UNIVERSITÉ DU QUÉBEC À MONTRÉAL

UNE NOUVELLE APPROCHE DE DÉTECTION DE COMMUNAUTÉS DANS LES RÉSEAUX MULTIDIMENSIONNELS

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

OUALID BOUTEMINE

AVRIL 2016

REMERCIEMENTS

En premier lieu, je remercie Dieu de m'avoir permis de mener ce travail de recherche à terme.

Ce mémoire n'aurait pas vu le jour sans la contribution de plusieurs personnes à qui j'aimerais adresser mes remerciements. Tout d'abord, je tiens à remercier mon directeur de recherche, Mohamed Bouguessa, qui a supervisé mon travail tout en me laissant une grande marge de liberté. Je le remercie pour son encadrement, sa patience, sa disponibilité et la pertinence de ses remarques tout au long de la réalisation de ce projet de maîtrise.

Je remercie également la faculté des sciences de l'UQAM pour la bourse d'exemption des frais de scolarité majorés que j'ai reçue durant mes études de maîtrise. J'adresse également ma gratitude à la Fondation de l'UQÀM pour m'avoir choisi parmi les récipiendaires de la bourse d'excellence FARE de la faculté des sciences.

À mes parents et sœurs, ainsi que toute ma famille et amis en Algérie, ma plus profonde reconnaissance pour votre soutien et vos encouragements.

Je ne saurais terminer sans remercier les professeurs du Département d'informatique et le personnel de l'UQAM qui ont contribué de près ou de loin à ma formation. J'exprime, enfin, mes plus vifs remerciements à mes collègues au sein du laboratoire LATECE, et en particulier, Choukri Djelalli.

DÉDICACE

À la mémoire de mon grand-père ...

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

L'analyse des graphes complexes, aussi appelés réseaux multidimensionnels ou réseaux multiplex, est l'un des nouveaux défis apparus en forage de données. Contrairement à la représentation classique de graphes où deux nœuds sont reliés par le biais d'une simple liaison, deux nœuds dans un réseau multidimensionnel se connectent par un ou plusieurs liens décrivant chacun une interaction spécifique dans une dimension particulière. Une des problématiques fondamentales étudiées dans ce domaine est la détection de communautés. Le but est de découvrir les sous-ensembles de nœuds densément connectés ou fortement interactifs, souvent, associés à des caractéristiques organisationnelles et fonctionnelles non connues à priori.

Bien qu'elle ait fait l'objet de nombreuses études dans le contexte unidimensionnel, la détection de communautés dans les réseaux multidimensionnels demeure une question de recherche ouverte. C'est d'une part en raison des complexités inhérentes à ce type de réseaux et d'autre part, la conséquence de l'absence d'une définition universellement reconnue pour le concept de communauté multidimensionnelle. En dépit du nombre croissant de travaux abordant cette problématique, certains aspects demeurent peu ou pas abordés dans la littérature. En effet, les approches existantes souffrent d'au moins un des problèmes suivants : (1) La difficulté de fixer des valeurs propres aux paramètres d'entrée, (2) la sensibilité aux dimensions non pertinentes, et (3) l'incapacité de découvrir les sous-espaces de dimensions pertinentes associés aux communautés détectées.

Afin de pallier les limites des approches existantes, nous présentons dans le cadre de ce mémoire une nouvelle approche de détection de communautés dans les réseaux multidimensionnels. Axée sur le principe de propagation d'étiquettes, l'approche développée vise l'identification automatique des structures denses dans les différents sous-espaces de dimensions, de même que leurs dimensions pertinentes via la maximisation d'une nouvelle fonction objective. L'efficacité de l'approche proposée est comparée à d'autres méthodes récentes par le biais d'une étude empirique détaillée sur différents réseaux synthétiques et réels. Les résultats obtenus démontrent la capacité de notre approche à identifier les communautés qui existent même dans des sous-espaces de faibles dimensions.

Mots clés : Détection de communautés, Réseaux multidimensionnels, Clustering.

CHAPITRE I

INTRODUCTION

1.1 Mise en contexte

Un réseau (aussi appelé graphe) est une représentation abstraite des interactions entre les entités d'un système complexe. En modélisant les acteurs comme des nœuds et les interactions comme des arêtes, plusieurs applications réelles notamment issues du Web social et de la biologie, peuvent être décrites au moyen d'un réseau. Fondée sur les outils de la théorie des graphes et la fouille de données, l'analyse de modèles d'interaction dans les réseaux complexes ne cesse de susciter l'attention dans les milieux académiques et industriels de nos jours (Tang *et al.*, 2012).

Une des tâches fondamentales de l'analyse des réseaux est la détection de communautés où le but est de découvrir les sous-ensembles de nœuds densément connectés souvent associés à des caractéristiques organisationnelles ou fonctionnelles non connues *à priori*. Les communautés, aussi appelées clusters ou modules, peuvent servir différentes tâches. L'analyse des réseaux sociaux et biologiques (Girvan et Newman, 2002) ; le marketing viral (Richardson et Domingos, 2002) ; la visualisation (Kang *et al.*, 2007) ou la découverte des fonctions biologiques au sein de réseaux d'interactions de protéines (Chen et Yuan, 2006; Rives et Galitski,

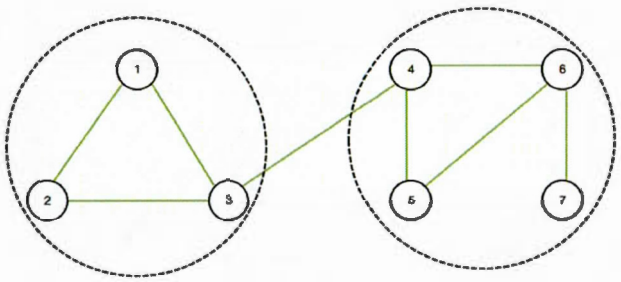


Figure 1.1: Un réseau avec deux communautés.

2003) sont des exemples d'application des algorithmes de détection de communautés. La figure 1.1 illustre un petit réseau constitué de deux communautés. Les nœuds dans chaque communauté sont densément connectés comparativement aux connexions inter communautés.

La détection de communautés a reçu beaucoup d'attention au cours de la dernière décennie et de nombreuses approches ont été proposées (Fortunato, 2010). Cependant, la majorité des approches existantes traitent principalement des réseaux unidimensionnels, à savoir, les réseaux où l'on ne trouve qu'au plus une connexion entre n'importe quelle paire de nœuds (Tang *et al.*, 2012). Bien que largement utilisée, cette représentation standard semble être inadéquate pour la description des interactions du monde réel qui sont plus complexes. En effet, les entités d'un réseau peuvent s'engager dans différents types d'interactions en même temps. À titre d'exemple, deux amis au sein d'un réseau social peuvent également travailler pour la même société. De même, dans un réseau de collaboration, deux scientifiques peuvent se connecter à travers une multitude de liaisons indiquant les conférences ou les revues dans lesquelles ils ont copublié des articles. Puisqu'ils ne permettent qu'une seule connexion entre deux nœuds, les réseaux unidimensionnels se révèlent donc incompatible pour décrire ce genre d'interactions.

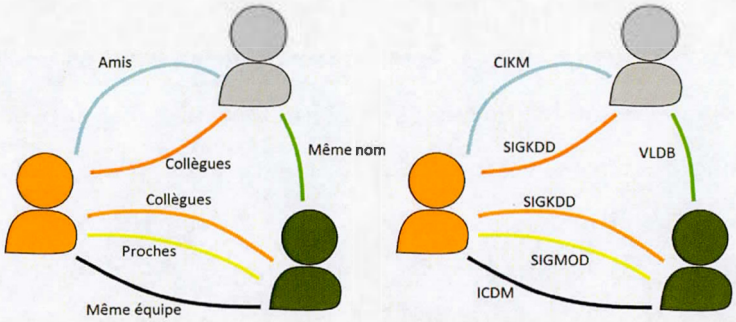


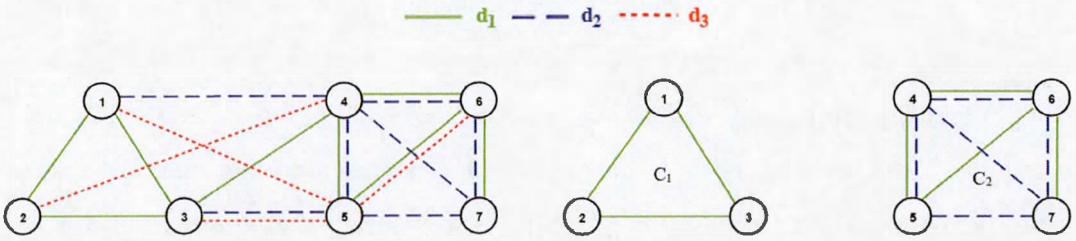
Figure 1.2: Un exemple de réseaux multidimensionnels.

Les réseaux multidimensionnels ont récemment été proposés comme une alternative pour mieux décrire les interactions au sein des systèmes complexes (Berlingerio *et al.*, 2013a). Dans ce type de réseaux, aussi appelés les réseaux multicouches (De Domenico *et al.*, 2015a), ou les réseaux multiplex (Mucha *et al.*, 2010), les relations sont qualifiées en fonction de leurs types de telle sorte qu'il existe différentes connexions entre chaque paire de nœuds. Autrement dit, contrairement à un réseau unidimensionnel, dans un réseau multidimensionnel, deux nœuds peuvent être reliés par un ou plusieurs liens (ou connexions). Chaque lien appartient à une dimension spécifique. Aux fins d'illustration, la figure 1.2, tirée de (Berlingerio *et al.*, 2013a), montre deux réseaux multidimensionnels de cinq dimensions chacun. Dans cette figure, chaque type de relation décrit une dimension d'interaction particulière. À noter qu'un réseau multidimensionnel peut être représenté par une série de réseaux unidimensionnels, et ce en réalisant une projection des nœuds à travers la même dimension.

La détection de communautés dans les réseaux multidimensionnels est une problématique de recherche relativement récente, bien que celle-ci ait été largement étudiée dans le contexte des réseaux unidimensionnels. Ici, il convient de rappeler qu'une définition de la notion de communauté multidimensionnelle est loin de faire

l'unanimité chez les chercheurs. Un certain nombre de travaux (Amelio et Pizzuti, 2014; Tang *et al.*, 2012) la considèrent comme un groupe de nœuds densément connectés à travers l'ensemble des dimensions du réseau. Ainsi, une communauté formée dans une dimension est censée exister sur chacune des dimensions restantes. Une telle définition semble, cependant, être trop contraignante, car, en réalité, on n'attribue pas toujours la même importance aux différents types d'interactions entre les entités (Nicosia et Latora, 2014). En effet, un nœud actif sur une dimension peut rester sans activité sur le reste des dimensions. Dans un travail récent, Nicosia et Latora (2014) montrent que seulement une petite fraction des nœuds du réseau s'implique activement à travers la totalité de ses dimensions. Par conséquent, le fait d'exiger l'existence d'une communauté sur toutes les dimensions peut entrainer une perte d'information substantielle quant à l'organisation réelle du système modélisé.

Quelques travaux (Boden *et al.*, 2012; Papalexakis *et al.*, 2013) ont, cependant, suggéré de considérer l'importance, relative, des dimensions dans la formation des communautés. Ainsi, l'existence d'une communauté est limitée à un sous-ensemble de dimensions. En d'autres termes, une communauté peut exister dans un sous-espace de dimensions plus petit que l'espace en entier. Spécifiquement, chaque communauté C_k se définie par un couple (V_k, D_k) où V_k représente l'ensemble des nœuds formant C_k et D_k l'ensemble des dimensions où C_k existe. Les nœuds formant la communauté C_k ont tendance à être plus densément connectés dans toutes les dimensions dans D_k qu'ailleurs dans le réseau. Les dimensions D_k sont appelées les *dimensions pertinentes* de C_k. Les dimensions restantes (c.-à-d., les dimensions qui n'appartiennent pas à D_k) sont appelées les *dimensions non pertinentes* de C_k. Ainsi, une dimension peut être pertinente pour zéro, une ou plusieurs communautés. Afin d'illustrer ce point, considérons le réseau multidimensionnel



(a) Un réseau de 3 dimensions. (b) La partition multidimensionnelle.

Figure 1.3: Un réseau multidimensionnel de 3 dimensions avec deux communautés encastrées dans différents sous-espaces de dimensions.

présenté par la figure 1.3a. Deux communautés peuvent être identifiées dans ce réseau. Telles qu'illustrées dans la figure 1.3b, une première communauté possible est $C_1 = (V_1, D_1) = (\{n_1, n_2, n_3\}, \{d_1\})$ tandis qu'une deuxième communauté possible est $C_2 = (V_2, D_2) = (\{n_4, n_5, n_6, n_7\}, \{d_1, d_2\})$. Les nœuds formant C_2 sont densément connectés à travers les dimensions d_1 et d_2 tandis que C_1 est seulement définie dans d_1. D'autre part, on peut observer à partir de la figure 1.3a que la dimension d_3 ne contribue pas à la formation des deux communautés C_1 et C_2, ce qui fait d'elle une dimension non pertinente. Pour résumer, on constate que d_1 est pertinente, à la fois, pour C_1 et C_2, d_2 est uniquement pertinente pour C_2 et d_3 n'est pertinente pour aucune communauté.

1.2 Motivations

La détection de communautés multidimensionnelles est une tâche difficile pour laquelle un certain nombre d'approches ont été proposées. Parmi les travaux existants, certains (Berlingerio *et al.*, 2011; Tang *et al.*, 2012; Cai *et al.*, 2005; Kun *et al.*, 2014) proposent des stratégies d'agrégation visant à transformer le réseau multidimensionnel en un réseau unidimensionnel. Ce dernier est ensuite analysé par des algorithmes standards de détection de communautés. D'autres travaux

(Tang *et al.*, 2012; Amelio et Pizzuti, 2014; Berlingerio *et al.*, 2013a) adoptent des stratégies consensuelles pour ressortir la partition latente en partant des partitions identifiées indépendemment sur chaque dimension. En revanche, quelques techniques récentes suggèrent d'adapter les algorithmes unidimensionnels classiques au contexte multidimensionnel. À titre d'exemple, une approche inspirée de l'algorithme WalkTrap (Pons et Latapy, 2005) a été présentée dans (Kuncheva et Montana, 2015) tandis qu'une extension de l'algorithme Infomap (Rosvall et Bergstrom, 2008) a été proposée dans (De Domenico *et al.*, 2015a). De même, une généralisation de la modularité de Newman (2006) a aussi été introduite dans (Mucha *et al.*, 2010).

En dépit de ces travaux, l'identification des communautés dans les réseaux multidimensionnels continue de défier les méthodes existantes. En fait, chacune des approches existantes souffre d'une ou de plusieurs des limitations suivantes :

1. De nombreuses approches rencontrent des difficultés lorsqu'elles sont appliquées à des réseaux ayant des dimensions non pertinentes. En effet, les dimensions non pertinentes n'offrent aucune structure de communautés apparente. Par conséquent, la présence de ce type de dimensions affecte les caractéristiques structurelles des autres dimensions ce qui rend la tâche de détection de communautés plus difficile. Plusieurs approches, notamment celles qui adoptent les stratégies d'agrégation, sont grandement affectées en présence de ce type de dimensions. Cet effet est d'autant plus prononcé si les communautés existent dans des sous-espaces de faibles dimensions, c.-à-d., des communautés avec un nombre très restreint de dimensions pertinentes comparativement au nombre total de dimensions du réseau.
2. La majorité des approches existantes ne permettent pas l'identification explicite des dimensions pertinentes associées aux communautés détectées.

En dépit de son importance pour de nombreuses applications, cet aspect demeure peu étudié dans la littérature. Bien que certaines approches (Papalexakis *et al.*, 2013; Cai *et al.*, 2005) offrent une estimation numérique pour la pertinence des dimensions aux communautés détectées, un mécanisme de discrimination explicite entre les ensembles pertinents et non pertinents n'a toujours pas été élaboré.

3. La grande majorité des approches existantes souffrent de leurs dépendances à un certain nombre de paramètres d'entrée. Dans ce contexte, il est clair que la précision de la détection de communautés est étroitement liée à un réglage adéquat de ces paramètres. Si l'utilisateur fournit des valeurs inappropriées, la précision sera grandement diminuée. En outre, de nombreuses méthodes exigent que le nombre de communautés soit fixé à l'avance. Cependant, il est rarement possible pour l'utilisateur de fournir le nombre exact de communautés dans un réseau. En fait, fournir des valeurs exactes des paramètres d'entrée, incluant le nombre de communautés, exige des connaissances *à priori* du réseau à analyser. Or, en pratique, de telles connaissances ne sont pas toujours disponibles.

1.3 Contributions

Afin de pallier les limitations mentionnées précédemment, nous proposons dans le cadre de ce mémoire une nouvelle approche de détection de communautés dans les réseaux multidimensionnels. L'approche proposée (ci-après nommée MDLPA pour MultiDimensional Label Propagation Algorithm) est entièrement automatisée et ne nécessite aucun réglage de paramètres ou connaissances préalables pour identifier les communautés et leurs dimensions pertinentes.

Le travail présenté dans ce mémoire est marqué par les éléments suivants :

1. Nous considérons l'identification de communautés dans les réseaux multidimensionnels comme une tâche d'optimisation d'une fonction objective. La fonction proposée implique la sélection des dimensions pertinentes dans l'évaluation de la qualité des partitions recherchées. Le processus d'optimisation développé s'inspire du principe de propagation d'étiquettes en se basant sur un mécanisme de pondération permettant de guider la recherche de la partition optimale.

2. L'approche proposée ne nécessite pas de paramétrage dans le sens où elle détecte les communautés et leurs dimensions pertinentes d'une manière entièrement automatisée. L'utilisateur n'aura pas à effectuer un réglage de paramètres ou à fournir le nombre de communautés à identifier.

3. Les expérimentations menées sur les réseaux synthétiques et réels démontrent la performance de notre approche par rapport à d'autres techniques récentes qui ont l'avantage d'utiliser des connaissances préalables (tel que le nombre de communautés) sur les réseaux analysés. En outre, les résultats obtenus montrent que l'approche proposée offre une performance compétitive dans les situations qui impliquent l'existence de communautés dans des sous-espaces de faibles dimensions.

1.4 Structure et organisation du mémoire

Le reste de ce document est organisé comme suit : le chapitre 2 présente une revue des principaux travaux qui se rattachent à la détection de communautés dans les réseaux multidimensionnels. En particulier, nous passons en revue cinq familles différentes d'algorithmes. Pour chaque famille, nous mettons l'accent sur quelques techniques représentatives et nous discutons de leurs avantages et inconvénients. Le chapitre 3 décrit l'approche proposée en détails. Spécifiquement, dans ce chapitre, nous présentons la fonction objective utilisée et les considérations

théoriques qui ont conduit à sa définition. Par la suite, nous décrivons la procédure d'optimisation développée ainsi que le mécanisme de pondération qui s'y rattache. Finalement, nous consacrons la dernière partie du chapitre 3 à l'analyse de la complexité de l'approche proposée. Le chapitre 4 présente une évaluation empirique détaillée de notre approche. Précisément, dans ce chapitre, nous mènons des expérimentations sur différents réseaux synthétiques et réels où nous comparons la performance de l'approche à d'autres techniques récentes. Finalement, le chapitre 5 conclut ce mémoire.

CHAPITRE II

REVUE DE LA LITTÉRATURE

Ce chapitre présente un survol des principaux algorithmes de détection de communautés dans les réseaux multidimensionnels. Spécifiquement, il donne précisément une vue d'ensemble sur leurs modes de fonctionnement, les types de communautés qu'ils ciblent ainsi que leurs avantages et inconvénients.

D'une manière générale, les travaux qui abordent la problématique de détection de communautés dans les réseaux multidimensionnels peuvent être classés en cinq grandes familles. En fonction du modèle exploratoire, on distingue les approches à base : (1) d'agrégation de dimensions ; (2) d'intégration de partitions ; (3) d'intégration de caractéristiques structurelles ; (4) d'exploration simultanée de dimensions ; et (5) de décomposition tensorielle. Dans ce qui suit, nous allons porter notre attention sur quelques travaux représentatifs dans chaque famille d'algorithmes.

2.1 Les approches à base d'agrégation de dimensions

Ces approches supposent qu'une communauté existe sur toutes les dimensions du réseau. Ainsi, il est possible d'identifier la partition finale en procédant à une transformation du réseau multidimensionnel à un réseau unidimensionnel. L'idée de base consiste à effectuer une agrégation des dimensions sur un réseau unidi-

mensionnel. Ce dernier peut ensuite être analysé par un algorithme de détection de communautés classique.

Le principe consiste à remplacer l'ensemble des liens reliant chaque paire de nœuds par une seule connexion pondérée sur le graphe agrégé. Dans notre revue de littérature, nous avons identifié deux types de stratégies d'agrégation : l'agrégation naïve, où l'on ne tient pas compte de la pertinence des dimensions dans le processus d'agrégation, et l'agrégation à base d'apprentissage où la pertinence des dimensions dans la formation des communautés est considérée.

2.1.1 L'agrégation naïve

Dans cette catégorie d'approches, l'agrégation s'effectue d'une manière *ad hoc* sans prise en compte du degré d'implication des dimensions dans la formation des communautés. Généralement, on peut distinguer trois schémas d'agrégation (Kanawati, 2013) :

- ***L'agrégation binaire*** (Berlingerio *et al.*, 2011) : cette technique consiste à remplacer l'ensemble des connexions reliant une paire de nœuds par une simple liaison.
- ***L'agrégation fréquentielle*** (Berlingerio *et al.*, 2011) : le but ici est de remplacer l'ensemble des connexions entre chaque paire de nœuds par une seule connexion à laquelle on associe un poids représentant le nombre de liens reliant les deux nœuds en question.
- ***L'agrégation par similarité*** (Berlingerio *et al.*, 2011) : cette technique est similaire à l'agrégation fréquentielle sauf que le poids de connexion sur le graphe agrégé est estimé en fonction de la similarité entre les voisinages des deux nœuds connectés. Généralement, un poids élevé reflète une grande proportion de voisins partagés à travers l'ensemble des dimensions du ré-

seau (Loe et Jensen, 2015).

En dépit de leur simplicité d'implémentation, les approches basées sur l'agrégation naïve (ce qui inclus l'agrégation : binaire, fréquentielle et par similarité) souffrent de trois limitations :

1. La sensibilité aux dimensions non pertinentes. En effet, la considération des liens appartenant aux dimensions non pertinentes dans le calcul des poids peut masquer la partition latente du réseau analysé. En outre, l'efficacité de ces approches est grandement affectée si le nombre de dimensions non pertinentes est élevé.
2. La perte d'information induite par la compression. L'identification des dimensions pertinentes associées aux communautés détectées n'est pas possible avec ces stratégies.
3. La dépendance à l'algorithme de détection de communautés appliqué sur le réseau agrégé. Ici, il est évident qu'il peut y avoir des résultats différents en fonction du choix de l'algorithme et de son paramétrage.

2.1.2 L'agrégation par apprentissage

Ces techniques se distinguent des approches précédentes par leur capacité à prendre en charge la pertinence des dimensions dans le processus d'agrégation. Généralement, on estime un coefficient qui mesure l'importance de la dimension en question dans la formation des communautés. Ce coefficient est ensuite utilisé dans le calcul des poids de connexions sur le graphe agrégé.

Une première approche qui adopte l'agrégation par apprentissage a été présentée dans (Cai *et al.*, 2005). Cette approche estime, pour chaque dimension, un coefficient de pertinence en utilisant la régression linéaire sous contraintes. Le

principe se base sur la minimisation d'erreur quadratique moyenne des distances séparant les dimensions de la représentation agrégée optimale. Les coefficients obtenus sont ensuite exploités dans le calcul des poids des connexions sur le graphe agrégé à travers la stratégie d'agrégation fréquentielle. Ainsi, la partition finale est principalement définie par les dimensions ayant des partitions semblables. Bien qu'elle limite l'impact des dimensions non pertinentes, cette approche exige que les contraintes d'affiliations des nœuds soient fournies à l'avance.

Kun *et al.* (2014) proposent une technique d'agrégation inspirée de l'apprentissage ensembliste par le *boosting*, intitulée LBGA (Locally Boosted Graph Aggregation). Cette approche combine une fonction de qualité avec un algorithme de clustering dans un système de récompense. Ce dernier vise à sélectionner les meilleures arêtes pour la formation des communautés sur le graphe agrégé. L'idée est d'utiliser la métrique de qualité pour évaluer les partitions obtenues vis-à-vis les arêtes sélectionnées pour le calcul des poids. On attribue ensuite un score à chaque arête sélectionnée en fonction de la qualité de la partition obtenue. L'objectif est de pénaliser les arêtes qui limitent la séparation des communautés sur le graphe agrégé. Ce système de récompense permet à l'algorithme de générer des graphes plus modulaires où les communautés sont plus apparentes. LBGA est sensible à certains paramètres, notamment, le taux d'apprentissage.

2.2 Les approches fondées sur l'intégration de partitions

Contrairement aux approches précédentes, les techniques fondées sur l'intégration de partition visent à combiner les partitions identifiées séparément sur les dimensions du réseau. Principalement, on distingue deux types de stratégies d'intégration de partitions : l'intersection et le consensus. Dans la première, l'objectif est de repérer les zones de chevauchement entre les communautés unidimension-

nelles tandis que la deuxième vise à dégager une partition de consensus à partir des partitions identifiées sur les dimensions du réseau. Dans ce qui suit, nous allons présenter trois approches qui adoptent ces deux stratégies d'intégration.

2.2.1 ABACUS : Intersections par fouille de motifs fermés fréquents

Proposée par (Berlingerio *et al.*, 2013b), ABACUS (frequent pAttern mining-BAsed Community discovery in mUltidimensional networkS) est une approche visant l'identification des chevauchements entre les communautés unidimensionnelles identifiées sur les dimensions d'un réseau multidimensionnel. C'est une approche qui exploite les techniques de fouille de motifs fermés fréquents en partant des adhésions aux communautés identifiées sur chacune des dimensions. L'idée est de représenter chaque nœud sous forme de transaction, ou itemset, portant sur ses affiliations aux communautés unidimensionnelles. La base transactionnelle résultante est ensuite évaluée à travers un algorithme de fouille d'itemset fermés fréquents (Borgelt, 2003). Dans ce modèle, l'ensemble de support d'un itemset fermé fréquent définit une zone de chevauchement et ainsi, une communauté multidimensionnelle. La problématique de détection de communautés se réduit alors à un problème de fouille d'itemsets fermés fréquents. ABACUS dépend du support minimal, un paramètre qui définit le nombre minimal de nœuds nécessaires dans un chevauchement pour qu'il soit retenu comme une communauté. Ce paramètre est critique pour l'algorithme puisqu'il détermine le nombre de communautés qu'il retourne.

2.2.2 Consensus par clustering d'ensembles

Les techniques de clustering d'ensembles visent à générer une partition de consensus en partant d'un ensemble de partitions issues du même jeu de données. Ces dernières sont généralement identifiées par le biais d'un algorithme non dé-

terministe (exécuté à plusieurs reprises), ou à travers diverses approches. Dans le contexte multidimensionnel, Tang *et al.* (2012) proposent d'adopter la même stratégie en considérant les partitions identifiées sur les dimensions du réseau.

Les techniques de clustering d'ensembles exploitent les appartenances aux communautés comme une métrique de distance internœuds. Le principe consiste à évaluer la similarité entre une paire de nœuds en fonction de la proportion d'affiliations partagées sur l'ensemble des partitions. La matrice de similarité estimée est ensuite évaluée à travers un algorithme de partitionnement classique de sorte que les nœuds suffisamment proches finissent dans les mêmes communautés de consensus. Plusieurs travaux se sont penchés sur cette problématique et de nombreuses approches ont été proposées telles que : CSPA (Cluster-Based Similarity Partitioning Algorithm), HGPA (Hypergraph Partition Algorithm), et MCLA (Meta-Clustering Algorithm) (Strehl et Ghosh, 2003).

Les techniques de clustering d'ensembles ont été adoptées pour la détection de communautés dans le contexte multidimensionnel en se basant sur l'hypothèse d'existence d'une partition sur chacune des dimensions du réseau (Tang *et al.*, 2012). Toutefois, comme les dimensions non pertinentes n'offrent pas de structures de communautés apparentes, la partition de consensus peut en être affectée significativement. De ce fait, ces stratégies risquent de ne pas produire les résultats souhaités en présence de ce type de dimensions.

2.2.3 Consensus par optimisation multiobjectif

Dans (Amelio et Pizzuti, 2014), les auteurs introduisent MultiMOGA (Multidimensional Multi-Objective Genetic Algorithm), une approche consensuelle basée sur l'optimisation multiobjectif. L'approche adopte une stratégie visant à évaluer

les dimensions selon un classement préétabli. Plus spécifiquement, la partition identifiée à partir de la dimension du rang i est utilisée comme référence pour optimiser deux fonctions objectives sur la prochaine dimension : la qualité de partitionnement et le coût de partage SC, une métrique qui définit la similarité entre la partition de référence et celle nouvellement obtenue. Pour y parvenir, l'approche proposée adopte la modularité de Newman (2006) et l'information mutuelle normalisée comme fonctions de qualité dans un algorithme génétique multiobjectif. La limitation majeure de MultiMOGA réside dans l'absence d'une fonction de classement adéquate. L'utilisateur doit spécifier l'ordre selon lequel les dimensions sont évaluées. Par ailleurs, l'approche suppose l'existence d'une partition sur chaque dimension du réseau. Cette hypothèse s'avère non réaliste compte tenu du fait que dans un réseau multidimensionnel, il peut y avoir des dimensions non pertinentes qui ne relatent aucune structure de communauté.

2.3 Les approches basées sur l'intégration de caractéristiques

Les techniques d'intégration de caractéristiques sont semblables aux stratégies précédentes dans le sens où elles abordent les dimensions d'une façon individuelle. Cependant, la différence réside dans le type d'information extraite et agrégée. Ces approches visent en effet à combiner les caractéristiques structurelles séparément identifiées à partir des dimensions du réseau. Le principe consiste à effectuer un changement de représentation vers un espace Euclidien où les nœuds s'identifient à travers un système de coordonnées. La transformation est faite de sorte que les membres d'une communauté soient proches dans le nouvel espace. Ce changement de représentation permet aux algorithmes de clustering classiques (tel que l'algorithme k-means par exemple) de mettre en évidence la partition latente en examinant les coordonnées des points. Globalement, cette famille d'approches est considérée comme une extension des techniques de partitionnement spectral dans les réseaux unidimensionnels.

Tang *et al.* (2009a) introduisent la méthode PMM (Principal Modularity Maximization), une approche qui combine les caractéristiques spectrales des matrices de modularité des dimensions. Dans un premier temps, PMM construit pour chaque dimension sa matrice de modularité. Par la suite, les ℓ vecteurs propres ayant les valeurs propres les plus élevées sont sélectionnés en utilisant la décomposition en valeurs singulières. La représentation générée est ensuite évaluée à travers l'algorithme k-means. D'autres stratégies similaires (Tang *et al.*, 2009b; Dong *et al.*, 2012, 2014) ont récemment été proposées. À titre d'exemple, Dong *et al.* (2014) introduisent la méthode SC-ML (Spectral Clustering on Multilayer Graphs), une technique qui exploite les spectres des matrices Laplaciennes normalisées. L'inconvénient majeur de ces approches réside dans le besoin de spécifier le nombre de communautés recherchées (en raison de l'utilisation de l'algorithme k-means).

2.4 L'exploration simultanée de dimensions

Les approches d'exploration simultanée de dimensions ont été proposées pour permettre aux techniques conventionnelles de supporter la détection de communautés sur les réseaux multidimensionnels. Ces approches se distinguent des techniques précédentes par leur capacité à explorer les dimensions simultanément.

Mucha *et al.* (2010) introduisent une version multidimensionnelle de la modularité de Newman (2006), qui est considérée, à la fois, comme une métrique de qualité et une fonction objective dans plusieurs techniques conventionnelles. La métrique multidimensionnelle proposée, aussi appelée la modularité généralisée ou la modularité multicouche, donne la possibilité aux techniques d'optimisation classiques d'opérer sur les réseaux multidimensionnels. Nous pouvons citer le cas de l'approche introduite dans (Carchiolo *et al.*, 2011), où les auteurs proposent une stratégie d'optimisation inspirée de la méthode Louvain (Blondel *et al.*, 2008).

Dans un travail récent, De Domenico *et al.* (2015a) introduisent Multiplex Infomap, une approche qui s'inspire de l'algorithme unidimensionnel Infomap (Rosvall et Bergstrom, 2008). L'approche modélise le flux d'information sous forme d'une marche aléatoire. Cette dernière vise à minimiser une version modifiée de l'équation de carte de Infomap. LART (Locally Adaptive Random Walks) (Kuncheva et Montana, 2015) est une autre technique d'exploration simultanée de dimensions qui s'inspire de l'algorithme unidimensionnel WalkTrap (Pons et Latapy, 2005). Spécifiquement, LART utilise des probabilités de transition qui permettent à un marcheur aléatoire d'explorer le réseau dans la même dimension ou à travers différentes dimensions. La mise à jour des probabilités s'effectue selon les similarités topologiques des dimensions au niveau des nœuds. Le but est de favoriser les sauts entre les membres d'une même communauté. Une métrique de similarité est ensuite utilisée afin de regrouper les nœuds de façon hiérarchique. LART et Multiplex Infomap peuvent découvrir les communautés qui existent dans les différents sous-espaces de dimensions. Toutefois, la limitation majeure de ces deux algorithmes réside dans le réglage de la longueur de marche aléatoire et le taux de relâchement. Ces deux paramètres peuvent grandement affecter la précision des deux approches si les valeurs fournies par l'utilisateur sont incorrectes.

Hmimida et Kanawati (2015) introduisent mux-LICOD, une version multidimensionnelle de l'approche LICOD (Leaders Identification for Community Detection in complex networks) (Kanawati, 2014). mux-LICOD détecte les communautés en identifiant un sous-ensemble de nœuds centraux dans le réseau. Ces derniers sont repérés en utilisant une mesure appelée le degré de centralité. L'idée consiste à sélectionner les nœuds pour lesquels la centralité est 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. Les nœuds sont en-

suite affectés aux nœuds centraux les plus proches selon une métrique de distance géodésique. mux-LICOD peut estimer le nombre de communautés d'une manière automatique. Cependant, c'est un algorithme qui pâtit de sa dépendance à deux paramètres, en l'occurrence, le seuil de similarité et le seuil de voisinage. Une valeur non appropriée de ces seuils affecte grandement la qualité des résultats.

Boden *et al.* (2012) présentent l'approche MiMAG (Mining Multi-layered Attributed Graphs), une technique basée, en partie, sur l'algorithme Quick (Liu et Wong, 2008). L'approche vise à identifier les 0.5-quasi-cliques dans les différents sous-espaces de dimensions, à savoir, les communautés où la densité des liens dépasse 0.5, pour chaque dimension pertinente. Le principe consiste à énumérer les communautés candidates dans un arbre en utilisant un parcours en profondeur. Par la suite, chaque communauté est évaluée pour la propriété de 0.5-quasi-clique. En dépit de sa capacité à identifier les dimensions pertinentes associées aux communautés détectées, la contrainte de quasi-clique adoptée par l'approche l'empêche de découvrir des communautés de faible densité. Par ailleurs, MiMAG est limitée à l'identification des communautés existantes dans des sous-espaces d'au moins deux dimensions et ne considère que des communautés de 8 nœuds ou plus.

2.5 La décomposition tensorielle

La détection de communautés dans les réseaux multidimensionnels a également été étudiée par le biais des techniques de décomposition tensorielle (Dunlavy *et al.*, 2011; Papalexakis *et al.*, 2013; Li *et al.*, 2014). En fait, un réseau multidimensionnel peut être représenté à l'aide d'un tenseur où chaque coupe correspond à une matrice d'adjacence d'une dimension. L'utilisation d'un tenseur pour la représentation d'un réseau multidimensionnel permet aux techniques de décomposition tensorielle (Kolda et Bader, 2009) d'identifier la partition latente.

Quelques travaux récents (Dunlavy *et al.*, 2011; Papalexakis *et al.*, 2013; Li *et al.*, 2014) ont adopté le modèle de décomposition tensorielle pour s'attaquer à la problématique de détection de communautés multidimensionnelles. Papalexakis *et al.* (2013) présentent GraphFuse, une technique qui vise à identifier les superpositions de lignes, colonnes et fibres d'un tenseur. L'approche mesure la pertinence des dimensions aux communautés détectées à travers un système de pondération. Toutefois, GraphFuse requiert un paramétrage adéquat vu que c'est une approche qui dépend du facteur de pénalité et du nombre de communautés recherchées.

En terminant, il est important de noter que parmi les algorithmes présentés dans ce chapitre, certains tels que Louvain multicouche (Carchiolo *et al.*, 2011), LART (Kuncheva et Montana, 2015), Multiplex Infomap (De Domenico *et al.*, 2015a), et ABACUS (Berlingerio *et al.*, 2013b)) produisent des partitions où les communautés peuvent partager les nœuds sur différents sous-espaces de dimensions. Autrement dit, un nœud peut appartenir, à la fois, à plusieurs communautés, et ce, dans différents sous-espaces de dimensions. Bien que ces approches peuvent découvrir des structures intéressantes, le grand nombre de communautés générées peut rendre l'interprétation des résultats difficile. Dans ce mémoire, nous nous concentrons sur les algorithmes qui produisent des partitions disjointes, c.-à-d., des partitions où un nœud appartient à une et une seule communauté. Nous croyons que ce type de partitionnement permet de fournir des structures de communautés qui sont facilement interprétables par l'utilisateur.

2.6 Conclusion

Dans ce chapitre, nous avons présenté un bref survol des principaux travaux qui portent sur la détection de communautés dans les réseaux multidimensionnels. À titre de rappel, nous présentons dans le tableau 2.1 un résumé des caractéris-

tiques majeures des approches existantes. Comme on peut le constater, aucune approche n'est complètement libre de paramétrage. En outre, l'identification explicite des dimensions pertinentes associées aux communautés détectées demeure une fonctionnalité manquante dans la majorité des approches déjà proposées. Afin de pallier ces limites, nous présentons dans le chapitre suivant, une nouvelle approche de détection de communautés les réseaux multidimensionnels.

Tableau 2.1: Liste des approches de détection de communautés dans les réseaux multidimensionnels.

Approche	Nécessite un paramétrage	Découvre les dimensions pertinentes
Agrégation binaire	Oui	Non
Agrégation fréquentielle	Oui	Non
Agrégation par similarité	Oui	Non
Combinaison par régression linéaire	Oui	Non
Combinaison par boosting (LBGA)	Oui	Non
Clustering d'ensembles	Oui	Non
ABACUS	Oui	Oui
MultiMOGA	Oui	Non
PMM	Oui	Non
SC-ML	Oui	Non
Louvain Multicouche	Oui	Non
LART	Oui	Non
Multiplex Infomap	Oui	Oui
MiMAG	Oui	Oui
Mux-LICOD	Oui	Non
GraphFuse	Oui	Non

CHAPITRE III

APPROCHE PROPOSÉE

Ce chapitre présente une nouvelle approche de détection de communautés dans les réseaux multidimensionnels. L'approche proposée est basée sur l'optimisation d'une fonction objective à travers un processus itératif qui s'inspire du principe de propagation d'étiquettes. La stratégie d'optimisation développée s'appuie sur un mécanisme de pondération qui permet de guider la recherche de la partition optimale. Dans ce qui suit, nous commençons par introduire quelques notations et définitions. Ensuite, nous présentons la fonction objective suivie par une description détaillée de la procédure d'optimisation développée ainsi que le mécanisme de pondération qui s'y rattache. Finalement, la dernière partie de ce chapitre est consacrée à l'analyse de la complexité de l'approche proposée.

3.1 Notation et concepts

Avant de décrire notre approche, nous présentons quelques notations et définitions. Comme dans (Berlingerio *et al.*, 2013a), nous utilisons les multigraphes pour représenter les réseaux multidimensionnels. Soit $G = (V, E, D)$ un multigraphe non orienté et non pondéré où V est un ensemble de n nœuds, D est un ensemble de nd dimensions et E est un ensemble de m arêtes, c.-à-d., l'ensemble de triplets (v, u, d) tel que $v, u \in V$ sont des nœuds et $d \in D$ est une dimension. Le triplet (v, u, d) indique que les nœuds v et u sont liés par une arête appartenant

à la dimension d. Ainsi, deux nœuds dans G peuvent être liés par, au plus, nd arêtes.

Notre objectif est d'identifier une partition disjointe du multigraphe G où chaque nœud $v \in V$ peut appartenir à une et une seule communauté $C_k = (V_k, D_k), k = 1...K$, tel que K est le nombre de communautés (non connu *à priori*) ; V_k est un sous-ensemble de V ; $D_k \subseteq D$ est un sous-ensemble de dimensions où les nœuds V_k sont densément (fortement) connectés. Rappelons que les dimensions appartenant à D_k sont appelées les dimensions pertinentes de la communauté C_k, tandis que les dimensions restantes, à savoir, $D - D_k$, sont appelées les dimensions non pertinentes pour C_k.

L'approche proposée vise à identifier des communautés disjointes qui satisfont les propriétés suivantes :

1. Une communauté $C_k = (V_k, D_k), k = 1...K$ est un sous-ensemble non vide de G.
2. Dans chaque communauté $C_k = (V_k, D_k)$, les nœuds dans V_k doivent être plus densément connectés à travers toutes les dimensions dans D_k qu'ailleurs dans le réseau. En d'autres termes, la densité de liens internes dans C_k doit être plus élevée que la densité de liens sortant de C_k.
3. Chaque ensemble D_k de C_k doit contenir un nombre suffisant de dimensions pertinentes qui permettent de distinguer les nœuds appartenant à C_k des autres nœuds de G. En d'autres termes, la densité de liens dans D_k doit être largement suppérieure à la densité de liens dans $D - D_k$.
4. Les sous-ensembles de dimensions $\{D_k\}, k = 1...K$ peuvent être ou ne pas être disjoints. De même, ces ensembles peuvent avoir des cardinalités dif-

férentes.

La raison de la première propriété est que l'on vise à partitionner le multigraphe G en un ensemble fini de communautés disjointes $C_k, k = 1 \ldots K$. La deuxième propriété assure que chaque communauté C_k se compose d'un ensemble de nœuds V_k densément connectés à travers D_k par rapport aux autres nœuds dans G. En d'autres termes, les nœuds qui forment C_k doivent être fortement liés à travers toutes les dimensions dans D_k. La troisième propriété stipule que chaque D_k doit seulement contenir les dimensions pertinentes qui aident à l'identification des membres de C_k. Finalement, la dernière propriété est basée sur le fait que, dans un réseau multidimensionnel, les communautés existent dans différents sous-espaces de dimensions. Dans cette optique, l'algorithme de détection doit également supporter le fait que les communautés peuvent exister dans les mêmes ou dans différents sous-espaces de dimensions. À titre illustratif, considérons la partition présentée dans la figure 1.3. Dans cette figure, on observe que les deux communautés $C_1 = (V_1, D_1) = (\{n_1, n_2, n_3\}, \{d_1\})$ et $C_2 = (V_2, D_2) = (\{n_4, n_5, n_6, n_7\}, \{d_1, d_2\})$ satisfont les quatre propriétés des communautés qu'on souhaite identifier.

3.2 Développement d'une fonction objective

Dans cette section, nous développons une fonction objective qui s'appuie sur le principe de propagation d'étiquettes pour la recherche d'une partition optimale de G. À cette fin, nous discutons d'abord, dans ce qui suit, le principe de propagation d'étiquettes ainsi que la fonction objective qui s'y rattache, et ce dans un contexte unidimensionnel. Par la suite, nous abordons la problématique de détection de communautés dans un contexte multidimensionnel, pour laquelle, nous présentons une nouvelle fonction objective. Une caractéristique notable de la fonction objective développée est qu'elle considère à la fois la pertinence des

dimensions et l'appartenance à une communauté dans le processus de recherche de la partition optimale dans l'espace multidimensionnel.

Les techniques qui se basent sur le principe de propagation d'étiquettes (Barber et Clark, 2009; Raghavan *et al.*, 2007; Leung *et al.*, 2009; Liu et Murata, 2010) représentent l'une des principales familles d'algorithmes de détection de communautés dans les réseaux unidimensionnels. L'approche de propagation d'étiquette, aussi appelée LPA (Label Propagation Algorithm), s'appuie sur une stratégie de recherche locale qui exploite la structure topologique du réseau pour identifier les communautés. Le principe consiste à attribuer à chaque nœud du réseau une étiquette numérique représentant la communauté à laquelle il appartient. Au démarrage, chaque nœud est placé dans une communauté à part. Puis, chaque étiquette associée à un nœud est mise à jour en fonction de l'étiquette dominante dans le voisinage du nœud en question. La dominance est dictée par une règle de propagation spécifique à la variante de l'algorithme. À titre d'exemple, la version basique de LPA (Raghavan *et al.*, 2007) sélectionne l'étiquette portée par la majorité des voisins. Le processus de réétiquetage se répète jusqu'à ce qu'une étiquette dominante soit attribuée à chaque nœud. Les communautés sont ensuite extraites à partir des nœuds portant la même étiquette. LPA est simple à mettre en œuvre, rapide et ne nécessite pas de paramétrage, incluant le nombre de communautés, ce qui le rend très pratique dans de nombreux contextes.

Dans un travail récent, Barber et Clark (2009) proposent une formulation mathématique du principe de LPA (Raghavan *et al.*, 2007). L'idée fondamentale se base sur le fait que les communautés identifiées par LPA sont le résultat d'un processus de maximisation d'une fonction objective. Formellement, pour un graphe

non orienté $G' = (V, E)$, la fonction objective optimisée par LPA est définie par :

$$F = \frac{1}{2} \sum_{v,u \in V} A_{vu} \delta(l_v, l_u) \tag{3.1}$$

où A_{vu} désigne un élément de la matrice d'adjacence symétrique de G', l_v et l_u les étiquettes associées aux nœuds v et u respectivement, et δ le symbole de Kronecker, une fonction qui retourne la valeur 1 si les étiquettes l_v et l_u sont égales, c.-à-d., quand v et u appartiennent à la même communauté. Dans le cas où v et u n'appartiennent pas à la même communauté, δ retourne la valeur 0. La fonction F mesure le nombre de liens qui contribuent à la formation des communautés dans un contexte unidimensionnel. Cette fonction est basée sur le fait que la règle de propagation de LPA mette à jour les étiquettes d'appartenance aux communautés de façon à augmenter le nombre de connexions au sein d'une communauté. Formellement, la procédure d'optimisation de LPA est définie par :

$$l'_v = \underset{\iota}{\operatorname{argmax}} \sum_{u \in V} A_{vu} \delta(l_u, \iota) \tag{3.2}$$

où l'_v dénote la nouvelle étiquette du nœud v. Dans le cas où deux ou plusieurs étiquettes maximiseraient la somme, la fonction *argmax* doit garder l'étiquette actuelle l_v pourvu qu'elle satisfasse déjà l'équation (3.2). Autrement, une étiquette aléatoire doit être sélectionnée à partir de l'ensemble dominant. Ici, il convient de noter que la convergence de cette procédure d'optimisation est garantie d'une part, par la nature monotone de la fonction F (3.1) elle même qui admet une borne supérieure finie représentée par le nombre de liens m du graphe G' et d'autre part par la nature asynchrone et aléatoire du processus itératif qui permet d'éviter les cycles infinis (Barber et Clark, 2009).

Notons que l'équation (3.1) peut être redéfinie pour supporter les réseaux multidimensionnels. L'idée consiste à considérer toutes les arêtes des dimensions impliquées dans la formation d'une communauté. Formellement, pour le multigraphe

G, F peut être redéfinie par :

$$F' = \frac{1}{2} \sum_{v,u \in V} \sum_{d \in D} A_{vu}^{(d)} \delta(l_v, l_u) \tag{3.3}$$

où $A_{vu}^{(d)}$ désigne un élément de la matrice d'adjacence associée à la dimension d de G. Autrement dit, $A^{(d)}$ représente la matrice d'adjacence d'un graphe unidimensionnel projeté selon la dimension d.

La recherche de communautés multidimensionnelles peut être guidée par la fonction F'. À noter que l'optimisation de F' peut être achevée en utilisant la règle de propagation définie par (3.2). Le principe consiste à considérer le nombre de connexions $\sum_{d \in D} A_{vu}^{(d)}$ entre le couple (v, u) comme un poids associé à un élément de la matrice A_{vu}. Ceci revient à l'application de LPA sur le réseau agrégé par la stratégie fréquentielle (Berlingerio *et al.*, 2011).

Pour un réseau unidimensionnel, Barber et Clark (2009) indiquent que la maximisation de F, définie par (3.1), ne garantit pas nécessairement un partitionnement optimal du réseau. Cela est dû au fait que le maximum global de l'équation (3.1) correspond à la solution triviale où tous les nœuds appartiennent à la même communauté. En fait, comme l'équation (3.2) ne produit que des changements locaux, le processus de recherche de l'optimum global de F est susceptible de converger vers un maximum local. Cette stratégie de recherche locale permet à LPA d'éviter le maximum trivial et faire ressortir ainsi une partition du réseau (Barber et Clark, 2009). Cependant, avec l'augmentation de la proportion de liens inter communautés, le maximum global indésirable devient beaucoup plus difficile à éviter. Pour le cas multidimensionnel, les mêmes observations demeurent valables puisque la fonction F', définie par (3.3), considère toutes les dimensions au même titre (alors que dans un contexte multidimensionnel, seule une partie des dimensions contribuent à la formation des communautés). Par conséquent, toute

procédure qui tente de maximiser l'équation (3.3) risque d'hériter les mêmes inconvénients liés à la maximisation de F tel que discuté dans ce paragraphe.

Afin d'éviter le maximum global indésirable de l'équation (3.3), une approche courante consiste à introduire des contraintes supplémentaires qui limitent l'espace de recherche. Intuitivement, puisque la densité de liens est beaucoup plus élevée dans les sous-espaces pertinents de dimensions, les nœuds formant une communauté se verront interagir principalement à travers un sous-ensemble commun de dimensions. Ainsi, nous pouvons identifier les dimensions pertinentes en repérant les dimensions les plus fréquemment utilisées au niveau des nœuds. Ces dimensions peuvent être exploitées comme une contrainte supplémentaire sur les arêtes qui entrent dans la formation d'une communauté. De ce fait, la pertinence des dimensions est considérée. L'objectif est d'identifier une partition où les membres d'une communauté se connectent à travers un nombre maximal de liens dans un sous-espace de dimensions commun. L'équation (3.3) peut ainsi être redéfinie comme suit :

$$F_{multi} = \frac{1}{2}\sum_{v,u\in V}\sum_{d\in D} A_{vu}^{(d)}\mathbb{I}_{D_v}(d)\mathbb{I}_{D_u}(d)\delta(l_v, l_u) \tag{3.4}$$

où $D_v \subseteq D$ et $D_u \subseteq D$ désignent, respectivement, les dimensions pertinentes pour les nœuds v et u. Dans l'équation (3.4), $\mathbb{I}$ représente la fonction caractéristique des sous-ensembles de D. $\mathbb{I}$ retourne la valeur 1 si la dimension du lien (v, u, d) fait partie du groupe de dimensions pertinentes D_v (ou D_u). $\mathbb{I}$ retourne la valeur 0 dans le cas contraire. Par conséquent, un lien n'est considéré que si la dimension à laquelle il appartient est réciproquement pertinente pour v et u. Notons que l'équation (3.4) demeure vraie pour les réseaux unidimensionnels, car lorsque $D = \{d\}$, $D_v = \{d\} \forall v \in V$, la fonction F_{multi} (équation (3.4)) reste toujours cohérente avec la fonction F (équation (3.1)). Autrement dit $F_{multi} = F$.

Tel que mentionné précédemment, une caractéristique notable de la fonction objective proposée est qu'elle prend en compte la pertinence des dimensions dans la formation des communautés. Ici, il convient de noter que toute procédure d'optimisation qui vise à maximiser la nouvelle fonction objective F_{multi} doit fournir un mécanisme pour la sélection des dimensions pertinentes D_v des nœuds. Ceci soulève la question suivante : comment définir la pertinence d'une dimension pour un nœud et dans quel sens ? La réponse à cette question exige une définition capable de discriminer entre les dimensions en fonction de leurs contributions en liens. De ce fait, la métrique de degré de centralité se voit la plus appropriée. Pour un nœud v, nous pouvons déterminer la pertinence d'une dimension d, par rapport aux autres dimensions du réseau, en mesurant le degré de centralité de v dans d, c-à-d., le nombre de liens sortant de v. Alternativement, nous pouvons également calculer la fraction des voisins directement accessibles dans d. Cette hypothèse a été utilisée dans les métriques de pertinence des dimensions proposées par Berlingerio *et al.* (2013a).

3.3 Procédure d'optimisation

Dans cette section, nous élaborons une procédure d'optimisation pour la fonction objective F_{multi} définie par (3.4). La procédure développée est basée sur un processus itératif qui s'inspire du principe de propagation d'étiquettes. Ce dernier est guidé par un mécanisme de pondération de voisins qui lui permet d'évaluer l'affinité entre les paires de nœuds en fonction du nombre et de la pertinence des dimensions qui connectent les nœuds en question. À cette fin, MDLPA procède en deux phases :

1. Dans la première phase, MDLPA associe un poids à chaque voisin u de chaque nœud v. Ce poids reflète, à la fois, le nombre et la pertinence des dimensions qui connectent v à son voisin u. Le calcul du poids en question

est effectué en deux étapes. D'abord, il est initialement estimé en fonction du nombre de dimensions pertinentes pour le nœud v. Par la suite, ce poids est réajusté afin de refléter la pertinence des dimensions en question pour le voisin u.

2. Dans la deuxième phase, les communautés sont identifiées en utilisant le processus itératif de propagation d'étiquettes. Chaque opération de réétiquetage implique les poids précédemment estimés dans la sélection de l'appartenance d'un nœud à une communauté de sorte que le nombre d'arêtes augmente au long du sous-espace de dimensions pertinentes de la communauté en question. Ce processus se répète jusqu'à ce que chaque nœud adopte une étiquette dominante.

Il convient de noter que la procédure d'optimisation que nous proposons est différente des stratégies de propagation classiques adoptées dans les réseaux unidimensionnels. Notre approche implique la sélection des dimensions pertinentes dans le processus d'optimisation. Le processus de pondération de voisins élaboré permet la recherche de communautés qui existent dans des sous-espaces de faible dimensionnalité, c-à-d., des sous-espaces ayant un nombre restreint de dimensions par rapport au nombre total de dimensions du réseau. Le processus d'optimisation est effectué de façon complètement automatique dans le sens où l'approche en question ne dépend d'aucun paramètre à fournir par l'utilisateur. Les détails de chaque phase sont présentés dans ce qui suit.

3.3.1 La première phase : initialisation

L'objectif de cette phase est de mesurer pour chaque nœud $v \in V$, des poids qui relatent l'affinité de v à chacun de ses voisins u. Ces poids, que nous appelons les poids d'attractions, visent à servir deux objectifs : (1) mesurer la contribution des

liens qui résulte de l'addition de v à la communauté du voisin u, et (2) identifier les dimensions pertinentes pour chaque nœud v, c.-à-d., les groupes D_v.

Le mécanisme de pondération élaboré dans cette première phase est basé sur l'hypothèse d'existence d'un sous-ensemble de dimensions pertinentes où la densité de connexions entre les membres d'une communauté est plus élevée. Ainsi, on s'attend à constater un plus grand nombre de voisins à travers les dimensions pertinentes comparativement aux dimensions non pertinentes. Cela permet de récupérer les ensembles de dimensions pertinentes D_v pour chaque nœud v, et donc son affinité à chacun de ses voisins u. Ceci se fait en estimant la fraction des voisins directement accessibles dans les dimensions reliant le couple (v, u). Plus la valeur de cette fraction est grande, plus les dimensions seront pertinentes pour v, et par conséquent, plus l'attraction de u sur v sera élevée et vice versa. Ainsi, les sous-ensembles de nœuds présentant des poids d'attraction élevés les uns sur les autres sont susceptibles de former une communauté.

Dans ce qui suit, nous introduisons le mécanisme de pondération utilisé pour l'estimation des poids d'attraction et la stratégie correspondante pour la sélection des dimensions pertinentes D_v. Tel que mentionné précédemment, le poids d'attraction exercée par le voisin u sur le nœud v est estimé en deux étapes. Initialement, le poids est calculé en fonction du nombre de dimensions pertinentes pour v uniquement. Par la suite, un ajustement est effectué afin de refléter la pertinence des dimensions de liaison pour le voisin u. Cet ajustement est réalisé en se basant sur le groupe D_u identifié dans la première étape. Les deux prochaines sections décrivent chaque étape.

3.3.1.1 Estimation initiale des poids d'attraction

Dans un travail récent, Berlingerio *et al.* (2013a) introduisent la métrique de pertinence exclusive de dimensions DR_{xor}, une fonction qui évalue la pertinence d'un ensemble de dimensions $S \subseteq D$ pour un nœud v en se basant sur la fraction des voisins exclusivement accessibles dans les sous-ensembles de S. Formellement, la DR_{xor} est définie par :

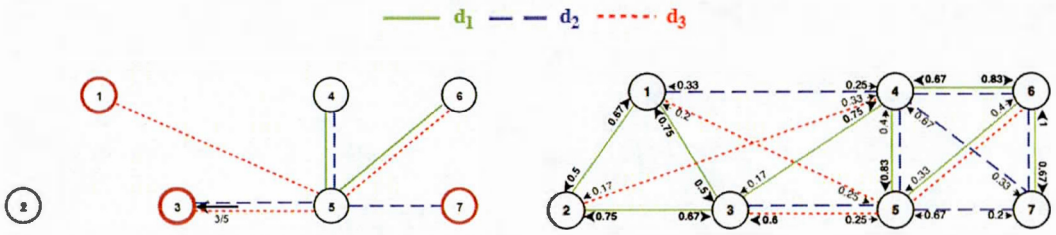
$$DR_{xor}(v, S) = \frac{\eta_{xor}(v, S)}{\eta(v)} \tag{3.5}$$

où $\eta(v)$ désigne l'ensemble de tous les voisins de v et est défini par $\eta(v) = \{u | \exists (v, u, d) \in E \wedge d \in D\}$ et $\eta_{xor}(v, S)$ est l'ensemble des voisins de v exclusivement accessibles dans les sous-ensembles de S et est défini par $\eta_{xor}(v, S) = \{u | \exists (v, u, s) \in E \wedge s \in S \wedge \forall d \in D - S, \nexists (v, u, d) \in E\}$. La DR_{xor} retourne des valeurs dans [0, 1] et atteint son maximum lorsque tous les voisins de v ne peuvent être atteints en dehors de S. Cette métrique offre un nombre de qualités appréciables la rendant appropriée pour l'estimation des poids d'attraction et ainsi la maximisation de la fonction F_{multi} (équation (3.4)). Premièrement, DR_{xor} peut mesurer la pertinence d'un ensemble de dimensions simultanément. De même, cette métrique favorise les grands ensembles de dimensions pertinentes. Ceci est attribué au fait qu'elle ne tient pas compte des voisins accessibles en dehors de S.

Soit $D(v, u) = \{d | \exists (v, u, d) \in E\}$, l'ensemble des dimensions reliant la paire (v, u). Nous définissons $\omega_0(v, u)$, le poids d'attraction initial d'un voisin u sur le nœud v (aussi appelé le score d'affinité initial du nœud v à son voisin u) par :

$$\omega_0(v, u) = DR_{xor}(v, D(v, u)) \tag{3.6}$$

Une valeur élevée de $\omega_0(v, u)$ indique un plus grand nombre de dimensions pertinentes pour v dans $D(v, u)$, et, initialement, une plus grande contribution



(a) Le poids d'attraction de n_3 sur n_5.

(b) Les poids d'attraction exercés sur l'ensemble des nœuds.

Figure 3.1: Poids d'attraction initiaux sur le réseau de la figure 1.3a.

en liens pour la communauté du voisin u. Afin de supporter les mises à jour ultérieures, une matrice d'adjacence non symétrique est utilisée pour stocker les poids ω. La figure 3.1a illustre le poids d'attraction initial du nœud n_3 sur le nœud n_5. Les voisins de n_5 qui respectent la condition d'exclusivité de η_{xor} contre $D(n_5, n_3)$, c.-à-d. $\{d_1, d_2\}$, sont soulignés en gras (spécifiquement n_1, n_3 et n_7). Les poids d'attraction calculés pour l'ensemble des nœuds sont illustrés sur la figure 3.1b. La direction des flèches indique l'affinité de la source à la destination.

Les poids ω_0 offrent une partition préliminaire dans le sens où ils permettent aux nœuds présentant une affinité élevée les uns aux autres d'achever rapidement un consensus sur une seule étiquette. Cependant, puisqu'ils ne reflètent la pertinence de $D(v, u)$ que pour v, les valeurs ω_0 peuvent engendrer une violation de la contrainte de réciprocité de l'équation (3.4). Pour illustrer ce cas, considérons les poids estimés dans la figure 3.1b. On peut constater que, à l'exception du nœud n_5, tous les nœuds montrent plus d'intérêt à rejoindre les voisins de la même communauté vis-à-vis la partition de la figure 1.3b. En revanche, la valeur élevée de $\omega_0(n_5, n_3)$ suggère que n_5 doit rejoindre la communauté C_1 de n_3. Or, les dimensions d_2 et d_3 ne sont pas pertinentes pour cette communauté. L'acquisition du

nœud n_5 dans ce cas va apporter des liens qui ne contribuent pas à sa formation. Ceci va se traduire par une baisse de densité de liens au sein du groupe. Par conséquent, un ajustement des poids ω_0 est nécessaire.

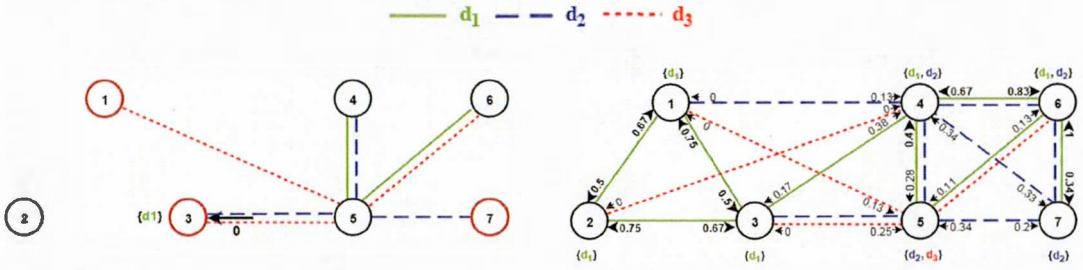
3.3.1.2 Ajustement des poids d'attraction

Tel que discuté dans la section précédente, $\omega_0(v,u)$ définit la contribution d'un nœud v à la communauté du voisin u en se basant sur le nombre de dimensions dans $D(v,u)$ qui sont uniquement pertinentes pour v. Toutefois, comme u peut exprimer une distribution différente de liens dans son voisinage, les dimensions $D(v,u)$ peuvent ne pas être pertinentes pour u. Dans le but d'éluder cette situation, le nombre de dimensions pertinentes dans $D(v,u)$ pour u doit être reflété dans $\omega_0(v,u)$. Une solution possible consiste à pénaliser la valeur ω_0 initialement estimée par la distance séparant l'ensemble $D(v,u)$ de l'ensemble de dimensions pertinentes D_u pour u. Cette distance peut être estimée en utilisant l'indice de Jaccard. Formellement, on définit le poids d'attraction ajusté $\omega(v,u)$ du voisin u sur v par :

$$\omega(v,u) = \omega_0(v,u) \times \frac{|D_u \cap D(v,u)|}{|D_u \cup D(v,u)|} \tag{3.7}$$

Plus la valeur de ω est grande, plus le nombre de dimensions pertinentes dans $D(v,u)$ est élevé pour v et u conjointement.

Les poids ω permettent de donner une mesure relative dans laquelle le terme $\sum_{d \in D} A_{vu}^{(d)} \mathbb{I}_{D_v}(d) \mathbb{I}_{D_u}(d)$ de l'équation (3.4) est indirectement estimé. La figure 3.2a illustre le poids d'attraction ajusté $\omega(n_5, n_3)$ du voisin n_3 sur le nœud n_5 après la considération de la pertinence des dimensions $\{d_2, d_3\}$ pour n_3. Puisque d_1 est la seule dimension pertinente (la plus fréquemment utilisée) pour ce dernier, le nombre de dimensions dans $D(n_5, n_3)$ qui sont à la fois pertinentes pour n_5 et n_3 serait égale à 0. Ainsi, la contribution du nœud n_5 pour la communauté de n_3



(a) Le poids d'attraction ajusté de n_3 sur n_5.

(b) Les poids d'attraction ajustés sur l'ensemble des nœuds.

Figure 3.2: Poids d'attraction ajustés sur le réseau de la figure 1.3a.

sera nulle. La figure 3.2b illustre les poids d'attraction ajustés sur l'ensemble des nœuds du réseau. Les nœuds présentant des valeurs élevées en interne ont une forte chance de former une communauté.

Il reste de définir la règle selon laquelle les dimensions pertinentes D_u sont sélectionnées. Rappelons que ω_0 reflète le nombre de dimensions pertinentes à v uniquement. Par conséquent, on peut identifier les dimensions pertinentes D_v en sélectionnant le groupe $D(v,u)$ pour lequel ω_0 est maximal. Toutefois, comme DR_{xOR} favorise les grands ensembles de dimensions pertinentes, le groupe sélectionné risque d'avoir des dimensions non pertinentes. Pour illustrer ce cas, considérons le nœud n_7 dans la figure 3.1b. Si l'on opte pour cette règle, $D(n_7, n_6)$ serait retenu pour D_{n_7} (puisqu'il définit le poids ω_0 le plus élevé parmi tous les voisins de n_7). Or, d_1 est moins fréquemment utilisée par n_7 et donc non pertinente. Pour éviter ce cas, on peut regrouper les voisins u selon $D(v,u)$ pour ensuite sélectionner l'ensemble supportant le poids d'attraction le plus élevé. Formellement, l'ensemble initial des dimensions pertinentes D_v est défini par :

$$D_v^{init} = \underset{S}{\operatorname{argmax}} \sum\nolimits_{u \in \eta(v)} \omega_0(v,u)\delta(D(v,u), S) \tag{3.8}$$

Dans le cas où deux ou plusieurs ensembles supporteraient le même poids combiné, l'union des groupes est utilisée. D_v peuvent également être exploités pour récupérer les sous-espaces pertinents D_K des communautés. La prochaine section introduit une stratégie consensuelle permettant d'identifier le sous-espace D_K associé à une communauté en partant des ensembles D_v (ici D_v représente les dimensions pertinentes des nœuds v appartenant à la même communauté). La figure 3.2b illustre les dimensions pertinentes sélectionnées D_v selon l'équation (3.8) ainsi que les poids d'attraction ajustés ω. Les membres des deux communautés C_1 et C_2 sont mieux séparés après la mise à jour, tel qu'illustré par les faibles poids sur les arêtes inter communautés. Le principal avantage du système de pondération présenté est qu'il fournit une mesure relative dans laquelle la contribution des dimensions pertinentes est indirectement estimée, seulement, en se basant sur la distribution locale des liens. Le résumé de la phase d'initialisation est décrit par l'algorithme 1.

3.3.2 La deuxième phase : identification des communautés

Conformément à la stratégie de pondération décrite dans la première phase, nous introduisons une nouvelle règle de propagation pour l'identification des communautés dans les réseaux multidimensionnels. La règle proposée exploite les poids ω estimés pour déterminer l'appartenance d'un nœud à une communauté de sorte que l'équation (3.4) soit maximisée. L'idée de base consiste à attribuer aux nœuds, l'étiquette portée par les voisins exerçant le poids d'attraction ω le plus élevé. Ainsi, la dominance d'une étiquette dans un voisinage dépend du poids d'attraction exercée par les voisins qui l'adoptent. Formellement, la règle de propagation d'étiquettes est définie par :

$$l'_v = \underset{\iota}{\operatorname{argmax}} \sum_{u \in \eta(v)} \omega(v,u)\delta(l_u, \iota) \tag{3.9}$$

Cette règle est basée sur le fait qu'une valeur maximale de ω correspond à la

Algorithme 1 : Initialisation

Données : Un multigraphe G=(V,E,D)

Résultat : ω_0, ω et les ensembles initiaux des dimensions pertinentes $\{D_v^{init}\}_{v \in V}$

début

  pour chaque $v \in V$ **faire**

    pour chaque $u \in \eta(v)$ **faire**

      calculer $\omega_0(v,u)$ selon l'équation (3.6);

    fin

    sélectionner D_v selon l'équation (3.8);

  fin

  pour chaque $v \in V$ **faire**

    pour chaque $u \in \eta(v)$ **faire**

      calculer $\omega(v,u)$ selon l'équation (3.7) ;

    fin

  fin

  retourner ω_0, ω et $\{D_v^{init}\}_{v \in V}$;

fin

valeur la plus élevée du terme $\sum_{d \in D} A_{uv}^{(d)} \mathbb{I}_{D_v}(d) \mathbb{I}_{D_u}(d)$ de l'équation (3.4). Dans le cas où deux ou plusieurs communautés concurrentes exerceraient le même poids d'attraction, la fonction *argmax* doit sélectionner une étiquette aléatoire à partir du groupe dominant indépendamment de l'affiliation (l'appartenance à une communauté) courante du nœud. Ce changement d'affiliation permettra à l'algorithme de chercher une meilleure solution en effectuant une marche aléatoire lorsque'un plateau de F_{multi} (3.4) est atteint.

En adoptant cette règle de propagation, nous pouvons identifier une partition optimale en utilisant le même processus itératif de LPA. Spécifiquement, chaque nœud v reçoit une étiquette l_v^{init} initiale unique. Par la suite, les étiquettes des nœuds sont mises à jour selon la règle (3.9). Avec chaque opération de réétiquetage, nous mettons à jour les groupes D_v ainsi que les poids d'attraction ω exercés par le nœud v sur ses voisins. Cette opération vise à sélectionner les dimensions qui sont à la fois pertinentes pour le nœud acquis v et sa nouvelle communauté. Une façon simple d'y parvenir est d'effectuer une intersection entre l'union des dimensions D_u du groupe gagnant, et celles qui le relient au nœud acquis v. Cette règle de sélection est basée sur le fait que les ensembles de dimensions $D(v,u)$ pour lesquels ω sont les plus élevés définissent déjà les dimensions qui sont à la fois pertinentes pour v et sa nouvelle communauté. L'intersection permettra ainsi d'élaguer les dimensions non pertinentes possiblement sélectionnées durant la phase d'initialisation. Formellement, nous définissons l'ensemble des dimensions pertinentes D_v d'un nœud v par rapport à la communauté gagnante comme suit :

$$D_v = \left[\cup_{u \in \eta(v) | l_u = l'_v} D(v,u)\right] \cap \left[\cup_{u \in \eta(v) | l_u = l'_v} D_u\right] \tag{3.10}$$

où le premier terme corresponds aux dimensions reliant le nœud v aux voisins gagnants et le deuxième terme l'ensemble des dimensions pertinentes pour ces mêmes voisins.

Une fois les dimensions pertinentes mises à jour, un ajustement des poids d'attraction exercés par le nœud v sera nécessaire. L'objectif est de corriger les poids précédemment estimés en vue de refléter la pertinence du nouveau groupe de dimension pour les voisins qui s'intéressent à rejoindre la communauté de v. Une fois le traitement de l'ensemble des nœuds du réseau complété, l'algorithme vérifie si le critère d'arrêt est satisfait, en d'autres termes, si chaque nœud est attribué à

une étiquette dominante. Si ce n'est pas le cas, un nouveau cycle de propagation se lance et les règles de mise à jour dans (3.7), (3.9) et (3.10) s'appliquent de nouveau jusqu'à ce que le critère d'arrêt soit satisfait. Étant basée sur le même processus de propagation dans le contexte unidimensionnel, la convergence de cette procédure d'optimisation est ainsi garantie (F_{multi} (3.4) est également monotone et bornée par le nombre de liens m du multigraphe). Par ailleurs, les contraintes exprimées par la fonction objective F_{multi} (3.4) permettent d'accélérer la convergence du processus de propagation. En effet, d'après les expérimentations menées sur les réseaux réels et synthétiques, nous avons constaté que plus de 97% des nœuds sont correctement classifiés (attribués à la communauté dominante) après la 3^{eme} itération du processus. Ce taux n'a été achevé qu'après la 5^{eme} itération sur les réseaux unidimensionnels (Raghavan *et al.*, 2007).

Les communautés commencent à se développer à partir des régions de haute densité progressivement. Dès que la frontière d'une autre communauté concurrente est atteinte, la règle de propagation décide de l'appartenance des nœuds en fonction de leurs contributions aux communautés candidates. Ainsi, la communauté qui reçoit le plus grand nombre de liens sur son sous-espace pertinent définit l'affiliation du nœud. Une fois le critère d'arrêt satisfait, l'algorithme s'arrête et la partition finale est récupérée à partir des nœuds portant les mêmes étiquettes l_v. Les sous-espaces D_k associés aux communautés détectées sont ensuite obtenus en fusionnant les groupes D_v des nœuds appartenant à la même communauté.

La figure 3.3 illustre la partition identifiée sur le réseau multidimensionnel de la figure 1.3a. Les étiquettes finales ainsi que les poids révisés sur la figure 3.3 ont été obtenus après deux cycles de propagation. Comme on peut le constater, les nœuds n_1, n_2, et n_3 portent tous la même étiquette C_1 et constituent ainsi une

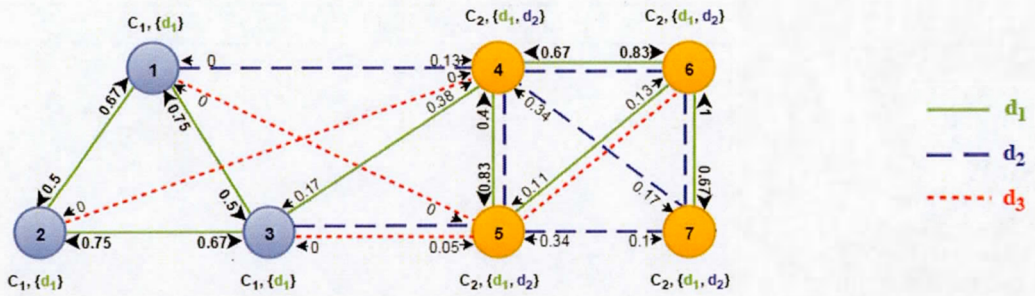


Figure 3.3: Partition finale identifiée par MDLPA sur le réseau de la figure 1.3a.

communauté dans la dimension d_1. De même, les nœuds n_5, n_6, n_7, et n_8 forment une communauté dans les dimensions d_1 et d_2. Nous pouvons également constater que, par rapport à ce qui est présenté à la figure 3.2, les dimensions pertinentes des nœuds n_5 et n_7 ont été mises à jour pour refléter les dimensions pertinentes de la communauté C_2. L'algorithme 2 illustre les étapes nécessaires pour identifier les communautés ainsi que leurs dimensions pertinentes.

3.4 Analyse de complexité

La complexité algorithmique de MDLPA dépend des deux phases de l'algorithme, à savoir, l'estimation des poids d'attraction initiaux et la propagation d'étiquettes.

3.4.1 Complexité de la première phase

Le temps d'exécution de cette phase est principalement déterminé par le nombre total de voisins dans le multigraphe G. Pour chaque nœud v, le nombre maximal de voisins $|\eta(v)|$ est égal à la somme des degrés de v dans chacune des dimensions de G : $deg(v) = \sum_{d \in D} deg_d(v)$. Autrement dit, quand chaque voisin est accessible par le biais d'une dimension unique.

Algorithme 2 : Identification des communautés

Données : Un multigraphe G=(V,E,D)

Résultat : Les communautés multidimensionnelles $P = \{(V_k, D_k)_{k=1...K}\}$

début

Initialiser ω_0, ω et $\{D_v^{init}\}_{v \in V}$ selon l'algorithme 1;

pour chaque $v \in V$ **faire**

Affecter une étiquette l_v unique pour v.

fin

// Identification de l_v et D_v associés à chaque nœud $v \in V$.

tant que $\exists v \in V$ *tel que* l_v *est différente de l'étiquette dominante sur* $\eta(v)$ **faire**

Mélanger la liste de nœuds V;

pour chaque $u \in \eta(v)$ **faire**

Mettre à jour l_v selon l'équation (3.9);

Mettre à jour D_v selon l'équation (3.10);

pour chaque $u \in \eta(v)$ **faire**

Mettre à jour $\omega(u, v)$ selon l'équation (3.7);

fin

fin

fin

regrouper $v \in V$ selon l_v dans K groupes $\{V_k\}_{k=1...K}$;

// Identification des dimensions pertinentes D_k pour chaque V_k

pour chaque $V_k \in P$ **faire**

pour chaque $v \in V_k$ **faire**

$D_k = D_k \cup D_v$;

fin

fin

retourner P;

fin

Pour chaque $u \in \eta(v)$, afin de calculer $\omega_0(v,u)$, nous devons effectuer $deg(v)$ vérifications contre $D(v,u)$ ce qui nécessite un temps d'exécution dans l'ordre de $deg(v)$. Puisque nous avons $deg(v)$ voisins, le nombre maximal possible de groupes $D(v,u)$ est égal à $deg(v)$, ce qui résulte, dans le pire cas, en un temps d'exécution de l'ordre de $O(deg^2(v))$ pour l'ensemble de tous les voisins de v. Par conséquent, le pire temps d'exécution pour un multigraphe connexe G est égal à $O(M_1)$, où $M_1 = \sum_{v \in V} deg^2(v)$, dénote le premier indice de Zagreb (Graovac *et al.*, 1972). Pour les graphes simples, plusieurs bornes supérieures ont été proposées pour M_1. Nous prenons la borne $2nm - n^2 + n$ donnée par Liu et Liu (2009). Comme nous supposons une seule arête entre chaque paire de nœuds, l'inégalité $M_1 \leq 2nm - n^2 + n$ reste applicable pour le multigraphe G puisque $m = \frac{1}{2} \sum_{v \in V} \sum_{d \in D} deg_d(v)$. Par conséquent, le calcul de ω_0 est effectué en $O(mn)$ au plus.

L'initialisation des étiquettes l_v est réalisée en $O(n)$ tandis que la sélection des dimensions pertinentes D_v nécessite un temps d'exécution dans l'ordre de $O(m)$. À chaque nœud v, les voisins u sont d'abord regroupés selon $D(v,u)$, ce qui nécessite $O(deg(v))$. Ensuite, le groupe de voisins ayant le poids maximal est sélectionné et les dimensions associées sont affectées à v, ce qui nécessite, dans le pire cas, un temps d'exécution de $O(deg(v))$, et donc, un temps global dans l'ordre de $O(m)$. Enfin, le calcul des ω pour chaque nœud v nécessite $O(deg(v))$ opérations, et donc, un temps global de $O(m)$. De ce fait, l'initialisation des étiquettes et l'estimation de ω est réalisée en un temps de l'ordre de $O(m+n)$. Par conséquent, la phase d'initialisation nécessite, au pire des cas, un temps d'exécution de l'ordre de $O(mn)$.

3.4.2 Complexité de la deuxième phase

Chaque cycle de propagation est effectué dans un temps presque linéaire au nombre d'arêtes m. Pour chaque $v \in V$, nous regroupons les voisins u selon leurs étiquettes. Par la suite, nous sélectionnons l'étiquette du groupe ayant le poids d'attraction le plus élevé. Ceci nécessite, dans le pire cas, un temps d'exécution de $O(deg(v))$. Cette complexité est aussi conservée pour la mise à jour des groupes D_v et ω, ce qui résulte en un pire temps d'exécution de $O(m)$. Bien que le nombre d'itérations nécessaires pour la convergence soit inconnu *à priori*, nos expérimentations sur les réseaux synthétiques et réels montrent que plus de 97% des nœuds sont correctement classifiés (attribués à la communauté dominante) après la 3^{eme} itération, peu importe le nombre de nœuds ou de dimensions (pour $nd > 1$). Ces résultats sont en fait conformes avec ceux obtenus par LPA sur les réseaux unidimensionnels (Raghavan *et al.*, 2007). La permutation de l'ensemble des nœuds dans V s'effectue en $O(n)$ tandis que la vérification de convergence est achevée en $O(n)$. Par conséquent, chaque cycle est effectué en $O(m+n)$ au pire des cas.

Pour conclure, la complexité de notre approche, dans son implémentation actuelle, est $O(mn)$, la rendant linéairement évolutive avec le nombre d'arêtes m, quel que soit le nombre de dimensions nd, pourvu que le nombre de nœuds n soit constant. À noter que lorsque nd est constant, nous pouvons modifier la stratégie d'estimation des poids ω_0 pour achever un temps linéaire en nombre d'arêtes m. Avec nd dimensions, nous obtenons $2^{nd} - 1$ groupes de dimensions possibles $D(v, u)$ entre n'importe quelle paire de nœuds. Par conséquent, l'estimation des ω_0 appliqués sur v nécessiterait un pire temps d'exécution de l'ordre de $O(2^{nd} deg(v))$. Il en résulte que la complexité globale sur un multigraphe est de $O(m)$ lorsque nd est constant. Ainsi, le pire temps d'exécution de l'algorithme est de l'ordre de $O(m+n)$. De ce fait, quand le nombre de dimensions nd est fixé, le temps

d'exécution de l'algorithme évolue quasi linéairement en nombre d'arêtes m. Par conséquent, l'efficacité computationnelle des algorithmes basés sur la propagation d'étiquettes est préservée.

3.5 Conclusion

Dans ce chapitre, nous avons présenté une nouvelle approche de détection de communautés dans les réseaux multidimensionnels. L'approche proposée se fonde sur l'optimisation d'une fonction objective à travers un mécanisme itératif qui se base sur le principe de propagation d'étiquettes. La stratégie d'optimisation développée comporte deux phases et exploite la pertinence des dimensions via un système de pondération permettant de guider la recherche de la partition optimale. Cette stratégie permet une identification efficace et relativement précise des communautés ainsi que leurs dimensions pertinentes. Les résultats des expérimentations présentées dans le chapitre suivant corroborent nos propos.

CHAPITRE IV

ÉVALUATION DE L'APPROCHE PROPOSÉE

Ce chapitre présente une évaluation empirique de MDLPA sur des réseaux synthétiques et réels. La performance de notre approche est comparée à cinq algorithmes connus dans la littérature. Dans ce qui suit, nous commençons d'abord par décrire le cadre expérimental ainsi que les indicateurs de performance et les cirières d'évaluation retenus. Par la suite, nous présentons les résultats obtenus.

4.1 Cadre expérimental

4.1.1 Algorithmes sélectionnés pour la comparaison

Afin de démontrer l'efficacité de MDLPA, cinq approches appartenant à trois catégories différentes d'algorithmes de détection de communautés ont été retenues pour la comparaison, à savoir : (1) l'agrégation de dimensions, (2) l'intégration de caractéristiques, et (3) le clustering d'ensembles. Pour les techniques basées sur l'agrégation de dimensions, nous avons implémenté deux techniques bien connues, soit l'agrégation binaire et l'agrégation fréquentielle (Berlingerio *et al.*, 2011). La partition finale est identifiée sur le réseau agrégé en utilisant l'algorithme LPA de base (Raghavan *et al.*, 2007). Quant aux techniques d'intégration de caractéris-

tiques, nous avons sélectionné PMM [1] Tang *et al.* (2012) et SC-ML [2] (Dong *et al.*, 2014). Dans nos expérimentations, nous évaluons également la performance des techniques de clustering d'ensembles [3] proposées par Strehl et Ghosh (2003). À cette fin, nous avons utilisé l'algorithme de Louvain [4] (Blondel *et al.*, 2008) comme l'algorithme de base pour identifier les communautés dans chaque dimension du réseau. Pour identifier la partition finale, nous avons considéré les trois stratégies consensuelles proposées par Strehl et Ghosh (2003), à savoir, Cluster-Based Similarity Partitioning Algorithm (CSPA), HyperGraph Partition Algorithm (HGPA) and Meta-CLustering Algorithm (MCLA). Le meilleur résultat obtenu parmi ces trois stratégies est retenue pour le clustering d'ensembles. Nous pensons que notre choix d'algorithmes couvre une variété d'approches de détection de communautés. Finalement, en ce qui concerne les approches qui se basent sur la décomposition tensorielle, nous avons initialement considéré GraphFuse (Papalexakis *et al.*, 2013). Toutefois, en raison de son temps d'exécution excessif, nous l'avons retiré. Pour un réseau 100-dimensionnel de 3000 nœuds, GraphFuse prend plus d'une semaine pour identifier les communautés. Les autres techniques telles que ABACUS (Berlingerio *et al.*, 2013b), Multiplex Infomap (De Domenico *et al.*, 2015a) et LART (Kuncheva et Montana, 2015) n'ont pas été sélectionnées pour la comparaison à cause de la nature chevauchante des communautés qu'elles découvrent.

---

1. L'implémentation de PMM est disponible à partir de `http://leitang.net/heterogeneous_network.html`

2. L'implémentation de SC-ML est disponible à partir de `http://lts4.epfl.ch/xdong/code`

3. L'implémentation des techniques de clustering d'ensembles est disponible à partir de `http://strehl.com/soft.html`

4. Louvain est une approche non paramétrique qui identifie les communautés dans les réseaux unidimensionnels en maximisant la modularité de Newman (2006)

4.1.2 Stratégie de réglage de paramètres et d'exécution des algorithmes

PMM ainsi que SC-ML nécessitent que le nombre de communautés à identifier soit fourni par l'utilisateur. Dans nos expérimentations, le nombre réel de communautés a été fourni pour ces deux approches. Par ailleurs, puisque PMM et SC-ML dépendent d'autres paramètres d'entrée, diverses valeurs ont été considérées. Plus spécifiquement, pour PMM, nous avons sélectionné le nombre de caractéristiques structurelles dans l'intervalle $[5, 14]$ en utilisant des incréments graduels de 1. Pour SC-ML, les valeurs du paramètre de régularisation ont été choisies à partir de l'intervalle $[0, 1]$ en utilisant des incréments graduels de 0.1. Rappelons que, comme mentionné dans le chapitre 2, PMM et SC-ML utilisent l'algorithme k-means pour récupérer la partition finale. Ce dernier est sensible à l'initialisation aléatoire des centres. Pour éviter le biais de l'initialisation, nous avons exécuté PMM et SC-ML à diverses reprises, et ce pour chaque valeur de paramètre sélectionnée. La combinaison de paramètres qui donne le meilleur résultat est conservée. Nous présentons aussi la performance moyenne et la plus basse de chaque algorithme.

Les deux stratégies d'agrégation binaire et fréquentielle ainsi que MDPLA et le clustering d'ensembles sont évaluées de la même manière que PMM et SC-ML, c.-à-d., à travers dix exécutions successives. Suite à ces exécutions, nous avons retenu la meilleure ainsi que la plus basse performance. Nous avons également calculé la performance moyenne à la suite de ces dix exécutions. À noter que ces essais répétitifs ont été effectués en raison de la nature non déterministe de MDLPA et des techniques d'agrégation. En effet, l'algorithme LPA utilisé par les techniques d'agrégation (binaire et fréquentielle) est sensible à l'ordre de traitement des nœuds. Notre approche souffre également de ce problème puisque la stratégie d'optimisation développée s'appuie sur le principe de propagation d'étiquettes. Cependant, comme le montreront les résultats, l'impact de ce problème

sur notre approche est minime. Finalement, tel que discuté précédemment, nous avons utilisé la méthode Louvain comme approche de base pour le clustering d'ensembles. Cette méthode est également non déterministe et peut générer différentes partitions en fonction de l'ordre dans lequel les nœuds sont traités. De ce fait, nous exécutons les techniques de clustering d'ensembles dix fois. Pour chaque exécution, la partition finale est celle qui rapporte le meilleur résultat parmi les trois stratégies consensuelles (CSPA, HGPA et MCLA). Comme les autres approches, nous rapportons le plus faible, le moyen et le meilleur résultat. Nous croyons que la stratégie d'exécution adoptée contribue à assurer une comparaison équitable des différentes approches considérées dans les expérimentations.

4.1.3 Critères d'évaluation

Afin d'évaluer la performance des algorithmes sélectionnés, nous avons considéré des critères d'évaluation internes et externes. Les critères internes sont utilisés lorsque les résultats de partitionnement sont évalués par rapport à une partition de référence, c.-à-d., une partition prédéfinie où l'on connaît à l'avance l'appartenance d'un nœud à une communauté (ground truth). Quant à eux, les critères externes s'appliquent quand les partitions identifiées sont évaluées d'une façon non supervisée, et ce en termes de quantités mesurables à partir des données du réseau. Dans nos expérimentations, nous avons utilisé les critères internes lorsque les partitions de référence sont disponibles (en particulier avec les réseaux synthétiques). En l'absence de partitions de référence (ce qui est généralement le cas des réseaux réels), nous avons utilisé les critères externes.

4.1.3.1 Critères internes

Parmi les métriques disponibles, nous avons choisi d'utiliser l'information mutuelle normalisée (NMI) (Manning *et al.*, 2008), l'indice de Rand ajusté(ARI)

(Hubert et Arabie, 1985) et l'indice de Fowlkes-Mellow (FM) (Fowlkes et Mallows, 1983). Les indices NMI, ARI et FM évaluent les résultats de détection en calculant la correspondance entre la partition de référence R et la partition obtenue O par les algorithmes. NMI est défini par :

$$NMI(R,O) = \frac{2I(R;O)}{H(R)+H(O)} \tag{4.1}$$

où $I(R;O) = H(R) - H(R|O)$ est l'information mutuelle entre R et 0, $H(R)$ et $H(O)$ représentent l'entropie de Shannon de R et O respectivement, et $H(R|O)$ l'entropie conditionnelle de R sachant O.

Les indices ARI et FM sont définis par :

$$ARI(O,R) = \frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)} \tag{4.2}$$

$$FM(O,R) = \frac{a}{\sqrt{(a+b)(a+c)}} \tag{4.3}$$

où a, b, c, et d dénotent, respectivement, le nombre de paires affectées à la même communauté dans R et O; dans la même communauté dans R mais pas dans O; dans la même communauté dans O mais pas dans R; et dans différentes communautés dans R et O. Plus R et O sont semblables, plus les valeurs retournées par ces métriques (NMI, ARI et FM) sont élevées. Lorsque R et O sont identiques, la valeur de chaque indice est égale à 1. Lorsque R et O sont bien différentes, la valeur de chaque indice est proche de 0. Nous pensons que la comparaison des valeurs retournées par ces trois métriques offre un critère objectif pour l'évaluation des résultats obtenus par les algorithmes comparés.

4.1.3.2 Critères externes

Les critères externes représentent des métriques utilisées pour l'évaluation de la qualité des résultats obtenus à partir des réseaux où les partitions de références

sont inconnues. Dans nos expérimentations, nous avons considéré des métriques externes locales et globales.

En résumé, telles que décrites dans (Loe et Jensen, 2015), les métriques locales se basent sur le fait qu'une communauté est censée démontrer de faibles interactions avec les autres communautés du réseau. Ainsi, il est possible d'évaluer la partition obtenue en évaluant chaque communauté à part. En revanche, une métrique globale considère à la fois les interactions intra communautés et inter communautés dans l'évaluation. Dans ce qui suit, nous présentons les métriques locales et globales que nous avons utilisées dans nos tests.

La première métrique externe que nous avons implémentée est celle de la redondance de liens (Berlingerio *et al.*, 2011). Pour une communauté C_k, la redondance $\rho(C_k)$ est définie par :

$$\rho(C_k) = \frac{1}{|DF_k||SP_k|} \sum_{(v,u)\in SP'_k} \sum_{d\in DF_k} A^{(d)}_{vu} \tag{4.4}$$

où DF_k dénote l'ensemble des dimensions retrouvées dans C_k, SP_k l'ensemble des paires (v, u) connectées par, au moins, une dimension dans DF_k, $SP'_k \subseteq P_k$ l'ensemble des paires connectées par, au moins, deux dimensions dans DF_k et $A^{(d)}_{vu}$ la matrice d'adjacence du multigraphe projeté sur la dimension d. La métrique définie par Berlingerio *et al.* (2011) mesure la redondance de connexions, à savoir, le ratio d'arêtes reliant des paires adjacentes dans, au moins, deux dimensions par rapport nombre maximal de liens entre toutes les paires de nœuds connectées. L'intuition derrière cette métrique se base sur le fait que les nœuds formant une communauté multidimensionnelle sont censés interagir sur différentes dimensions simultanément. La redondance ρ prend des valeurs dans l'intervalle $[0, 1]$. Plus le nombre de dimensions reliant les paires de nœuds est grand, plus la redondance

sera élevée (Berlingerio *et al.*, 2011). La métrique atteint son maximum lorsque l'ensemble des dimensions DF_k apparait entre chaque paire de nœuds adjacents. Pour une partition $P = \{C_1, ..., C_K\}$, la redondance peut être obtenue en calculant la moyenne de redondances dans les communautés C_K de P (Hmimida et Kanawati, 2015). Formellement, $\rho(P)$ peut être définie par :

$$\rho(P) = \frac{1}{|P|} \sum_{C_k \in P} \rho(C_k) \tag{4.5}$$

La deuxième métrique locale sélectionnée est celle de la densité multidimensionnelle de communauté (Berlingerio *et al.*, 2013b). Cette métrique mesure le nombre de liens au sein d'une communauté normalisé par le nombre maximum possible de liens entre ses membres (c.-à-d., les nœuds appartenant à la même communauté en question). Pour une communauté C_k, la densité multidimensionnelle de communauté $MCD(C_k)$ est définie par :

$$MCD(C_k) = \frac{\frac{1}{2} \sum_{v,u \in V_k} \sum_{d \in DF_k} A_{vu}^{(d)}}{|DF_k||V_k|\frac{|V_k|-1}{2}} \tag{4.6}$$

MCD retourne des valeurs dans $[0, 1]$ où une valeur maximale reflète une plus grande connectivité dans la communauté. Pour une partition $P = \{C_1, ..., C_K\}$, on peut calculer la moyenne des densités des communautés identifiées. Formellement, $MCD(P)$ est défini par :

$$MCD(P) = \frac{1}{|P|} \sum_{C_k \in P} MCD(C_k) \tag{4.7}$$

Ici, il est important de faire la distinction entre les ensembles DF_k, utilisés dans les équations (4.4) et (4.6), et l'ensemble D_k que notre approche vise à identifier. En fait, DF_k désigne l'ensemble de toutes les dimensions retrouvées dans une communauté C_k, tandis que D_k correspond au sous-ensemble de dimensions pertinentes de C_k. Ainsi, $D_k \subseteq DF_k$. Pour illustrer la différence,

considérons de nouveau le réseau multidimensionnel de la figure 1.3a. Rappelons que ce réseau contient deux communautés $C_1 = (V_1, D_1) = (\{v_1, v_2, v_3\}, \{d_1\})$ et $C_2 = (V2, D2) = (\{v_4, v_5, v_6, v_7\}, \{d_1, d_2\})$. Par ailleurs, comme on peut le constater à partir de la figure 1.3a, l'ensemble des dimensions retrouvées dans C_1 est $DF_1 = \{d1\}$, tandis que l'ensemble des dimensions retrouvées dans C_2 est $DF_2 = \{d_1, d_2, d_3\}$.

Dans nos expérimentations, les ensembles DF_k vont être utilisés dans le calcul de $\rho(C_k)$ et $MCD(C_k)$ sur les communautés détectées par les algorithmes considérés pour la comparaison (y compris le notre). De plus, puisque MDLPA a l'avantage d'identifier les dimensions pertinentes (ce qui est n'est pas le cas des autres algorithmes), nous avons également considéré les ensembles D_k dans le calcul de $\rho(C_k)$ et $MCD(C_k)$. Dans ce cas, nous remplaçons DF_k par D_k dans les deux équations (4.4) et (4.6). Ceci nous permet d'évaluer l'efficacité du mécanisme de sélection des dimensions pertinentes.

L'utilisation unique des métriques locales (la redondance et la densité multidimensionnelle de communauté) ne permet pas nécessairement d'établir, d'une façon objective et équitable, les mérites de chaque approche considérée dans la comparaison. En fait, une partition où les communautés se composent d'une seule paire de nœuds aura des scores élevés pour ces deux métriques (la redondance et la densité multidimensionnelle). Afin de balancer cet effet, une métrique globale qui tient compte des interactions inter communautés doit également être considérée pour évaluer les résultats des algorithmes de détection de communautés. À cette fin, Mucha *et al.* (2010) proposent une version généralisée de la modularité de Newman (2006) pour les réseaux multidimensionnels, intitulée : modularité multicouche. Formellement, pour un multigraphe G, la modularité multicouche Q

est définie par :

$$Q(G) = \frac{1}{2\mu} \sum_{v,u \in V} \sum_{d,r \in D} \left\{ \left(A_{vu}^{(d)} - \gamma_d \frac{k_v^d k_u^d}{2m_d} \right) \delta(d,r) + \delta(v,u) \sigma_v^{(d)(r)} \right\} \delta(g_v^{(r)}, g_u^{(d)}) \tag{4.8}$$

où μ dénote le facteur de normalisation, γ_d est le paramètre de résolution de modularité associé à la dimension d, δ représente le symbole de Kronecker, $\sigma_v^{(d)(r)}$ désigne le paramètre de couplage du nœud v dans la dimension d à son instance dans la dimension r, k_v^d, k_u^d dénote, respectivement, les degrés des nœuds v et u dans la dimension d, m_d est le nombre d'arêtes du multigraphe G dans la dimension d et, finalement, $g_v^{(r)}$ et $g_u^{(d)}$ représente les communautés des nœuds v et u dans les dimensions d et r respectivement. Notons que dans nos expérimentations, nous avons fixé la valeur de γ_d et celle de $\sigma_v^{(d)(r)}$ à 1, c'est la valeur par défaut telle que suggérée par Mucha *et al.* (2010). La modularité multicouche retourne des valeurs dans $[0, 1]$ où une valeur élevée (proche de 1) correspond à une bonne partition.

4.2 Expérimentations sur les réseaux synthétiques

4.2.1 Génération des réseaux

Nous utilisons le modèle de génération de réseaux artificiels défini dans Condon et Karp (2001). Le processus de génération est paramétrique au nombre de nœuds n, le nombre de communautés K, le nombre de dimensions nd, la dimensionnalité moyenne par communauté nd_r, la plage de densités internes $[\vartheta_{int_{min}}, \vartheta_{int_{max}}]$ et finalement la plage de densités externes $[\vartheta_{ext_{min}}, \vartheta_{ext_{max}}]$. En fonction des valeurs de paramètres fournis, les communautés sont générées aléatoirement à travers les dimensions sur deux étapes :

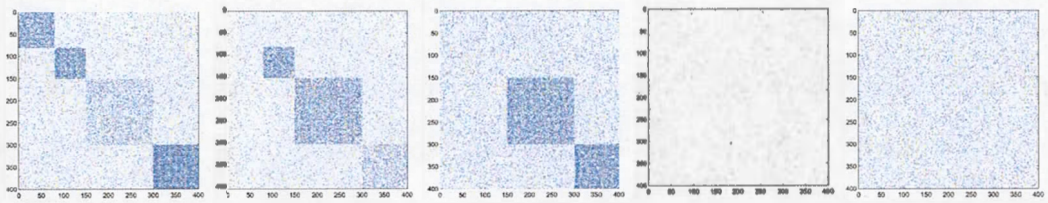
1. Dans un premier temps, les communautés sont générées sur $\frac{nd_r}{2}$ dimensions pertinentes, que l'on dénote D_{r1}. Les dimensions sont sélectionnées aléatoirement à partir de D. Pour chaque dimension sélectionnée $d \in D_{r1}$, la

matrice d'adjacence correspondante est divisée en K blocs B_{kd} de tailles variables, de sorte que chaque bloc correspond à une communauté projetée sur d. Les nœuds sont ensuite raccordés aléatoirement selon une probabilité uniformément sélectionnée à partir de l'intervalle $[\vartheta_{int_{min}}, \vartheta_{int_{max}}]$. Par la suite, les liaisons inter communautés sont générées selon une probabilité uniformément tirée de $[\vartheta_{ext_{min}}, \vartheta_{ext_{max}}]$.

2. Un deuxième sous-ensemble D_{r2} de nd_r dimensions pertinentes est ensuite sélectionné à partir de $D - D_{r1}$. Les blocs de communautés sont insérés aléatoirement sur les sous-ensembles de D_{r2} de sorte que la dimensionnalité moyenne soit proche de nd_r. Finalement, les matrices d'adjacences des dimensions restantes sont générées selon le modèle d'Erdos-Rényi avec des probabilités de génération de liens uniformément choisies de $[\vartheta_{ext_{min}}, \vartheta_{ext_{max}}]$.

Le modèle de génération que nous avons utilisé permet de produire des réseaux synthétiques où les interactions entre les nœuds diffèrent en fonction des dimensions et les appartenances aux communautés. Ce modèle permet donc de simuler diverses configurations de réseaux multidimensionnels, ce qui, à son tour, permet d'effectuer une évaluation plus objective des algorithmes considérés dans les tests.

La figure 4.1 montre un exemple de réseau synthétique construit par le générateur qui vient d'être décrit. Ce réseau est composé de 5 dimensions avec 400 nœuds répartis sur 4 communautés de différentes tailles. Le nombre de dimensions pertinentes par communauté varie entre 1 et 3. Notons que chaque figure dans cette illustration graphique représente la matrice d'adjacence associée à une dimension du réseau. Chaque bloc dans les matrices correspond à une communauté dans une dimension spécifique. Comme on peut le voir, les 4 communautés existent dans



(a) d_1 (b) d_2 (c) d_3 (d) d_4 (e) d_5

Figure 4.1: Matrices d'adjacence des dimensions d'un réseau synthétique de 5 dimensions.

différents sous-espaces de dimensions. Par exemple, la première communauté (définie par le premier bloc dans la figure 4.1a) existe dans la dimension d_1, tandis que la deuxième communauté (définie par le deuxième bloc dans la figure 4.1a et le premier bloc dans la figure 4.1b) existe dans d_1 et d_2. Finalement, tel qu'illustré par la figure 4.1d et la figure 4.1e, aucune structure de communauté n'existe dans les dimensions d_4 et d_5. Cet exemple présenté par la figure 4.1 est simple et a été fourni à titre illustratif seulement. Les réseaux générés pour l'évaluation sont, cependant, plus complexes. La prochaine section décrit les réseaux synthétiques générés ainsi que les résultats obtenus par les algorithmes comparés.

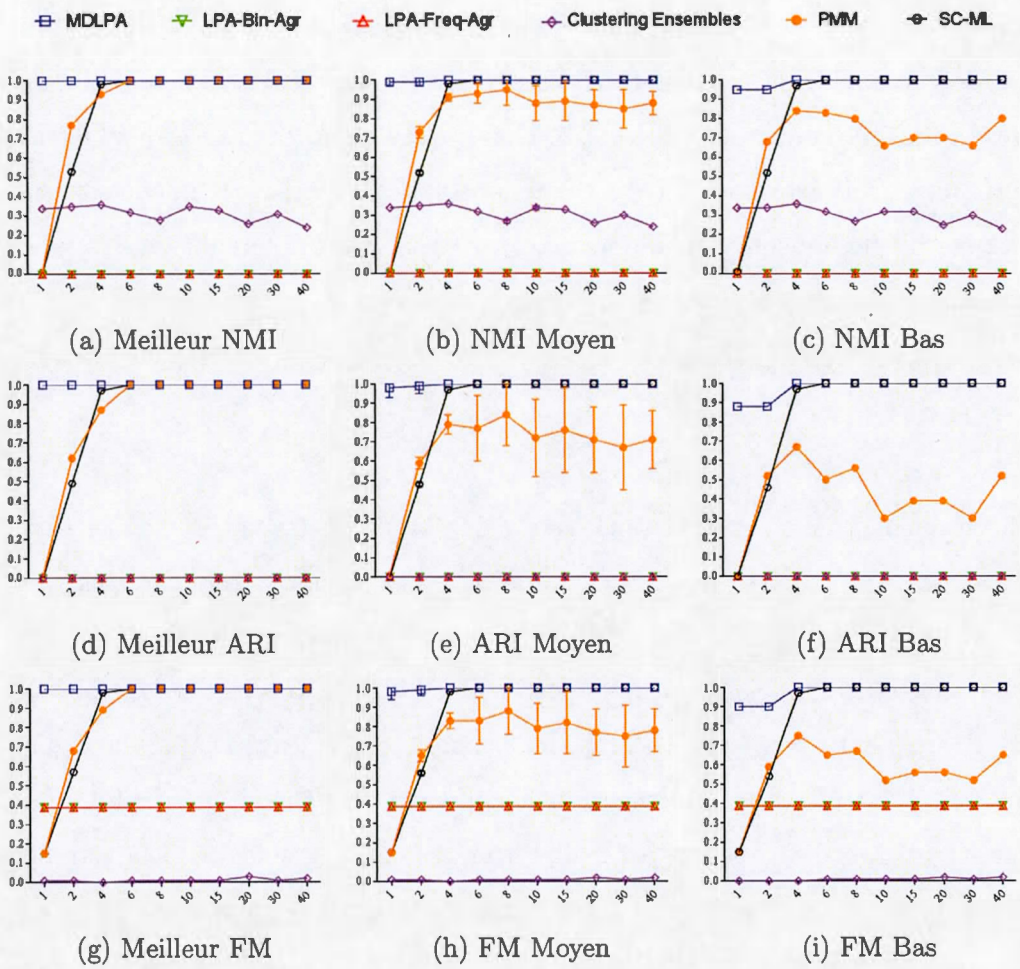
4.2.2 Résultats et discussions

L'objectif de cette section est d'analyser l'impact de la dimensionnalité moyenne nd_r sur : (1) la précision de détection de communautés et (2) la capacité de MDLPA à identifier les dimensions pertinentes réelles de chaque communauté. À cette fin, nous avons généré 10 réseaux multidimensionnels synthétiques. Chaque réseau contient $n = 3000$ nœuds et $nd = 100$ dimensions. La dimensionnalité moyenne nd_r variait entre 1 et 40 pour cent de la dimensionnalité globale de chaque réseau. Les communautés ont été générées en utilisant les paramètres suivants : les valeurs de densités de liens intra communautés ont été sélectionnées

à partir de $[0.2, 0.6]$ tandis que les densités inter communautés ont été choisies à partir de $[0, 0.022]$. Enfin, le nombre de communautés a été fixé à 7, alors que la taille de chaque communauté varie entre 10 et 20 pour cent de la taille n du réseau.

Le premier aspect étudié dans nos tests est la performance de détection des communautés par rapport au pourcentage de dimensions pertinentes. L'objectif est d'évaluer l'impact des dimensions non pertinentes sur la précision de détection des algorithmes. À cet effet, nous avons appliqué les différents algorithmes que nous avons considérés sur les 10 réseaux générés. La figure 4.2 illustre la performance des algorithmes concurrents, évaluée par NMI, ARI et FM. Ici, LPA-Bin-Agr et LPA-Freq-Agr dénotent la performance de LPA sur les réseaux agrégés en utilisant, respectivement, la stratégie d'agrégation binaire et fréquentielle. Rappelons que pour chaque réseau, nous avons exécuté chaque algorithme 10 fois et nous avons sélectionné les meilleurs, moyens et plus bas résultats.

Dans l'ensemble, les résultats révèlent que les méthodes basées sur l'agrégation de dimensions et le clustering d'ensembles ne sont pas en mesure de découvrir les vraies partitions. En ce qui concerne les techniques d'agrégations, la performance de LPA (appliqué sur le graphe agrégé) a été grandement affectée par les dimensions non pertinentes. L'algorithme retourne toujours la plus large composante connexe du réseau indépendamment de la stratégie d'agrégation ou la dimensionnalité moyenne nd_r.



(a) Meilleur NMI (b) NMI Moyen (c) NMI Bas

(d) Meilleur ARI (e) ARI Moyen (f) ARI Bas

(g) Meilleur FM (h) FM Moyen (i) FM Bas

Figure 4.2: Résultats sur les réseaux synthétiques. L'axe des X représente la dimensionalité moyenne tandis que celui des Y correspond aux valeurs des indices.

D'autre part, la précision des techniques de clustering d'ensembles ne s'est pas améliorée en dépit du nombre croissant de dimensions pertinentes. Cela peut être expliqué par le fait que les partitions identifiées (suite au processus de consensus) ont été affectées par les dimensions non pertinentes pour lesquelles l'algorithme de Louvain tend à sur estimer le nombre de communautés. Nous avons observé que, pour $nd_r = 40\%$, le clustering d'ensembles génère, en moyenne, 800 petites

communautés de taille allant de 1 à 7 nœuds. La précision était encore pire pour $nd_r \leq 6\%$, pour lequel le clustering d'ensembles produit environ 2500 communautés (majoritairement des singletons). Le fait que l'algorithme accorde la même importance aux dimensions du réseau rend la recherche d'une partition de consensus assez difficile, puisque les communautés existent dans différents sous espaces de dimensions (et non pas dans l'espace de dimensions en entier).

En revanche, comme montré sur la figure 4.2, MDLPA démontre une meilleure performance peu importe les variations de nd_r. Ces résultats sont attribués au mécanisme de sélection de dimensions qui permet à notre algorithme de focaliser la recherche sur les dimensions pertinentes et ainsi localiser les régions de haute densité à travers l'espace multidimensionnel du réseau. D'après la figure 4.2, MDLPA est plus précis même pour les plus petites valeurs de nd_r. En effet, pour $nd_r \leq 6\%$, la différence entre les algorithmes comparés devient plus apparente et l'impact des dimensions non pertinentes est beaucoup plus prononcé. Bien que légèrement affecté, MDLPA démontre plus d'immunité à la présence d'un grand nombre de dimensions non pertinentes. On attribue cela aux règles de sélection de dimensions pertinentes qui permettent d'identifier les ensembles D_k d'une manière appropriée. En outre, MDLPA est plus robuste que les algorithmes concurrents, tel que démontré par les faibles valeurs d'écart-type (barres d'erreurs sur les tracés dans les figures 4.2b, 4.2e et 4.2h). Dans l'ensemble, on peut remarquer qu'il n'y a pas une grande différence entre les meilleurs et plus bas résultats de MDPLA. Bien que la stratégie d'optimisation adoptée soit basée sur le principe de propagation d'étiquettes, l'algorithme a tendance à être plus stable que ses concurrents. Ceci est principalement dû à l'intégration des dimensions pertinentes dans la fonction objective de MDLPA.

Lorsque $nd_r \geq 6\%$, PMM et SC-ML atteignent leur meilleure performance. Bien que SC-ML semble être plus robuste, les deux approches arrivent à identifier les partitions originales lorsque le bon nombre de communautés est fourni. Cependant, lorsque $nd_r = 1$, SC-ML et PMM n'arrivent plus à identifier la partition originale. En effet, nous avons constaté que l'attribution des nœuds aux communautés a été était faite d'une manière aléatoire. Par ailleurs, nous avons constaté que le clustering d'ensembles semble être le plus affecté par les dimensions non pertinentes. En moyenne, l'algorithme produisait plus de 2 800 communautés indépendamment de la valeur de nd_r. Ce nombre élevé est principalement causé par les dimensions non pertinentes où la méthode de Louvain, appliquée aux différentes dimensions du réseau, surévalue le nombre de communautés. Ainsi, nous pouvons conclure que le clustering d'ensembles ne serait capable d'identifier les bonnes partitions qu'en présence des communautés dans l'ensemble des dimensions du réseau. En résumé, les résultats obtenus démontrent que les techniques d'agrégation de dimensions, de l'intégration de caractéristiques, ainsi que le clustering d'ensembles, rencontrent des difficultés en présence de dimensions non pertinentes. MDPLA demeure toutefois robuste même dans des situations qui impliquent la présence de communautés dans des sous-espaces trop réduits.

Après avoir analysé l'impact de la variation de la valeur de dimensionnalité moyenne des communautés, nous évaluons maintenant la capacité de l'approche proposée à identifier les dimensions pertinentes associées aux communautés détectées. Notons que dans cette évaluation, nous n'avons considéré que MDLPA, car les méthodes concurrentes n'offrent aucun mécanisme permettant de déterminer les dimensions pertinentes aux communautés détectées. Afin d'évaluer la similarité entre les dimensions réellement pertinentes (c-à-d., les dimensions utilisées pour l'injection des structures de communautés lors du processus de génération)

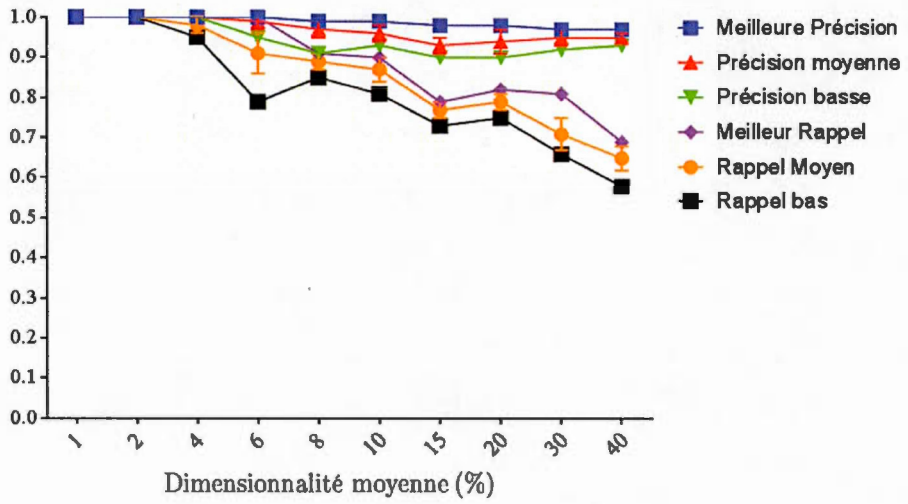


Figure 4.3: Précision de sélection des dimensions pertinentes D_k par rapport à la dimensionalité moyenne nd_r.

et celles identifiées par MDLPA, nous avons utilisé les métriques de précision et de rappel. En fait, pour chaque communauté, la précision mesure le rapport entre le nombre de dimensions réellement pertinentes choisies par l'algorithme et le nombre de dimensions identifiées. Le rappel correspond au nombre de dimensions pertinentes identifiées divisé par le nombre réel de dimensions pertinentes. La valeur rapportée pour une partition correspond à la moyenne sur l'ensemble des communautés détectées.

La figue 4.3 illustre les valeurs de précision et de rappel. Comme on peut le constater, les valeurs élevées de précision confirment l'efficacité du mécanisme de sélection de dimensions pertinentes de MDLPA. Par contre, on remarque que le rappel diminue presque linéairement avec nd_r. Ce comportement est principalement causé par la règle de sélection de dimensions pertinentes qui tend à défavoriser les dimensions dont les liens ne dominent suffisamment pas les voisinages des nœuds. En effet, nous avons constaté que les dimensions où la densité de liens

s'avérait relativement faible étaient parfois ignorées en faveur des dimensions plus denses, d'où la baisse constante dans les valeurs de rappel. En revanche, quand $nd_r \leq 6\%$, MDLPA sélectionne de façon correcte les sous-espaces de dimensions pertinentes.

4.3 Expérimentations sur les réseaux réels

Dans cette section, nous évaluons l'efficacité de notre approche sur quatre réseaux réels : (1) le réseau social du département d'informatique de l'Université Aarhus, (2) le réseau de collaboration de l'observatoire Pierre Auger, (3) le réseau multidimensionnel Foursquare, et (4) le réseau d'interactions de protéines de la mouche à fruit Drosophila Melanogaster. Contrairement aux réseaux synthétiques qui offrent un environnement contrôlé où le nombre de communautés est connu à l'avance, une connaissance préalable sur les structures de communautés dans les réseaux réels est souvent manquante. De ce fait, nous n'avons retenu que les algorithmes qui ne nécessitent pas une connaissance préalable sur le nombre de communautés à identifier. De même, en l'absence d'une partition de référence (ground truth), nous n'avons considéré que les critères externes pour l'évaluation. Une description de chaque réseau utilisé dans l'expérimentation ainsi que l'analyse des résultats obtenus par MDLPA et les autres algorithmes concurrents est présentée dans ce qui suit.

4.3.1 Le réseau social du département d'informatique de l'Université Aarhus

C'est un réseau qui contient cinq dimensions représentant des interactions sociales entre les employés du département d'informatique de l'Université Aarhus (Magnani *et al.*, 2013). Le réseau original se compose de 61 employés (assistants administratifs, professeurs, associés, doctorants et post-doctorants) affectés à 8 groupes de travail. Afin de rendre les tests plus conviviaux, nous avons sélectionné

Tableau 4.1: Les dimensions du réseau social du département d'informatique de l'Université Aarhus.

Dimension	Liens	Densité
1. Déjeuner ensembles	162	0.1222
2. Amitié sur Facebook	96	0.0724
3. Coauteurs	21	0.0158
4. Loisir	87	0.0656
5. Travailler ensemble	114	0.0860

un sous-ensemble de 52 nœuds en éliminant 6 nœuds pour lesquels les groupes de travail sont inconnus, 2 nœuds appartenant à plusieurs groupes et, finalement, un nœud spécial faisant un groupe de travail à part. Le tableau 4.1 résume les statistiques de chaque dimension du réseau.

La figure 4.4 montre les résultats obtenus par les différents algorithmes, tels qu'évalués par la redondance ρ, la densité multidimensionnelle de communauté MCD, et la modularité multicouche Q. Notons que, tel que discuté précédemment, nous utilisons les deux ensembles DF_k (l'ensemble des dimensions retrouvées dans la communauté C_k), et D_k (l'ensemble des dimensions pertinentes identifiées pour C_k) pour le calcul de la redondance ρ et la densité MCD pour notre algorithme. L'objectif est d'illustrer l'impact des dimensions pertinentes sur la qualité des résultats. Dans la figure 4.4, MDLPA (DF) indique que ρ et MCD ont été calculé en considérant les dimensions retrouvées dans la communauté DF_k, tandis que MDLPA (RD) indique que les deux métriques ont été calculées en utilisant les dimensions pertinentes D_k identifiées par MDLPA. En revanche, nous n'avons utilisé que les dimensions retrouvées dans les communautés retournées par les algorithmes concurrents dans le calcul de la redondance ρ et la densité MCD (à

	Redondance ρ			MCD			Modularité Multicouche			Nombre moyen de coms
Algorithme	bas	moy ± écart.t	haut	bas	moy ± écart.t	haut	bas	moy ± écart.t	haut	
MDLPA(DF)	0.41	0.44 ± 0.04	0.51	0.42	0.47 ± 0.04	0.55	0.63	0.65 ± 0.01	0.66	7
MDLPA(RD)	0.48	0.57 ± 0.09	0.72	0.53	0.60 ± 0.06	0.71				
LPA-Bin-Agr	0.25	0.27 ± 0.04	0.38	0.07	0.15 ± 0.09	0.28	0.30	0.34 ± 0.07	0.53	2
LPA-Freq-Agr	0.37	0.50 ± 0.09	0.55	0.28	0.42 ± 0.10	0.48	0.53	0.55 ± 0.00	0.55	5
Clustering-Ens	0.76	0.80 ± 0.03	0.83	0.67	0.69 ± 0.01	0.71	0.35	0.36 ± 0.02	0.41	30

Figure 4.4: Résultats obtenus sur le réseau social du département d'informatique de l'Université Aarhus.

rappeler que ces algorithmes, n'offrent aucun mécanisme de sélection de dimensions pertinentes).

D'après la figure 4.4, on constate que l'approche de clustering d'ensembles atteint les valeurs les plus élevées de redondance et de densité MCD. Cela peut être attribué au fait que le clustering d'ensembles retourne un nombre relativement élevé de communautés de petite taille. En moyenne, cette approche produisait 30 communautés dont les tailles varient entre 2 et 3 nœuds. Il est à rappeler que ρ et MCD évaluent la qualité du partitionnement au niveau local, et ce en considèrent que les connexions intra communautés. Par conséquent, un grand nombre de petites communautés va se traduire par de grandes valeurs de la redondance ρ et la densité MCD. Par ailleurs, en ce qui concerne la modularité multicouche, le clustering d'ensembles rapporte un score relativement faible. Cela est dû au fait que l'algorithme retourne un grand nombre de petites communautés. En fait, le nombre élevé des communautés identifiées par le clustering d'ensembles est attribué à la dimension coauteurs qui ne contribue que par 21 arêtes (voir le tableau 4.1). Le résultat de clustering d'ensembles sur ce jeu de données confirme sa sensibilité aux dimensions non pertinentes, telle qu'observée dans la section précédente (expérimentations sur les réseaux synthétiques).

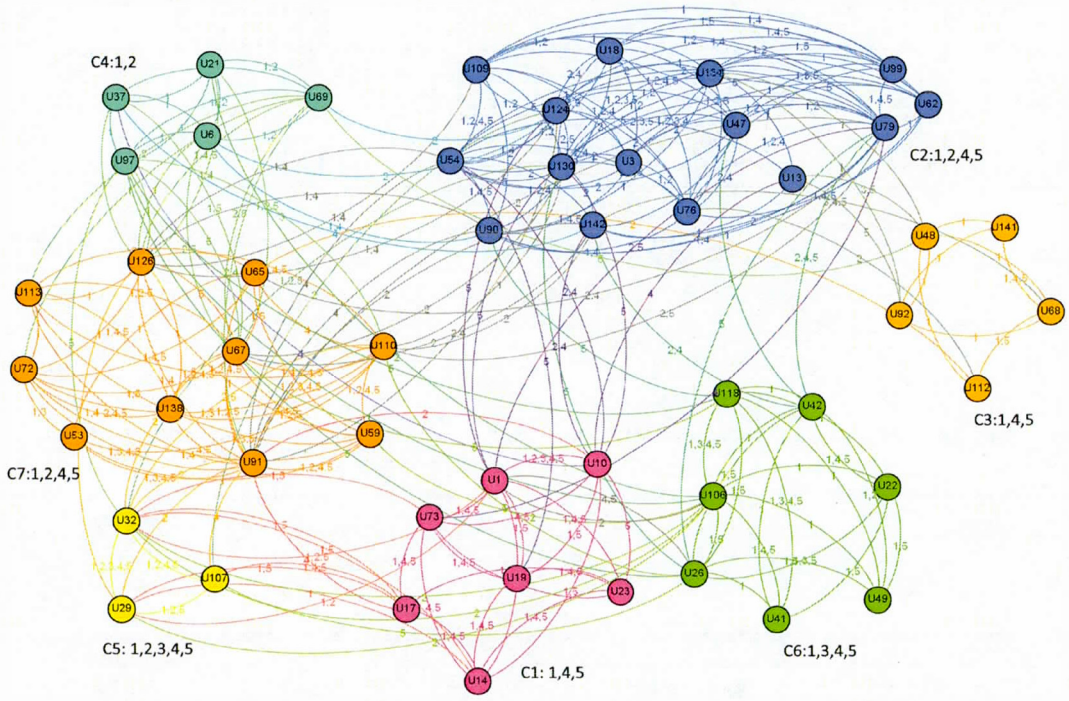


Figure 4.5: La partition identifiée par MDLPA sur le réseau social du département d'informatique de l'Université Aarhus.

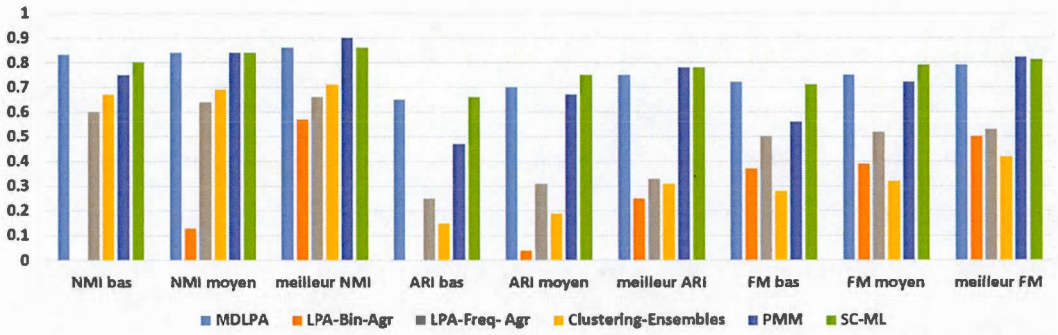
À partir des résultats illustrés sur la figure 4.4, on peut constater que MDLPA fournit les meilleures valeurs de modularité. Contrairement aux autres algorithmes, la combinaison des valeurs des trois métriques (ρ, MCD et Q) rapportées par MDLPA suggère que notre algorithme identifie de bonnes structures de communautés. De même, à partir de la figure 4.4 on constate l'amélioration des valeurs de redondance ρ et MCD sur les sous-espaces de dimensions pertinentes (voir les résultats de MDLPA-DF versus ceux de MDLPA-RD). L'amélioration des scores de ces deux métriques sur les sous-espaces sélectionnés par l'approche proposée suggère qu'elle peut éliminer d'une manière effective les dimensions dont la contribution est insignifiante à la formation des communautés.

Pour avoir une idée sur le résultat de MDLPA, la figure 4.5 illustre une partition composée de 7 communautés (ainsi que leurs dimensions pertinentes) identifiées par notre algorithme. Le réseau a été projeté sur une seule dimension pour une meilleure lisibilité. En fait, pour des raisons de simplicité, nous avons remplacé les arêtes reliant n'importe qu'elle paire de nœuds par une seule arête portant les identifiants des dimensions de liaison (1 : Déjeuner, 2 : Facebook, 3 : Coauteurs, 4 : Loisir et 5 : Travail). Comme nous pouvons le constater, MDPLA découvre des communautés qui existent dans différents sous-espaces de dimensions.

Nous avons également comparé la performance des algorithmes sélectionnés en considérant les affiliations aux groupes de travail comme une partition de référence possible. L'hypothèse retenue est que les employés du même groupe tendent à développer plus de relations avec les membres du même groupe qu'avec les membres des autres groupes. L'objectif est donc d'évaluer s'il est possible d'identifier les groupes de travail du département en partant des interactions entre les membres.

La figure 4.6 illustre la performance des algorithmes comparés par rapport à la partition latente présumée. Notons que pour ce cas bien particulier, nous avons considéré PMM et SC-ML dans la comparaison vu que le nombre de communautés (groupes de travail) est disponible. Dans ces expérimentations, nous avons fixé le nombre de communautés à identifier par PMM et SC-ML à 7, ce qui correspond au nombre de groupes de travail dans le réseau étudié.

Les résultats de MDLPA, PMM et SC-ML suggèrent qu'effectivement, la structure latente des communautés est corrélée avec les groupes de travail du département. Bien que SC-ML semble être légèrement meilleur, les résultats de ces trois approches sont relativement comparables. Ceci est vrai quand le bon nombre



	NMI			ARI			FM		
Algorithm	**bas**	**moy ± écart.t**	**haut**	**bas**	**moy ± écart.t**	**haut**	**bas**	**moy ± écart.t**	**haut**
MDLPA	0.83	0.84 ± 0.01	0.86	0.65	0.70 ± 0.04	0.75	0.72	0.75 ± 0.03	0.79
LPA-Bin-Agr	0.00	0.13 ± 0.18	0.57	0.00	0.04 ± 0.08	0.25	0.37	0.39 ± 0.04	0.50
LPA-Freq- Agr	0.60	0.64 ± 0.03	0.66	0.25	0.31 ± 0.03	0.33	0.50	0.52 ± 0.01	0.53
Clustering-Ens	0.67	0.69 ± 0.01	0.71	0.15	0.19 ± 0.04	0.31	0.28	0.32 ± 0.04	0.42
PMM	0.75	0.84 ± 0.05	0.90	0.47	0.67 ± 0.12	0.78	0.56	0.72 ± 0.10	0.82
SC-ML	0.80	0.84 ± 0.02	0.86	0.66	0.75 ± 0.04	0.78	0.71	0.79 ± 0.03	0.81

Figure 4.6: Résultats obtenus sur le réseau social du département d'informatique de l'Université Aarhus vis-à-vis la partition présumée.

de communautés est fourni par l'utilisateur à PMM et SC-ML. MDLPA, en revanche, peut identifier les communautés sans aucune intervention de l'utilisateur. Par ailleurs, comme illustré dans la figure 4.6, on constate que les résultats des approches fondées sur l'agrégation de dimensions et le clustering d'ensembles sont moins bons comparativement à ceux de MDLAP, PMM et SC-ML.

Une inspection attentive de la figure 4.5 suggère que la partition retournée par notre algorithme n'est pas très différente des groupes de travail des employés. Dans la figure 4.5, la communauté C_6 correspond à l'unité de travail G_6 alors que la communauté C_1 correspond à l'unité G_1, avec l'exception du nœud $U17$ qui, normalement, devrait appartenir au groupe de travail G_5 (communauté C_5). Nous avons constaté que $U17$ interagit plus fréquemment avec les membres de G_1 ce qui justifie son attribution à C_1. La même observation est vraie pour les nœuds non affectés à leurs groupes originaux, notamment dans le cas de G_3 et G_4 (C_3

et C_4 respectivement). Par ailleurs, les dimensions pertinentes sélectionnées par notre approche permettent de mieux comprendre les canaux d'interaction fondamentaux au sein des communautés détectées. À titre d'exemple, les membres de G_1 ont tendance à moins collaborer à la rédaction d'articles scientifiques et évitent complètement les interactions sur Facebook. La même remarque s'applique sur les membres de G_6 qui ont tendance à éviter le contact direct sur Facebook. En revanche, les employés de l'unité G_5 s'impliquent activement sur les cinq dimensions, ce qui indique des relations professionnelles et sociales bien établies.

4.3.2 Le réseau de collaboration de l'observatoire Pierre Auger

Nous avons analysé le réseau de collaboration des scientifiques de l'observatoire Pierre Auger (De Domenico *et al.*, 2015a), un réseau de 514 chercheurs travaillant sur 16 tâches à l'observatoire Pierre Auger pour l'étude des rayons cosmiques à haute énergie. Chaque tâche représente une dimension de collaboration dans laquelle deux chercheurs se connectent s'ils corédigent un rapport scientifique. Le réseau se caractérise par la présence de nombreuses petites composantes connexes et une grande proportion de dimensions non pertinentes, c.-à-d., des dimensions avec faibles densités telles qu'illustrées par le tableau 4.2.

Tableau 4.2: Les dimensions du réseau de collaboration de l'observatoire Pierre Auger.

Dimension/Tâche	Liens	Densité	Dimension/Tâche	Liens	Densité
1.Neutrinos	60	0.0005	9.Spectre	80	0.0006
2.Détecteur	550	0.0042	10.Photons	21	0.0002
3.Améliorations	5433	0.0412	11.Atmosphérique	51	0.0004
4.Anistropie	76	0.0006	12.Reconstruction SD	211	0.0016
5.Source ponctuelle	105	0.0008	13.Interaction	53	0.0004
6.Décomposition	191	0.0014	14.Exotiques	18	0.0001
7.Horizontale	61	0.0005	15.Magnétiques	38	0.0003
8.Reconstruction	184	0.0014	16.Astrophysique	21	0.0002

La figure 4.7 illustre la partition identifiée par MDLPA sur ce réseau. Afin de faciliter l'interprétation, nous avons projeté le réseau sur une seule dimension en remplaçant l'ensemble des dimensions reliant n'importe quelle paire par une seule arête. La partition retournée par MDLPA révèle une organisation spécialisée dans laquelle chaque groupe de scientifiques se concentre sur un nombre restreint de tâches. En effet, nous avons trouvé que la taille des sous-espaces de dimensions pertinentes varie entre 1 à 6 dimensions, ce qui correspond à une dimensionnalité moyenne de 6 à 38 pour cent de l'espace multidimensionnel. Par ailleurs, la majorité des communautés détectées se focalise sur une seule tâche. À titre d'exemple, les communautés C_1 et C_2 représentent des équipes travaillant sur la tâche Améliorations tandis que C_3 et C_4 correspondent à des groupes spécialisés, respectivement, dans les détecteurs et la reconstruction hybride. Notons que ces quatre communautés (à savoir, C_1, C_2, C_3 et C_4) comptent pour 38% de la taille du réseau. Il est également à noter qu'en revanche, plusieurs communautés identifiées se révèlent multidisciplinaires. La communauté C_5 par exemple, correspond

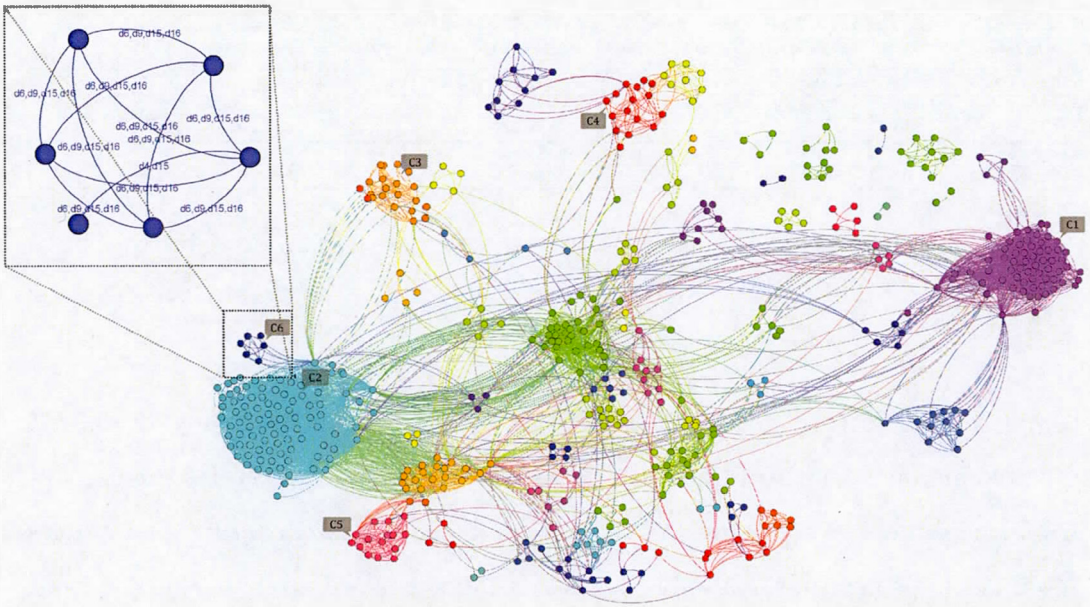


Figure 4.7: Une partition identifiée par MDLPA sur le réseau de collaboration de l'observatoire Pierre Auger.

à une équipe de 14 chercheurs qui travaillent sur les détecteurs et la reconstruction hybride tandis que les membres de C_6 se spécialisent dans la décomposition de masse, l'analyse de spectre, et les scénarios magnétiques et astrophysiques. MDLPA signale également la présence de plusieurs communautés déconnectées de la composante connexe du réseau (certaines sont d'une taille aussi petite que 2 nœuds).

La figure 4.8 montre les résultats des algorithmes comparés. Tel qu'illustré, on peut voir que le clustering d'ensembles n'arrive pas à identifier une structure de communautés adéquate. L'algorithme retourne un nombre de communautés proche de celui du nombre de nœuds. À l'exception de quelques groupes, la plupart des communautés identifiées représentent un ou deux chercheurs. De même, parmi les communautés restantes, plusieurs correspondent à des paires de nœuds

	Redondance ρ			MCD			Modularité Multicouche			Nombre moyen de coms
Algorithme	bas	moy ± écart.t	haut	bas	moy ± écart.t	haut	bas	moy ± écart.t	haut	
MDLPA (DF)	0.39	0.41 ± 0.02	0.44	0.69	0.70 ± 0.01	0.72	0.83	0.84 ± 0.01	0.85	66
MDLPA (RD)	0.63	0.67 ± 0.02	0.69	0.78	0.81 ± 0.01	0.83				
LPA-Bin-Agr	0.30	0.36 ± 0.03	0.39	0.60	0.65 ± 0.03	0.69	0.77	0.77 ± 0.00	0.77	46
LPA-Freq-Agr	0.35	0.40 ± 0.03	0.42	0.62	0.66 ± 0.02	0.68	0.77	0.77 ± 0.00	0.77	45
Clustering-Ens	1.00	1.00 ± 0.00	1.00	0.52	0.65 ± 0.07	0.76	0.68	0.68 ± 0.00	0.68	493

Figure 4.8: Résultats obtenus sur le réseau de collaboration de l'observatoire Pierre Auger.

complètement déconnectées. Ceci explique la différence entre les valeurs de densité et la redondance qui, théoriquement, devraient être égales à 1 pour les communautés formées d'une seule paire. Les résultats obtenus par le clustering d'ensembles sont principalement dus à la faible densité des dimensions du réseau. En effet, pour ce réseau particulier, la dimensionnalité moyenne était proche de 10 pour cent ce qui correspond à une moyenne de 1.6 tâche par équipe. La haute spécialisation des équipes de recherche fait en sorte que l'activité sur les autres dimensions soit faible. Ceci explique le grand nombre de petites communautés identifiées par la méthode de Louvain qui, à son tour, affecte la qualité des résultats obtenus par le clustering d'ensembles.

La figure 4.8 suggère que LPA-Bin-Agr et LPA-Freq-Agr rapportent de bons résultats sur le réseau de l'observatoire Pierre Auger. Ces résultats peuvent être attribués à la nature modulaire de ce réseau (De Domenico *et al.*, 2015a) qui fait en sorte que les deux stratégies d'agrégation binaire et fréquentielle produisent un graphe agrégé hautement modulaire. En revanche, comme on peut le constater à travers la figure 4.8, la performance de MLDPA, telle que mesurée par ρ, MCD et Q, est meilleure que celle de LPA-Bin-Agr et LPA-Freq-Aggr. Par ailleurs, les améliorations enregistrées sur les deux métriques de redondance ρ, et de densité MCD sur les sous-espaces sélectionnés par MDLPA (voir les résultats de MDLPA

(DF) versus MDLPA (RD)) confirment l'efficacité de son mécanisme de sélection de dimensions pertinentes et sa capacité à fournir des connaissances supplémentaires. Cela permet à l'analyste d'avoir une idée sur les centres d'intérêt au sein des différents groupes.

4.3.3 Le réseau multidimensionnel Foursquare

Foursquare est un réseau géosocial qui permet aux utilisateurs de partager leurs positions avec leurs amis à travers un système d'enregistrement sur divers types d'endroits tel que restaurants, monuments, hôtels ou aéroports. Chaque endroit est identifié dans la plateforme de Foursquare à travers une page qui offre aux membres la possibilité d'interagir et de partager leurs avis. Aux fins d'évaluation, nous avons construit un réseau multidimensionnel de 3 488 utilisateurs (nœuds) qui interagissent sur 4 dimensions. Dans un premier temps, nous avons utilisé l'API de Twitter pour récupérer un gazouillis qui indique des coordonnées GPS d'un endroit aléatoire dans la ville de New York. Cela nous permet d'identifier l'utilisateur qui a posté ce gazouillis. Par la suite, nous avons exploré les informations publiquement accessibles à partir du profil Twitter de l'utilisateur en question pour accéder à son profil Foursquare. L'information disponible sur le profil Foursqure est utilisée pour récupérer la liste d'amis, les endroits aimés et les avis sur les sites visités. Pour chaque ami, nous avons collecté les mêmes informations par le biais d'un crawler que nous avons implémenté.

Afin de construire le réseau multidimensionnel, nous avons considéré 4 types d'interactions : (1) Amitié sur Foursquare : comme son nom l'indique, la dimension d'amitié sur Foursquare connecte deux utilisateurs (nœuds) s'ils sont amis sur Foursquare ; (2) Relation *follower* ou *following* sur Twitter : par le biais de cette dimension, nous visons à créer un lien additionnel entre deux amis sur Foursquare

Tableau 4.3: Les dimensions du réseau multidimensionnel Foursquare.

Dimension	Liens	Densité
1. Amitié sur Foursquare	19865	0.0033
2. *follower* ou *following* sur Twitter	2133	0.0004
3. Visiter et aimer le même endroit	6994	0.0012
4. Feedback sur un endroit visité	2380	0.0004

s'ils disposent d'une relation de *follower* ou *following* sur Twitter. Bien évidemment, ici, nous ne considérions que les utilisateurs ayant un compte Twitter relié à leur profil Foursquare ; (3) Visiter et aimer le même endroit : à travers cette dimension, deux utilisateurs sont connectés si et seulement s'ils ont visité et aimé le même endroit ; et finalement (4) Feedback sur un endroit visité : deux utilisateurs sont connectés à travers cette dimension s'ils laissent un commentaire sur un endroit qu'ils ont tous les deux visités. Le tableau 4.3 résume les statistiques des quatres dimensions.

La figure 4.9 représente une partition identifiée par MDLPA sur le réseau Foursquare. Dans notre investigation du résultat obtenu, nous avons constaté que les dimensions 2 (relation follower ou following sur Twitter) et 4 (feedback sur un endroit visité) n'offrent aucune structure de communautés apparentes. Ceci est attribué à la faible densité de liens au sein de ces deux dimensions (voir le tableau 4.3). D'autre part, nous avons constaté que la dimension 3 (visiter et aimer le même endroit) contribue à la formation de la plus grande communauté identifiée par MDLPA (cette communauté est identifiée par l'étiquette C_3 dans la figure 4.9). Nous avons également constaté que la dimension 1 (amitié sur Foursquare) contribue à la formation de la plupart des communautés identifiées. En effet, un grand nombre de communautés détectées par MDLPA représentent

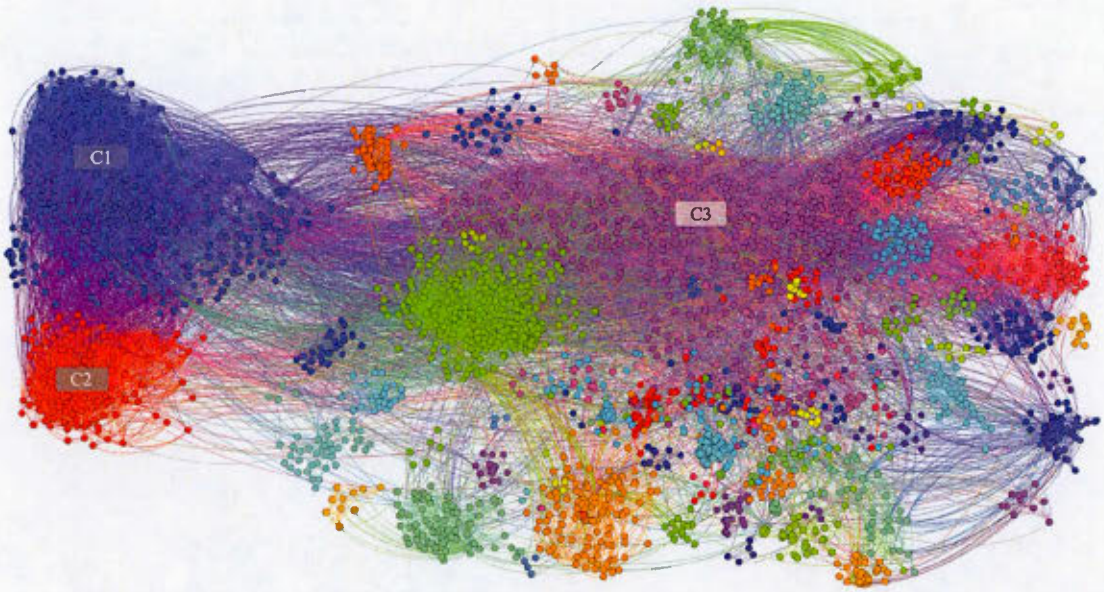


Figure 4.9: La partition identifiée par MDLPA sur le réseau multidimensionnel Foursquare.

des groupes d'amis sur Foursquare. Par exemple, la communauté C_1 dans la figure 4.9 représente 359 amis dans Foursquare, tandis que C_2 correspond à un autre groupe de 151 amis.

La figure 4.10, illustre les meilleurs, les moyens et les plus bas résultats obtenus par chaque algorithme considéré dans la comparaison. Comme nous pouvons le constater, le clustering d'ensembles génère un nombre important de petites communautés (des communautés avec un à trois nœuds au maximum). Similairement aux résultats précédents, le fait que ce dernier génère un très grand nombre de petites communautés contribuera à l'augmentation des valeurs de la redondance ρ et de la densité MCD. En revanche, le clustering d'ensembles rapporte les valeurs de modularité les plus basses. Nous avons également constaté que les deux stratégies basées sur l'agrégation ont tendance à générer des partitions principalement do-

	Redondance ρ			MCD			Modularité Multicouche			Nombre moyen de coms
Algorithme	bas	moy ± écart.t	baut	bas	moy ± écart.t	haut	bas	moy ± écart.t	haut	
MDLPA (DF)	0.29	0.32 ± 0.03	0.37	0.39	0.40 ± 0.02	0.45	0.54	0.57 ± 0.01	0.57	105
MDLPA (RD)	0.52	0.56 ± 0.04	0.61	0.43	0.44 ± 0.02	0.49				
LPA-Bin-Agr	0.10	0.17 ± 0.04	0.23	0.57	0.62 ± 0.03	0.65	0.32	0.34 ± 0.01	0.35	68
LPA-Freq-Agr	0.29	0.32 ± 0.03	0.36	0.60	0.63 ± 0.03	0.66	0.35	0.36 ± 0.01	0.38	73
Clustering-Ens	0.80	0.86 ± 0.05	0.94	0.13	0.14 ± 0.01	0.15	0.25	0.25 ± 0.00	0.26	2639

Figure 4.10: Résultats obtenus sur le réseau multidimensionnel Foursquare.

minées par une grande communauté et de nombreuses petites communautés. Ceci explique les valeurs relativement élevées de MCD atteintes par ces approches. Enfin, tel qu'illustré par la figure 4.10, notre algorithme rapporte les valeurs les plus élevées de modularité et atteint des valeurs bien acceptables de ρ et de MCD. En outre, les améliorations constatées sur les valeurs de la redondance et la densité (voir les résultats de MDLPA (DF) versus MDLPA (RD)) confirment, encore une fois, l'efficacité du mécanisme de sélection des dimensions pertinentes de notre algorithme.

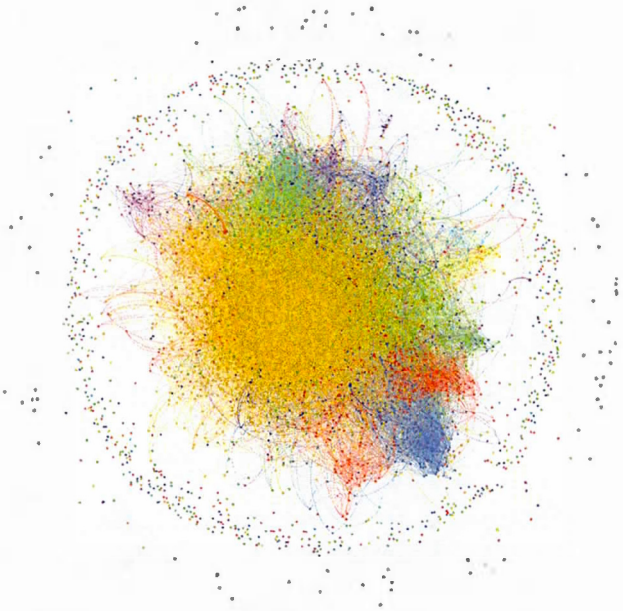
4.3.4 Le réseau d'interactions de protéines de Drosophila Melanogaster

La découverte des fonctions biologiques des protéines hautement interactives constitue un domaine d'application intéressant des algorithmes de détection de communautés (Chen et Yuan, 2006). Plusieurs travaux ont démontré que les molécules qui interagissent souvent ensemble sont associées au même phénomène biologique. Ainsi, chaque module fonctionnel peut être identifié à l'aide d'une communauté distincte. Dans cette section, nous proposons d'analyser un réseau multidimensionnel de 8 215 protéines (nœuds) de la mouche à fruit Drosophila Melanogaster (De Domenico *et al.*, 2015b; Stark *et al.*, 2006). Le réseau est constitué de sept dimensions telles qu'illustrées par le tableau 4.4.

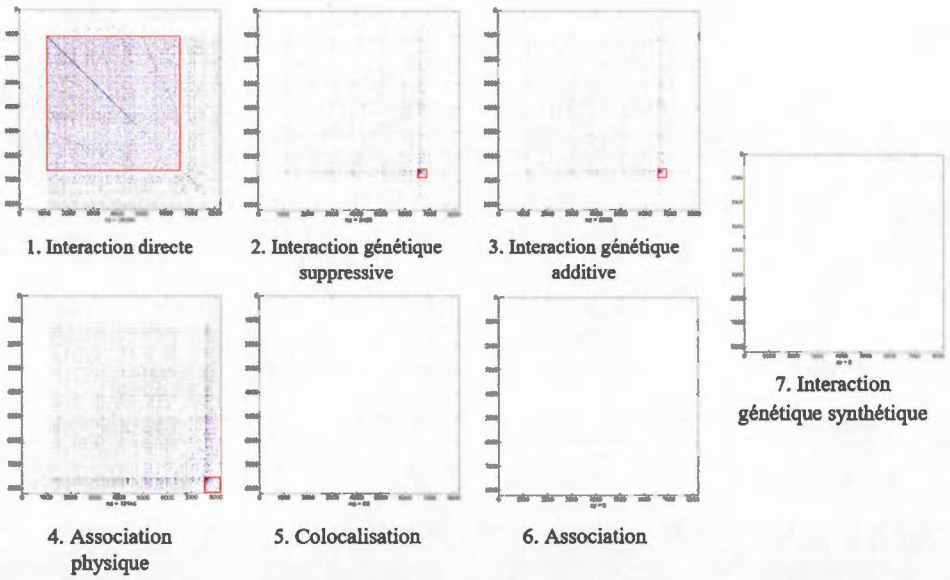
Tableau 4.4: Les dimensions du réseau d'interactions de protéines de Drosophila Melanogaster.

Dimension/Type	# Liens	Densité
1. Interaction Directe	14142	0.0002
2. Interaction Génétique Suppressive	1733	0.0000
3. Interaction Génétique Additive	1330	0.0000
4. Association Physique	6573	0.0001
5. Colocalisation	33	0.0000
6. Association	4	0.0000
7. Interaction Génétique Synthétique	4	0.0000

La figure 4.11a montre la partition identifiée par MDLPA. La ceinture autour de la composante connexe géante représente les petites communautés de protéines. Combinées, ces communautés comptent pour 1 099 protéines, ce qui correspond à 13 pour cent de la taille du réseau. Ceci explique le nombre relativement élevé des communautés détectées par LPA-Bin-Agr, LPA-Freq-Agr et MDLPA (le nombre de communautés identifiées par chaque algorithme est spécifié dans la figure 4.12). Le clustering d'ensembles semble être plus affecté par les dimensions non pertinentes et génère souvent des singletons (parmi 8 215 nœuds du réseau, le clustering d'ensembles produit, en moyenne, 8 211 communautés). Les matrices d'adjacence illustrées dans la figure 4.11b suggèrent la présence de trois dimensions non pertinentes (la dimension 5 : colocalisation, la dimension 6 : association, et la dimension 7 : les interactions génétiques synthétiques). En effet, il intéressant de mentionner que ces trois dimensions ensemble ne contribuent que 41 liens à travers l'ensemble du réseau. Notons également que la bande blanche entourant les régions ombrées des matrices d'adjacence représente la ceinture des singletons dans la figure 4.11a.



(a) La partition identifiée par MDLPA.



(b) Les matrices d'adjacence des dimensions du réseau.

Figure 4.11: La partition identifiée par MDLPA sur le réseau d'interaction de protéines de Drosophila Memanogaster.

Tel qu'illustré sur les matrices d'adjacence dans la figure 4.11b, MDLPA découvre trois communautés qui se caractérisent par différents sous-ensembles de dimensions pertinentes. Une inspection visuelle minutieuse de cette figure permet d'observer l'existence de trois blocs de communautés sur les quatre premières matrices d'adjacence (associées aux dimensions 1, 2, 3, et 4). Dans la figure 4.11b, le large bloc encadré en rouge sur la première matrice d'adjacence correspond à une communauté de 5 019 protéines qui interagissent à travers la première dimension. Les deux petits blocs encadré dans les matrices d'adjacence associées aux dimensions 2 et 3 (dimension d'interactions génétiques additives et celle d'interactions suppressives) représentent une autre communauté composée de 122 protéines. Enfin, le petit bloc dense dans la partie inférieure droite de la matrice d'adjacence de la dimension d'association physique est liée à une communauté de 275 protéines.

En ce qui concerne la performance selon les métriques externes, nous observons à partir de la figure 4.12 que MDLPA dépasse LPA-Bin-Agr et LPA-Freq-Agr sur la redondance et obtient des résultats comparables sur les métriques de densité et modularité multicouche. Quant au clustering d'ensembles, l'approche montre, de nouveau, le même comportement déjà observé sur réseaux précédents. En effet, le clustering d'ensembles génère un nombre de communautés élevé qui avoisine la taille du réseau analysé. Tel que discuté précédemment, cela est causé principalement par la présence des dimensions non pertinentes qui affectent le processus d'identification des communautés.

Bien que nous n'ayons pas considéré la description biologique des communautés identifiées, notre approche découvre des communautés statistiquement pertinentes. En outre, les dimensions pertinentes sélectionnées par l'algorithme fournissent des informations supplémentaires sur les canaux d'interaction entre les

	Redondance ρ			MCD			Modularité Multicouche			Nombre moyen de coms
Algorithm	bas	moy ± écart.t	haut	bas	moy ± écart.t	haut	bas	moy ± écart.t	haut	
MDLPA (DF)	0.24	0.27 ± 0.02	0.29	0.67	0.68 ± 0.01	0.69	0.76	0.76 ± 0.00	0.77	1143
MDLPA (RD)	0.55	0.58 ± 0.04	0.63	0.69	0.70 ± 0.01	0.70				
LPA-Bin-Agr	0.06	0.07 ± 0.02	0.09	0.63	0.68 ± 0.01	0.76	0.74	0.78 ± 0.02	0.81	1893
LPA-Freq-Agr	0.06	0.12 ± 0.03	0.15	0.64	0.64 ± 0.01	0.65	0.79	0.81 ± 0.00	0.81	2097
Clustering-Ens	0.84	0.87 ± 0.05	0.91	0.41	0.43 ± 0.03	0.45	0.67	0.67 ± 0.00	0.67	8211

Figure 4.12: Les résultats obtenus sur le réseau d'interactions de protéines de Drosophila Melanogaster.

protéines du même module. Nous croyons que les communautés identifiées peuvent caractériser des phénomènes biologiques ou suggérer de nouvelles hypothèses pouvant être vérifiées par la consultation de l'expertise biologique existante.

4.4 Conclusion

Les résultats obtenus suggèrent que l'identification de la bonne partition n'est possible qu'en combinant les différentes sources d'interaction d'une manière efficace. Ainsi, à moins d'avoir des dimensions structurellement semblables, l'application d'une approche de clustering d'ensembles peut ne pas donner les résultats escomptés. De même, l'adoption des stratégies d'agrégation peut engendrer une disparition de la partition latente. L'approche proposée est, en revanche, capable de combiner et d'exploiter les sources de connectivité hétérogènes d'une manière plus efficace. En outre, nous avons pu constater que le fait d'ignorer les dimensions non pertinentes contribuera non seulement à l'amélioration de la précision de partitionnement, mais aussi à la compréhension des facteurs de formation des groupes.

CHAPITRE V

CONCLUSION ET PERSPECTIVES

Dans ce mémoire, nous avons abordé la problématique de détection de communautés dans les réseaux multidimensionnels. Nous avons présenté les limitations des approches existantes, à savoir, la dépendance aux paramètres d'entrée, la sensibilité aux dimensions non pertinentes et l'incapacité d'identifier les dimensions pertinentes associées aux communautés détectées. Pour pallier ces problèmes, nous avons introduit une nouvelle approche de détection de communautés dans les réseaux multidimensionnels. Outre son efficacité, l'approche proposée offre la possibilité de discrimination entre les dimensions pertinentes et non pertinentes sans aucune intervention de l'utilisateur. L'évaluation expérimentale confirme la capacité de l'approche à identifier des structures de communautés relatant des caractéristiques fonctionnelles et organisationnelles non connues *à priori*.

Bien que nous ayons pu répondre aux objectifs de ce mémoire, quelques pistes d'amélioration demeurent envisageables. Par exemple, notre approche identifie des communautés disjointes où un nœud appartient à une et une seule communauté. L'extension de notre approche afin de supporter le partitionnement flou (où un nœud peut appartenir à plusieurs communautés) lui confère plus de flexibilité en permettant aux communautés détectées de partager les nœuds sur différents niveaux de l'espace multidimensionnel. Une autre possibilité d'amélioration consiste

à étendre la métrique de pertinence pour qu'elle puisse supporter les arêtes pondérées. Une telle addition est bénéfique pour les systèmes où les degrés de relations s'expriment à travers des poids.

En conclusion, nous pensons que l'approche proposée peut également être appliquée pour l'identification des communautés sur les réseaux dynamiques. En effet, nous pourrions découvrir les communautés évoluant au fil du temps, de même que leurs périodes formation, en prenant pour dimensions les différentes instances de la séquence ordonnée du réseau. Dans les perspectives de nos travaux, nous envisageons d'effectuer une recherche plus approfondie dans cette direction. En terminant, il convient de noter que la qualité des résultats obtenus par notre approche fait d'elle un outil viable dans différents contextes pratiques.

RÉFÉRENCES

Amelio, A. et Pizzuti, C. (2014). A cooperative evolutionary approach to learn communities in multilayer networks. In *Parallel Problem Solving from Nature - PPSN XIII*, volume 8672 222–232. Lecture Notes in Computer Science, Springer.

Barber, M. J. et Clark, J. W. (2009). Detecting network communities by propagating labels under constraints. *Physical Review E*, *80*(2), 026129.

Berlingerio, M., Coscia, M. et Giannotti, F. (2011). Finding and characterizing communities in multidimensional networks. Dans *Proceedings of the IEEE International Conference on Advances in Social Networks Analysis and Mining (ASONAM'2011)*, 490–494.

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

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

Blondel, V. D., Guillaume, J.-L., Lambiotte, R. et Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics : Theory and Experiment*, *2008*(10), P10008.

Boden, B., Günnemann, S., Hoffmann, H. et Seidl, T. (2012). Mining coherent subgraphs in multi-layer graphs with edge labels. Dans *Proceedings of the 18th*

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'2012), 1258–1266.

Borgelt, C. (2003). Efficient implementations of apriori and eclat. Dans *Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations (FIMI'2003)*.

Cai, D., Shao, Z., He, X., Yan, X. et Han, J. (2005). Mining hidden community in heterogeneous social networks. Dans *Proceedings of the 3rd ACM SIGKDD International Workshop on Link Discovery (LinkKDD'2005)*, 58–65.

Carchiolo, V., Longheu, A., Malgeri, M. et Mangioni, G. (2011). Communities unfolding in multislice networks. In *Complex Networks* 187–195. Communications in Computer and Information Science, Springer.

Chen, J. et Yuan, B. (2006). Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics*, *22*(18), 2283–2290.

Condon, A. et Karp, R. M. (2001). Algorithms for graph partitioning on the planted partition model. *Random Structures and Algorithms*, *18*(2), 116–140.

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

De Domenico, M., Nicosia, V., Arenas, A. et Latora, V. (2015b). Structural reducibility of multilayer networks. *Nature Communications*, *6*.

Dong, X., Frossard, P., Vandergheynst, P. et Nefedov, N. (2012). Clustering with multi-layer graphs : A spectral perspective. *IEEE Transactions on Signal Processing*, *60*(11), 5820–5831.

Dong, X., Frossard, P., Vandergheynst, P. et Nefedov, N. (2014). Clustering on multi-layer graphs via subspace analysis on grassmann manifolds. *IEEE Transactions on Signal Processing*, *62*(4), 905–918.

Dunlavy, D. M., Kolda, T. G. et Kegelmeyer, W. P. (2011). Multilinear algebra for analyzing data with multiple linkages. *Graph Algorithms in the Language of Linear Algebra, J. Kepner and J. Gilbert, eds., Fundamentals of Algorithms, SIAM, Philadelphia*, 85–114.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3), 75–174.

Fowlkes, E. B. et Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, *78*(383), 553–569.

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

Graovac, A., Gutman, I., Trinajstić, N. et Živković, T. (1972). Graph theory and molecular orbitals. *Theoretica Chimica Acta*, *26*(1), 67–78.

Hmimida, M. et Kanawati, R. (2015). Community detection in multiplex networks : A seed-centric approach. *Networks and Heterogeneous Media*, *10*(1), 71–85.

Hubert, L. et Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218.

Kanawati, R. (2013). Détection de communautés dans les grands graphes d'interactions (multiplexes) : état de l'art.

Kanawati, R. (2014). Seed-centric approaches for community detection in complex networks. In *Social Computing and Social Media* 197–208. Springer.

Kang, H., Getoor, L. et Singh, L. (2007). Visual analysis of dynamic group membership in temporal social networks. *ACM SIGKDD Explorations Newsletter*, *9*(2), 13–21.

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

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

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

Leung, I. X. Y., Hui, P., Lio, P. et Crowcroft, J. (2009). Towards real-time community detection in large networks. *Physical Review E*, *79*(6), 066107.

Li, X., Ng, M. K. et Ye, Y. (2014). Multicomm : Finding community structure in multi-dimensional networks. *IEEE Transactions on Knowledge and Data Engineering*, *26*(4), 929–941.

Liu, G. et Wong, L. (2008). Effective pruning techniques for mining quasi-cliques. In *Machine Learning and Knowledge Discovery in Databases* 33–49. Lecture Notes in Computer Science, Springer.

Liu, M. et Liu, B. (2009). New sharp upper bounds for the first zagreb index. *Communications in Mathematical and in Computer Chemistry*, *62*(3), 689–698.

Liu, X. et Murata, T. (2010). Advanced modularity-specialized label propagation algorithm for detecting communities in networks. *Physica A : Statistical Mechanics and its Applications*, *389*(7), 1493–1500.

Loe, C. W. et Jensen, H. J. (2015). Comparison of communities detection algorithms for multiplex. *Physica A : Statistical Mechanics and its Applications*, *431*, 29–45.

Magnani, M., Micenkova, B. et Rossi, L. (2013). Combinatorial analysis of multiple networks. *arXiv preprint arXiv :1303.4986*.

Manning, C. D., Raghavan, P., Schütze, H. *et al.* (2008). *Introduction to information retrieval*. Cambridge University Press.

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

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*(23), 8577–8582.

Nicosia, V. et Latora, V. (2014). Measuring and modelling correlations in multiplex networks. *arXiv preprint arXiv :1403.1546*.

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

Pons, P. et Latapy, M. (2005). Computing communities in large networks using random walks. In *Computer and Information Sciences-ISCIS 2005*, volume 3733 284–293. Lecture Notes in Computer Science, Springer.

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

Richardson, M. et Domingos, P. (2002). Mining knowledge-sharing sites for viral marketing. Dans *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge discovery and Data Mining (KDD'2002)*, 61–70.

Rives, A. W. et Galitski, T. (2003). Modular organization of cellular networks. *Proceedings of the National Academy of Sciences*, *100*(3), 1128–1133.

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

Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A. et Tyers, M. (2006). Biogrid : a general repository for interaction datasets. *Nucleic Acids Research*, *34*(suppl 1), D535–D539.

Strehl, A. et Ghosh, J. (2003). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, *3*, 583–617.

Tang, L., Wang, X. et Liu, H. (2009a). Uncoverning groups via heterogeneous interaction analysis. Dans *Proceedings of the 9th IEEE International Conference on Data Mining (ICDM'2009)*, 503–512.

Tang, L., Wang, X. et Liu, H. (2012). Community detection via heterogeneous interaction analysis. *Data Mining and Knowledge Discovery*, *25*(1), 1–33.

Tang, W., Lu, Z. et Dhillon, I. S. (2009b). Clustering with multiple graphs. Dans *Proceedings of the 9th IEEE International Conference on Data Mining (ICDM'2009)*, 1016–1021.