nœuds de degré k et γ appelé exposant d'invariance d'échelle, un réel strictement positif, avec $2 \leq \gamma \leq 3$. Ce type de graphe est qualifié de *réseau invariant d'échelle* (scale-free network). Autrement dit, les nœuds se connectent de manière non uniforme. Certains nœuds (une minorité) attirent les nouveaux nœuds en formant de nouvelles connexions. Ces nœuds que l'on pourrait qualifier d'attracteurs, caractérisés par une forte centralité de degré, sont appelés *hubs* et sont typiques des graphes de terrains.

Effet petit-monde L'effet du petit monde est une expérience menée en 1967 par les psychologues Travers et Milgram [133] qui met en exergue l'hypothèse que la longueur de la chaîne des connaissances sociales requise pour lier une personne arbitrairement choisie à n'importe quelle autre sur terre est généralement courte.

En 2011, le site social Facebook publie une analyse de sa topologie et indiquant qu'il y a en moyenne cinq degrés de séparation entre ses membres (quatre, si l'on se réfère uniquement aux Etats-Unis). Ces expériences montrent qu'un petit nombre d'intermédiaires est suffisant pour connecter n'importe quelle personne à une autre.

Densité faible et cœfficient de clustering élevé Dans un réseau complexe, chaque nœud est connecté à un certain nombre de sommets mais rarement à tous les sommets. Le degré moyen des graphes de terrains est très faible et indépendant du nombre de sommets du graphe (Albert *et al.* [5]). Les graphes de terrain ont un cœfficient de clustering élevé. C'est-à-dire que deux sommets voisins d'un nœud auront tendance à être liés. Ainsi, des nœuds dans certaines régions denses du graphe seront connectés à de nombreux triangles. Cependant, Melançon [98] montre que la densité d'un graphe varie en fonction du domaine d'application en donnant des exemples réels.

Structures communautaires En 2002, Girvan et Newman [50] ont montré que la présence au sein des réseaux complexes de groupes de nœuds fortement connectés entre eux et faiblement avec le reste du graphe est une caractéristique des réseaux complexes, ils donnent le nom de communautés à ces groupes de nœuds fortement connectés.

Considérons un réseau complexe représenté par un graphe $G = (V, E)$, la problématique de la détection de communautés dans sa forme générale consiste à trouver une partition $P = (C_1, ..C_k)$ de l'ensemble des sommets V en k classes, de telle sorte que les sommets dans une communauté soient fortement connectés entre eux et faiblement avec le reste du graphe.

Nous nous intéressons dans ce mémoire plus particulièrement aux algorithmes de détection de communautés chevauchantes ainsi qu'à l'utilisation des propriétés caractéristiques des réseaux complexes pour la détection du chevauchement.

D'un autre côté, il existe une grande diversité de réseaux complexes et cela rend leur classification selon leurs propriétés communes difficiles, mais nous pouvons retenir quatre groupes principaux : les réseaux sociaux, les réseaux d'informations, les réseaux technologiques, et les réseaux biologiques. Nous présentons dans la suite les types de réseaux complexes que nous avons traités dans ce mémoire.

1.2.1 Les réseaux sociaux

L'analyse des réseaux sociaux est définie comme étant l'étude des entités sociales (les personnes dans les organisations qu'on appelle acteurs) ainsi que leurs interactions et leurs relations. Ces interactions et relations peuvent être représentées par un graphe ou un réseau, dans lequel chaque nœud représente un acteur et chaque lien est une relation. Ces interactions peuvent être très variées, comme des liens d'amitié ou de parenté, des activités professionnelles ou personnelles communes, ou encore le partage des mêmes opinions.

Nous pouvons étudier la position, le rôle et le prestige de chaque acteur social. Nous pouvons rechercher aussi les différents types de sous-graphes comme par exemple les communautés formées par des groupes d'acteurs ayant des intérêts communs, en isolant le groupe d'individus ayant une densité élevée.

Le réseau social peut être aussi une source permettant l'élaboration de recommandations : trouver un expert dans un domaine donné, suggérer des produits à vendre, proposer un ami, etc. Cette élaboration peut être fondée sur des algorithmes d'exploration de chemins, d'analyse de degrés, etc.

D'un autre côté, la compréhension des mécanismes propagation des discours, des opinions, des rumeurs dans les réseaux sociaux devient un enjeu de société. L'analyse de sentiment (opinion mining) s'appuie sur diverses techniques telles que le traitement du langage naturel, la recherche d'informations et l'exploration de données structurées et non structurées. Des chercheurs d'horizons différents ont présenté plusieurs modèles pour étudier la formation, la propagation et l'agrégation des opinions selon différents points de vue (Mohammadinejad *et al.* [100]).

Nous explorons dans nos travaux actuels la combinaison de méthodes classiques d'exploration d'opinion avec l'analyse des réseaux complexes et son impact sur la formation et la propagation d'opinion afin de construire un modèle d'opinion cohérent (Folly *et al.* [43]).

1.2.2 Les réseaux d'affiliation et les réseaux de collaboration

Les réseaux d'affiliation sont une classe particulière de réseaux complexes, dont les nœuds sont divisés en deux ensembles X et Y, et seules les connexions entre deux nœuds dans des ensembles différents sont autorisées.

De nombreux réseaux complexes du monde réel peuvent être modélisés naturellement par un graphe bipartite. Citons pour exemple a) le réseau acteurs-films, où chaque acteur est lié aux films dans lesquels il/elle a joué, b) les réseaux d'affiliation bibliographique où les auteurs sont liés aux articles qu'ils ont signés. Ces réseaux sont souvent appelés réseaux d'affiliation ou réseaux à deux modes.

Les réseaux d'affiliation peuvent généralement être compressés par projection à un ensemble. Cela signifie que le réseau qui en résulte contient des nœuds uniquement de l'un des deux ensembles, une paire de nœuds X (ou, alternativement, Y) est connectée seulement si les deux nœuds ont au moins un nœud Y comme voisin commun (voir figure 4.15). Par exemple, le réseau acteurs-films se transforme en sa projection sur l'ensemble des acteurs où deux acteurs sont liés ensemble s'ils ont agi ensemble dans au moins un film. Ce réseau projeté est pondéré, le poids d'un un lien entre deux nœuds correspond au nombre de leurs voisins communs dans le graphe bipartite.

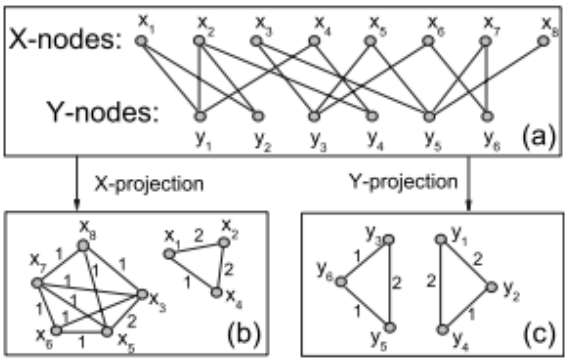


FIGURE 1.1 – Illustration de la projection d'un réseau d'affiliation (source : Zhou *et al.* [152]).

Les réseaux projetés sont souvent appelés réseaux de collaboration et sont considérés par plusieurs chercheurs comme étant des graphes de similarité pondérés car ils expriment des comportements similaires extraits du réseau d'affiliation.

Nous avons évalué certains de nos algorithmes sur des réseaux projetés à partir de certains réseaux d'affiliation. Par exemple, notre approche de recommandation décrite dans le deuxième chapitre a été évaluée sur des réseaux de co-citation et de couplage bibliographique extraits à partir des réseaux de citation bibliographique ainsi que sur des réseaux de co-achats extraits des données d'achats (utilisteurs-items).

1.2.3 Les réseaux multicouches et les systèmes réels

Comme nous venons de l'évoquer ci-dessus, les réseaux sont souvent modélisés, analysés comme étant des graphes de terrains ayant des propriétés connues. Cependant, les systèmes du monde réel seront mieux représentés par un ensemble de réseaux ou de couches en interaction. Ces réseaux multicouches ont été récemment étudiés par les chercheurs du domaine des réseaux complexes (Kivelä et al., 2014). De même, les chercheurs de divers domaines d'application étudient ces systèmes ; nous citons la biologie, les sciences humaines numériques, la sociologie, l'analyse financière, la géographie et le journalisme. Nous proposons dans la troisième chapitre de ce mémoire une extension des algorithmes et métriques des réseaux complexes aux réseaux multicouches. Cette étude appliquée initialement aux réseaux biologiques composés de trois couches (protéines, gènes et métabolites), a mené à une proposition d'un cadre générique ainsi que de nouveaux algorithmes d'exploration et de fouille de graphes multicouches. Le but de cette étude est proposer aux experts des domaines des mesures et des méthodes leur permettant de comprendre, d'évaluer ou de compléter leurs données privées en les confrontant avec des données ouvertes lorsque les deux sont modélisées par des réseaux multicouches (Malek *et al.* [91])).

1.3 Contenu du mémoire

Ce mémoire est composé d'une introduction de trois chapitres et d'une conclusion. Le deuxième chapitre intitulé **exploration des réseaux complexes pour**

la recommandation traite la problématique de recommandation dans les réseaux complexes avec une approche fondée sur la combinaison des informations topologiques et nodales qui explore et combine trois types d'information extraits à partir d'un réseau complexe. Le premier type concerne l'information stockée au niveau de chaque nœud comme par exemple le profil utilisateur et/ou ses préférences et qui peut être adossée à une ontologie de domaine. Le deuxième type d'information est celle présente dans la structure même du réseau ; nous utilisons par exemple des méthodes permettant d'explorer l'arbre couvrant ainsi que des algorithmes de parcours en largeur ou en profondeur associés à des heuristiques. Le troisième type d'information concerne les propriétés du réseau et consiste à utiliser les mesures locales comme la centralité de degré, l'intermédiarité et le prestige pour guider les algorithmes de navigation dans le graphe.

Le troisième chapitre est intitulé **détection des communautés disjointes et chevauchantes**. Nous discutons dans ce chapitre la notion de la structure communautaire dans les graphes de terrain ainsi que les critères permettant de qualifier les communautés. Nous présentons également les différentes approches de détection de communautés disjointes dans un graphe de terrain ainsi que les mesures d'évaluation internes et externes proposées dans la littérature. Ensuite, les différentes approches de detection de communautés chevauchantes sont présentées ainsi que les mesures d'évaluation internes et externes qui sont obtenues par adaptation des mesures proposée pour les communautés disjointes. Enfin, une méthode de détection des communautés chevauchante à partir de communautés disjointes pré-calculées en utilisant une version stable de propagation des labels, nommé *core detection label propagation*(CDLP) et décrit dans Attal et Malek [8], est présentée. L'approche proposée donne des résultats relativement bons et compétitifs en terme de qualité mais dépendent de la topologie du réseau.

Dans le quatrième chapitre intitulé **analyse des réseaux multicouches pour des données ouvertes et privées**, nous présentons une méthodologie d'analyse de réseau multicouche afin de fournir aux experts des mesures et des méthodes pour comprendre, évaluer et compléter leurs données privées en les comparant et/ou combinant avec des données ouvertes lorsque les deux sont modélisés par des réseaux multicouches.

Les principales contributions sont (Malek *et al.* [91]) :

1. Proposition d'un nouveau formalisme de réseau multicouche qui permet de réaliser une analyse fine en considérant deux niveaux : le niveau intra-couche et celui inter-couche.
2. Définition *du réseau multicouche privé* qui correspond au graphe induit élaboré à partir des données privées, ce réseau sera analysé et comparé à l'ensemble du réseau.
3. Définition de la notion *du réseau égocentrique privé* : la notion de réseau égocentrique qui se définit autour d'un nœud égo proposé dans (Djemili *et al.* [40], Marsden [94]) est étendue à un réseau égocentrique autour d'un réseau privé multicouche. Le réseau égocentrique privé peut être utilisé pour évaluer la force de connectivité des données privées par rapport à l'ensemble du réseau à travers les différentes couches. Le réseau égocentrique privé peut également

aider à centrer l'étude du réseau privé dans l'espace de ses voisins et à travers les couches notamment dans le cadre de réseaux ouverts à très grande échelle.

4. Définition *des métriques d'accessibilité inter-couches* d'un sous-réseau donné : cette mesure est basée sur le réseau égocentrique privé et permet d'apprécier la force de connectivité des données privées à travers les couches.

Nous illustrons notre méthodologie à travers une application biologique. Le réseau multicouche ouvert est construit à partir de bases de données ouvertes constituées des interactions pondérées entre des molecules. Ces interaction peuvent être de types : protéines-protéines, métabolites-métabolites ou protéines-métabolites. Ce travail a été réalisé dans la cadre du projet ANR BLIZAAR : `https://anr.fr/Projet-ANR-15-CE23-0002`.

Et finalement, nous terminons le mémoire avec une conclusion et des perspectives.

Chapitre 2

Exploration des réseaux complexes pour la recommandation

Sommaire

2.1 Introduction

Un réseau complexe peut être une source permettant l'élaboration de recommandations comme par exemple : trouver un expert dans un domaine donné, suggérer des produits à vendre, proposer un ami, etc. Cette élaboration peut être fondée sur des algorithmes d'exploration de graphe comme par exemple : l'exploration de chemins et d'analyse de degrés,etc.

Nous abordons la problématique de recommandation dans les réseaux complexes avec une approche fondée sur la combinaison des informations topologiques et nodales qui explore et combine trois types d'information extraits à partir d'un réseau complexe. Le premier type concerne l'information stockée au niveau de chaque nœud comme par exemple le profil utilisateur et/ou ses préférences et qui peut être adossée à une ontologie de domaine. Le deuxième type d'information est celle présente dans la structure même du réseau ; nous utilisons par exemple des méthodes permettant d'explorer l'arbre couvrant ainsi que des algorithmes de parcours en largeur ou en profondeur associés à des heuristiques. Le troisième type d'information concerne les

propriétés du réseau et consiste à utiliser les mesures locales comme la centralité de degré, l'intermédiarité et le prestige pour guider les algorithmes de navigation dans le graphe.

La suite de ce chapitre est organisée ainsi : nous décrivons dans la première partie les approches de recommandation d'expert par exploration de graphes et de réseaux sociaux, nous détaillons notre approche avec une proposition d'un algorithme exhaustif et d'un algorithme guidé par une heuristique (Kadima et Malek [65]) et nous présentons une application dans le domaine des réseaux bibliographiques de co-citations entre auteurs. La deuxième partie du chapitre concerne une approche de recommandation d'items qui intègre une ontologie de domaine. La construction du réseau de collaboration entre utilisateurs à partir d'un graphe bipartite les reliant aux items est présentée. Les algorithmes de recommandation fondés sur les parcours en largeur et profondeur avec utilisation d'heuristiques sont détaillés (Sulieman [128], Sulieman *et al.* [129]), une étude expérimentale est ensuite effectuée sur deux réseaux (benchmarks) de collaboration qui sont MovieLens et Amazon.

2.2 Recommandation d'expert par exploration de graphes et de réseaux sociaux

Étant donné une tâche particulière et un ensemble d'experts, le problème de l'identification des experts consiste à trouver un ensemble d'experts compétents. Nous considérons le cas lorsque des experts sont organisés en réseaux qui correspondent aux réseaux sociaux ou aux structures organisées d'entreprise où le réseau peut capturer les organisations hiérarchiques. Dans une communauté de recherche, le réseau capture la collaboration antérieure entre scientifiques. L'identification de l'expert peut s'effectuer avec la propagation des scores qui est définie ainsi (Aggarwal [3]) : *Étant donné une requête Q qui consiste en une liste de compétences et un graphe social G, il s'agit d'identifier un sous-ensemble de candidats qui ont la compétence spécifiée par la requête et ceci en utilisant le graphe pour propager des scores en vue de proposition de classement des experts.*

2.2.1 Travaux reliés

L'expertise d'un candidat peut être déduite des compétences d'autres acteurs avec qui il est connecté. Certains algorithmes populaires conçus initialement pour le tri des pages web tels que PageRank (Brin et Page [27]) et Hits (Kleinberg [73]) peuvent être utilisés pour le problème de localisation des experts. Dans Campbell *et al.* [30], les auteurs ont proposé d'utiliser le réseau de communication par email pour affiner l'identification des experts. Dans ce réseau, les personnes qui ont reçu de nombreuses demandes par courrier sont définies comme les *autorités* et les experts et les gens qui sont en mesure de transmettre des questions à de nombreux experts sont définis comme *centres (hubs)*. Dans Zhang *et al.* [149] et Lu *et al.* [86], les auteurs ont utilisé pageRank et Hits sur des réseaux communautaires de réponse aux questions ; un utilisateur A qui a répondu à la question d'un autre utilisateur B signifie souvent que A a plus connaissances sur le sujet que B, etc.

De nombreux systèmes ont utilisé les connexions entre individus pour identifier

les experts. Nous citons le système ArnetMiner développé par Tang *et al.* [132] pour la recherche universitaire. Étant donné une requête, le système renvoie une liste adéquate d'experts, le système suggère également les meilleures conférences et communications liées à la requête. Nous mentionnons également d'autres exemples de systèmes sociaux comme le *Spree* par Metze *et al.* [99] et l'*expert recommender* par McDonald et Ackerman [96].

D'un autre côté, les algorithmes de recherche dans les graphes ont été utilisés pour la recommandation d'experts dans les réseaux sociaux. Nous citons les stratégies suivantes (Zhang et Ackerman [148]) :

Exploration en largeur qui diffuse la requête à chaque acteur dans le réseau social en suivant une exploration en largeur.

Recherche aléatoire qui choisit au hasard un voisin le long duwuel se propager la requête.

Recherche de meilleure connexion proposé par Adamic *et al.* [1] et qui utilise la distribution des degrés au sein du réseau social.

Les algorithme de faibles et de forts liens qui sont fondés sur le fait que les liens entre les deux individus peuvent avoir différentes degrés de forces. La force du lien varie et n'est pas toujours symétrique.

La recherche fondée sur la distance de Hamming qui choisit parmi les voisins ceux qui ont les moindre d'amis en commun avec l'acteur actuel.

La recherche fondée sur l'information qui choisit l'acteur ayant le profil le plus similaire à la requête.

Les stratégies de recherche dans un graphe ont été appliquées et évaluées sur les données des mails (Campbell *et al.* [30]) et plus particulièrement sur la base de Enron (Zhang et Ackerman [148]). Parmi les critères d'évaluation nous citons : le nombre de personnes trouvées par requête, la profondeur de la chaîne de recherche, etc. Les expériences ont montré que la recherche fondée sur l'information n'est pas plus performante que les stratégies fondées sur les liens sortants comme le parcours en largeur ou bien la distance de Hamming. Les stratégies de liens faibles peuvent parfois être très utiles pour trouver les nouvelles informations.

Dans Zhang *et al.* [149], la recommandation est formalisée comme étant un problème de tri dans un réseau social hétérogène. La recherche aléatoire est utilisée pour naviguer dans ce type de réseau.

D'un autre côté, le problème particulier lié à l'expertise scientifique à travers l'analyse de la bibliographie a été étudié dans la littérature. Newman [104] a traité le thème de la collaboration scientifique et a étudié les relation entre les auteurs. Il a également étudié les caractéristiques de ce réseau ainsi que sa structure.

Dans Grossman [55], les auteurs ont étudié la structure du réseau social constitué des papiers publiés dans le domaine des mathématiques. les nœuds de ce réseau représentent les auteurs. Un lien entre deux nœuds-auteurs correspond à un papier co-écrit pas les deux auteurs. L'évolution de ce type de réseau a été également analysée.

La figure 2.1 montre un réseau de citations concernant les auteurs principaux des publications extraites de l'ensemble de données de *InfoVis contest* (Weimao Ke *et al.* [138]).

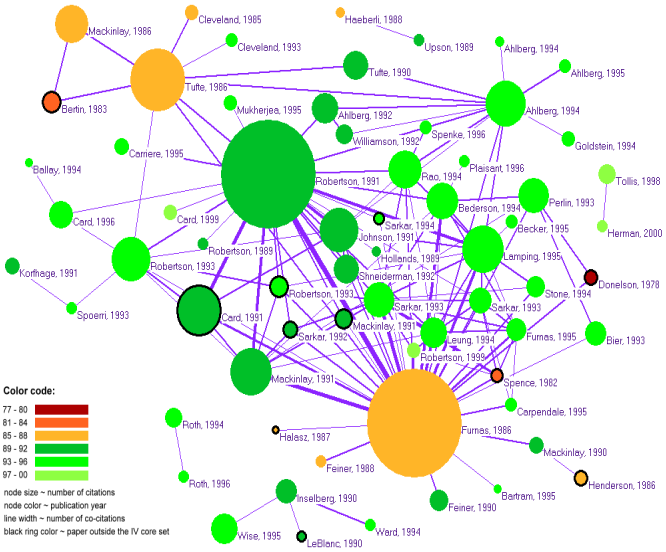


FIGURE 2.1 – Réseau de citations des auteurs des papiers principaux publiés avant 2004, extraits de l'ensemble de données *InfoVis contest* (Weimao Ke *et al.* [138]).

2.2.2 Approche de recommandation d'experts

Nous présentons une approche de recommandation d'experts dans un réseau complexe (social, professionnel, ou de collaboration) (Kadima et Malek [65], Malek et Sulieman [89]). Le réseau est composé d'un ensemble de personnes ayant des interactions entre elles. Ces interactions peuvent être sociales professionnelles ou bien des interactions de collaboration. Notre but est de proposer un algorithme de recommandation efficace. L'application d'une telle approche peut être dans la recherche d'une expertise ou bien d'une collaboration scientifique ou professionnel. Selon la demande d'un acteur d'origine X, le système doit proposer (recommander) un ou plusieurs acteurs $\{Z1, Z2, .., Z_n\}$ répondant le mieux possible aux critères demandés (exemple : recherche d'une personne ayant des compétences données pour un poste, etc.). Nous proposons un algorithme/approche de recommandation qui utilise les trois types d'information suivants :

— L'information stockée sur la personne (l'acteur ou le nœud du graphe) d'une façon décentralisée au niveau de chaque nœud. Cette connaissance peut être représentée en utilisant une ontologie décrivant les profils utilisateurs.

— L'information décrite par la structure du réseau même. Autrement dit, en explorant les liens partants de l'acteur origine x et en utilisant des algorithmes d'exploration de graphe comme les chemins les moins coûteux, nous pouvons délimiter le champ de recherche de l'ensemble $\{Z1, Z2, .., Z_n\}$. Notre contribution consiste à utiliser l'arbre couvrant du graphe.

— L'information explicitée par les mesures liées aux acteurs intermédiaires passant par les chemins retenus. Cette heuristique donne plus d'importance aux

chemins ayant des acteurs plus prestigieux.

Nous décrivons dans la suite notre système de recommandation, nous proposons ensuite une réalisation du système pour la recommandation d'auteurs dans le domaine de la bibliographie scientifique. Nous exposons ensuite les résultats.

2.2.3 Système de recommandation d'experts

Le principe du système est de proposer (recommander) un ou plusieurs acteurs répondant au mieux aux critères demandés à partir d'une requête posée par un utilisateur X qui est lui même acteur dans le réseau social.

Idée de l'algorithme

Comme mentionné ci-dessous, l'idée étant de proposer un algorithme de recherche qui combine la sémantique, la structure & les propriétés des réseaux sociaux (Kadima et Malek [65]) :

La partie sémantique La partie sémantique consiste à calculer la mesure de pertinence entre la requête posée par l'acteur X et le profil stocké dans un nœud donné :

- R_X est la requête posée par le sommet X sous forme d'un ensemble de termes T_i : $R_X = \{T_1, T_2.., T_n\}$. T_i étant un terme donné, P_i le poids associé.
- Pro_Z est le profil associé à un sommet donné Z donné également par un ensemble de termes pondérés : $Pro_Z = \{(T_1, P_1), (T_2, P_2).., (T_m, P_m)\}$.
  Nous utilisons dans la suite l'opérateur . pour accéder aux termes ou au profils de la manière suivante : $Pro_Z.T_k = T_k$ et $Pro_Z.P_k = P_k$
- Nous définissons *la pertinence* (relevance en anglais) entre la requête R_X et le profil du sommet Pro_Z par :

$$rel(R_X, Pro_Z) = \frac{\sum\limits_{j \in inter(R_X, Pro_Z)} Pro_Z.P_j}{\sum\limits_{j=1}^{m} Pro_Z.P_j + |R_X \setminus Pro_Z|}$$

  .
  avec : $inter(R_X, Pro_Z) = \{k \in \{1, ..m\}, Pro_Z.T_k \in R_X\}$
  la fonction *inter* calcule les termes en commun entre la requête et le profil.
  La fonction $\setminus$ correspond à la différence ensembliste. Cette mesure est une adaptation de la mesure de similarité de Jaccard dans un contexte de termes pondérés.

La partie topologique Il s'agit de trouver l'arbre couvrant maximum à partir d'un graphe valué en utilisant une version adaptée de l'algorithme de Kruskal Kruskal [76].

Le but est d'améliorer la recherche en effectuant une navigation optimisée dans l'arbre couvrant au lieu d'explorer le graphe ou une partie du graphe. L'arbre couvrant maximum sera l'arbre couvrant le plus représentatif dans le graphe.

Intermédiarités des nœuds Soient deux nœuds *non adjacents* k et j ayant au moins un chemin les reliant, si le nœud i se trouve sur un ou plusieurs des chemins les plus courts qui les relient alors *i est un acteur intermédiaire*. Nous calculons l'intermédiairé de l'acteur i selon la formule :

$$C_B(i) = \sum_{j<k} \frac{p_{jk}(i)}{p_{jk}}$$

p_{jk} étant le nombre des chemins les plus court entre j et k, et $p_{jk}(i)$ le nombre des chemins les plus court entre j et k passant par i.

L'utilisation de la mesure de l'intermédiarité va nous permettre de privilégier certains chemins de recherche par rapport à d'autres.

L'algorithme de recommandation

Nous proposons un algorithme avec l'entrée et la sortie suivantes :

Entrée : une requête posée par l'acteur X formulée par une suite de mots clés (termes).

Sortie : une suite pondérée d'auteurs $\{(Z_1, P_1), (Z_2, P_2).., (Z_n, P_n)\}$ correspondant au mieux à la requête *ainsi que* : la chaîne sémantique reliant les deux acteurs (X, Z_i) : une chaîne sémantique reliant deux acteurs X, Z_i est constituée de la liste de mots clefs (termes) se trouvant dans la suite des sommets reliant X à Z_i.

Etapes de l'algorithme

1. Calculer et stocker les intermédiarités des nœuds.
2. Trouver l'arbre couvrant maximum selon les poids des arêtes.
3. Extraire de l'arbre couvrant une liste de sommets triée à recommander en utilisant l'algorithme exhaustif ou l'algorithme guidé détaillés par la suite.

Algorithme exhaustif pour la recherche de toutes les solutions

1. Recherche dans A (l'arbre couvrant) des sommets Z_i à recommander à X à partir de sa requête (voir figure 2.2), et ceci en effectuant un parcours en largeur dans A : trouver un ensemble $\{Z_1, Z_2, .., Z_n\}$ tel que $rel(R_X, Pro_{Z_i}) >= seuil$.
2. Nous associons à chaque Z_i proposé un poids qui exprime l'importance de la recommandation et qui permet par la suite de trier les sommets (rating) ; ce poids dépend de la valeur de pertinence entre la requête et le profil de Z_i ainsi que de l'intermédiarité des nœuds se trouvant sur le chemin de la solution : Soient $[Y_1, Y_2, .., Y_l]$ la liste des sommets se trouvant sur la chaîne reliant X à Z_i.
   P_i étant le poids à associer au sommet Z_i, P_i est calculé par :
   $P_i = rel(R_X, Pro_{Z_i}) * \frac{\sum_{j=1}^{l}(C_B(j)}{l}$ **si** $l > 1$

   $P_i = rel(R_X, Pro_{Z_i})$ **sinon**

Exemple du déroulement de l'algorithme exhaustif La figure 2.2 montre un exemple du déroulement de l'algorithme de recommandation dans un arbre couvrant. X étant l'auteur qui soumet la requête et qui correspond à la racine de l'arbre. Le résultat final sera une liste d'auteurs triée selon leurs poids. leurs distance de la racine X est également calculée (voir tableau 2.1).

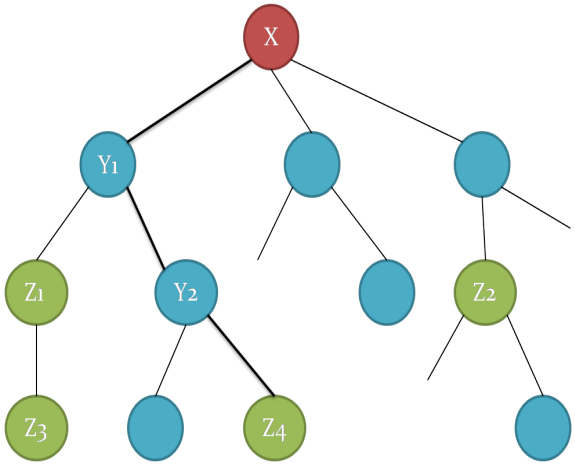


FIGURE 2.2 – Exemple du déroulement de l'algorithme de recommandation dans un arbre couvrant : X étant l'auteur qui soumet la requête ; une recherche des sommets pertinents est effectuée, soient $[Z_4, Z_3, Z_1, Z_2]$ cette liste : un poids qui dépend de la mesure d'intermédiartié des acteurs sur le chemin (X, Z_i) est affecté à chaque sommet retenu ; la liste de sommets à recommander est $[(Z_4, P_4), (Z_3, P_3), (Z_1, P_1), (Z_2, P_2)]$ (Kadima et Malek [65]).

Auteur	Poid	Distance
Z_4	P_4	3
Z_3	P_3	3
Z_1	P_1	2
Z_2	P_2	2

Tableau 2.1 – Résultat du déroulement de l'algorithme sur l'exemple illustré dans la figure (2.2), la liste des auteurs à recommander est triée selon leurs poids. leurs distances de la racine de l'arbre est également calculée

Algorithme guidé pour la recherche d'une solution Nous proposons une deuxième version moins coûteuse permettant de trouver une solution permettant de trouver plus rapidement le chemin de la recherche dans l'arbre couvrant au lieu d'effectuer un parcours en largeur.

Nous utilisons une heuristique permettant de choisir le sommet à visiter parmi un ensemble de sommets candidats et nous appliquons ensuite un algorithme de type A*, permettant de passer à chaque étape par le sommet Y minimisant l'heuristique :

$$h(Y) = (seuil - rel(R_X, Pro_Y)) * (1 - C_B(Y))$$

jusqu'à ce qu'on arrive à un sommet Z à recommander pour lequel nous avons :

$$rel(X, Z) >= seuil.$$

Nous démontrons théoriquement que notre heuristique est monotone, nous démontrons également qu'elle reconnaît la solution, car elle prend la valeur nulle pour la solution.

D'un autre côté, nous montrons par l'expérimentation que cette version converge plus rapidement vers la solution et permet d'explorer de 11 à 49 % de l'arbre couvrant en comparaison avec la version exhaustive.

2.2.4 Application : réseau de co-citations bibliographiques

Nous montrons dans cette section l'application de notre approche sur un réseau d'experts composé d'un ensemble d'auteurs de références bibliographiques. Le réseau est un graphe de similarité dont les nœuds sont les auteurs. Un lien est créé entre deux auteurs s'ils sont similaires par rapport à leur citation. Deux auteurs sont similaires s'ils citent un certain nombre d'articles en commun *et/ou* s'ils sont cités par un certain nombre d'articles. Nous analysons le site de références bibliographiques *libra.msra.cn* : *communauté datamining*.

Nous procédons tout d'abord par l'extraction du *graphe de citations* entre auteurs qui est un graphe orienté, le graphe de similarité entre auteurs (graphe non orienté) est ensuite construit à partir du graphe de citations (Malek et Sulieman [89]).

Graphe de citations

Par simplification, nous nous sommes limités à l'ensemble des publications effectuées à partir de l'année 2005. A partir de l'analyse des citations, le graphe de citations est extrait.

Le graphe de citations entre auteurs est un graphe dirigé dans lequel : les nœuds sont les auteurs et les liens dirigés sont les citations entre auteurs pondérées par leur nombre.

Nous stockons au niveau de chaque acteur-auteur un vecteur *pondéré* de mots clefs qui constituera le profil utilisateur. Ce vecteur est extrait en effectuant une analyse simplifiée du texte constituant les titres des articles.

Extraction du graphe de similarité entre auteurs

Le graphe de similarité est un graphe non dirigé extrait à partir du graphe de citations précédent : les nœuds de ce graphe sont les auteurs. Une arête entre deux auteurs exprime la similarité entre deux auteurs.

Nous rappelons que deux auteurs sont similaires s'ils citent un certain nombre d'articles en commun *et/ou* s'ils sont cités par une certain nombre d'articles. Le graphe de similarité est construit à partir de deux mesures qui sont : le couplage bibliographique ainsi que la co-citation détaillés à la suite.

Les mesures de citations bibliographiques

Nous traitons dans cette partie les mesures de citation bibliographiques. Une publication sous n'importe quelle forme contient une partie de citations de références bibliographiques. Quand un papier i cite une autre papier j, un lien dirigé est créé entre les deux papiers dans le sens i vers j. Ce lien peut donner une indication de relations entre auteurs, papiers, etc. Les citations sont modélisées par une matrice L appelée la matrice de citation ; le terme L_{ij} vaut 1 si i cite j, sinon il vaut 0. Nous présentons deux mesures de citations : la co-citation et le couplage bibliographique.

La co-citation La co-citation est une mesure de similarité entre deux documents qui exprime le fait que si les papiers i et j sont cités par le papier k alors ils sont liés. De même, si les papiers i et j sont cités par plusieurs papiers alors ils sont *similaires*. Nous pouvons extraire la matrice de co-citation à partir de la matrice de citations L mentionnée la dessous. La co-citation est donc une mesure qui est définie par :

$$C_{ij} = \sum_{k=1}^{n} L_{ki} L_{kj}$$

Remarquer bien que la mesure de co-citation est symétrique.

Le couplage bibliographique Le couplage bibliographique est une mesure de similarité entre deux documents qui exprime le fait que si les papiers i et j citent le papier k alors ils sont liés. De même, si les papiers i et j citent plusieurs papiers alors ils sont *similaires*. Nous pouvons extraire la matrice du couplage bibliographique à partir de la matrice de citations L mentionnée la dessous. Le couplage bibliographique est donc une mesure qui est définie par :

$$B_{ij} = \sum_{k=1}^{n} L_{ik} L_{jk}$$

Remarquer bien que la mesure du couplage bibliographique est symétrique.

Le graphe de similarité entre auteurs Nous définissons le graphe de similarité à partir de la somme des deux matrices C et B qui représentent simultanément la co-citation et le couplage bibliographique. Si A_i est un des auteurs du papier i et A_j un des auteurs du papier j, un lien de similarité est créé entre les deux auteurs A_i et A_j ssi $[B + C]_{ij} >= seuil$ (voir figure 4.4).

Expérimentations

La figure 2.4 montre quelques résultats visuels du graphe et de l'arbre couvrant obtenu, le tableau 2.2 montre les caractéristiques de graphe de similarité.

Nous présentons un exemple d'une requête : nous supposons que l'auteur "Francesco Masulli " soumet une requête composée des trois termes : <T_1= Ranking, T_2= Clustering, T_3= Data mining> : en appliquant l'algorithme exhaustif, les résultats mentionnés dans le tableau 2.3 montrent une liste d'auteurs recommandés triés selon leurs poids avec leur distance (nombre de liens) par rapport à l'auteur racine.

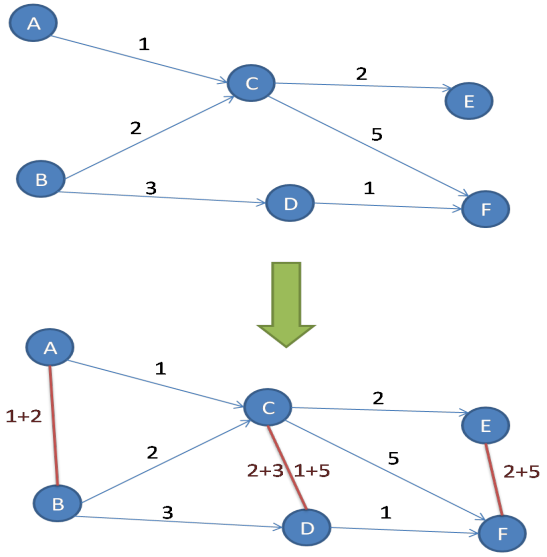


FIGURE 2.3 – Le graphe de citations (liens orientés bleus) et le graphe de similarité (lien non orientés rouges).

Nombre de nœuds	7065
Nombre de liens	1 009 940
Densité	0.04

Tableau 2.2 – Caractéristiques du graphe de similarité entre auteurs.

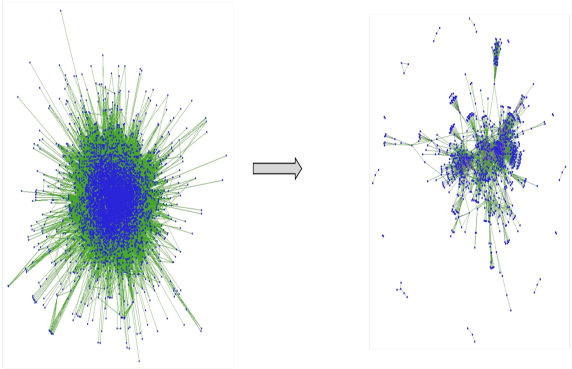


FIGURE 2.4 – Deux versions du graphe de similarité extraits du site microsoft, avec deux seuils différents (Kadima et Malek [65]).

Nous avons évalué la version guidée par rapport à la version exhaustive en procédant ainsi sur dix expériences : nous avons élaboré un ensemble de requêtes à tester par un auteur X (qui devient la racine de l'arbre couvrant) en utilisant les termes trouvés dans la communauté. Ensuite, pour chaque requête nous avons appliqué les deux versions de l'algorithme et relevé les mesures suivantes :

Le rang de l'auteur trouvé par l'algorithme guidé par rapport à l'algorithme exhaustif. Nous rappelons que la version exhaustive propose pour la même requête une liste triée d'auteurs à recommander.

Le nombre de sommets parcourus par l'algorithme guidé.

Auteur	Distance
Mikolaj Morzy	2
Steven Warner	2
Bob Garcia	2
Wendy Gersten	3
Manuel Lozano	2
Matthias Schonlau	2
Lyane T Watson	2
Carl Wunsch	2
Yang Seok Kim	2
David W Aha	2

Tableau 2.3 – Résultats de la requête <T_1= Ranking, T_2= Clustering, T_3= Data mining> : auteurs à recommander triés par leurs poids, la distance séparant chaque auteur recommandé de l'auteur racine est aussi calculée (Malek et Sulieman [89]).

Nous avons trouvé que pour 8 expériences sur 10 le rang numéro 1 a été trouvé par la version guidée tandis que pour les 2 expériences restantes le rang numéro 2 a été trouvé (voir tableau 2.4). L'arbre couvrant n'a pas été recherché en totalité par la version guidée, l'espace de recherche a été réduit de 11% à 49%.

2.3 Recommendation d'items avec taxonomie du domaine

L'approche de recommandation fondée sur la structure et le contenu a été généralisée dans Sulieman [128] en intégrant une taxonomie du domaine. Nous considérons un ensemble de données constitué d'un ensemble d'utilisateurs, un ensemble d'items ainsi que les interactions entre eux, ces interactions sont modélisées par un graphe bipartite.

L'objectif est de proposer des algorithmes de recommandation d'items à des utilisateurs en utilisant un réseau de collaboration entre utilisateurs. Ce réseau sera extrait à partir du réseau d'interaction (le graphe bipartite).

Ces algorithmes prennent en compte des informations topologiques extraites du graphe combinées avec des informations sémantiques extraites de la taxonomie du domaine (Sulieman *et al.* [129]).

Cette approche a été appliquée et validée sur deux domaines d'applications :

1. recommendation des livres sur Amazon,
2. recommendation de films (Movie lens dataset)

2.3.1 Extraction d'information topologique

Etant donnés deux ensembles : un ensemble d'utilisateurs U et un ensemble d'items I. Les éléments de U sont connectés aux éléments de I constituant un graphe bipartite G_{UI}, un lien entre l'utilisateur u et l'élément i signifie que l'utilisateur u a acheté et/ou évalué l'item i (le poids du lien étant la note en cas de notation). G_{UI} est un *graphe bipartite* (Cormen *et al.* [36]) défini sur l'ensemble des nœuds $U \cup I$

Requête	Algorithme exhaustif
	Auteurs à recommander
1	Andrew Emili
2	G V Belle
3	Hans A Kestler
4	Jimin Pei
5	John F Canny
6	C Wang
7	J Michael Brady
8	Peter G Neumann
9	Peter Eades
10	Liang Chen

Requête	Algorithme guidé	
	Auteur à recommander	**Graphe exploré**
1	Andrew Emili (1)	39.25%
2	G V Belle (1)	21.13%
3	*Yuichi Asahiro (2)*	13.86%
4	Jimin Pei (1)	20.02%
5	John F Canny (1)	11.77%
6	C Wang (1)	49.13%
7	J Michael Brady (1)	41.14%
8	*Elizabeth J O neil (2)*	24.88%
9	Peter Eades (1)	30.95%
10	Liang Chen (1)	16.67%

Tableau 2.4 – Comparaison entre la recherche en largeur et l'algorithme A* : chaque ligne est le résultat d'une requête envoyée par l'auteur racine. Les auteurs recommandés par A* sont indiqués avec leurs rangs trouvés par l'algorithme exhaustif. Nous remarquons que pour 8 requêtes sur 10 l'auteur trouvé par A* est celui ayant eu le rang numéro 1 par l'algorithme exhaustif (Malek et Sulieman [89]).

avec $U \cap I = \emptyset$. Pour chaque élément u dans U, on note $I(u)$ l'ensemble de tous les éléments de I tel que (u, i) est dans $E(G_{UI})$. Intuitivement, $I(u)$ est l'ensemble de tous les articles achetés et/ou notés par u.

Un graphe de collaboration entre utilisateurs (*users collaboration network*) est ensuite calculé à partir du graphe G_{UI} en effectuant une projection sur U.

Definition 1 (Graphe projeté) *Le* graphe projeté sur U *: $G = (V(G), E(G))$ défini à partir d'un graphe bipartite $G_{UI} = (U, I, E(G_{UI}))$ est un graphe non dirigé pondéré (Sulieman* et al. *[129]) :*

- *$V(G) = U$,*
- *(u, u') est une arête de $E(G)$ ssi $I(u) \cap I(u') \neq \emptyset$,*
- *$w(u, u')$ est le poids de (u, u'), est déterminé à partir du nombre et/ou ou des notations des items inclus dans $I(u) \cap I(u')$.*

Example 1 *La figure 2.5(a) montre un graphe bipartite utilisateurs-items $G_{UI} = (U, I, E(G_{UI}))$,dans lequel $V_{UI} = U \cup I$ où $U = \{A, B, C, D, E\}$ et $I = \{1, 2, 3, 4\}$*

sont respectivement l'ensemble des utilisateurs et l'ensemble d'items. La figures 2.5(a) et (b) montrent respectivement les projections utilisateurs et items associée à G_{UI}. *Noter que dans le graphe projeté sur U, le poids de chaque arête* (u, u') *noté* $w_U(u, u')$ *est le nombre d'items achetés et/ou évalué par* u *et* u'.

De même, dans le graphe projeté sur I, le poids de chaque arête(i, i'), *noté* $w_I(i, i')$ *correspond au nombre d'utilisateurs qui ont acheté et/ou évalué* i *et* i'.

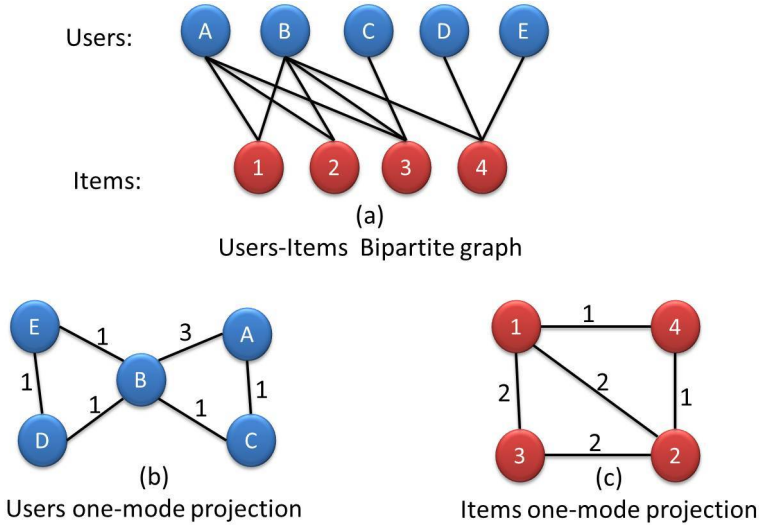


FIGURE 2.5 – (a) Exemple d'un graphe bipartit utilisteurs-items, (b) Graphe de collaboration entre utilisateurs, (c) graphe de projection sur les items (Sulieman *et al.* [129]).

Trois mesures de centralité sont utilisées dans les algorithmes de recommandation proposés à savoir la centralité du degré, la centralité de proximité et la centralité d'ntermédiarité, la combinaison de la centralité du degré avec la centralité de l'intermédiarité sera également étudiée.

En fait, étant donné l'une de ces mesures de centralité γ et un nombre positif n, les n sommets ayant les valeurs γ les plus élevées dans G sont considérés. Ces n top sommets constituent les sommets par lesquels sera débuté l'exploration du réseau pour effectuer la recommandation.

D'un autre côté, la fonction de pondération des arêtes du graphe projeté G est définie comme suit.

Definition 2 *Pour chaque arête* (u, u') *in* $E(G)$, *le* poids *de* (u, u'), *noté* $w(u, u')$ *est défini par (Sulieman* et al. *[129]) :*

$$w(u, u') = \frac{1}{|I(u) \cap I(u')|} \sum_{i \in I(u) \cap I(u')} (r(u, i) + r(u', i)).$$

Intuitivement, la définition 2 signifie que plus il y a d'items achetés et/ou notés par les deux utilisateurs u et u', plus ces deux utilisateurs sont connectés (plus le poids de l'arête (u, u') est élevé).

2.3.2 Extraction d'information sémantique

Les informations concernant les utilisateurs et les items sont définies à l'aide d'une taxonomie de domaine. Cette taxonomie est donnée sous forme d'arbre permettant au système de recommandation d'utiliser les informations relatives aux étiquettes et aux niveaux des nœuds (Sulieman *et al.* [129]).

Definition 3 (Taxonomie) *Soit T un arbre dans lequel les nœuds sont les termes du domaine, et les arêtes représentent la hiérarchie entre ces termes. On note $P(T)$ l'ensemble de toutes les paires de la forme (τ, λ), où τ est un terme de T et λ le niveau de τ en T. Si T a n niveaux, le terme de niveau 0 est le terme le plus général, tandis que les termes de niveau $n-1$ sont les termes les plus spécifiques du domaine. De plus, nous supposons que chaque élément traité dans le système de recommandation est associé à une feuille* unique *de T.*

Example 2 *La figure 2.6 représente partiellement une taxonomie définie pour les livres. Selon la définition 3 cette taxonomie est formalisée par :*

$$P(T) = \{(Thing, 0), (Book, 1), (HumanScience, 2), (Philosophy, 3), (JavaScript, 5), \ldots\}.$$

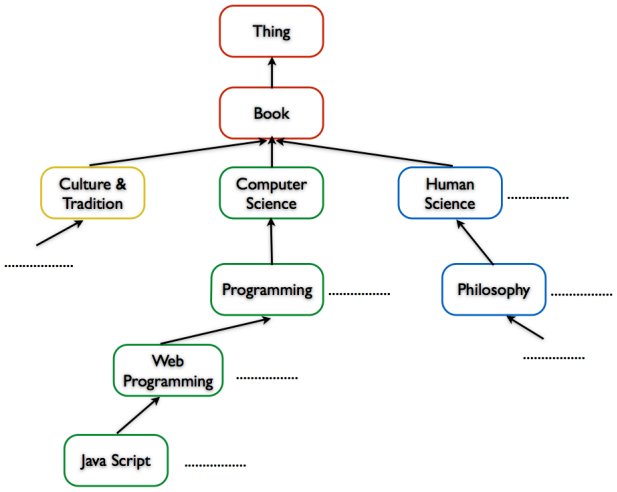


FIGURE 2.6 – Petit extrait de la taxonomie *livre*.

La taxonomie T et son ensemble de paires associé $P(T)$ sont utilisés pour associer des éléments et des utilisateurs à des ensembles de paires, appelés *profils* et définis comme suit.

Definition 4 (Profil d'item) *Soit T une taxonomie et $P(T)$ son ensemble de paires associé. Le* profil d'élément *d'un élément i, noté $IP(i)$ est l'ensemble des paires (λ, τ) dans $P(T)$ qui sont associées aux nœuds dans T qui constituent le chemin depuis la racine de T jusqu'à l'unique feuille de T associée à i.*

Example 3 *Dans le contexte de l'exemple 2, la figure 2.7 montre le chemin dans la taxonomie T concernant le livre i 'Java Script'. Selon la definition 4, $IP(x)$ est spécifié ainsi :*

$$\begin{aligned} IP(x) = \quad & \{(Thing, 0), (Book, 1), (ComputerScience, 2), \\ & (Programming, 3), (WebProgramming, 4), \\ & (JavaScript, 5)\}. \end{aligned}$$

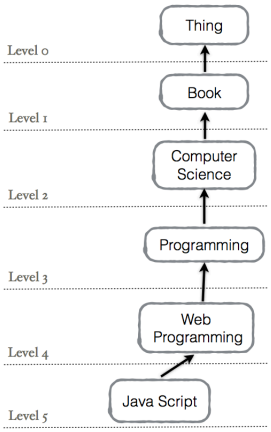


FIGURE 2.7 – Exemple du profil du livre *Java Script* (Sulieman *et al.* [129]).

De la même manière, chaque utilisateur peut être associé à un profil défini à l'aide de la taxonomie T et de son ensemble de paires associé $P(T)$. En supposant que chaque utilisateur u est associé à un ensemble d'items $I(u)$ (représentant les items que u a achetés ou évalués), la notion de profil utilisateur est définie ci-dessous.

Definition 5 (Profil d'utilisteur) *Soit T une taxonomie et $P(T)$ son ensemble de paires associé. Le* profil utilisateur *d'un utilisateur u, noté $UP(u)$, est l'union des profils d'élément de tous les éléments de $I(u)$, soit : $UP(u) = \bigcup_{x \in I(u)} IP(x)$.*

Example 4 *Revenant à l'exemple 2, la figure 2.8 montre un exemple de profil utilisateur. On peut voir sur cette figure que l'utilisateur u a acheté ou noté trois articles (livres) appelés $Book_i$ ($i = 1, 2, 3$) et dont les chemins correspondants dans T sont également illustrés dans la figure 2.8. Donc, en supposant que les profils de $Book_i$ ($i = 1, 2, 3$) ont été calculés, selon la définition 5, $UP(u)$ est donné par :*

$$UP(u) = IP(Book_1) \cup IP(Book_2) \cup IP(Book_3)$$

Sur la base des définitions précédentes, nous proposons une mesure de pertinence sémantique item-utilisateur pour estimer la proximité de l'item à recommander par rapport à un utilisateur donné. Pour ce faire, nous introduisons une mesure de similarité σ entre deux sous-ensembles P_1 et P_2 de $P(T)$ comme suit :

$$\sigma(P_1, P_2) = \frac{1}{\mu} \left(\sum_{(\tau, \lambda) \in P_1 \cap P_2} \lambda \right)$$

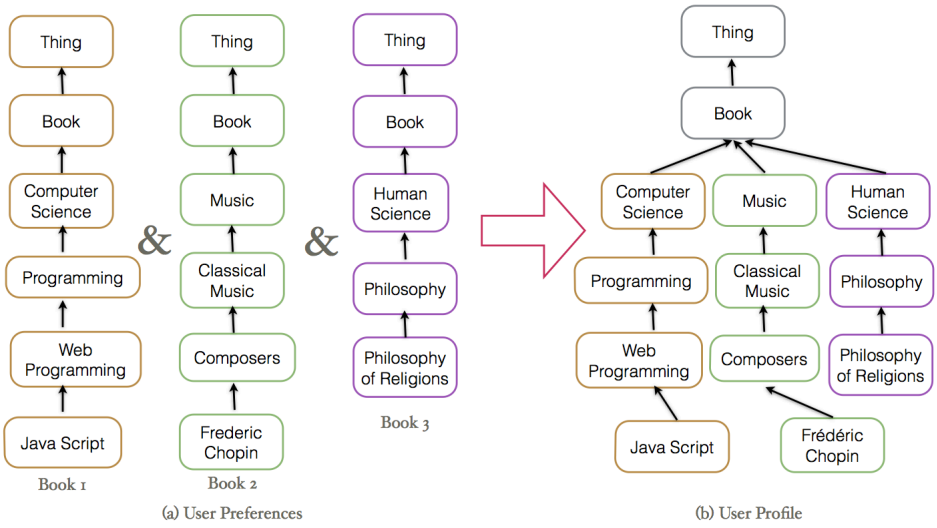


FIGURE 2.8 – Exemple d'un profil utilisateur (Sulieman *et al.* [129]).

avec $\mu = \min\left(\sum_{(\tau,\lambda)\in P_1} \lambda, \sum_{(\tau,\lambda)\in P_2} \lambda\right)$. En utilisant σ, nous définissons maintenant la pertinence entre un utilisateur u et un élément i par la similitude entre leurs profils.

Definition 6 *Soit u un utilisateur et i un élément. La mesure de pertinence de l'élément utilisateur entre u et i, notée $sim(u,i)$, est définie comme suit :*

$$sim(u,i) = \sigma(UP(u), IP(i))$$

Comme on peut le voir dans la définition 6, $sim(u,x)$ n'est pas une mesure de similarité standard car elle s'applique à deux arguments de types différents (à savoir un utilisateur et un item). En fait, $sim(u,i)$ quantifie le fait que des items *similaires* à i peuvent être trouvés dans $UP(u)$. Notez à cet égard que si u a acheté ou noté i, alors $IP(i)$ est un sous-ensemble de $UP(u)$, auquel cas $sim(u,i)$ est maximal, c'est-à-dire, égal à 1.

Example 5 *Concernant l'exemple 4 la figure 2.8 montre un exemple d'un utilisateur u qui a acheté ou évalué les items $Book_1$, $Book_2$ and $Book_3$.*

Soit l'item i ayant le profil :

$$IP(i) = \{(Books, 0), (ComputerScience, 1), (Programming, 2), (ObjectOriented, 3)\}$$

nous avons :

1. $UP(u) \cap IP(i) = \{(Books, 0), (ComputerScience, 1), (Programming, 2)\}$
2. $\sum_{(\tau,\lambda)\in UP(u)\cap IP(x)} \lambda = 3$
3. $\sum_{(\tau,\lambda)\in UP(u)} \lambda = 38$
4. $\sum_{(\tau,\lambda)\in IP(i)} \lambda = 6$

Le résultat est : $sim(u,i) = 3/6 = 0.5$.

2.3.3 Les algorithmes d'exploration

Dans cette section, deux algorithmes principaux qui explorent le réseau de collaboration des utilisateurs G, l'un en profondeur d'abord et l'autre en largeur d'abord (Cormen *et al.* [36]) sont décrits ; ces algorithmes sont appelés respectivement *social-sémantique* depth-*first search (SSDFS)* et *social-semantic* width-*first search (SSBFS)* (Sulieman *et al.* [129]).

L'objectif étant de ne pas explorer entièrement le graphe, des heuristiques basées sur les deux types d'informations suivants sont intégrés dans les algorithmes :

1. L'information topologique (appelée information sociale par Sulieman [128]), calculées sur le réseau de collaboration des utilisateurs (voir Définition 1), et constituée de :
   (a) La centralité de l'utilisateur, qui peut être la centralité de degré, la centralité de proximité ou la centralité d'intermédiarité.
   (b) Les liens de collaboration (ou liens sociaux) présentés par les poids des arêtes (voir Définition 2), notés $e.weight$ dans les algorithmes, pour chaque arête e dans $E(G)$.
2. L'information sémantique extraite de la taxonomie du domaine :
   (a) le profil de l'item (voir Definition 4),
   (b) le profil d'utilisateur (voir Definition 5),
   (c) la mesure de pertinence sémantique item-utilisteur (voir Definition 6).

Les algorithmes SSDFS et SSBFS sont détaillés dans Algorithm 1 et Algorithm 3, respectivement Ces deux algorithmes prennent comme paramètres d'entrée :
(i) le profil $IP(i)$ de l'item i à recommander,
(ii) un entier positif n qui correspond au nombre de sommets de départ de l'algorithme d'exploration.
(iii) un seuil de pertinence item-utilisateur δ,
(iv) un seuil de poids d'arête θ.
La sortie de chaque algorithme est une *liste d'utilisateurs* auxquels l'item i est à recommander.

Les algorithmes 1 et 3 suivent les étapes suivantes :

1. Chaque sommet est d'abord étiqueté à *non visité* et le la liste d'utilisateurs est initialisée à la liste vide.
2. La centralité de chaque sommet dans $V(G)$ est calculée et le vecteur ayant les N valeurs de centralité les plus élevées est construit.
3. Le graphe G est exploré en profondeur d'abord selon l'algorithme 1 et en largeur d'abord selon l'algorithme 3, en partant des *top* N sommets et en arrêtant de visiter les sommets dans un chemin lorsque soit le poids de l'arête est inférieur à θ ou la pertinence rentre le sommet courant (utilisateur courant) et l'élément i est inférieure à δ.
4. La liste des utilisateurs à recommander est retournée (variable *user_list*).

Rappelant que plusieurs mesures de centralité sont considérées, la centralité mentionnée dans l'algorithme 1 et l'algorithme 3 peut être soit la centralité de degré soit la centralité de la proximité soit la centralité de l'intermédiarité.

Algorithm 1 Algorithm SSDFS (Sulieman *et al.* [129]).

Input
(i) Item profile $IP(i)$
(ii) A positive integer n
(iii) A user-item similarity threshold θ
(iv) An edge weight threshold δ

1: **for all** vertices v in $V(G)$ **do**
2: $v.label = unvisited$
3: **end for**
4: $user\_list$ = empty list
5: Compute the centrality of every vertex in $V(G)$
6: Assign the vector N to the n vertices of $V(G)$ with the highest centrality
7: **for all** v in N **do**
8: Call Vertex-edge-based-visit(G, v)
9: **end for**
10: **Return** ($user\_list$)

Algorithm 2 SSDFS-visit

1: **if** $v.label = unvisited$ **then**
2: $v.label = visited$
3: **if** $sim(v, i) > \delta$ **then**
4: Add v to the current value of $user\_list$
5: **for all** $e = (v, v')$ or $e = (v', v)$ in $E(G)$ **do**
6: **if** $e.weight > \theta$ **then**
7: Call Vertex-edge-based-visit(G, v')
8: **end if**
9: **end for**
10: **end if**
11: **end if**

Algorithm 3 Algorithm SSBFS (Sulieman *et al.* [129])

Input
(i) An item i having $IP(i)$
(ii) A positive integer n
(iii) A user-item similarity threshold δ
(iv) An edge weight threshold θ

```
1: for all vertices v in V(G) do
2:    v.label = unvisited
3: end for
4: user_list = empty list
5: Compute the centrality of every vertex in V(G)
6: Assign the vector N to the n vertices of V(G) with the highest centrality
7: Q = empty queue
8: for all v in N do
9:    Enqueue(Q, v)
10: end for
11: while Q ≠ ∅ do
12:    v = Dequeue(Q)
13:    for all e = (v, v') or e = (v', v) in E(G) do
14:       if v'.label = unvisited then
15:          v'.label = visited
16:          if sim(v', i) > δ then
17:             if e.weight > θ then
18:                Add v' to the current value of user_list
19:                Enqueue(Q, v')
20:             end if
21:          end if
22:       end if
23:    end for
24: end while
25: Return (user_list)
```

De plus, une mesure de centralité *hybride* basée sur les centralités de degré et d'intermédiarité a été utilisée également. Dans ce cas, un nouveau seuil σ est introduit et les algorithmes sont modifiés comme suit :

— Les N sommets sont insérés selon l'ordre décroissant de leur degré de centralité ;

— Lors du test du poids d'une arête, une condition impliquant la mesure d'intermédiarité et son seuil associé σ est ajoutée. Plus précisément, la condition de l'algorithme 2 et 3 est changée en :
if $e.weight > \theta$ et $v'.betweenness > \sigma$ **then** .

2.3.4 Etude expérimentale

Dans cette section, nous rapportons les expériences menées afin de valider cette approche. Nous donnons d'abord les caractéristiques des ensembles de données sur

lesquels les algorithmes ont été testés et comparés avec les algorithmes classiques de la littérature à savoir l'algorithme collaboratif par item (Linden *et al.* [84], Sarwar *et al.* [121, 122]) et l'algorithme de recommandation hybride (Burke [28, 29]). Nous terminons la section avec des commentaires sur les résultats.

Les données

Deux ensembles de données pour tester cette approche sont utilisés : *MovieLens*[1] et *Amazon.com*[2].

collaboration Network	$\|V(G)\|$	$\|E(G)\|$	Dia	D	CC
MovieLens	999	46,689	4	0.094	0.436
Amazon.com	51,220	6,893,029	17	0.002	0.878

FIGURE 2.9 – Les caractéristiques des deux réseaux de collaboration : MovieLens et Amazon.com. $|V(G)|$ est le nombre de sommets, $|E(G)|$ est le nombre d'arêtes, Dia est le diamètre du réseau, D est la densité du réseau et CC est le coefficient du clustering.

Concernant les données *MovieLens*, le graphe bipartite utilisateur-film G_{UI} est composé de deux ensembles disjoints de sommets : l'ensemble des utilisateurs U et l'ensemble des films M (ce qui signifie que dans ce cas, les items sont des films). Un lien (u, m) connecte un utilisateur u de U avec un film u de M si u a évalué m.

Le réseau de collaboration des utilisateurs G est le réseau projeté sur les utilisateurs à partir du graphe G_{UI}. La figure 2.10 montre le réseau G (obtenu à l'aide du logiciel Gephi), la première ligne de tableau 2.9 donne les caractéristiques du réseau. La figure 2.11 montre la distribution en degrés de ce réseau.

La taxonomie des films basée sur les genres est celle proposée dans Schickel-Zuber [123]. Dans cette taxonomie T, les genres de films sont organisés selon une hiérarchie allant du niveau 1 (le concept le plus général décrivant les genres de films), jusqu'au niveau 4 (contenant les concepts les plus spécifiques décrivant les genres de films). La taxonomie T est représentée sur la figure 2.12.

Le deuxième ensemble de données est celui *d'Amazon* disponibles sur la page Web de l'Université de Stanford[3] :

Les sommets du graphe bipartite utilisateur-item G_{UI} sont divisés en deux ensembles disjoints, l'ensemble des utilisateurs U et l'ensemble des éléments I, une arête connecte un utilisateur u dans U avec un article i de I si u a acheté i. Ainsi, dans ce cas, nous ne considérons pas les notations.

Le réseau de collaboration (ou de co-achat) est le graphe projeté sur l'ensemble des utilisateurs extrait à partir du graphe. Les caractéristiques de ce réseau sont données dans la deuxième ligne du tableau 2.9.

La taxonomie des items est celle proposée avec l'ensemble des données Amazon.

1. http ://movielens.org
2. http ://snap.stanford.edu/data/amazon-meta.html
3. http ://snap.stanford.edu/data/amazon-meta.html

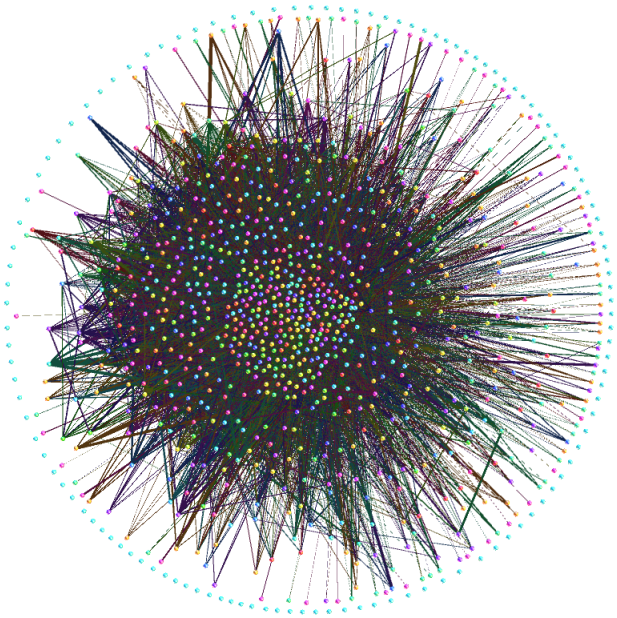


FIGURE 2.10 – Le réseau de collaboration MoviesLens extrait du graphe biparti utilisateur-film (Sulieman *et al.* [129]).

Evaluation des algorithme et métriques

Nous comparons notre approche avec deux algorithmes standards de la littérature à savoir l'algorithme de *filtrage collaboratif centré sur les items* et l 'algorithme de *recommandation hybride* (Adomavicius et Tuzhilin [2], Burke [28], Sarwar *et al.* [122]).

La mesure de similarité cosinus telle que définie ci-dessous est utilisée pour évaluer la similitude entre les items :

$$Cosine(u, u') = \frac{\sum_{i \in I} r(u,i).r(u',i)}{\sqrt{\sum_{i \in I} r(u,i)^2 . \sum_{i \in I} r(u',i)^2}} \tag{2.1}$$

u et u' sont deux utilisateurs et $r(u,i)$ (respectivement $r(u',i)$) est l'évaluation de u (respectivement u') de l'item i.

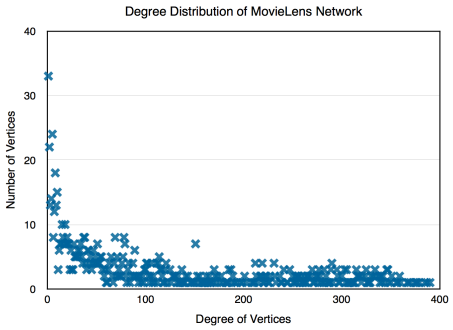


FIGURE 2.11 – Distribution des degrés du réseau de collaboration MovieLens (Sulieman *et al.* [129]).

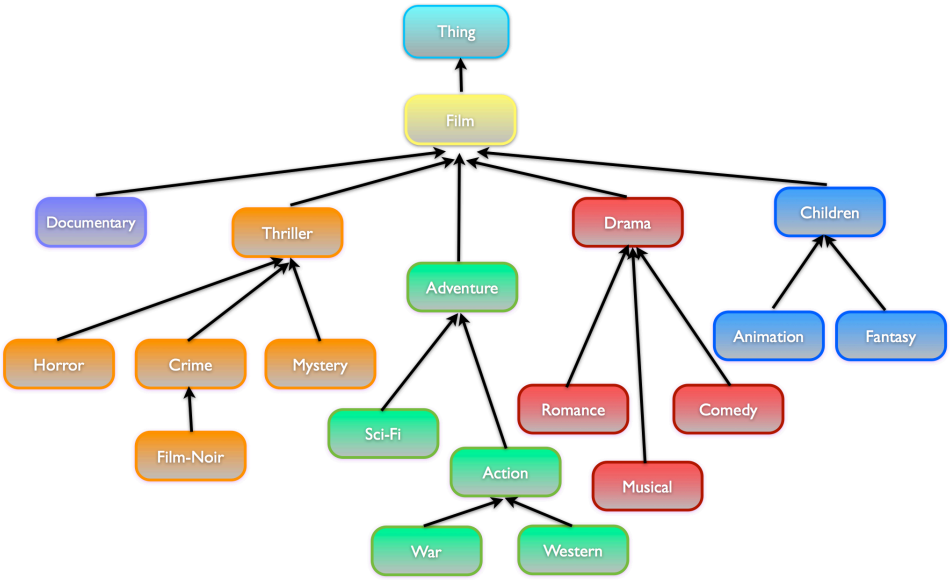


FIGURE 2.12 – Arborescence de la taxonomie *Film*.

Noter que concernant l'ensemble de données d'Amazon, comme les notes ne sont pas explicitement prises en compte, pour chaque u dans U et chaque i dans I, nous définissons $r(u, i) = 1$ si u acheté i et $r(u, i) = 0$ sinon.

L'approche est également comparée avec le système de recommandation hybride pondéré (Bennett et Lanning [20]). Il s'agit d'appliquer successivement l'algorithme de filtrage collaboratif basé sur les items avec l'utilisation de la similarité cosinus, et l'algorithme basé sur le contenu avec la mesure de pertinence sémantique item-utilisateur comme indiqué dans la definition 6 pour calculer la pertinence entre les utilisateurs et l'item d'entrée.

Des mesures d'évaluation standards comme la *precision*, le *rappel* et la *F-measure* (Herlocker *et al.* [57] sont aussi utilisées.

Cependant, afin de clarifier la signification de ces mesures dans le contexte des algorithmes de recommandation, nous rappelons la définition des notions standards de vrais et faux positifs ainsi que de celles de vrais et faux négatifs comme suit : considérant un ensemble de données Δ contenant un ensemble d'utilisateurs U et un ensemble d'éléments I, soit R une liste d'utilisateurs recommandés pour l'élément i (sortie d'algorithmes).

- Un utilisateur u dans Δ est dit vrai positif s'il a acheté et/ou noté i, et si u est dans R.
- Un utilisateur u dans Δ est dit faux positif s'il n'a pas acheté et/ou noté i, et si u est dans R.
- Un utilisateur u dans Δ est dit vrai négatif s'il n'a pas acheté et/ou noté i, et si u n'est pas dans R.
- Un utilisateur u dans Δ est dit faux négatif s'il a acheté et/ou noté i, et si u n'est pas dans R.

En notant TP (respectivement FP) le nombre de vrais (respectivement faux) positifs dans R et de même TN le nombre de vrais négatifs dans R, la précision, le

rappel et la F-mesure sont définis comme suit :

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F\text{-}Measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Afin d'évaluer les performances de calcul de l'approche nous considérons la *Couverture des données* qui correspond au pourcentage des *sommets du graphe* qui ont été visités pendant le calcul.

Résultats expérimentaux

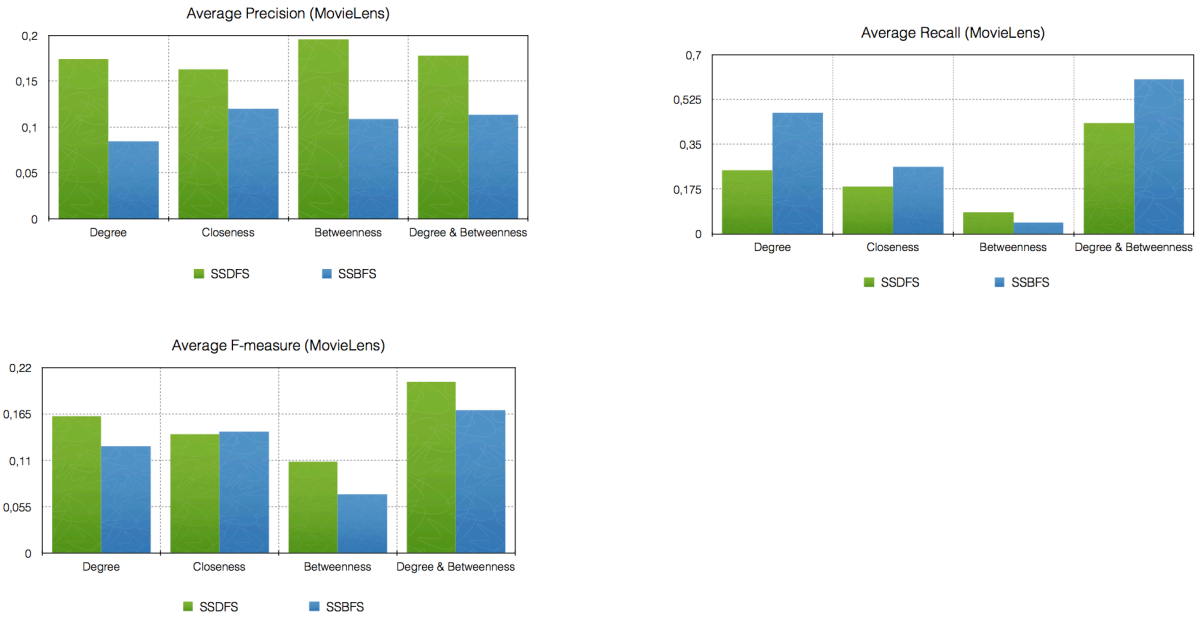


FIGURE 2.13 – Résultats des algorithmes algorithmes SSDFS, SSBFS sur les données de MovieLens : precision, rappel et F-score vs. mesures de centralité (Sulieman *et al.* [129]).

54 articles différents et 100 films ont été choisis respectivement dans le deux ensembles de données Amazon.com et MovieLens, les valeurs des paramètres : δ, θ, N ont été fixés pour chaque jeu de données après application des deux algorithmes SSDFS et SSBFS de façon à maximiser les mesures F tout en minimisant les couvertures des graphes projetés sur les utilisateurs.

La figure 2.13 montre les valeurs moyennes de la précision, du rappel et de F-measure pour les 100 requêtes soumises à la fois aux algorithmes SSDFS et SSBFS avec les mesures de centralité évoquées précédemment. On peut voir sur cette figure que l'algorithme SSDFS avec la centralité d'intermédiarité donne la meilleure valeur de précision, de même, l'algorithme SSDFS avec les centralités de degré et d'intermédiarité donne la meilleure valeur de rappel.

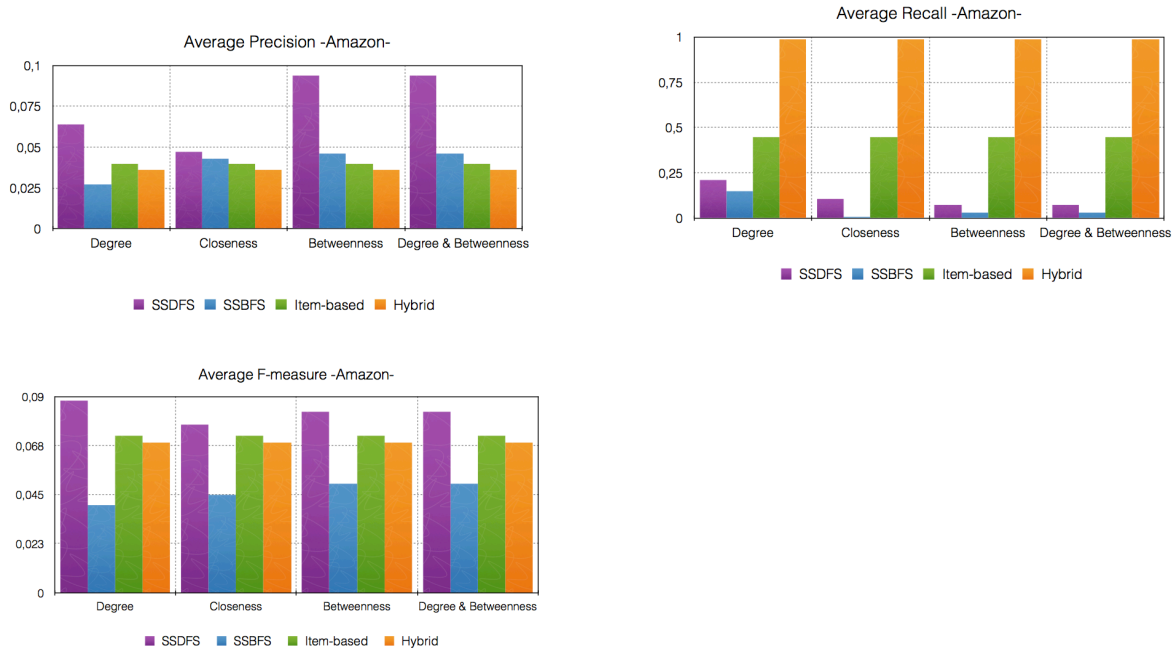


FIGURE 2.14 – Résultats des algorithmes SSDFS, SSBFS, Filtrage collaborative centré sur les items et algorithmes hybrides sur les données de Amazon.com : precision, rappel et F-score vs. mesures de centralité(Sulieman *et al.* [129]).

Concernant la mesure F, l'algorithme SSDFS avec la centralité combinée de degré et d'intermédiarité donne le meilleur résultat.

Concernant l'algorithme SSBFS, la figure 2.13, nous permet d'affirmer que des remarques similaires à celles de l'algorithme SSDFS sont valables concernant le rappel et la mesure F. Cependant, dans ce cas, la centralité de proximité donne la meilleure valeur de précision.

La figure 2.14 affiche les valeurs moyennes de précision, de rappel et de mesure F pour les 54 requêtes soumises aux algorithmes SSDFS, SSBFS, le filtrage collaboratif centré sur les items et l'algorithme hybride. Concernant l'algorithme SSDFS, l'utilisation des centralités de degré et d'intermédiarité donne la meilleure valeur de précision, de même, la centralité de degré donne la meilleure valeur de rappel. Nous soulignons également que l'algorithme SSDFS avec différentes centralités donne toujours une meilleure précision et des meilleures valeurs de mesure F que l'algorithme SSBFS et les algorithmes de recommandation centrés sur les items et l'algorithme hybride.

En ce qui concerne l'algorithme SSBFS, l'utilisation des centralités de degré et d'intermédiarité donne la meilleure précision, la meilleure centralité de degré et les meilleures valeurs de rappel et de F-measure. D'un autre côté, l'algorithme de recommandation hybride donne la meilleure valeur de rappel mais pourtant SSDFS a la meilleure mesure F.

En considérant maintenant les métriques de couverture du graphe, la figure 2.15 montre que, pour l'ensemble de données MovieLens, les algorithmes SSDFS et SSBFS explorent une petite partie du graphe, par rapport à l'algorithme de fil-

trage collaboratif centré sur les items et l'algorithme hybride. En effet :

— Concernant l'algorithme SS*D*FS, le rapport des sommets visités est compris entre 16% dans le cas de centralité d'intermédiarité et 53% dans le cas des centralités de degré et d'intermédiarités.

— Concernant l'algorithme SS*B*FS, ce rapport est de 86% dans le cas des centralités de degré et d'intermédiarité et de 92 % dans le cas de la centralité de degré.

Il est important de noter à cet égard que le filtrage collaboratif centré sur les items ainsi que les algorithmes de recommandation hybrides recherchent tout l'ensemble de données.

Des résultats similaires ont été également observés pour l'ensemble de données Amazon.com.

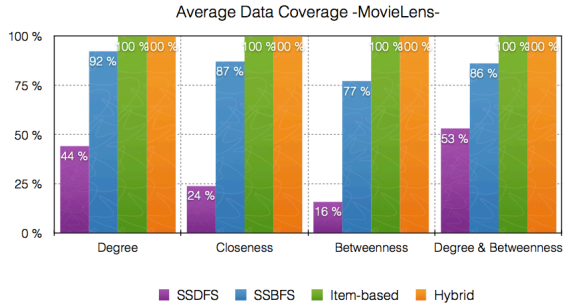


FIGURE 2.15 – Taux de couverture des données MovieLens.

Pour résumer les résultats expérimentaux, les points suivants sont à souligner :

1. Les quatre algorithmes affichent des valeurs de précision assez similaires, bien que le filtrage collaboratif centré sur les items et les algorithmes hybrides aient les meilleures valeurs de rappel (car ils visitent le graphe entier) et l'algorithme SS*D*FS donne les meilleures valeurs de mesure F.
2. Par rapport au filtrage collaboratif centré sur les items et aux algorithmes hybrides, les algorithmes SS*D*FS et SS*B*FS affichent des performances nettement meilleures en termes de temps d'exécution.
3. En général, SS*D*FS fonctionne mieux que SS*B*FS en termes de précision et d'efficacité.

2.4 Synthèse du chapitre

Nous avons présenté dans la première partie de ce chapitre un système de recommandation d'experts dans un réseau social professionnel. Notre réseau est composé d'un ensemble de personnes ayant des liens professionnels. Selon la demande d'un acteur d'origine X, le système doit proposer (recommander) un ou plusieurs acteurs $\{Z_1, Z_2, .., Z_n\}$ répondant au mieux que possible aux critères demandés. Nous avons appliqué notre algorithme sur la recommandation d'auteurs de papiers scientifiques dans un graphe de similarités entre auteurs.

La recommandation dépend d'un côté, de la valeur de pertinence entre les profils des auteurs et la requête soumise et d'un autre côte de l'intermédiarité des nœuds-auteurs se trouvant sur les chemins de la solution. Pour effectuer une recherche dans le graphe, l'arbre couvrant le plus représentatif est extrait et ensuite il est exploré.

Le premier algorithme est exhaustif, il est fondé sur la recherche en largeur dans l'arbre couvrant, jusqu'à ce qu'on trouve un auteur à recommander. Le deuxième algorithme utilise l'approche A* pour explorer l'arbre couvrant. Nous définissons une heuristique admissible qui dépend de la pertinence entre un profil et la requête ainsi de l'intermédiarité des nœuds-auteurs. Les expériences ont montré que la version guidée trouve souvent la meilleure recommandation tout en améliorant les performances de l'exploration. En comparant les deux algorithmes nous remarquons que l'arbre couvrant n'a pas été recherché en totalité par la version guidée, l'espace de recherche a été réduit de 11% à 49%.

Suite à cette première proposition, nous avons envisagé l'idée de l'intégration de l'ontologie de domaine dans le processus de la recommandation, le but étant d'utiliser cette ontologie dans l'élaboration des requêtes posées afin d'aider l'utilisateur à reformuler sa requête ou à la compléter ainsi que dans la représentation ontologique du profil utilisateur ou des ses préférences (Malek *et al.* [90]). Par conséquent, la mesure de pertinence entre la requête et le profil utilisateur doit intégrer les concepts ainsi que la structure de l'ontologie ce qui permettra d'avoir des recommandations plus précises.

Les travaux de thèse de Sulieman [128] ont traité ce sujet par la proposition d'une nouvelle approche de système de recommandation appliquée sur les réseaux de collaboration. Dans ce système, appelé système de recommandation social-sémantique, deux types d'information sont pris en compte : les informations extraites de la topologie du réseau comme les mesures de centralité et les poids des liens ainsi que les informations sémantiques qui sont définies à l'aide d'une taxonomie de domaine sur les items et qui ont permis de proposer une mesure de pertinence fine et précise entre élément et utilisateur.

Deux algorithmes principaux de recommandation fondés respectivement sur la recherche en profondeur et la recherche en largeur ont été proposés. Différentes mesures de centralité (degré, proximité, intermédiarité et hybride) ont été utilisées pendant l'exploration du graphe afin d'éviter de parcourir complètement le graphe (Sulieman *et al.* [129]). Ces algorithmes ont été comparés à deux algorithmes classiques de recommandation (filtrage collaboratif centré sur les items et recommandation hybride) en terme de précision (précision, rappel et mesure F) et d'efficacité (couverture du graphe et temps d'exécution). Les expériences montrent que les algorithmes proposés présentent des valeurs de précision similaires à celles des deux algorithmes standards, tout en les surpassant en termes d'efficacité puisqu'une partie restreinte du réseau est visitée.

Chapitre 3

Détection des communautés disjointes et chevauchantes

Sommaire

3.1 Introduction

La plupart des réseaux représentant des systèmes complexes contiennent des *communautés*. Une communauté est un groupe de nœuds fortement interconnectés entre eux mais faiblement liés au reste du graphe.

D'un point de vue général, les méthodes de clustering visent à synthétiser et résumer des observations (ou objets) en les regroupant de telle sorte que les objets

d'un même groupe (appelé cluster) soient plus similaires (dans un certain sens) les uns aux autres qu'à ceux d'autres groupes (ou clusters) (Jain *et al.* [61]). Plusieurs mesures de similarité (ou dissemblance) ont été proposées dans la littérature afin de tester la qualité d'une partition. Les techniques de clustering sont très diverses et elles ont été continuellement développées pendant plus d'un demi-siècle en se basant souvent sur les techniques d'optimisation. Ces algorithmes sont généralement classés en deux catégories principales : le clustering par partitionnement et le clustering hiérarchique.

Lorsque des objets sont connectés via un réseau représenté par un graphe, nous parlons des communautés. Une structure de communauté se compose de plusieurs nœuds qui montrent des connexions internes denses par rapport au reste du réseau. L'identification des communautés cachées dans la structure d'un grand réseau est un problème difficile qui a suscité un intérêt considérable.

Malgré l'ambiguïté de la définition de communauté, de nombreuses méthodes de detection de communautés ont été proposées. Des revues sur la détection de communautés disjointes sont présentées dans Danon *et al.* [38], Fortunato [44], Fortunato et Lancichinetti [45], Yang et Leskovec [145].

Cependant, certains nœuds peuvent appartenir à plusieurs communautés en même temps. Par exemple, une personne a généralement des liens avec plusieurs groupes sociaux comme la famille, les amis et les collègues ; un chercheur peut être actif dans plusieurs domaines. La problématique *de détection de communauté chevauchante* consiste à trouver donc un ensemble des communautés (Lancichinetti *et al.* [78]), dans lequel un nœud peut appartenir à plus d'une communauté.

Nous nous intéressons dans ce chapitre à la problématique de la détection de communautés disjointes et chevauchantes dans les réseaux complexes. Nous discutons dans la première partie la notion de la structure communautaire dans les graphes de terrain ainsi que les critères permettant de qualifier les communautés.

Nous présentons par la suite les différentes approches de détection de communautés disjointes dans un graphe de terrain ainsi que les mesures d'évaluation internes et externes proposées dans la littérature. Les mesures externes sont fondées sur la comparaison entre une partition réelle (observée et évaluée par des experts) et la partition trouvée par l'algorithme. Les mesures internes reposent sur des mesures permettant une estimation intrinsèque de la qualité de la partition.

Dans la troisième partie de ce chapitre, les différentes approches de detection de communautés chevauchantes sont présentées ainsi que les mesures d'évaluation internes et externes qui sont obtenues par adaptation des mesures proposées pour les communautés disjointes.

Dans la dernière partie, une méthode de détection des communautés chevauchante à partir de communautés disjointes pré-calculées en utilisant une version stable de propagation des labels (*appelée core detection label propagation*(CDLP) et décrite dans Attal et Malek [8]) est présentée. L'algorithme sélectionne les nœuds candidats pour le chevauchement et utilise *des fonctions d'appartenance* pour décider de l'affectation ou non d'un nœud candidat à chacune de ses communautés voisines. Ces fonctions d'appartenance sont soit fondées sur des mesures globales qui sont la densité et le coefficient de clustering (Attal *et al.* [9]) ou sur les moyennes

des mesures locales qui sont les centralités d'intermédiarité et de proximité. Nous comparons notre algorithme avec les algorithmes les plus utilisés en détection de communautés chevauchantes et qui sont considérés comme des algorithmes de référence pour leurs catégories (Attal *et al.* [10]). L'approche proposée donne des résultats relativement bons et compétitifs en terme de qualité mais dépendent de la topologie du réseau.

3.2 La structure communautaire

En 2002, Girvan et Newman [50] ont montré que la présence au sein de graphes sociaux de groupes de nœuds fortement connectés entre eux et faiblement avec le reste du graphe est une caractéristique des réseaux complexes, le nom de communautés a été donné à ces groupes de nœuds fortement connectés.

La structure communautaire est présente dans plusieurs systèmes de réseaux et plusieurs domaines comme la biologie, l'informatique, l'économie, la politique, etc.

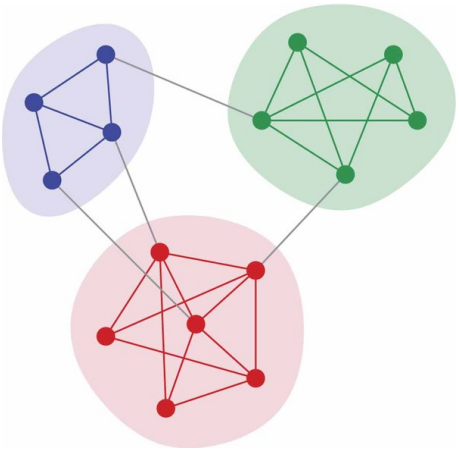


FIGURE 3.1 – Exemple d'une structure de communautés dans un graphe simple.

Les domaines d'application de la détection de communautés sont très variés (Benno Schwikowski1 et Fields [21, 22], Jeong et Al. [62], JF et Al. [63]). Par exemple, une interaction protéine-protéine nait lorsque deux ou plusieurs protéines se lient entre elles, le plus souvent pour mener à bien leurs fonctions biologiques. La détection de communautés peut ainsi aider à comprendre certains dysfonctionnements dans l'organisme.

De même, en sciences humaines, et plus particulièrement en sociologie et en anthropologie (branche des sciences qui étudie l'être humain sous tous ses aspects), la détection de communautés peut servir pour comprendre le comportement de groupes d'individus et la structure d'une société (Law et Hassard [82], Urry [134]).

Il existe également un lien important entre les communautés et la nature des nœuds. Les nœuds centraux des communautés partageant un nombre important de liens avec d'autres et ont ainsi une fonction importante de contrôle et de stabilité. Par ailleurs, les nœuds se trouvant aux frontières de ces communautés jouent un rôle

important pour la communication et l'échange d'information au sein du réseau.

La détection de communautés peut être très utile en calcul parallèle également. Elle permet de déterminer une meilleure façon de distribuer les charges sur les unités de calculs de telle sorte à ce que la communication soit minimale, le but étant d'améliorer la performance des calculs.

3.2.1 Notion de communauté dans un graphe

Selon Fortunato [44], il n'y a pas de définition précise de ce que l'on appelle *une communauté*. Intuitivement, une idée serait d'obtenir un sous-graphe d'un graphe initial tel que les sommets soient plus densément connectés qu'avec le reste du graphe.

Formulation du problème de détection de communautés disjointes

Considérons un réseau complexe représenté par un graphe G = (V,E), la problématique de la détection de communautés dans sa forme générale consiste à trouver une partition $P = (c_1, .., c_k)$ de l'ensemble des sommets V en k classes, de telle sorte que les sommets dans une communautés soient fortement connectés et faiblement avec le reste du graphe. Dans son article sur la détection de communautés, Fortunato (2010) propose une formulation de la problématique de la détection de communautés fondée sur la mesure de Mancoridis *et al.* [92] reposant sur la difference entre les ratios de connections internes et externes des groupes.

Ainsi, la problématique de la détection de communautés peut être vue comme une fonction à optimiser. En considérant un partition $C = (c_1.., c_k)$ en k parties de sommets disjoints, la mesure de Mancoridis est définie comme suit :

$$MQ = \frac{1}{k}\sum_i s(c_i, c_i) - \frac{1}{k(k-1)\sum_{i,j\neq i} s(c_i, c_j)} \tag{3.1}$$

avec $s(c_i, c_j) = \frac{|E(c_i,c_j)|}{|c_i||c_j|}$, $E(c_i, c_j)$ étant l'ensemble des liens présents à la fois dans la communauté c_i et c_j.

En posant $\Delta_{Int}(C) = \frac{1}{k}\sum_i s(c_i, c_i)$ et $\Delta_{Ext}(C) = \frac{1}{k(k-1)\sum_{i,j\neq i} s(c_i,c_j)}$, nous obtenons $MQ = \Delta_{Int}(C) - \Delta_{Ext}(C)$.

$\Delta_{Int}(C)$ représente la cohésion interne des groupes $c_1, ..., c_k$ (ou intra-classe) alors que $\Delta_{Ext}(C)$ représente la cohésion externe des groupes (ou inter-classes). Il s'agit de maximiser la somme de ratios sur l'ensemble des classes.

Fortunato [44] propose également des critères permettant de qualifier les communautés de *bonnes qualités* :

la réciprocité complète : les voisins de deux nœuds au sein d'une même communauté sont sensiblement les mêmes,

la joingnabilité : deux nœuds d'une même communauté doivent pouvoir être proches topologiquement l'un de l'autre,

le degré de sommet : les communautés doivent contenir des nœuds ayant un degré moyen important,

un fort coefficient de clustering au sein des communautés : une communauté doit contenir des triangles et des triades,

3.2.2 La modularité

La modularité est une mesure de qualité de partitionnement pour la détection de communauté créée par Newman et Girvan [108].

En considérant une partition P de l'ensemble des nœuds d'un graphe $G = (V, E)$, la modularité est une fonction prenant comme paramètre une partition $P = \{c_1, ..., c_k\}$ de communautés : $Q : P \to [-1, 1]$ définie par

$$Q(P) = \frac{1}{2m} \sum_i (A_{ij} - \frac{k_i k_j}{2m}) \delta(l_i, l_j)$$

avec A_{ij} la matrice d'adjacence, k_i le degré du nœud i, m le nombre de liens dans le graphe, l_i l'identifiant de la communauté à laquelle appartient le nœud i, l_j l'identifiant de la communauté à laquelle appartient le nœud j, et $\delta(l_i, l_j) = 1$ ssi les nœuds i et j sont dans la même communauté, sinon $\delta(l_i, l_j) = 0$.

Cette mesure est la somme, sur toutes les communautés des différences entre la proportion de liens à l'intérieur d'une communauté (ce qui équivaut à $\frac{A_{ij}}{2m}$) et la proportion de liens (soit $(\frac{k_i k_j}{2m})$) que devrait avoir une communauté dans un graphe aléatoire dont la distribution de degrés est la même que celui du graphe originel. Un tel graphe aléatoire est nommé le *modèle nul*. La modularité prend une valeur entre -1 et 1. On peut considérer qu'un graphe a une structure de communautés significative quand une partition obtient un score de modularité supérieur à 0.3. L'inconvénient de la modularité est sa limite de résolution. En effet, si l'on est confronté à des communautés de tailles différentes à l'intérieur d'un même graphe, certaines communautés, même bien définies, pourront ne pas être distinguées dans la partition de modularité optimale. On ne pourra pas détecter des communautés ayant une taille inférieure à $\sqrt{m}$, m étant le nombre d'arêtes.

De nombreux algorithmes de détection de communautés visent à maximiser cette métrique.

3.3 Detection des communautés disjointes

3.3.1 Approches de detection des communautés disjointes

Nous présentons les approches principales de la detection des communautés selon la littérature (Attal [7]) :

Approches divisives

Les approches divisives consistent à considérer la topologie entière du graphe et à effectuer une coupe pour obtenir un partitionnement. Une coupe d'un graphe est une partition des sommets en deux sous-ensembles. Les coupes ont lieu sur des liens connectant des régions denses du graphe. La méthode la plus connue est celle de la bissection (Fiedler [42] et Pothen *et al.* [112]) qui coupe le graphe en deux, puis opère de manière itérative sur les sous-graphes résultants.

Nous présentations dans la suite l'approche spectrale ainsi que l'approche utilisant la centralité d'intermédiarité.

Approche spectrale La méthode spectrale, issue de l'algèbre linéaire, établit notamment l'existence d'une base orthonormale de vecteurs propres pour tout endomorphisme symétrique sur un espace vectoriel complexe de dimension finie. Elle consiste en l'étude de matrices particulières, portant notamment sur les vecteurs propres de matrices définies positives.

Cela induit certaines propriétés comme la positivité des valeurs propres, que l'on peut ordonner de la manière suivante $\lambda_1 = 0 \leq \lambda_2 \leq ... \leq \lambda_n$. On peut également par des méthodes algébriques ou numériques, calculer les valeurs et les vecteurs propres (Lanczos [81]). Une étude théorique de la méthode spectrale a été traitée par Chung [31] et Von Luxburg [136].

Dans l'analyse spectrale des réseaux, la matrice Laplacienne est calculée ainsi :

$$L = D - A \tag{3.2}$$

où D est la matrice diagonale des degrés et A est la matrice d'adjacence.

La matrice Laplacienne permet l'obtention d'information sur la topologie du graphe ainsi que d'affirmer la possibilité d'effectuer une partition $P = \{P_1, P_2\}$ du graphe G selon un réel r. L'une des premières méthodes exploitant ce champ a été la méthode de Barnes et Hoffman [16] suivi d'autres études, nous citons : Pothen *et al.* [112] Barnard et Simon [15] Barnard [14].

Shi et Malik [127] et Ng *et al.* [109] ont eu l'idée d'utiliser l'information stockée dans d'autres vecteurs propres pour améliorer la qualité de partitionnement. L'idée consistait à utiliser l'algorithme k-means dans l'espace propre afin de trouver k clusters et par transposition sur le graphe, k communautés.

Donetti et Muñoz [41] ont proposé d'utiliser les vecteurs propres associés aux K plus petites valeurs propres non-nulles, K étant un entier ne pouvant excéder le nombre de valeurs propres. A chaque itération de l'algorithme, K est incrémenté de un et un algorithme de clustering hiérarchique est appliqué. Le meilleur partitionnement du dendrogramme est celui qui a la plus grande modularité.

Dans Newman [106] et Newman [107], la matrice de modularité a été considérée pour y appliquer la méthode spectrale, tout en effectuant des méthodes de raffinage (où des nœuds changent dynamiquement de communautés).

Approche par centralité d'intermédiarité L'un des premiers algorithmes modernes pour la détection de communautés fut proposé par Girvan et Newman [50].

Girvan et Newman étendent le calcul de la centralité d'intermédiarité (pondéré) aux arêtes d'un graphe. Si un lien se trouve fréquemment sur les plus courts chemins entre les nœuds du graphe, alors il est naturel de penser qu'il ne fait pas partie d'une communauté mais plutôt qu'il relie des communautés distinctes. En retirant progressivement le lien qui a la plus forte centralité d'intermédiarité, on obtient un découpage en blocs du réseau, les composantes connexes restantes sont les communautés on obtient ainsi une décomposition hiérarchique du réseau.

Approches agglomératives et multi-niveaux

Il s'agit des approches ascendantes qui considère initialement chaque noeud unique comme une communauté, une méthodes itérative est appliquée par la suite pour fusionner les communautés selon un critère de qualité.

En 2008, Blondel *et al.* [24] proposent de fusionner localement un nœud avec le voisin dont le résultat augmentera le plus une fonction de qualité, en l'occurence la modularité. Le processus (cf Algorithme 4) se poursuit de manière récursive sur le graphe résultant à chaque nouvelle fusion jusqu'à ce qu'il n'y ait plus d'augmentation de la modularité. Il s'agit de la version locale de la méthode de Clauset *et al.* [32].

Algorithm 4 L'algorithme de Louvain.

Entrée : Un graphe $G = (V, E)$

1: **A répéter jusqu'à l'obtention d'un score local optimal**
2: **Phase 1** : partitionnement du réseau de manière gloutonne en utilisant la modularité
3: **1)** Assigner à chaque nœud une communauté spécifique
4: **2)** Pour chaque nœud i du réseau
   — Pour chaque voisin j de i, choisir le voisin pour lequel l'assignation du nœud i dans une communauté augmenterait le plus la modularité
   — Répéter le processus jusqu'à ce qu'il n'y ait plus d'amélioration de modularité
5: **Phase 2** : agglomération des sous-graphes en nouveaux nœuds
6: **1)** Chaque communauté C_i est considéré comme un nouveau nœud i
7: **2)** Les arêtes entre les nouveaux nœuds i et j sont la réunion des arêtes entre les nœuds appartenant respectivement à C_i et C_j au sein du graphe précédent

La méthode permet de produire des communautés de bonne qualité sur de petits réseaux ainsi que l'obtention de dendrogrammes. Cependant, pour de grands graphes (de plusieurs millions de nœuds et d'arêtes), l'optimisation d'une mesure globale conduit à une propagation d'erreur.

De ce fait, la qualité risque de se détériorer en fonction de la taille des réseaux. L'algorithme agissant localement, la méthode est instable, ne produisant jamais le même résultat d'un lancement à l'autre. Il a été montré que l'ordre jouait un rôle important sur la qualité des communautés détectées. La figure 3.2 montre un exemple de fonctionnement de la méthode de Louvain.

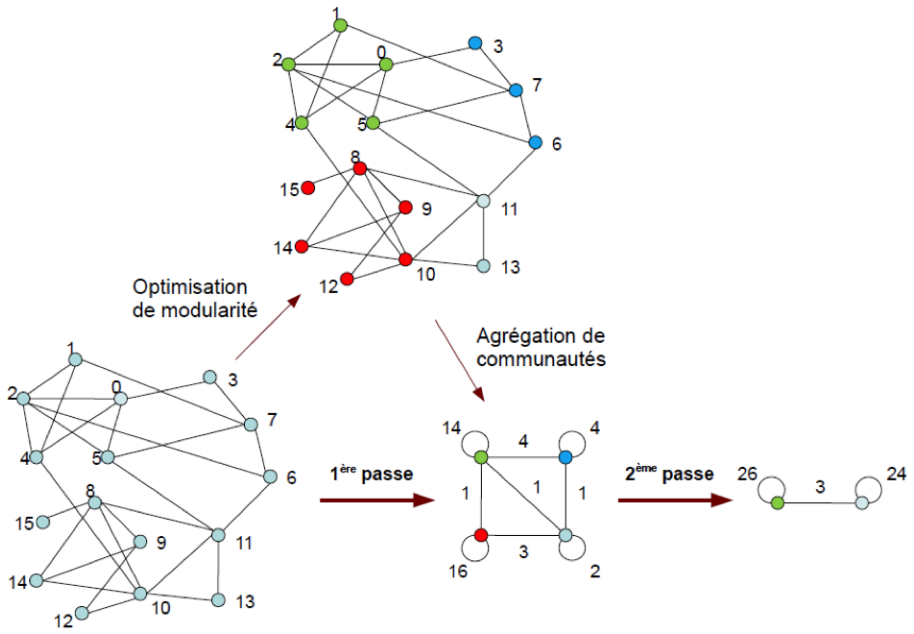


FIGURE 3.2 – Exemple d'application de la méthode de Louvain sur un graphe à 16 sommets (Extrait de Blondel *et al.* [24]).

Approches fondées sur la détection de leaders

Dans un réseau, certains nœuds peuvent être considérés plus importants que d'autres. Cela peut s'exprimer par une mesure de centralité ou un cœfficient de clustering plus fort pour certains nœuds du graphe par rapport à d'autres. C'est une des caractéristiques des graphes de terrains avec des nœuds situés au centre de structures communautaires et d'autres nœuds qui leur sont liés. De nombreux algorithmes ont ainsi proposé de détecter les nœuds les plus importants que l'on peut qualifier de leaders et d'attribuer les autres nœuds à ces derniers pour former des communautés (Shah et Zaman [125] Kanawati [66]).

Approches fondées sur la perturbation du réseau

Les méthodes de perturbation des réseaux ont été créées pour l'amélioration d'algorithmes déterministes. En effectuant une modification topologique du graphe, certaines structures communautaires peuvent être plus facilement détectées.

Gfeller *et al.* [48] proposent de construire une suite de graphes G_1,G_2,...,G_n en modifiant la pondération des arêtes pour chaque couple de nœuds (x, y) via une loi de distribution uniforme. L'idée est que si des communautés existent au sein d'un graphe, une faible modification topologique du graphe en utilisant une nouvelle pondération des arêtes ne devrait pas modifier la structure des communautés détectées.

Karrer *et al.* [69] ont proposé un algorithme perturbatif qui enlève une certaine portion d'arêtes α et qui la remet entre certaines paires de sommets (x, y) avec une probabilité $\frac{d(x)d(y)}{2m}$, que l'on retrouve dans le modèle nul. L'objectif est de conserver la distribution des degrés des nœuds au sein du réseau.

Rosvall et Bergstrom [118] ont proposé *Infomod*. Il s'agit d'une méthode fondée sur la théorie de l'information, plus exactement sur la quantité d'information qu'une partition a vis-à-vis du graphe originel.

Rosvall et Bergstrom [119] ont proposé l'algorithme *Infomap*. Il s'agit d'une méthode fondée sur la théorie de l'information avec un processus d'encodage et sur la marche aléatoire. Le procédé d'Infomap consiste à trouver des structures communautaires en étudiant la marche aléatoire et la ressemblance de motifs.

De Meo *et al.* [39] ont proposé CONCLUDE (pour Complex Network Cluster Detection). L'algorithme utilise à la fois l'importance des arêtes du graphe et un algorithme de clustering qui est appliquée suite à une projection des nœuds dans un espace euclidien pour trouver les communautés.

Approche par propagation de labels

La méthode de propagation de labels (Raghavan *et al.* [113]), notée LPA, est basée sur la transmission d'un label d'un nœud à ses voisins. Un état d'équilibre est atteint lorsque chaque nœud a son label égal à celui de la majorité de ses voisins.

A chaque étape, chaque nœud met à jour son label selon les labels de ses voisins, en utilisant un vote. Le label du nœud u prendra le label majoritaire de ses voisins. En notant c_u le label du nœud u, et par $N^l(u)$ l'ensemble du voisinage du nœud u avec le label l, l'affectation d'un label au nœud u est donnée par la formule suivante :

$$c_u = \arg\max_l |N^l(u)| \tag{3.3}$$

A la fin du processus, les nœuds ayant le même label représentent une communauté. Cette méthode peut être effectuée de manière *synchrone* ou *asynchrone*. La méthode asynchrone signifie que la mise à jour d'un label d'un nœud est connue par tous les autres nœuds du graphe immédiatement. Son label est transmis pour la mise à jour des labels des autres nœuds. Ce n'est pas le cas du mode synchrone, où la mise à jour des labels utilise les labels des nœuds à la précédente propagation. La complexité de cet algorithme que cela soit en mode synchrone ou asynchrone est en $\mathcal{O}(k \times (n+m))$, où $k \in \mathbb{N}$ représente le nombre d'itérations de l'algorithme, spécifié par l'utilisateur, n étant le nombre de nœuds et m celui des arêtes.

Cependant, l'algorithme de propagation de labels présente l'inconvénient d'être instable, ne donnant que rarement le même résultat après plusieurs lancements. Il est aussi caractérisé par un problème intrinsèque conduisant dans certains cas à de très grandes communautés (monstres) (Attal [7]).

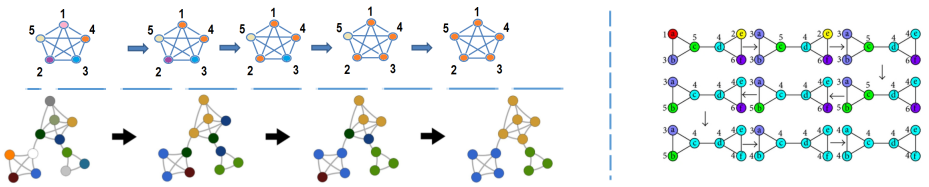


FIGURE 3.3 – Exemples de propagation de labels (Attal [7]).

L'algorithme est non déterministe et souffre d'une forte instabilité. En considérant l'exemple du club de Karaté (Zachary [146]), on peut s'apercevoir sur la Figure

3.4 que le nombre de communautés change selon les différentes propagations de labels.

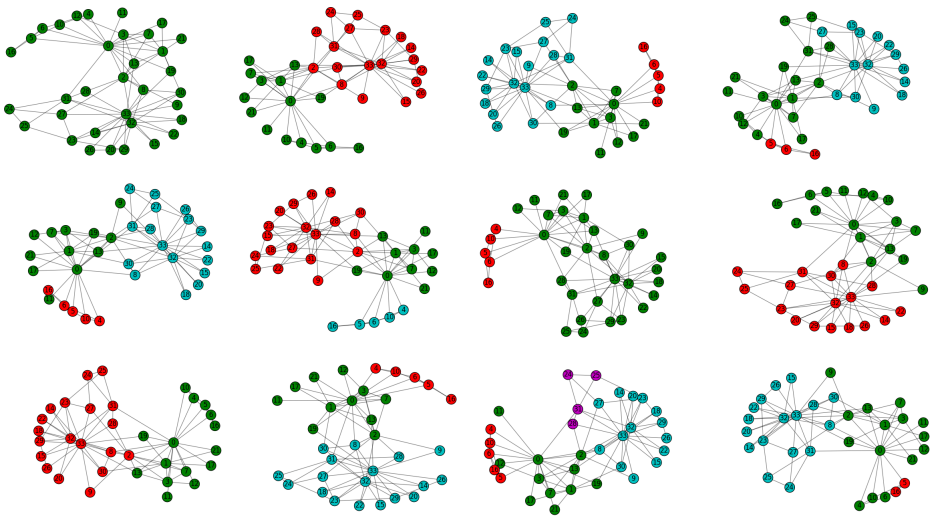


FIGURE 3.4 – La propagation de labels est un algorithme non déterministe et instable (Attal [7]). Application sur le graphe de Karaté (Zachary [146]).

Pour résoudre le problème de l'équidistribution des labels majoritaires pour le processus de vote, Xie et Szymanski [143] ont proposé une propagation de labels avec opérateurs afin de stabiliser et de rendre déterministe l'algorithme de propagation de labels (LPA) tout en améliorant la qualité de partitionnement. Cela consiste à stocker, propager et trier les labels de chaque nœud en utilisant quatre opérateurs qui sont la *propagation*, *l'inflation*, *la coupure* et une phase de mise à jour de stabilisation du LPA.

Zhang *et al.* [147] ont proposé une version modifiée du LPA avec la capacité à la prédiction de transition de percolation (LPAp). Les effets de la phase de prédiction au sein du LPA permettent de retarder la formation de communautés géantes.

3.3.2 Mesures externes et internes pour l'évaluation des communautés disjointes

Pour apprécier la qualité de partitionnement obtenu par les algorithmes de détection de communautés, nous utilisons deux types de mesures : les mesures externes et internes (Attal [7]).

1. Les mesures externes sont fondées sur la comparaison entre une partition réelle (observée et évaluée par des experts) et la partition trouvée par l'algorithme.

2. Les mesures internes permettent d'avoir une estimation de la qualité de partitionnement. Ces mesures sont principalement basées sur les densités et les liens entre communautés.

Mesures externes pour la détection de communautés disjointes

Nous présentons trois mesures : l'indice de Rand (et sa version ajustée), l'information mutuelle normalisée et la pureté.

L'indice de Rand et la version ajustée (Rand [114])
Considérons deux partitions P_1 et P_2 d'un même ensemble $S \subseteq V$. Pour évaluer la ressemblance entre deux partitions nous mesurons le taux de bonnes assignations de ces paires d'observations. Pour chaque couple d'observations, on peut compter quatre types d'assignations possibles :

— N_{11} le nombre de paires de nœuds classées ensemble selon P_1 et P_2
— N_{10} le nombre de paires de nœuds classées ensemble selon P_1 et séparées selon P_2
— N_{01} le nombre de paires de nœuds séparées selon P_1 et classées ensemble selon P_2
— N_{00} le nombre de paires de nœuds séparées à la fois dans P_1 et dans P_2

L'indice de Rand est donné par la formule suivante :

$$Rand(P_1, P_2) = \frac{N_{11} + N_{00}}{N_{11} + N_{10} + N_{01} + N_{00}} \tag{3.4}$$

Cette mesure varie entre 0 et 1 et prend la valeur maximale en cas de parfaite correspondance. *L'indice de Rand ajusté* (ARI) (Hubert et Arabie [59]) a été proposé dans l'objectif d'avoir une espérance nulle lorsque les partitions sont tirées aléatoirement :

$$ARI(P_1, P_2) = \frac{2(N_{11}N_{00} - N_{01}N_{10})}{(N_{01} + N_{00})(N_{11} + N_{01}) + (N_{00} + N_{10})(N_{10} + N_{11})} \tag{3.5}$$

$$\tag{3.6}$$

Sa valeur est comprise entre 0 et 1. Elle prend la valeur 1 lorsque les deux partitions sont identiques. Si la valeur est proche de 0, les deux partitions sont très différentes.

La pureté (Manning *et al.* [93]) La mesure de la pureté d'une communauté $c_i \in P_1$, avec $P_1 = \{c_1, ..., c_m\}$, par rapport à une partition donnée P_2 (avec $P_2 = \{c'_1, ..., c'_n\}$) est définie par :

$$purete(c_i, P_2) = \max_{1 \leq j \leq n} \frac{|c_i \cap c'_j|}{|c_i|} \tag{3.7}$$

Cette fonction calcule le taux de recouvrement maximal entre la communauté c_i et les communautés se trouvant dans la partition P_2. On définit la pureté de la partition P_1 par rapport à la partition P_2 par la somme pondérée de la pureté des communautés de P_1 par rapport à P_2 :

$$purete(P_1, P_2) = \sum_{i=1}^{m} w_{c_i} \times purete(c_i, P_2) \tag{3.8}$$

où $w_{c_i} = \frac{|c_i|}{\sum_{i=1}^{m} |c_i|}$

L'information mutuelle normalisée (NMI) (Kullback [77]) : est une mesure qui permet de représenter le degré de dépendance entre deux partitions P_1 et P_2. Elle est basée sur la théorie de l'information. La probabilité qu'un nœud choisi au hasard dans une partition P_1 appartienne à la communauté k est $P(k) = \frac{n_k}{n}$ où n_k est le nombre de nœuds dans la communauté k et n est le nombre total de nœuds. L'entropie de Shannon est définie comme : $H(P_1) = -\sum_{k=1}^{|P_1|} \frac{n_k}{n} log_2 \frac{n_k}{n}$, où $|P_1|$ est le nombre de communautés dans la partition P_1. L'entropie de Shannon de la partition P_1 représente la quantité d'information fournie par cette même partition sur la topologie du graphe en termes de communautés.

L'information mutuelle $I(P_1, P_2)$ évalue le niveau d'inter-dépendance entre deux partitions d'un graphe. L'information mutuelle est $I(P_1, P_2) = \sum_{i=1}^{P_1} \sum_{j=1}^{P_2} \frac{n_{ij}}{n} log_2(\frac{n_{ij}}{n_i n_j})$ où n_i est le nombre de nœuds de la communauté i de la partition P_1, n_j est le nombre de nœuds de la communauté j de la partition P_2 et n_{ij} est le nombre des nœuds en commun entre les deux communautés i et j. La variation de l'information $V(P_1, P_2)$ est $V(P_1, P_2) = H(P_1) + H(P_2) - 2I(P_1, P_2)$. L'information mutuelle normalisée est définie par : $NMI(P_1, P_2) = \frac{I(P_1,P_2)}{\sqrt{H(P_1)+H(P_2)}}$. Sa valeur peut varier entre 0 et 1. Plus la valeur est proche de 1, plus les deux partitions sont identiques.

Le score F_1 (van Rijsbergen [135]), couramment utilisé dans la recherche d'information, peut être vu comme une mesure d'exactitude de la précision et du rappel entre deux partitions. En considérant deux communautés c et c', le score F_1 est définit par :

$$F_1(c, c') = 2\frac{|c \cap c'|}{|c| + |c'|} \tag{3.9}$$

Plus les communautés c et c' partagent de nœuds en commun, plus le score F_1 est proche de 1. Un score nul signifie que les deux communautés ne partagent aucun nœud en commun. Cette méthode est fondée sur la combinaison de la précision ($precision(c, c') = \frac{|c \cap c'|}{|c|}$) et du rappel ($rappel(c, c') = \frac{|c_1 \cap c'|}{|c'|}$).

Mesures internes pour la détection de communautés disjointes

Les mesures internes permettent d'avoir une estimation de la qualité intrinsèque de partitionnement sans prendre en compte les véritables partitions. Ces mesures sont soit basées sur des structures aléatoires respectant la topologie du graphe initial, soit basées sur les densités et les liens entre communautés. Nous citons la modularité, la mesure de Mancoridis *et al.* [92] (voir sections 3.2.1, 3.2.2) et la conductance.

La conductance (Kannan *et al.* [68]) est fondée sur la densité des communautés et le nombre de liens sortant de celles-ci. Une structure communautaire est supposée avoir beaucoup de liens en son sein et un nombre faible de liens sortants.

Soit c une communauté d'un graphe G, la conductance de cette communauté est définie par $\varphi_{(c,G)} = \frac{l^c_{out}}{[2l^c_{int}+l^c_{out}]}$, l^c_{out} étant le nombre de liens et l^c_{int} celui de liens intérieurs à la communauté c.

En considérant une partition $P = \{c_1, ..., c_k\}$ en k parties de nœuds disjoints, la conductance de G se définit comme suit :

$$\Phi_G = \frac{1}{k}[\sum_{c=1}^{k} \varphi_{(c,G)}] \tag{3.10}$$

$$= \frac{1}{k}[\sum_{c=1}^{k} \frac{l^c_{out}}{[2l^c_{int} + l^c_{out}]}] \tag{3.11}$$

La conductance peut avoir une valeur comprise entre 0 et 1.

3.4 Détection de communautés chevauchantes

Il est possible que certains nœuds appartiennent à plusieurs communautés.

En considérant un graphe $G = (V, E)$, la détection de communautés chevauchantes consiste à trouver des groupes de nœuds fortement connectés entre eux et faiblement avec le reste du graphe, avec des nœuds pouvant appartenir à plusieurs communautés. Plus formellement, il s'agit de trouver une *couverture* (Lancichinetti *et al.* [79]) $C = \{c_1, ..., c_k\}$, k n'étant pas à spécifier avec $c_i \bigcap c_j = \emptyset$ ou $c_i \bigcap c_j \neq \emptyset$, et $\cup_{i=1}^{k} c_i = V$.

De nombreuses études précédentes (Gregory [53], Kelley [70], Lancichinetti *et al.* [79], Lee *et al.* [83], Reichardt et Bornholdt [116], Sales-Pardo *et al.* [120], Wang *et al.* [137], Xie *et al.* [142], Hajiabadi *et al.* [56], Huang *et al.* [58]) ont montré que le chevauchement est une caractéristique importante dans des nombreux réseaux complexes du monde réel.

3.4.1 Approches de détection de communautés chevauchantes

Xie *et al.* [142] classent les algorithmes de détection des communautés en cinq classes selon la façon dont les communautés sont identifiées à savoir : les algorithmes de clique de percolation, le partitionnement des liens, l'expansion et l'optimisation locales, la détection floue et les algorithmes dynamiques et fondés sur les agents (Attal [7], Attal *et al.* [10]).

Les algorithmes de percolation de cliques

Les algorithmes de percolation de cliques (CPM) sont basés sur l'hypothèse qu'une communauté consiste en des ensembles de sous-graphes entièrement connectés qui se chevauchent. La détection des communautés se fait donc en cherchant des cliques adjacentes. Tout d'abord, toutes les cliques de taille k sont identifiées. Ensuite, un nouveau graphe est construit de telle sorte que chaque nœud représente l'une de ces k-cliques. Les algorithmes CPM sont adaptés pour les réseaux qui ont des parties connectées denses.

Palla *et al.* [111] ont proposé *Cfinder*, le premier algorithme pour la détection de communautés chevauchantes fondé sur la recherche de motifs locaux par percolation de cliques (CPM : Clique Percolation Method). Les auteurs observent qu'une communauté peut-être définie comme une chaîne de k-cliques adjacentes. Cette méthode

permet la détection de communautés couvrantes où un sommet peut appartenir à plusieurs k-cliques.

Les expérimentations menées pas Palla *et al.* [111] ont montré que la valeur $k = 4$ donnait les résultats en termes de qualité de partitionnement les plus probants. Selon la paramétrisation de k, Cette méthode peut être appliquée à des graphes de plusieurs centaines de milliers d'arêtes. La Figure 3.5 montre un exemple extrait du site de Palla *et al.* [111].

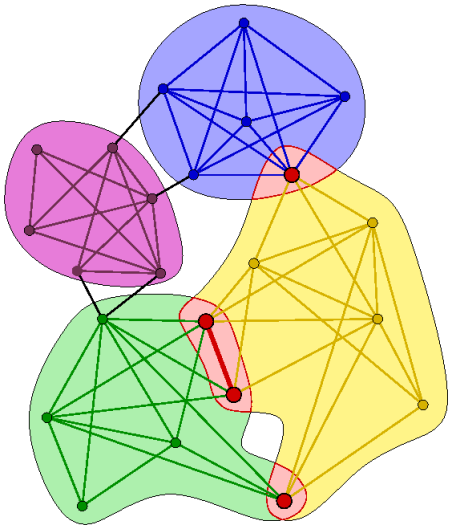


FIGURE 3.5 – Exemple d'application de la méthode CFinder (Extrait du site de Palla *et al.* [111]).

Shen *et al.* [126] proposent EAGLE (agglomerativE hierarchicAl clusterinG based on maximaL cliquE). EAGLE est une méthode agglomérative fondée sur la détection de cliques maximales et l'optimisation locale d'une fonction de qualité permettant l'élaboration d'un dendrogramme. L'algorithme opère en deux étapes sur le graphe. Premièrement, les cliques maximales sont trouvées pour former les premières communautés. La seconde étape consiste à fusionner les communautés ayant la plus grande similarité selon la mesures de la modularité de Newman.

Partitionnement de liens

Cette catégorie d'algorithmes est fondée sur le principe de regrouper des liens au lieu de nœuds. Un nœud du graphe est appelé chevauchant si les liens connectés à celui-ci appartiennent à plusieurs clusters. Dans Ahn *et al.* [4] les liens sont partitionnés en appliquant un regroupement hiérarchique qui utilise une similarité entre les liens calculée avec l'indice de Jaccard. Bien que l'idée de partitionnement des liens pour la détection de chevauchement semble être intuitive, l'obtention de la qualité n'est pas bien argumentée (Lancichinetti *et al.* [80]) car ces algorithmes utilisent une définition ambiguë de la communauté.

Expansion locale et Optimisation

Ces algorithmes utilisent le principe de l'expansion locale et de l'optimisation de la croissance d'une communauté naturelle ou partielle Lancichinetti *et al.* [78]. Une fonction de bénéfice local qui évalue la qualité d'un groupe de nœuds densément connectés est définie.

Baumes *et al.* [19] a proposé un processus en deux étapes. Premièrement, l'algorithme RankRemoval est utilisé pour classer les nœuds selon un critère. Ensuite, le processus supprime de manière itérative les nœuds hautement classés jusqu'à l'obtention de petits cœurs de cluster disjoints. Ces noyaux sont utilisés comme communautés de graines pour la deuxième étape du processus, appelée scan itératif (IS). Dans cette étape, les cœurs sont étendus en ajoutant ou en supprimant des nœuds jusqu'à ce qu'une fonction de densité locale ne puisse pas être améliorée.

Lancichinetti *et al.* [78] proposent la méthode LFM qui élargit une communauté à partir d'un nœud de graine aléatoire pour former une communauté en utilisant une fonction de fitness basée sur le degré interne et externe de la communauté ainsi qu'un paramètre de résolution contrôlant la taille des communautés. Après avoir trouvé une communauté, LFM sélectionne au hasard un autre nœud qui n'est pas encore affecté à une communauté pour étendre une nouvelle communauté. Les résultats de LFM dépendent du paramètre de résolution. La complexité dans le pire des cas est $\mathrm{O}(n^2)$.

Lee *et al.* [83] ont proposé GCE (Greedy Clique Expansion). Cet algorithme est très similaire à EAGLE. Il identifie dans un premier temps les cliques comme des graines au sein du graphe. Ces graines, qui sont des cœurs de communautés s'agrandissent par optimisation d'une fonction de qualité locale. Les auteurs choisissent la fonction de fitness de Lancichinetti *et al.* [79], à savoir pour une communauté S, $F_S = \frac{k_{in}^S}{(k_{in}^S + k_{out}^S)^\alpha}$ où α est un paramètre ajustable, k_{in}^S (nombre d'arêtes internes dans S) et k_{out}^S (nombre d'arêtes ayant une extrémité hors de S). Cette mesure privilégie une expansion vers des nœuds maximisant les liens internes et minimisant les liens externes.

Lancichinetti *et al.* [80] ont proposé OSLOM (Order Statistics Local Optimization Method). Il s'agit d'une méthode de raffinage agissant sur une partition déjà fournie. OSLOM utilise les algorithmes de Louvain ou d'Infomap pour créer une première partition. Des nœuds sont ensuite ajoutés ou retirés pour arriver à un état stable où il n'est plus intéressant de modifier la structure topologique des communautés. Le principe consiste à comparer les structures communautaires entre le modèle nul et la partition réelle en se fondant sur le nombre de liens dans les communautés et sur le degré du nœud étudié. La complexité la plus défavorable est $\mathrm{O}(n^2)$.

Cette stratégie est basée sur des observations réelles sur la formation de clusters autour de sommets de haut degré dans des réseaux du monde réel avec une distribution de degrés en loi de puissance. L'algorithme utilisé pour développer un ensemble de graines est une version de clustering personnalisé de l'algorithme PageRank (PPR) et est nommé NISE (Neighbourhood-Inflated Seed Expansion). Les ré-

sultats expérimentaux semblent trouver les bonnes communautés qui se chevauchent dans les réseaux du monde réel. Les auteurs montrent que l'algorithme NISE surpasse les autres méthodes de détection communautaire qui se chevauchent.

Détection de communautés floue

Les algorithmes de détection de communautés floue utilisent les poids des liens entre toutes les paires de nœuds et les communautés. Dans ces algorithmes, un vecteur d'appartenance (Gregory [54]) est calculé pour chaque nœud. Un des inconvénients de ces algorithmes est la nécessité de déterminer la dimension k du vecteur d'appartenance. Cette valeur peut être spécifiée en tant que paramètre d'entrée à l'algorithme ou calculée à partir des données. Wang *et al.* [137] ont combiné des méthodes de détection disjointes avec des algorithmes d'optimisation locaux. Tout d'abord, une partition est obtenue à partir d'un algorithme de détection de communauté disjointe. Les communautés tentent ensuite d'ajouter ou de supprimer des nœuds. L'algorithme utilise la différence calculée (appelée variance) de deux scores de fitness sur une communauté, soit pour inclure un nœud, soit pour le supprimer.

Zhang *et al.* [151] proposent un algorithme spectral pour la détection de communautés chevauchantes. Le principe est de calculer un certain nombre de vecteurs propres liés à la matrice Laplacienne représentant le graphe, puis d'appliquer sur cet espace propre un algorithme flou de clustering : le Fuzzy C means (FCM) et de retranscrire les résultats sur le graphe pour obtenir les recouvrements. La méthode nécessite de spécifier le nombre de recouvrements et nécessite le calcul des valeurs et des vecteurs propres.

Algorithmes dynamiques et basés sur les agents

Cette famille d'algorithmes concerne les algorithmes de propagation de labels (Raghavan *et al.* [113], Xie *et al.* [142]), dans lesquels les nœuds avec le même label forment une communauté, ont été étendus à la détection de communauté chevauchantes en permettant à un nœud d'avoir plusieurs labels.

La première méthode connue utilisant la propagation de labels a été proposée par Gregory [53], à savoir COPRA. Les auteurs proposent d'utiliser un vecteur pour maintenir les labels les plus courants en utilisant un seuil de probabilité. Le résultat est qu'un nœud peut appartenir à une ou plusieurs communautés.
Dans COPRA (Gregory [53]), chaque nœud met à jour ses coefficients d'appartenance en faisant la moyenne des coefficients de tous ses voisins à chaque pas de temps de manière synchrone. Un paramètre v est utilisé pour contrôler le nombre maximum de communautés auxquelles un nœud peut s'associer. La complexité temporelle est $O(v * m * log(v * m/n))$ par itération, n est le nombre de nœuds et m est le nombre d'arêtes.

SLPA (Xie *et al.* [142]) est un processus général de propagation d'information basé sur le principe *speaker-listener*. Il répartit les labels entre les nœuds selon des règles d'interaction par paire. Contrairement à Gregory [53], Raghavan *et al.* [113], où un nœud oublie les connaissances acquises lors des itérations précédentes, SLPA fournit à chaque nœud une mémoire pour stocker les informations reçues (c'est-à-dire des labels). Le degré d'appartenance est interprété comme étant la probabilité

d'observer un label dans la mémoire d'un nœud. L'un des avantages du SLPA est qu'il ne nécessite aucune connaissance du nombre de communautés. La complexité temporelle est O $(t * m)$, linéaire avec le nombre d'arêtes m, où t est un nombre maximum prédéfini d'itérations.

De nombreuses techniques utilisant la propagation de labels ont ensuite été définies, telles que SPAEM (*propagation de labels dynamique*) dans Ren *et al.* [117], BMLPA (*algorithme de propagation multi-labels équilibré*) dans Wu *et al.* [141], MLPA (*algorithme de propagation multi-labels*) dans Dai *et al.* [37], etc.

Dans Wu et Zhang [140], les auteurs soulignent que le degré de connexion peut refléter la possibilité d'appartenance pour un nœud à ses communautés voisines, ils ont donc proposé un COPRA basé sur la connexion nommé COPRA-CD. Dans COPRA-CD, tous les nœuds sont initialisés avec un identifiant de communauté unique et un coefficient d'appartenance réglé à 1, chaque nœud met à jour son identifiant de communauté par l'union des labels de ses voisins, le coefficient d'appartenance correspondant est obtenu en normalisant la somme des coefficients d'appartenance des communautés sur tous les voisins. Après plusieurs itérations, les communautés qui sont totalement contenues par d'autres sont supprimées et les communautés déconnectées sont divisées. Les résultats expérimentaux montrent des améliorations de la qualité et une meilleure stabilité pour les réseaux flous.

Autres méthodes pour la détection de communautés chevauchantes

Il existe d'autres méthodes pour la détection de communautés chevauchantes, fondées sur la topologie du graphe et le nombre de connexions qu'un nœud a vis-à-vis des communautés avoisinantes.

Nepusz *et al.* [102] ont modélisé la détection de la communautés chevauchantes par un problème d'optimisation non linéaire qui peut être résolu par la méthode du recuit simulé.

Gregory [52] étend l'algorithme de clustering divisive de Girvan et Newman [50]en permettant à un nœud de se diviser en plusieurs copies avec CONGA.

Kovács *et al.* [74] a proposé une approche fondée sur la centralité ainsi que les fonctions d'influence.

D'autres méthodes ont été développées pour lier la détection communautaire à une méthode de factorisation matricielle non négative dans Jin *et al.* [64]. Des algorithmes évolutifs ont été utilisés pour trouver une partition avec chevauchement dans Wen *et al.* [139] et Zhang *et al.* [150].

3.4.2 Mesures externes et internes pour l'évaluation des communautés chevauchantes

Certaines mesures utilisées dans l'évaluation des communautés disjointes ont été revues pour le cas du chevauchement (Attal [7]).

Mesures externes pour l'évaluation des communautés chevauchantes

L'information mutuelle normalisée pour le chevauchement a été proposée par Lancichinetti *et al.* [79] . Considérons deux couvertures C et C' de V. $C = \{c_1, ..., c_k\}$, et $C' = \{c_1, ..., c'_k\}$, La probabilité de tirer un élément et qu'il appartienne à un élément (communauté) (c_i) de la couverture C est $\frac{n_{c_i}}{n}$, où n_{c_i} est le nombre d'éléments dans c_i. L'entropie de Shanon de la partition C est ainsi définie par :

$$H(C) = -\sum_{c_i \in C} \frac{n_{c_i}}{n} log_2 \frac{n_{c_i}}{n} \tag{3.12}$$

Lancichinetti *et al.* [79] ont proposé une version du NMI pour la comparaison de couvertures C et C' comme étant :

$$NMI_{LFK} = 1 - \frac{1}{2}(\frac{H(C|C')}{C} + \frac{H(C|C')}{C'}) \tag{3.13}$$

avec $H(C|C')$ étant l'entropie conditionnelle. Cette quantité mesure l'entropie restante provenant de la couverture C, si l'on connait parfaitement la seconde couverture C'. $H(C|C') = 0$ si et seulement si la couverture C est complètement déterminée par la couverture C'. La valeur de NMI_{LFK} est comprise entre 0 et 1. Plus cette valeur est proche de 1, plus les deux couvertures sont identiques.

McDaid *et al.* [95] ont proposé une extension du NMI en ajoutant une pénalité si les deux couvertures sont trop différentes.

L'indice d'Omega (Collins et Dent [35]) est fondé sur des paires de nœuds qui sont dans les mêmes communautés selon les deux couvertures.

Considérons deux couvertures C et C', l'indice d'omega est défini de la manière suivante :

$$\Omega(C, C') = \frac{o_u(C, C') - o_e(C, C')}{1 - o_e(C, C')} \tag{3.14}$$

où $o_u(C, C')$ représente la fraction de paires de nœuds qui apparaissent ensemble dans les mêmes communautés à la fois dans C et C'.
$o_u(C, C') = \frac{2}{n(n-1)} \sum_j |t_j(C) \cap t_j(C')|$ où $t_j(C)$ est l'ensemble des paires de nœuds qui apparaissent dans exactement j communautés dans la couverture C. $\frac{n(n-1)}{2}$ est Le nombre de paires de nœuds. Le terme $o_e(C, C')$ est la valeur espérée de cette fraction dans le modèle nul.

$$o_e(C, C') = \frac{4}{n^2(n-1)^2} \sum_j |t_j(C)||t_j(C')| \tag{3.15}$$

L'indice d'Omega prend sa valeur entre 0 et 1. Une valeur proche de 1 signifie que les deux couvertures sont identiques.

Mesures internes pour l'évaluation des communautés chevauchantes

Plusieurs extensions de la modularité ont été proposées pour le cas du chevauchement (Nepusz *et al.* [102], Nicosia *et al.* [110], Shen *et al.* [126], Zhang *et al.* [151]).

Elles se basent toutes sur un cœfficient d'appartenance d'un nœud i à une communauté c, que l'on note $a_{i,c}$. On considère par la suite une couverture de communautés chevauchantes $C = \{c_1, ..., c_{|C|}\}$ et un vecteur d'appartenance pour un nœud i à chacune des communautés $(a_{i,c_1}, ..., a_{i,c_{|C|}})$. Le cœfficient d'appartenance d'un nœud i vérifie deux contraintes qui sont $0 \leq a_{i,c} \leq 1, \forall i \in V, \forall c \in C$ et par $\sum_{c \in C} a_{i,c} = 1$.

Nicosia *et al.* [110] ont proposé une extension à la modularité pour le cas du chevauchement entre communautés. Les auteurs la définissent de la manière suivante :

$$Q_{Ov}^{Ni} = \frac{1}{m} \sum_{c} \sum_{i,j \in V} [\beta_{ij,c} A_{ij} - \beta_{ij,c}^{out} \beta_{ij,c}^{in} \frac{k_i^{out} k_j^{in}}{m}] \tag{3.16}$$

où $\beta_{ij,c}$ est le cœfficient d'appartenance du lien ij à la communauté c, $\beta_{ij,c}^{out}$ est le cœfficient d'appartenance espéré de tous les liens possibles ij du nœud i au nœud j à l'intérieur de la communauté c et $\beta_{ij,c}^{in}$ est le cœfficient d'appartenance de n'importe quel lien ij pointant vers un nœud j dans la communauté c. Les termes k_i^{out} et k_i^{in} sont respectivement le degré sortant et entrant du nœud i. Il est à noter que $\beta_{ij,c}$ est une fonction de la forme $\beta_{ij,c} = F(a_{i,c}, a_{j,c})$ avec $a_{i,c}$ cœfficient d'appartenance du nœud i par rapport à la communauté c. $F(a_{i,c}, a_{j,c}) = \frac{1}{(1+e^{-f(a_{i,c})})(1+e^{-f(a_{j,c})})}$ où $f(a_{i,c})$ est une fonction linéaire échelonnable du type $f(x) = 2px - p$, $p \in \mathbb{R}$. Les auteurs ont proposé d'utiliser la fonction $f(x) = 60x - 30$ pour le paramétrage de la modularité chevauchante car elle donnait des résultats très encourageants.

La valeur de la modularité de Nicosia varie entre 0 et 1. Une valeur proche de 0 implique que l'algorithme n'a détecté qu'une seule communauté (le graphe en lui même) alors qu'une valeur proche de 1 signifie la présence de structures communautaires possiblement chevauchantes.

3.4.3 Petite Discussion

Les méthodes chevauchantes présentent des complexités plus élevées que les méthodes disjointes ; une partie des algorithmes chevauchants repose sur le fait de lancer plusieurs fois un algorithme disjoint et d'observer pour un nœud les communautés auxquelles il appartient le plus (comme COPRA). Ces algorithmes demandent un espace de stockage élevé.

Les méthodes chevauchantes requièrent également un degré de paramétrage plus élevé. De nombreux algorithmes nécessitent par exemple de fournir un nombre de communautés auxquelles un nœud pourrait appartenir.

Une grande partie des algorithmes de détection de communautés chevauchantes sont instables car ils sont fondés sur des algorithmes non déterministes disjoints.

Cependant, les méthodes par propagation de labels ont été l'objet de nombreuses études et constituent, pour l'étude des grands graphes, l'option à privilégier.

3.5 Détection de communautés chevauchantes en utilisant des fonctions d'appartenance

Nous proposons une méthode de détection des communautés chevauchante à partir de communautés disjointes pré-calculées en utilisant une version stable de propagation des labels : *core detection label propagation*(CDLP) décrite dans Attal et Malek [8]. L'algorithme sélectionne les nœuds candidats pour le chevauchement et utilise *des fonctions d'appartenance* pour décider de l'affectation ou non d'un nœud candidat à chacune de ses communautés voisines. Ces fonctions d'appartenance sont soit basées sur des mesures globales qui sont la densité et le coefficient de clustering (Attal *et al.* [9]) ou sur les moyennes des mesures locales qui sont les centralités d'intermédiarité et de proximité. La figure 3.6 montre le diagramme de notre méthode (Attal *et al.* [10]).

Même si notre algorithme est basé sur une approche de propagation de labels, nous ne pouvons pas le classer dans la catégorie des *Algorithmes dynamiques et basés sur les agents* (voir la section 3.4.1) car le processus d'identification des communautés qui se chevauchent ne se fait pas simultanément avec le propagation des labels. Il est fondé sur l'utilisation des fonctions d'appartenance proposées appliquées aux nœuds candidats qui sont situés dans les limites de communautés disjointes. Ces nœuds candidats pourraient faire se chevaucher des communautés disjointes.

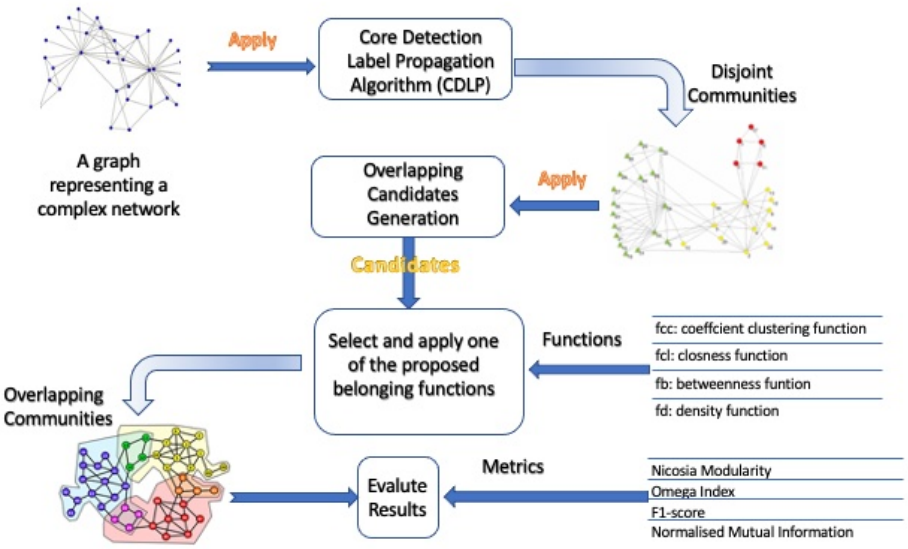


FIGURE 3.6 – Diagramme représentant notre approche de détection de communautés chevauchantes en utilisant des fonctions d'appartenance (Attal *et al.* [10]).

3.5.1 Propagation de labels avec détection de cœurs (CDLP)

Les algorithmes de détection des communautés chevauchantes décrits dans cette section reposent sur l'algorithme de propagation de labels (LPA). Cet algorithme présente l'avantage d'avoir une complexité assez faible et est adapté aux grands graphes. Il a néanmoins les inconvénients suivants (voir section 3.3.1) :

— une mauvaise propagation qui peut mener à de trop grandes communautés (le problème des communautés géantes),
— une instabilité,
— de mauvaises propagations qui peuvent donner des communautés ayant le même label.

Une façon de stabiliser l'algorithme consiste à appliquer plusieurs fois le LPA et à observer les nœuds qui apparaissent souvent ensemble. Dans Attal et Malek [8], nous avons proposé une méthode pour stabiliser la propagation des labels. La méthode consiste à lancer $\mathcal{N}$ fois l'algorithme non déterministe et à créer une matrice $P_{ij}^{\mathcal{N}} = [p_{ij}]_{n\times n}^{\mathcal{N}}$ telle que p_{ij} représente la fréquence d'apparition des nœuds i et j dans les mêmes communautés. Un cœur de communautés est l'ensemble des nœuds se trouvant fréquemment ensemble dans une même communauté après plusieurs lancements d'un algorithme non déterministe. Un nouveau graphe $G' = (V, E')$ est alors créé en utilisant un seuil $\alpha \in [0, 1]$. Les paires de noeuds de la matrice $P_{ij}^{\mathcal{N}}$ ayant un poids plus petit que α sont exclues de l'ensemble des arêtes de G'. Les composants connectés créés dans le graphe G' sont les communautés. La figure 3.7 montre un exemple simple d'obtention du graphe $G' = (V, E')$ avec $N = 4$ et $\alpha = 0.5$. Le choix de α est fait en prenant le score de modularité le plus élevé. Cet algorithme est appelé *propagation de labels avec détection de cœurs* (CDLP). Un des avantages du CDLP est qu'en variant α, la hiérarchie des communautés peut être trouvée et sera représentée par un dendrogramme. Les expérimentations effectuées par Attal et Malek [8] ont montré que les faibles valeurs de α implique une augmentation de la taille des communautés et les valeurs élevées de α génèrent de très petites communautés.

La méthode de stabilisation matricielle a été également empiriquement appliquée sur l'algorithme de Louvain (Blondel *et al.* [24]) dans Seifi *et al.* [124] qui est une méthode agglomérative utilisant une optimisation locale de la modularité.

3.5.2 Détection des communautés chevauchantes en utilisant des fonctions d'appartenance

Une nouvelle méthode pour détecter des communautés chevauchantes à partir de communautés disjointes pré-calculées obtenues en utilisant la *propagation de labels avec détection du cœur* (CDLP) (Attal et Malek [8]) a été proposée initialement dans Attal *et al.* [9] et complétée dans Attal *et al.* [10] . L'algorithme sélectionne les nœuds candidats au chevauchement et utilise *les fonctions d'appartenance* pour décider l'affectation ou non d'un nœud candidat à chacune de ses communautés voisines. Ces fonctions d'appartenance sont définies en utilisant des mesures globales telles que la densité et le coefficient de clustering (Attal *et al.* [9]) ou bien sur les moyennes de certaines mesures locales (des nœuds) comme les centralité d'intermédiarité et de proximité.

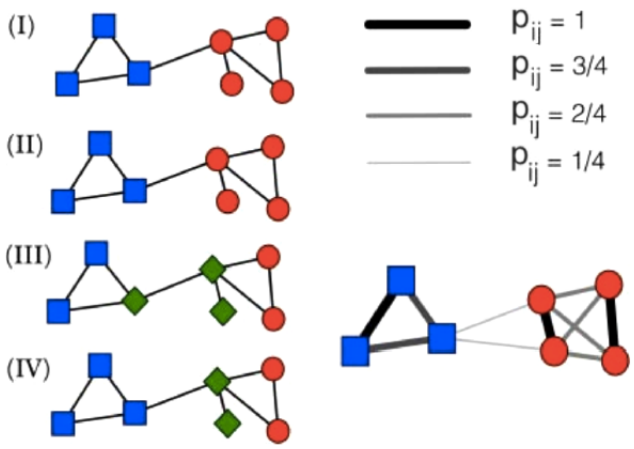


FIGURE 3.7 – Exemple des résultats obtenus en appliquant l'algorithme CDLP : communautés obtenues avec $N = 4$ et $\alpha = 0.5$.

Nous illustrons cette méthode sur un petit graphe $G = (V, E)$, avec

$$V = \{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8, v_9, v_{10}, x\}$$

et E, l'ensemble des arêtes, (voir Figure 3.8). La notation c est utilisée pour désigner une communauté et C pour désigner un ensemble de communautés. Les poids sur les connexions sont extraits de la matrice $P_{ij}^{\mathcal{N}}$ après avoir calculé 100 propagations de labels. En choisissant $\alpha > 0,5$ (qui correspond au score de modularité le plus élevé), nous obtenons trois communautés, $c_1^x = \{v_1, v_2, v_3, v_4\}$, $c_2^x = \{v_9, v_{10}\}$, $c_3^x = \{v_5, v_6, v_7, v_8\}$ avec le nœud x dont l'appartenance à différentes communautés est étudiée. Nous nous concentrons sur le nœud x, qui est le plus susceptible de chevaucher différentes communautés.

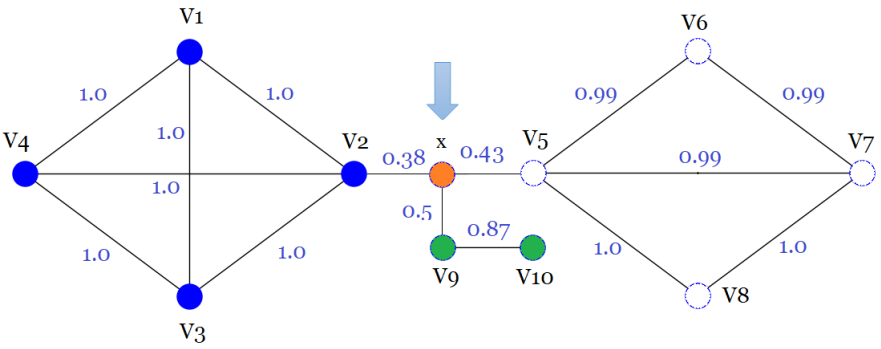


FIGURE 3.8 – Un graphe avec 3 communautés évidentes et un nœud (x) dont l'appartenance à une communauté spécifique est discutable.

Considérons l'ensemble $\{c_1^x, ..., c_K^x\}$ des communautés voisines du nœud x, c'est-à-dire toutes les communautés qui ont un lien avec x.

Pour générer toutes les combinaisons de chevauchement possibles entre le nœud x et ses K communautés voisines, nous définissons une fonction $gen{-}candidate(\{c_1^x, ..., c_K^x\}, j) \rightarrow$

C_j, où C_j est un ensemble de communautés représentant toutes les combinaisons possibles de j communautés, parmi les K communautés voisines de x. La cardinalité de C_j est $\binom{k}{j}$.

On obtient les combinaisons suivantes à partir du graphe G (Fig. 3.8) :
$C_1 = [c_1^x, c_2^x, c_3^x] = [\{v_1, v_2, v_3, v_4\}, \{v_9, v_{10}\}, \{v_5, v_6, v_7, v_8\}]$ avec $\binom{3}{1} = 3$ possibilités

$C_2 = [\{c_1^x, c_2^x\}, \{c_1^x, c_3^x\}, \{c_2^x, c_3^x\}]$
$= [\{\{v_1, v_2, v_3, v_4\}, \{v_9, v_{10}\}\}, \{\{v_1, v_2, v_3, v_4\}, \{v_5, v_6, v_7, v_8\}\}, \{\{v_9, v_{10}\}, \{v_5, v_6, v_7, v_8\}\}]$

avec $\binom{3}{2} = 3$ possibilités

$C_3 = [\{c_1^x, c_2^x, c_3^x\}] = [\{\{v_1, v_2, v_3, v_4\}, \{v_9, v_{10}\}, \{v_5, v_6, v_7, v_8\}\}]$ avec $\binom{3}{3} = 1$ possibilité de 1 .

La fonction $gen-candidate$ et les résultats des différentes combinaisons possibles est utilisée par la suite pour la définition des fonctions d'appartenance.

Première fonction d'appartenance fondée sur la densité

L'idée est de proposer une fonction d'appartenance basée sur l'intuition suivante : *un nœud connecté à un ensemble de communautés ayant des densités élevées pourrait les chevaucher.*

La densité d'une communauté c est donnée par $d : c \longmapsto [0, 1]$, avec $d(c) = \frac{2*|E|}{|V|*(|V|-1)}$, $|V|$ étant le nombre des sommets de la communauté et $|E|$ est le nombre des arêtes reliant les paires de V.

Dans la figure 3.8 les communautés $c_1^x = \{v_1, v_2, v_3, v_4\}$, $c_2^x = \{v_9, v_{10}\}$ et $c_3^x = \{v_5, v_6, v_7, v_8\}$ ont une densité élevée : $d(c_1^x) = 1$, $d(c_2^x) = 1$ and $d(c_3^x) = 0.83$.

Nous considérons la densité des communautés associées au poids des arêtes qui les lient au nœud x pour trouver les communautés qui se chevauchent (Attal *et al.* [9]). Nous définissons une fonction d'appartenance qui permet de trouver les combinaisons qui maximisent la quantité obtenue en multipliant les densités de communautés par les poids extraits de la matrice de stabilisation ($P_{ij}^{\mathcal{N}}$), la fonction d'appartenance basée sur la densité est : $f_d\{x, C\} \longmapsto \mathbb{R}_+$:

$$f_d(x, C) = \frac{1}{|C|} \sum_{c \in C} w_{c,x} \times d(c) \tag{3.17}$$

$w_{c,x}$ représente les poids extraits de la matrice de stabilisation ($P_{ij}^{\mathcal{N}}$) des liens reliant le nœud x aux différentes communautés c dans C, C étant une combinaison dans C_j, $d(c)$ étant la densité de la communauté c.

En utilisant les informations contenues dans la matrice de stabilisation, un nœud lié à un ensemble de communautés à haute densité avec un poids important a plus de chance de se dupliquer dans ces communautés qu'un nœud lié à un ensemble

de communautés de faible densité avec un faible poids selon la formule ci-dessus. Le chevauchement peut également être refusé. Cela pourrait être le cas si les liens reliant le nœud x aux autres communautés ont un poids faible ou si f_d est faible, avec une faible densité. Nous proposons donc que le chevauchement se fasse si et seulement si $f_d(x, C) \geq \frac{1}{|C*|} \sum_{S \in C*} d(S)$, avec

$$C* = argmax_{C \in C_j \wedge j \in \{1,...,K\}}(f_d(x, C))$$

et $d(S)$ étant la densité du sous-graphe S (une communauté dans $C*$).

Il est également possible de contrôler le nombre de communautés qui se chevauchent. Par exemple, si nous souhaitons augmenter le nombre minimum des communautés qui se chevauchent de 1 à L, le domaine des variations de j est modifié à $j \in \{L, ..., K\}$ au lieu de $j \in \{1, ..., K\}$). Cela force le nœud x à appartenir simultanément à au moins L communautés.

Les autres fonctions d'appartenance

Trois autres fonctions d'appartenance ont été également proposées :

— Une fonction fondée sur les coefficient de clustering des communautés :

$$f_{cc}(x, C) = \frac{1}{|C|} \sum_{c \in C} w_{c,x} \times CC(c) \tag{3.18}$$

$w_{c,x}$ étant le poids du lien entre le nœud x et la communauté c. $CC(c)$ étant le coefficient de clustering de la communauté c.

— Une fonction fondée sur la moyenne des centralités d'intermédiarité dans les différentes communautés :

$$f_b(x, C) = \frac{1}{|C|} \sum_{c \in C} w_{c,x} \times [|c| - g(c)] \tag{3.19}$$

$g(c)$ étant la somme des centralités d'intermédiarité dans la communauté c, |c| étant le nombre de nœuds dans c. C est l'ensemble des communautés disjointes pré-calculées et $w_{c,x}$ est le poids du lien entre le nœud x et la communauté c.

— Une fonction fondée sur la moyenne des centralités de proximité dans les différentes communautés :

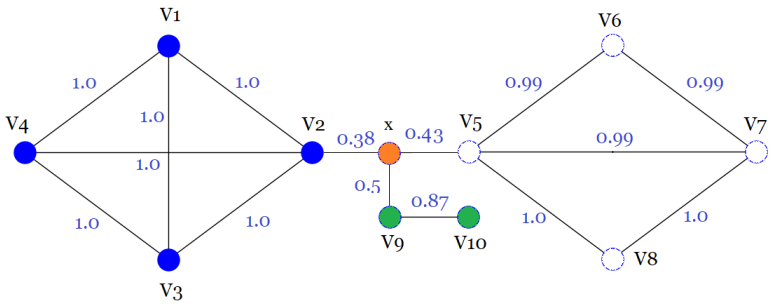
$$f_{cl}(x, C) = \frac{1}{|C|} \sum_{c \in C} w_{c,x} \times [|c| - cl(c)] \tag{3.20}$$

$cl(c)$ étant la somme des centralités de proximité dans la communauté c, $|c|$ étant le nombre de nœuds dans c. C est l'ensemble des communautés disjointes pré-calculées et $w_{c,x}$ est le poids du lien entre le nœud x et la communauté c.

La figure 3.9 montre un exemple d'utilisation de ces quatre fonctions d'appartenance. Cette illustration prend en compte l'ensemble des communautés résultant de

la propagation de labels de base avec l'algorithme (CDLP) qui utilise la matrice de fréquence, le but étant de savoir si x peut appartenir à plusieurs communautés.

La meilleure configuration pour obtenir une détection de communauté se chevauchant sur x est donné en gras pour chaque fonction d'appartenance.



Combinations	f_{cc}	f_d	f_b	f_{cl}
$\{x, \{c_1^x\}\}$	**0.38**	0.38	1.346	0.908
$\{x, \{c_2^x\}\}$	0	0.5	0.9	0.645
$\{x, \{c_3^x\}\}$	0.357	0.357	**1.512**	**1.0621**
$\{x, \{c_1^x, c_2^x\}\}$	0.19	0.44	1.226	0.78
$\{x, \{c_1^x, c_3^x\}\}$	0.368	0.368	1.429	0.98
$\{x, \{c_2^x, c_3^x\}\}$	0.178	0.428	1.206	0.85
$\{x, \{c_1^x, c_2^x, c_3^x\}\}$	0.246	**0.618**	1.253	0.87

FIGURE 3.9 – Soit les trois communautés disjointes suivantes :$c_1^x = \{v_1, v_2, v_3, v_4\}$, $c_2^x = \{v_9, v_{10}\}$, $c_3^x = \{v_5, v_6, v_7, v_8\}$. Les fonctions f_d, f_{cc},f_b et f_{cl} sont calculées pour le nœud x, x peut appartenir à plusieurs communautés selon la fonction d'appartenance. les résultats en gras montrent son affectation aux différentes communautés.

3.5.3 Résultats expérimentaux

Pour évaluer nos algorithmes, nous utilisations certaines mesures internes et externes définies pour le problème de détection de communauté chevauchantes. La version chevauchante de la *modularité*(Nicosia *et al.* [110]) décrite dans 3.4.2 est utilisée comme mesure interne. Comme mesures externes, les version étendues de l' *information mutuelle normalisées*, l'*indice de Rand ajusté* ainsi que le F_1 *score* sont utilisées (voir section 3.4.2).

Le tableau 3.1 montre les caractéristiques des réseaux (benchmarks) utilisés pour les expérimentations : le réseau du club de karaté de Zachary (Zac)(Zachary [146]), le réseau du club de football (Foot) (Girvan et Newman [49]), le réseau du livre politique (Pol) (Krebs [75]), le réseau des dauphins (Dol) (Lusseau *et al.* [87]) et un réseau de collocation scientifique (co-authoring network : NS) (Newman [106]) .

Nous montrons dans cette section les résultats obtenus sur le réseau du club de karate de Zachary (les autres résultats sont détaillés dans Attal *et al.* [10]).

Figures 3.10, 3.11, 3.12 et 3.13 montrent les résultats expérimentaux obtenus selon les différentes fonctions d'appartenance. L'algorithme CDLP détecte 2 communautés lorsque $\alpha \geq 0.6$ et 4 communautés lorsque $\alpha \geq 0.7$ et $\alpha \geq 0.8$. Figures

caractéristiques des réseaux d'évaluation				
Réseaux	$\|V\|$ et $\|E\|$	Densité	Diamètre	Transitivité
Zachary	34 \ 78	0.139	5.0	0.256
Foot	115 \ 615	0.094	4.0	0.407
Dol	62 \ 159	0.084	8.0	0.309
Pol	105 \ 441	0.081	7.0	0.348

Tableau 3.1 – Caractéristiques des réseaux d'évaluation.

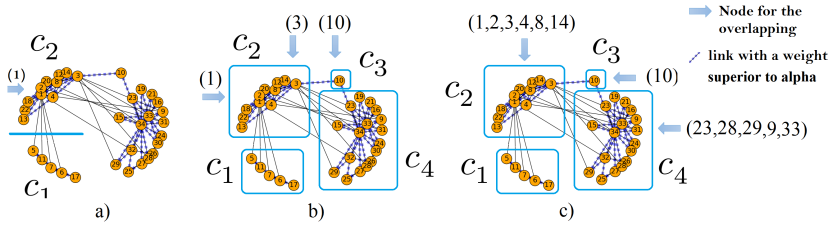


FIGURE 3.10 – Réseau de Zachary avec la fonction f_d, a) $\alpha \geq 0.6$, b) $\alpha \geq 0.7$ c) $\alpha \geq 0.8$.

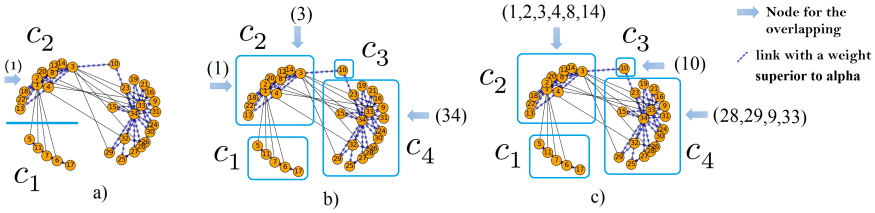


FIGURE 3.11 – Réseau de Zachary avec la fonction f_{cc}, a) $\alpha \geq 0.6$, b) $\alpha \geq 0.7$ c) $\alpha \geq 0.8$.

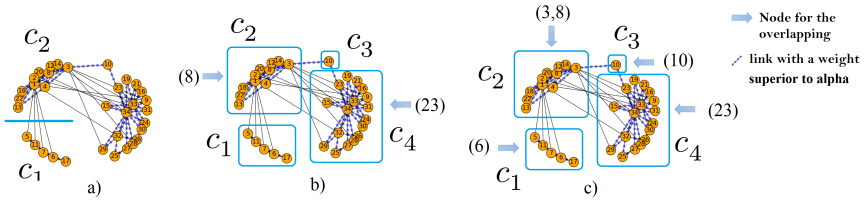


FIGURE 3.12 – Réseau de Zachary avec la fonction f_b, a) $\alpha \geq 0.6$, b) $\alpha \geq 0.7$ c) $\alpha \geq 0.8$.

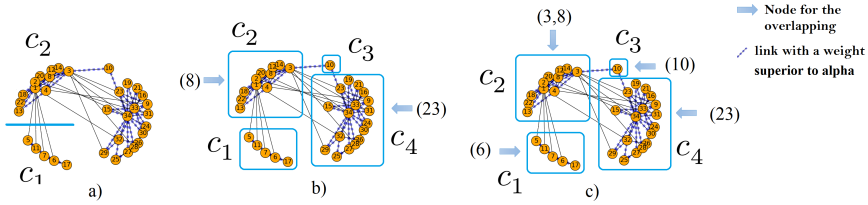


FIGURE 3.13 – Réseau de Zachary avec la fonction f_{cl}, a) $\alpha \geq 0.6$, b) $\alpha \geq 0.7$ c) $\alpha \geq 0.8$.

3.10 et 3.11, montrent que le nœud 10 appartient à deux communautés quand on utilise f_d. Quand $\alpha \geq 0.8$, le nœud 10 devient chevauchant selon la fonction f_{cc}.

Lorsque $\alpha \geq 0.8$, le nœud 3 qui est connu pour être dans des communautés qui

se chevauchent dans la littérature, est répliqué dans deux communautés : c_3 et c_4, selon f_d, mais seulement dans c_4 selon f_{cc}.

Le nœud 3 est également détecté par f_b et f_{cl} mais avec une valeur α élevée. Le nœud 1 est répliqué dans une communauté dans c_2 avec f_d et f_{cc}. Le tableau 3.2 montre que plus α est grand plus le nombre des candidats au chevauchement est élevé. La qualité des résultats est meilleure pour f_d que f_{cc}. La plus grande valeur de modularité est obtenue pour $\alpha \geq 0.7$ pour toutes les fonctions. Le constat est le même en ce qui concerne la valeur la plus élevée du NMI et de l'indice Ω.

En ce qui concerne la fonction d'appartenance utilisant la centralité de proximité, la valeur la plus élevée de modularité est obtenue pour un nombre de communautés égal à trois.

Réseau du club de Karaté Zachary : résultats en utilisant les fonctions f_d, f_{cc}, f_{cl} and f_b						
	Cand	EBC				
$\alpha \geq 0.6$	47.058%	17.95%				
$\alpha \geq 0.7$	41.17%	16.0%				
$\alpha \geq 0.8$	55.88%	26.92%				
$\alpha \geq 0.9$	55.88%	26.92%				
f_d	CandOv	Q_{Ov}^{Nic}	Ω	F_1	NMI	#
$\alpha \geq 0.6$	2.94% (1)	0.399	0.064	0.65	0.2365	2
$\alpha \geq 0.7$	8.8235% (3)	0.621	**0.711**	**0.86**	**0.518**	4
$\alpha \geq 0.8$	32.352% (11)	0.42	0.4923	0.75	0.3488	5
$\alpha \geq 0.9$	32.352% (11)	0.42	0.4923	0.75	0.3488	5
f_{cc}	CandOv	Q_{Ov}^{Nic}	Ω	F_1	NMI	#
$\alpha \geq 0.6$	2.941% (1)	0.3986	0.064	0.65	0.2365	2
$\alpha \geq 0.7$	8.823% (3)	0.6210	**0.711**	**0.86**	**0.518**	3
$\alpha \geq 0.8$	32.353% (11)	0.42	0.4923	0.75	0.3488	5
$\alpha \geq 0.9$	32.353% (11)	0.42	0.4923	0.75	0.3488	5
f_b	CandOv	Q_{Ov}^{Nic}	Ω	F_1	NMI	#
$\alpha \geq 0.6$	2.941% (1)	0.064	0.0645	0.65	0.2365	2
$\alpha \geq 0.7$	5.882% (2)	0.68	0.66	**0.86**	0.44	5
$\alpha \geq 0.8$	11.76% (4)	0.62	0.49	0.75	0.31	6
$\alpha \geq 0.9$	11.76% (4)	0.62	0.49	0.75	0.31	6
f_{cl}	CandOv	Q_{Ov}^{Nic}	Ω	F_1	NMI	#
$\alpha \geq 0.6$	2.94%b (1)	0.384	0.25	0.13	0.26	3
$\alpha \geq 0.7$	5.88% (2)	**0.73**	0.66	**0.86**	0.45	3
$\alpha \geq 0.8$	14.70% (5)	0.68	0.549	0.85	0.36	4
$\alpha \geq 0.9$	14.70% (5)	0.59	0.55	0.75	0.35	5

Tableau 3.2 – Cand : pourcentage des candidats potentiels pour le chevauchement, EBC : pourcentage des liens entre communautés, CandOv : pourcentage des nœuds qui se chevauchent, Q_{Ov}^{Nic} : modularité de Nicosia, Ω : indice d'omega, F_1 : F_1-score, NMI : information mutuelle normalisée, # : nombre de communautés.

Observations générales

Les fonctions basées sur la densité et le coefficient de clustering sont calculées sur des sous-graphes connectés représentant des communautés disjointes pré-calculées tandis que les centralités de l'intermédiarité et de proximité sont calculées sur les nœuds regroupés dans ces communautés.

Les résultats expérimentaux montrent que les performances des fonctions d'appartenance dépendaient de la topologie des communautés. Les fonctions f_d et f_{cc} sont plus adaptées quand les communautés ont des densités élevées pour lesquels le taux d'attribution à différentes communautés pour un de nœud candidat au chevauchement est important. Les fonctions f_b et f_{cl} semblent plus performants pour

des graphes représentant des communautés ayant des faibles densités, où certains chevauchements sont difficiles à détecter.

3.5.4 Etude comparative

Nous comparons notre algorithme avec les algorithmes les plus utilisés en détection de communautés chevauchantes qui sont considérés comme des algorithmes de référence pour leurs catégories (Attal *et al.* [10]). Ces catégories (ou classes) ont été décrites dans la section 3.4. Ces algorithmes sont :

- CFinder (Palla *et al.* [111]) : représentant de la catégorie *cliques de percolation*,
- Oslom (Lancichinetti *et al.* [80]) : appartenant à la classe *expansion locale et optimisation*,
- COPRA ($\nu = 2$ and $\nu = 3$), ν étant le nombre des communautés auxquelles les nœuds pourraient appartenir (Gregory [53]) et SLPA (Xie *et al.* [144]) qui sont deux algorithmes basés sur la propagation des labels, appartenant à la classe *algorithmes dynamiques et basés sur les agents*,
- CONGA (Gregory [52]) : une extension de l'algorithme de Girvan and Newman's qui est un algorithme divisive.

Le tableau 3.3 montre les résultats obtenus avec les paramètres qui ont permis d'obtenir le score interne le plus élevé (Q_{Ov}^{Nic} : modularité de Nicosia). L'approche proposée donne des résultats relativement bons et compétitifs en terme de qualité mais dépendent de la topologie du réseau. Pour le réseau des dauphins, les différentes fonctions donnent des mesures NMI égales à 1. Lorsque les graphes ont des densités communautaires élevées, les fonctions utilisant le coefficient de clustering ou la densité donnent de meilleurs résultats que ceux utilisant la centralité d'intermédiarté ou de proximité, ce que l'on peut observer pour le réseau du football. La performance de notre approche dépasse COPRA et nous constatons une meilleure stabilisation pour les réseaux de Zachary, Dolphin et les livres politiques. D'un autre côté, les algorithmes basés sur la propagation des labels impliquent un nombre plus élevé de communautés, l'algorithme CDLP avec utilisation des fonctions f_d and f_{cc} permet de réduire ce nombre de communautés grâce à l'utilisation de la matrice de fréquences qui permet de stabiliser la propagation des labels.

3.6 Synthèse du chapitre

Nous avons présenté dans ce chapitre les différentes approches de détection de communautés disjointes et chevauchantes dans un graphe de terrain ainsi que les mesures d'évaluation internes et externes proposées dans la littérature et leur adaptation au contexte du chevauchement.

Nous avons proposé ensuite une méthode pour la detection des communautés chevauchantes à partir de communautés disjointes pré-calculées obtenues en utilisant la méthode du *core detection label propagation* (CDLP) décrite dans (Attal et Malek [8]). L'algorithme sélectionne les nœuds candidats pour le chevauchement et utilise des *fonctions d'appartenance* pour décider de l'affectation ou non d'un nœud candidat à chacune de ses communautés voisines. Nous avons proposé plusieurs

Analyse comparative						
Networks	F_1	Ω	NMI	Q_{Ov}^{Nic}	#	%
Zac #2						
CFinder	0.48	0.35	0.18	0.52	3	5.88%
OSLOM	**0.86**	**0.84**	**0.80**	**0.748**	2	2.94%
CONGA	0.65	0.113	0.274	0.441	2	2.94%
$COPRA_2$*	0.281	0.266	0.228	0.414	11.3	5.58%
$COPRA_3$*	0.684	0.359	0.347	0.452	6.4	12.64%
SLPA*	0.86	0.633	0.564	0.608	2.12	2.20%
CDLPOV f_d*	0.852	0.711	0.518	0.621	4	8.82%
CDLPOV f_{cc}*	0.852	0.711	0.518	0.621	4	8.82%
CDLPOV f_b*	0.857	0.65	0.44	0.68	5	2.94%
CDLPOV f_{cl}*	**0.86**	0.66	0.45	0.68	3	5.88%
Dol #2						
CFinder	0.57	0.35	0.26	0.66	4	3.72%
OSLOM	**1.0**	0.914	0.852	0.742	2	1.61%
CONGA	0.85	0.892	0.821	0.746	2	3.22%
$COPRA_2$*	0.933	0.788	0.751	0.693	10.8	0.52%
$COPRA_3$*	0.893	0.767	0.701	0.677	3.7	7.73%
SLPA*	0.56	0.754	0.632	0.742	3.44	2.00%
CDLPOV f_d*	**1.0**	**1.0**	**1.0**	**0.796**	2	0.0%
CDLPOV f_{cc}*	**1.0**	**1.0**	**1.0**	**0.796**	2	0.0%
CDLPOV f_b*	0.9	0.93	0.89	0.7	2	3.22%
CDLPOV f_{cl}*	1	1	1	0.7	2	3.22%
Foot #12						
CFinder	0.701	0.64	0.55	0.51	13	6.9%
OSLOM	0.814	0.704	0.55	0.847	2	1.90%
CONGA	0.823	0.321	0.423	0.451	11	60.0%
$COPRA_2$*	0.933	0.788	0.705	0.693	10.8	0.52%
$COPRA_3$*	0.944	0.747	0.712	0.668	11.2	2.52%
SLPA*	0.748	0.684	0.612	**0.715**	10.30	1.69%
CDLPOV f_d*	0.854	0.865	0.751	0.699	11	0.0%
CDLPOV f_{cc}*	0.854	0.865	0.685	0.699	11	0%
CDLPOV f_b*	0.86	0.37	0.39	0.565	11	9.56%
CDLPOV f_{cl}*	0.86	0.37	0.39	0.565	11	9.56%
Pol #4						
CFinder	**0.855**	**0.740**	**0.79**	**0.884**	4	(9)
OSLOM	**0.954**	**0.802**	**0.759**	0.696	12	0.0%
CONGA	0.688	0.651	0.49	0.779	4	4.16%
$COPRA_2$*	0.687	0.637	0.385	0.825	3	1.05%
$COPRA_3$*	0.702	0.649	0.416	0.827	2.8	6.47%
SLPA*	0.755	0.648	0.497	0.83	3.40	12.5%
CDLPOV f_d*	0.784	0.654	0.495	0.844	3	1.90%
CDLPOV f_{cc}*	0.5788	0.667	0.504	0.834	2	0%
CDLPOV f_b*	0.8	0.63	0.47	0.82	2	1.9%
CDLPOV f_{cl}*	0.8	0.65	0.49	0.83	2	0.952%

Tableau 3.3 – (*) algorithmes fondés sur la propagation des labels, F_1 : F_1-score, Ω :indice d'omega, NMI : information mutuelle normalisée, Q_{Ov}^{Nic} : modularité de Nicosia, # : nombre de communautés et % : pourcentage du chevauchement.

fonctions d'appartenance (Attal *et al.* [10]) qui sont soit basées sur des mesures topologiques globales comme la densité et le coefficient de clustering (f_d et f_{cc}) soit sur les moyennes des mesures de centralités d'intermédiarité et de proximité des nœuds constituant les communautés pré-calculées (f_b et f_{cl}).

Les résultats expérimentaux sont relativement bons et compétitifs en comparaison avec d'autres méthodes de détection de communautés chevauchantes en terme de qualité mais dépendent bien de la topologie du graphe. La performance de notre approche dépasse COPRA et nous constatons une meilleure stabilisation pour les réseaux de Zachary, Dolphin et les livres politiques. D'un autre côté, les algorithmes basés sur la propagation des labels impliquent un nombre plus élevé de communau-

tés, l'algorithme CDLP avec utilisation des fonctions f_d and f_{cc} permet de réduire ce nombre de communautés grâce à l'utilisation de la matrice de fréquences qui permet de stabiliser la propagation des labels.

Chapitre 4

Analyse des réseaux multicouches pour des données ouvertes et privées

Sommaire

4.1 Introduction

Les réseaux complexes constituent aujourd'hui un outil important permettant de décrire et d'analyser des systèmes complexes qui peuvent être représentés sous forme de graphes mathématiques. Il existe de nombreuses applications qui illustrent cette utilisation comme les sciences sociales, biologiques, physiques, de l'information et de l'ingénierie (Gosak *et al.* [51], Kivelä *et al.* [72], Newman [105]).

Les études ont montré que les réseaux complexes réels présentent des propriétés intéressantes nous citons : la distribution des degrés en loi de puissance, la propriété

du petit monde, l'existence de nœuds jouant des rôles centraux et/ou l'existence de structures modulaires (Newman [105]).

Récemment, de plus en plus de travaux étudient les réseaux avec plusieurs types de liens aussi appelés *réseaux de réseaux*. Des variantes de ces systèmes ont été examinées il y a des décennies dans des disciplines telles que la sociologie et l'ingénierie, mais ce n'est que récemment qu'elles ont été unifiées, avec d'autres nomenclatures, dans le cadre de réseaux multicouches définis par Kivelä *et al.* [71].

Parallèlement, de nombreux aspects des systèmes réels sont de plus en plus régulièrement détectés, mesurés et décrits, ce qui donne lieu à de nombreux ensembles de données privés mais aussi ouvertes. Par données privées, nous entendons les données collectées en interne dans une entreprise ou une institution. D'un autre côté, les données ouvertes font référence à l'idée que certaines données devraient être librement accessibles à tous pour être utilisées et re-publiées à volonté, sans restrictions de droit d'auteur, de brevets ou d'autres mécanismes de contrôle.

Dans de nombreux domaines, les référentiels publics d'ensembles de données ouverts constituent une excellente opportunité pour les experts pour contextualiser leurs données générées de manière privée par rapport aux données disponibles publiquement dans leur domaine.

Dans ce chapitre, nous présentons une méthodologie d'analyse de réseau multicouche afin de fournir aux experts des mesures et des méthodes pour comprendre, évaluer et compléter leurs données privées en les comparant et/ou combinant avec des données ouvertes lorsque les deux sont modélisés par des réseaux multicouches.

Les principales contributions sont (Malek *et al.* [91]) :

1. Proposition d'un nouveau formalisme de réseau multicouche qui permet de réaliser une analyse fine en considérant deux niveaux : le niveau intra-couche et celui inter-couche.
2. Définition *du réseau multicouche privé* qui correspond au graphe induit élaboré à partir des données privées, ce réseau sera analysé et comparé à l'ensemble du réseau.
3. Définition de la notion *du réseau égocentrique privé* : la notion de réseau égocentrique qui se définit autour d'un nœud égo proposé dans (Djemili *et al.* [40], Marsden [94]) est étendue à un réseau égocentrique autour d'un réseau privé multicouche. Le réseau égocentrique privé peut être utilisé pour évaluer la force de connectivité des données privées par rapport à l'ensemble du réseau à travers les différentes couches. Le réseau égocentrique privé peut également aider à centrer l'étude du réseau privé dans l'espace de ses voisins et à travers les couches notamment dans le cadre de réseaux ouverts à très grande échelle.
4. Définition *des métriques d'accessibilité inter-couches* d'un sous-réseau donné : ces mesures sont basées sur le réseau égocentrique privé et permettent d'apprécier la force de connectivité des données privées à travers ces couches.

Nous illustrons notre méthodologie via une application biologique. Le réseau multicouche ouvert est construit à partir de bases de données ouvertes constituées des interactions pondérées entre des molecules. Ces interaction peuvent être de types : protéines-protéines, métabolites-métabolites ou protéines-métabolites. Les données privées correspondent à un ensemble de protéines et de métabolites collectés ex-

périmentalement [1] et sont identiques à un sous-ensemble de nœuds dans le réseau multicouche ouvert.

En appliquant cette méthodologie aux données biologiques, nous montrons comment elle peut aider les biologistes à compléter, évaluer et interpréter leurs données privées en utilisant le réseau ouvert : des interactions pondérées entre les molécules collectées privées sont ajoutées en utilisant le réseau ouvert. Le réseau égocentrique privé est analysé et les métriques d'accessibilité des couches sont calculées et discutées. La connectivité entre les molécules inter-couches et entre les couches est calculée et la distribution des molécules identifiées dans le réseau ouvert est observée et interprétée, les accessibilités entre les couches sont calculées en plus les chemins les plus courts qui sont biologiquement significatifs sont également analysés et classés.

Ce travail a été réalisé dans la cadre du projet ANR BLIZAAR : `https://anr.fr/Projet-ANR-15-CE23-0002` ; un prototype a été développé pour illustrer et expérimenter cette méthodologie en collaboration avec des chercheurs du laboratoire LIST, les résultats obtenus ont été publiés dans le journal *Applied Network Science* (Malek *et al.* [91]).

4.2 Elements d'analyse des réseaux multicouches

Nous présentons tout d'abord un nouveau formalisme de réseau multicouche ainsi que des exemples montrant comment nous étendons les définissions des mesures globales et locales des réseaux multicouches. Nous proposons également une définition formelle d'un (*réseau égocentrique multicouche*, du *réseau privé multicouche* et du ıréseau égocentrique privé. Nous montrons ensuite comment nous pouvons utiliser ces notions pour définir *les accessibilités entre couches* et au niveau *inter-couche* (Malek *et al.* [91]).

4.2.1 Notations, Propriétés et métrique

Nous représentons un réseau multicouche par un tuple qui contient un ensemble de sommets, un ensemble d'arêtes intra-couches et un ensemble d'arêtes inter-couches.

Soit $\mathbb{N} = (\mathbb{V}, \mathbb{E}, \mathbb{C})$ un graphe contenant l couches (voir figure 4.1)

1. $\mathbb{V} = \{V_1, .., V_i, ..V_l\}$ est l'ensemble de sommets contenus dans les couches, l étant le nombre de couches $l > 1$, V_i est l'ensemble des nœuds de la couche numéro i, $V_i = \{v_1^i, ..v_{n_i}^i\}$, $n_i =\mid V_i \mid$
2. $\mathbb{E} = \{E_1, .., E_i, ..E_l\}$ est l'ensemble des liens intra-couche : E_i est l'ensemble de liens dans la couche i, avec $\mid E_i \mid = m_i$. $E_i = \{(v_j^i, v_k^i) \mid v_j^i \in V_i, v_k^i \in V_i\}$ [2]
3. $\mathbb{C} = \{C_{i_1 j_1}, ..C_{i_b j_b} \mid i_k \neq j_k\}$ est l'ensemble des liens inter-couche, b étant le nombre des composantes biparties. $C_{ij} = \{(v_k^i, v_{k'}^j) \mid v_k^i \in V_i, v_{k'}^j \in V_j\}$, avec $\mid C_{ij} \mid = c_{ij}$.

1. Ces données ont été rendues disponibles par des chercheurs en biologie au laboratoire LIST dans le cadre du projet ANR BlIZAAR : `https://anr.fr/Projet-ANR-15-CE23-0002`

2. un lien est noté (v, u) si le graphe est orienté sinon il est noté $\{v, u\}$ si le graphe n'est pas orienté

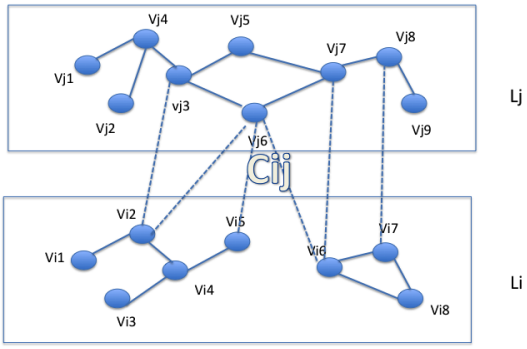


FIGURE 4.1 – Exemple d'un réseau à 2 couches.

Cette représentation permet de proposer une adaptation des métriques globales et locales en prenant en compte les deux types de composantes : intra-couches et inter-couches. Nous pouvons ensuite agréger ces métriques afin de proposer une métrique pour l'ensemble du réseau. Par exemple, nous proposons la métrique suivante pour la densité :

- Densité intra-couche pour la couche i : $D_i = \frac{m_i}{\frac{n_i*(n_i-1)}{2}}$
- Densité inter-couche pour la composante bipartie C_{ij} : $D_{ij} = \frac{c_{ij}}{n_i*n_j}$
- La densité globale est : $D = \frac{\sum_{C_{ij}\in\mathbb{C}} c_{ij}+\sum_{l\in\{1..l\}} m_l}{\sum_{C_{ij}\in\mathbb{C}} n_i*n_j+\sum_{l\in\{1..l\}} \frac{n_i*(n_i-1)}{2}}$

De même, la centralité de degré peut être généralisée au niveau inter-couche et à l'ensemble du réseau. La centralité de degré ainsi que la connectivité d'un sommetv_j^i appartenant à a couche V_i sont données par :

- Centralité de degré intra-couche : $CD(v_j^i) = \frac{deg_i(v_j^i)}{n_i-1}$ avec $n_i = \mid V_i \mid$ et $deg_i(v_j^i)$ est le degré of v_k^i dans la couche i.
- Connectivité d'un sommet inter-couche : nous définissons la connectivité d'un sommet dans une composante bipartie C_{ik} par $CN_k(v_j^i) = \frac{deg_{C_{ik}}(v_k^i)}{n_k}$ avec $n_k = \mid V_k \mid$ et $deg_{C_{ik}}(v_k^i)$ étant le degré de v_k^i dans C_{ik}
- Connectivité globale : nous proposons de généraliser la définition de la connectivité d'un sommet au réseau multicouche ainsi : $CN(v_j^i) = \frac{deg_i(v_j^i)+\sum_k CN_k(v_j^i)}{n_i-1+\sum_k n_k}$

4.2.2 Réseaux égocentriques multicouches

Dans un réseau complexe (et plus particulièrement un réseau social en ligne), le réseau égocentrique se définit autour d'un nœud égo u est un sous-réseau contenant l'ego u et les altères (les voisins) ainsi que l'ensemble des liens du réseau égo. Dans la littérature, deux cas de réseaux personnels en ligne sont identifiés en fonction de la distance du altère de l'ego : niveau 1 et niveau k. Soit $G = (V, E)$ un graphe représentant un réseau et u un sommet, le réseau égocentrique à 1 niveau de u : $G^u = (V^u, E^u)$ est défini par le graphe (voir figure 4.2) :

- $V^u = \{x \in V \mid (u, x) \in E\} \cup \{u\}$
- $E^u = \{(x, y) \in E \mid x \in V^u \wedge y \in V^u\}$

Nous proposons une extension de cette définition aux réseaux multicouches qui permet d'accéder aux altères situés dans la même couche ainsi qu'à ceux appartenant

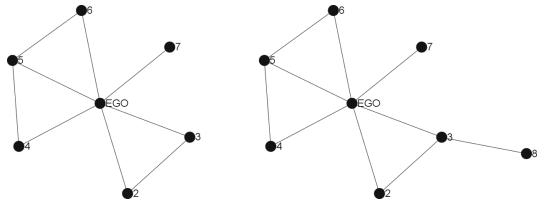


FIGURE 4.2 – Réseau égocentrique à 1 et 2 niveaux (Malek *et al.* [91]).

aux couches connectées à celle de l'ego (voir figure 4.3).

$u \in V_i$, $\mathbb{N}^u = G(V^u, E^u)$

— $V^u = \{x \in V_i \mid (u,x) \in E\} \cup \{u\} \cup_k \{y \in V_k \mid (u,y) \in C_{ik}\}$

— $E^u = \{(x,y) \in E_i \mid x \in V^u \wedge y \in V^u\} \cup_k \{(u,y) \in C_{ik}\}$

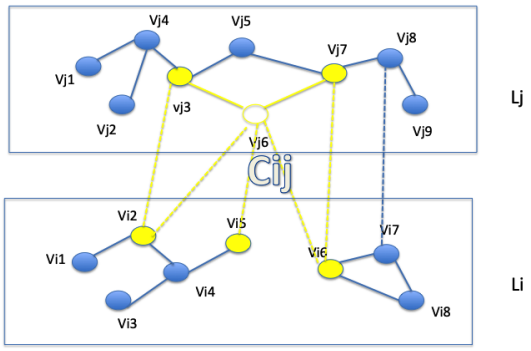


FIGURE 4.3 – Exemple d'un réseau égocentrique à 2 couches (Malek *et al.* [91]).

4.2.3 Réseau multicouche et réseaux égocentrique privé

Comme mentionné précédemment, le but de cette étude est de fournir aux experts d'un domaine donné des mesures et des méthodes pour comprendre, évaluer et compléter leurs données privées en les comparant et/ou en les combinant avec des données ouvertes lorsque les deux sont modélisées par des réseaux multicouches.

Dans notre cas, les données privées sont un sous-ensemble de nœuds qui sont identifiés dans le réseau ouvert. Les interactions entre ces nœuds privés sont extraites du réseau ouvert.

Nous proposons donc d'étudier le graphe induit élaboré à partir des données privées. Celui-ci doit être construit, analysé et comparé à l'ensemble du réseau (ouvert) (voir figure 4.4).

Soit $\mathbb{N}$ un réseau multicouche (extrait des données ouvertes) : $\mathbb{N} = (\mathbb{V}, \mathbb{E}, \mathbb{C})$ contenant l couches. Soit PV un ensemble de sommets $PV = \{PV_1, ..PV_z\}$ tel que : $PV_i \subset V_i$ (données privées). Nous définissons le réseau multicouche privé par $\mathbb{N}[PV] = (\mathbb{PV}, \mathbb{PE}, \mathbb{PC})$ avec

1. $\mathbb{PE} = \{PE_1, ..PE_i, ..PE_z\}$ est l'ensemble des liens intra-couche :
   PE_i étant l'ensemble des liens appartenant à la couche numéro i et donné par : $PE_i = \{(pv^i_j, pv^i_k) \in E_i \mid pv^i_j \in PV_i, pv^i_k \in PV_i\}$
2. $\mathbb{PC} = \{PC_{i_1 j_1}, ..PC_{i_b j_b} \mid i_k \neq j_k\}$ est l'ensemble des liens inter-couche :
   $PC_{ij} = \{(pv^i_k, pv^j_{k'}) \in C_{ij} \mid pv^i_k \in PV_i, pv^j_{k'} \in PV_j\}$.

Dans la figure 4.4, le graphe représente le réseau multicouche $\mathbb{N}$ extrait des données ouvertes, les sommets rouges représentent les données privées et le graphe rouge illustre le réseau multicouche privé $\mathbb{N}[PV]$.

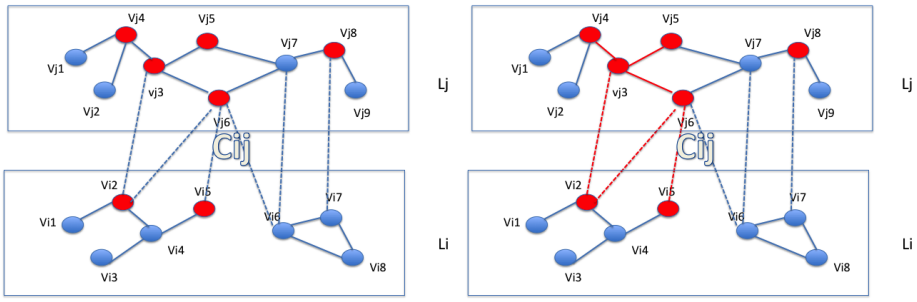


FIGURE 4.4 – Réseau multicouche et données privées : le graphe représente le réseau ouvert, les sommets rouges (à gauche) sont des données privées. Le graphe rouge (à droite) représente le réseau privé (Malek *et al.* [91]).

Nous étendons maintenant la définition du réseau égocentrique (qui est défini autour d'un nœud égo donné (Marsden [94], Djemili *et al.* [40]) à un réseau égocentrique autour d'un réseau privé multicouche.

Nous définissons le réseau égocentrique privé comme suit : Soit $\mathbb{N}[PV] = (\mathbb{PV}, \mathbb{PE}, \mathbb{PC})$ un réseau multicouche privé : nous définissons le réseau égocentrique privé : $\mathbb{N}^{PV} = G(V^{PV}, E^{PV})$

— $V^{PV} = \bigcup_{u \in PV}\{x \in V_i \mid (u,x) \in E_i\} \bigcup \{u\} \bigcup_k \{y \in V_k \mid (u,y) \in C_{ik} \mid C_{ik} \in \mathbb{C}\}$

— $E^{PV} = \bigcup_{u \in PV}\{(x,y) \in E_i \mid x \in V^u \wedge y \in V^u\} \bigcup_k \{(u,y) \in C_{ik} \mid C_{ik} \in \mathbb{C}\}$

Dans la figure 4.5, les nœuds rouges représentent les données privées et le graphe contenant les nœuds et les arêtes rouges et jaunes illustre le réseau égocentrique privé $\mathbb{N}^{PV}$

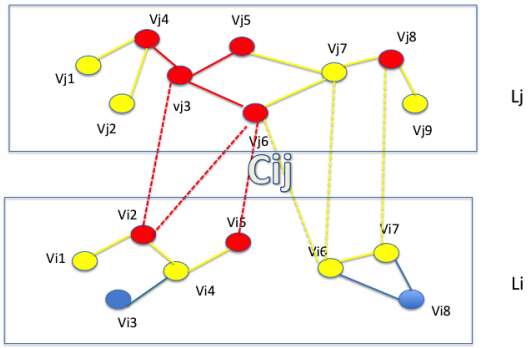


FIGURE 4.5 – Réseau complexe multicouche : le réseau égocentrique privé est représenté par les nœuds et les arêtes rouges et jaunes (Malek *et al.* [91]).

4.2.4 Accessibilité de couche et inter-couche d'un sous-réseau

Nous définissons l'accessibilité d'un graphe pour une couche donnée comme suit.

Soit $\mathbb{N} = (\mathbb{V}, \mathbb{E}, \mathbb{C})$ un réseau multicouche contenant l couches, $G = (V, E)$ est un sous graphe de $\mathbb{N}$ et i est une couche donnée.

— *Reachability(G,i) est donnée par le sous graphe* $G_i = (V_i', E_i')$ tel que :

— $V'_i = \{v'^i_j \in V \cap V_i\}$
— $E'_i = \{(v'^i_j, v'^i_k) \in E \cap E_i\}$

Afin d'apprécier la force de connexion entre les nœuds privés à travers la couche, nous appliquons l'accessibilité sur le réseau égocentrique privé calculé pour une couche donnée i vers une autre couche j. Soit $\mathbb{N}^{PV_i} = G(V^{PV_i}, E^{PV_I})$ le réseau privé égocentrique calculé à partir de la couche i, $Reachability(\mathbb{N}^{PV_i}, j)$ est l'accessibilité à partir du graphe privé de la couche i vers une autre couche j et est donné par le graphe $\mathbb{N}'^{PV_i}_j$. Soit V'_j l'ensemble des nœuds de $\mathbb{N}'^{PV_i}_j$, nous pouvons maintenant apprécier le ratio de nœuds privés accessibles sur la couche j en calculant la précision et le rappel comme suit (voir figures 4.6 and 4.7) :

$$precisionR(i,j) = |\frac{V'_j \cap PV_i}{V'_j}|$$

$$recallR(i,j) = |\frac{V'_j \cap PV_i}{PV_i}|$$

$precisionR(i,j)$ donne le rapport des nœuds privés appartenant à la couche j qui sont accessibles depuis la couche i à tous les nœuds accessibles de la couche j.
$recallR(i,j)$ est le rapport des nœuds privés de la couche j accessibles depuis la couche i à tous les nœuds privés appartenant à la couche j.

Nous définissons également un graphe d'accessibilité inter-couche pour une partie bipartite donnée comme suit.

Soit $\mathbb{N} = (\mathbb{V}, \mathbb{E}, \mathbb{C})$ un réseau multicouche contenant l couches, $G = (V, E)$ est un sous graphe de $\mathbb{N}$ and C_{ij} est une composant bipartie

— $InterReachability(G, C_{ij})$ *est donné par le sous-graphe* $G_{ij} = (V'_{ij}, E'_{ij})$
  — $V'_{ij} = \{v'^i_k \in V \cap V_i\} \cup \{v'^i_{k'} \in V \cap V_j\}$
  — $E'_{ij} = \{(v'^i_k, v'^j_{k'}) \in E \cap C_{ij}\}$

Pour une composante bipartie donnée C_{ij}, nous pouvons appliquer l'accessibilité inter-couche à partir du réseau multicouche induit privé ou du réseau égocentrique privé.

Par exemple, soit $\mathbb{N}[PV]$ un réseau multicouche privé, $InterReachability(\mathbb{N}[PV], C_{ij})$ étant le graphe C'_{ij} on peut évaluer les arêtes bipartites atteignables en calculant le rapport $\frac{c'_{ij}}{c_{ij}}$ where $c'_{ij} =\mid C'_{ij} \mid$ and $c_{ij} =\mid C_{ij} \mid$

4.3 Travaux reliés

Plusieurs travaux se sont récemment intéressés aux réseaux ayant plusieurs types de liens aussi appelés *réseaux de réseaux*. Des variantes de ces systèmes ont été examinées il y a des décennies dans des disciplines telles que la sociologie et l'ingénierie, mais ce n'est que récemment qu'elles ont été unifiées, avec d'autres nomenclatures, dans le cadre de réseaux multicouches définis par Kivelä *et al.* [71].

Dans Kivelä *et al.* [71], une revue complète du domaine des réseaux multicouches est présentée : les types de réseaux, les caractéristiques des nœuds et des couches, la notion d'aspect ainsi que la nature du couplage entre couches sont détaillés.

De nombreuses études abordent actuellement des thèmes liés aux réseaux multicouches (Interdonato *et al.* [60]) comme la structure et la dynamique des réseaux

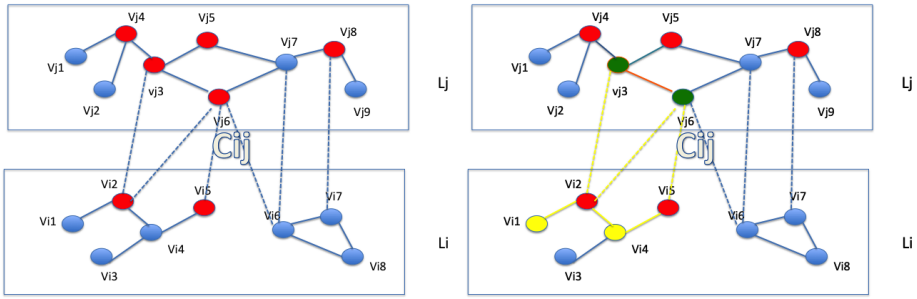


FIGURE 4.6 – Accessibilité de la couche i à la couche j : les nœuds rouges sont privés, les nœuds jaunes et verts sont des nœuds accessibles calculés à partir de la couche i, les nœuds verts sont les nœuds accessibles qui appartiennent au réseau privé de la couche j, precisionR (i, j) = 1 et rappelR (i, j) = 0.4 : cela signifie que tous les nœuds accessibles depuis la couche i sont des nœuds privés mais seulement 40 % des nœuds privés de la couche j sont accessibles depuis la couche i (Malek *et al.* [91]).

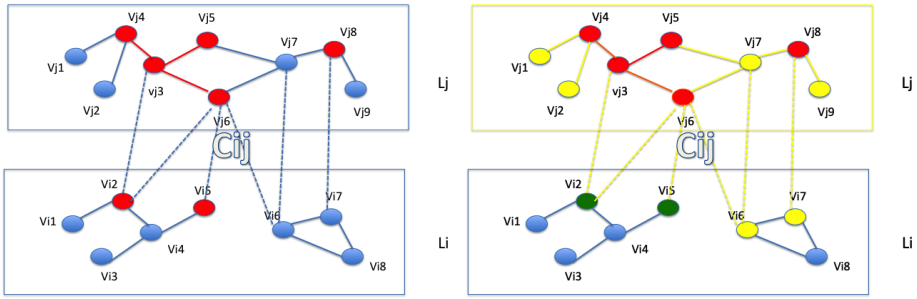


FIGURE 4.7 – Accessibilité de la couche j à la couche i : les nœuds rouges sont privés, les nœuds jaunes et verts sont des nœuds accessibles calculés à partir de la couche j, les nœuds verts sont les nœuds accessibles qui appartiennent au réseau privé de la couche i, precisionR (j, i) = 0.5 et rappelR (j, i) = 1 : cela signifie que 50% des nœuds accessibles depuis la couche j sont des nœuds privés mais tous les nœuds privés de la couche i sont accessibles depuis la couche j (Malek *et al.* [91]).

multicouches (Boccaletti *et al.* [25] Magnani et Rossi [88], Aleta et Moreno [6]) ainsi que la détection de communautés dans les réseaux multicouches (Liu *et al.* [85]) et leur visualisation (McGee *et al.* [97]).

De nombreux travaux montrent également que des experts dans de multiples domaines comme les humanités numériques (McGee *et al.* [97]), la biologie (Gosak *et al.* [51]), la techno-anthropologie, etc. présentent leurs données à l'aide des réseaux multicouches et sont conscients de la forte nécessité de disposer d'outils d'analyse de leur données (Kivelä *et al.* [72]).

Dans Kivelä *et al.* [71], un formalisme général du type le plus général de réseau multicouche a été proposé, un graphe sous-jacent qui représente ce réseau multicouche est défini, où un nœud est représenté par un tuple contenant trois identifiants : celui du nœud même, celui la couche à laquelle appartient ce et celui de l'aspect. De plus, deux types d'arêtes sont proposées : les arêtes intra-couches et les arêtes inter-couches.

Notre formalisme des réseaux multicouches permet de réaliser une analyse fine en considérant deux niveaux : le niveau intra-couche et celui inter-couche. Nous avons

montré ci-dessus des exemples de la manière dont nous pouvons étendre la définition des mesures globales et locales comme la densité et les centralités au niveau inter-couches. Les mesures pour l'ensemble des réseaux sont ensuite calculées en agrégeant les mesures inter et intra couches.

Dans d'autres travaux (Battiston *et al.* [18]), un réseau monoplex est construit en agrégeant les données des différentes couches d'un réseau multicouche, la définition classique du degré de nœud est alors appliquée au réseau monoplex résultant. Cependant, l'agrégation du réseau entraîne une perte d'informations. Dans d'autres travaux, la distinction des couches est maintenue et le degré de nœud est représenté par un vecteur. Il est également possible de définir le degré et le voisinage en termes de nœud focal par rapport à un sous-ensemble du couches (Berlingerio *et al.* [23]).

D'autre part, nous avons défini des *métriques d'accessibilité des couches et inter-couches* d'un sous-réseau donné. Cette mesure est basée sur le réseau égocentrique privé et aide à apprécier la force de connectivité des données privées entre les couches.

Dans Kivelä *et al.* [71], la mesure de l'interdépendance des nœuds est définie comme étant le rapport du nombre des chemins les plus courts dans lesquels deux couches ou plus sont utilisées pour le nombre total de chemins les plus courts. Cette mesure permet de quantifier la valeur ajoutée par la multiplexicité à l'accessibilité des nœuds. L'interdépendance d'un réseau multicouche est calculée comme l'interdépendance moyenne des nœuds.

4.4 Application biologique

Le but de cette application est d'étudier certains ensembles de données biologiques collectées dans des échantillons expérimentalement (des échantillons de cannabis). Les molécules identifiées (protéines et métabolites) sont mesurées à partir des données biologiques collectées dans différentes expériences *omiques* : transcriptomique, protéomique et métabolomique. Dans leurs expériences, les biologistes mesuraient à plusieurs moments, des contigs : chacun quantifie des gènes, des spots : chacun quantifie une ou plusieurs protéines, et métabolites. Chaque gène produit généralement une protéine mais parfois plusieurs protéines [3].

À ce stade, nous n'avons que des nœuds (mais pas d'arêtes), correspondant aux molécules mesurées dans les expériences. Pour obtenir des relations, les biologistes utilisent en général la base de données ouverte STRING (outil de recherche pour la récupération des gènes/protéines avec leurs interactions) (Szklarczyk *et al.* [130]), qui est la principale base de données d'interactions protéine-protéine (et aussi gène-gène) ainsi que la base de données STITCH (outil de recherche pour les interactions de produits chimiques). STITCH (Szklarczyk *et al.* [131]) est une base de données jumelle incluant les liens entre les métabolites ainsi qu'entre les protéines et les métabolites (voir figure 4.8). Chaque interaction dans les deux bases de données est basée sur la présence d'une co-expression expérimentale (comportement similaire dans plusieurs expériences publiques disponibles), d'une co-ocuurence textuelle (co-occurence dans la même phrase), d'un chemin biologique (participant au même réseau biologique connu), etc. Un score combiné agrégeant tous ces types d'interactions dont la valeur

3. Ces données ont été rendues disponibles par des chercheurs en biologie au laboratoire LIST dans le cadre du projet ANR BlIZAAR : `https://anr.fr/Projet-ANR-15-CE23-0002`

est comprise entre 0 et 1000 est ajouté aux deux bases de données (voir tableaux 4.1, 4.2 et 4.3).

Le réseau multicouche ouvert est construit à partir des bases de données ouvertes STRING et STITCH (voir figure 4.8). Des interactions pondérées (liens) entre protéines-protéines, métabolites-métabolites et protéines-métabolites sont créées en fonction de la valeur du score combiné. Les données privées sont l'ensemble des protéines et métabolites identifiés collectés expérimentalement en laboratoire par les biologistes comme mentionné ci-dessus et correspondent à un ensemble de nœuds dans l'ensemble des données ouvertes comme expliqué dans la figure 4.4.

Une fois le réseau de régulation est construit, il sera analysé. La complexité du réseau doit être réduite, en sélectionnant des interactions significatives. Les biologistes doivent identifier les chemins les plus courts des nœuds clés (molécules), les trier via des mesures de centralité entre deux molecules (nœuds) données et suivre le chemin d'un récepteur vers les facteurs de transcription et vice versa.

D'un point de vue biologique, nous rappelons que :

1. Les biologistes sont souvent intéressés par trouver des voisins de molécules (et plus particulièrement des protéines), d'où la nécessité d'analyser le réseau égocentrique privé.
2. Les biologistes doivent extraire et analyser la transduction du signal et les voies métaboliques du réseau Le chemin le plus court est biologiquement significatif car il est énergétiquement le plus favorable pour détecter les interactions de transduction du signal ainsi que les voies métaboliques.

   La transduction du signal représente une série d'interactions entre différentes entités biologiques telles que des protéines, des produits chimiques ou des macromolécules afin d'étudier comment la transmission du signal est effectuée soit de l'extérieur vers l'intérieur de la cellule, soit à l'intérieur de la cellule.

   De même, les voies métaboliques sont liées à une série de réactions chimiques se produisant dans une cellule à différents moments, contenant des informations sur une série d'événements biochimiques et la manière dont ils sont corrélés que nous considérons.

protein1	protein2	coexpression	experimental	database	textmining	combined_score
AT1G01010.1	AT1G02220.1	102	0	0	222	298
AT1G01010.1	AT1G02230.1	291	0	0	176	415
AT1G01010.1	AT1G02250.1	0	0	0	202	202

Tableau 4.1 – Exemples d'interactions entre proteins extraits de la base de données ouverte STRING utilisée pour construire la couche de protéines.

Pour analyser le réseau biologique, nous procédons comme suit (Malek *et al.* [91]) :

1. Analyse de chaque couche (couches protéines et métabolites)
   (a) Le réseau est construit à partir de la base de données ouverte (STRING et STITCH). Le réseau privé est également construit à partir de l'ensemble des molécules identifiées (protéines et métabolites) collectés à partir de l'expérience sur les échantillons de cannabis où les données biologiques

chemical1	chemical2	similarity	experimental	database	textmining	combined_score
CIDm00024759	CIDs00024759	0	0	900	0	900
CIDs91758695	CIDs00107694	0	0	0	230	230
CIDs91758695	CIDs11013287	0	0	0	230	230

Tableau 4.2 – Exemples d'interactions entre métabolites extraits de la base de données ouverte STITCH utilisée pour construire la couche de métabolites.

chemical	protein	experimental	prediction	database	textmining	combined_score
CIDs91758425	AT1G09340.1	0	0	0	250	250
CIDs91758425	AT2G42600.1	0	0	0	300	300
CIDs91758423	AT1G04070.1	0	0	0	153	153

Tableau 4.3 – Exemples extraits de la base de données ouverte STITCH utilisée pour construire des interactions métabolites-protéines .

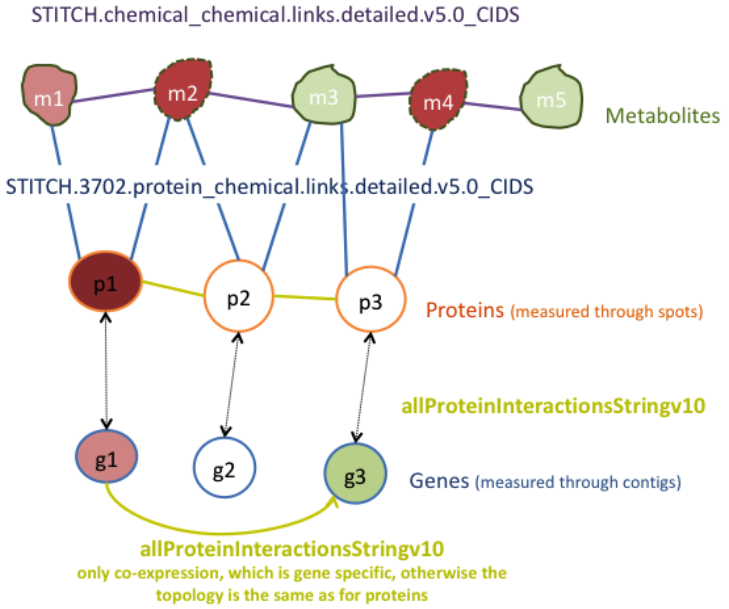


FIGURE 4.8 – Construction du réseau à partir des bases de données ouvertes : les interactions protéine-protéine sont extraites de la base de données STRING. Les métabolites-métabolites et métabolites-protéines sont extraites de la base de données STITCH (Malek *et al.* [91]).

sont collectées dans différentes expériences "omiques" : transcriptomique, protéomique et métabolomique.

(b) Les mesures globales et locales sont comparées et discutées pour les réseaux ouverts et privés, nous appliquons l'algorithme de Louvain (Blondel *et al.* [24]) afin de détecter les communautés, la distribution des molécules privées (identifiées par l'expérience) est étudiée en fonction des communautés détectées.

2. Analyse de réseau multicouche :

(a) Le composant bipartite contenant les interactions protéines-métabolites est construit à partir de la base de données ouverte STITCH. le réseau

multicouche correspondant aux donnée ouvertes, le réseau privé multicouche ainsi que le réseaux égocentrique privé sont également construits.

(b) Les mesures globales et locales de ces réseaux sont comparées et discutées.

(c) Les accessibilités des couches des métabolites aux protéines et des protéines aux métabolites sont calculées et discutées.

(d) Les chemins les plus courts entre paires de protéines privées sont ensuite analysés et classés en fonction de leur emplacement dans le réseau ouvert (privé, égocentrique ou extra-égocentrique). L'ensemble de données ouvertes KEGG [67] est également utilisé pour décrire les chemins.

Les bases de données biologiques ouvertes sont très volumineuses par rapport à l'expérience biologique. Dans notre cas, les protéines de l'expérience ne représentent que 0,58% à 0,74% du nombre total de protéines dans le réseau entier. De même, les métabolites de l'expérience représentent moins de 0,05% du nombre total de métabolites dans tout le réseau.

4.4.1 Analyse de la couche de protéines

Le tableau 4.4 montre la distribution des valeurs de l'attribut combined_score. Les scores combinés sont les poids des interactions entre deux protéines selon la base de données ouverte STRING.

Min	1 Qu	Median	Mean	3 Qu	Max
150.0	184.0	238.0	325.1	377.0	999.0

Tableau 4.4 – La distribution du score combiné selon la base de données STRING.

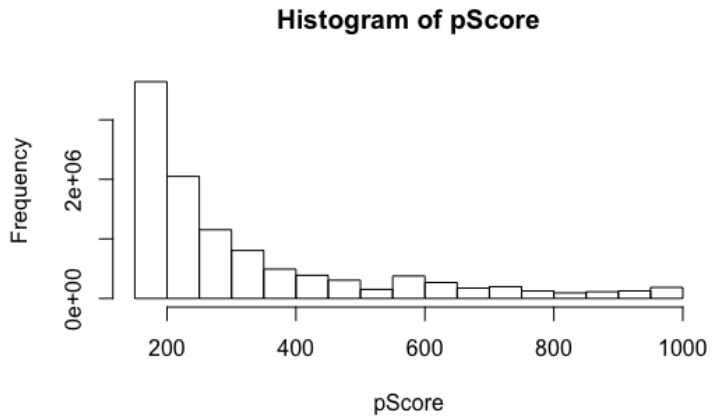


FIGURE 4.9 – La distribution du score combiné extrait de la base de données ouverte STRING.

La figure 4.9 montre que la distribution des valeurs du scores combiné est similaire au comportement de réseau invariant d'échelle (scale free network).

1. **Construction du réseau**

   Nous construisons la couche de protéines en considérant le score combiné comme seuil : si nous prenons le score minimal (150) toutes les interactions sont considérées sinon une partie du réseau est considérée selon le percentile choisi. Lorsque la valeur du score combinée est incrémentée, certains nœuds seront déconnectés. Ces nœuds sont supprimés du réseau. Le réseau privé est construit également.

   Les protéines identifiées dans l'expérience ne forment que 0,58% à 0,74% du nombre total de protéines dans le réseau ouvert (voir tableau 4.5).

   Les expériences montrent que la variation de densité entre réseaux ouvert et privé est minime, ce qui signifie que les protéines identifiées (dans l'expérience) sont réparties d'une façon presque équilibrée dans le réseau ouvert.

2. **Distribution de degrés** :

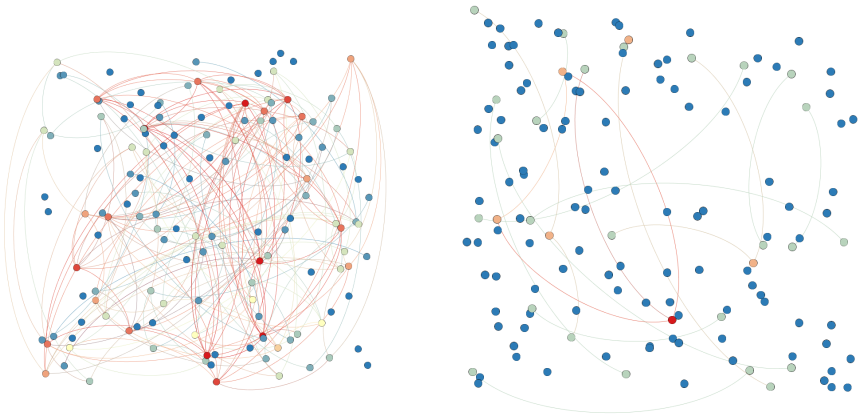


FIGURE 4.10 – Réseaux privés de protéines : celui de gauche correspond au score combiné minimal (150), celui de droite correspond à des valeurs de score supérieures à 500. Les couleurs chaudes des nœuds indiquent des centralités de haut degré. Les couleurs chaudes des liens indiquent des poids élevés (Malek *et al.* [91]).

   Les résultats montrent que les valeurs moyennes des degrés des protéines identifiées ne varient pas beaucoup par rapport aux autres protéines. Ceci est cohérent avec l'observation sur les densités de réseaux comme mentionné ci-dessus (voir tableau 4.5)

3. **Détection de communautés** : Nous appliquons l'algorithme de Louvain à la couche des proteins [4] [67]. Huit communautés sont détectées pour le score combiné minimal. On remarque que les valeurs de la précision ne varient pas beaucoup selon les communautés. Cela signifie que les protéines identifiées sont presque distribuées de manière équilibrée dans les communautés.

Le tableau 4.5 présente un résumé des résultats et des observations concernant l'analyse de la couche des protéines.

4. nous considérons la composante principale connectée du réseau

	Réseau ouvert	**Protéines identifiées**	**Réseau privé**	**Observations**
Nœuds#	24283	142	142	les protéines identifiées dans l'expérience ne présentent que 0,58 % de l'ensemble des protéines
Densité(moy)	0.018		0.019	Les réseaux ouverts et privés ont presque la même densité
Degrés(moy)	438.0	423.60	2.6	Les valeurs de degré des protéines identifiées ne varient pas beaucoup par rapport aux autres protéines
Communautés	8 communitiés	precision $\in [0,48\%; 0,77\%]$		Les protéines identifiées sont presque distribuées de manière équilibrée dans les communautés.

Tableau 4.5 – Analyse de la couche de protéines : résumé des résultats pour le score combiné minimal.

4.4.2 Analyse de la couche des métabolites

Le tableau 4.6 montre la distribution des valeurs du score combiné pour les métabolites. Les scores combinés expriment les poids des interactions entre deux métabolites selon la base de données ouverte STITCH.

Min	1 Qu	Median	Mean	3 Qu	Max
2.0	175.0	231.0	299.8	354.0	999.0

Tableau 4.6 – La distribution du score combiné selon la base de données STITCH.

La figure 4.11 montre que la distribution des valeurs du score combinée est similaire au comportement d'un réseau à échelle invariant.

1. **Construction du réseau**

   Comme pour la couche des protéines, nous construisons la couche métabolites en considérant le score combiné comme seuil, si nous prenons le score minimal (2) l'ensemble du réseau complet est considéré sinon une partie du réseau est considérée selon le percentile choisi.

   Nous remarquons que les métabolites identifiés dans l'expérience présentent moins de 0,05% du nombre total de métabolites dans le réseau entier.

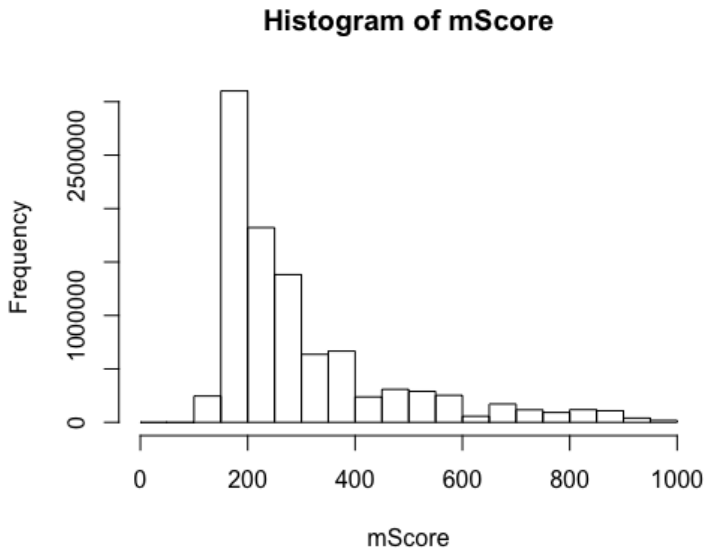


FIGURE 4.11 – La distribution du score combiné extrait de la base de données ouverte STITCH.

Cependant, les réseaux privés de métabolites extraits de l'expérience présentent une densité élevée par rapport aux réseaux ouverts, ce qui signifie que les métabolites de l'expérience sont fortement connectés par paires (voir tableau 4.7).

2. **La distribution des degrés** :
   Les résultats montrent que les métabolites identifiés ont des centralités de degré très élevés. Cela signifie qu'ils sont fortement connectés connectés par paires selon la base de données ouverte STITCH (voir tableau 4.7).

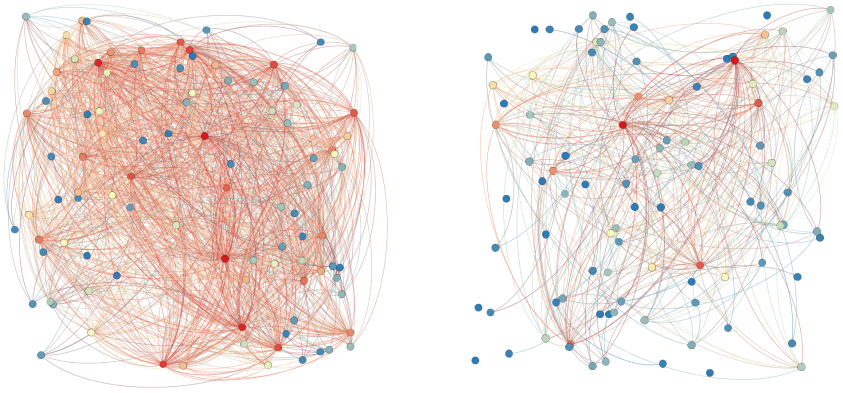


FIGURE 4.12 – Réseaux privés de métabolites : celui de gauche correspond au score combiné minimal (2), celui de droite correspond aux scores supérieurs à 500. Les couleurs chaudes des nœuds indiquent des centralités de haut degré. Les couleurs chaudes des liens indiquent des poids élevés (Malek *et al.* [91]).

3. **Détection de communautés** :

   On applique l'algorithme de Louvain à la couche de métabolites (Blondel *et al.* [24]), on obtient une modularité de 0,46. 39 communautés sont détectées à partir de la composante principale connectée pour le score combiné minimal :
   — 22 ont leurs cardinalités entre 3 et 28
   — 11 ont leurs cardinalités entre 1000 et 10000
   — 6 ont leurs cardinalités entre 15000 et 30000

   On constate qu'une majorité des métabolites sont présents dans seulement deux communautés. Ces résultats sont corrélés avec la valeur de densité élevée du réseau privé de métabolites et signifient que les métabolites sont fortement connectés et forment principalement deux sous-réseaux hautement connectés.

Le tableau 4.7 présente un résumé des résultats et des observations concernant l'analyse de la couche de métabolites.

	Réseau ouvert	**Métabolites identifiés**	**Réseau privé**	**Observations**
Nœuds#	205398	97	97	Les métabolites identifiés dans l'expérience présentent moins de 0,05% de l'ensemble des métabilites
Densité(moy)	0.00023		0.26009	Le réseau privé a une densité élevée par rapport au réseau ouvert.
Degrés(moy)	47.06	776.8	24.97	Les métabolites identifiés ont des centralité de degrés élevées. Cela signifie qu'ils sont fortement connectés à d'autres métabolites
Communautés	39 communities	80% appartiennent à 2 communautés		Une majorité des métabolites identifiés sont fortement connectés et forment deux sous-réseaux hautement connectés

Tableau 4.7 – Analyse de la couche de métabolites : résumé des résultats pour le score combiné minimal.

4.4.3 Analyse du réseau multicouche métabolites-proteines

La figure 4.13 montre que la distribution des valeurs des scores combinés extraits de la base de données STITCH concernant la partie bipartite est similaire au comportement de réseau à échelle invariant.

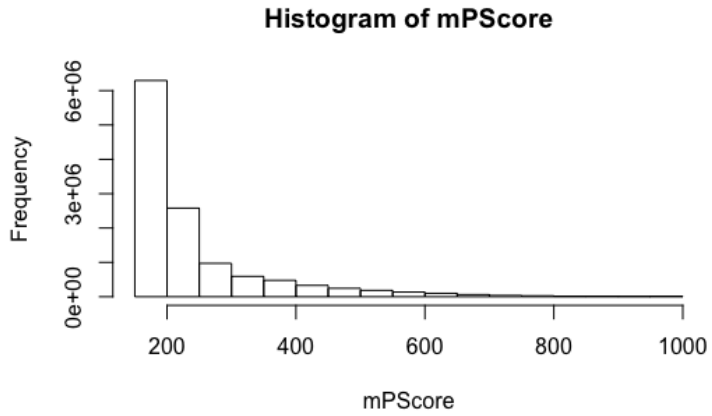


FIGURE 4.13 – La distribution du score combiné extrait de la base de données ouverte STITCH.

1. **Construction de réseaux** : nous construisons la partie bipartite protéine-métabolite en considérant le score combiné comme seuil, si nous prenons le score minimum, le réseau entier est construit sinon une partie du réseau est considérée selon le percentile choisi. Le réseau à 2 couches est ensuite construit en considérant les deux couches de protéines et de métabolites. Le réseau privé à 2 couches extrait de l'expérience ainsi que le réseau égocentrique privé sont également construits.

   Les résultats montrent que :
   — Le ratio de molécules identifiées (protéines et métabolites) dans l'expérience est de 0,1% par rapport aux réseaux ouverts à 2 couches mais diminue à [0,9 %, 1,42 %] dans le réseau égocentrique privé.
   — la densité des réseaux privés obtenue à partir de l'expérience est 100 à 165 plus grande que celle du réseau ouvert mais elle n'est que 1,65 plus grande que celle du réseau égocentrique.
2. **Mesures d'accessibilités aux protéines et métabolites** :

   Le tableau 4.8 montre que les réseaux privés égocentriques des métabolites atteignent (voir section 4.2.4) un ensemble de protéines contenant de 51% à 67% des protéines identifiées malgré une très faible précision.

   De même, le tableau 4.9 montre que les réseaux privés égocentriques de protéines atteignent (voir la section 4.2.4) un ensemble de métabolites qui contient 53% à 84% des métabolites identifiés malgré une très faible précision.

   Ces résultats montrent qu'une majorité des métabolites identifiés (dans l'expérience) sont accessibles à partir de toutes les protéines identifiées et vice-versa. Cela aidera le biologiste à classer les molécules en molécules voisines et distantes.

Percentiles	precision	recall
min	0.0064	0.67
median	0.006	0.59
mean	0.006	0.51

Tableau 4.8 – L'accessibilité des protéines identifiés à partir des réseaux privés égocentriques des métabolites selon les percentiles de score combinés.

Percentiles	precision	recall
min	0.0056	0.84,
median	0.007	0.71
mean	0.007	0.53

Tableau 4.9 – L'accessibilité des métabolites identifiés à partir des réseaux privés égocentriques des proteins selon les percentiles de score combinés.

Nous présentons dans les sections suivantes deux méthodes d'analyse des voies biologiques entre les pairs des protéines : la première est basée sur l'analyse des chemins les plus courts entre des paires de protéines privées et la seconde est basée sur l'utilisation du réseau d'affiliation extrait de la base de données ouverte KEGG (Kanehisa et Goto [67]).

4.4.4 Analyse des chemins les plus courts entre protéines

Le chemin le plus court est biologiquement significatif car énergétiquement le plus favorable. Le réseau privé de protéines est composé de 142 protéines (voir le tableau 4.5), les chemins les plus courts sont calculés entre 20164 paires de protéines. Le tableau 4.10 montre le nombre et le pourcentage de chemins les plus courts classés selon leur longueur

Longueur du chemin	0	1	2	3	4	5
Nombre	142	374	13790	5178	664	16
Pourcentage de paires	0,7 %	1,85 %	68,39%	25,78%	3,29 %	0,08 %

Tableau 4.10 – Nombre et longueur des chemins les plus courts entre les paires de protéines privées.

On peut ainsi proposer une classification des chemins les plus courts obtenus en trois classes en fonction de leur localisation dans l'ensemble des réseaux de protéines :

1. Chemins les plus courts dont les longueurs sont inférieures ou égales à un : ces chemins appartiennent au réseau protéique privé.
2. Chemins les plus courts dont les longueurs sont inférieures ou égales à trois : ceux-ci peuvent soit atteindre les réseaux égocentriques, soit appartenir complètement au réseau privé.
3. Les chemins les plus courts dont la longueur est supérieure à quatre : ceux-ci peuvent soit atteindre les réseaux ouverts, soit appartenir complètement à l'égocentrique ou au privé

Nous remarquons que la majorité des chemins les plus courts trouvés appartiennent au réseau égocentrique (ou privé) donc ils sont dans les voisins des nœuds du réseau privé. Seuls quelques-uns d'entre eux (3,37% des paires) peuvent être en dehors des réseaux égocentriques. Ces quelques longs chemins les plus courts peuvent être isolés et étudiés par des biologistes afin de comprendre les interactions moléculaires dans ces chemins. Le tableau 4.11 montre deux chemins des plus courts de longueur 5 composés de protéines et de métabolites : l'un a des nœuds en dehors du réseau égocentrique et l'autre est complètement à l'intérieur du réseau égocentrique et l'autre.

Nœuds du chemin	Chemin1	Chemin 2
1	"AT1G05450.2" "private"	"AT1G05450.2" "private"
2	"AT1G17030.1" "ego"	"AT1G17030.1" "ego"
3	CIDs70789281" "open"	" "CIDs00119211" "ego"
4	AT4G12920.1" "ego"	CIDs06435808" "ego"
5	"AT1G67290.1" "ego"	"AT3G07450.1" "ego"
6	"AT1G04540.1" "private"	"AT1G03390.1" "private"

Tableau 4.11 – Deux chemins les plus courts de longueur 5 : le premier contient des nœuds qui atteignent la zone non égocentrique et le second est complètement à l'intérieur du réseau égocentrique. Les molécules dont les noms commencent par *AT* sont des protéines et celles dont les noms commencent par *CID* sont des métabolites.

4.4.5 Analyse des chemins des protéines à l'aide de la base de données KEGG

KEGG (Kyoto Encyclopedia of Genes and Genomes) est une base de données ouverte qui intègre des informations génomiques, chimiques et systémiques fonctionnelles. En particulier, les catalogues de gènes concernant des génomes complètement séquencés sont liés aux fonctions systémiques de niveau supérieur de la cellule, de l'organisme et de l'écosystème (Kanehisa et Goto [67]).

Des efforts importants ont été entrepris pour créer manuellement une base de connaissances pour ces fonctions systémiques en capturant et en organisant les connaissances expérimentales sous des formes calculables ; à savoir, sous la forme de réseaux moléculaires appelés cartes de voies KEGG.

De l'ensemble de données KEGG, nous extrayons 4692 protéines qui sont associées à des identificateurs de chemin. Chaque identifiant est également lié à un nom de chemin qui caractérise les molécules (voir tableau 4.12 et 4.13). Il y a 238 chemins.

	KEGGPathway_ID	TAIRNoIso
1	KEGG :00190	AT1G01050
2	KEGG :04712	AT1G01060
3	KEGG :04712	AT1G01060

Tableau 4.12 – Exemples de paires d'identificateurs de chemins et de protéines extraites de la base de données KEGG.

	Identificateur du chemin	Nom du chemin
1	KEGG :00010	Glycolysis / Gluconeogenesis
2	KEGG :00020	Citrate cycle (TCA cycle)
3	KEGG :00030	Pentose phosphate pathway

Tableau 4.13 – Exemples de descriptions des chemin extraites de la base de données KEGG.

Notre objectif est d'utiliser cet ensemble de données afin de caractériser l'ensemble des protéines privées extraites de l'expérience. Nous représentons l'affiliation de protéines privées par un réseau d'affiliation extrait de la base de données KEGG et modélisé par un graphe bipartite contenant l'ensemble des protéines privées connectées à des identifiants de chemins (see figure 4.14). Les réseaux bipartites sont une classe particulière de réseaux complexes, dont les nœuds sont divisés en deux ensembles X et Y, et seules les connexions entre deux nœuds dans des ensembles différents sont autorisées. Les réseaux bipartites peuvent généralement être compressés par projection à un ensemble. Cela signifie que le réseau qui en résulte contient des nœuds uniquement de l'un des deux ensembles, une paire de nœuds X (ou, alternativement, Y) est connectée seulement si les deux nœuds ont au moins un nœud Y comme voisin commun (voir figure 4.15).

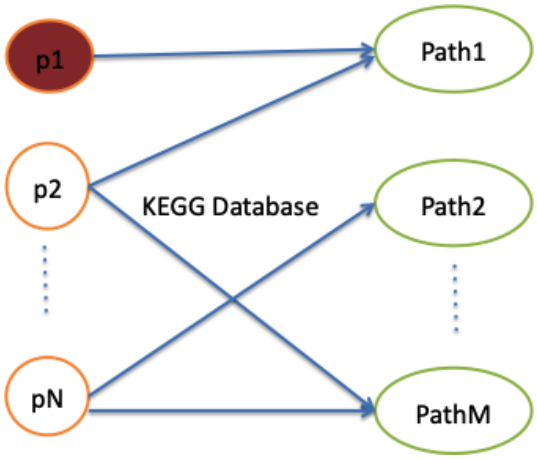


FIGURE 4.14 – Le réseau d'affiliation privé extrait de la base de données KEGG : des protéines privées connectées à des identifiants de chemins (Malek *et al.* [91]).

Nous considérons les réseaux de chemins obtenus par projection sur l'ensemble des chemins. Le tableau 4.14 montre les caractéristiques de ce réseau.

Nodes#	Edges#	Density
127	561	0.07

Tableau 4.14 – Le réseau projeté sur l'ensemble de chemins obtenu à partir du réseau d'affiliation privé.

Afin de caractériser l'ensemble des protéines privées, nous procédons comme suit : nous appliquons tout d'abord l'algorithme de Louvain (Blondel *et al.* [24]) au réseau projeté sur les chemins pour détecter les communautés. Les chemins appartenant à

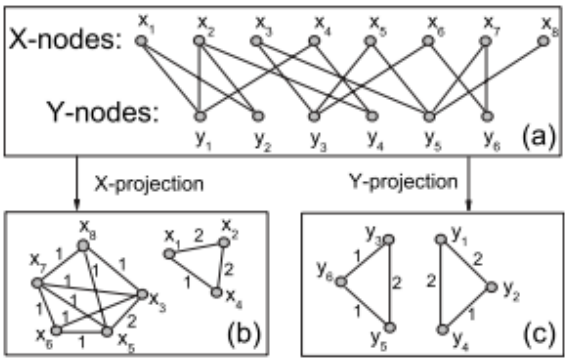


FIGURE 4.15 – Illustration de la projection d'un réseau bipartite (source : Zhou *et al.* [152]).

la même communauté sont similaires dans le sens où ils sont associés à des protéines en commun.

Nous analysons ensuite les chemins de l'ensemble des protéines privées pour chaque communauté. Soit $PathE$ l'ensemble de chemins concernant les protéines privées. Soit $PathC_i$ l'ensemble de chemins inclus dans la communauté numéro i. Pour chaque communauté i nous calculons :

1. La précision de l'expérience donnée par $\frac{PathC_i \cap PathE}{PathC_i}$: mesure le taux de voies de protéines privées parmi les voies incluses dans la communauté.
2. Le rappel de l'expérience donné par $\frac{PathC_i \cap PathE}{PathE}$: mesure le taux de voies de protéines privées parmi toutes les voies de protéines privées.
3. L'indice de jaccard donné par $\frac{PathC_i \cap PathE}{PathC_i \cup PathC_i}$: mesure le taux de protéines privées incluses dans cette communauté et ayant une affiliation.

Le tableau 4.15 montre des résultats concernant les communautés trouvées, certaines d'entre elles correspondent à des chemins isolés (de cardinalité un) qui n'ont pas d'affiliation à des protéines privées (la précision de l'expérience est égale à 0) ou qui ont une affiliation isolée (la précision de l'expérience est de 1).

Chaque communauté peut être décrite par un ensemble de noms de chemin (voir le tableau 4.13). La communauté 9 a le meilleur indice de jaccard sur l'ensemble des protéines privées.

Les chemins constituant cette communauté ont comme description la liste de molecules suivante : ["Oxidative phosphorylation", "N-Glycan biosynthesis" , "Porphyrin and chlorophyll metabolism" ,"Ribosome biogenesis in eukaryotes", "RNA transport" ,"RNA degradation" "Spliceosome" ,"Ubiquitin mediated proteolysis" , "Protein processing in endoplasmic reticulum","Circadian rhythm"]

Nous souhaitons dans l'avenir étudier ces chemins afin de montrer si la liste des métabolites et protéines retrouvés sur les chemins est biologiquement significative. Nous souhaitons également les comparer aux chemins les plus courts et à leurs relations avec le réseau égocentrique privé.

4.4.6 Discussion des résultats

Nous discutons dans cette section les résultats obtenus suite à l'application de notre méthodologie sur les réseaux biologiques. Nous présentons les observations et

CommunityNb	Cardinality	ExperimentPrecision	ExpermientRecall	jaccardInd
1	1	1.00	0.04	0.04
2	14	0.21	0.11	0.08
3	1	1.00	0.04	0.04
4	1	1.00	0.04	0.04
5	17	0.29	0.18	0.12
6	1	0.00	0.00	0.00
7	17	0.18	0.11	0.07
8	1	0.00	0.00	0.00
9	18	**0.56**	**0.36**	**0.28**
10	30	0.13	0.14	0.07
11	4	0.00	0.00	0.00
12	1	0.00	0.00	0.00
13	1	0.00	0.00	0.00
14	16	0.00	0.00	0.00
15	1	0.00	0.00	0.00
16	1	0.00	0.00	0.00
17	1	0.00	0.00	0.00
18	1	0.00	0.00	0.00

Tableau 4.15 – Détection de communautés avec l'algorithme de Louvain appliqué au réseau projeté aux chemins, la modularité est de 0,57.

les résultats liés à une couche et ceux obtenus pour l'ensemble du réseau.

Observations et résultats obtenus à partir de l'analyse d'une couche

— Dans un premier temps, le taux des molécules identifiées est calculé par rapport au réseau ouvert. Cela aide les biologistes à positionner les molécules identifiées par rapport aux données ouvertes. Dans notre cas, les bases de données biologiques ouvertes est très volumineuse par rapport à l'expérience biologique.
Les protéines de l'expérience ne représentent que 0,58% à 0,74% du nombre total de protéines dans tout le réseau. Les métabolites de l'expérience représentent moins de 0,05% du nombre total de métabolites dans tout le réseau.
— Les mesures de densité des réseaux ouverts et privés sont calculées et comparées, ce qui donne une indication sur les forces des connexions dans l'ensemble des molécules identifiées par rapport aux données ouvertes. Dans notre cas, nous remarquons que les réseaux de protéines ouverts et privés ont presque la même densité, par contre, le réseau de métabolites privés a une densité élevée par rapport au réseau ouvert (voir tableau 4.5 et 4.7).
— Afin d'avoir une idée de la distribution des molécules identifiées dans le réseau, nous appliquons l'algorithme de Louvain (Blondel *et al.* [24]) pour détecter des communautés, la distribution des molécules privées (identifiées à partir de l'expérience) est étudiée en fonction des communautés détectées. Dans notre cas, huit communautés sont détectées pour la couche protéique. Le taux de distribution des protéines identifiées dans ces communautés est

dans l'intervalle $[0,48\%; 0,77\%]$, cela signifie que les protéines identifiées sont presque distribuées de manière équilibrée dans les communautés. Notez que ces taux sont comparables au taux global de protéines identifiées dans le réseau. D'autre part, 39 communautés sont détectées pour la couche de métabolites. 80% des métabolites identifiés n'appartiennent qu'à 2 communautés, ce qui signifie qu'une majorité des métabolites identifiés sont fortement connectés et forment deux sous-réseaux hautement connectés. Les biologistes ont confirmé ces résultats en identifiant deux catégories connues de métabolites (voir les tableaux 4.5 et 4.7).

Observations et résultats obtenus à partir de l'analyse deux couches

— Le calcul de l'accessibilité des couches des métabolites aux protéines et des protéines aux métabolites permet d'apprécier le rapport des interactions immédiates entre molécules privées (identifiées dans leur expérience) en comparaison avec les données ouvertes. Le tableau 4.8 montre que les réseaux privés de métabolites égocentriques atteignent (voir section 4.2.4) un ensemble de protéines contenant 51% à 67 % des protéines identifiées malgré une très faible précision (0,6 % à 0,64 %).
De même, le tableau 4.9 montre que les réseaux privés de protéines égocentriques atteignent (voir section 4.2.4) un ensemble de métabolites contenant 53 % à 84% des métabolites identifiés malgré une très faible précision (0,56% à 0,7%).

— L'analyse des chemins les plus courts qui sont biologiquement significatifs reliant des paires de protéines privées pourrait être très utile pour les biologistes. Nous proposons de les classer en fonction de leur emplacement dans le réseau ouvert (privé, égocentrique ou extra-égocentrique). Dans notre cas, nous remarquons que la majorité des chemins les plus courts trouvés appartiennent au réseau égocentrique (ou privé) donc ils sont dans les voisins des nœuds du réseau privé.
Seuls quelques-uns d'entre eux (3,37 % des paires de protéines) peuvent être en dehors des réseaux égocentriques. Ces quelques longs chemins les plus courts peuvent être isolés et étudiés afin de comprendre les interactions moléculaires dans ces chemins.

— En utilisant la base de données KEGG (Kanehisa et Goto [67]), nous avons proposé de caractériser l'ensemble des protéines privées identifiées à partir de l'expérience par la description des chemins présentée sous forme de liste de noms de molecules. Nous souhaitons par la suite étudier ces chemins afin de montrer si la liste des métabolites et protéines retrouvés sur les chemin est biologiquement significative. Nous souhaitons également les comparer pour trouver les chemins les plus courts et leurs relations avec le réseau égocentrique privé.

4.5 Synthèse du chapitre et perspectives

Nous avons présenté une méthodologie comprenant des mesures et des méthodes qui aident les experts d'un domaine donné à comprendre, évaluer et compléter leurs

données privées en les comparant et/ou en les combinant avec des données ouvertes, lorsque les deux sont modélisées par des réseaux multicouches (Malek *et al.* [91]).

Nous avons proposé un nouveau formalisme pour les réseaux multicouches qui permet de réaliser une analyse fine en considérant deux niveaux : le niveau intra-couche et celui inter-couche.

Nous avons introduit les notions de réseau privé multicouche et de réseau égocentrique privé qui se définit autour du réseau privé multicouche. Le réseau égocentrique privé est utilisé pour évaluer la force de connectivité entre les différentes couches de données privées par rapport au réseau ouvert. Nous avons montré comment nous pouvons utiliser ces notions pour définir les métriques d'accessibilité de couche et inter-couche d'un sous-réseau donné.

Nous avons illustré notre méthodologie à travers une application biologique où les interactions entre molécules (protéines et métabolites) sont extraites à partir de bases de données ouvertes. Les données privées correspondent à un ensemble de protéines et de métabolites collectés expérimentalement et correspondent à un sous-ensemble de nœuds dans le réseau multicouche. Les résultats expérimentaux actuels sont pertinents du point de vue biologique et permet aux biologistes de comparer et d'évaluer les molécules identifiées et les réseaux privés dans le contexte des réseaux ouverts.

Ce travail a été réalisé dans la cadre du projet ANR BLIZAAR : `https://anr.fr/Projet-ANR-15-CE23-0002` ; un prototype a été développé pour illustrer et expérimenter cette méthodologie en collaboration avec des chercheurs du laboratoire LIS (Malek *et al.* [91]).

Dans notre méthodologie, deux types de calculs et de traitement sont proposés et étudiés : des calculs et des traitements intra-couches et d'autres qui agrègent les mesures et les traitements inter et intra-couches pour explorer la globalité du réseau. Par exemple, calculer les mesures des réseaux ouverts et privés pour les différentes couches comme les densités et les degrés permet d'apprécier la force des connexions internes entre les données privées et de comparer avec celle des connexions à d'autres données ainsi qu'aux mesures globales. De même, l'application d'algorithmes de détection de communautés dans les différentes couches permet d'apprécier la distribution des nœuds privés dans le réseau ouvert. Le calcul des accessibilités à travers les couches permet également d'apprécier le rapport des interactions immédiates entre les molécules privées ainsi que l'ensemble des données (nœuds) ouvertes accessibles à partir des noeuds identifiées.

Ces travaux nous ouvrent les perspectives liées à la problématique de la détection des communautés (Fortunato [44]) dans un réseau multicouche et plus particulièrement à l'utilisation des métriques d'accessibilité des couches et inter-couches afin de proposer des algorithmes permettant les comparaisons et la caractérisation des communautés des différentes couches. Nous souhaitons étudier cette approche plus particulièrement pour les algorithmes de détection des communautés chevauchantes que nous avons proposés dans le chapitre précédent.

Chapitre 5

Conclusion et Perspectives

5.1 Conclusion

Nous nous sommes intéressés dans ce mémoire à l'analyse des réseaux complexes modélisés par des graphes non orientés et plus particulièrement à : a) l'exploration des réseaux complexes pour la recommandation b) la détection des communautés chevauchantes et c) l'analyse des réseaux multicouches pour des données ouvertes et privées.

5.1.1 Exploration des réseaux complexes pour la recommandation

Nous avons présenté dans un premier temps, un système de recommandation d'experts dans un réseau social professionnel. Notre réseau est composé d'un ensemble de personnes ayant des liens professionnels. Selon la demande d'un acteur d'origine X, le système doit proposer (recommander) un ou plusieurs acteurs $\{Z_1, Z_2, .., Z_n\}$ répondant au mieux que possible aux critères demandés. Nous avons appliqué notre algorithme sur la recommandation d'auteurs de papiers scientifiques dans un graphe de similarités entre auteurs.

La recommandation dépend d'un côté, de la valeur de pertinence entre les profils des auteurs et la requête soumise et d'un autre côte de l'intermédiarité des nœuds-auteurs se trouvant sur les chemins de la solution. Pour effectuer une recherche dans le graphe, l'arbre couvrant le plus représentatif est extrait et ensuite il est exploré.

Le premier algorithme est exhaustif, il est fondé sur la recherche en largeur dans l'arbre couvrant, jusqu'à ce qu'on trouve un auteur à recommander. Le deuxième algorithme utilise l'approche A* pour explorer l'arbre couvrant. Nous définissons une heuristique admissible qui dépend de la pertinence entre un profil et la requête ainsi de l'intermédiarité des nœuds-auteurs. Les expériences ont montré que la version guidée trouve souvent la meilleure recommandation tout en améliorant les performances de l'exploration. En comparant les deux algorithmes nous remarquons que l'arbre couvrant n'a pas été recherché en totalité par la version guidée, l'espace de recherche a été réduit de 11% à 49%.

Suite à cette première proposition, nous avons envisagé l'idée de l'intégration de l'ontologie de domaine dans le processus de la recommandation, le but étant d'utiliser cette ontologie dans l'élaboration des requêtes posées afin d'aider l'utilisateur à re-

formuler sa requête ou à la compléter ainsi que dans la représentation ontologique du profil utilisateur ou des ses préférences (Malek *et al.* [90]). Par conséquent, la mesure de pertinence entre la requête et le profil utilisateur doit intégrer les concepts ainsi que la structure de l'ontologie ce qui permettra d'avoir des recommandations plus précises.

Les travaux de thèse de Sulieman [128] ont traité ce sujet par la proposition d'une nouvelle approche de système de recommandation appliquée sur les réseaux de collaboration. Dans ce système, appelé système de recommandation social-sémantique, deux types d'information sont pris en compte : les informations extraites de la topologie du réseau comme les mesures de centralité et les poids des liens ainsi que les informations sémantiques qui sont définies à l'aide d'une taxonomie de domaine sur les items et qui ont permis de proposer une mesure de pertinence fine et précise entre élément et utilisateur.

Deux algorithmes principaux de recommandation fondés respectivement sur la recherche en profondeur et la recherche en largeur ont été proposés. Différentes mesures de centralité (degré, proximité, intermédiarité et hybride) ont été utilisées pendant l'exploration du graphe afin d'éviter de parcourir complètement le graphe (Sulieman *et al.* [129]). Ces algorithmes ont été comparés à deux algorithmes classiques de recommandation (filtrage collaboratif centré sur les items et recommandation hybride) en terme de précision (précision, rappel et mesure F) et d'efficacité (couverture du graphe et temps d'exécution). Les expériences montrent que les algorithmes proposés présentent des valeurs de précision similaires à celles des deux algorithmes standards, tout en les surpassant en termes d'efficacité puisqu'une partie restreinte du réseau est visitée.

5.1.2 Détection des communautés chevauchantes

Nous avons présenté les différentes approches de détection de communautés disjointes et chevauchantes dans un graphe de terrain ainsi que les mesures d'évaluation internes et externes proposées dans la littérature et leur adaptation au contexte du chevauchement.

Nous avons proposé ensuite une méthode pour la detection des communautés chevauchantes à partir de communautés disjointes pré-calculées obtenues en utilisant la méthode du *core detection label propagation* (CDLP) décrite dans (Attal et Malek [8]). L'algorithme sélectionne les nœuds candidats pour le chevauchement et utilise des *fonctions d'appartenance* pour décider de l'affectation ou non d'un nœud candidat à chacune de ses communautés voisines. Nous avons proposé plusieurs fonctions d'appartenance (Attal *et al.* [10]) qui sont soit basées sur des mesures topologiques globales comme la densité et le coefficient de clustering (f_d et f_{cc}) soit sur les moyennes des mesures de centralités d'intermédiarité et de proximité des nœuds constituant les communautés pré-calculées (f_b et f_{cl}).

Les résultats expérimentaux sont relativement bons et compétitifs en comparaison avec d'autres méthodes de détection de communautés chevauchantes en terme de qualité mais dépendent bien de la topologie du graphe.

5.1.3 Analyse des Réseaux multicouches pour des données ouvertes et privées

Nous avons présenté une méthodologie comprenant des mesures et des méthodes qui aident les experts d'un domaine donné à comprendre, évaluer et compléter leurs données privées en les comparant et/ou en les combinant avec des données ouvertes, lorsque les deux sont modélisées par des réseaux multicouches (Malek *et al.* [91]).

Nous avons proposé un nouveau formalisme pour les réseaux multicouches qui permet de réaliser une analyse fine en considérant deux niveaux : le niveau intra-couche et celui inter-couche.

Nous avons introduit les notions de réseau privé multicouche et de réseau égocentrique privé qui se définit autour du réseau privé multicouche. Le réseau égocentrique privé est utilisé pour évaluer la force de connectivité entre les différentes couches de données privées par rapport au réseau ouvert. Nous avons montré comment nous pouvons utiliser ces notions pour définir les métriques d'accessibilité de couche et inter-couche d'un sous-réseau donné.

Nous avons illustré notre méthodologie à travers une application biologique où les interactions entre molécules (protéines et métabolites) sont extraites à partir de bases de données ouvertes. Les données privées correspondent à un ensemble de protéines et de métabolites collectés expérimentalement et sont identiques à un sous-ensemble de nœuds dans le réseau multicouche. Les résultats expérimentaux actuels sont pertinents du point de vue biologique et permet aux biologistes de comparer et d'évaluer les molécules identifiées et les réseaux privés avec les réseaux ouverts. Ce travail a été réalisé dans la cadre du projet ANR BLIZAAR : `https://anr.fr/Projet-ANR-15-CE23-0002`.

Dans notre méthodologie, deux types de calculs et de traitements sont proposés et étudiés : des calculs et des traitement intra-couches et d'autres qui agrègent les mesures et les traitements inter et intra-couches pour explorer la globalité du réseau. Par exemple, calculer les mesures des réseaux ouverts et privés pour les différentes couches comme les densités et les degrés permet d'apprécier la force des connexions internes entre les données privées et de comparer avec celle des connexions à d'autres données ainsi qu'aux mesures globales. De même, l'application d'algorithmes de détection de communautés dans les différentes couches permet d'apprécier la distribution des nœuds privés dans le réseau ouvert. Le calcul des accessibilités à travers les couches permet également d'apprécier le rapport des interactions immédiates entre les molécules privées ainsi que l'ensemble des données (nœuds) ouvertes accessibles à partir des noeuds identifiées.

Ces travaux nous ouvrent les perspectives liées à la problématique de la détection des communautés (Fortunato [44]) dans un réseau multicouche et plus particulièrement à l'utilisation des métriques d'accessibilité des couches et inter-couches afin de proposer des algorithmes permettant les comparaisons et la caractérisation des communautés des différentes couches.

5.2 Perspectives

5.2.1 Travaux actuels : analyse des sentiments dans les médias sociaux

L'analyse de sentiment (opinion mining) s'appuie sur diverses techniques telles que le traitement du langage naturel, la recherche d'informations et l'exploration de données structurées et non structurées. Le processus d'exploration d'opinion implique trois étapes principales qui sont la récupération d'opinion, la classification d'opinion et la synthèse d'opinion (Ravi et Ravi [115]). Des chercheurs d'horizons différents ont présenté plusieurs modèles pour étudier la formation, la propagation et l'agrégation des opinions selon différents points de vue (Mohammadinejad *et al.* [100]).

Nous explorons dans nos travaux actuels la combinaison de méthodes classiques d'exploration d'opinion avec l'analyse des réseaux complexes et son impact sur la formation et la propagation d'opinion afin de construire un modèle d'opinion cohérent.

Afin d'étudier l'impact des utilisateurs influents (nœuds influents) sur la formation et la propagation de l'opinion, nous intégrons dans un premier temps plusieurs facteurs d'influence extraits du réseau dans le processus d'exploration d'opinion. Ces facteurs sont généralement calculés en utilisant différentes mesures de centralité comme le degré, la proximité, l'intermédiarité, la centralité PageRank, etc.

De plus et afin de comprendre comment les opinions interagissent et sont influencées les unes par les autres à travers la structure et comment les propriétés du réseau complexe contribuent à ce processus, nous proposons une méthode d'agrégation des opinions afin de converger vers un modèle d'opinion cohérent pour des groupes d'utilisateurs/acteurs appartenant à des types particuliers de sous-réseaux. Ces sous-réseaux sont soit des réseaux égocentriques autour d'influenceurs, soit des communautés obtenues en appliquant des algorithmes de détection de communauté disjointes et chevauchantes (Folly *et al.* [43]).

Nous définissons et étudions ensuite la notion de la stabilité d'opinion au sein de ces réseaux égocentriques autour des influenceurs et au sein des communautés, notre objectif étant de détecter la modification d'opinion pour les deux types de sous-réseaux. La stabilité d'opinion peut s'exprimer par le fait de partager une majorité de préférences communes concernant un sujet donné au sein d'un groupe d'utilisateurs. Les premières expérimentations ont comme objectif d'observer comment l'opinion peut être stable dans les réseaux égocentriques et quels sont les facteurs qui aident à détecter une modification de l'opinion. Nous étudions ainsi la variation d'opinion selon plusieurs mesures nodales et topologiques comme les centralités des influenceurs (egos), les poids des liens et la distance géodésique de l'influenceur, etc. De même, nous étudions la stabilité de l'opinion dans les communautés en fonction de plusieurs mesures globales et locales comme la densité, le coefficient de clustering et la moyenne des centralités, etc. Nous observons la modification de l'opinion spécialement pour les nœuds frontières entre les communautés. Cela nous aide à apprécier l'apport de la structure communautaire du réseau et sa contribution à étendre ou à limiter la propagation de l'opinion.

Nous avons choisi le thème *vaccination COVID* pour nos premières expérimentations. Après avoir déterminer deux influenceurs ayant des opinions opposées sur le sujet, nous avons construit le réseau de propagation des opinions positives et négatives à partir des réseaux égocentriques autour de ces deux influenceurs. Nous sommes en trains d'étudier les facteurs influant la modification et la stabilité de l'opinion au sein de ce réseau.

5.2.2 Travaux à venir : explicabilité et réseaux complexes

Les racines de l'IA remontent à plusieurs décennies, il existe un consensus clair sur l'importance primordiale des systèmes intelligents dotés d'apprentissage, de raisonnement et de capacité d'adaptation. Les méthodes d'apprentissage automatique atteignent aujourd'hui des niveaux de performance élevées permettant de résoudre des tâches de plus en plus complexes. Il existe aujourd'hui un besoin émergent de comprendre comment ces décisions sont rendues par des méthodes d'IA lorsque les décisions de tels systèmes affectent la vie des humains. Les paradigmes sous-jacents à ce problème révèlent du domaine dit de l'IA explicable (XAI) reconnu comme une caractéristique cruciale pour le déploiement pratique des modèles d'IA (Barredo Arrieta *et al.* [17]). L'explicabilité a pour objectif de faciliter la compréhension de divers aspects d'un modèle conduisant à des informations qui peuvent être utilisées par plusieurs acteurs tels que : le data scientiste, le manager, l'expert du domaine, l'utilisateur du modèle, etc. L'explicatibilité peut être considérée comme une caractéristique active d'un modèle, désignant toute action prise par un modèle dans le but de clarifier ou de détailler ses fonctions internes (Barredo Arrieta *et al.* [17], Morichetta *et al.* [101]) De même, l'interprétabilité est la capacité d'expliquer ou de donner le sens de manière compréhensible à un humain. Les méthodes XAI sont classées selon divers critères (Barredo Arrieta *et al.* [17]) : i) intrinsèque ou post-hoc méthodologique, ii) objectif de la méthode d'interprétation, iii) modèle spécifique ou agnostique, iv) interprétation locale ou globale.

i Les modèles intrinsèques signifient que le modèle est interprétable à sa base comme les modèles linéaires et les arbres de décisions. Post-hoc fait référence à l'application de méthode d'interprétations au moment du test après avoir le modèle d'apprentissage (comme pour l'apprentissage profond, les modèles SVM et les forets aléatoires, etc.).

ii Différentes approches peuvent être utilisées pour interpréter les résultats comme : le résumé des caractéristiques statistiques, l'importance des caractéristiques, la visualisation ou bien la simplification en utilisant des arbres de décisions, en recherchant des prototypes, etc.

iii La méthode d'interprétabilité peut être spécifique au modèle si elle ne s'applique qu'à un modèle particulier, ou indépendant du modèle. Généralement l'interprétabilité post-hoc est indépendante du modèle.

iv L'interprétabilité est dite globale si elle peut expliquer le comportement global du modèle ou locale si elle explique un aspect particulier. L'interprétabilié globale est souvent très difficile à réaliser.

Notre objectif est d'étendre la méthodologie d'explicabilité au domaine de l'analyse des réseaux complexes afin de produire des explications émergeantes à partir

des informations topologiques extraites de la structure du graphe comme les mesures topologiques globales (densité, coefficient de clustering, diamètre) et les mesures locales (différents types de centralités : degrés, intermédiarités, pageRnak, etc.). Cette méthodologie s'avère très nécessaire lorsque l'analyse du réseau complexe fait partie d'un système de décision qui permet de réaliser une tâche particulière comme par exemple les systèmes de recommandation, l'analyse des sentiments ou la prédiction des liens. Ceci peut être bien utile également pour la compréhension des structures communautaires détectées dans le réseau.

Dans la plupart des méthodes d'apprentissage automatique les entités traitées dans une méthode d'interprétabilité sont liées aux caractéristiques (attributs), aux données, aux prototypes extraits ainsi qu'à des connaissances simples extraites ayant souvent la forme de règles ou d'arbre décision. Lorsqu'il s'agit des réseaux complexes, des nouvelles entités doivent être prises en compte comme la nature des liens dans le graphe (lien social, lien de collaboration ou d'interaction, etc.) ainsi que les informations topologiques extraites à partir du graphe. Ces informations peuvent être combinées également avec les informations nodales afin de proposer des actions explicatives complètes qui intègrent des informations topologiques avec des informations sémantiques sur les acteurs.

Pour illustrer cette idée, nous nous basons sur notre travail actuel (Folly *et al.* [43]) : nous construisons un réseau de propagation de tweets, de retweets et de réponses qui correspond à l'union de deux réseaux égocentriques autour de deux influenceurs ayant deux avis opposés sur un sujet donné. Nous proposons par la suite un système d'analyse de sentiments qui intègre les éléments d'analyse du réseau de propagation. Nous souhaitons analyser les communautés obtenues et comprendre les opinions émergeantes à partir de ces communautés non seulement en fonction des profils utilisateurs mais aussi en fonction d'éléments topologiques extraits du graphe de propagation des opinions opposée des deux influenceurs. Nous souhaitons détecter également des indicateurs concernant la stabilité des opinions ainsi que ceux liés à leurs changements dans le réseau des utilisateurs apparentant aux deux réseaux égocentriques d'opinions opposées.

En intégrant une méthodologie d'expliacabilité adéquate, nous souhaitons rendre plus compréhensibles également les résultats concernant la polarité de l'opinion trouvée au niveau des utilisateurs et au niveau des groupe (les communautés). De même le modèle doit être capable d'expliquer les changements d'opinion détectés en lien avec les informations extraites du réseau de propagation et les séquences d'actions entreprises (tweets, retweets, réponses) menant à ce changement.

Bibliographie

[1] Lada A. Adamic, Orkut Buyukkokten et Eytan Adar : A social network caught in the web. *First Monday*, 8(6), 2003.

[2] G. Adomavicius et A. Tuzhilin : Toward the Next Generation of Recommender Systems : A Survey of the State-of-the-Art and Possible Extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749, 2005.

[3] C.C. Aggarwal, éditeur. *Social Network Data Analytics*, chapitre A survey of algorithms and systems for expert location in social networks, pages 215–241. Springer Science and Business Media, 2011.

[4] Yong-Yeol Ahn, James P Bagrow et Sune Lehmann Jørgensen : Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761–764, 2010. ISSN 0028-0836.

[5] Réka Albert, Hawoong Jeong et Albert-László Barabási : Internet : Diameter of the world-wide web. *Nature*, 401(6749):130–131, 1999.

[6] Alberto Aleta et Yamir Moreno : Multilayer networks in a nutshell. *Annual Review of Condensed Matter Physics*, 10(1):45–62, 2019. URL `https://doi.org/10.1146/annurev-conmatphys-031218-013259`.

[7] Jean-philippe Attal : *Nouveaux algorithmes pour la détection de communautés disjointes et chevauchantes basés sur la propagation de labels et adaptés aux grands graphes.* Theses, Université de Cergy Pontoise, janvier 2017. URL `https://tel.archives-ouvertes.fr/tel-01534480`.

[8] Jean-Philippe Attal et Maria Malek : A new label propagation with dams. *In 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1292–1299. IEEE, 2015.

[9] Jean-Philippe Attal, Maria Malek et Marc Zolghadri : Overlapping community detection using core label propagation and belonging function. *In International Conference on Neural Information Processing*, pages 165–174. Springer, 2016.

[10] Jean-philippe Attal, Maria Malek et Marc Zolghadri : Overlapping community detection using core label propagation algorithm and belonging functions. *Applied Intelligence*, mars 2021. URL `https://hal.archives-ouvertes.fr/hal-03180441`.

[11] David A Bader, Shiva Kintali, Kamesh Madduri et Milena Mihail : Approximating betweenness centrality. *In Algorithms and Models for the Web-Graph*, pages 124–137. Springer, 2007.

[12] Albert-László Barabási et Réka Albert : Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[13] Albert-László Barabási et Eric Bonabeau : Scale-free networks. *Scientific American*, 288(5):50–59, 2003.

[14] Stephen T Barnard : Pmrsb : Parallel multilevel recursive spectral bisection. *In Proceedings of the 1995 ACM/IEEE conference on Supercomputing*, page 27. ACM, 1995.

[15] Stephen T Barnard et Horst D Simon : Fast multilevel implementation of recursive spectral bisection for partitioning unstructured problems. *Concurrency : Practice and experience*, 6(2):101–117, 1994.

[16] ER Barnes et AJ Hoffman : On bounds for eigenvalues of real symmetric matrices. *Linear Algebra and its Applications*, 40:217–223, 1981.

[17] Alejandro Barredo Arrieta, Natalia Diaz-Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila et Francisco Herrera : Explainable artificial intelligence (xai) : Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020. ISSN 1566-2535. URL `https://www.sciencedirect.com/science/article/pii/S1566253519308103`.

[18] Federico Battiston, Vincenzo Nicosia et Vito Latora : Structural measures for multiplex networks. *Phys. Rev. E*, 89:032804, Mar 2014. URL `https://link.aps.org/doi/10.1103/PhysRevE.89.032804`.

[19] Jeffrey Baumes, Mark Goldberg et Malik Magdon-Ismail : Efficient identification of overlapping communities. *In* Paul Kantor, Gheorghe Muresan, Fred Roberts, Daniel D. Zeng, Fei-Yue Wang, Hsinchun Chen et Ralph C. Merkle, éditeurs : *Intelligence and Security Informatics*, pages 27–36, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-32063-0.

[20] J. Bennett et S. Lanning : The Netflix Prize. *In Proceedings of the KDD Cup Workshop 2007*, pages 3–6, New York, 2007. ACM.

[21] Peter Uetz Benno Schwikowski1 et Stanley Fields : A network of protein-protein interactions in yeast. *PubMed, Nat Biotechnol*, Volume 18:1173–1178, December 2000.

[22] Peter Uetz Benno Schwikowski1 et Stanley Fields : Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *PNAS, Proceedings of the National Academy of Sciences*, Volume 18:5934–5939, December 2004.

[23] Michele Berlingerio, Michele Coscia, Fosca Giannotti, Anna Monreale et Dino Pedreschi : Multidimensional networks : Foundations of structural analysis. *World Wide Web*, 16(5-6):567–593, novembre 2013. ISSN 1386-145X.

[24] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte et Etienne Lefebvre : Fast unfolding of communities in large networks. *Journal of statistical mechanics : theory and experiment*, 2008(10):P10008, 2008.

[25] S. Boccaletti, G. Bianconi, R. Criado, C.I. del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña-Nadal, Z. Wang et M. Zanin : The structure and dynamics of multilayer networks. *Physics Reports*, 544(1):1–122, Nov 2014. ISSN 0370-1573. URL `http://dx.doi.org/10.1016/j.physrep.2014.07.001`.

[26] Ulrik Brandes : A faster algorithm for betweenness centrality*. *Journal of mathematical sociology*, 25(2):163–177, 2001.

[27] Sergey Brin et Larry Page : The anatomy of a large-scale hypertextual web serach engine. *In 7th international conference on World Wide Web, AUSTRIA*, 1998.

[28] Robin Burke : Hybrid recommender systems : Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, novembre 2002. ISSN 0924-1868.

[29] Robin Burke : Hybrid Web Recommender Systems The Adaptive Web. *In* Peter Brusilovsky, Alfred Kobsa et Wolfgang Nejdl, éditeurs : *The Adaptive Web*, volume 4321 de *Lecture Notes in Computer Science*, chapitre 12, pages 377–408. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2007. ISBN 978-3-540-72078-2.

[30] Christopher S. Campbell, Paul P. Maglio, Alex Cozzi et Byron Dom : Expertise identification using email communications. *In CIKM*, pages 528–531, 2003.

[31] Fan R. K. Chung : *Spectral Graph Theory (CBMS Regional Conference Series in Mathematics, No. 92)*. American Mathematical Society, décembre 1996. ISBN 0821803158.

[32] Aaron Clauset, Mark EJ Newman et Cristopher Moore : Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

[33] Aaron Clauset, Cosma Rohilla Shalizi et Mark EJ Newman : Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[34] Aaron Clauset, Cosma Rohilla Shalizi et Mark EJ Newman : Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[35] Linda M Collins et Clyde W Dent : Omega : A general formulation of the rand index of cluster recovery suitable for non-disjoint solutions. *Multivariate Behavioral Research*, 23(2):231–242, 1988.

[36] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest et Clifford Stein : *Introduction to Algorithms*. The MIT Press, third edition édition, juillet 2009. ISBN 0262033844. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0262033844.

[37] Qiguo Dai, Maozu Guo, Yang Liu, Xiaoyan Liu et Ling Chen : Mlpa : Detecting overlapping communities by multi-label propagation approach. *In Evolutionary Computation (CEC), 2013 IEEE Congress on*, pages 681–688. IEEE, 2013.

[38] Leon Danon, Jordi Duch, Albert Diaz-Guilera et Alex Arenas : Comparing community structure identification, 2005. URL doi:10.1088/1742-5468/2005/09/P09008.

[39] Pasquale De Meo, Emilio Ferrara, Giacomo Fiumara et Alessandro Provetti : Enhancing community detection using a network weighting strategy. *Information Sciences*, 222:648–668, 2013.

[40] Sarah Djemili, Claudia Marinica, Maria Malek et Dimitris Kotzinos : Personal Networks of Scientific Collaborators : A Large Scale Experimental Analysis of Their Evolution. *In Information Search, Integration, and Personlization. Communications in Computer and Information Science.*, volume 760. Springer, Cham, 2017. URL https://hal.archives-ouvertes.fr/hal-01707186.

[41] L. Donetti et M. A. Muñoz : Detecting network communities : a new systematic and efficient algorithm. *Journal of Statistical Mechanics : Theory and Experiment*, 10:12, octobre 2004.

[42] Miroslav Fiedler : Algebraic connectivity of graphs. *Czechoslovak mathematical journal*, 23(2):298–305, 1973.

[43] Kossi Folly, Maria Malek et Dimitris Kotzinos : Social networks analysis for opinion model extraction. *In Networks 2021 : first combined meeting of the International Network for Social Network Analysis (Sunbelt XLI), and the Network Science Society (NetSci 2021).*, Indiana, United States, juillet 2021. URL https://hal.archives-ouvertes.fr/hal-03199000.

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

[45] Santo Fortunato et Andrea Lancichinetti : Community detection algorithms : A comparative analysis : Invited presentation, extended abstract. *In Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools*, VALUETOOLS '09, pages 27 :1–27 :2, ICST, Brussels, Belgium, Belgium, 2009. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering). ISBN 978-963-9799-70-7. URL http://dl.acm.org/citation.cfm?id=1698822.1698858.

[46] Linton C Freeman : Centrality in social networks conceptual clarification. *Social networks*, 1(3):215–239, 1978.

[47] Robert Geisberger, Peter Sanders et Dominik Schultes : Better approximation of betweenness centrality. *In Proceedings of the Meeting on Algorithm Engineering & Expermiments*, pages 90–100. Society for Industrial and Applied Mathematics, 2008.

[48] David Gfeller, Jean-Cédric Chappelier et Paolo De Los Rios : Finding instabilities in the community structure of complex networks. *Physical Review E*, 72(5):056135, 2005.

[49] M. Girvan et M. E. J. Newman : Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[50] Michelle Girvan et Mark EJ Newman : Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12): 7821–7826, 2002.

[51] Marko Gosak, Rene Markoviç, Jurij Dolensek, Marjan Slak Rupnik, Marko Marhl, Andraz Stozer et Matjaz Perc : Network science of biological systems at different scales : A review. *Physics of Life Reviews*, 24:118 – 135, 2018. ISSN 1571-0645. URL `http://www.sciencedirect.com/science/article/pii/S1571064517301501`.

[52] Steve Gregory : An algorithm to find overlapping community structure in networks. *In Knowledge discovery in databases : PKDD 2007*, pages 91–102. Springer, 2007.

[53] Steve Gregory : Finding overlapping communities in networks by label propagation. *New Journal of Physics*, 12(10):103018, 2010.

[54] Steve Gregory : Fuzzy overlapping communities in networks. *CoRR*, abs/1010.1523, 2010. URL `http://arxiv.org/abs/1010.1523`.

[55] J.W. Grossman : The evolution of the mathematical research collaboration graph. *Congressus Numeratium*, 2002.

[56] Mahdi Hajiabadi, Hadi Zare et Hossein Bobarshad : IEDC : an integrated approach for overlapping and non-overlapping community detection. *Knowl.-Based Syst.*, 123:188–199, 2017. URL `https://doi.org/10.1016/j.knosys.2017.02.018`.

[57] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen et John T. Riedl : Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, janvier 2004. ISSN 1046-8188.

[58] Faliang Huang, Xuelong Li, Shichao Zhang, Jilian Zhang, Jinhui Chen et Zhi-nian Zhai : Overlapping community detection for multimedia social networks. *IEEE Trans. Multimedia*, 19(8):1881–1893, 2017. URL `https://doi.org/10.1109/TMM.2017.2692650`.

[59] Lawrence Hubert et Phipps Arabie : Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[60] Roberto Interdonato, Matteo Magnani, Diego Perna, Andrea Tagarelli et Davide Vega : Multilayer network simplification : Approaches, models and methods. *Comput. Sci. Rev.*, 36:100246, 2020. URL `https://doi.org/10.1016/j.cosrev.2020.100246`.

[61] A. K. Jain, M. N. Murty et P. J. Flynn : Data clustering : A review. *ACM Comput. Surv.*, 31(3):264–323, septembre 1999. ISSN 0360-0300. URL `https://doi.org/10.1145/331499.331504`.

[62] H. Jeong et Al. : Lethality and centrality in protein networks. *Nature, Internationnal weekly journal of science*, Volume 411:41–42, May 2001.

[63] Rual JF et Al. : Towards a proteome-scale map of the human protein-protein interaction network. *Nature, the weekly, international, interdisciplinary journal of science.*, Volume 437:1173–1178, October 2005.

[64] Di Jin, Bogdan Gabrys et Jianwu Dang : Combined node and link partitions method for finding overlapping communities in complex networks. *Scientific reports*, 5:8600, 2015.

[65] Hubert Kadima et Maria Malek : Toward ontology-based personalization of a recommender system in social network. *International Journal of Computer Information Systems and Industrial Management (IJCISIM)*, 5, 2013.

[66] Rushed Kanawati : Licod : Leaders identification for community detection in complex networks. *In Privacy, Security, Risk and Trust (PASSAT) and IEEE Third Inernational Conference on Social Computing (SocialCom)*, pages 577–582. IEEE, 2011.

[67] Minoru Kanehisa et Susumu Goto : KEGG : Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 28(1):27–30, 01 2000. ISSN 0305-1048. URL `https://doi.org/10.1093/nar/28.1.27`.

[68] Ravi Kannan, Santosh Vempala et Adrian Vetta : On clusterings : Good, bad and spectral. *Journal of the ACM (JACM)*, 51(3):497–515, 2004.

[69] Brian Karrer, Elizaveta Levina et Mark EJ Newman : Robustness of community structure in networks. *Physical Review E*, 77(4):046119, 2008.

[70] Stephen Kelley : *The existence and discovery of overlapping communities in large-scale networks*. Thèse de doctorat, RENSSELAER POLYTECHNIC INSTITUTE, 2009.

[71] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno et Mason A. Porter : Multilayer networks. *J. Complex Networks*, 2(3):203–271, 2014. URL `https://doi.org/10.1093/comnet/cnu016`.

[72] Mikko Kivelä, Fintan McGee, Guy Melançon, Nathalie Henry Riche et Tatiana von Landesberger : Visual analytics of multilayer networks across disciplines (dagstuhl seminar 19061). *Dagstuhl Reports*, 9(2):1–26, 2019. URL `https://doi.org/10.4230/DagRep.9.2.1`.

[73] Jon M. Kleinberg : Authoritative sources in a hyperlinked environment. *Journal of the AM*, 46:604–632, 1999.

[74] István A Kovács, Robin Palotai, Máté S Szalay et Peter Csermely : Community landscapes : an integrative approach to determine overlapping network module hierarchy, identify key nodes and predict network dynamics. *PloS one*, 5(9):e12528, 2010.

[75] Valdis Krebs : Books about us politics. *unpublished, http ://www. orgnet. com*, 2004.

[76] Joseph B. Kruskal : On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7 (1):48–50, 1956. URL `https://app.dimensions.ai/details/publication/pub.1018477579andhttps://www.ams.org/proc/1956-007-01/S0002-9939-1956-0078686-7/S0002-9939-1956-0078686-7.pdf`.

[77] S Kullback : Statistics and information theory. *J. Wiley and Sons, New York*, 1959.

[78] Andrea Lancichinetti, Santo Fortunato et János Kertész : Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015, 2009.

[79] Andrea Lancichinetti, Santo Fortunato et János Kertész : Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015, 2009.

[80] Andrea Lancichinetti, Filippo Radicchi, José J Ramasco et Santo Fortunato : Finding statistically significant communities in networks. *PloS one*, 6(4):e18961, 2011.

[81] Cornelius Lanczos : *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*. United States Governm. Press Office Los Angeles, CA, 1950.

[82] J Law et J Hassard : *Actor network theory and after*. Blackwell publishers, 1999.

[83] Conrad Lee, Fergal Reid, Aaron McDaid et Neil Hurley : Detecting highly overlapping community structure by greedy clique expansion. *In SNAKDD Workshop*, page 4533 ?Äì42, 2010.

[84] Greg Linden, Brent Smith et Jeremy York : Amazon.com recommendations : Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, janvier 2003. ISSN 1089-7801.

[85] Weiyi Liu, Toyotaro Suzumura, Hongyu Ji et Guangmin Hu : Finding overlapping communities in multilayer networks. *PLOS ONE*, 13(4):1–22, 04 2018. URL https://doi.org/10.1371/journal.pone.0188747.

[86] Yao Lu, Xiaojun Quan, Xingliang Ni, Wenyin Liu et Yinlong Xu : Latent link analysis for expert finding in user-interactive question answsering services. *In 2009 fifth International Conference on Semantics, Knowledge and Grid*, 2009.

[87] David Lusseau, Karsten Schneider, Oliver J Boisseau, Patti Haase, Elisabeth Slooten et Steve M Dawson : The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, 54(4):396–405, 2003.

[88] Matteo Magnani et Luca Rossi : Pareto distance for multi-layer network analysis. *In Proceedings of the 6th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction*, SBP'13, pages 249–256, Berlin, Heidelberg, 2013. Springer-Verlag. ISBN 9783642372094.

[89] Maria Malek et Dalia Sulieman : Exhaustive and guided algorithms for recommendation in a professional social network. *In 7th conference on Application of Social Network Analysis*, september 2010.

[90] Maria Malek, Dalia Sulieman et Hubert Kadima : Integration of semantic user profile within a social recommendation system : semantic relevance measures characterization. *International Journal of complex systems in sciences, Acte de IV International Conference Net-Works 2011 Complex Networks : Structure, Applications and Related Topics*, 5, 2011.

[91] Maria Malek, Simone Zorzan et Mohammad Ghoniem : A methodology for multilayer networks analysis in the context of open and private data : biological application. *Applied Network Science*, 5(1), juillet 2020. URL https://hal.archives-ouvertes.fr/hal-02905392.

[92] Spiros Mancoridis, Brian S Mitchell, Chris Rorres, Yih-Farn Chen et Emden R Gansner : Using automatic clustering to produce high-level system organizations of source code. *In IWPC*, volume 98, pages 45–52, 1998.

[93] Christopher D Manning, Prabhakar Raghavan et Hinrich Schütze : Flat clustering. *Introduction to information retrieval*, pages 350–374, 2008.

[94] Peter V. Marsden : Egocentric and sociocentric measures of network centrality. *Social Networks*, 24(4):407–422, 2002.

[95] Aaron F McDaid, Derek Greene et Neil Hurley : Normalized mutual information to evaluate overlapping community finding algorithms. *arXiv preprint arXiv :1110.2515*, 2011.

[96] David W. McDonald et Mark S. Ackerman : Expertise recommender : a flexible recommendation system and architecture. *In CSCW '00 : Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 231–240, New York, NY, USA, 2000. ACM. URL http://dx.doi.org/10.1145/358916.358994.

[97] Fintan McGee, Mohammad Ghoniem, Guy Melançon, Benoît Otjacques et Bruno Pinaud : The state of the art in multilayer network visualization. *Comput. Graph. Forum*, 38(6):125–149, 2019. URL `https://doi.org/10.1111/cgf.13610`.

[98] Guy Melançon : Just how dense are dense graphs in the real world ? : a methodological note. *In Proceedings of the 2006 AVI workshop on BEyond time and errors : novel evaluation methods for information visualization*, pages 1–7. ACM, 2006.

[99] Florian Metze, Christian Bauckhage et Tansu Alpcan : The "spree" exeprt finding system. *In ICSC07 : Proceedings on the international conference on Semantic Computing., IEEE computer Society, Washington, USA.*, 2007.

[100] Amir Mohammadinejad, Reza Farahbakhsh et Noel Crespi : Consensus opinion model in online social networks based on influential users. *IEEE Access*, 7:28436–28451, 2019.

[101] Andrea Morichetta, Pedro Casas et Marco Mellia : Explain-it : Towards explainable ai for unsupervised network traffic analysis. *In Proceedings of the 3rd ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks*, Big-DAMA '19, page 22 ?28, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450369992. URL `https://doi.org/10.1145/3359992.3366639`.

[102] Tamás Nepusz, Andrea Petróczi, László Négyessy et Fülöp Bazsó : Fuzzy communities and the concept of bridgeness in complex networks. *Physical Review E*, 77(1):016107, 2008.

[103] M. E. J. Newman : The structure and function of complex networks. *SIAM Review*, 45:167–256, Mar 2003. URL `http://arxiv.org/abs/cond-mat/0303516`.

[104] M. E. J. Newman : Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Science of the United States (PNAS)*, 101:5200–5205, 2004.

[105] Mark EJ Newman : The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.

[106] Mark EJ Newman : Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3):036104, 2006.

[107] Mark EJ Newman : Spectral methods for community detection and graph partitioning. *Physical Review E*, 88(4):042822, 2013.

[108] Mark EJ Newman et Michelle Girvan : Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[109] Andrew Y. Ng, Michael I. Jordan et Yair Weiss : On spectral clustering : Analysis and an algorithm. *In ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS*, pages 849–856. MIT Press, 2001.

[110] Vincenzo Nicosia, Giuseppe Mangioni, Vincenza Carchiolo et Michele Malgeri : Extending the definition of modularity to directed graphs with overlapping communities. *Journal of Statistical Mechanics : Theory and Experiment*, 2009(03):P03024, 2009.

[111] Gergely Palla, Imre Derényi, Illés Farkas et Tamás Vicsek : Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.

[112] Alex Pothen, Horst D Simon et Kang-Pu Liou : Partitioning sparse matrices with eigenvectors of graphs. *SIAM journal on matrix analysis and applications*, 11(3):430–452, 1990.

[113] Usha Nandini Raghavan, Réka Albert et Soundar Kumara : Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3):036106, 2007.

[114] William M Rand : Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[115] Kumar Ravi et Vadlamani Ravi : A survey on opinion mining and sentiment analysis : Tasks, approaches and applications. *Knowledge-Based Systems*, 89:14–46, 2015. ISSN 0950-7051. URL `https://www.sciencedirect.com/science/article/pii/S0950705115002336`.

[116] Jörg Reichardt et Stefan Bornholdt : Statistical mechanics of community detection. *Physical Review E*, 74(1):016110, 2006.

[117] Wei Ren, Guiying Yan, Xiaoping Liao et Lan Xiao : Simple probabilistic algorithm for detecting community structure. *Physical Review E*, 79(3):036111, 2009.

[118] Martin Rosvall et Carl T Bergstrom : An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences*, 104(18):7327–7331, 2007.

[119] Martin Rosvall et Carl T Bergstrom : Mapping change in large networks. *PloS one*, 5(1):e8694, 2010.

[120] Marta Sales-Pardo, Roger Guimera, André A Moreira et Luís A Nunes Amaral : Extracting the hierarchical organization of complex systems. *Proceedings of the National Academy of Sciences*, 104(39):15224–15229, 2007.

[121] Badrul Sarwar, George Karypis, Joseph Konstan et John Riedl : Analysis of recommendation algorithms for e-commerce. *In Proceedings of the 2nd ACM conference on Electronic commerce*, EC '00, pages 158–167, New York, NY, USA, 2000. ACM. ISBN 1-58113-272-7.

[122] Badrul M. Sarwar, George Karypis, Joseph A. Konstan et John Reidl : Item-based collaborative filtering recommendation algorithms. *In World Wide Web*, pages 285–295, 2001. URL `http://citeseer.ist.psu.edu/sarwar01itembased.html`.

[123] Vincent Schickel-Zuber : *Ontology Filtering Inferring Missing User's preferences in eCommerce Recommender Systems*. Thèse de doctorat, École Polytechnique Fédérale de Lussanne, Switzerland, October 2007.

[124] Massoud Seifi, Ivan Junier, Jean-Baptiste Rouquier, Svilen Iskrov et Jean-Loup Guillaume : Stable community cores in complex networks. *In Complex Networks*, pages 87–98. Springer, 2013.

[125] Devavrat Shah et Tauhid Zaman : Community detection in networks : The leader-follower algorithm. *In Workshop on Networks Across Disciplines : Theory and Application*, pages 1–8, 2010.

[126] Huawei Shen, Xueqi Cheng, Kai Cai et Mao-Bin Hu : Detect overlapping and hierarchical community structure in networks. *Physica A : Statistical Mechanics and its Applications*, 388(8):1706–1712, 2009.

[127] Jianbo Shi et Jitendra Malik : Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000.

[128] Dalia Sulieman : *Systèmes de recommandation sociaux et sémantiques*. Theses, Université de Cergy Pontoise, janvier 2014. URL `https://tel.archives-ouvertes.fr/tel-01017586`.

[129] Dalia Sulieman, Maria Malek, Hubert Kadima et Dominique Laurent : Toward social-semantic recommender systems. *Int. J. Inf. Syst. Soc. Chang.*, 7(1):1–30, 2016. URL `https://doi.org/10.4018/IJISSC.2016010101`.

[130] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, Lars J Jensen et Christian von Mering : STRING v11 : protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613, 11 2018. ISSN 0305-1048. URL `https://doi.org/10.1093/nar/gky1131`.

[131] Damian Szklarczyk, Alberto Santos, Christian von Mering, Lars Juhl Jensen, Peer Bork et Michael Kuhn : STITCH 5 : augmenting protein ?chemical interaction networks with tissue and affinity data. *Nucleic Acids Research*, 44(D1):D380–D384, 11 2015. ISSN 0305-1048. URL `https://doi.org/10.1093/nar/gkv1277`.

[132] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang et Zhong Su : Arnetminer : extraction and mining of academic social networks. *In KDD*, 2008.

[133] Jeffrey Travers et Stanley Milgram : An experimental study of the small world problem. *Sociometry*, pages 425–443, 1969.

[134] John Urry : Mobile sociology. *BJS, The Britsih Journal of Sociology*, Volume 51:185ñ203, March 2008.

[135] CJ van Rijsbergen : Information retrieval. 1979, 1979.

[136] Ulrike Von Luxburg : A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[137] Jianxin Wang, Jun Ren, Min Li et Fang-Xiang Wu : Identification of hierarchical and overlapping functional modules in ppi networks. *IEEE transactions on nanobioscience*, 11(4):386–393, 2012.

[138] Weimao Ke, K. Borner et L. Viswanath : Major information visualization authors, papers and topics in the acm library. *In IEEE Symposium on Information Visualization*, pages r1–r1, 2004.

[139] Xuyun Wen, Wei-Neng Chen, Ying Lin, Tianlong Gu, Huaxiang Zhang, Yun Li, Yilong Yin et Jun Zhang : A maximal clique based multiobjective evolutionary algorithm for overlapping community detection. *IEEE Transactions on Evolutionary Computation*, 21(3):363–377, 2017.

[140] Xiaolan Wu et Chengzhi Zhang : Multi-label propagation for overlapping community detection based on connecting degree. *In* Albert Ali Salah, Yasar Tonta, Alkim Almila Akdag Salah, Cassidy R. Sugimoto et Umut Al, éditeurs : *Proceedings of the 15th International Conference on Scientometrics and Informetrics, Istanbul, Turkey, June 29 - July 3, 2015*. ISSI Society, 2015.

[141] Zhi-Hao Wu, You-Fang Lin, Steve Gregory, Huai-Yu Wan et Sheng-Feng Tian : Balanced multi-label propagation for overlapping community detection in social networks. *Journal of Computer Science and Technology*, 27(3):468–479, 2012.

[142] Jierui Xie, Stephen Kelley et Boleslaw K Szymanski : Overlapping community detection in networks : The state-of-the-art and comparative study. *Acm computing surveys (csur)*, 45(4):43, 2013.

[143] Jierui Xie et Boleslaw K Szymanski : Labelrank : A stabilized label propagation algorithm for community detection in networks. *In Network Science Workshop (NSW), 2013 IEEE 2nd*, pages 138–143. IEEE, 2013.

[144] Jierui Xie, Boleslaw K Szymanski et Xiaoming Liu : Slpa : Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. *In Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 344–349. IEEE, 2011.

[145] Jaewon Yang et Jure Leskovec : Structure and overlaps of communities in networks. *SNAKDD*, 2012.

[146] W.W. Zachary : An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473, 1977.

[147] Aiping Zhang, Guang Ren, Yejin Lin, Baozhu Jia, Hui Cao, Jundong Zhang et Shubin Zhang : Detecting community structures in networks by label propagation with prediction of percolation transition. *The Scientific World Journal*, 2014, 2014.

[148] Jun Zhang et Mark S. Ackerman : Searching for expertise in social networks : a simulation of potential strategies. *In GROUP*, pages 71–80, 2005.

[149] Jun Zhang, Mark S. Ackerman et Lada Adamic : Expertise networks in online communities : structure and algorithms. *In WWW '07 : Proceedings of the 16th international conference on World Wide Web*, pages 221–230, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-654-7. URL `http://dx.doi.org/10.1145/1242572.1242603`.

[150] Lei Zhang, Hebin Pan, Yansen Su, Xingyi Zhang et Yunyun Niu : A mixed representation-based multiobjective evolutionary algorithm for overlapping community detection. *IEEE Transactions on Cybernetics*, 47(9):2703–2716, 2017.

[151] Shihua Zhang, Rui-Sheng Wang et Xiang-Sun Zhang : Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Physica A : Statistical Mechanics and its Applications*, 374(1):483–490, 2007.

[152] Tao Zhou, Jie Ren, Matú š Medo et Yi-Cheng Zhang : Bipartite network projection and personal recommendation. *Phys. Rev. E*, 76:046115, Oct 2007. URL `https://link.aps.org/doi/10.1103/PhysRevE.76.046115`.