UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS

Département d'informatique et d'ingénierie

Thèse de doctorat

DÉTECTION DE COMMUNAUTÉS DANS DES RÉSEAUX COMPLEXES

préparé par: Abir Messaoudi

Prof. Omar Boussaïd, université Lumière Lyon 2 .... Membre externe
Prof. Belkacem Chikhaoui, TÉLUQ ..... Membre externe
Prof. Mohand-Saïd Allili, président ..... Membre interne
Prof. Ana-Maria Cretu ..... Membre interne
Prof. Rokia Missaoui ..... Directrice de recherche

Remerciements

C'est avec fierté que je dépose cette thèse. Ce fut une aventure parsemée de découvertes et de rencontres enrichissantes. Heureusement, différentes personnes significatives ont jalonné mon cheminement et m'ont aidé à croire en mes capacités à mener ce travail de longue haleine. Je dois les remercier et leur exprimer ma plus profonde reconnaissance.

Je souhaite remercier en premier lieu ma directrice de thèse Prof. Rokia Missaoui, qui grâce à sa rigueur, ses commentaires constructifs, ses idées et ses suggestions, a grandement contribué à la réalisation de ce projet. Elle a patiemment lu, relu, commenté et annoté les nombreux textes que je lui ai soumis. À chaque étape, elle a su me guider, me conseiller et m'encourager. Je lui suis également reconnaissante pour le temps conséquent qu'elle m'a accordée, ses qualités pédagogiques et scientifiques, sa franchise, sa sympathie, son aide dans le cheminement de mes études et la peine qu'elle s'est donnée tout au long de ce travail afin de faire de ce document ce qu'il représente

Je tiens à remercier Prof.Mohand Said Allili, Prof. Ana- Maria Cretu, Prof. Omar Boussaïd et Prof. Belkacem Chikhaoui pour l'honneur qu'ils m'ont fait en acceptant de faire partie de mon jury de thèse, pour le temps qu'ils ont consacré à ma thèse et pour leurs précieux conseils.

Je tiens à exprimer ici ma reconnaissance et mes sincères remerciements à toute l'équipe du LAboratoire de Recherche sur l'Information Multimédia (LARIM) et particulièrement Mohamed-Hamza Ibrahim et Pedro Ruas pour toutes nos discussions et leurs conseils qui m'ont accompagnés tout au long de mes recherches.

Je désire en outre remercier les chercheurs qui ont implémenté divers outils en Python et qui les ont mis à la disposition du grand public. La qualité et la disponibilité de la documentation m'ont permis d'apprendre et de me servir de ces outils pour mener diverses expérimentations.

Et bien évidemment, je ne voudrais pas terminer sans une pensée toute particulière pour

ma famille, mes proches et mes amis.
Tous les mots ne sauraient exprimer la gratitude, l'amour, le respect, la reconnaissance à mes parents Nejia et Béchir, à qui je dois la réussite, pour l'éducation qu'ils m'ont prodigué ; avec tous les moyens et au prix de toutes les sacrifices qu'ils ont consentis à mon égard. Leurs prières et leurs Bénédiction m'ont été d'un grand secours tout au long de ma vie.
Je remercie chaleureusement mon mari, Issam, pour avoir su trouver les mots justes dans les moments les plus difficiles, pour l'encouragement , et pour le soutien moral et matériel. Cher mari, j'aimerais bien que tu trouves dans ce travail l'expression de mes sentiments de reconnaissance les plus sincères car grâce à ton aide (même dans tes moments les plus durs) et à ta patience avec moi que ce travail a pu voir le jour. Que dieu le tout puissant t'accorde une bonne santé et un avenir meilleur.
Je remercie également ma soeur Balkys et mes frères Badis et Anis. Je ne saurais exprimer ma profonde reconnaissance pour vos soutiens continus dont vous avez toujours fait preuve. Vous m'avez toujours encouragée et incitée à faire de mon mieux.
Finalement, je dédie ce modeste travail et ma profonde gratitude à ma petite Lyna, âgée du même âge que ma thèse, qui m'avait accompagnée pendant chaque étape de cette thèse. Lyna, quoique nous avons vécu des moments difficiles ensemble, le simple fait que tu existes rendait ma vie agréable.

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations, sigles et acronymes

AFC Analyse formelle de concepts

FCA *Formal Concept Analysis*

DC Détection de communautés

RM Réseaux multicouches

MR *MapReduce*

COIN *COncept INterestingnes* : Une approche pour détecter des communautés dans les réseaux à un seul mode de données

CoDeBi : Une approche pour de détection de communautés dans les réseaux à deux modes

Résumé

La détection de communautés est l'un des thèmes de recherche les plus prolifiques en analyse de réseaux sociaux. Une communauté est alors décrite par un ensemble de nœuds intensément liés entre eux mais faiblement liés au reste du réseau.
Dans le cadre de cette thèse de doctorat, nous avons exploité l'analyse formelle de concepts (AFC) en tant que théorie de regroupement conceptuel ainsi que des mesures de pertinence des concepts (stabilité et séparation) pour proposer et valider deux méthodes distinctes de détection de communautés : une pour les réseaux à un seul mode (un seul type de nœuds) et une autre pour ceux à deux modes (graphe biparti).
Dans un deuxième temps, nous avons étendu ce travail à des réseaux plus complexes pour identifier des communautés au sein de deux graphes bipartis G_1 et G_2, pris conjointement et non individuellement, avec $G_1 = (U_1, V_1, E_1)$ et $G_2 = (U_2, V_2, E_2)$, où l'un des ensembles de nœuds U_1 ou V_1 de $G1$ est égal à l'un des ensembles U_2 ou V_2, et E_1 (respectivement E_2) est un ensemble de liens entre des éléments de U_1 et des éléments de V_1 (respectivement U_2 et V_2). Ce cas peut être généralisable à plus de deux graphes interreliés (multicouches). Pour traiter ce volet, nous avons appliqué des opérations d'assemblage de contextes en AFC, empruntées à des études d'analyse de concepts formels.

Enfin, pour tenir compte des réseaux sociaux complexes, il était nécessaire de s'intéresser aux réseaux d'information dits hétérogènes. Nous nous sommes intéressés en particulier aux données tridimensionnelles et nous avons adapté deux mesures de pertinence de concepts du contexte dyadique de FCA au le contexte triadique. Ce dernier exprime un réseau de la forme $G = (K_1, K_2, K_3, Y)$ où K_1, K_2 et K_3 sont trois ensembles distincts de nœuds et Y une relation ternaire vérifiant $Y \subseteq K_1 \times K_2 \times K_3$.

Chapitre 1

Introduction

1.1 Motivation et contexte

A l'ère du Web et des médias sociaux, l'analyse des réseaux est un domaine de recherche très actif qui fait appel à plusieurs disciplines et théories pour décrire et analyser des systèmes complexes dans les domaines des sciences sociales, biologiques, physiques, logistiques et informatiques. En effet, les réseaux sont partout autour de nous pusique le cerveau est un réseau de cellules nerveuses connectées par des axones et les cellules sont elles-mêmes des réseaux de molécules reliées par des réactions biochimiques. En écologie, les écosystèmes peuvent être représentés sous forme de réseaux d'espèces. Ces derniers peuvent être organisés d'une manière simplifiée en producteurs primaires (plantes), les consommateurs (animaux), les bio-réducteurs (micro-organismes), etc.

Par ailleurs, les réseaux sociaux (*Facebook, Twitter, LinkedIn* et bien d'autres) ainsi que les réseaux d'ordinateurs ou de transport ne sont que quelques autres exemples de tels réseaux. Même le langage que nous utilisons pour nous exprimer et pour transmettre nos idées est un réseau constitué de mots reliés par des relations syntaxiques. Dans un contexte social, chacun des membres est un nœud de multiples réseaux interconnectés, définis par des relations de différentes natures. Ces réseaux contiennent un volume important de données cachant une mine d'informations et de connaissances.

Certes, l'existence des réseaux n'est pas récente. Cependant, ce qui demeure récent aujourd'hui, c'est le jumelage des réseaux avec un nombre grandissant de technologies. L'étude des réseaux est devenue l'un des domaines phares du 21ème siècle et a déjà permis d'aborder et de couvrir une grande variété de problématiques dont les principales sont l'identification des nœuds centraux, la détection et l'évolution de communautés, la

prédiction de nouveaux liens, la propagation de l'influence, le maintien versus la rupture d'un réseau, etc. Dans la plupart des recherches actuelles, ces réseaux sont habituellement supposés être homogènes, modélisés par un graphe où les nœuds sont des objets d'un même type d'entité et les liens appartiennent également à un même type. Cependant, dans le monde réel, la plupart des objets se trouvent dans des réseaux complexes avec plusieurs couches interconnectées.

Au fur et à mesure que la recherche sur des systèmes complexes prend de l'ampleur, il devient de plus en plus essentiel de dépasser l'analyse de simples graphes et d'étudier des cadres plus complexes mais surtout plus réalistes. C'est la problématique que nous nous proposons de traiter dans le présent projet de recherche en nous intéressant à la détection de communautés dans des réseaux simples puis ceux multicouches par la proposition de nouvelles approches bâties sur la théorie de l'analyse formelle de concepts (AFC).

1.2 Objectifs et contributions

La détection de communautés est l'un des thèmes de recherche les plus prolifiques en analyse de réseaux sociaux. Une communauté est alors décrite par un ensemble de nœuds intensément liés entre eux mais faiblement liés au reste du réseau.

Les réseaux sociaux ont fréquemment des structures complexes comme ceux à deux modes représentés par des graphes bipartis. Plusieurs travaux sur la détection de communautés mettent l'accent soit sur l'identification de groupes disjoints ou chevauchants en procédant d'abord à la projection des données à deux modes (deux dimensions/nœuds distincts) en deux tables à un seul mode qui sont ensuite analysées. Cependant, cela entraîne une perte d'information et aboutit à des communautés mal définies. En outre, les travaux sur des réseaux multidimensionnels (trois dimensions et plus) ou à des réseaux interreliés appelés multicouches sont relativement limités. Dans le cadre de cette thèse de doctorat, nous avons exploité l'analyse formelle de concepts et des mesures de pertinence des concepts (stabilité et séparation) pour proposer et valider deux méthodes distinctes de détection de communautés : une pour les réseaux à un seul mode (un seul type de nœuds) et une autre pour ceux à deux modes (graphe biparti). Nous avons également analysé des réseaux plus complexes représentant des structures multicouches ou multidimensionnelles. Cela a débouché vers trois publications scientifiques [63, 91, 92].

Dans le premier cas, nous avons proposé un algorithme appelé COIN qui détermine d'abord les concepts formels lesquels représentent spécifiquement les cliques et les ponts

pour identifier les premiers et éliminer les seconds. Ensuite, nous exploitons l'indice de stabilité pour couper les liens (ponts) entre communautés et finalement effectuer une fusion (*percolation*) entre les cliques adjacentes. Des tests empiriques sur des réseaux sociaux connus (et communément utilisés dans la littérature) montrent que COIN peut rapidement et d'une manière plus précise déceler les communautés d'un réseau à un mode que l'algorithme de centralité d'intermédiarité des liens de Girvan et Newman et les méthodes de Louvain, Walktrap ou Infomap.

Dans le deuxième cas, nous avons défini une approche appelée CoDeBi de détection de communautés chevauchantes et même imbriquées dans les graphes bipartis en exploitant l'indice de stabilité et la séparation. L'analyse Silhouette est également utilisée pour raffiner le processus de délimitation des communautés. Des tests préliminaires sur des réseaux réels montrent que cette approche permet d'identifier correctement des communautés chevauchantes.

Ensuite, nous avons étendu ce travail à des réseaux plus complexes pour identifier des communautés au sein de deux graphes bipartis G_1 et G_2, pris conjointement et non individuellement, avec $G_1 = (U_1, V_1, E_1)$ et $G_2 = (U_2, V_2, E_2)$, où l'un des ensembles de nœuds U_1 ou V_1 de $G1$ est égal à l'un des ensembles U_2 ou V_2, et E_1 (respectivement E_2) est un ensemble de liens entre des éléments de U_1 et des éléments de V_1 (respectivement U_2 et V_2). Ce cas peut être généralisable à plus de deux graphes interreliés (multicouches). Pour traiter ce volet, nous avons appliqué des opérations d'assemblage de contextes en AFC, empruntées à des études d'analyse de concepts formels [56, 140, 137].

Enfin, nous avons pris le cas d'un réseau d'information hétérogène ayant des données tridimensionnelles et nous avons adapté deux mesures de pertinence de concepts du contexte dyadique de FCA au contexte triadique. Ce dernier réseau d'intérêt est de la forme $G = (K_1, K_2, K_3, Y)$ où K_1, K_2 et K_3 sont trois ensembles distincts de nœuds et Y une relation ternaire vérifiant $Y \subseteq K_1 \times K_2 \times K_3$.

1.3 Organisation de la thèse

La thèse est organisée en sept chapitres. Le chapitre 2 est dédié aux rappels et fondements théoriques. Il contient des rappels sur l'analyse et la variété de réseaux sociaux ainsi que sur l'analyse formelle de concepts. Le chapitre 3 définit les communautés et leur intérêt et passe en revue les travaux existants en matière de détection de tels groupes. Le chapitre 4 présente une nouvelle approche de détection de communautés au sein des

réseaux à un seul mode alors que le chapitre 5 décrit une autre approche pour la détection de communautés dans des réseaux à deux modes. Le chapite 6 décrit l'approche proposée pour la détection de communautés dans les réseaux multicouches et particulièrement deux réseaux à deux modes superposés. Le chapite 7 détaille notre approche pour la detection pour la détection de communautés dans les réseaux tridimensionnels. Enfin, le chapitre 8 résume les contributions actuelles, présente les travaux à mener à court terme aussi bien des perspectives de recherches.

Chapitre 2

Rappels et fondements théoriques

Ce chapitre est une présentation des différentes notions reliées à l'analyse des réseaux sociaux et de l'analyse formelle de concepts.

2.1 Définition et types des réseaux

Un réseau désigne généralement un ensemble d'entités (acteurs, chercheurs, animaux, plantes, cellules, etc.), que l'on nomme nœuds lesquels sont reliés entre eux par un ensemble de relations appelées liens. Les réseaux sont principalement divisés en quatre catégories : les réseaux*sociaux*, comme Twitter, Facebook et Google, les réseaux *d'information* comme les réseaux sémantiques et WordNet, les réseaux de *citations* et les réseaux *technologiques* comme Internet, les circuits électriques et électroniques, les réseaux de télécommunication, les réseaux routiers et ferroviaires, et finalement les réseaux *biologiques* comme les chaînes alimentaires, les espèces animales et les réseaux de neurones.

2.1.1 Définition

Un réseau à un seul mode est communément défini par un graphe $G = (V, E)$ dans lequel V est un ensemble des nœuds et E un ensemble d'arcs (liens) du réseau de la forme : $\{(v_i, v_j) \mid v_i, v_j \in V \text{ and } v_i \neq v_j\}$

Deux grandes familles de réseaux peuvent être distinguées : les réseaux *homogènes* et les réseaux *hétérogènes* [123].

Definition 1. *Réseau hétérogène/ homogène. Un réseau est appelé* hétérogène *si le nombre des types nœuds est $|V| > 1$ ou celui des types de relations $|E| > 1$; sinon, il s'agit d'un réseau d'information* homogène.

Notons aussi que ces réseaux peuvent être orientés ou non. En présence d'une orientation, un lien $e_1 = (v_i, v_j)$ entre v_i et v_j n'implique pas nécessairement l'existence d'un lien $e_2 = (v_j, v_i)$ et la représentation des liens est exprimée par une flèche. Pour les réseaux *non orientés*, si les nœuds v_i et v_j sont connectés, il existe également un lien entre v_i et v_j.

2.1.2 Types des réseaux

La littérature comporte diverses notions et variantes de réseaux complexes que nous allons décrire brièvement dans ce qui suit.

Réseaux homogènes

Ce sont des réseaux qui ne contiennent que des liens connectant des nœuds ayant le même type. Il s'agit alors d'un réseau à un seul mode de données. Un exemple classique peut représenter un réseau bibliographique formé par des auteurs reliés entre eux par des liens d'affinité. Bien que ce type de structure est simple et facile à manipuler, mais leur analyse n'est pas triviale. Ainsi, soient v_i et v_j deux nœuds du réseau, s'il existe un lien directe entre ces nœuds, alors les nœuds v_i et v_j sont dits adjacents, ou encore connectés ou voisins.

Réseaux hétérogènes

Un réseau d'information est dit hétérogène s'il contient plusieurs types de nœuds et/ou de liens. Les réseaux hétérogènes ont deux caractéristiques importantes : une structure complexe et une sémantique riche [123].

Les réseaux de données à deux modes sont un cas spécial des réseaux hétérogènes et sont largement utilisés pour représenter deux types de nœuds (ex. personnes et événements) et une relation qui lie un nœud d'un type avec un autre nœud de l'autre type (ex. participation de chercheurs à des événements scientifiques). Ils sont représentés par des graphes bipartis et sont également appelés des réseaux d'affiliation.

Les réseaux multi-relationnels appelés aussi multidimensionnels ou multi-modes ou composites possèdent un seul type d'objets mais plusieurs types de liens. Ce type de

réseaux existe largement dans les réseaux sociaux, tel que Facebook où les utilisateurs peuvent être liés les uns aux autres par le biais de plusieurs connexions telles que la publication de vidéos, la participation à des jeux, ou le partage de photos.

Les graphes *k-parti* contiennent plusieurs types de nœuds et des liens entre deux nœuds de deux types distincts. Les réseaux *en étoile* ont un nœud central, appelé hub ou concentrateur connectant les autres nœuds.

Dans les réseaux *complexes*, les connexions entre les nœuds ne sont ni purement régulières, ni totalement aléatoires [3]. En effet, nous pouvons observer que la plupart des réseaux du monde réel, et particulièrement les réseaux sociaux, ont souvent des structures topologiques non triviales qui représentent les interactions entre des individus. Les réseaux complexes englobent les réseaux invariants d'échelle *scale-free*, les réseaux petit monde *small world*, les réseaux aléatoires *random* et les réseaux réguliers selon leur distance géodésique moyenne, leur coefficient de clustering et leur densité. Les réseaux *réguliers* sont tels que chaque nœud possède un nombre identique de liaisons. La densité du réseau est souvent faible, alors que le coefficient de clustering est, quant à lui, relativement élevé. dans les réseaux invariants d'échelle, une faible proportion de nœuds (appelés des nœuds concentrateurs ou *hubs*) a beaucoup plus de connexions que les autres, et une forte proportion de nœuds est faiblement connectée. De telles structures sont observées principalement dans les réseaux de citations d'articles scientifiques et de pages Web. Plusieurs modèles ont été proposés pour la génération des réseaux invariants d'échelle. Le plus connu d'entre eux est le modèle Barabasi-Albert [7].

Les réseaux *petit monde* ont une distance géodésique moyenne qui croît d'une manière logarithmique en fonction du nombre de nœuds et possèdent deux propriétés : une distance moyenne relativement faible et un coefficient de clustering élevé.

Les réseaux *multicouches* [67] sont définis comme étant soit un réseau d'information hétérogène dont la structure comporte un ensemble de nœuds et de liens dont le nombre de types de sommets et/ou d'arcs dépasse 1 ou bien une superposition/alignement de plusieurs réseaux interconnectés. Pour la deuxième catégorie, cela représente deux ou plusieurs réseaux superposés qui partagent des nœuds en commun. Par exemple, l'existence d'un premier réseau social (ou professionnel) décrivant des individus avec leurs caractéristiques personnelles (ou professionnelles) et un deuxième réseau décrivant leurs liens d'amitié (ou de collaboration). C'est aussi le cas de la coexistence du réseau *Facebook* avec celui de *LinkedIn*. Ces réseaux sont reliés par des liens d'ancrage.

La notation utilisée pour les graphes en général peut être facilement étendue pour décrire des structures comportant, en plus des nœuds et des liens, des couches. Ainsi, un réseau multicouche est défini comme un quadruple [136] $M = (V_M, E_M, V, L)$ où V est l'ensemble des nœuds du réseau, $V_M \subseteq V \times L_1 \times \ldots \times L_d$ est l'ensemble des combinaisons nœud-couche, c'est-à-dire l'ensemble des couches dans lesquelles un nœud $v \in V$ est présent. L'ensemble $EM \subseteq V_M \times V_M$ est une collection de liens contenant l'ensemble des paires de combinaisons possibles de nœuds dans V_M. L'ensemble $L = \{L_a\}_{a=1}^{d}$ représente les d couches élémentaires du réseau.

Dans le cadre de cette thèse, nous allons mettre l'accent sur les réseaux multicouches (à deux couches et plus) et les réseaux tridimensionnels dont un exemple de chacun est illustré par la figure 2.1. Cette dernière illustre des réseaux bibliographiques : (a) à un mode de coauteurs, (b) à deux modes de participation d'auteurs à des conférences, (c) un réseau multicouches dont la première couche présente un réseau auteurs-conférences signifiant que des auteurs publient dans des conférences, la deuxième couche présente un réseau articles-conférences et la troisième couche présente un réseau articles-thèmes et (d) un réseau tridimensionnel représentant une relation ternaire entre des auteurs, des conférences et des thèmes.

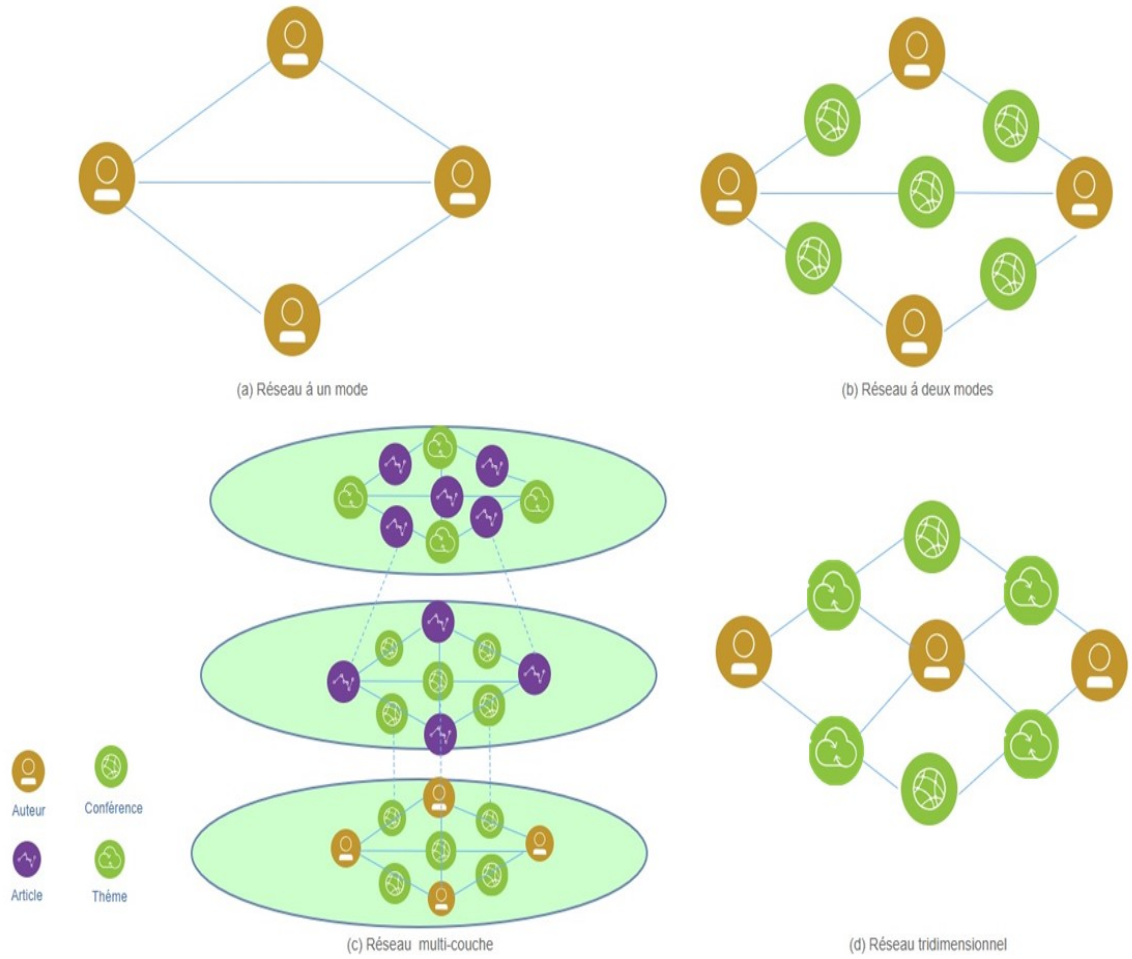


FIGURE 2.1: Quelques types de réseaux bibliographiques

2.1.3 Analyse des réseaux

On distingue deux grandes familles de méthodes d'analyse de ces données : les méthodes traditionnelles et les méthodes d'extraction des connaissances. Les premières s'appuient uniquement sur des propriétés structurelles locales ou globales du réseau pour caractériser les nœuds et la structure alors que les secondes appliquent les principes de la fouille de données pour découvrir des motifs pertinents. Un cas de cette seconde catégorie est la détection de communautés qui nous intéresse et qui est expliquée dans le chapitre 3.

Les méthodes traditionnelles

Il existe plusieurs mesures pour caractériser localement et globalement les réseaux. Les mesures *locales* apportent des informations sur le voisinage d'un nœud et s'intéressent uniquement aux propriétés structurelles des nœuds et des liens et les mesures *globales* décrivent l'ensemble de la structure en mettant en évidence certaines propriétés statistiques [125].

Mesures locales

L'article de Freeman nommé *(Centrality in social networks : Conceptual clarification)* [51] représente sans doute l'une des contributions les plus importantes dans ce domaine. En effet, les mesures de centralité apportent une information sur la connectivité des nœuds et permettent de déterminer les nœuds les plus significatifs, influents, c'est-à-dire ceux qui sont les plus actifs et qui ont le plus de liens avec les autres nœuds. Les quatre mesures de centralité les plus utilisées aujourd'hui sont [19] : *La centralité de degré (Degree centrality)* [51] représente le nombre de liens incidents à un nœud. Cette mesure est basée sur l'intuition qu'un nœud est plus important au sein d'un graphe si le nombre total des nœuds avec lesquels il interagit directement est important. Cette centralite pour le nœud i est définie par l'équation 2.1 et représente le nombre de voisins (nœuds adjacents) de i.

$$d_i = \sum_{j=1}^{n} A_{ij} \tag{2.1}$$

oe n est le nombre total de nœuds, A_{ij} est un élément de la matrice prenant la valeur 1 si le nœud i a une relation directe avec le nœud j, sinon, c'est la valeur 0.

L'ampleur de d_i dépend de la taille du réseau. Plus le réseau est large, plus il y a une grande probabilité de trouver des nœuds ayant un degré élevé. Dans certaines situations, il peut être plus approprié d'utiliser une centralité de degré qui est indépendante de la taille du réseau. Dans ce cas, la centralité de degré est normalisée selon l'équation suivante :

$$d'_i = \frac{\sum_{j=1}^{n} A_{ij}}{n-1} \tag{2.2}$$

La centralité de proximité (Closeness centrality) [51] est habituellement définie comme la somme des distances géodésiques d'un nœud à tous les autres. Etant donné deux nœuds

i et j, la distance géodésique représente le plus court chemin entre ces deux nœuds. Cette mesure considère que le nœud, ayant une distance globalement proche des autres nœuds du graphe (peut contacter les autres sans dépendre d'acteurs intermédiaires), est un nœud important, occupant une position stratégique (ou avantageuse).

La centralité de proximité d'un nœud i correspond alors à l'inverse de la somme des distances géodésiques entre i et les autres nœuds et est obtenue par l'équation 2.3

$$c_i = \frac{1}{\sum_{j=1}^{n} d_{ij}} \tag{2.3}$$

oe d_{ij} représente le plus court chemin, *i.e.* le nombre minimum d'arêtes reliant les nœuds i et j.

La centralité intermédiaire d'un nœud (Betweenness centrality) [51] est égale au nombre de fois où ce nœud est sur le chemin le plus court (intermédiaire) entre deux autres nœuds quelconques du graphe. L'idée de cette mesure est que, dans un graphe, un nœud est considéré d'autant plus important s'il est nécessaire de le traverser pour aller d'un nœud quelconque à un autre. Cette mesure est donnée par l'équation 2.4 :

$$b_i = \sum_{j}^{n} \sum_{k}^{n} b_{jk}(i) \tag{2.4}$$

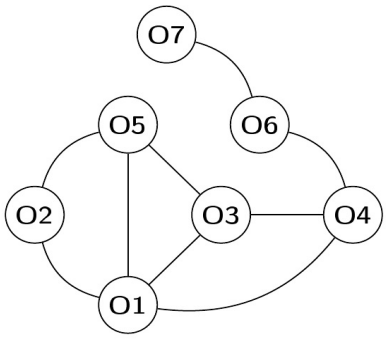
$i \neq j \neq k$, n est le nombre total de nœuds dans le graphe et $b_{jk}(i)$ est une intermédiarité partielle du nœud i. Si i se trouve entre les nœuds j et k, alors $b_{jk}(i)$ est égale à 1, sinon, $b_{jk}(i)$ vaut 0.

La centralité spectrale (eigenvector) [17, 18] est proposée par Bonacich comme une extension de la centralité de degré. L'intuition de cette mesure suggère que la centralité d'un nœud dépend de la centralité des nœuds auxquels il est connecté. Ainsi, un nœud est d'autant plus important qu'il est connecté à des nœuds qui sont eux même importants. Bonacich propose donc un indicateur prenant en compte les centralités des voisins. La centralité de vecteur propre d'un nœud i est définie par l'équation 2.5 :

$$e_i = \frac{1}{\lambda} \sum_{j=1}^{n} A_{ij} e_j \tag{2.5}$$

qui indique que la centralité de vecteur propre du nœud i est proportionnelle à la somme des centralités de ses voisins immediats. La constante λ est la plus grande valeur du vecteur propre, A_{ij} est un élément de la matrice d'adjacence et e_j est la centralité de vecteur propre du nœud j.

La transitivité ou le coefficient de regroupement (Clustering Coefficient) d'un sommet est proposée par Watts et Strogatz [138]. Essentiellement, cette propriété stipule que si $u \Leftrightarrow v$ et $v \Leftrightarrow w$, alors la probabilité que $u \Leftrightarrow w$ est plus élevée. Le coefficient de regroupement local d'un sommet dans un graphe quantifie la proximité de ses voisins par rapport à une clique (graphe complet). Autrement dit, il mesure à quel point ses voisins sont proches d'une clique. Plus le coefficient est grand, plus le voisinage est proche d'une clique. Structuralement, ceci signifie qu'il y a plus de triangles dans le graphe.
On définit le coefficient de regroupement d'un nœud i par :
$C_i = \frac{Nombredeconnexionsentrelesvoisinsdei}{Nombremaximaldeconnexionspossiblesentrelesvoisinsdei}$
Le figure 2.2 illustre les calculs de quelques mesures locales que nous avons présentées ci-dessus.



	O1	O2	O3	O4
Degré (I)	4	2	3	3
coefficient de clustering	$\frac{3}{6}$	1	$\frac{2}{3}$	$\frac{1}{3}$
proximité	$\frac{9}{6}$	$\frac{13}{6}$	$\frac{10}{6}$	$\frac{9}{6}$
Intermediarité	$\frac{5}{15}$	0	$\frac{1,5}{15}$	$\frac{8}{15}$

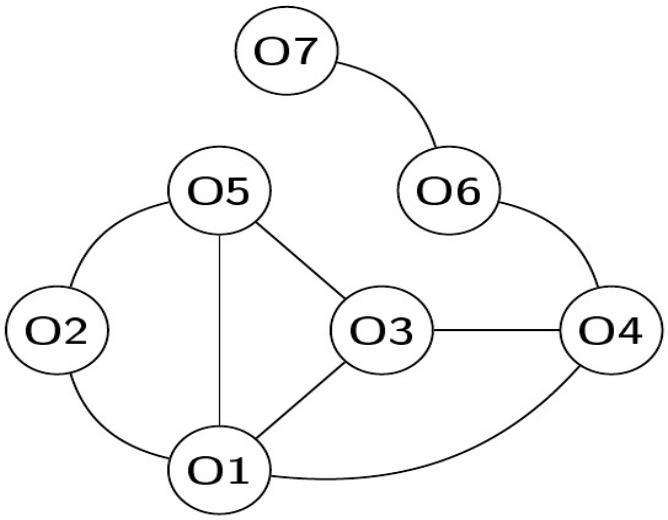
FIGURE 2.2: Quelques mesures locales

Mesures globales
Il existe une multitude de mesures, les plus utilisées sont :

La densité d'un graphe est l'un des attributs les plus fondamentaux d'un réseau. Elle est définie par le rapport entre le nombre de liens existants présents divisé par le nombre total de liens possibles. La densité d'un graphe varie entre 0 (graphe composé de nœuds isolés) et 1 (graphe complet). Ainsi, un graphe est dit dense lorsque le nombre de liens est égal ou proche du nombre maximal de liens possibles.
Degré moyen présente la moyenne de la centralité de degré des nœuds du réseau, définie ci-dessus. Cette mesure est mieux adaptée que la densité pour comparer deux réseaux de tailles très différentes. Mentionnons qu'il existe une relation entre ces deux mesures qui est donnée par :

$Degré\ moyen = Densité * (n - 1)$ où n est le nombre de nœuds.

*Le diamètre*d'un graphe signifie la plus grande distance possible qui peut se trouver entre deux nœuds quelconques. La distance étant définie par la longueur du plus court chemin entre deux nœuds.
Le Coefficient de clustering est donné par la moyenne des coefficients de clustering des sommets (définie au dessus). Cette mesure peut être vue aussi comme la proportion de triangles (groupes de trois nœuds connectés) parmi toutes les triades connectées (groupes de trois nœuds dont au moins deux connectés). Autrement dit, ce coefficient est le rapport entre le nombre de boucles de longueur trois et le nombre de chemins de longueur deux.
La figure 2.3 illustre les calculs de quelques mesures globales que nous avons présentées ci-dessus.



Mesures Globales	
La densité	0.40
Le diamètre	4
Le cœfficient de clustering	0.70
Le degré moyen	2.6

FIGURE 2.3: Quelques mesures globales

Cliques sont les sous-graphes les plus fortement connexes au sein d'un graphe, une clique est tant définie comme étant un sous graphe maximal complet (dont les sommets sont tous connectés les uns aux autres) comprenant au minimum trois sommets [2]. L'identification de la clique maximale d'un graphe est un problème d'optimisation classé comme NP- complet. C'est ainsi que de nombreuses variantes des cliques ont été conçues au fil des années et citons parmi eux [11] :

N-clique [88] est sous-graphe maximal tel que la distance de chaque paire à ses sommets ne soit pas supérieure à n.

N-clans est un type particulier des n-cliques, imposant aux n-cliques présents dans le graphe un diamètre maximal égal à n.

k-core est un sous-graphe dans lequel tous les sommets sont reliés à au moins k autres sommets de ce sous-graphe (tolérance sur le nombre de liens absents).

k-plex [119] est un sous-graphe maximal dans lequel tous les sommets envoient au minimum x − k liens vers les autres sommets de ce sous-graphe.

La figure 2.4 illustre un graphe contenant diverses structures, dont ces dernières sont

présentées par le tableau 2.1.

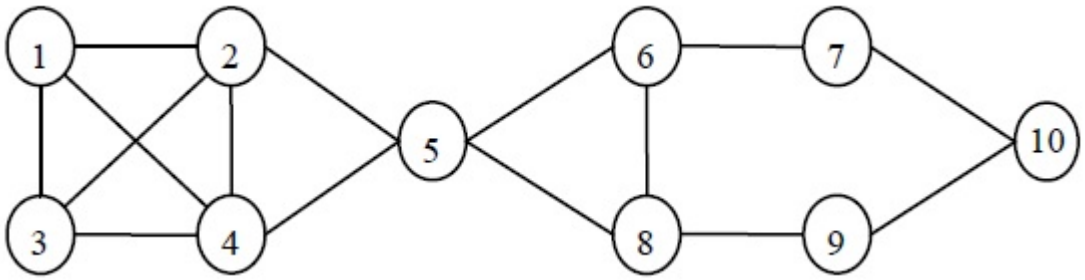


FIGURE 2.4: Exemple d'un graphe contenant diverses structures

Cliques	variantes
$Trois\ \ cliques$	{1,2,3,4},{2,4,5},{5,6,8}.
$Quatre\ \ 2-cliques$	{1,2,3,4,5},{2,4,5,6,8},{5,6,7,8,9},{6,7,8,9,10}.
$Trois\ \ 2-clans$	{1,2,3,4,5},{2,4,5,6,8},{6,7,8,9,10}.
$Cinq\ \ 2-clubs$	{1,2,3,4,5},{2,4,5,6,8},{5,6,7,8},{5,6,8,9},{6,7,8,9,10}.
$Un\ \ 3-core$	{1,2,3,4}.
$Trois\ \ 2-plex$	{1,2,3,4,5},{2,4,5,6,8},{6,7,8,9,10}.

Tableau 2.1: Exemple de cliques et ses variantes

2.2 Analyse formelle de concepts

L'analyse formelle de concepts [55] est un formalisme de représentation et de découverte de la connaissance, basé sur la formalisation des concepts et la hiérarchie de concepts. Elle est considérée comme une méthode de regroupement conceptuel *(conceptual clustering)* puisqu'elle permet de produire des groupes homogènes à la fois chevauchants (*overlapping clusters*) et conceptuels du fait que chaque groupe représenté par un concept formel est décrit non seulement par son extension, mais également par son intention (sa description).

Définition 2.1. Soit $\mathbb{K} = (\mathcal{G}, \mathcal{M}, \mathcal{I})$ *un contexte formel* où $\mathcal{G}$, $\mathcal{M}$ et $\mathcal{I}$ sont respectivement un ensemble d'objets, une collection d'attributs et une relation binaire entre $\mathcal{G}$ et $\mathcal{M}$. L'expression $(g, m) \in I$ ou encore gIm signifie que l'objet g possède l'attribut m.

Le tableau 2.2 représente un contexte très étudié dans les réseaux à deux modes appelé "les femmes du Sud" *(Southern Women)* ou encore l'ensemble des données de "Davis". Ce contexte a été tant utilisé comme référence pour tester les algorithmes de détection de communautés des graphes bipartis car il possède une réalité de terrain [53]. Il comporte 18 femmes (formant l'ensemble $\mathcal{G}$) qui ont participé aux 14 événements différents (formant l'ensemble $\mathcal{M}$).

$\mathbb{K} = (G, M, I)$	**E1**	**E2**	**E3**	**E4**	**E5**	**E6**	**E7**	**E8**	**E9**	**E10**	**E11**	**E12**	**E13**	**E14**
EVELYN	X	X	X	X	X	X		X	X					
LAURA	X	X	X		X	X	X	X						
THERESA		X	X	X	X	X	X	X	X					
BRENDA	X		X	X	X	X	X	X						
CHARLOTTE			X	X	X		X							
FRANCES			X		X	X		X						
ELEANOR					X	X	X	X						
PEARL						X		X	X					
RUTH					X		X	X	X					
VERNE							X	X	X			X		
MYRNA								X	X	X		X		
KATHERINE								X	X	X		X	X	X
SYLVIA							X	X	X	X		X	X	X
NORA						X	X		X	X	X	X	X	X
HELEN							X	X		X	X	X		
DOROTHY								X	X					
OLIVIA									X		X			
FLORA									X		X			

Tableau 2.2: Exemple de contexte formel du réseau "Femmes du Sud" $\mathbb{K}$

Comme on le voit dans la tableau 2.2, le contexte formel est représenté sous forme d'un tableau où les objets sont en lignes et les attributs en colonnes. L'intersection des lignes et des colonnes représente la relation binaire $\mathcal{I}$ entre les objets et les attributs.

Définition 2.2. Un *concept formel* c est une paire d'ensembles $c := (A, B)$ avec $A \subseteq G$, $B \subseteq M$, $A = B'$ et $B = A'$, où A' est l'ensemble des attributs partagés par les objets dans A et B' est l'ensemble des objets ayant tous leurs attributs dans B. Les sous-ensembles A et B sont appelés respectivement l'extension et l'intention du concept c. Les valeurs de A' et B' sont obtenues comme suit : $A' := \{m \in M \mid gIm \ \forall g \in A\}$ et $B' := \{g \in G \mid gIm \ \forall \ m \in B\}$.

En partant du contexte précédent, la paire $(\{EVELYN, LAURA\}, \{1, 2, 3, 5, 6, 8\})$ [1] forme un concept. En effet, $A = \{EVELYN, LAURA\} \subseteq \mathcal{G}$, $B = \{1, 2, 3, 5, 6, 8\} \subseteq \mathcal{M}$. L'ensemble des attributs partagés par les objets dans A, c'est-à-dire par les objets $EVELYN$ et $LAURA$, coïncide avec $\{1, 2, 3, 5, 6, 8\}$, donc $B = A'$. De même, l'ensemble des objets possédant les attributs dans B, c'est-à-dire les attributs $1, 2, 3, 5, 6et8$, coïncide avec $\{EVELYN, LAURA\}$, et donc $A = B'$.
Un sous-ensemble X est fermé si $X'' = X$. Un concept objet pour l'objet g est une paire de la forme $\gamma(g) = (g'', g')$ alors que le concept attribut pour l'attribut m est $\mu(m) = (m', m'')$. Les sous-ensembles fermés de $\subseteq G$ sont les extensions alors que les sous-ensembles fermés de $\subseteq M$ sont les intentions de $\mathbb{K}$.

Définition 2.3. Un *treillis de concepts* $\underline{\mathfrak{B}}(G, M, I)$ [55] est un treillis résultant de l'ordre partiel existant entre les concepts du contexte $\mathbb{K} = (\mathcal{G}, \mathcal{M}, \mathcal{I})$.

$$(A, B) \leq (C, D) \Leftrightarrow A \subseteq C \text{ et } D \subseteq B$$

(A, B) est alors un sous-concept ou prédecesseur de (C, D) alors que ce dernier est un successeur de (A, B). (A, B) est dit prédécesseur immédiat de (C, D) si en plus de l'inégalité précédente, lorsqu'il existe (E, F) tel que $(A, B) \leq (E, F) < (C, D)$, alors $(A, B) = (E, F)$. Autrement dit, il n'existe pas de concept entre (A, B) et (C, D).

Nous pouvons construire le treillis selon différents algorithmes et outils. Plusieurs algorithmes ont été conçus pour générer les concepts ainsi que leur treillis tels que [78] : Bordat, Next Closure, Close by One, Lindig, Chein, Nourine, Norris, Godin, Dowling. Ces algorithmes utilisent différentes techniques pour générer le treillis, certainement avec différentes complexités temporelles, mais les performances varient selon la taille et la densité des contextes .

Notons aussi que quelques algorithmes ont été conçus pour générer les concepts uniquement, à savoir In-Close [2] [6], NextConcept[54], Close-By-One [75] et Krajca[70]. Le temps d'exécution asymptotique de ces derniers ne dépend que linéairement de la taille du treillis de concepts*(Concept lattice)*.

1. Nous utiliserons la notation simplifiée $(EVELYNLAURA, 123568)$ à la place de la notation $(\{EVELYN, LAURA\}, \{1, 2, 3, 5, 6, 8\}))$.
2. https ://sourceforge.net/projects/inclose/

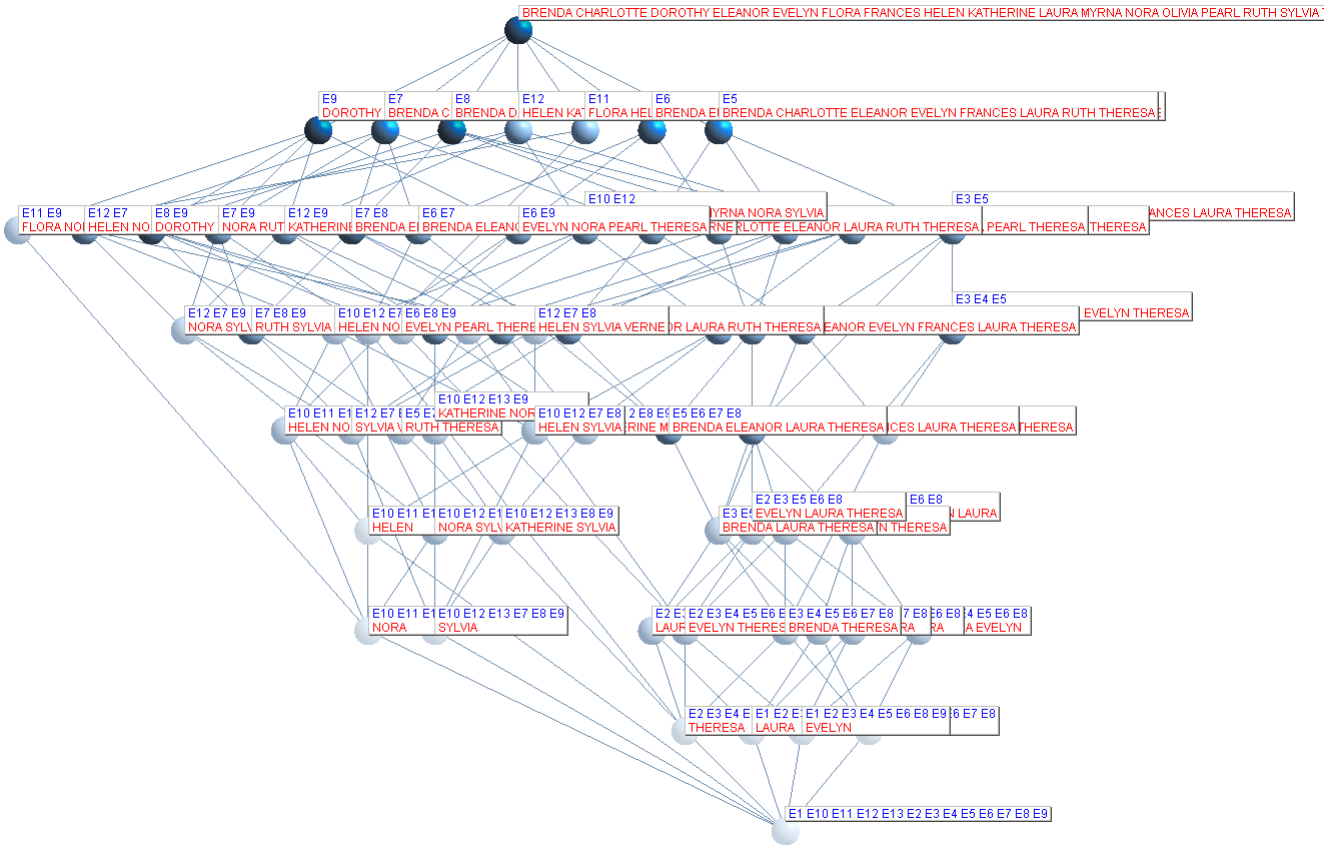


FIGURE 2.5: Exemple de treillis de concepts pour le contexte du tableau 2.2

Dans un treillis de concepts, les nœuds représentent des concepts et les arcs représentent une relation d'ordre. Comme la distance entre les nœuds n'a aucune signification particulière, nous pouvons positionner les nœuds d'une manière qui rend le treillis plus visible. Dans la figure 2.2, on est en présence d'un treillis à étiquetage complet (*full labeling*), ce qui signifie que l'on affiche systématiquement l'extension et l'intention des divers concepts. Certains outils offrent un étiquetage réduit (*reduced labeling*) où un objet est indiqué uniquement la première fois qu'il est rencontré à partir du bas du treillis, et dualement, un attribut est indiqué uniquement la première fois qu'il est rencontré à partir du haut du treillis.

2.3 Les mesures des concepts formels

Les concepts formels jouent un rôle important dans la découverte des connaissances. Toutefois, ils peuvent être démesurément nombreux. Récemment, plusieurs approches ont été introduites pour traiter le problème : soit par la sélection des concepts avec des extensions dépassant un certain seuil [73], ou la construction d'une sous structure dite *Iceberg lattice*, [127] ou la détection des concepts plus pertinents.

Une idée qui devient de plus en plus omniprésente est d'évaluer la pertinence des concepts au moyen d'indices d'intérêt comme : la stabilité *(Stability)* [74], la robustesse *(Robustness)* [135], la probabilité du concept *(Concept Probability)* [68], la séparation *(Separation)* [68], la fréquence *(Frequency)* [90],etc. Dans ce que suit, nous rappelons la notion de l'indice de stabilité et le coefficient de séparation qui nous intéressent en particulier.

2.3.1 Stabilité

Bien que plusieurs mesures d'intérêt aient été introduites pour sélectionner les concepts pertinents, l'indice de stabilité s'est avéré être le plus approprié [76] .

La stabilité intensionnelle σ(c) du concept $c = (A, B)$ mesure la force de dépendance entre l'intention B et les objets de l'extension A. Plus précisément, elle exprime la probabilité de maintenir B fermé lorsqu'un sous-ensemble d'objets dans A est supprimé avec une même probabilité. Aussi, un concept stable résiste au bruit et ne s'effondre pas lorsque certains objets sont supprimés de son extension.

σ(c) est définie comme suit :

$$\sigma(c) = \frac{\mid \{e \in \mathcal{P}(A) | e' = B\} \mid}{2^{|A|}} \tag{2.6}$$

En considérant les communautés, la stabilité aide à identifier les concepts pertinents qui représentent des groupes cohésifs.

2.3.2 Séparation

La séparation $\alpha(c)$ d'un concept formel $c = (A, B)$ [68] est calculée comme suit

$$\alpha(c) = \frac{\mid A \mid \cdot \mid B \mid}{\sum_{g \in A} \mid g' \mid + \sum_{m \in B} \mid m' \mid - \mid A \mid \cdot \mid B \mid} \tag{2.7}$$

où g' est l'ensemble des attributs de l'objet g de A et m' est l'ensemble des objets associés à l'attribut m dans B.

L'indice de séparation estime la spécificité de la relation objet-attribut d'un concept par rapport au contexte formel. Il est défini comme une partie de la zone couverte par un concept formel parmi tous les éléments non nuls dans les lignes et les colonnes correspondant au concept formel. Par conséquent, l'intérêt principal de cet indice est la qualité des attributs de ce concept. Plus les attributs sont exclusifs et rares pour ce concept, plus la séparation est grande. De même, s'il s'agit des attributs populaires partagés par tous les objets, ces attributs seraient considérés comme bruyants.

2.4 Indice de silhouette

Il y a une autre mesure de qualité d'un regroupement *(clustering)*,qui n'est pas propre à l'analyse formelle de concepts, c'est le cœefficient *(Indice)* de silhouette . Le coefficient de silhouette $\mathcal{S}(o)$ est une mesure interne introduite par Rousseeuw dans [117]. Concrètement, pour un objet donné o_i dans le jeu de données, notons C_a le groupe*(cluster)* auquel il a été affecté et $a(o_i)$ la distance moyenne entre o_i et tous les autres objets du même *(cluster)*.

Considérons maintenant tout groupe*(cluster* C_c différent de C_a et calculons la distance moyenne entre o_i et tous les autres objets dans C_c. Soit $b(o_i)$ la distance moyenne la plus basse de o_i vers tous les points de tout autre groupe, dont o_i n'est pas membre. Le cœefficient de silhouette pour l'objet o_i est donné par :

$$\mathcal{S}\left(o_i\right) = \frac{b\left(o_i\right) - a\left(o_i\right)}{max\{a\left(o_i\right), b\left(o_i\right)\}} \tag{2.8}$$

où $a(o_i)$ est la distance moyenne entre o_i et les autres objets de son groupe et $b(o_i)$ est la distance moyenne entre o_i et les objets du groupe le plus proche (n'incluant pas o_i) $b(o_i)$ et $a(o_i)$ sont exprimés comme suit :

$$b(o_i) = min_{C_c}\left(\frac{1}{|C_c|}\sum_{o_j \in C_c} d(o_j, o_i)\right) C_c \neq C_a \tag{2.9}$$

$$a(o_i) = \frac{1}{|C_a|}\sum_{o_j \in C_a} d(o_j, o_i) \tag{2.10}$$

La valeur de $\mathcal{S}(o)$ est comprise dans l'intervalle [-1,1] où ; une valeur négative indique que l'objet o_i est affecté au mauvais cluster, tandis qu'une valeur de 0 indique que o_i est très proche de la limite de décision entre deux clusters voisins. Une valeur proche de 1 reflète le fait que l'objet se trouve dans le meilleur cluster et loin des clusters voisins. Le coefficient de silhouette d'un groupe est ensuite calculé en tant que moyenne des silhouettes de ses éléments.

2.5 Opérations sur les contextes

Dans [56, 140, 137], quelques opérations algébriques de manipulation de contextes ont été proposées. Nous allons exploiter les trois opérations suivantes : l'apposition, la subposition et la concaténation. L'apposition (resp. la subposition) consiste à assembler verticalement (resp. horizontalement) deux contextes ayant le même ensemble G d'objets (resp. ensemble M d'attributs) mais deux ensembles disjoints d'attributs M_1 et M_2 (resp. d'objets G_1 et G_2) pour produire un nouveau contexte formel qui pourra ensuite être analysé par la procédure décrite dans le chapitre 5. Dans cette perspective, notons que la démarche pourra être étendue à plus de deux graphes interreliés (multicouches) en produisant d'abord le contexte issu de la composition de p contextes interreliés.

2.5.1 Apposition

Cette opération consiste à assembler verticalement deux contextes ayant le même ensemble d'objets. Prenons les deux contextes $\mathbb{K}_1 := (G_1, M_1, I_1)$ et $\mathbb{K}_2 := (G_2, M_2, I_2)$, Si $G = G_1 = G_2$, alors $\mathbb{K}$ est l'apposition de ces contextes tel que $\mathbb{K} := \mathbb{K}_1 \mid \mathbb{K}_2 = (G, M_1 \cup M_2, I_1 \cup I_2)$.

2.5.2 Subposition

D'une manière duale, la subposition consiste à assembler horizontalement deux contextes ayant le même ensemble d'attributs, pour les mêmes contextes précités $\mathbb{K}_1 := (G_1, M_1, I_1)$ et $\mathbb{K}_2 := (G_2, M_2, I_2)$, Si $M = M_1 = M_2$ alors $\mathbb{K}$ est la subposition de ces contextes tel que $\mathbb{K} := \mathbb{K}_1 \cup \mathbb{K}_2 = (G_1 \cup G_2, M, I_1 \cup I_2)$.

2.5.3 Concaténation

La concaténation consiste à assembler deux contextes dont l'ensemble des attributs dans le premier présente l'ensemble des objets dans le deuxième.
Formellement, la concaténation de deux contextes $\mathbb{K}_1 := (G_1, M_1, I_1)$ et $\mathbb{K}_2 := (G_2, M_2, I_2)$ tel que $M_1 = G_2$ est exprimée par $\mathbb{K} := \mathbb{K}_1 \odot \mathbb{K}_2 := (\mathbb{K}_1 \mid \mathbb{K}_3) \cup (\mathbb{K}_4 \mid \mathbb{K}_2)$ où $\mathbb{K}_3 := (G_1, M_2, I_1 \circ I_2)$ et $\mathbb{K}_4 := (M_1, M_1, I_1 \star I_2)$. L'opération de composition $I_1 \circ I_2$ consiste à déterminer les attributs de M_2 à associer à chaque objet o_i de G_1 de sorte que $o_i I_1 p_j I_2 q_k$ soit vrai pour un attribut p_j de M_1 avec $q_k \in M_2$. En effet, cela permet de déterminer d'une manière transitive les propriétés de M_2 que possède par transitivité chaque objet de G_1 en plus de ceux de l'ensemble M_1 décrits dans le contexte $\mathbb{K}_1$.

2.6 Analyse triadique de concepts

L'analyse triadique de concepts a été introduite par Lehmann et Wille [81] comme une extension à l'analyse formelle de concepts. On part d'un contexte triadique $\mathbb{K} := (K_1, K_2, K_3, Y)$ avec K_1, K_2 et K_3 représentant respectivement un ensemble d'objets, un ensemble d'attributs et un ensemble de conditions, et $Y \subseteq K_1 \times K_2 \times K_3$ une relation ternaire entre les trois ensembles. Le triplet (a_1, a_2, a_3) dans Y signifie que l'objet a_1 possède l'attribut a_2 sous la condition a_3 (voir le tableau 2.3). Le but d'une telle analyse est d'identifier des concepts et des implications triadiques.

$\mathbb{K}$	P	N	R	K	S
1	abd	abd	ac	ab	a
2	ad	bcd	abd	ad	d
3	abd	d	ab	ab	a
4	abd	bd	ab	ab	d
5	ad	ad	abd	abc	a

Tableau 2.3: Un contexte triadique $\mathbb{K} := (K_1, K_2, K_3, Y)$

L'exemple du tableau 2.3 représente un cube de données à trois dimensions (chercheurs, événements, et rôles). Il y a cinq chercheurs du groupe K_1 notés de 1 à 5 qui

assistent à des événements (**P**, **N**, **R**, **K** et **S**) du groupe K_2 avec différents rôles de l'ensemble K_3. Ces rôles $a, b, cetd$ représentent respectivement auteur, organisateur, orateur principal (conférencier) et membre du PC .

Définition 2.4. Un concept triadique d'un contexte $\mathbb{K} := (K_1, K_2, K_3, Y)$ est un triplet (A_1, A_2, A_3) avec $A_1 \subseteq K_1$, $A_2 \subseteq K_2$, $A_3 \subseteq K_3$ et $A_1 \times A_2 \times A_3 \subseteq Y$. Il représente un cuboïde plein de 1. Les sous-ensembles A_1, A_2 et A_3 sont appelés respectivement l'extension, l'intention et le mode (*modus*) du concept.

Du tableau 2.3, on peut extraire plusieurs concepts triadiques comme $(12345, PRK, a)$ et $(14, PN, bd)$. Par contre, le triplet $(135, PN, d)$ n'est pas un concept car il est non maximal puisque son extension peut être augmentée pour aboutir au concept $(12345, PN, d)$ tout en respectant la relation ternaire. En regardant la figure 2.6, on voit qu'il y a deux concepts triadiques associés au nœud n° 13 et ayant la même extension $\{1, 4\}$. Il s'agit de $(14, PNK, b)$ et $(14, PN, bd)$.

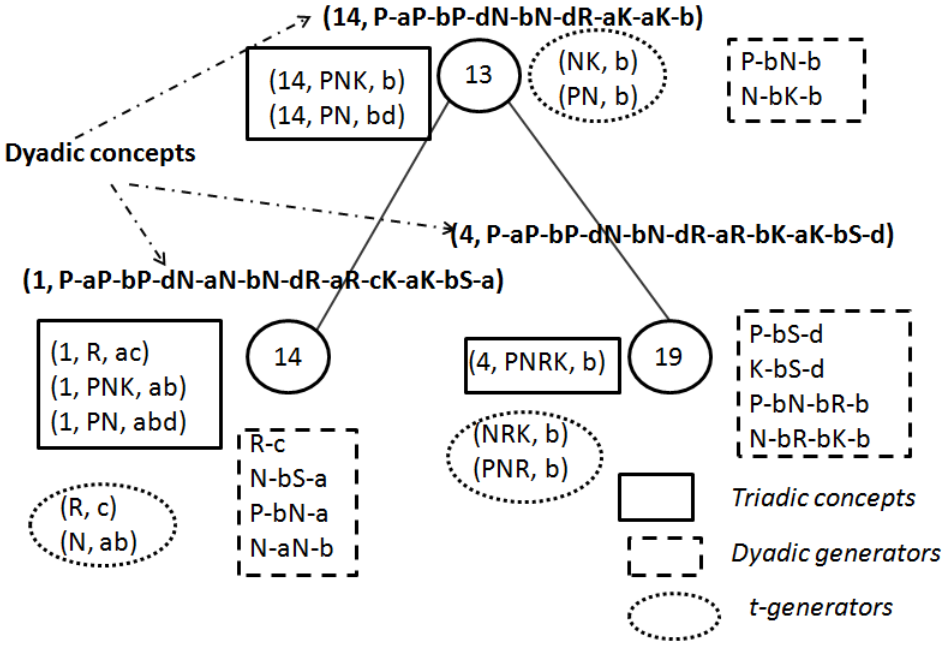


FIGURE 2.6: Concepts triadiques et dyadiques

Chapitre 3

Détection de communautés

3.1 Les communautés : définition et intérêt

Une communauté est un ensemble de nœuds qui se connectent davantage les uns aux autres, en partant du principe qu'ils partagent les mêmes ressources, ont des propriétés similaires, des intérêts ou des affinitéss, ou parfois sont réunis par des circonstances, des situations externes ou des frontières géographiques. Dans une communauté *disjointe*, un nœud appartient à une seule communauté. La communauté disjointe est également appelée affectation précise *(crisp assignment)*, dans laquelle une relation binaire est établie entre un nœud et une communauté. Un nœud peut appartenir tout au plus à une communauté [49]. Cependant, dans une communauté *chevauchante* [141], un nœud peut appartenir à plusieurs communautés. Ceci est également connu sous le nom d'affectation floue *(fuzzy assignment)* des nœuds. Une communauté *imbriquée* se trouve au sein de communautés plus grandes. Par exemple, une communauté sociale d'amis ayant le même encadreur est une communauté imbriquée dans la communauté d'amis ayant étudié dans le même programme, laquelle est imbriquée dans la communauté d'amis de la même université.

La découverte de communautés d'individus qui ont tendance à communiquer davantage entre eux qu'avec le reste est avantageuse dans plusieurs domaines comme par exemple le domaine financier, le marketing, la sécurité, la médecine et l'informatique. En effet, dans le domaine du marketing, la connaissance des communautés donne une image plus précise sur la structure de collaboration, aidant ainsi à effectuer des actions ciblées et ajuster les recommandations. Dans le domaine médical, cela permet d'étudier les phénomènes de propagation, d'identifier le processus de transmission des maladies

infectieuses telles que la grippe ou les maladies transmissibles. Il est aujourd'hui reconnu [46] que les principaux vecteurs de transmission des maladies sont les contacts de proximité. Dans le domaine de la sécurité, cela pourra servir à étudier la diffusion des informations ou des rumeurs. Zanette [143] a montré que les relations sociales et les liens de confiance et d'influence que possède un individu avec son entourage sont les principaux facteurs impliqués dans la diffusion de l'information et des rumeurs. Typiquement, il en va de même pour la propagation des virus en informatique ou en médecine. Dans cet ordre d'idées, Jure affirme dans [83] que les choix et les comportements des individus sont majoritairement déterminés par les relations sociales avec leur entourage et les communautés auxquelles ils appartiennent. Ainsi, des études récentes ont montré que les états physiologiques, spirituels ou comportementaux des individus sont majoritairement dépendants de leurs liens sociaux avec leur entourage et leur appartenance à des groupes. Notons à titre d'exemple, le tabagisme [31], l'obésité [30] et le bonheur [50].

3.2 État de l'art

La détection de communautés est souvent traitée comme un problème de regroupement *(clustering)* en fouille de données ou apprentissage machine et comme une situation de partitionnement de graphes en théorie des graphes. On peut alors détecter des communautés disjointes, chevauchantes ou même imbriquées. Elle est basée principalement par la notion d'homophilie laquelle se réfère à la tendance des personnes a avoir des liens avec des personnes qui leur sont similaires et s'avère très utilisée au sein des réseaux sociaux puisqu'elle a des implications importantes sur la façon dont l'information circule.

Fortunato a montré en [49] que la détection de communauté est un problème NP-difficile à résoudre. Concrètement, plus la structure du réseau est complexe, plus l'optimum sera difficile à détecter dans un temps raisonnable. Ainsi, plusieurs chercheurs ont eu recours à des méthodes d'optimisation issues de la théorie des graphes, de l'apprentissage machine, de l'analyse formelle de concepts et des statistiques dans le but d'avoir des solutions presque optimales en terme du temps d'exécution et de l'espace mémoire. Cependant, il peut ne pas y avoir de forte corrélation entre les communautés trouvées par un algorithme et la vérité du terrain car la formation de communautés réelles peut être le résultat de nombreuses règles en interaction et potentiellement non mesurables.

En se basant sur une revue des méthodes de détection de communautés [65, 49, 29, 131, 122, 100, 4], nous proposons de classer les méthodes de détection de communautés en quatre catégories non exclusives :

— les méthodes de classification hiérarchiques

— les méthodes centrées sur la structure globale ou locale des réseaux

— les méthodes d'optimisation d'une fonction objective

— les méthodes alternatives.

3.2.1 Méthodes hiérarchiques

Deux approches opposées sont largement expérimentées : les approches agglomératives (ascendantes) et les approches divisives (descendantes).

Approches d'agglomératives

Dans cette catégorie, on part des objets et on constitue itérativement des groupes par fusion si la similarité est suffisamment élevée. Le choix de la mesure de similarité (fonction de qualité) dépend du contexte et des exigences de l'application. Plusieurs méthodes ont été proposées dont : l'algorithme glouton de Newman *Fast Greedy*, WalkTrap, Louvain et Agarwal. La première [96] a pour principe de considérer l'ensemble des sommets du réseau comme étant des communautés individuelles. Ensuite, les communautés seront jointes deux à deux en utilisant la modularité comme métrique de jointure entre elles. A chaque étape, l'algorithme calcule la variation de la modularité pour chaque paire de communautés et joint la paire qui donne une meilleure augmentation de la modularité. Bien que l'algorithme Fast Greedy ait fonctionné relativement bien, il s'est averé assez lent. C'est pourquoi Clauset et al. [33] ont proposé la modularité gloutonne rapide *Fast greedy modularity(FGM)* qui effectue la même optimisation gloutonne que l'algorithme de Newman et donne donc des résultats identiques mais plus rapidement. Les auteurs dans [33] ont exploité certaines structures de données plus sophistiquées notamment (tas-max *(max-heaps)*, arbres binaires.). La méthode WalkTrap [107] se base sur le principe de la marche aléatoire pour mesurer la similarité entre les nœuds. Plus précisément, la distance entre deux nœuds i et j est exprimée par la différence entre le comportement de deux marcheurs aléatoires commençant respectivement aux nœuds i et j et effectuant une marche de longueur fixe. La méthode Louvain [15] implante une méthode d'optimisation gloutonne locale de la modularité. À l'état initial, chaque nœud présente une communauté. Pour chaque nœud i, on évalue le gain de la modularité si on le déplace

dans la communauté de ses voisins directs. Ainsi, on déplace i dans la communauté du voisin s'il maximise le gain de la modularité. si aucun gain n'est trouvé, le nœud reste dans sa communauté. La méthode d'Agarwal [1] mesure la similarité entre les nœuds à travers la mesure de centralité d'intermédiarité.

Approches divisives

Dans cette catégorie, on part du graphe entier pour procéder à la formation de groupes de plus en plus fins en coupant les liens reliant des sommets faiblement similaires. Les méthodes existantes se distinguent par le choix des liens à éliminer et par les poids accordés aux liens. L'exemple le plus populaire de ces approches est l'algorithme de Girvan-Newman (GN)[95] qui sera détaillé dans la section suivante et qui comporte des variantes [139, 145, 110, 112].

3.2.2 Méthodes centrées sur la structure globale ou locale des réseaux

Dans ce groupe de méthodes, les nœuds sont regroupés en communautés en fonction soit des propriétés topologiques locales partagées ou de la structure globale du réseau.

— *Propriétés topologiques locales partagées*

Les définitions locales se concentrent sur le sous-graphe à l'étude, y compris éventuellement son voisinage immédiat tout en négligeant le reste du graphe. Les communautés extraites sont principalement les sous-graphes maximaux qui ne peuvent pas être agrandis avec l'ajout de nouveaux nœuds ou liens sans perdre la propriété qui les définit. L'exemple le plus simple consiste à assimiler une communauté à une clique maximale. La technique la plus populaire est celle de percolation de cliques (CPM) de Palla et al. [101]. Cependant, ceci est difficile à envisager car le problème de cliques maximales est NP difficile[16, 23]. Ainsi, des variantes de cliques ont été envisagées pour détecter les communautés [103]. Il s'agit du calcul des $n-$cliques [2, 88], $n-$clans et $n-$clubs [93], k-plex [119], k-cores [118, 9], LS-sets [87], bicliques [82, 102] et BiTector [43].

— *Approches centrées réseau*

Ces approches exigent l'examen de la structure globale du réseau pour décomposer le graphe en communautés. Étant donné un graphe G ayant n nœuds à regrouper en *clusters*, une approche simple consiste à construire une matrice de similarité entre les nœuds du graphe en utilisant une mesure de similarité topologique. Différentes

mesures de similarité topologiques ont été définies. Un premier exemple de cette approche est l'algorithme de Girvan-Newman [95] où l'heuristique appliquée repose sur le principe que les liens inter-communautés ont forcement une centralité d'intermédiarité élevée. Ainsi, à chaque itération, on supprime le lien dont la centralité d'intermédiarité est maximale. Partant du même principe, l'approche de Radicchi [110] supprime le lien dont le coefficient de clustering est maximal. D'autres approches basées sur les techniques de propagation de labels exploitent la propriété de la densité des liens intra-communautés. Différents algorithmes exploitent cette propriété différemment [35, 36, 57, 107, 128, 111, 142].

3.2.3 Méthodes d'optimisation d'une fonction objective

Ces méthodes identifient les communautés en maximisant une fonction de qualité. La modularité proposée initialement dans [96] est la fonction de qualité la plus utilisée et la plus connue. Par hypothèse, les valeurs élevées de modularité Q indiquent de bonnes partitions. Ainsi, la partition correspondant à la valeur maximale de la modularité devrait être la meilleure. La maximisation de la modularité demeure un problème NP-difficile [20]. Aussi, des méthodes d'optimisation ont été proposées. Elles sont basées sur les techniques d'algorithmique génétique [84, 105], de recuit-simulé [58, 113], d'optimisation appelée *extremal optimization* [44], de coupes normalisées COPRA [57], d'optimisation spectrale [97], de propagation d'étiquette *(label propagation)* [111], ou d'optimisation statistique Oslom [80].

Fortunato [49] a montré que les algorithmes fondés sur l'optimisation de la modularité souffrent d'un problème de limite de résolution dans le sens qu'ils ne peuvent pas distinguer des communautés plus petites d'une certaine taille limite.

3.2.4 Méthodes alternatives [100]

Contrairement à la famille précédente des méthodes basées sur la théorie des graphes, cette catégorie de méthodes [100] inclut des méthodologies variées : approche basée sur la théorie de l'information, l'approche probabiliste et l'analyse formelle de concepts. Dans la première catégorie Rosvall et al proposent Infomap [115] qui identifie les communautés en fonction du flux d'informations dans les réseaux. Infomap comporte deux étapes principales : un algorithme de recherche glouton déterministe, puis une approche de recuit simulé pour affiner les résultats obtenus. Dans sa phase de recherche gloutonne,

l'algorithme commence par compresser la liste des nœuds visités par des marches aléatoires. Selon les auteurs, l'objectif est la détermination de la partition M minimisant la taille des marches aléatoires et la reconnaissance d'une communauté est définie par la rapidité de circulation de l'information entre les nœuds du réseau. La deuxième approche fait référence principalement aux modèles basés sur l'inférence statistique. Skvoretz et Faust [124] ainsi que Field et al. [48] ont exploré la capacité du modèle $p*$ (une famille de modèles structurels définis par Wasserman) à révéler les propriétés importantes du réseau. Brièvement, il s'agit de tester la robustesse d'une partition à priori d'un graphe et d'utiliser cette partition comme variable dans un modèle de régression.

L'autre méthode alternative consiste à utiliser l'analyse formelle de concepts pour identifier des communautés sémantiquement intéressantes. Freeman fut le premier à utiliser l'AFC pour la découverte de communautés dans les réseaux à un seul mode [52]et ceux à deux modes [53]. Dans le premier cas, il propose de déterminer d'abord les cliques maximales contenues dans le réseau pour ensuite construire le treillis de concepts d'un contexte formel qui décrit les nœuds et les cliques dans lesquels ils se trouvent. Ensuite, les cliques intermédiaires sont identifiées. Ces dernières sont des cliques maximales dont au moins deux chemins allant de leur position courante dans le treillis vers l'infimum ne sont pas de même longueur. Ensuite, on procède à l'élimination des arêtes partant de ces nœuds pour obtenir des groupes disjoints. Une extension de cette méthode a été proposée par Falzon [47] puis Selmane [121] qui a intégré la fonction de modularité de Newman [95].

Roth et *al.* [116] utilisent la méthode *(iceberg)* d'élagage du treillis afin de réduire le nombre de concepts et de se limiter aux concepts ayant une extension fréquente supérieure à un seuil donné. Ce type de filtrage peut éliminer des concepts de faible support mais présentant un réel intérêt. Kuznetsov *et al.* [74] proposent une méthode fondée sur la notion de stabilité des concepts. Elle requiert le calcul de la stabilité exacte de tous les concepts du treillis initial puis la sélection des concepts dont la stabilité dépasse un seuil fixé. l'inconvénient de cette méthode est que quelques objets appartenant uniquement à des concepts dont la stabilité est au-dessous du seuil vont être ignorés. Cette limite est inacceptable dans de nombreux cas où tous les individus doivent appartenir à au moins une communauté. Plantié et Crampes dans [38] ont proposé une approche qui sélectionne les concepts ayant autonomie élevée calculée à partir de l'indice de Jaccard et d'un seuil fixé. Ici, outre les seuils qui doivent être fixés, d'où une part arbitraire, le coefficient de Jaccard ne semble pas être bien adapté au contexte formel.

3.3 Détection des communautés dans les réseaux multicouches

Les approches existantes de détection de communautés dans les réseaux multicouches peuvent être regroupées en deux catégories [60, 66] :

(i) La première famille comporte des méthodes qui sont issues de l'adaptation des algorithmes standards de détection de communautés monocouches existants. Il s'agit de transformer le réseau multicouche en un réseau monocouche afin de l'analyser ensuite avec les algorithmes classiques.

(ii) La deuxième famille explore simultanément les couches pour la détection des communautés. Il s'agit de généraliser les approches de détection de communautés dans les graphes simples aux graphes multiplexes.

3.3.1 Les méthodes basées sur l'agrégation

Dans la littérature, deux types de stratégies d'agrégation sont distingués pour la première famille des méthodes :

Agrégation des couches

L'idée principale consiste à effectuer une agrégation des dimensions sur un réseau unidimensionnel. Il s'agit de remplacer l'ensemble des liens reliant chaque paire de nœuds par une seule connexion pondérée sur le graphe agrégé.
Dans notre revue de la littérature, nous avons pu distinguer deux schémas d'agrégation :

(a) Les approches dites naïves ne tenant pas compte de la pertinence des dimensions, c'est-à-dire du degré d'implication des dimensions, notamment l'agrégation binaire [12, 130], l'agrégation fréquentielle [132] et l'agrégation par similarité [108]. L'avantage évident de ces approches d'intégration est la simplicité de leur mise en œuvre. Toutefois, ces dernières sont sensibles aux dimensions non pertinentes et induisent une perte de l'information de la multiplicité des liens.

(b) Les approches d'agrégation à base d'apprentissage où la pertinence des dimensions est considérée dans la formation des communautés. Notons par exemple l'approche LBGA (*Locally Boosted Graph Aggregation*) [71], inspirée de l'apprentissage ensembliste par le 'boosting'. Cette approche combine une fonction de qualité avec un algorithme de regroupement pour sélectionner les meilleures arêtes.

Les approches à base d'apprentissage, quoiqu'elles génèrent des communautés plus intéressantes, sont sensibles à certains paramètres, notamment, le taux d'apprentissage.

Agrégation des partitions

Cette technique vise à appliquer un algorithme de détection de communautés classique pour chaque couche du réseau en isolant les autres. De la sorte, une combinaison des partitions retrouvées séparément serait appliquée à l'aide des techniques de *clustering* classiques. Deux types de stratégies d'intégration de partitions ont été proposés : l'intersection et le consensus. Dans la première [14, 120], il s'agit de repérer les zones de chevauchement entre les communautés unidimensionnelles tandis que la deuxième s'intéresse à dégager une partition de consensus à partir des partitions identifiées sur les dimensions du réseau.

Quelques approches ont adopté cette stratégie d'intégration. Citons CSPA (*Cluster-Based Similarity Partitioning Algorithm*), HGPA (*Hypergraph Partition Algorithm*) et MCLA (*Meta-Clustering Algorithm*) [133, 126]

Toutefois, ces dernières sont affectées significativement par les dimensions non pertinentes. De ce fait, les auteurs dans [5] introduisent MultiMOGA, une nouvelle approche consensuelle basée sur l'optimisation. Cette approche adopte une stratégie visant à évaluer les dimensions selon un classement préétabli.

3.3.2 Exploration simultanée des couches

Très peu de travaux ont abordé le problème d'exploration simultanée des couches. L'un des premiers travaux dans ce contexte est celui de Tang [132], dans lequel les auteurs ont proposé un modèle unifié pour identifier de nouvelles approches d'agrégation. Ainsi à la dernière étape, l'approche utilise l'algorithme de *clustering* K-means.

Par ailleurs, quelques approches ont généralisé les techniques conventionnelles de détection de communautés dans les réseaux simples aux réseaux multicouches. Vu le rôle prédominant de l'optimisation de la modularité dans le contexte des graphes simples, une nouvelle formule de la modularité multiplexe est ainsi proposée dans [94]. Plus tard, une autre version [24] inspirée de l'algorithme de Louvain à été proposée.

D'une manière similaire, les auteurs dans [10, 60] généralisent les mesures classiques, notamment le degré d'un nœud et le degré de centralité dans les graphes simples au cas des graphes multiplexes. Dans [10] les auteurs ont mis l'accent sur l'importance de l'im-

plication d'un nœud dans plus d'une couche en utilisant une fonction d'entropie. Ainsi, cette redéfinition des mesures permet d'appliquer des types d'algorithmes ayant montré des performances intéressantes sur des réseaux simples, notamment les algorithmes centrés graines et les algorithmes centrés diffusion.
Dans [60], l'idée fondamentale consiste à sélectionner les nœuds ayant une centralité supérieure à celle d'une majorité de voisins similaires. La similarité d'un voisin est mesurée en fonction de la fraction de voisins qu'il partage avec le nœud évalué. Par conséquent, les nœuds sont affectés aux nœuds centraux les plus proches selon une métrique de distance géodésique. Cet algorithme dépend des seuils de similarité et de voisinage, des valeurs non appropriées des seuils affectent la qualité des résultats.
Plus récemment, De Domenico et al dans [41] s'inspirent de l'algorithme unidimensionnel Infomap et proposent Multiplex Infomap. Une autre technique d'exploration simultanée de couches s'inspirant de l'algorithme unidimensionnel WalkTrap a été introduite dans [72].
Des outils d'algèbre linéaire ont été également généralisés pour présenter les réseaux multidimensionnels, particulièrement, la factorisation matricielle dite aussi la décomposition en valeurs singulières (SVD). Ces méthodes, basées sur l'intégration de caractéristiques, visent à combiner les caractéristiques structurelles séparément identifiées à partir des dimensions du réseau. Ces le cas des approches de [134, 42] qui peuvent être également considérées comme une généralisation des techniques de partitionnement spectral dans les réseaux unidimensionnels.
D'une manière similaire, les techniques de décomposition tensorielle [45, 104, 86] ont été adoptées pour la détection de communautés dans les réseaux multicouches.
En fait, un réseau multicouche peut être représenté à l'aide d'un tenseur où chaque coupe correspond à une matrice d'adjacence d'une dimension. Ainsi, les techniques développées [69] pour la décomposition de tenseurs peuvent être utilisées pour identifier la partition latente.
En terminant, il est important de noter que nombreuses approches rencontrent des difficultés lorsqu'elles sont appliquées à des réseaux ayant des dimensions non pertinentes. En outre, l'identification explicite des dimensions pertinentes demeure peu étudiée dans la littérature, mais les approches qui la proposent souffrent de leur dépendance d'un certain nombre de paramètres d'entrée. Donc, un réglage non approprié de ces paramètres influence grandement leur précision. En outre, la plupart des méthodes existantes exigent que le nombre de communautés soit fixé à l'avance.

Chapitre 4

Détection de communautés dans de données réseaux à un mode

4.1 Introduction

Dans ce chapitre, nous présentons une nouvelle stratégie appelée COIN qui exploite les mesures d'intérêt d'un concept *(COncept INterestingnes)* pour détecter des communautés dans les réseaux à un seul mode de données. Ainsi, COIN exploite des caractéristiques conceptuelles pertinentes héritées de l'analyse formelle de concepts afin de découvrir des structures locales importantes. Notre solution pour les réseaux à un mode est structurée en trois étapes, à savoir : (i) construire le contexte formel associé au réseau et générer les concepts formels, (ii) extraire les concepts formels identiques et utiliser les indices de stabilité pour détecter les communautés fortes et éliminer les ponts dits bruyants entre les communautés, et (iii) fusionner le reste des cliques pertinentes et adjacentes.

4.2 Construction du contexte formel et de ses concepts formels

Cette section donne un bref aperçu des notions utiles à la compréhension de l'algorithme proposé en utilisant un exemple illustratif extrait du réseau *LinkedIn* et contenant 15 membres d'un même laboratoire. Tel que montré dans la figure 4.1, le réseau est un graphe non orienté $\mathcal{U} = (\mathcal{G}, \mathcal{I})$, où $\mathcal{G}$ est un ensemble de 15 nœuds (membres) et $\mathcal{I}$ est

un ensemble d'arcs (g_i, g_j) reliant deux membres g_i et g_j dans $\mathcal{G}$ s'ils ont un lien dans *LinkedIn*.

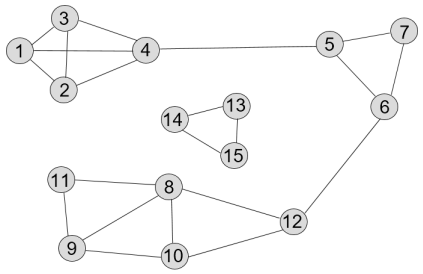


FIGURE 4.1: Un extrait du réseau LinkedIn

Dans COIN, la première tâche consiste à créer le contexte formel $\mathbb{K} = (\mathcal{G}, \mathcal{G}, \mathcal{I})$ associé au réseau social $\mathcal{U} = (\mathcal{G}, \mathcal{I})$ où $\mathcal{G}$ est un ensemble de nœuds et $\mathcal{I}$ est un ensemble de liens entre les nœuds. Cela se fait en calculant la matrice d'adjacence symétrique comme suit [59] :

$$\mathbb{K} := (\mathcal{G}, \mathcal{G}, \mathcal{I}) = \begin{cases} (g_i, g_j) = 1 & \text{si } \exists\ (g_i, g_j) \in \mathcal{I},\ i \neq j \\ (g_i, g_j) = 1 & \text{si } \ i = j \\ (g_i, g_j) = 0 & \text{sinon.} \end{cases} \tag{4.1}$$

Dans l'équation 4.1, nous affectons 0 à l'élément de $\mathbb{K}$ de la ligne i et de la colonne j si le nœud g_i n'est pas connecté au nœud g_j dans le graphe $\mathcal{U}$. Sinon, nous lui attribuons 1. Notez que la valeur 1 est attribuée aux éléments diagonaux. Le treillis de concepts associé à ce contexte $\mathbb{K}$ est donné par la figure 4.2.

4.3 Identification des concepts formels identiques

À cette étape, nous identifions les cliques fortes et les ponts non pertinents. Nous commençons par identifier les concepts dans lesquels l'intention est égale à l'extension. Ces concepts sont appelés *concepts identiques*. Nous utilisons $\tilde{\mathcal{C}}$ pour désigner l'ensemble de tous ces concepts identiques. Ainsi, nous avons montré et prouvé dans [63] qu'un concept identique présente une k-clique. Par exemple, le concept identique $\tilde{c} = (\{5, 6, 7\}, \{5, 6, 7\})$ du treillis de la figure 4.2 capture une 3-clique avec les nœuds $l = \{5, 6, 7\}$ du réseau de la figure 4.1.

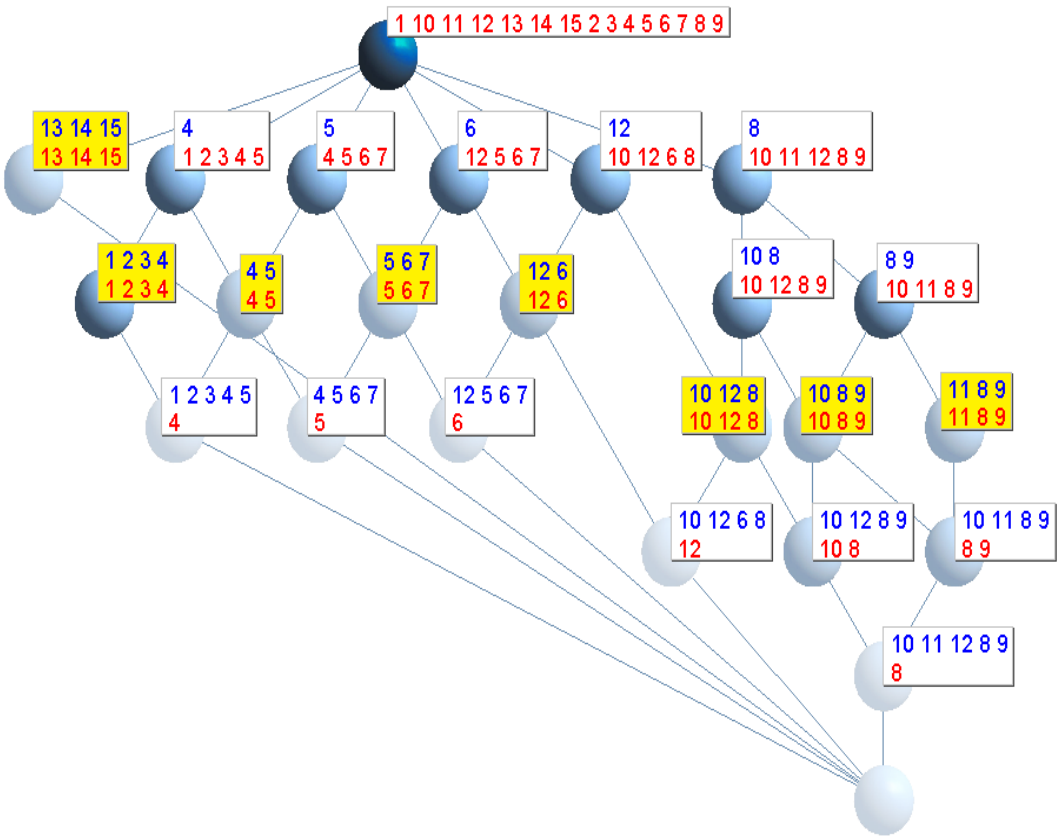


FIGURE 4.2: Le treillis de concepts $\mathcal{L}(\mathbb{K})$.

$\mathcal{G}$	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
3	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
4	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
5	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0
6	0	0	0	0	1	1	1	0	0	0	0	1	0	0	0
7	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0
9	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0
10	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0
11	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0
12	0	0	0	0	0	1	0	1	0	1	0	1	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
14	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
15	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1

Tableau 4.1: Le contexte formel $\mathbb{K}$ pour le réseau de la figure 4.1.

Definition 2 (Concept identique)**.** *Un concept formel $\tilde{c} = (A, B)$ d'extension A et d'intention B est appelé un concept identique si $A = B$, c'est-à-dire, son extension et son intention sont identiques.*

Proposition 1. *Soit un graphe $\mathcal{U}$ et son treillis de concepts correspondant $\mathcal{L}(\mathbb{K})$, un concept identique $c = (A, B) \in \mathcal{L}$ avec $A = B$ et $|A| = k > 2$, représente en fait une k-clique $l = \{g_i : g_i \in A\}$ dans $\mathcal{U}$.*

Démonstration. Comme un concept identique est un carré maximal rempli de 1 dans le contexte formel, il représente une matrice carrée de taille k comme sous-ensemble de la matrice d'adjacence et donc une k-clique. Supposons maintenant que $l = \{g_i\}_{i=1}^{k}$ est une k-clique de $\mathcal{U}$ avec $k > 2$. À partir de la définition d'une clique [2], pour n'importe quelle paire de nœuds g_i, g_j dans l, il existe un arc (g_i, g_j) dans $\mathcal{U}$ qui les relie. En utilisant l'équation 4.1, la matrice d'adjacence construite $k \times k$ pour le contexte $\mathbb{K}(l, l, \mathcal{I}_l)$ et exprimant la clique l, définit clairement une matrice dont toutes les cellules sont à 1. Une telle matrice coïncide avec le concept identique $\tilde{c} = (\{g_i\}_{i=1}^{k}, \{g_i\}_{i=1}^{k})$ dans lequel l'extension A et l'intention B contiennent uniquement les nœuds de l. Ceci implique qu'une k-clique dont l'ensemble de nœuds est $l = \{g_i : g_i \in A\}$ est équivalente à un concept identique $\tilde{c} = (A, B)$ tel que $A = B = \{g_i\}_{i=1}^{k}$. □

Par conséquent, à partir de la proposition 1, nous pouvons extraire les cliques du réseau $\mathcal{U}$ en identifiant leurs concepts identiques correspondants dans $\mathcal{L}$. Par exemple, les concepts identiques $\tilde{\mathcal{C}}$ apparaissent en gris dans la figure 4.2 et représentent toutes les cliques et les ponts comme le montre la figure 4.3-1.

Maintenant, le rôle de l'indice de stabilité entre en jeu. Aussi, nous mesurons la quantité de bruit existant dans une clique en calculant l'indice de stabilité du concept identique correspondant. Le bruit dans une clique mesure le niveau de cohésion de ses membres et leur degré de séparation par rapport aux objets externes à la clique. Maintenant, il s'agit de voir comment la cohésion d'une clique et sa séparation des autres cliques proviennent du bruit du concept identique correspondant. À un haut niveau, cela peut être illustré comme suit. D'abord, la stabilité mesure le bruit d'un concept identique en estimant comment son contenu dépend du retrait de chacun de ses membres. En fait, cette mesure de bruit quantifie la connectivité des membres au sein de la clique. Ainsi, une présence significative de bruit dans la clique signifie que plusieurs membres sont non cohésifs et doivent être écartés de la clique. En outre, la stabilité mesure the specificité des liens entre les objets et donc évalue la manière dont les membres de la clique sont influencés par les liens qui existent entre chaque membre et les autres noeuds du graphe. Ainsi, une clique est dite *pertinente* si elle contient une très petite quantité de bruit, c'est-à-dire si ses objets ont une grande cohésion et une faible séparation. Par conséquent, la valeur de la stabilité d'un concept identique donne approximativement la probabilité que sa clique correspondante soit une communauté potentielle.

Nous avons également prouvé dans [63] que si un concept identique a une valeur de stabilité très élevée (maximale), les nœuds de la clique correspondante sont hautement cohésifs et complètement séparables des autres nœuds du réseau, ce qui indique que cette clique est *maximale et isolée* du reste du réseau.

Proposition 2. *Soit un graphe $\mathcal{U}$ et son treillis de concepts correspondant $\mathcal{L}(\mathbb{K})$, un concept identique $\tilde{c} = (A, B) \in \mathcal{L}$ avec $A = B$ et $|A| = k > 2$, représente une k-clique maximale isolée $l = \{g_i : g_i \in A\}$ dans $\mathcal{U}$ où $\tilde{c}$ a la valeur la plus élevée de l'indice de stabilité :*

$$\sigma(\tilde{c}) = \frac{2^{|A|} - 1}{2^{|A|}} \tag{4.2}$$

Démonstration. La proposition est vérifiée si nous prouvons que : (i) $l = \{g_i\}_{i=1}^{k}$ est représenté par un concept identique, et (ii) ce concept identique a l'indice de stabilité le plus élevé.

(i) Supposons que $l = \{g_i\}_{i=1}^{k}$ est une clique isolée maximale de taille k dans $\mathcal{U}$. Puisque chaque k-clique isolée maximale a toutes les propriétés d'une k-clique, alors, à partir de la proposition 1, il est facile de démontrer que l a une matrice d'adjacence qui définit une matrice $k \times k$ remplie de 1 et donc l est équivalent à un concept identique $\tilde{c} = (A, B)$ avec $A = B = \{g_i\}_{i=1}^{k}$.
(ii) Etant donnée une clique isolée maximale, nous savons qu'aucun arc ne relie aucun objet g_i de l'intérieur de l à aucun autre objet $g_a \in \mathcal{G} \setminus l$ à l'extérieur de l. Ainsi, la matrice remplie de 1 et relative à l définit une sous-matrice de l'ensemble du contexte formel à un mode $\mathbb{K}$, où tous les éléments de $\mathbb{K}$ qui définissent les relations entre les objets $\{g_i\}_{i=1}^{k}$ à l'extérieur de l sont à zéro. À partir de la définition de la matrice qui définit l, nous avons :

$$\forall e \in \mathcal{P}(A), e \neq \emptyset \Rightarrow e' = A = B \tag{4.3}$$

Autrement dit, à l'exception de l'ensemble vide, tous les autres éléments de l'ensemble $\mathcal{P}(A)$ satisfont la condition de stabilité dans le numérateur de l'équation (2.6). Comme conséquence, ce numérateur est égal à la taille de $\mathcal{P}(A)$ après avoir exclu uniquement l'ensemble vide. Ainsi, nous avons :

$$\sigma(\tilde{c}) = \frac{|\mathcal{P}(A)| - 1}{|\mathcal{P}(A)|} = \frac{2^{|A|} - 1}{2^{|A|}} \tag{4.4}$$

Cela signifie que la stabilité du concept identique $\tilde{c} = (A, B)$ avec $A = B = \{g_i\}_{i=1}^{k}$ est égale à $\frac{2^{|A|}-1}{2^{|A|}}$, et par conséquent, augmente avec la taille de sa k-clique correspondante. □

Par exemple, le concept identique $c = (\{13, 14, 15\}, \{13, 14, 15\})$ dans la figure 4.2 capture une 3-clique isolée maximale $l = \{13, 14, 15\}$, montrée dans la figure 4.1, car sa valeur de stabilité est égale à $\sigma(\tilde{c}) = \frac{2^3-1}{2^3} = 0.875$. Ainsi cette clique présente une communauté autonome.

Par ailleurs, il nous a été permis de constater dans [63] qu'un concept identique de taille 2 a une faible valeur de la stabilité (cf. la proposition 3 ci-après), et pourrait alors identifier une 2-clique que nous dénommerons *bruyante* et donc un pont qui ne fait probablement pas partie d'une communauté potentielle.

Proposition 3. *Soit un graphe $\mathcal{U}$ et son treillis de concepts correspondant $\mathcal{L}(\mathbb{K})$, un concept identique $c = (A, B) \in \mathcal{L}$, avec $A = B = \{g_i, g_j\}$, $|A| = 2$, représente un pont*

(g_i, g_j) dans $\mathcal{U}$, et $\tilde{c}$ a l'indice de stabilité suivant :

$$\sigma(\tilde{c}) = \frac{1}{4} \tag{4.5}$$

Démonstration. La proposition est vraie une fois que nous prouvons que : (i) un pont est représenté par un concept identique avec une extension et une intention impliquant uniquement les deux objets du pont, et (ii) ce concept identique a une valeur de stabilité de $\frac{1}{4}$.
(1) Soit $b = (g_i, g_j)$ un pont entre deux composants $\mathcal{T}_i$ et $\mathcal{T}_j$ de $\mathcal{U}$ tels que $g_i \in \mathcal{T}_i$ et $g_j \in \mathcal{T}_j$. À partir de l'équation (4.1), la matrice d'adjacence 2×2 de b définit une matrice J_b remplie de 1. Puisque chaque objet du pont b appartient à un composant différent et donc a un lien avec un autre élément de son composant, alors sa matrice J_b est également une sous-matrice de tout le contexte formel à un mode $\mathbb{K}$ tel que nous avons les deux propriétés suivantes :

(i) $(g_i, g_p) = 0 \ \forall g_i \in \mathcal{T}_i$ et $g_p \in \mathcal{T}_j \setminus \{g_j\}$

(ii) $(g_j, g_p) = 0 \ \forall g_j \in \mathcal{T}_j$ et $g_p \in \mathcal{T}_i \setminus \{g_i\}$

La matrice d'adjacence J_b du pont peut être utilisée pour extraire, à partir de $\mathbb{K}$, un concept identique $\tilde{c} = (\{g_i, g_j\}, \{g_i, g_j\})$ où son intention et son extension contiennent les deux objets (nœuds) du pont.
(2) $\mathcal{P}(\tilde{c}) = \{\emptyset, \{g_i\}, \{g_j\}, \{g_i, g_j\}\}$ est l'ensemble des parties du concept identique $\tilde{c} = (\{g_i, g_j\}, \{g_i, g_j\})$. Sur la base de la définition et des propriétés de la matrice d'adjacence modifiées J_b du pont, seul le sous-ensemble $\{g_i, g_j\} \in \mathcal{P}(\{g_i, g_j\})$ satisfait la condition de stabilité au numérateur de l'équation (2.6), alors que les autres sous-ensembles dans $\{\emptyset, \{g_i\}, \{g_j\}\}$ ne les vérifient pas. Cela implique que la stabilité d'un concept identique de la forme : $\tilde{c} = (\{g_i, g_j\}, \{g_i, g_j\})$ est égale à $\frac{1}{2^{|\{g_i, g_j\}|}} = \frac{1}{2^{|2|}} = \frac{1}{4}$.

□

Comme les concepts identiques de taille 2 représentent un pont entre deux composants, ils sont systématiquement écartés. Par exemple, dans la figure 4.2, le concept identique $c = (\{4, 5\}, \{4, 5\})$ ayant une stabilité $\sigma(\tilde{c}) = 0.25$ représente le pont $(4, 5)$ dans le graphe de la figure 4.1. Plus précisement, nous notons que l'objet 4 appartient à la communauté $\mathcal{T}_i = \{1, 2, 3, 4\}$ alors que l'objet 5 est un élément de $\mathcal{T}_j = \{5, 6, 7\}$. Rappelons que pour le calcul de la stabilité, nous avons opté pour l'utilisation d'une

méthode approximative qui offre une approximation efficace et précise de la stabilité à l'aide de la technique d'échantillonnage à faible divergence [62].

4.4 Fusion de concepts

Une fois que tous les concepts identiques de taille $k \geq 2$ sont identifiés, ceux représentant des ponts sont éliminés et le processus de fusion de groupes est envisagé. En effet, on fusionne chaque paire de cliques voisines partageant au moins $a_{ij} - 1$ objets, où a_{ij} est le plus petit nombre d'objets dans l'extension de deux concepts $\tilde{c}_i$ et $\tilde{c}_j$. La procédure de notre algorithme COIN est présentée par l'algorithme 1.

4.5 Analyse de la complexité

La première étape de l'algorithme 1 (lignes 1 à 10) consiste à approximer la stabilité (σ) de chaque concept identique et distinguer les deux types de concepts identiques (décrits auparavant) en se basant sur l'estimation de la stabilité. Cette étape a une complexité temporelle de $O(|\tilde{\mathcal{C}}| \times \xi)$, où $\tilde{\mathcal{C}}$ est l'ensemble des concepts identiques et ξ est le temps nécessaire pour approximer l'indice de stabilité d'un concept identique.

La deuxième étape (lignes 11à 20) consiste à appliquer la fusion d'une manière itérative pour chaque paire de cliques voisines. Cette étape a une complexité temporelle de $O(|\tilde{\mathcal{C}}|^2)$. Ainsi la complexité totale de cet algorithme est de $O(|\tilde{\mathcal{C}}| \times \xi + |\tilde{\mathcal{C}}|^2)$.

La figure 4.3 illustre le fonctionnement de COIN sur un exemple illustratif : (1) extraction des concepts identiques qui représentent des cliques et des ponts, (2) utilisation de l'indice de stabilité approximatif des concepts identiques pour couper des ponts bruyants, par exemple (4,5), (6,12) et pour détecter des cliques isolées maximales, par exemple (13,14,15) et finalement (3) fusion des cliques pertinentes restantes, par exemple, (11,8,9), (10,8,9) et (12,10,8).

4.6 Expérimentation

COIN s'est montré efficace et précis et peut ainsi mieux déceler les communautés d'un réseau à un mode que l'algorithme de centralité d'intermédiarité des liens de Girvan et Newman [98], la méthode de Louvain [15], Walktrap [106], Infomap [115] ou FGM [33] .

Algorithme 1 ComDet : COIN

Input : Ensemble de tous les concepts identiques $\tilde{\mathcal{C}}$ du treillis de concepts $\mathcal{L}(\mathbb{K})$.
Output : Ensemble de communautés $\mathcal{D}$ du réseau.

$\mathcal{D} \leftarrow \mathcal{R} \leftarrow \emptyset$
// Etape 1: Extraire les cliques isolées et supprimer les ponts
pour chaque concept $\tilde{c}_i = (A_i, B_i) \in \tilde{\mathcal{C}}$ **faire**
    $\sigma(\tilde{c}_i) \leftarrow$ la stabilité approximative $\tilde{c}_i$ //Utilisant la méthode LDS
    si $\sigma(\tilde{c}_i) = \frac{2^{|A_i|}-1}{2^{|A_i|}}$ **alors**
// $\tilde{c}_i$est une clique maximale isolée
        $\mathcal{D} \leftarrow \mathcal{D} \cup \{\tilde{c}_i\}$ // $\tilde{c}_i$ est une communauté
        $\tilde{\mathcal{C}} \leftarrow \tilde{\mathcal{C}} \setminus \{\tilde{c}_i\}$
    sinon si $|A_i| = 2$ et $\sigma(\tilde{c}_i) = \frac{1}{4}$ **alors**
// Couper $\tilde{c}_i$ qui représente un pont bruyant non-trivial
        $\tilde{\mathcal{C}} \leftarrow \tilde{\mathcal{C}} \setminus \{\tilde{c}_i\}$
    fin si
fin pour
Etape 2 : Fusionner les cliques pertinentes adjacentes
pour $\tilde{c}_i = (A_i, B_i), \tilde{c}_j = (A_j, B_j) \in \tilde{\mathcal{C}}$ **faire**
    $a_{ij} \leftarrow \min(|A_i|, |A_j|)$
    si $|A_i \cap A_j| \geqslant a_{ij} - 1$ **alors**
        $\tilde{c}_{ij} \leftarrow \text{Fusionner}(\tilde{c}_i, \tilde{c}_j)$
        $\tilde{\mathcal{C}} \leftarrow \tilde{\mathcal{C}} \setminus \{\tilde{c}_i, \tilde{c}_j\}$
        $\tilde{\mathcal{C}} \leftarrow \tilde{\mathcal{C}} \cup \{\tilde{c}_{ij}\}$
    fin si
fin pour
$\mathcal{D} \leftarrow \mathcal{D} \cup \tilde{\mathcal{C}}$
retourner $\mathcal{D}$

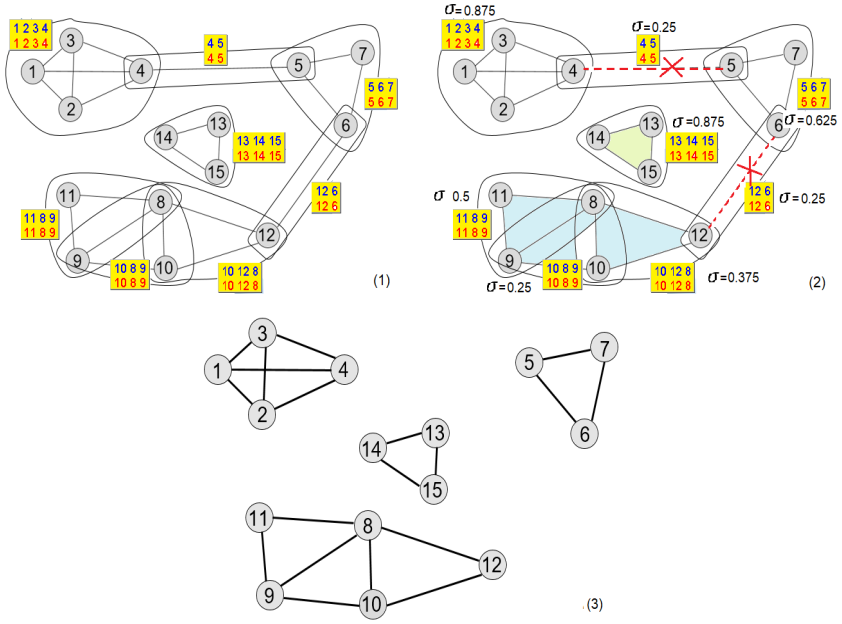


FIGURE 4.3: Le fonctionnement de COIN sur un exemple illustratif

Une brève description des réseaux sociaux testés est mentionnée dans le tableau 4.2 où $|\mathcal{G}|$ est le nombre de nœuds d'objets, $|\mathcal{I}|$ est le nombre d'arêtes et $|\hat{\mathcal{D}}|$ est le nombre de communautés de la réalité du terrain **(ground truth)**.

Pour mesurer la précision, nous avons calculé l'information mutuelle normalisée **(Normalized mutual information)** comme suit [39] :

$$NMI(\mathcal{D}, \hat{\mathcal{D}}) = \frac{-2\sum_{i=1}^{|\hat{\mathcal{D}}|}\sum_{j=1}^{|\mathcal{D}|} n_{ij}\log\left(\frac{n_{ij}n}{n_i n_j}\right)}{\sum_{i=1}^{|\hat{\mathcal{D}}|} n_i \log\left(\frac{n_i}{n}\right) + \sum_{j=1}^{|\mathcal{D}|} n_j \log\left(\frac{n_j}{n}\right)} \tag{4.6}$$

Où $n = |\mathcal{G}|$ est le nombre de nœuds, $\mathcal{N}$ est une matrice de confusion où les lignes correspondent à *la réalité de terrain* $\hat{\mathcal{D}}$ et les colonnes correspondent aux *communautés prédites* $\mathcal{D}$ trouvées par un algorithme de détection de communauté donné. Chaque élément $n_{ij} \in \mathcal{N}$ est le nombre de nœuds de la i-éme réalité de terrain *i-th* qui apparaissent dans la j-éme communauté prédite de *j-th*.
$n_i = \sum_j N_{ij}$ est la somme sur la ligne i de $\mathcal{N}$ et $n_j = \sum_i N_{ij}$ est la somme sur la colonne j.

Nos résultats sont montrés dans le tableau 4.2 où la valeur entre (.) montre le nombre de communautés prédites.

Méthodes	*réseaux sociaux*			
	Karaté	**Football**	**Dolphins**	**PolBooks**
Louvain	0.597 (4)	0.775 (10)	0.526 (5)	0.591 (4)
CPM	0.294 (3)	0.903 (13)	0.38 (4)	0.455 (6)
COIN	**0.837** (2)	**0.978** (12)	**1.00** (2)	**0.885** (3)
GN	0.789 (2)	0.807 (10)	0.890 (2)	0.543 (5)
WalkTrap	0.546 (5)	0.363 (10)	0.465 (8)	0.569 (4)
FGM	0.706 (3)	0.513 (6)	0.454 (4)	0.568 (4)
InfoMAP	0.712 (3)	0.890 (12)	0.557 (6)	0.567 (6)

Tableau 4.2: Le score NMI des algorithmes de détection de communautés dans les réseaux sociaux testés.

4.7 Discussion

En termes de précision, les résultats du tableau 4.2 illustrent le fait que COIN est le plus précis par rapport aux autres algorithmes, ayant le meilleur score NMI sur les quatre réseaux sociaux testés.

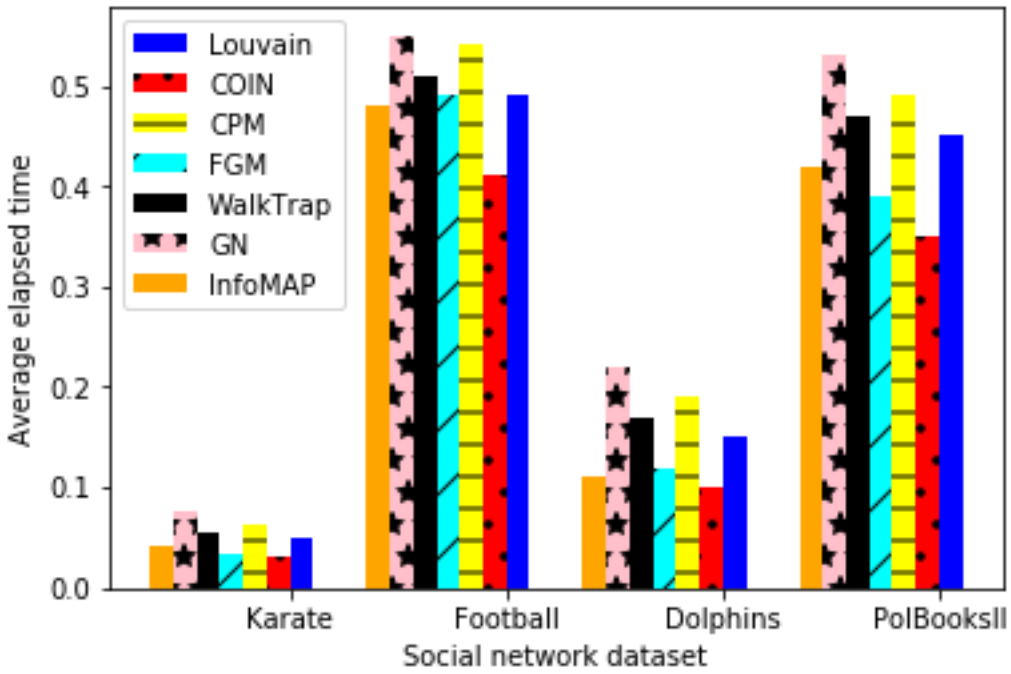


FIGURE 4.4: Temps moyen écoulé τ des algorithmes de détection de communauté sur les réseaux sociaux testés

InfoMAP et CPM se rapprochent de COIN sur l'ensemble de données *Football*, mais considérablement plus loin pour l'ensemble de données *Dolphins* et *PolBooks*. Pour l'ensemble de données *Karaté* et *PolBooks*, WalkTrap était légèrement moins précis que Louvain, mais plus précis que CPM. Ce dernier est plus performant que FGM et Louvain sur l'ensemble de données *Football*. Remarquablement, CPM a des résultats médiocres sur les jeux de données *Karaté* et *Dolphins* mais InfoMAP et Louvain surpassent clairement FGM sur le jeu de données *Football*. En termes de temps de calcul, les résultats de la figure 4.4 sont prometteuses et montrent que COIN domine tous les autres algorithmes testés sur tous les réseaux sociaux considérés. En pratique, cela est dû au fait que l'ensemble des concepts identiques $\tilde{\mathcal{C}}$ est très petit comparé aux ensembles de nœuds et de liens. Quelques visualisations des résultats des communautés prédites par COIN sont illustrées par les figures 4.5, 4.6, 4.7 et 4.8.

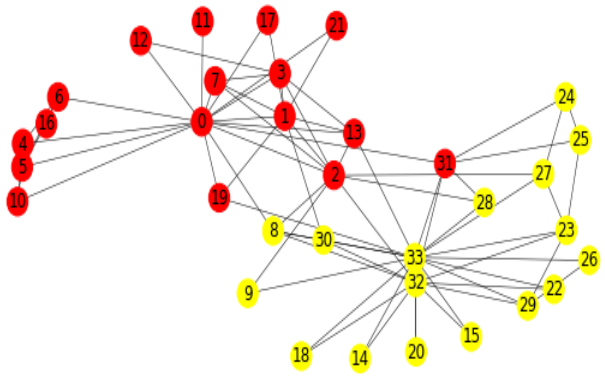


FIGURE 4.5: Les communautés prédites du réseau de *Karate* obtenues par l'algorithme COIN.

4.8 Conclusion

Dans ce chapitre, nous avons proposé COIN qui est une procédure en deux étapes exploitant l'analyse formelle de concepts et leur stabilité pour détecter efficacement les communautés dans les réseaux sociaux à un mode de données. Tout d'abord, tous les concepts identiques qui capturent des cliques et des ponts sont extraits du réseau. Ensuite, l'indice de stabilité est utilisé pour identifier les communautés isolées, autonomes, couper les ponts non pertinents entre les communautés et enfin, fusionner les cliques pertinentes et adjacentes pour obtenir les communautés finales. Nous nous concentrons

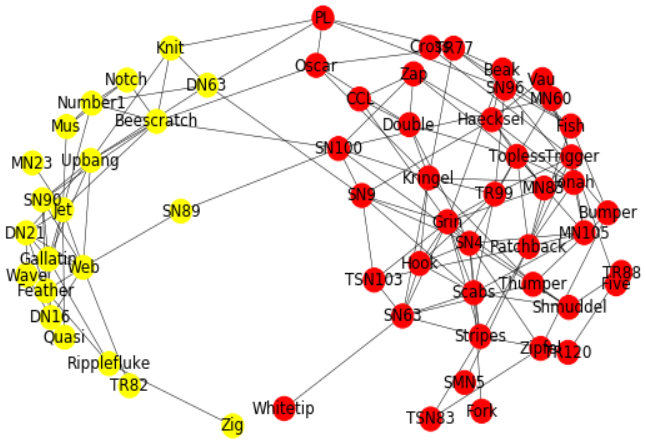


FIGURE 4.6: Les communautés prédites du réseau de *Dolphin* obtenues par l'algorithme COIN.

ici sur la détection des communautés disjointes, mais nous envisageons d'étendre la méthode à la détection des communautés chevauchantes comme nous le faisons dans les réseaux à deux modes dans le chapitre suivant.

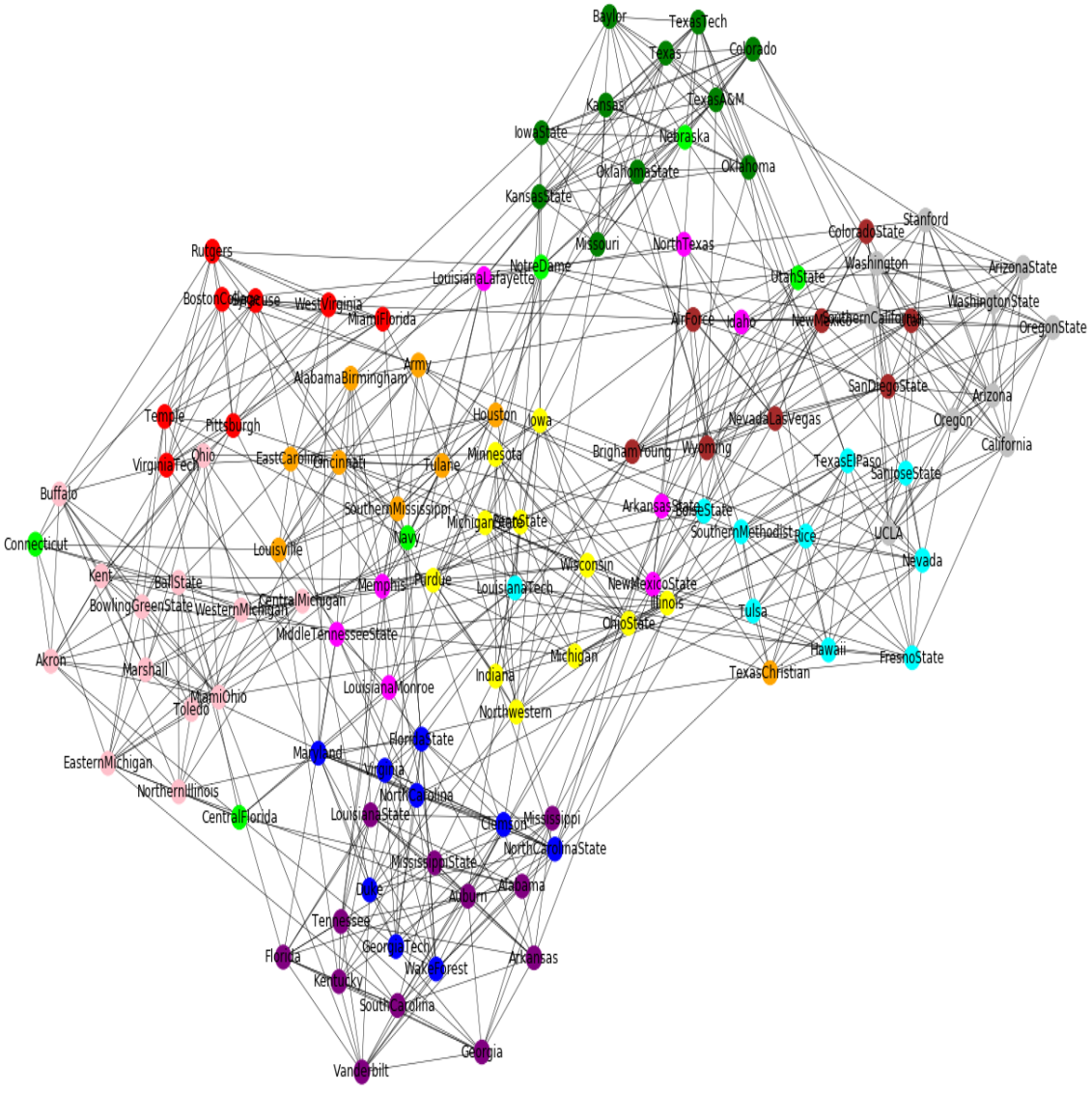


FIGURE 4.7: Les communautés prédites du réseau de *Football* obtenues par l'algorithme COIN.

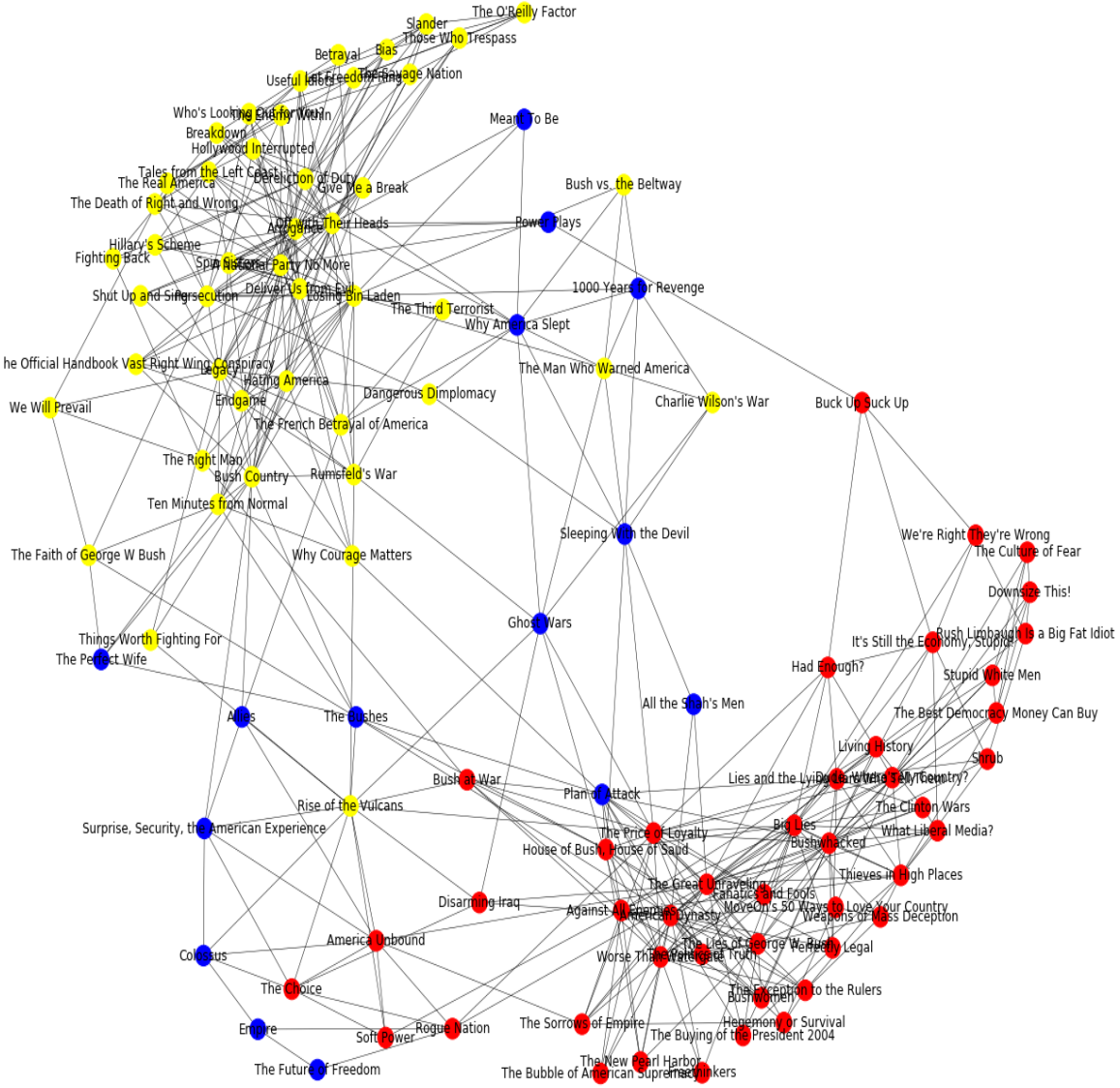


FIGURE 4.8: Les communautés prédites du réseau de *PolBooks* obtenues par l'algorithme COIN.

Chapitre 5

Détection de communautés dans les réseaux à deux modes

5.1 Introduction

Comme présenté dans l'introduction, notre approche CoDeBi de détection de communautés dans les réseaux (non orientés et non pondérés) à deux modes comporte trois étapes, à savoir : (i) calcul de l'ensemble des concepts à partir du contexte formel décrivant les données du réseau, (ii) calcul de l'autonomie des concepts en tant que moyenne harmonique des indices de stabilité et de séparation, puis sélection des concepts ayant les scores d'autonomie les plus élevés et couvrant l'ensemble des nœuds du réseau, et (iii) raffinement des concepts sélectionnés en utilisant l'indice Silhouette pour obtenir les communautés finales imbriquées et chevauchantes.

5.2 Génération des concepts formels

Au début, nous construisons le contexte formel $\mathbb{K}$ (voir le tableau 2.2) du réseau à deux modes en calculant la matrice d'incidence de la manière suivante :

$$\mathbb{K} = (\mathcal{G}, \mathcal{M}, \mathcal{I}) = \begin{cases} (g_i, m_j) = 1 & \text{si } g_i \in \mathcal{G}, m_j \in \mathcal{M}, \exists\ (g_i, m_j) \in \mathcal{I}, \\ (g_i, m_j) = 0 & \text{sinon.} \end{cases} \tag{5.1}$$

Ensuite, l'ensemble des concepts est calculé. En prenant l'exemple de femmes du sud, l'un des concepts générés est $(\{EVELYN, LAURA\}, \{1, 2, 3, 5, 6, 8\})$. Nous notons que la plupart des applications en analyse formelle de concepts tiennent compte non seulement des concepts, mais également de la relation d'ordre qui existe entre eux. Dans notre cas, nous utilisons uniquement les concepts sans avoir besoin de construire le treillis de concepts dans sa totalité. Cela réduit évidemment les coûts. Pour générer les concepts d'une manière efficace, nous avons exploité l'algorithme *In-Close* [6].

5.3 Sélection de concepts cohésifs et séparables

Une fois que l'ensemble des concepts $\mathcal{C}$ est déterminé, il s'agit de sélectionner ceux qui sont les plus pertinents et susceptibles de former une communauté. Nous considérons un concept $c = (A, B) \in \mathcal{C}$ comme une communauté de base (*core community*) s'il respecte les deux critères qualitatifs de définition de communauté et d'une manière générale de groupe (*cluster*), à savoir, la cohésion interne et la séparation avec le reste du réseau. Maintenant, étant donné le concept c, comment pouvons-nous mesurer la cohésion et la séparation des objets associés au concept ?
Ici, les indices de stabilité et de séparation interviennent pour quantifier la force des liens entre les objets à l'intérieur de c et la faiblesse des liens entre les objets de l'extension de c et tous les autres objets dans les autres concepts. En fait, des indices de stabilité et de séparation élevés d'un concept indiquent que les objets à l'intérieur de ce concept sont très cohésifs, c'est-à-dire qu'ils ont des liens forts entre eux et sont très séparables des objets contenus dans d'autres groupes, c'est-à-dire qu'ils ont des liens faibles avec tous les autres objets qui existent en dehors de leur concept. Pratiquement, l'indice de stabilité d'un concept indique la probabilité de préserver son intention après avoir écarté un sous-ensemble arbitraire d'objets du contexte. Par exemple, la stabilité du concept $c = (\{EVELYN, LAURA\}, \{1, 2, 3, 5, 6, 8\})$ est égale à $\frac{1}{4} = 0.25$. La stabiltié exacte d'un concept est calculée selon l'équation 2.6 avec une complexité temporelle de $O(|\mathcal{G}|^2 \cdot |\mathcal{M}| \cdot |\mathcal{C}|^2|)$ [116, 144] où $|\mathcal{G}|$, $|\mathcal{M}|$, et $|\mathcal{C}|$ représentent respectivement le nombre d'objets, d'attributs et de concepts. C'est la raison pour laquelle nous avons opté pour une méthode efficace et robuste de calcul approximatif de la stabilité qui se base sur la méthode d'échantillonnage à faible divergence (*low-discrepancy sampling*) [62].

Tel qu'indiqué auparavant, nous utilisons l'indice de séparation selon l'équation 2.7 pour estimer la spécificité de la relation objet-attribut d'un concept par rapport à tout

le contexte formel, contrairement à la stabilité qui s'intéresse au concept et à ses sous-ensembles uniquement. Pratiquement, la séparation est définie comme une partie de la zone couverte (non nulle) par un concept formel parmi toutes les rangées (objets) et les colonnes (attributs) du concept formel. Pour le même concept mentionné dans l'exemple ci-dessus, la séparation de $c = (\{EVELYN, LAURA\}, \{1, 2, 3, 5, 6, 8\})$ est égale à $\frac{12}{45} = 0.26$. Nous calculons l'autonomie de chaque concept comme la moyenne harmonique de sa stabilité et sa séparation. Pour ce même concept, l'autonomie est égale à $2 * \frac{0.25 \times 0.26}{0.25 + 0.26} = 0.258$. Cette faible valeur d'autonomie va probablement conduire à écarter le concept de l'ensemble des communautés noyau. La stabilité, la séparation et l'autonomie de tous les concepts du contexte formel $\mathbb{K}$ du tableau 2.2 sont données par le tableau 5.1.

Une fois que l'autonomie de tous les concepts est calculée, les concepts sont classés dans un ordre décroissant des valeurs de leur autonomie. L'ensemble des communautés de base couvre les concepts les mieux classés dont l'union de leur extension couvre la collection d'objets du contexte. Tout concept dont l'extension contient uniquement des objets déjà traités n'est pas pris en compte. Ainsi, l'ensemble des concepts sélectionnés définit les communautés de base et son identification ne fait nullement appel à un seuil. Dans notre exemple, les communautés principales sont présentées dans le tableau 5.2.

5.4 Raffinement des communautés de base

Cette étape vise à identifier les communautés finales qui se chevauchent en utilisant l'indice Silhouette $\mathcal{S}(o)$ décrit par l'équation 2.8 pour affiner les communautés de base. Cela nous permet de vérifier que chaque objet se trouve dans les communautés les plus appropriées. Dans le cas contraire, des déplacements d'objets d'un groupe à un autre sont prévus. Pour calculer cet indice, nous n'avons besoin que des communautés (groupes) de base et des paires des distances entre les objets.
$\mathcal{S}(o_i)$ mesure la distance entre l'objet o_i, le centroïde du groupe auquel il a été attribué et le centroïde le plus proche d'un autre groupe. La distance de Jaccard est utlisée pour mesurer la dissimilarité entre les objets. Ainsi, si un objet se trouve à une distance plus proche d'une autre communauté C_j que de la sienne, il faut le réaffecter à C_j et mettre à jour les deux communautés. Sinon, il reste dans sa communauté d'origine mais inclus également dans C_j si l'indice Silhouette vaut 0. Ce processus est appliqué de manière séquentielle pour tous les objets jusqu'à ce qu'aucune amélioration supplémentaire ne

$\mathcal{C}$	$\mathcal{A}$	$\mathcal{B}$	$\sigma(c)$	$\alpha(c)$	$\zeta(c)$
Concept 1	{EVE}	{E1,E2,E3,E4,E5,E6,E8,E9}	0.5	0.14	0.22
Concept 2	{LAU}	{E1,E2,E3,E5,E6,E7,E8}	0.5	0.13	0.21
Concept 3	{THE}	{E2,E3,E4,E5,E6,E7,E8,E9}	0.5	0.12	0.2
Concept 4	{BRE}	{E1,E3,E4,E5,E6,E7,E8}	0.5	0.13	0.21
Concept 5	{SYL}	{E7,E8,E9,E10,E12,E13,E14}	0.5	0.13	0.21
Concept 6	{NOR}	{E6,E7,E9,E10,E11,E12,E13,E14}	0.5	0.16	0.24
Concept 7	{HEL}	{E7,E8,E10,E11,E12}	0.5	0.13	0.2
Concept 8	{EVE,LAU}	{E1,E2,E3,E5,E6,E8}	0.25	0.27	0.26
Concept 9	{EVE,THE}	{E2,E3,E4,E5,E6,E8,E9}	0.25	0.25	0.25
Concept 10	{EVE,BRE}	{E1,E3,E4,E5,E6,E8}	0.25	0.26	0.26
Concept 11	{LAU,THE}	{E2,E3,E5,E6,E7,E8}	0.25	0.23	0.24
Concept 12	{LAU,BRE}	{E1,E3,E5,E6,E7,E8}	0.25	0.24	0.24
Concept 13	{THE,BRE}	{E3,E4,E5,E6,E7,E8}	0.25	0.23	0.24
Concept 14	{THE,RUT}	{E5,E7,E8,E9}	0.5	0.17	0.25
Concept 15	{THE,NOR}	{E6,E7,E9}	0.25	0.15	0.19
Concept 16	{VER,SYL}	{E7,E8,E9,E12}	0.5	0.18	0.26
Concept 17	{KAT,SYL}	{E8,E9,E10,E12,E13,E14}	0.5	0.27	0.35
Concept 18	{SYL,NOR}	{E7,E9,E10,E12,E13,E14}	0.25	0.29	0.27
Concept 19	{SYL,HEL}	{E7,E8,E10,E12}	0.25	0.21	0.23
Concept 20	{NOR,HEL}	{E7,E10,E11,E12}	0.25	0.27	0.26
Concept 21	{EVE,LAU,THE}	{E2,E3,E5,E6,E8}	0.12	0.32	0.18
Concept 22	{EVE,LAU,BRE}	{E1,E3,E5,E6,E8}	0.12	0.33	0.18
Concept 23	{EVE,THE,BRE}	{E3,E4,E5,E6,E8}	0.12	0.31	0.18
Concept 24	{EVE,THE,PEA}	{E6,E8,E9}	0.5	0.2	0.29
Concept 25	{EVE,THE,RUT}	{E5,E8,E9}	0.25	0.2	0.22
Concept 26	{LAU,THE,BRE}	{E3,E5,E6,E7,E8}	0.12	0.28	0.17
Concept 27	{THE,BRE,CHA}	{E3,E4,E5,E7}	0.5	0.34	0.41
Concept 28	{VER,SYL,NOR}	{E7,E9,E12}	0.25	0.24	0.24
Concept 29	{VER,SYL,HEL}	{E7,E8,E12}	0.25	0.24	0.25
Concept 30	{MYR,KAT,SYL}	{E8,E9,E10,E12}	0.5	0.29	0.36
Concept 31	{KAT,SYL,NOR}	{E9,E10,E12,E13,E14}	0.25	0.43	0.32
Concept 32	{SYL,NOR,HEL}	{E7,E10,E12}	0.12	0.28	0.17
Concept 33	{NOR,OLI,FLO}	{E9,E11}	0.75	0.27	0.4
Concept 34	{EVE,THE,BRE,CHA}	{E3,E4,E5}	0.25	0.36	0.3
Concept 35	{EVE,THE,PEA,NOR}	{E6,E9}	0.38	0.21	0.27
Concept 36	{LAU,THE,BRE,CHA}	{E3,E5,E7}	0.25	0.32	0.28
Concept 37	{LAU,THE,BRE,ELE}	{E5,E6,E7,E8}	0.5	0.32	0.39
Concept 38	{THE,RUT,VER,SYL}	{E7,E8,E9}	0.56	0.26	0.35
Concept 39	{VER,MYR,KAT,SYL}	{E8,E9,E12}	0.38	0.29	0.33
Concept 40	{VER,SYL,NOR,HEL}	{E7,E12}	0.12	0.25	0.17
Concept 41	{MYR,KAT,SYL,NOR}	{E9,E10,E12}	0.25	0.33	0.29
Concept 42	{MYR,KAT,SYL,HEL}	{E8,E10,E12}	0.38	0.34	0.36
Concept 43	{NOR,HEL,OLI,FLO}	{E11}	0.38	0.24	0.29
Concept 44	{EVE,LAU,THE,BRE,FRA}	{E3,E5,E6,E8}	0.53	0.4	0.46
Concept 45	{LAU,THE,BRE,ELE,RUT}	{E5,E7,E8}	0.44	0.32	0.37
Concept 46	{LAU,THE,BRE,ELE,NOR}	{E6,E7}	0.44	0.24	0.31
Concept 47	{THE,RUT,VER,SYL,NOR}	{E7,E9}	0.34	0.23	0.28
Concept 48	{VER,MYR,KAT,SYL,NOR}	{E9,E12}	0.19	0.27	0.22
Concept 49	{VER,MYR,KAT,SYL,HEL}	{E8,E12}	0.19	0.28	0.22
Concept 50	{MYR,KAT,SYL,NOR,HEL}	{E10,E12}	0.19	0.32	0.24
Concept 51	{EVE,LAU,THE,BRE,CHA,FRA}	{E3,E5}	0.31	0.3	0.31
Concept 52	{EVE,LAU,THE,BRE,FRA,ELE}	{E5,E6,E8}	0.38	0.36	0.37
Concept 53	{LAU,THE,BRE,CHA,ELE,RUT}	{E5,E7}	0.38	0.3	0.33
Concept 54	{VER,MYR,KAT,SYL,NOR,HEL}	{E12}	0.09	0.18	0.12
Concept 55	{EVE,LAU,THE,BRE,FRA,ELE,PEA}	{E6,E8}	0.47	0.29	0.36
Concept 56	{EVE,LAU,THE,BRE,FRA,ELE,RUT}	{E5,E8}	0.36	0.28	0.31
Concept 57	{EVE,LAU,THE,BRE,CHA,FRA,ELE,RUT}	{E5}	0.28	0.17	0.21
Concept 58	{EVE,LAU,THE,BRE,FRA,ELE,PEA,NOR}	{E6}	0.41	0.16	0.23
Concept 59	{LAU,THE,BRE,ELE,RUT,VER,SYL,HEL}	{E7,E8}	0.81	0.3	0.43
Concept 60	{EVE,THE,PEA,RUT,VER,MYR,KAT,SYL,DOR}	{E8,E9}	0.93	0.33	0.49
Concept 61	{LAU,THE,BRE,CHA,ELE,RUT,VER,SYL,NOR,HEL}	{E7}	0.69	0.17	0.28
Concept 62	{EVE,THE,PEA,RUT,VER,MYR,KAT,SYL,NOR,DOR,OLI,FLO}	{E9}	0.87	0.21	0.33
Concept 63	{EVE,LAU,THE,BRE,FRA,ELE,PEA,RUT,VER,MYR,KAT,SYL,HEL,DOR}	{E8}	0.94	0.19	0.32

Tableau 5.1: Concepts produits et leur stabilité, séparation et autonomie

C_1	{EVELYN,THERESA,PEARL,RUTH,VERNE,MYRNA,KATHERINE,SYLVIA,DOROTHY }
C_2	{EVELYN,LAURA,THERESA,BRENDA,FRANCES}
C_3	{LAURA,THERESA, BRENDA, ELEANOR, RUTH, VERNE, SYLVIA,HELEN}
C_4	{THERESA,BRENDA,CHARLOTTE}
C_5	{NORA,OLIVIA,FLORA}

Tableau 5.2: Les communautés de base

puisse être obtenue. Dans notre exemple, les communautés finales sont présentées dans le tableau 5.4. L'utilisation de l'indice Silhouette (cf. chapitre 2) a permis de déplacer, en premier lieu, la femme EVELYN de la communauté 1 à la communauté 4, et qui est la plus appropriée selon la réalité de terrain [53]. L'indice silhouette pour EVELYN est calculé comme suit : $\mathcal{S}(EVELYN) = \frac{0.74-0.85}{max\{0.85,0.74)\}}$ où :
$a(EVELYN) = 0.85$ présente la distance qui sépare EVELYN de sa communauté actuelle C_1. Quant à $b(EVELYN) = 0.74$, cela présente la distance qui sépare EVELYN de la communauté la plus proche où elle n'existe pas, notament la communauté C_4.
Comme $\mathcal{S}(EVELYN)$ présente une valeur négative, EVELYN est déplacée de sa communauté actuelle vers la communauté C_4. Le tableau 5.3 montre les paires des distances qui séparent EVELYN des autres membres de sa propre communauté et des autres communautés auxquelles elle n'appartient pas. Un traitement similaire s'applique aux autres femmes.

La figure 5.1 présente une visualisation des communautés finales produites pour le réseau des femmes du sud américain.

5.5 Algorithme

L'algorithme 2 donne le pseudo-code de notre procédure. En premier lieu, il génère l'ensemble des concepts $\mathcal{C}$ de $\mathbb{K}$. Ensuite, il calcule l'autonomie $\zeta(c)$ de chaque concept c dans l'ensemble $\mathcal{C}$ (lignes 1 à 4) pour trier ensuite les concepts dans un ordre décroissant de l'autonomie (ligne 6). Puis les communautés de base sont placées dans Λ en sélectionnant les concepts ayant les valeurs $\zeta(c)$ les plus élevées jusqu'à ce que l'ensemble des $\mathcal{C}$ sélectionnés couvre tous les objets (et par conséquent tous leurs attributs) de $\mathbb{K}$ (lignes 7-14). À une étape ultérieure, l'algorithme raffine l'ensemble Λ en calculant $\mathcal{S}(o)$ de chaque objet $o \in A$ dans chacune des communautés identifiées. Si $\mathcal{S}(o)$ est inférieur à 0, o n'est pas dans la bonne communauté c et est ensuite déplacé vers la communauté

EVELYN	C_1	C_3	C_4	C_5
LAURA		0.73		
THERESA	0.75	0.75	0.75	
BRENDA		0.73	0.73	
CHARLOTTE			0.75	
ELEANOR		0.83		
PEARL	0.84			
RUTH	0.85	0.85		
VERNE	0.88	0.88		
MYRNA	0.88			
KATHERINE	0.88			
SYLVIA	0.89	0.89		
NORA				0.87
HELEN		0.92		
DOROTHY	0.87			
OLIVIA				0.9
FLORA				0.9
distance moyenne	a=0.85	0.82	b=0.74	0.89

Tableau 5.3: Distance entre EVELYN et les autres femmes

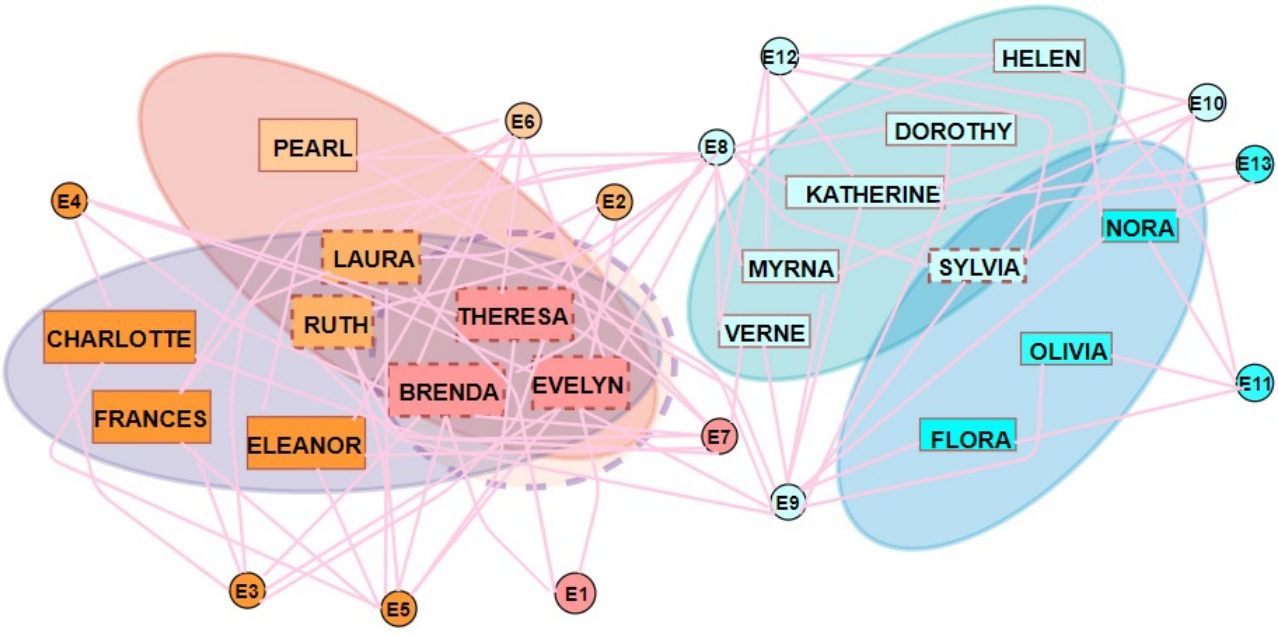


FIGURE 5.1: Visualisation des communautés finales pour le réseau des femmes du sud

Algorithme 2 *CoDeBi*

Entrée : Contexte formel $\mathbb{K} = (\mathcal{G}, \mathcal{M}, \mathcal{I})$
Sortie : Communautés chevauchantes et imbriquées ($\tilde{\mathcal{C}}$)

```
   //Générer l'ensemble des concepts
 1: C ← générer les concepts de K
 2: Γ ← D ← ∅
 3: pour chaque concept c = (A, B) ∈ C faire
 4:     t_c ← Calculer l'autonomie ζ(c)
 5:     Γ ← Γ ∪ {(c, t_c)}
 6: fin pour
   // Trier les concepts dans Γ par ordre décroissant de t_c
 7: D ← Tri(Γ)
   // Sélectionner les communautés de base
 8: Λ ← O ← ∅
 9: tant que O ≠ G faire
10:     c(A, B) ← D.pop()
11:     si (A ⊄ O) alors
12:         Λ ← ∪{c}  // Considérer c comme une communauté de base
13:         O ← O ∪ {A}
14:     fin si
15: fin tant que
   // Affiner les communautés dans Λ pour obtenir les communautés
   finales C̃
16: pour chaque communauté c = (A, B) ∈ Λ faire
17:     pour o ∈ A faire
18:         s_o ← S(o)
19:         si s_o < 0 alors
   //Déplacer o de c vers la communauté la plus proche
20:             A ← A \ {o}
21:             c_1(A_1, B_1) ← Trouver la communauté de base la plus proche de o
22:         sinon si s_o = 0 alors
   //Ajouter o de c à la communauté la plus proche
23:             c_1(A_1, B_1) ← Trouver la communauté de base la plus proche de o
24:             A_1 ← A_1 ∪ {o}
25:         fin si
26:     fin pour
27: fin pour
28: retourner(Λ)
```

$\tilde{C}1$	({ VERNE, MYRNA, KATHERINE, SYLVIA, DOROTHY, *HELEN*})
$\tilde{C}2$	({ EVELYN, LAURA, THERESA, BRENDA, *PEARL*,*RUTH* })
$\tilde{C}3$	({THERESA,BRENDA,*EVELYN*})
$\tilde{C}4$	({THERESA, BRENDA, CHARLOTTE, *EVELYN*, *RUTH*, *LAURA*, *FRANCESELEANOR*})
$\tilde{C}5$	({NORA, OLIVIA, FLORA, *SYLVIA*})

Tableau 5.4: Communautés finales

de Λ la plus proche. Si la valeur $\mathcal{S}(o)$ est égale à 0, o va apparaître également dans une autre communauté de Λ la plus proche. Sinon, o est maintenu dans sa communauté c. Après avoir affiné toutes les communautés de base Λ, l'algorithme les communautés finales (ligne 28).

5.6 Complexité algorithmique

Analyse de complexité

Le calcul des concepts formels nécessite $O(|\mathcal{G}|^2 \times |\mathcal{M}| \times |\mathcal{C}|$, où $|\mathcal{G}|$, $|\mathcal{M}|$, et $|\mathcal{C}|$ représentent respectivement le nombre d'objets, d'attributs et de concepts. Pour calculer l'autonomie du concept $c = (A, B)$, on a besoin de calculer l'index de stabilité et la séparation. Une approximation de la première mesure est possible à l'aide de la méthode d'échantillonnage à faible divergence (*low-discrepancy sampling*) [62] avec une complexité de $O(|\mathcal{S}|)$, où $|\mathcal{S}|$ est le nombre d'échantillons extraits de $\mathcal{P}(A)$. Puisque la complexité de calcul de la séparation d'un concept est $O(|\mathcal{G}| \times |\mathcal{M}|)$, le calcul de l'autonomie pour tous les concepts est $O(|\mathcal{C}| \times (|\mathcal{S}| + (|\mathcal{G}| \times |\mathcal{M}|))$. Le tri des concepts est $O(|\mathcal{C}| \times Log(|\mathcal{C}|))$. Finalement, puisque l'analyse Silhouette nécessite la comparaison de chaque membre d'un groupe avec les objets des autres groupes, sa complexité est $O(|\tilde{\mathcal{C}}| \times |\mathcal{M}| \times |\mathcal{G}|^2)$, où $\tilde{\mathcal{C}}$ représente l'ensemble des communautés générées. La complexité globale est donc dominée par celle des étapes 1 et 3 de l'algorithme.

5.7 Expérimentation

Pour valider la stratégie *CoDeBi*, nous avons mené deux séries de tests afin d'analyser sa performance et sa précision par rapport à quatre autres algorithmes de détection de communautés utilisant des réseaux du monde réel avec une structure de communautés

imbriquées. Bien que la première série de tests concerne de petites collections de données, la seconde couvre de plus grands ensembles de données et vise à tester plus de variantes de *CoDeBi*.

5.7.1 Première série de tests

L'algorithme a été mis en œuvre en Python et les expériences ont été exécutées sur un ordinateur personnel fonctionnant sous Windows et doté d'un processeur Intel Core i7 à 3.4 GHz et d'une mémoire de 16 Go. Les réseaux utilisés pour l'expérimentation sont décrits dans le tableau 5.5 et les deux premiers possèdent une réalité de terrain *(ground truth)*. Nous comparons ensuite la précision et les performances de notre approche avec les algorithmes de détection suivants : (i) Oslom [80], (ii) Crampes et Plantié [38], (iii) Jay [64], et (iv) Bitector [43]. De plus, nous considérons trois métriques [28] pour évaluer la précision des algorithmes, à savoir l'indice Omega [34] et l'information mutuelle normalisée chevauchantes (ONMI) [28] pour les réseaux possédant une réalité de terrain et la modularité des liens d'appartenance *(link-belonging modularity)* [79] pour le reste.

Nom	$\mathcal{G}$	$\mathcal{M}$	$\mathcal{I}$	$\lvert\mathcal{C}\rvert$	**Description de la nature du groupe**
Femmes du sud (S-W)	18	14	89	63	réseau de participation de femmes du sud américain à des événements
Zoo	17	101	746	377	réseau représentant différents types d'animaux et leurs caractéristiques dans un zoo
Customer-Product (C-P)	1143	865	2008	1545	réseau des clients qui ont commandés des produits en ligne sur gazella.com
Sénateurs x Comités (Senat)	59	189	890	548	réseau des relations entre les sénateurs de la 124ème législature des USA et le comité législatif
DBpediaLanguages (PL)	169	316	9022	5681	réseau sémantique des langues officielles parlées par des personnes vivant dans des pays différents
Star alliance (Star)	58	28	579	226	Compagnies aériennes et leurs destinations en l'an 2000

Tableau 5.5: Les réseaux utilisés pour l'expérimentation

Résultats

Nous pouvons observer sur la figure 5.2 que notre algorithme est testé sous deux variantes : l'une dans laquelle la troisième étape (raffinement) est incluse (c'est-à-dire les trois étapes) et l'autre dans laquelle la troisième étape est exclue.

On peut constater que la précision est satisfaisante par rapport aux méthodes testées pour les deux types de jeux de données (avec ou sans réalité de terrain). Notre algorithme est suivi par Bitector pour le premier type de jeux de données (les femmes du Sud et Zoo) et par Oslom pour le second type. En effet, cette dernière approche

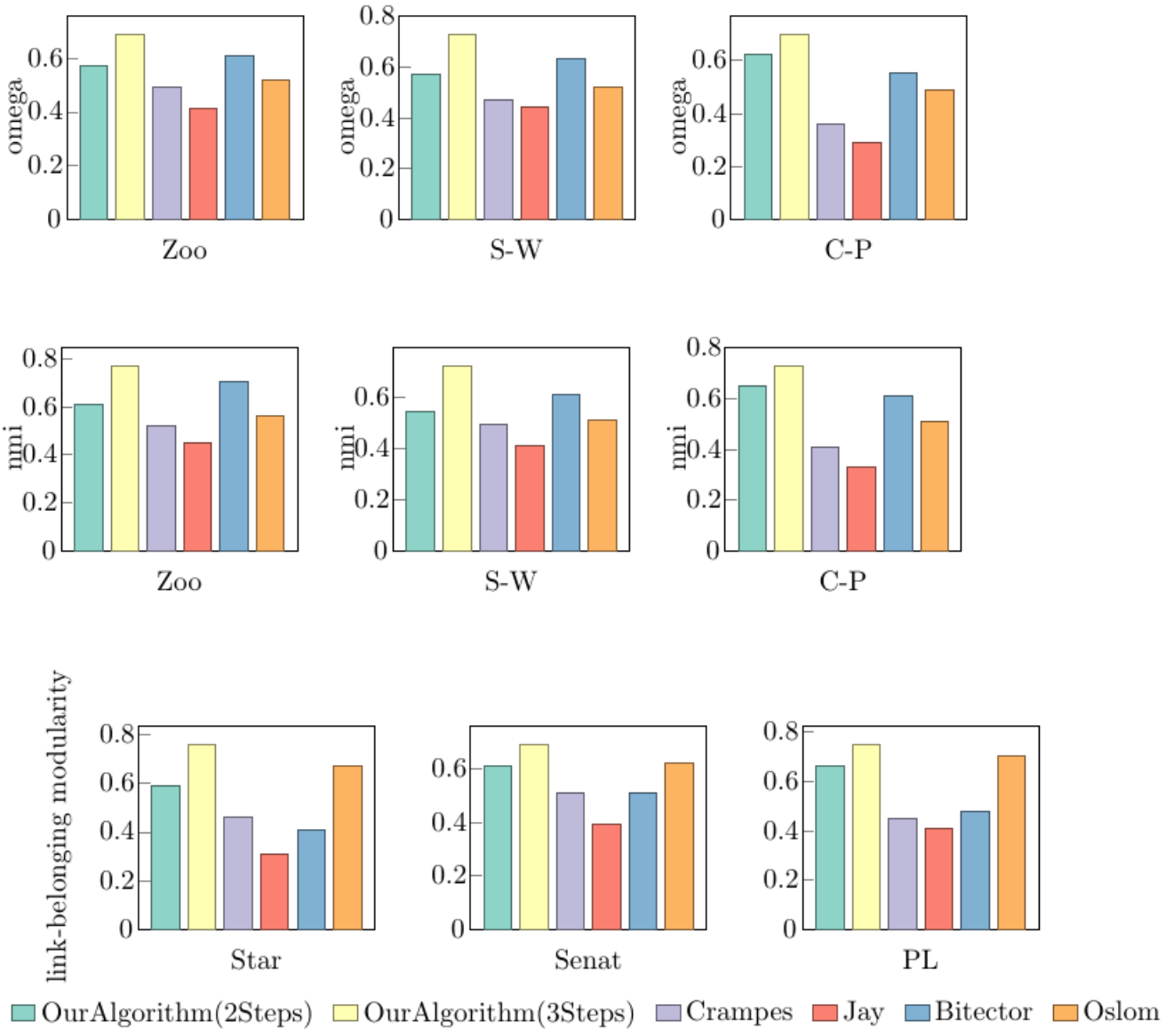


FIGURE 5.2: Évaluation de la précision des algorithmes à l'aide de NMI, OMEGA et (*link-belonging modularity*)

n'a pas si bien fonctionné sur les réseaux avec une réalité de terrain, bien qu'elle ait mieux réussi sur les autres ensembles de données. De plus, Oslom n'a pas pu détecter les petites communautés, contrairement à Bitector BiTector [43] qui trouve de petites communautés et fonctionne mieux avec des réseaux épars. Il commence par énumérer toutes les bicliques maximales, ce qui est coûteux dans les réseaux denses. Les algorithmes les moins performants sont les procédures de Crampes [38] et de Jay [64] et la méthode la moins précise semble être l'algorithme de Jay [64] probablement parce que certains objets étaient ignorés en utilisant un seuil sur les métriques utilisées. Pour le temps d'exécution, notre approche est présentée selon les deux variantes (cf. figure 5.3). Le temps est donné en secondes et représente la moyenne de dix exécutions de chacun des algorithmes évalués. La première variante de notre algorithme qui se limite au deux premières étapes obtient globalement de meilleurs temps d'exécution que les méthodes étudiées. Quant à la deuxième variante, elle dépasse les algorithmes existants quand il s'agit de contextes denses mais elle est égale ou légèrement moins coûteuse que les autres pour les contextes épars. L'algorithme le plus coûteux semble être celui de Jay [64].

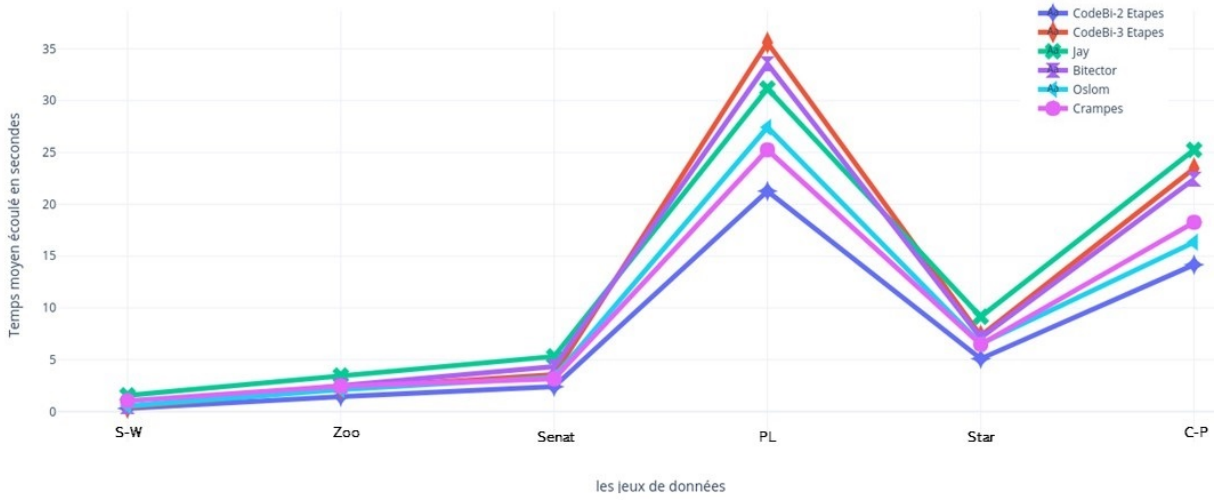


FIGURE 5.3: Temps d'exécution en secondes des algorithmes de détection de communautés

5.7.2 Deuxième série de tests

Comme indiqué précédemment, l'objectif de cette deuxième série de tests est de comparer la précision et le temps d'exécution de *CoDeBi* avec les quatre algorithmes

précités, mais en analysant des ensembles de données plus volumineux et trois variantes de l'algorithme principal, à savoir (i) les deux premières étapes de la procédure *CodeBi*, (ii) toute la procédure avec le calcul exact du coefficient Silhouette, (iii) toute la procédure avec le calcul approximatif du coefficient Silhouette. La dernière variante donne une estimation approximative du coefficient Silhouette en sélectionnant d'une manière aléatoire un échantillon d'objets dans chaque communauté de base. Pour cette série, les expériences sont effectuées sur un serveur SPARC-M7, 32 cores 4.133 GHz avec 64 Go de RAM, en utilisant notre code en Python.

Jeux de données utilisés

Les caractéristiques de ces jeux de données sont résumées par le tableau (5.6). Le premier jeu de données nommé *Breast-cancer* représente des femmes et des caractéristiques des noyaux cellulaires présents dans l'image numérisée d'une Biopsie à l'aiguille fine d'une masse mammaire.
Le deuxième jeu de données, *Nursery*, est dérivé d'un modèle de décision hiérarchique mis au point à l'origine pour classer les applications des écoles maternelles. Ce dernier contient des exemples de demandes d'inscription et des caractéristiques connexes, l'occupation des parents, la structure familiale, la situation financière, sociale et sanitaire de la famille, etc..
Le troisième jeu de données est un extrait de *Facebook*. Les données de ce dernier ont été collectées auprès des participants à l'enquête présentant les utilisateurs de Facebook et les caractéristiques de leurs profils. Le quatrième jeu de données *WikiElec* représente un réseau de participation des utilisateurs à l'élection de l'administration de Wikipédia. Le dernier réseau utilisé *Ca-HepTh* est un réseau scientifique de collaboration qui relie des auteurs et des articles soumis à la catégorie physique des Hautes Energies (*High Energy Physics*). Pour chaque jeu de données, nous fournissons le nombre d'objets dans $\mathcal{G}$ (nombre de noeuds du premier type), le nombre d'attributs dans $\mathcal{M}$ (nombre de noeuds du second type), le nombre de liens $\mathcal{I}$ entre les deux ensembles (types de nœuds) et le nombre de concepts formels générés $|\mathcal{C}|$. Afin d'exprimer tous les jeux de données de manière cohérente, nous avons converti les attributs catégoriques et numériques en attributs binaires. Ces jeux de données sont disponibles au public sur le site Web de

Stanford [1] ou le référentiel UCI de l'apprentissage automatique (*UC Irvine Machine Learning Database Repository*) [2].

Jeu de données	$\mathcal{G}$	$\mathcal{M}$	$\mathcal{I}$	$\lvert\mathcal{C}\rvert$
Breast-Cancer	699	110	6 990	9 860
Nursery	12 960	31	116 640	147 577
Facebook	3813	276	88 234	48 032
WikiElec	7000	2800	100 021	32 054
Ca-HepTh	9694	183	25 998	18 710

Tableau 5.6: Description des jeux de données

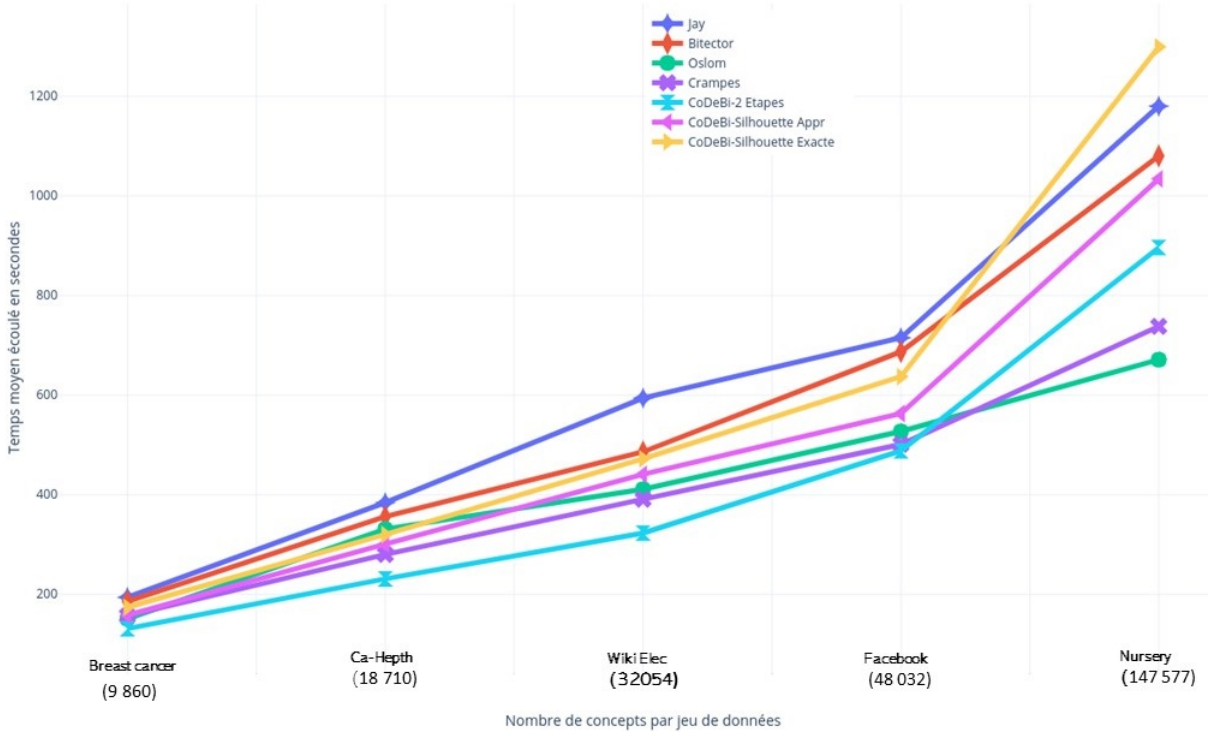


FIGURE 5.4: Temps d'exécution en secondes des algorithmes de détection de communautés

Chaque méthode et chaque variante de *CoDeBi* est exécutée dix fois sur tous les réseaux bipartis, puis les valeurs moyennes sont données par le tableau 5.7. Au terme de

1. `http://snap.stanford.edu/data/index.html#citnets`.
2. `http://archive.ics.uci.edu/ml/index.php`.

Méthodes	Breast-cancer	Nursey	Facebook	WikiElec	Ca-HepTh
Jay	0.132	0.089	0.105	0.168	0.092
Crampes	0.245	0.231	0.186	0.321	0.215
Bitector	0.316	0.267	0.251	0.271	0.447
Oslom	0.478	0.321	0.167	0.372	**0.487**
CodeBi (2étapes)	0.406	0.453	0.522	0.477	0.331
CodeBi (Silhouette exacte)	**0.681**	**0.594**	**0.729**	**0.701**	**0.583**
CodeBi (Silhouette appr)	0.517	**0.484**	**0.603**	**0.531**	0.462

Tableau 5.7: Évaluation de la précision des algorithmes à l'aide de NMI

ces expérimentations, nous pouvons faire quelques constats. La qualité de partitionnement, en comparaison avec la vérité terrain, que nous obtenons à travers notre méthode *CoDeBi* et ses variantes est souvent meilleure que celle des autres algorithmes. Les résultats des tests confirment la plupart des faits observés dans les résultats de la première série de tests.

Comme indiqué précédemment, l'approche proposée surpasse les autres dans presque tous les cas. Elle est suivi par Oslom. Cette dernière fonctionne mieux dans les grands réseaux mais souffre toujours du grave problème lié à l'optimisation de la modularité, car elle ne permet pas de localiser de petites communautés. Ce manquement est bien clair dans le jeu de données Facebook qui comporte plusieurs petites communautés imbriquées. Les algorithmes de Jay et Crampes sont probablement les méthodes les moins précises, car certains objets peuvent être ignorés à l'aide des seuils sur les métriques utilisées. Quant à Bitecor, il apparaît comme ayant plus de mal à trouver la structure communautaire à mesure que le réseau grandit.

Sur la figure 5.4 , le temps d'exécution est présenté pour les trois variantes de CodeBi et les méthodes sélectionnées, et inclut le calcul du coût du concept.

Nous avons utilisé l'algorithme efficace Data Peeler pour générer des concepts pour ces ensemble de données. Par exemple, il a fallu 0,45 seconde pour générer le 147577 concepts du jeu de données de *Nursey*.

Notons que dans certains cas, plus précisément lorsque le nombre de concepts augmente, notre approche renvoie un temps d'exécution légèrement plus coûteux que des

autres méthodes. Cela est particulièrement vrai lorsque la troisième étape avec le calcul exact de silhouette est utilisée.

Nous observons que le calcul exact du coefficient Silhouette est coûteux pour les grands réseaux car le temps d'exécution de l'analyse silhouette est dominé par le calcul exhaustif de la distance entre chaque paire d'objets. C'est la raison pour laquelle nous avons opté pour une silhouette approximative.

Les résultats des temps d'exécution que nous obtenons dans la figure 5.4 montrent que l'algorithme Oslom est le plus rapide, suivi de Crampes alors que ce dernier s'est avéré peu précis. Par ailleurs, notre algorithme *CoDeBi* avec les deux premières étapes vient en troisième position. Bitector et Jay et *CoDeBi* sont les plus coûteux. Le coût élevé de notre algorithme est dû à la troisième étape de raffinement utilisant Silhouette mais apportant plus de précision dans la détection de communautés.

Au terme de ces tests, nous pouvons dire que notre approche offre un bon compromis entre la qualité de la détection et les temps d'exécution.

5.8 Conclusion

Dans ce chapitre, nous avons proposé un nouvel algorithme de détection de communautés imbriquées et chevauchantes dans les réseaux de données à deux modes.

Notre méthode ne nécessite ni un nombre prédéfini de groupes, ni des seuils sur les métriques utilisées. Elle peut automatiquement identifier les communautés et leur description à travers les caractéristiques des concepts formels.

Elle s'appuie à la fois sur la sémantique associée à la structure inhérente aux données et sur un ensemble de métriques adéquates permettant de sélectionner des communautés pertinentes et cohérentes.

En effet, elle évalue la pertinence des concepts à l'aide de métriques basées sur l'analyse formelle de concepts telles que la stabilité et la séparation. Enfin, elle utilise l'indice de Silhouette pour mieux délimiter les communautés.

Pour valider la procédure *CoDeBi*, nous avons mené deux séries de tests afin de comparer sa précision et son temps d'exécution par rapport à quatre autres algorithmes de détection de communauté sur des réseaux du monde réel. Alors que la première série de tests concerne de petites collections de données, la seconde couvre de plus grands ensembles de données et vise à tester plus de variantes de *CoDeBi*. Les études

empiriques montrent que de notre approche est effective et efficace pour l'identification de communautés cheveauchantes et imbriquées dans des réseaux à deux modes de données.

Le processus de raffinement (3ème étape de la procédure) indique une meilleure précision dans la délimitation des communautés extraites. Cependant, il s'est avéré coûteux, ce qui pourrait être pénalisant dans de très grands réseaux. Toutefois, l'approche pourrait être améliorée davantage en utilisant le paradigme MapReduce [32].

Chapitre 6

Détection des communautés dans les réseaux multicouches

6.1 Introduction

Permettant une seule connexion entre deux nœuds, les réseaux unidimensionnels (à un seul mode) se révèlent donc incompatibles pour décrire les interactions au sein des systèmes complexes. En effet, les entités d'un réseau peuvent s'engager dans différents types d'interactions en même temps. À titre d'exemple, dans un réseau de collaboration scientifique, deux auteurs peuvent se connecter à travers une multitude de liaisons indiquant les conférences, les thèmes et les revues où ils ont co-publié des articles. De même, la plupart des utilisateurs ont aujourd'hui plusieurs comptes sur des réseaux sociaux *(Facebook, LinkedIn, Google plus, Twitter, etc.)*. Les réseaux multicouches ont récemment été proposés comme une alternative pour mieux décrire l'hétérogénéité des types de relations. La détection de communautés dans les réseaux multicouches dits aussi multiplexes demeure une question de recherche ouverte bien qu'elle ait été largement étudiée dans le contexte des réseaux unidimensionnels.

Dans ce chapitre, nous adaptons notre solution de détection de communautés proposée dans le chapitre 5 au contexte multicouche. Pour ce faire, nous allons appliquer des opérations d'assemblage de contextes en AFC en nous inspirant principalement de celles présentées dans [56, 140, 137]. En fait, ces opératons s'appliquent sur des contextes formels individuels pour en construire des contextes globaux. Dans ce qui suit, nous commençons par introduire les représentations usuelles des réseaux multicouches. Ensuite,

nous illustrons les trois opérations de base : l'apposition, la subposition et la concaténation préalablement définies dans la section 2.5. Après, nous détaillons la méthode par quelques illustrations de réseaux sociaux bibliographiques. Enfin, nous concluons ce chapitre et proposons quelques perspectives de recherche.

6.2 Représentation des réseaux multicouches

Un réseau multicouche, précédemment défini dans la sous-section 2.1.2, est souvent représenté [66, 67] par une série de réseaux unidimensionnels. Chaque couche contient un même ensemble de nœuds V mais correspond à un type différent de relation.
Kivelä et al. [67] ont présenté plusieurs jeux de données de réseaux complexes qui ont été utilisés dans la littérature. Un extrait de cet ensemble est présenté dans le tableau 6.1. Nous notons qu'un nombre placé entre parenthèses à côté du nom du réseau indique le nombre de réseaux différents existants, lorsqu'il est supérieur à 1. Les nombres entre parenthèses à côté des champs "nœuds" et "couches" indiquent le nombre de nœuds et de couches dans ces réseaux. S'il s'agit de deux réseaux tout au plus, les tailles des différents réseaux sont ainsi séparées par une virgule à l'intérieur de parenthèses.

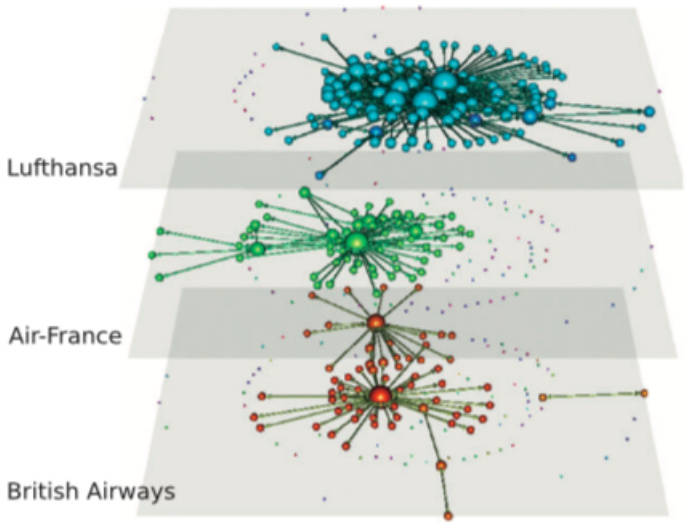


FIGURE 6.1: Visualisation d'un réseau multicouche du transport aérien [26]

Prenons l'exemple d'un réseau de transport aérien européen (cf. la figure 6.1) [26], les nœuds représentent les localisations des aéroports en Europe, les liens représentent

Nom du réseau social	Noeuds	Couches	Références
Enron e-mail	Utilisateurs (500)	Mots clés (500)	[129]
IBM	Individuels (3679)	Mots clés (1000)	[129]
Transport Aérien Europe (3)	Les aéroports (308–3108)	Compagnies aériennes (15–530)	[26, 37, 25]
Transport souterrain- Londonien	Stations (314)	Lignes (14)	[114, 37]
Flickr (2)	Utilisateurs (1000, 1186895)	Contacts, activités partagées (11, 4)	[40, 13]
Réseau terroriste mondial	Groupes terroristes (2509)	Pays cible (124)	[12]
DBLP-citations (2)	Auteurs (6848, 10305)	catégories de publications (617)	[85, 99]
DBLP -auteurs	Auteurs (424455)	Co-auteur, co-citations (3)	[22]
Baboon	Individuels (12)	Types d'interaction (3)	[8]
Friendfeed	Utilisateurs (7629)	Services (3)	[89]
Netflix	Films (13581)	Catégories d'évaluation (3)	[61]
Extrarandom.pl	Utilisateurs (4404)	Activités partagées (11)	[21]

Tableau 6.1: Exemples de jeux de données multiplexes utilisés dans la littérature

les itinéraires et chaque couche contient les vols d'une compagnie aérienne donnée.
Une des barrières qui nous est apparue est que ces représentations standards des réseaux multicouches semblent être inadéquates pour la description des interactions du monde réel qui sont plus complexes et peuvent dépasser une série de réseaux homogènes.
Pour ajouter plus de réalisme à l'étude des réseaux sociaux multicouches, il est nécessaire de s'intéresser à la superposition des couches hétérogènes. Prenons l'exemple de la figure 2.1 présentant un réseau multicouche scientifique où la première couche exprime les publications des auteurs dans des conférences, la deuxième couche exprime un réseau articles-conférences et la troisième couche montre un réseau articles-thèmes.
Dans ce chapitre, nous proposons une nouvelle approche du problème de la détection de communautés dans les réseaux multicouches, en particulier deux réseaux à deux modes superposés en se basant sur l'AFC.
Curieusement, au meilleur de notre connaissance, il n'existe pas de travaux sur la détection de communautés dans les RM en se basant sur l'AFC.

6.3 Opérations d'assemblage de contextes en AFC

Pour le traitement des réseaux multicouches, nous allons partir de deux contextes formels $\mathbb{K}_1 := (G_1, M_1, I_1)$ et $\mathbb{K}_2 := (G_2, M_2, I_2)$ représentant des réseaux à deux modes de données où l'un des ensembles de nœuds G_1 ou M_1 de $\mathbb{K}_1$ est égal à l'un des ensembles G_2 ou M_2 de $\mathbb{K}_2$, et, I_1 (respectivement I_2) est un ensemble de liens entre des éléments de G_1 et des éléments de M_1 (respectivement G_2 et M_2).
En effet, les cas d'étude peuvent être résumés autour de trois points :

(i) Deux contextes partageant le même ensemble d'objets ($G_1 = G_2$), c.-à-d.. le même groupe d'individus est décrit par deux ensembles distincts de propriétés et sont issus de deux réseaux différents tel que $M_1 \cap M_2 = \emptyset$.

(ii) Respectivement, deux contextes partageant le même ensemble d'attributs ($M_1 = M_2$) c.-à-d. deux groupes distincts d'individus possédant le même ensemble de propriétés avec $G_1 \cap G_2 = \emptyset$.

(iii) Deux contextes dont l'ensemble M_1 des attributs du premier représente l'ensemble des objets du deuxième ($M_1 = G_2$) c.-à-d. les propriétés des individus de premier réseau sont elles-mêmes décrites dans le deuxième réseau par un autre ensemble d'attributs.

6.4 Exemples illustratifs

Nous illustrons quelques contextes, les trois opérations avec des exemples de leurs compositions ainsi que les communautés produites en appliquant la procédure CoDeBi décrite dans le chapitre 5. Tout d'abord, considérons trois réseaux à deux modes décrits comme suit : le premier réseau, *Chercheurs- Evénements*, indique une participation de chercheurs à des événements scientifiques. Il est décrit par le contexte $\mathbb{K}_1$ (voir le tableau 6.2).

$\mathbb{K}_1$	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13
1	×		×		×			×				×	
2				×					×		×		×
3	×	×		×					×			×	
4	×	×	×									×	
5	×		×		×							×	
6		×				×	×			×			
7				×		×			×		×		×
8		×				×	×			×			
9		×				×	×				×		
10		×				×	×			×	×		
11		×				×	×			×	×		
12		×				×	×			×	×		
13				×					×		×		×
14				×		×			×		×		×

Tableau 6.2: Chercheurs-Événements

Le deuxième contexte $\mathbb{K}_2$ (cf. tableau 6.3) est le réseau *Evénements-Thèmes* d'association des événements à leurs thèmes principaux. Les abréviations SNA, DB, BIG, AI, KDD et DM, MS, DA, IP, Rob, IS, MLDM, VA, PreMod désignent respectivement analyse des réseaux sociaux, bases de données, *(Big Data)*, intelligence artificielle, découverte de connaissances à partir de bases de données, fouille de données, modélisation et simulation, analyse des données, traitement d'image, robotique, systèmes d'information, apprentissage automatique et exploration de données, analyse visuelle, et Modélisation prédictive.

La figure 6.2 illustre la présence de deux réseaux 1 et 2 reliés aux contextes $\mathbb{K}_1$ et $\mathbb{K}_2$ respectivement. Notons qu'outre les deux modes et les relations entre eux, la figure signale les communautés extraites par CoDeBi pour chacun des réseaux séparément. Les communautés imbriquées et chevauchantes sont représentées par des cercles en pointillés.

Le troisième contexte, décrit par $\mathbb{K}_3$ (cf. 6.4), représente un réseau d'association des chercheurs aux thèmes des événements auxquels ils participent. Il est obtenu par l'opération de composition sur les deux contextes $\mathbb{K}_1$ et $\mathbb{K}_2$.

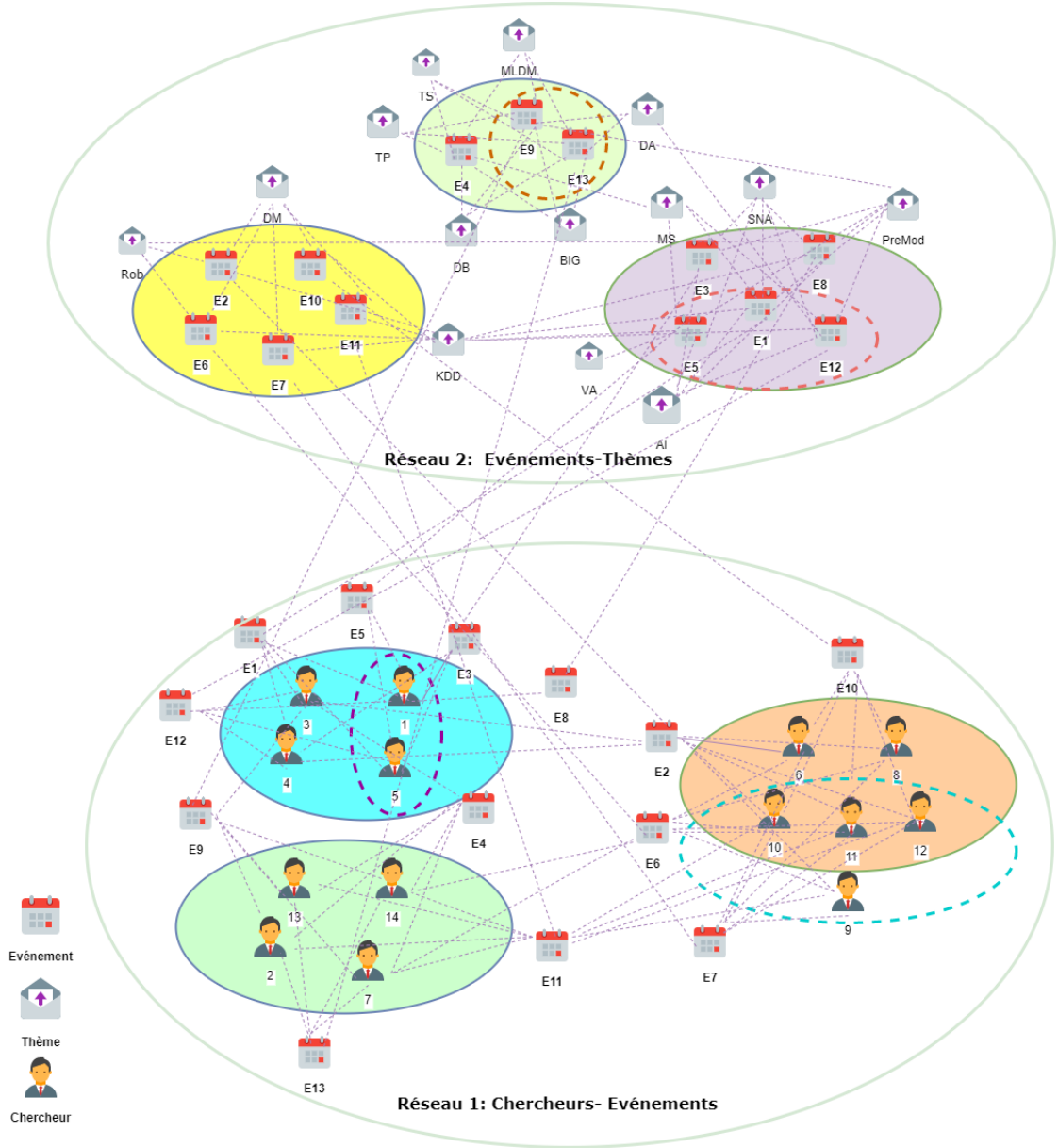


FIGURE 6.2: Un réseau multicouche ayant deux réseaux interreliés : la première couche représente un réseau (Chercheurs-Evénements) et la deuxième un réseau (Événements-Thèmes).

$\mathbb{K}_2$	SNA	DB	BIG	AI	KDD	DM	MS	DA	TP	Rob	TS	MLDM	VA	PreMod
E1	×			×	×		×							×
E2					×	×				×				
E3	×			×										×
E4		×	×				×		×			×		
E5	×			×	×		×						×	×
E6					×	×				×				
E7					×	×					×			
E8	×			×	×					×				×
E9		×	×					×	×			×		×
E10					×	×					×			
E11					×	×								
E12	×			×	×		×	×						×
E13		×	×					×	×			×		

Tableau 6.3: Événements-Thèmes

Ainsi donc, les trois contextes $\mathbb{K}_1 := (G_1, M_1, I_1)$, $\mathbb{K}_2 := (G_2, M_2, I_2)$ et $\mathbb{K}_3 := (G_3, M_3, I_3)$ sont tels que $M_1 = G_2$ et $M_2 = G_3$.

$\mathbb{K}_3$	SNA	DB	BIG	AI	KDD	DM	MS	DA	TP	Rob	TS	MLDM	VA	PreMod
1	×			×	×		×	×		×			×	×
2		×	×		×	×	×	×	×			×		×
3	×	×	×	×	×	×	×	×	×	×		×		×
4	×			×	×	×	×	×		×				×
5	×			×	×		×	×					×	×
6					×	×				×	×			
7		×	×		×	×	×	×	×	×		×		×
8					×	×				×	×			
9					×	×				×	×			
10					×	×				×	×			
11					×	×				×	×			
12					×	×				×	×			
13		×	×		×	×	×	×	×			×		×
14		×	×		×	×	×	×	×	×		×		×

Tableau 6.4: Chercheurs-Thèmes par composition de $\mathbb{K}_1$ et $\mathbb{K}_2$

La prise en compte du premier et du troisième réseau engendre un réseau ayant le même ensemble d'individus et deux ensembles d'attributs disjoints. Cela correspond à l'opération d'apposition de $\mathbb{K}_1$ et $\mathbb{K}_3$ aboutissant au contexte $\mathbb{K}_4$ indiqué dans le tableau 6.5.

L'ensemble des communautés extraites à partir de nouveau contexte $\mathbb{K}_4$ est représenté par le tableau 6.6.

La lecture du tableau 6.6 montre quatre groupes homogènes de chercheurs selon les événements scientifiques qu'ils fréquentent et les thèmes de ces événements. À titre d'exemple, on peut découvrir :

$\mathbb{K}_4$	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	SNA	DB	BIG	AI	KDD	DM	MS	DA	TP	Rob	TS	MLDM	VA	PreMod
1	x		x		x			x				x		x			x	x		x	x		x			x	x
2				x					x		x		x		x	x		x	x	x	x	x			x		x
3	x	x		x					x			x		x	x	x	x	x	x	x	x	x	x		x		x
4	x	x	x									x		x			x	x	x	x	x		x				x
5	x		x		x							x		x			x	x		x	x					x	x
6		x				x	x			x								x	x				x	x			
7				x		x			x		x		x		x	x		x	x	x	x	x	x		x		x
8		x				x	x			x								x	x				x	x			
9		x				x	x				x							x	x				x	x			
10		x				x	x			x	x							x	x				x	x			
11		x				x	x			x	x							x	x				x	x			
12		x				x	x			x	x							x	x				x	x			
13				x					x		x		x		x	x		x	x	x	x	x			x		x
14				x		x			x		x		x		x	x		x	x	x	x	x	x		x		x

Tableau 6.5: Apposition de $\mathbb{K}_1$ et $\mathbb{K}_3$

$\tilde{C_1}$	({ 6,8, 10, 11, 12}, {E2, E6, E7 ,E10, KDD, TS, DM, Rob})
$\tilde{C_2}$	({ 2, 3, 7, 13, 14 }, { E4, E9, KDD, DM, DA, PreMod, MS, DB, TP, MLDM, BIG })
$\tilde{C_3}$	({6,8,9, 10,11,12 }, {E2, E6, E7, KDD, TS, DM, Rob })
$\tilde{C_4}$	({1, 4,5}, { E1, E3, E12, KDD, DA, PreMod, AI, MS, SNA})

Tableau 6.6: Les communautés extraites du contexte $\mathbb{K}_4$

$\tilde{C_1}$	({6, 8, 9, 10, 11, 12 }, {Rob, KDD, DM, TS})
$\tilde{C_2}$	({ 2, 3, 7, 13, 14, 4] }, {DA, PreMod, MS, KDD, DM })
$\tilde{C_3}$	({3, 6, 7, 8, 9, 10, 11, 12, 14 }, {Rob, KDD, DM})
$\tilde{C_4}$	({1, 2, 3, 4, 5, 7, 13, 14}, {DA, PreMod, MS, KDD})

Tableau 6.7: les communautés extraites du contexte $\mathbb{K}_3$

(i) un élargissement de communautés existantes comme la communauté $\tilde{C_2}$= ({2, 3, 7, 13, 14}, {KDD, DM, DA, PreMod, MS, DB, TP, E4, MLDM, BIG, E9}) du contexte $\mathbb{K}_4$ suite à l'ajout du chercheur 3 à la communauté ({2, 7, 13, 14}, {E4, E9, E11, E13}) existante déja dans le réseau 1 (contexte $\mathbb{K}_1$) séparément. Cela est dû au fait que même si ce chercheur ne se trouve pas dans le même groupe d'événements que les quatre autres chercheurs, il fréquente des conférences ayant des thèmes similaires à ceux des événements $E4, E9, E11$, et $E13$.

(ii) L'existence d'un même groupe de chercheurs dans chacun des réseaux séparément mais certes ayant une description davantage enrichie dans $\mathbb{K}_4$ que dans $\mathbb{K}_1$, telle que la communauté $\tilde{C_1}$ = ({6,8, 10 11, 12}, {E2, E6, E7, E10, **KDD, TS, DM, Rob**}). Cette description nous indique que ces cinq chercheurs assistent non seulement aux événements ({E2, E6, E7, E10}) mais qu'ils sont aussi intéressés par les thèmes KDD, TS, DM et Rob de ces événements.

D'une manière duale à l'apposition, on peut effectuer l'opération de subposition de contextes sur $\mathbb{K}_2$ et $\mathbb{K}_3$ qui possèdent le même ensemble d'attributs (thèmes) et deux ensembles d'objets différents. Il en résulte le contexte $\mathbb{K}_5$ décrit par le tableau 6.8. Cette opération aurait été plus significative si on disposait de deux réseaux sociaux décrivant par exemple quelques membres du réseau professionnel *LinkedIn* à l'instant initial t_0 puis à l'instant t_1 avec uniquement l'ajout de nouveaux membres. Chaque membre est décrit par un même ensemble d'attributs comme le genre, la tranche d'âge, le dernier diplôme obtenu et la profession. En identifiant les communautés entre les deux instants, on est en mesure de voir leur évolution dans le temps.

$\mathbb{K}_5$	SNA	DB	KDD	BIG	AI	DM	MS	DA	TP	Rob	TS	MLDM	VA	PreMod
1	×			×	×		×	×		×			×	×
2		×	×		×	×	×	×	×			×		×
3	×	×	×	×	×	×	×	×	×	×		×		×
4	×			×	×	×	×	×		×				×
5	×			×	×		×	×					×	×
6					×	×				×	×			
7		×	×		×	×	×	×	×	×		×		×
8					×	×				×	×			
9					×	×				×	×			
10					×	×				×	×			
11					×	×				×	×			
12					×	×				×	×			
13		×	×		×	×	×	×	×			×		×
14		×	×		×	×	×	×	×	×		×		×
E1	×			×	×		×							×
E2					×	×				×				
E3	×			×										×
E4		×	×				×		×			×		
E5	×			×	×		×						×	×
E6					×	×				×				
E7					×	×					×			
E8	×			×	×					×				×
E9		×	×					×	×			×		×
E10					×	×					×			
E11					×	×								
E12	×			×	×		×	×						×
E13		×	×					×	×			×		

Tableau 6.8: Subposition de $\mathbb{K}_2$ et $\mathbb{K}_3$

La prise en compte des deux réseaux 1 et 2 peut être traitée par une opération de concaténation de contextes et plus particulièrement l'opération de composition. Ce nouveau contexte de composition représente les liens transitifs entre des individus et les caractéristiques attachées à leurs propres attributs. Notons que le contexte résultant n'est autre que le contexte $\mathbb{K}_3$ (voir la table 6.4). En combinant $\mathbb{K}_1$ et $\mathbb{K}_3$ par l'opération d'apposition, nous avons obtenu le contexte $\mathbb{K}_4$ ainsi que ses communautés.

6.5 Conclusion

Dans ce chapitre, nous avons utilisé *CoDeBi* pour identifier des communautés au sein des graphes multicouches, particulièrement deux graphes bipartis superposés. Pour cela, nous avons dû faire appel à trois opérations de manipulation de contextes en AFC [56, 140, 137] pour obtenir un contexte enrichi représentant la prise en compte de deux réseaux et qui est ensuite traité par *CodeBi*. La démarche est généralisable à plusieurs réseaux interreliés.

Nous envisageons de mener une étude empirique sur des réseaux réels et synthétiques en passant tout d'abord par une implémentation parallèle avec le paradigme *Map Reduce* afin d'optimiser le temps de calcul.

Nous soulignons que de nombreux *(frameworks)* ont vu le jour afin d'implémenter le *(MapReduce)* dont le plus connu est *(Hadoop)* développé par *Apache Software Foundation*. On rappelle que MapReduce (MR) est un paradigme de programmation destiné à la création de programmes parallèles qui traitent des grandes quantités de données, et qui s'exécutent sur des *clusters* de machines. Un programme MR s'écrit sous la forme de deux fonctions map et reduce, et les données sont représentées sous la forme de paires (clé, valeur).

Chapitre 7

Détection des communautés dans les réseaux tridimensionnels

7.1 Introduction

Dans ce chapitre, nous nous proposons d'étendre l'approche proposée dans le chapitre 5 à l'identification de communautés dans des réseaux tridimensionels décrits par les contextes tridimensionnels en Analyse Formelle de Concepts. À cet effet, nous définissons dans la première section les réseaux tridimensionnels et leurs modélisation. Ainsi, nous rappelons brièvement le contexte formel triadique, déjà présentée dans la sous-section 2.6, puis nous adaptons les extensions triadiques de la FCA et redéfinissons des mesures d'intérêt des concepts déjà vues dans le contexte dyadique et notamment la stabilité et la séparation.

7.2 Réseaux tridimensionnels

L'émergence des sites de réseautage social dans diverses disciplines telles que la recherche scientifique (*Academia, ResearchGate*) ou les médias sociaux *(Delicious, Flickr, YouTube*) a fait naître de nouveaux types de réseaux. Cela inclut les réseaux multidimensionnels dont un cas particulier est le réseau trimensionnel. Ce dernier comporte trois types de noeuds dont les instances sont liées entre elles par une relation ternaire. À titre d'exemple, on peut disposer d'un graphe triparti représenté par un contexte triadique décrivant des usagers d'une folksonomie qui annotent des ressources (ex. articles, photos,

vidéos, ...) à l'aide d'étiquettes (*tags*).

7.3 Contexte formel triadique

Nous rappelons qu'un contexte triadique $\mathbb{K} := (K_1, K_2, K_3, Y)$ décrit K_1, K_2 et K_3 comme un ensemble d'objets, un ensemble d'attributs et un ensemble de conditions respectivement avec $Y \subseteq K_1 \times K_2 \times K_3$ une relation ternaire entre les trois ensembles (cf. chapitre 2). Le triplet (a_1, a_2, a_3) dans Y signifie que l'objet a_1 possède l'attribut a_2 sous la condition a_3.
L'exemple donné dans le tableau 2.3 est un contexte triadique qui concerne un groupe K_1 composé de cinq chercheurs participant aux événements P, N, R, K et S de K_2 avec quatre diffèrents rôles a, b, c et d de K_3, notamment auteur, organisateur, conférencier invité et membre d'un comité de programme.
Prenons la valeur ac au croisement de la ligne 1 et de la colonne R. Elle signifie que le chercheur 1 assiste à l'événement R avec deux rôles a et c, c'est-à-dire en tant qu'auteur et conférencier invité.

En utilisant l'opération de dérivation définie dans la sous-section 2.6, $(PN, b)^{(1)} = \{1, 4\}$ signifie que les chercheurs 1 et 4 assistent aux événements P et N comme étant des organisateurs. Si nous calculons maintenant $\{1, 4\}^{(1)} = ((PN, b)^{(1)})^{(1)}$, nous obtenons la fermeture de (PN, b), présentant un ensemble de paires pouvant être factorisées comme suit : $((PN, b)^{(1)})^{(1)} = \{(PNK, b), (PN, bd)\}$. Cela nous informe que les deux chercheurs 1 et 4 partagent en fait ces deux caractéristiques : (PNK, b) et (PN, bd), c.-à-d. qu'ils agissent en tant qu'organisateurs (valeur b) des trois événements P, N et K d'une part, et d'autre part, ils sont à la fois organisateurs et membres d'un comité de programme pour P and N.

7.4 Mesures d'intérêt des concepts

Rappelons que notre première étape vers la détection de communauté dans les réseaux tridimensionnels consiste à exploiter l'analyse triadique des concepts. Afin d'identifier des communautés dans un esprit similaire à la procédure CoDeBi décrite au chapitre 5, la procédure adaptée implique :

(i) la génération de concepts triadiques,

(ii) le calcul des indices de stabilité et de séparation triadiques

(iii) le raffinement et l'identification des communautés d'une manière similaire au cas des réseaux à deux modes de données.

7.4.1 Génération des concepts triadiques

Le concept formel triadique d'un contexte $\mathbb{K} := (K_1, K_2, K_3, Y)$, précédemment défini dans la sous-section 2.6, est un triplet maximal (A_1, A_2, A_3) avec $A_1 \subseteq K_1$, $A_2 \subseteq K_2$, $A_3 \subseteq K_3$ et $A_1 \times A_2 \times A_3 \subseteq Y$. Il représente ainsi un cuboïde plein de 1. Afin de générer les concepts triadiques, nous avons utilisé un algorithme efficace et rapide nommé *Data Peeler* [27].

7.4.2 Stabilité triadique

La stabilité *extensionnelle* du concept triadique (A_1, A_2, A_3) est définie comme suit [1] :

$$Stab^{(1)}(A_1, A_2, A_3) := \frac{\mid \{X \subseteq A_1 | X^{(1)} \supseteq (A_2, A_3)\} \mid}{2^{|A_1|}} \tag{7.1}$$

Cette stabilité montre le degré de dépendance de la relation binaire entre les ensembles A_2 et A_3 par rapport à des sous-ensembles de l'extension A_1.

Considérons le contexte triadique du tableau 2.3, la stabilité du concept triadique $(\{3,4,5\}, \{R,K\}, \{a,b\})$ est égale à $\frac{3}{8} = 0.375$ car seuls les ensembles $\{3,5\}^{(1)}$, $\{4,5\}^{(1)}$ et $\{3,4,5\}^{(1)}$ contiennent la paire $(\{R,K\}, \{a,b\})$.

7.4.3 Séparation triadique

Nous définissons la séparation $\alpha(c)$ d'un concept triadique $c = (A_1, A_2, A_3)$ comme suit où g' est l'ensemble des paires attributs-conditions (m,p) appartenant à l'objet g tandis que $(m,p)'$ retourne les objets ayant les attributs-conditions $(m,p) \in A_2 \times A_3$:

$$\alpha(c) = \frac{\mid A_1 \mid \times \mid A_2 \mid \times \mid A_3 \mid}{\sum_{g \in A_1} \mid g' \mid + \sum_{(m,p) \in A_2 \times A_3} \mid (m,p)' \mid - \mid A_1 \mid \times \mid A_2 \mid \times \mid A_3 \mid} \tag{7.2}$$

1. Cette équation est légèrement différente de la formule générale définie par [77] car nous considérons que le symbole $\supseteq$ entre $X^{(1)}$ (A_2, A_3) est plus adéquat que $=$.

L'indice de séparation pour ce même concept $(\{3,4,5\},\{R,K\},\{a,b\})$ est $\frac{(3\times2\times2)}{((9+10+11)+(5+4+5+4)-(3\times2\times2))} = \frac{1}{3}$.
Ainsi, son score d'autonomie est égal à $2 \times \frac{0.375\times0.333}{0.375+0.333} = 0.35$. Par conséquent, les trois chercheurs 3, 4 et 5 partagent des caractéristiques communes puisqu'ils participent ensemble aux événements R et K en tant qu'auteurs et organisateurs.

Une communauté potentielle est $(\{1,3,5\},\{P,R,K,S\},\{a\})$ dont la stabilité égale à $\frac{4}{8} = 0.50$ (car la dérivation de chacun des quatre sous-ensembles (sur un total de huit) contient $(\{P,R,K,S\},\{a\})$). Ces sous-ensembles sont : {3}, {1, 3}, {3, 5}, {1, 3, 5}.
L'indice de séparation pour ce même concept est $\frac{(3\times2\times2)}{((9+10+11)+(5+4+5+4)-(3\times2\times2))} = .333$. La valeur d'autonomie est alors égale à $2 \times \frac{0.5\times0.324}{0.5+0.324} = 0.39$. Ce concept a un score d'autonomie légèrement supérieur au précédent et couvre les chercheurs qui assistent aux événements P, R, K, et S en tant qu'auteurs uniquement.
En adaptant notre procédure *CoDeBi* et en utilisant l'exemple du tableau 2.3, nous obtenons les communautés suivantes :
$\tilde{C}_1 = (\{3,4\},\{K,P,R\},\{a,b\})$, $\tilde{C}_2 = (\{1,5\},\{P,N,R,K,S\},\{a\})$,
et $\tilde{C}_3 = (\{2,5\},\{R\},\{a,b,d\})$. Notons que les concepts $(\{3,4,5\},\{R,K\},\{a,b\})$ et $(\{3,4,5\},\{R,K\},\{a,b\})$ ont un score d'autonomie supérieur à 0.343 lequel est associé à $(\{2,5\},\{R\},\{a,b,d\})$ mais ils ont été écartés car leurs extensions sont incluses dans l'ensemble d'objets collectés après la prise en compte de $\tilde{C}_1$ et $\tilde{C}_2$.

7.5 Conclusion

Dans ce chapitre, nous nous sommes intéressés aux réseaux sociaux complexes à représentation tripartie. Ainsi, nous avons exploité l'Analyse Triadique de Concepts et adapté les mesures de stabilité et de séparation pour exploiter le contexte triadique dans un esprit similaire au contexte dyadique. Notons que la démarche pourra être étendue pour la détection de communautés dans des réseaux d'une plus grande dimensionnalité. Nous prévoyons mener plus tard une étude empirique sur des réseaux réels et synthétiques en passant tout d'abord par une implémentation parallèle afin d'optimiser le temps de calcul.

Chapitre 8

Conclusion et perspectives

8.1 Conclusion

Le but de cette recherche est d'utiliser l'analyse formelle de concepts (AFC) et une de ses variantes comme cadre théorique de référence pour la détection de communautés dans les réseaux homogènes et hétérogènes dont les réseaux multicouches et multidimensionnels présentent des cas particuliers. Cela a donné lieu à trois publications scientifiques de notre contribution [63, 91, 92].

Dans la présente thèse, nous avons d'abord rappelé les notions de base de l'AFC et de celles de l'analyse de réseaux sociaux. Un survol de la littérature sur les approches existantes de détection de communautés aussi bien simples que complexes nous a permis non seulement de comprendre les travaux existants dans ce domaine, mais également de faire les observations suivantes, à savoir :

(i) La plupart des recherches fixent des seuils sur les métriques utilisées, requièrent un nombre prédéfini de communautés, ou dépendent d'un certain nombre de paramètres à l'entrée (graines, etc..).

(ii) Absence de travaux sur la détection des communautés dans les réseaux multidimensionnels ou mutlicouches en se basant sur l'AFC.

(iii) Les réseaux sociaux multicouches sont souvent présentés sous forme d'une série de réseaux homogènes alors que les interactions du monde réel sont beaucoup plus complexes..

Dans l'objectif de remédier à ces limites, nous avons proposé en premier lieu deux solutions COIN et CoDeBi pour la détection des communautés pouvant être chevauchantes

et imbriquées dans les réseaux à un et deux modes. Tout d'abord, différemment des travaux fondateurs qui ont exploité la théorie des graphes pour la modélisation des réseaux sociaux, nous avons porté un intérêt particulier à l'analyse formelle de concepts.

Nous avons tout d'abord proposé deux algorithmes de détection de communautés : COIN pour les réseaux à un seul mode de doinnées et CoDeBi pour les réseaux à deux modes. Les deux procédures ne requièrent ni un nombre de groupes prédéfini, ni des seuils sur les métriques utilisées. Elles peuvent automatiquement identifier les communautés cohésives et séparables aussi bien que leur description grâce à l'exploitation de l'AFC, ses caractéristiques et ses métriques appropriées telles que la stabilité et la séparation. CoDeBi utilise ensuite le coefficient Silhouette pour affiner les communautés. Quant à COIN, il finit par une étape de fusion des cliques pertinentes et adjacentes. Par ailleurs, nous avons implémenté notre approche en langage Python, et nous avons conduit des expérimentations sur des jeux de données réels de diverses tailles qui ont confirmé l'intérêt et l'efficacité de l'approche. Néanmoins, nous pensons que des progrès restent à faire au niveau du temps d'exécution de la méthode. En effet, à l'ère des données massives (*big data*) et des technologies du *cloud computing*, la parallélisation des approches devient nécessaire.

Notre autre contribution a consisté à étendre CoDeBi à l'identification de communautés au sein des graphes multicouches, particulièrement deux graphes bipartis superposés. Cette dernière a nécessité d'appliquer des opérations de manipulation de contextes en AFC, empruntées à des études d'analyse de concepts formels [56, 140, 137].

Pour traiter le volet de l'analyse de communautés dans les réseaux sociaux tridimensionnels sous la forme $G = (K_1, K_2, K_3, Y)$ où K_1, K_2 et K_3 sont trois ensembles distincts de nœuds et Y une relation ternaire vérifiant $Y \subseteq K_1 \times K_2 \times K_3$, nous avons exploité l'Analyse triadique de Concepts aussi bien de la stabilité des concepts, et nous avons adapté la métrique de la séparation pour exploiter le contexte triadique dans un esprit similaire à celle du contexte dyadique. Notons que la démarche pourra être étendue pour la détection de communautés dans des réseaux multidimensionnels ayant n-aires relations.

À court terme, nous prévoyons tester les deux dernières procédures sur des réseaux réels et synthétiques en passant tout d'abord par une conception et une implémentation parallèle de nos solutions afin d'optimiser le temps de calcul. L'inconvénient que nous

souhaitons souligner est le manque d'outils d'évaluation des communautés multicouches aussi bien que l'absence de réalité du terrain.

8.2 Perspectives de recherches

Plusieurs voies de recherche peuvent être explorées afin d'approfondir les résultats apportés dans le cadre de cette étude. Cela inclut l'exploration de domaines d'application et la prise en compte de l'AFC floue. En effet, l'analyse des réseaux sociaux, en particulier la détection de communautés, peut d'une part, aider à mieux comprendre les phénomènes et les processus qui se déroulent dans le monde social, et d'autre part, extraire de la connaissance servant dans plusieurs champs d'application tels que la recherche d'information, la navigation en ligne ou bien la recommandation.

Pour la recherche d'information et la navigation en ligne, il est possible de regrouper des pages web selon leur contenu en communautés afin de faciliter le processus d'indexation et retrouver d'une manière plus précise et efficace les documents les plus pertinents pour l'usager.

L'intérêt du processus de la recommandation est de proposer et prédire aux utilisateurs des réseaux sociaux du contenu adapté à leurs préférences sans pour autant exploiter des données relatives à leur profil et qu'ils ne souhaitent pas souvent partager. La détection de communautés peut alors servir pour recommander d'établir de nouveaux liens d'amitié, des services ou des contenus qui ont été utilisés ou bien évalués par le groupe auquel appartient un utilisateur. Dans le contexte de réseaux bibliographiques, on peut penser à la recommandation de nouvelles collaborations scientifique [109]. Dans le cadre de réseaux commerciaux d'achats tels que Amazon5 et Netflix, une communauté peut être vue comme une généralisation de l'approche classique du filtrage collaboratif. Cela suppose qu'on peut recommander à une personne des produits bien évalués par les membres de sa communauté, autrement dit, les individus appartenant à une même communauté auront probablement les mêmes préférences sur un autre ensemble d'items. Similairement, les produits peuvent être aussi regroupés en communautés selon le motif de leur achat, ce qui permet de recommander à un client des produits similaires à ceux qu'il a choisis.

Une autre perspective de recherche est inspirée par le constat que l'AFC classique exploite davantage des relations binaires. Cependant, certaines applications concrètes ont des données d'entrée obtenues par des mesures, des observations et des jugements

humains et peuvent donc donner lieu à de l'imprécision, l'incertitude et la gradualité. Aussi, l'adaptation de notre solution à des contextes formels flous est une piste à explorer.

Bibliographie

[1] Agarwal, G., and Kempe, D. Modularity-maximizing graph communities via mathematical programming. *The European Physical Journal B 66*, 3 (2008), 409–418.

[2] Alba, R. D. A graph-theoretic definition of a sociometric clique. *Journal of Mathematical Sociology 3*, 1 (1973), 113–126.

[3] Albert, R., and Barabási, A.-L. Statistical mechanics of complex networks. *Reviews of modern physics 74*, 1 (2002), 47.

[4] Alzahrani, T. Complex information networks–detecting community structure in bipartite networks.

[5] Amelio, A., and Pizzuti, C. A cooperative evolutionary approach to learn communities in multilayer networks. In *International Conference on Parallel Problem Solving from Nature* (2014), Springer, pp. 222–232.

[6] Andrews, S. In-close2, a high performance formal concept miner. In *International Conference on Conceptual Structures* (2011), Springer, pp. 50–62.

[7] Barabási, A.-L., and Albert, R. Emergence of scaling in random networks. *science 286*, 5439 (1999), 509–512.

[8] Barrett, L., Henzi, S. P., and Lusseau, D. Taking sociality seriously : the structure of multi-dimensional social networks as a source of information for individuals. *Philosophical Transactions of the Royal Society B : Biological Sciences 367*, 1599 (2012), 2108–2118.

[9] Batagelj, V., and Zaversnik, M. An o (m) algorithm for cores decomposition of networks. *arXiv preprint cs/0310049* (2003).

[10] Battiston, F., Nicosia, V., and Latora, V. Metrics for the analysis of multiplex networks. *arXiv preprint arXiv :1308.3182* (2013).

[11] Beauguitte, L. Cliques, communautés et dérivées.

[12] Berlingerio, M., Coscia, M., and Giannotti, F. Finding and characterizing communities in multidimensional networks. In *2011 International Conference on Advances in Social Networks Analysis and Mining* (2011), IEEE, pp. 490–494.

[13] Berlingerio, M., Coscia, M., Giannotti, F., Monreale, A., and Pedreschi, D. Multidimensional networks : foundations of structural analysis. *World Wide Web 16*, 5-6 (2013), 567–593.

[14] Berlingerio, M., Pinelli, F., and Calabrese, F. Abacus : frequent pattern mining-based community discovery in multidimensional networks. *Data Mining and Knowledge Discovery 27*, 3 (2013), 294–320.

[15] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. Fast unfolding of communities in large networks. *Journal of statistical mechanics : theory and experiment 2008*, 10 (2008), P10008.

[16] Bomze, I. M., Budinich, M., Pardalos, P. M., and Pelillo, M. The maximum clique problem. In *Handbook of combinatorial optimization*. Springer, 1999, pp. 1–74.

[17] Bonacich, P. Factoring and weighting approaches to status scores and clique identification. *Journal of mathematical sociology 2*, 1 (1972), 113–120.

[18] Bonacich, P. Some unique properties of eigenvector centrality. *Social networks 29*, 4 (2007), 555–564.

[19] Borgatti, S. P. 2-mode concepts in social network analysis. *Encyclopedia of complexity and system science 6* (2009), 8279–8291.

[20] Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., and Wagner, D. On modularity clustering. *IEEE transactions on knowledge and data engineering 20*, 2 (2008), 172–188.

[21] Bródka, P., Kazienko, P., Musiał, K., and Skibicki, K. Analysis of neighbourhoods in multi-layered dynamic social networks. *International Journal of Computational Intelligence Systems 5*, 3 (2012), 582–596.

[22] Brodka, P., Stawiak, P., and Kazienko, P. Shortest path discovery in the multi-layered social network. In *2011 International Conference on Advances in Social Networks Analysis and Mining* (2011), IEEE, pp. 497–501.

[23] Bron, C., and Kerbosch, J. Algorithm 457 : finding all cliques of an undirected graph. *Communications of the ACM 16*, 9 (1973), 575–577.

[24] Carchiolo, V., Longheu, A., Malgeri, M., and Mangioni, G. Communities unfolding in multislice networks. In *Complex Networks*. Springer, 2011, pp. 187–195.

[25] Cardillo, A., Gómez-Gardenes, J., Zanin, M., Romance, M., Papo, D., Del Pozo, F., and Boccaletti, S. Emergence of network features from multiplexity. *Scientific reports 3* (2013), 1344.

[26] Cardillo, A., Zanin, M., Gómez-Gardenes, J., Romance, M., del Amo, A. J. G., and Boccaletti, S. Modeling the multi-layer nature of the european air transport network : Resilience and passengers re-scheduling under random failures. *The European Physical Journal Special Topics 215*, 1 (2013), 23–33.

[27] Cerf, L., Besson, J., Robardet, C., and Boulicaut, J.-F. Closed patterns meet n-ary relations. *ACM Transactions on Knowledge Discovery from Data (TKDD) 3*, 1 (2009), 3.

[28] Chakraborty, T., Dalmia, A., Mukherjee, A., and Ganguly, N. Metrics for community analysis : A survey. *ACM Computing Surveys (CSUR) 50*, 4 (2017), 54.

[29] Chikhi, N. F. *Calcul de centralité et identification de structures de communautés dans les graphes de documents*. PhD thesis, Université de Toulouse, Université Toulouse III-Paul Sabatier, 2010.

[30] Christakis, N. A., and Fowler, J. H. The spread of obesity in a large social network over 32 years. *New England journal of medicine 357*, 4 (2007), 370–379.

[31] Christakis, N. A., and Fowler, J. H. The collective dynamics of smoking in a large social network. *New England journal of medicine 358*, 21 (2008), 2249–2258.

[32] Chunduri, R. K., and Cherukuri, A. K. Haloop approach for concept generation in formal concept analysis. *JIKM 17*, 3 (2018), 1850029.

[33] Clauset, A., Newman, M. E., and Moore, C. Finding community structure in very large networks. *Physical review E 70*, 6 (2004), 066111.

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

[35] Cordasco, G., and Gargano, L. Label propagation algorithm : a semi-synchronous approach. *International Journal of Social Network Mining 1*, 1 (2012), 3–26.

[36] Corlette, D., and Shipman III, F. M. Link prediction applied to an open large-scale online social network. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia* (2010), ACM, pp. 135–140.

[37] Cozzo, E., Kivelä, M., De Domenico, M., Solé, A., Arenas, A., Gómez, S., Porter, M. A., and Moreno, Y. Clustering coefficients in multiplex networks. *arXiv preprint arXiv :1307.6780* (2013).

[38] Crampes, M., and Plantié, M. Détection de communautés dans les graphes bipartis. In *IC 2012* (2012), p. 125.

[39] Danon, L., Diaz-Guilera, A., Duch, J., and Arenas, A. Comparing community structure identification. *Journal of Statistical Mechanics : Theory and Experiment 2005*, 09 (2005), P09008.

[40] De, M. D. L. P. O., HermannMaurer, P. A. A., and Imber, J. B. Knowledge management, information systems, e-learning, and sustainability research.

[41] De Domenico, M., Lancichinetti, A., Arenas, A., and Rosvall, M. Identifying modular flows on multilayer networks reveals highly overlapping organization in interconnected systems. *Physical Review X 5*, 1 (2015), 011027.

[42] Dong, X., Frossard, P., Vandergheynst, P., and Nefedov, N. Clustering on multi-layer graphs via subspace analysis on grassmann manifolds. *IEEE Transactions on signal processing 62*, 4 (2013), 905–918.

[43] Du, N., Wang, B., Wu, B., and Wang, Y. Overlapping community detection in bipartite networks. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01* (2008), IEEE Computer Society, pp. 176–179.

[44] Duch, J., and Arenas, A. Community detection in complex networks using extremal optimization. *Physical review E 72*, 2 (2005), 027104.

[45] Dunlavy, D. M., Kolda, T. G., and Kegelmeyer, W. P. Multilinear algebra for analyzing data with multiple linkages. In *Graph algorithms in the language of linear algebra*. SIAM, 2011, pp. 85–114.

[46] Durland, M. M., and Fredericks, K. A. An introduction to social network analysis. *New Directions for Evaluation 2005*, 107 (2005), 5–13.

[47] Falzon, L. Determining groups from the clique structure in large social networks. *Social networks 22*, 2 (2000), 159–172.

[48] Field, S., Frank, K. A., Schiller, K., Riegle-Crumb, C., and Muller, C. Identifying positions from affiliation networks : Preserving the duality of people and events. *Social Networks 28*, 2 (2006), 97–123.

[49] Fortunato, S. Community detection in graphs. *Physics reports 486*, 3-5 (2010), 75–174.

[50] Fowler, J. H., and Christakis, N. A. Dynamic spread of happiness in a large social network : longitudinal analysis over 20 years in the framingham heart study. *Bmj 337* (2008), a2338.

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

[52] Freeman, L. C. Cliques, galois lattices, and the structure of human social groups. *Social networks 18*, 3 (1996), 173–187.

[53] Freeman, L. C., and White, D. R. Using galois lattices to represent network data. *Sociological methodology* (1993), 127–146.

[54] Ganter, B., and Kuznetsov, S. O. Stepwise construction of the dedekind-macneille completion. In *International Conference on Conceptual Structures* (1998), Springer, pp. 295–302.

[55] Ganter, B., and Wille, R. *Formal Concept Analysis : Mathematical Foundations*. Springer-Verlag New York, Inc., 1999.

[56] Ganter, R., and Wille, R. Formal concept analysis : Mathematical foundations springer-verlag berlin germany.

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

[58] Guimera, R., Sales-Pardo, M., and Amaral, L. A. N. Modularity from fluctuations in random graphs and complex networks. *Physical Review E 70*, 2 (2004), 025101.

[59] Hao, F., Min, G., Pei, Z., Park, D.-S., and Yang, L. T. k-clique community detection in social networks based on formal concept analysis. *IEEE Systems Journal 11*, 1 (2017), 250–259.

[60] Hmimida, M., and Kanawati, R. Community detection in multiplex networks : A seed-centric approach. *NHM 10*, 1 (2015), 71–85.

[61] Horvát, E.-A., and Zweig, K. A. One-mode projection of multiplex bipartite graphs. In *Proceedings of the 2012 international conference on advances in social networks analysis and mining (ASONAM 2012)* (2012), IEEE Computer Society, pp. 599–606.

[62] Ibrahim, M.-H., and Missaoui, R. An efficient approximation of concept stability using low-discrepancy sampling. In *International Conference on Conceptual Structures* (2018), Springer, pp. 24–38.

[63] Ibrahim, M. H., Missaoui, R., and Messaoudi, A. Detecting communities in social networks using concept interestingness. In *Proceedings of the 28th Annual International Conference on Computer Science and Software Engineering, CASCON 2018, Markham, Ontario, Canada, October 29-31, 2018* (2018), I. Onut, A. Jaramillo, G. Jourdan, D. C. Petriu, and W. Chen, Eds., ACM, pp. 81–90.

[64] Jay, N., Kohler, F., and Napoli, A. Analysis of social communities with iceberg and stability-based concept lattices. In *International Conference on Formal Concept Analysis* (2008), Springer, pp. 258–272.

[65] Kanawati, R. Détection de communautés dans les grands graphes d'interactions (multiplexes) : état de l'art.

[66] Kim, J., and Lee, J.-G. Community detection in multi-layer graphs : A survey. *ACM SIGMOD Record 44*, 3 (2015), 37–48.

[67] Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., and Porter, M. A. Multilayer networks. *Journal of complex networks 2*, 3 (2014), 203–271.

[68] Klimushkin, M., Obiedkov, S. A., and Roth, C. Approaches to the selection of relevant concepts in the case of noisy data. In *ICFCA* (2010), vol. 20, Springer, pp. 255–266.

[69] Kolda, T. G., and Bader, B. W. Tensor decompositions and applications. *SIAM review 51*, 3 (2009), 455–500.

[70] Krajca, P., Outrata, J., and Vychodil, V. Parallel recursive algorithm for fca. In *CLA* (2008), vol. 2008, Citeseer, pp. 71–82.

[71] Kun, J., Caceres, R., and Carter, K. Locally boosted graph aggregation for community detection. *arXiv preprint arXiv :1405.3210* (2014).

[72] Kuncheva, Z., and Montana, G. Community detection in multiplex networks using locally adaptive random walks. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015* (2015), ACM, pp. 1308–1315.

[73] Kuznetsov, S. Interpretation on graphs and complexity characteristics of the problems of finding regularities of a certain type. *Nauchn.-Tekhn. Inform., Ser 2* (1989), 23–28.

[74] Kuznetsov, S., Obiedkov, S., and Roth, C. Reducing the representation complexity of lattice-based taxonomies. In *International Conference on Conceptual Structures* (2007), Springer, pp. 241–254.

[75] Kuznetsov, S. O. A fast algorithm for computing all intersections of objects from an arbitrary semilattice. *Nauchno-Tekhnicheskaya Informatsiya Seriya 2-Informatsionnye Protsessy i Sistemy*, 1 (1993), 17–20.

[76] Kuznetsov, S. O., and Makhalova, T. P. Concept interestingness measures : a comparative study. In *CLA* (2015), vol. 1466, pp. 59–72.

[77] Kuznetsov, S. O., and Makhalova, T. P. On stability of triadic concepts. In *Proceedings of the Thirteenth International Conference on Concept Lattices and Their Applications, Moscow, Russia, July 18-22, 2016.* (2016), pp. 245–253.

[78] Kuznetsov, S. O., and Obiedkov, S. A. Comparing performance of algorithms for generating concept lattices. *Journal of Experimental & Theoretical Artificial Intelligence 14*, 2-3 (2002), 189–216.

[79] Lancichinetti, A., and Fortunato, S. Community detection algorithms : a comparative analysis. *Physical review E 80*, 5 (2009), 056117.

[80] Lancichinetti, A., Radicchi, F., Ramasco, J. J., and Fortunato, S. Finding statistically significant communities in networks. *PloS one 6*, 4 (2011), e18961.

[81] Lehmann, F., and Wille, R. A Triadic Approach to Formal Concept Analysis. In *Proceedings of the Third International Conference on Conceptual Structures : Applications, Implementation and Theory* (1995), pp. 32–43.

[82] Lehmann, S., Schwartz, M., and Hansen, L. K. Biclique communities. *Physical review E 78*, 1 (2008), 016108.

[83] Leskovec, J., Adamic, L. A., and Huberman, B. A. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB) 1*, 1 (2007), 5.

[84] Li, J., and Song, Y. Community detection in complex networks using extended compact genetic algorithm. *Soft computing 17*, 6 (2013), 925–937.

[85] Li, X., Ng, M. K., and Ye, Y. Har : hub, authority and relevance scores in multi-relational data for query search. In *Proceedings of the 2012 SIAM International Conference on Data Mining* (2012), SIAM, pp. 141–152.

[86] Li, X., Ng, M. K., and Ye, Y. Multicomm : Finding community structure in multi-dimensional networks. *IEEE Transactions on Knowledge and Data Engineering 26*, 4 (2013), 929–941.

[87] Luccio, F., and Sami, M. On the decomposition of networks in minimally interconnected subnetworks. *IEEE Transactions on Circuit Theory 16*, 2 (1969), 184–188.

[88] Luce, R. D. Connectivity and generalized cliques in sociometric group structure. *Psychometrika 15*, 2 (1950), 169–190.

[89] Magnani, M., and Rossi, L. Formation of multiple networks. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction* (2013), Springer, pp. 257–264.

[90] Mannila, H., Toivonen, H., and Verkamo, A. I. E cient algorithms for discovering association rules. In *KDD-94 : AAAI workshop on Knowledge Discovery in Databases* (1994), pp. 181–192.

[91] Messaoudi, A., Missaoui, R., and Ibrahim, M. H. Detecting overlapping communities in two-mode data networks using formal concept analysis. In *Extraction et Gestion des connaissances, EGC 2019, Metz, France, January 21-25, 2019* (2019), M. Rousset and L. Boudjeloud-Assala, Eds., vol. E-35 of *RNTI*, Éditions RNTI, pp. 189–200. Best academic paper award.

[92] Missaoui, R., Messaoudi, A., Ibrahim, M. H., and Abdessalem, T. *Advances in Knowledge Discovery and Management*. Springer, 2020, ch. Detecting Communities in Multilayer Networks using Formal Concept Analysis. Rakia Jaziri and Arnaud Martin and Lydia Boudjeloud and Marie-Christine Rousset and fabric Guillet (Ed.). Accepté.

[93] Mokken, R. J. Cliques, clubs and clans. *Quality and quantity 13*, 2 (1979), 161–173.

[94] Mucha, P. J., Richardson, T., Macon, K., Porter, M. A., and Onnela, J.-P. Community structure in time-dependent, multiscale, and multiplex networks. *science 328*, 5980 (2010), 876–878.

[95] Newman, M. E. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the national academy of sciences 101*, suppl 1 (2004), 5200–5205.

[96] Newman, M. E. Fast algorithm for detecting community structure in networks. *Physical review E 69*, 6 (2004), 066133.

[97] Newman, M. E. Modularity and community structure in networks. *Proceedings of the national academy of sciences 103*, 23 (2006), 8577–8582.

[98] Newman, M. E., and Girvan, M. Finding and evaluating community structure in networks. *Physical review E 69*, 2 (2004), 026113.

[99] Ng, M. K.-P., Li, X., and Ye, Y. Multirank : co-ranking for objects and relations in multi-relational data. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* (2011), ACM, pp. 1217–1225.

[100] Norbert, T., and Félicité, G. D. Détection des communautés dans les réseaux orientés à l'aide des concepts formels. *Revue Africaine de la Recherche en Informatique et Mathématiques Appliquées* (2016).

[101] Palla, G., Barabási, A.-L., and Vicsek, T. Quantifying social group evolution. *Nature 446*, 7136 (2007), 664.

[102] Palla, G., Derényi, I., Farkas, I., and Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature 435*, 7043 (2005), 814.

[103] Papadopoulos, S., Kompatsiaris, Y., Vakali, A., and Spyridonos, P. Community detection in social media. *Data Mining and Knowledge Discovery 24*, 3 (2012), 515–554.

[104] Papalexakis, E. E., Akoglu, L., and Ience, D. Do more views of a graph help ? community detection and clustering in multi-graphs. In *Proceedings of the 16th International Conference on Information Fusion* (2013), IEEE, pp. 899–905.

[105] Pizzuti, C. Boosting the detection of modular community structure with genetic algorithms and local search. In *Proceedings of the 27th annual ACM symposium on applied computing* (2012), ACM, pp. 226–231.

[106] Pons, P., and Latapy, M. Computing communities in large networks using random walks. In *International symposium on computer and information sciences* (2005), Springer, pp. 284–293.

[107] Pons, P., and Latapy, M. Computing communities in large networks using random walks. *J. Graph Algorithms Appl. 10*, 2 (2006), 191–218.

[108] Potgieter, A., April, K. A., Cooke, R. J., and Osunmakinde, I. O. Temporality in link prediction : Understanding social complexity. *Emergence : Complexity & Organization (E : CO) 11*, 1 (2009), 69–83.

[109] Pujari, M., and Kanawati, R. Link prediction in multiplex bibliographical networks. *International Journal of Complex Systems in Science 3*, 1 (2013), 77–82.

[110] Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., and Parisi, D. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences 101*, 9 (2004), 2658–2663.

[111] Raghavan, U. N., Albert, R., and Kumara, S. Near linear time algorithm to detect community structures in large-scale networks. *Physical review E 76*, 3 (2007), 036106.

[112] Rattigan, M. J., Maier, M., and Jensen, D. Graph clustering with network structure indices. In *Proceedings of the 24th international conference on Machine learning* (2007), ACM, pp. 783–790.

[113] Reichardt, J., and Bornholdt, S. Statistical mechanics of community detection. *Physical Review E 74*, 1 (2006), 016110.

[114] Rombach, M. P., Porter, M. A., Fowler, J. H., and Mucha, P. J. Core-periphery structure in networks. *SIAM Journal on Applied mathematics 74*, 1 (2014), 167–190.

[115] Rosvall, M., and Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences 105*, 4 (2008), 1118–1123.

[116] Roth, C., Obiedkov, S., and Kourie, D. G. On succinct representation of knowledge community taxonomies with formal concept analysis. *International Journal of Foundations of Computer Science 19*, 02 (2008), 383–404.

[117] Rousseeuw, P. J. Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics 20* (1987), 53–65.

[118] Seidman, S. B. Network structure and minimum degree. *Social networks 5*, 3 (1983), 269–287.

[119] Seidman, S. B., and Foster, B. L. A graph-theoretic generalization of the clique concept. *Journal of Mathematical sociology 6*, 1 (1978), 139–154.

[120] Seifi, M. *Cœurs stables de communautés dans les graphes de terrain.* PhD thesis, Paris 6, 2012.

[121] Selmane, S., Bentayeb, F., Missaoui, R., and Boussaid, O. An efficient method for community detection based on formal concept analysis. In *Foundations of Intelligent Systems - 21st International Symposium, ISMIS 2014, Roskilde, Denmark, June 25-27, 2014. Proceedings* (2014), pp. 61–72.

[122] Selmane, S. A. *Détection et analyse des communautés dans les réseaux sociaux : approche basée sur l'analyse formelle de concepts.* PhD thesis, Lyon 2, 2015.

[123] Shi, C., Li, Y., Zhang, J., Sun, Y., and Philip, S. Y. A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering 29*, 1 (2017), 17–37.

[124] Skvoretz, J., and Faust, K. Logit models for affiliation networks. *Sociological Methodology 29*, 1 (1999), 253–280.

[125] Stattner, E. *Contributions à l'étude des réseaux sociaux : propagation, fouille, collecte de données.* PhD thesis, Université des Antilles-Guyane, 2012.

[126] Strehl, A., and Ghosh, J. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research 3*, Dec (2002), 583–617.

[127] Stumme, G., Taouil, R., Bastide, Y., Pasquier, N., and Lakhal, L. Computing iceberg concept lattices with titanic. *Data & knowledge engineering 42*, 2 (2002), 189–222.

[128] Šubelj, L., and Bajec, M. Robust network community detection using balanced propagation. *The European Physical Journal B 81*, 3 (2011), 353–362.

[129] Sun, J., Papadimitriou, S., Lin, C.-Y., Cao, N., Liu, S., and Qian, W. Multivis : Content-based social network exploration through multi-way visual analysis.

In *Proceedings of the 2009 SIAM International Conference on Data Mining* (2009), SIAM, pp. 1064–1075.

[130] Suthers, D., Fusco, J., Schank, P., Chu, K.-H., and Schlager, M. Discovery of community structures in a heterogeneous professional online network. In *2013 46th Hawaii International Conference on System Sciences* (2013), IEEE, pp. 3262–3271.

[131] Talbi, M. *Une nouvelle approche de détection de communautés dans les réseaux sociaux*. PhD thesis, Université du Québec en Outaouais, 2013.

[132] Tang, L., and Liu, H. Community detection and mining in social media. *Synthesis lectures on data mining and knowledge discovery 2*, 1 (2010), 1–137.

[133] Tang, L., Wang, X., and Liu, H. Community detection via heterogeneous interaction analysis. *Data mining and knowledge discovery 25*, 1 (2012), 1–33.

[134] Tang, W., Lu, Z., and Dhillon, I. S. Clustering with multiple graphs. In *2009 Ninth IEEE International Conference on Data Mining* (2009), IEEE, pp. 1016–1021.

[135] Tatti, N., Moerchen, F., and Calders, T. Finding robust itemsets under subsampling. *ACM Transactions on Database Systems (TODS) 39*, 3 (2014), 20.

[136] Tomasini, M. An introduction to multilayer networks. *BioComplex Laboratory, Florida Institute of Technology, Melbourne, USA* (2015), 1–14.

[137] Valtchev, P., and Missaoui, R. Building concept (galois) lattices from parts : generalizing the incremental methods. In *International Conference on Conceptual Structures* (2001), Springer, pp. 290–303.

[138] Watts, D. J., and Strogatz, S. H. Collective dynamics of 'small-world'networks. *nature 393*, 6684 (1998), 440.

[139] Wilkinson, D. M., and Huberman, B. A. A method for finding communities of related genes. *proceedings of the national Academy of sciences 101*, suppl 1 (2004), 5241–5248.

[140] Wille, R. Conceptual structures of multicontexts. In *International Conference on Conceptual Structures* (1996), Springer, pp. 23–39.

[141] Xie, J., Kelley, S., and Szymanski, B. K. Overlapping community detection in networks : The state-of-the-art and comparative study. *Acm computing surveys (csur) 45*, 4 (2013), 43.

[142] Xie, J., and Szymanski, B. K. Community detection using a neighborhood strength driven label propagation algorithm. *arXiv preprint arXiv :1105.3264* (2011).

[143] Zanette, D. H. Dynamics of rumor propagation on small-world networks. *Physical review E 65*, 4 (2002), 041908.

[144] Zhi, H.-l. On the calculation of formal concept stability. *Journal of Applied Mathematics 2014* (2014).

[145] Zhou, T., Liu, J.-G., and Wang, B.-H. Comment on"scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *arXiv preprint physics/0511084* (2005).