UNIVERSITÉ DU QUÉBEC À MONTRÉAL

UNE NOUVELLE APPROCHE DE DÉTECTION DES ANOMALIES DANS LES RÉSEAUX MULTIDIMENSIONNELS

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

AMANI CHOUCHANE

JANVIER 2018

REMERCIEMENTS

La réalisation de ce mémoire a été possible grâce au concours de plusieurs personnes à qui je veux témoigner toute ma reconnaissance.

Tout d'abord, je tiens à remercier mon directeur de recherche, Mohamed Bouguessa, pour son encadrement, son assistance et sa rigueur, qui m'ont beaucoup aidés dans la réalisation de ce travail.

Je remercie également la faculté des sciences de l'UQAM et la mission universitaire de la Tunisie au nord de l'Amérique pour les bourses d'exemption des frais de scolarité majorés que j'ai reçues durant mes études de maîtrise.

De plus, j'adresse toute ma gratitude à la fondation de l'UQAM pour m'avoir choisi parmi les récipiendaires de la bourse du bureau du registraire et de la bourse d'excellence FARE de la faculté des sciences.

J'adresse enfin une reconnaissance particulière à mes parents, ma famille et mes amis qui m'ont apporté leur support moral tout au long de mes études et à tous ceux qui ont contribué de près ou de loin à l'accomplissement de ce travail.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

L'analyse des réseaux d'informations complexes est un domaine de recherche récent, issu du croisement de deux univers connexes qui sont la théorie des graphes et le forage de données. De ce fait, les réseaux d'informations complexes peuvent être modélisés par des graphes multidimensionnels où les nœuds sont interconnectés par un ou plusieurs types de liens, de sorte que chaque type de lien représente une dimension d'analyse. À présent, les travaux de recherche dans les réseaux multidimensionnels couvrent quelques problématiques telles que la détection de communautés et l'élaboration de métriques de centralité. Toutefois, dans l'ensemble, les recherches ne se sont pas attardées sur la problématique de détection d'anomalies.

Principalement, la détection d'anomalies repose sur l'identification des observations rares et atypiques dans les jeux de données. Généralement, dans un graphe, une anomalie peut être associée à un nœud, à un lien ou à un sous-ensemble de nœuds du graphe. Du fait qu'elle ne détient pas une définition universelle et qu'elle est souvent liée à un contexte d'application bien spécifique, la détection d'anomalie reste peu développée dans les réseaux multidimensionnels.

Dans le cadre de ce mémoire, nous proposons une première tentative de détection d'anomalies dans les réseaux multidimensionnels, basée sur deux étapes. En premier, nous proposons une nouvelle fonction pour calculer un score d'anomalie pour chaque nœud du réseau. Les anomalies reçoivent des scores faibles tandis que les nœuds normaux reçoivent des scores élevés. Par la suite, nous modélisons la distribution des scores en utilisant le modèle de mélange beta. Les anomalies sont identifiées automatiquement par le composant beta qui détient les valeurs des scores d'anomalies les plus faibles. Nous testons notre approche moyennant un ensemble d'expérimentations sur des données synthétiques et réelles dans le but d'évaluer empiriquement ses performances.

Mots clés : Détection d'anomalie, Réseaux multidimensionnels, Analyse de liens.

INTRODUCTION

0.1 Mise en contexte

Un réseau d'informations est un ensemble d'entités interconnectées (Boccaletti *et al.*, 2006). Les réseaux sociaux, les réseaux biologiques et les réseaux de transports sont quelques exemples de systèmes complexes du monde réel. Pour les analyser, il est communément d'usage de combiner les outils de forages de données et de la théorie des graphes. De ce fait, un réseau peut être décrit par un graphe où les entités sont modélisées par des nœuds et les interactions entre ces entités sont représentées par des liens. Par exemple, dans un réseau de transport aérien, les villes se modélisent par des nœuds et les vols entre les villes se modélisent par des liens.

Plusieurs autres applications réelles peuvent être représentées sous forme de graphes. Ces applications génèrent un grand volume de données et leurs analyses ne cessent de susciter l'intérêt de la communauté scientifique et de l'industrie. À ce jour, plusieurs travaux ont été réalisés pour caractériser les réseaux, découvrir les phénomènes existants et dévoiler les patrons d'interactions cachés. Parmi ces travaux, nous citons la spécification de métriques d'analyse comme les métriques de centralité et l'élaboration d'algorithmes de détection de communautés.

Les réseaux d'informations ont été largement étudiés dans le contexte monodimensionnel où les entités sont interconnectées par un seul type de lien. Toutefois, les données réelles sont représentées par des réseaux d'informations complexes. Compacter la complexité d'interactions en un seul type de lien réduit la richesse

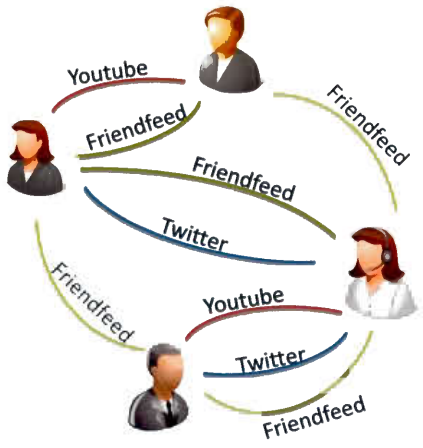


Figure 0.1: Exemple d'un réseau multidimensionnel réel *Friendfeed.*

de communication et conduit, éventuellement, à une perte considérable d'informations. Prenons l'exemple d'un réseau social formé par un ensemble d'individus reliés par des liens d'interactions sociales. La relation entre deux acteurs au sein du même réseau est au delà d'un simple lien d'interaction. Effectivement, les relations peuvent être de type amitié, travail, famille ou collaboration. Pour mieux préserver la richesse de communication, les réseaux multidimensionnels ont été proposés comme une alternative au réseau monodimensionnel (Boccaletti *et al.*, 2014). La figure 0.1 montre l'exemple du réseau social *Friendfeed.* Les acteurs de ce réseau peuvent se conneter via *Friendfeed*, *Twitter* et *Youtube.* Cette modélisation multidimensionnelle permet de mieux décrire les interactions entre les entités du réseau.

Plusieurs recherches se sont particulièrement concentrées sur l'étude des réseaux multidimensionnels. Des travaux ont été proposés pour répondre à certaines problématiques comme la détection de communautés et le développement de métriques d'analyse de la structure topologique du réseau (Battiston *et al.*, 2014), (Boccaletti *et al.*, 2014). D'autres travaux se sont focalisés sur l'analyse des processus dynamiques comme la prélocation (Hackett *et al.*, 2016). En général, des

problématiques diverses ont été couvertes dans les réseaux multidimensionnels. Cependant, dans notre investigation de la littérature, nous n'avons pas trouvé d'approches de détection d'anomalies spécifiques aux réseaux multidimensionnels.

La détection d'anomalie est une branche de la fouille de données qui s'occupent de l'identification des observations rares et atypiques dans les données (Tan *et al.*, 2006), (Chandola *et al.*, 2009). En d'autres mots, une telle démarche consiste à identifier les entités qui sont considérablement différentes de l'ensemble des entités du réseau. Les entités détectées sont appelées des anomalies, et peuvent aussi être appelées, selon le contexte de l'application étudiée, des *outliers*, des exceptions, des entités aberrantes ou des surprises. La détection d'anomalies tire son importance de la diversité des domaines qu'elle couvre et des conséquences avantageuses qu'elle apporte. Des travaux ont été proposés pour identifier, dans un réseau social, les utilisateurs malveillants (*spammers*) qui attestent un comportement étrange et nuisent à l'échange entre les utilisateurs légitimes (Savage *et al.*, 2014). D'autres travaux se sont intéressés à l'étude de la fraude d'opinions et ce pour détecter les utilisateurs qui postent de fausses opinions positives dans des sites comme Amazon ou Yelp pour promouvoir la vente d'un produit ou d'un service (Akoglu *et al.*, 2013). Dans les réseaux informatiques, des travaux ont été faits pour identifier les cyberattaques et les intrusions (Ding *et al.*, 2012).

La détection d'anomalies peut aussi être implémentée comme une phase de prétraitement de données pour localiser, dans un réseau, les nœuds qui n'appartiennent pas à une communauté. Dans un réseau, une communauté est un ensemble d'entités ayant une interconnexion forte. À ce propos, une majorité des algorithmes de détection de communautés sont parvenus à un consensus en identifiant comme des anomalies les nœuds qui n'appartiennent pas aux communautés (Xu *et al.*, 2007), (Huang *et al.*, 2010), (Huang *et al.*, 2013). Ces nœuds aberrants doivent être identifiés et éliminés puisqu'ils disposent d'un patron d'interaction ir-

régulier par rapport au reste des nœuds du réseau et peuvent, potentiellement, nuire au processus de détection de communautés.

0.2 Motivations

La détection d'anomalies dans les réseaux d'informations multidimensionnels est une problématique d'actualité. La portée de la détection d'anomalies ainsi que les travaux limités à ce propos dans les réseaux multidimensionnels ont motivé notre intérêt à élaborer une méthode qui détecte les nœuds atypiques dans un réseau multidimensionnel. Bien que la détection d'anomalies ait été bien étudiée dans les réseaux monodimensionnels (Akoglu *et al.*, 2010), (Aggarwal, 2017), aucune approche n'a été proposée à ce sujet dans les réseaux multidimensionnels.

Le manque d'approches qui se rattachent à la détection d'anomalies dans le contexte des réseaux multidimensionnels est possiblement lié, d'une part, à la structure topologique non triviale des réseaux multidimensionnels, et, d'une autre part, à l'absence d'une définition formelle du concept d'anomalies dans les réseaux multidimensionnels. Effectivement, une définition universelle d'une anomalie est loin d'être établie. Généralement, selon un contexte d'étude bien précis, l'anomalie se définit par une déviation à une norme précisée au préalable. De ce fait, une anomalie correspond à un état anormal ou à un comportement inattendu.

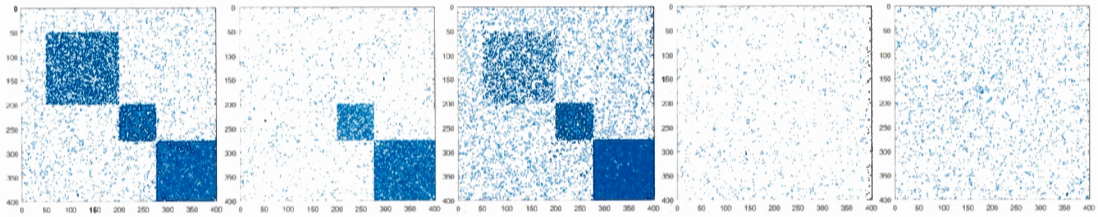
Il est à noter que dans le contexte des réseaux monodimensionnels, un certain nombre d'approches sont parvenus à un consensus et considèrent des anomalies les nœuds qui n'appartiennent pas aux communautés, à savoir les régions denses du réseau (Xu *et al.*, 2007), (Huang *et al.*, 2010), (Huang *et al.*, 2013). Autrement dit, les anomalies correspondent aux nœuds ayant un patron de connexion atypique, c.-à-d., des nœuds isolés ou des nœuds qui n'appartiennent à aucune région dense identifiable dans le réseau. Une telle observation est plausible puisqu'en général, la densité des liens entre les nœuds d'un réseau varie d'une zone à

une autre, ce qui implique l'existence de groupes de nœuds dans lesquels la densité de communication interne est élevée et la densité de communication externe est faible (Bouguessa *et al.*, 2014), (Girvan et Newman, 2002).

Dans le cadre de notre étude, nous nous intéressons à identifier les nœuds qui n'appartiennent pas à des régions denses dans les réseaux multidimensionnels. Une région dense peut être considérée comme une communauté. Dans le contexte de réseaux multidimensionnels, une communauté, appelée aussi un *cluster* est une région composée d'un groupe de nœuds fortement connectés entre eux et faiblement connectés avec le reste des nœuds du graphe dans un sous-espace ou dans l'ensemble des dimensions. Par conséquent, les nœuds normaux sont des nœuds qui forment des régions denses dans un sous-ensemble ou dans toutes les dimensions du réseau. Par contre, les nœuds anomalies sont des nœuds qui se connectent d'une manière aléatoire aux nœuds du réseaux à travers toutes les dimensions, sans appartenir à aucune structure de communauté identifiable. De ce fait, ils sont des nœuds qui n'appartiennent pas à des groupes de nœuds denses.

Pour des fins d'illustration, nous avons généré un réseau synthétique multidimensionnel composé de 400 nœuds interconnectés suivant 5 dimensions. Il est à noter que le réseau généré incorpore 3 communautés et 50 nœuds atypiques, qui n'appartiennent à aucune région dense sur les 5 dimensions du réseau. La Figure 0.2 montre les matrices d'adjacence du réseau synthétique généré. Chaque matrice d'adjacence correspond à une dimension spécifique du réseau. Dans chaque matrice, les régions foncées démontrent la présence de groupes de nœuds ayant une interconnexion forte tandis que les régions blanches, faiblement granulées, témoignent l'absence de communautés.

Tel qu'illustré par la Figure 0.2, les dimensions d_1, d_2 et d_3 contiennent des blocs de nœuds denses. Ces dimensions sont considérées pertinentes puisqu'elles montrent



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

Figure 0.2: Les matrices d'adjacence associées à un réseau synthétique multidimensionnel de 400 nœuds.

l'existence de communautés dans le réseau. Par contre, les dimensions d_4 et d_5 sont des dimensions non pertinentes puisque les matrices d'adjacence associées à ces dimensions ne montrent aucune structure dense du réseau. Une dimension est dite non pertinente lorsqu'elle contient des connexions aléatoires et ne montre pas les structures de communautés qui existent dans un réseau (voir les figures 0.2d, 0.2e). Les anomalies sont les nœuds qui n'appartiennent pas aux trois communautés sur toutes les dimensions du réseau. Plus précisément, selon les matrices d'adjacences de la Figure 0.2, les nœuds atypiques sont localisés en haut de chacune des 5 matrices. Dans ce travail, nous nous focalisons sur l'identification de ce type de nœuds.

Bien qu'au moment de la rédaction de ce mémoire, aucune méthode n'a été proposée pour identifier ce type d'anomalies dans les réseaux multidimensionnels, une méthode simple d'aborder cette problématique consiste à réduire le problème de détection d'anomalies à la configuration monodimensionnelle par la transformation d'un réseau multidimensionnel en un réseau monodimensionnel. Après agrégation des dimensions, une méthode de détection d'anomalies spécifique aux réseaux monodimensionnels peut être appliquée. Cependant, une telle approche est naïve. En effet, l'activité d'un nœud varie d'une dimension à une autre. L'agrégation

de toutes les dimensions du réseau entraîne une perte informationnelle substantielle et supprime la richesse de la configuration multidimensionnelle d'un réseau complexe. De plus, l'agrégation d'un réseau multidimensionnel produit un réseau monodimensionnel étroitement connecté. De telle sorte, l'identification des nœuds ayant des connexions aléatoires sera moins évidente, voire impossible.

Une autre alternative consiste à explorer individuellement les dimensions du réseau multidimensionnel. Afin d'identifier les nœuds atypiques, une technique de détection d'anomalies pour les réseaux monodimensionnels peut être appliquée sur chaque dimension. Ensuite, un consensus entre les anomalies identifiées sur chaque dimension doit être établi. Cependant, cette approche peut rencontrer des difficultés lorsque le réseau étudié contient des dimensions non pertinentes qui se caractérisent par la présence de connexions aléatoires entre les nœuds du réseau. Dans une telle situation, les dimensions non pertinentes rendent difficile le processus de détection d'anomalies puisqu'elles ne contiennent pas de structures denses qui permettent la distinction entre les nœuds normaux et anormaux.

0.3 Contributions

Ce mémoire présente une première tentative de détection des anomalies dans les réseaux multidimensionnels [1]. L'approche proposée permet l'identification automatique des nœuds atypiques (c.-à-d. les anomalies) qui ont des connexions aléatoires et qui n'appartiennent pas aux régions denses à travers toutes les dimensions d'un réseau multidimensionnel. Dans ce cadre, notre travail est marqué par les éléments suivants :

- Partant de l'hypothèse qu'une anomalie est un nœud qui se connecte aux

1. Ce travail a été accepté pour publication dans les actes de *2017 IEEE International Conference on Data Science and Advanced Analytics* (Chouchane et Bouguessa, 2017).

autres nœuds du réseau d'une façon éparse à travers toutes les dimensions, nous élaborons une fonction qui mesure la force de connexion entre les nœuds du réseau. Nous utilisons cette fonction dans le calcul d'un score qui évalue le degré d'anomalie d'un nœud dans un réseau multidimensionnel. Les nœuds densément connectés dans un sous-ensemble, ou dans tout l'ensemble de dimensions reçoivent un score élevé tandis que les anomalies reçoivent un score faible. Ici, il convient de noter que dans la conception de cette fonction, nous analysons un réseau multidimensionnel en tant que tel, sans le transformer en un réseau monodimensionnel et sans examiner indépendamment ses dimensions.

- En se basant sur le degré d'anomalie estimé, nous concevons une méthode probabiliste, basée sur le modèle de distribution beta, pour séparer automatiquement les anomalies des nœuds normaux. La méthode proposée nous permet d'identifier systématiquement les nœuds atypiques. De ce fait, aucune intervention humaine n'est nécessaire pour spécifier un seuil ou un nombre d'anomalies à identifier au préalable.

- Nous testons et validons l'approche proposée au moyen d'expérimentations sur des données synthétiques et réelles. Dans nos tests, nous examinons différents scénarios pour évaluer notre approche. Les résultats obtenus suggèrent que la méthode proposée parvient à bien identifier les anomalies même en présence de plusieurs dimensions non pertinentes dans le réseau.

0.4 Structure et organisation du mémoire

L'organisation de ce document est énoncée dans ce qui suit : le chapitre 2 couvre une revue de la littérature des travaux de détection d'anomalies dans les données vectorielles et dans les graphes monodimensionnels. Le chapitre 3 décrit l'approche proposée. Plus précisément, dans ce chapitre, nous présentons la fonction élaborée

pour calculer les scores d'anomalies pour chaque nœud du réseau et nous expliquons l'intuition derrière sa conception. Par la suite, nous détaillons le modèle probabiliste développé pour identifier automatiquement les anomalies. Le chapitre 4 présente une évaluation empirique de notre approche. Nous détaillons les différents scénarios des expérimentations effectuées sur des données synthétiques et réelles. Finalement, le chapitre 5 conclut ce mémoire.

CHAPITRE I

REVUE DE LA LITTÉRATURE

Ce chapitre présente un aperçu de quelques travaux existants qui traitent la problématique de détection d'anomalies. Nous commençons le chapitre par la présentation de quelques notions de base du concept d'anomalie et nous soulignons l'importance de son application dans plusieurs domaines. Nous présentons, par la suite, un survol de quelques techniques de détection d'anomalies appliquées sur les données vectorielles et sur les données représentées par des graphes. Pour les données vectorielles, nous catégorisons les techniques de détection d'anomalies selon leur mode de fonctionnement en trois types : (1) méthodes à base de statistiques, (2) méthodes à base de distance et (3) méthodes à base de densité. Pour les données représentées par des graphes, nous présentons quelques avancées algorithmiques qui traitent la notion d'anomalies dans les graphes monodimensionnels. Ici, il convient de souligner qu'au meilleur de notre connaissance, nous n'avons pas trouvé de techniques qui étudient explicitement la détection d'anomalies dans les graphes multidimensionnels. Par ce fait, le but de ce chapitre est de donner au lecteur une idée générale sur les recherches effectuées dans la détection d'anomalies.

1.1 Concepts et notions

La détection d'anomalies est une branche du forage de données qui s'occupe de l'identification des enregistrements atypiques ou des occurrences rares dans les données (Tan *et al.*, 2006). En d'autres termes, la détection d'anomalies consiste à trouver les objets qui sont différents ou inconsistants par rapport à la majorité des objets d'un jeu de données. Dans la littérature, les objets atypiques détectés sont dits anomalies, et sont aussi appelés, selon le contexte d'application, exceptions, surprises ou *outliers* (Aggarwal, 2017).

Initialement, la détection d'anomalies s'est développée dans les données à vecteur de caractéristiques. Formellement, la première définition d'anomalie revient à Hawkins en 1980 : "Une anomalie est une observation qui diffère tellement d'autres observations au point d'éveiller des soupçons qu'elle soit générée par un mécanisme différent" [Notre traduction] (Hawkins, 1980). Pour illustrer, nous prenons l'exemple de la figure 1.1. Dans cette figure, nous constatons deux groupes de points normaux $N1$ et $N2$. Les observations $o1$ et $o2$, sont, remarquablement, éloignées de la majorité des observations qui se trouvent dans les régions $N1$ et $N2$, et sont, par conséquent, considérées des anomalies.

Étant donnée la force d'expressivité des graphes et leur capacité à représenter des relations complexes entre les entités du monde réel, la notion d'anomalie s'est généralisée au cas des données représentées par des graphes. En effet, "dans un graphe, une anomalie peut être définie par les objets qui sont rares et qui diffèrent significativement de la majorité des objets de référence." [Notre traduction] (Akoglu *et al.*, 2015). Ici, un objet de référence est un objet qui se caractérise par un comportement ou par un état normal attendu.

La définition d'anomalie prend plus de sens lorsqu'elle est reliée à un contexte ou à une application bien spécifique. En effet, dans la littérature, nous trouvons des

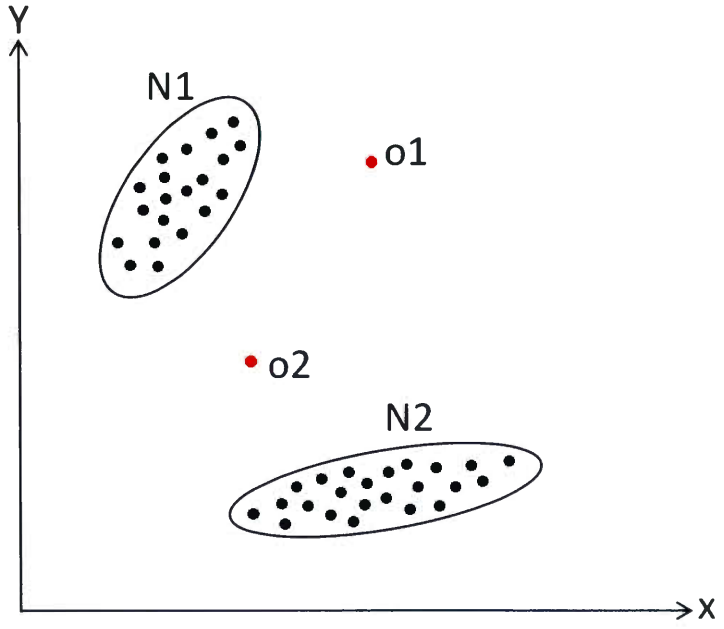


Figure 1.1: Exemple d'anomalies dans un espace bidimensionnel.

travaux qui se rattachent à plusieurs domaines d'application (Tan *et al.*, 2006). Particulièrement, la détection d'anomalies a été beaucoup appliquée pour dans la détection de fraude. Dans ce contexte, les compagnies de cartes bancaires, par exemple, identifient les comportements suspects et frauduleux en se basent sur le patron d'achat habituel de leurs clients.

Par ailleurs, la détection d'anomalies s'emploie aussi dans la détection des intrusions dans les réseaux d'ordinateurs. En fait, de nos jours, les cyberattaques sont de plus en plus en abondance. Quelques attaques sont désignées pour faire dysfonctionner ou submerger les ordinateurs, d'autres sont faites pour le vol d'informations. Plusieurs de ces attaques peuvent être identifiées en contrôlant et en surveillant les comportements atypiques des utilisateurs des systèmes informatiques.

Dans le domaine de santé publique, les hôpitaux et les cliniques médicales émettent des statistiques aux organisations nationales pour analyser l'état de santé de la population et évaluer l'efficacité des traitements offerts. Dans ce cadre, la détection d'anomalies prend sens lorsqu'il y a des personnes atteintes d'une maladie particulière même après vaccination. Une situation pareille peut remettre en question le traitement proposé. À cela s'ajoute, dans le web social, parmi les applications de détection d'anomalies utilisées, nous trouvons des applications qui permettent de repérer les utilisateurs malintentionnés comme les spammeurs et les anarqueurs qui publient de fausses opinions positives dans les sites d'achat en ligne tels que Amazon.

Dans la suite de ce chapitre, nous présentons quelques techniques déployées pour répondre aux problématiques précitées. En premier lieu, nous commençons par présenter les méthodes de détection d'anomalies dans les données vectorielles. Ensuite, nous présentons celles qui sont utilisées dans les données représentées par des graphes. Encore une fois, bien que la problématique de détection d'anomalies a été traitée dans les graphes monodimensionnels, elle reste encore non explorée dans les graphes multidimensionnels.

1.2 Détection d'anomalies dans les données vectorielles

Les techniques de détection d'anomalies dans les données vectorielles peuvent être catégorisées en trois types : (1) méthodes à base de statistiques, (2) méthodes à base de distance et (3) méthodes à base de densité. Dans ce qui suit, nous présentons une brève description de chaque catégorie et nous les illustrons par quelques exemples de méthodes représentatives.

1.2.1 Méthodes à base de statistiques

Les approches à base de statistiques consistent à élaborer des modèles statistiques probabilistes flexibles qui représentent la distribution des jeux de données testés comme les modèles gaussiens (Yamanishi *et al.*, 2004) et les modèles de régression (Aggarwal, 2005), (Li et Han, 2007). Le degré d'anomalie d'un objet particulier est évalué par rapport à sa conformité au modèle établi. Particulièrement, dans (Yamanishi *et al.*, 2004), un modèle de mélange gaussien est proposé pour représenter la distribution des données testées. Chaque objet reçoit un score d'anomalie qui caractérise sa déviation au modèle. Un score élevé dénote une forte probabilité que l'objet en question soit une anomalie.

Par ailleurs, dans la littérature, nous trouvons d'autres méthodes statistiques de détection d'anomalies comme les histogrammes (Fawcett et Provost, 1999) et les fonctions à noyaux (Bishop, 1994). Spécifiquement, les techniques à base d'histogrammes consistent à élaborer un profil fréquentiel des données, et sont appliquées dans plusieurs domaines comme la fraude (Fawcett et Provost, 1999), l'intrusion (Yamanishi *et al.*, 2004) et les cyberattaques (Krügel *et al.*, 2002). Les fonctions à noyau, à leur tour, offrent une approximation de la densité de distribution des données (Yeung et Chow, 2002). Par ce fait, une observation qui appartient à la zone la moins dense de la distribution des données est identifiée comme une anomalie.

1.2.2 Méthodes à base de distance

Les méthodes à base de distance consistent à calculer la disparité entre les objets d'un ensemble de données. Pour mesurer l'hétérogénéité des objets, plusieurs métriques peuvent être employées comme la distance euclidienne et la distance de Manhattan. Un objet est considéré une anomalie s'il est remarquablement distant

de la majorité des objets.

Spécifiquement, les techniques à base de distance comme k plus proches voisins (Ramaswamy *et al.*, 2000) et *KNN-pondéré* (Angiulli et Pizzuti, 2002) assignent un score d'anomalie à chaque objet en se basant sur ses k plus proches voisins. De cette manière, étant distants, les anomalies (*outliers*) reçoivent des scores élevés et les objets normaux (*inliers*) reçoivent des scores faibles. Les anomalies sont discernées en triant les scores dans un ordre ascendant et en sélectionnant les observations ayant les scores les moins élevés.

1.2.3 Méthodes à base de densité

Les méthodes à base de densité mesurent le degré d'anomalie d'un objet en considérant la densité locale de son voisinage. Spécifiquement, nous citons l'exemple de calcul du score d'anomalie LOF (*Local Outlier Factor*) (Breunig *et al.*, 2000). Le fondement de LOF a été inspiré de la méthode de partitionnement à base de densité DBSCAN qui identifie à la fois les communautés et les *outliers* (Ester *et al.*, 1996). Dans LOF, la densité locale de chaque objet se calcule en respect de ses k plus proches voisins. L'ensemble des distances d'un objet particulier à ses k plus proches voisins sont utilisées dans le calcul de sa densité locale. Les densités locales de tous les objets sont, ensuite, évaluées pour déterminer les régions de densité similaires et les objets *outliers* qui détiennent des densités locales remarquablement faibles par rapport à leurs voisinages. Plusieurs autres variantes de la méthode LOF ont été proposées comme COF (*Connectivity-based Outlier Factor*) (Tang *et al.*, 2002) et LoOP (Kriegel *et al.*, 2009).

Encore une fois, nous rappelons que les techniques présentées dans cette section ne représentent pas une liste exhaustive des méthodes de détection d'anomalies dans les données vectorielles. Nous trouvons plusieurs autres techniques dans la littérature. Particulièrement, dans (Chandola *et al.*, 2009), une revue détaillée

des méthodes de détection des *outliers* est élaborée. Par ailleurs, d'autres revues comme (Zimek *et al.*, 2012), se sont concentrées sur l'étude d'anomalies dans des réseaux à dimensionnalité élevée. De plus, nous trouvons dans (Chandola *et al.*, 2012), une revue des travaux de détection d'anomalies qui s'intéressent à la détection des événements et des changements dans les données. Ici, il convient de noter qu'aucune des revues précitées ne discute la problématique de détection d'anomalies dans les données représentées par des graphes. À cet effet, dans la suite de ce chapitre, nous attirons l'attention du lecteur à quelques techniques de détection d'anomalies spécifiques aux graphes.

1.3 Détection d'anomalies dans les graphes

Récemment, un grand intérêt a été porté à l'élaboration de techniques qui traitent les anomalies dans les graphes, et ce vu leur expressivité et leur capacité à représenter la complexité d'interaction du monde réel (Akoglu *et al.*, 2015). Dans cette section, nous présentons quelques travaux de détection d'anomalies spécifiques aux données représentées par des graphes. Précisément, les techniques présentées dans la suite de cette section, discernent les anomalies à partir de la structure topologique d'un graphe monodimensionnel.

La topologie d'un réseau est porteuse d'informations implicites décisives pour repérer les anomalies. À titre d'exemple, dans les réseaux sociaux, plusieurs types d'anomalies peuvent avoir lieu (Savage *et al.*, 2014). Particulièrement, les spammeurs sont incapables de cacher un certain type de métadonnées telles que leurs patrons d'interactions, c.-à-d., les liens qu'ils établissent et qui sont, entre autres, révélateurs de leur comportement irrégulier. Pour identifier ces acteurs malveillants du réseau, l'utilisation unique des données est insuffisante et il est nécessaire de considérer la structure topologique du réseau. Dans (Fathaliani et Bouguessa, 2015), un score d'anomalie qui évalue la proportion des liens émis et des liens

reçus a été proposé pour identifier ce type d'utilisateurs malintentionnés.

À cela s'ajoute, dans les graphes, la détection de communautés qui a été appliquée pour la détection des intrusions dans les réseaux (Ding *et al.*, 2012). En effet, la détection de communautés (appelée aussi partitionnement) compte parmi les problématiques les plus étudiées dans les graphes. Explicitement, la détection de communautés consiste à découvrir la structure sous-jacente d'un réseau, à savoir, l'identification des nœuds fortement connectés entre eux (c.-à-d., les *clusters*). Dans ce contexte, les intrusions sont les nœuds présents dans une communauté, mais qui n'y appartiennent pas réellement. D'autres techniques de détection de communautés se sont concentrées sur l'identification des nœuds superflus (*outliers*) qui sont marginalement connectés aux communautés. Ces nœuds forment un bruit dans le réseau et leur élimination peut, potentiellement, améliorer les résultats d'analyse. Un nombre d'algorithmes a été proposé à ce sujet, à savoir SCAN (Xu *et al.*, 2007), gSkeletonClu (Huang *et al.*, 2013) et SHRINK (Huang *et al.*, 2010).

L'algorithme SCAN détecte les communautés et les nœuds qui y sont marginalement connectés en se basant sur une métrique de similarité structurale qui se calcule entre les nœuds du graphe. Cette métrique tient compte de la connectivité entre les nœuds, et ce en examinant leurs voisinages. Plus les nœuds partagent des voisins, plus la valeur de la métrique de similarité est élevée. Pour déterminer les communautés et les *outliers*, un seuil minimal de similarité doit être fixé. De cette manière, les nœuds qui ont plusieurs voisins en commun se groupent dans une même communauté. Par ailleurs, les nœuds superflus sont les nœuds qui ne se sont pas affectés à des communautés et qui ont des valeurs de similarité faibles.

L'algorithme SCAN nécessite un paramétrage pour pouvoir identifier les communautés et les nœuds superflus. Pour éviter cette intervention supervisée, l'algo-

rithme gSkeletonClu a été proposé. Il convient de préciser que, tout comme SCAN, l'algorithme gSkeletonClu est un algorithme de partitionnement à base de densité qui détermine aussi bien les communautés et les *outliers* selon une métrique de similarité. Toutefois, l'algorithme gSkeletonClu permet une sélection automatique de la valeur du seuil minimal, et ce en maximisant une mesure de validité comme la modularité. La modularité est une métrique fréquemment utilisée pour mesurer la qualité de partitionnement d'un graphe. Quelques algorithmes comme gSkeletonClu utilisent cette métrique comme une fonction objective pour optimiser le partitionnement d'un graphe. Ici, il est utile de noter que, dans gSkeletonClu, l'extraction des communautés et des *outliers* peut se faire aussi avec un seuil spécifié à l'avance par l'utilisateur.

Dans la littérature, nous trouvons, également, SHRINK qui est un algorithme de partitionnement hiérarchique. Cet algorithme identifie les communautés et les *outliers* sans la nécessité de paramétrage en reposant sur le principe de réduction. Plus précisément, l'algorithme commence par identifier les pairs de nœuds denses, c.-à-d. les nœuds dont la similarité structurale est maximale par rapport à leurs voisinages. Ensuite, une fusion se fait entre les pairs de nœuds denses identifiés pour construire itérativement des microcommunautés. De ce fait, une microcommunauté peut être un nœud isolé ou un sous-graphe d'un ou plusieurs pairs de nœuds denses connectés. Au fur et à mesure des itérations, un arbre hiérarchique d'emboîtement des microcommunautés se construit.

1.4 Conclusion

Dans ce chapitre, nous avons présenté quelques avancées algorithmiques traitant la problématique de détection d'anomalie dans les données vectorielles et dans les données représentées par des graphes. Au mieux de notre investigation de la littérature, nous n'avons pas trouvé de méthodes qui traitent les anomalies dans les

graphes multidimensionnels. Cela nous a motivé à initier une première tentative de détection d'anomalies dans les graphes multidimensionnels. Dans notre travail, nous nous intéressons à identifier les nœuds isolés qui n'appartiennent à aucune région dense à travers toutes les dimensions d'un réseau multidimensionnel. Dans le chapitre suivant, nous détaillons le fondement de notre approche.

CHAPITRE II

APPROCHE PROPOSÉE

Ce chapitre présente une nouvelle approche de détection d'anomalies dans les réseaux multidimensionnels. Les nœuds identifiés par notre approche sont les nœuds qui n'appartiennent pas aux régions denses à travers toutes les dimensions du réseau. Pour identifier ce type de nœuds, nous procédons en deux étapes. Dans un premier temps, nous examinons la topologie du réseau pour estimer un score d'anomalie pour chaque nœud, et ce en nous basant sur une fonction qui relate la force de connexion entre les nœuds dans une configuration multidimensionnelle. Les nœuds normaux reçoivent des scores élevés tandis que les nœuds atypiques reçoivent des scores faibles. Dans un second temps, nous développons une méthode statistique qui se base sur le modèle de distribution beta pour modéliser les scores estimés dans la première phase et déceler automatiquement les anomalies. Dans le présent chapitre, nous commençons par introduire la notation utilisée. Ensuite, nous présentons en détail les deux phases qui concourent à l'identification des anomalies.

2.1 Notations

Avant de décrire le fondement de notre approche, il convient de présenter d'abord la notation utilisée pour analyser un réseau multidimensionnel (Boccaletti *et al.*, 2014). Dans la conception de notre approche, nous utilisons les multigraphes pour

modéliser un réseau multidimensionnel. Un multigraphe G non orienté et non pondéré se définit par le triplet (V, E, D) où V est un ensemble de n nœuds, E est un ensemble de m arêtes et D est un ensemble de c dimensions. Une arête $e \in E$ est un triplet (u, v, d) où $u, v \in V$ sont des nœuds et $d \in D$ est une dimension. Le triplet (u, v, d) spécifie que les nœuds u et v sont connectés par une arête qui appartient à la dimension d. Par conséquent, dans un multigraphe G, une paire de nœuds se connecte par, au plus, c arêtes. Comme mentionné précédemment, notre approche opère en deux phases : (1) estimation d'un score qui relate le degré d'anomalie pour chaque nœud et (2) modélisation de la distribution statistique des scores pour discriminer automatiquement les anomalies des nœuds normaux. Les détails qui se rattachent à chaque phase sont décrits dans ce qui suit.

2.2 Phase 1 : Estimation des scores d'anomalie

Dans cette section, nous élaborons une méthode qui estime un score d'anomalie pour chaque nœud, et ce afin de détecter les nœuds atypiques dans un réseau multidimensionnel. D'abord, nous développons une fonction qui relate la force de connexion entre les paires de nœuds. La fonction proposée se base sur le fait que les nœuds normaux sont étroitement connectés entre eux à travers un sous-ensemble ou toutes les dimensions, alors que les anomalies sont faiblement connectées aux nœuds du réseau à travers toutes les dimensions. Ensuite, pour estimer un score d'anomalie qui permet la discrimination entre les nœuds atypiques et les nœuds normaux, nous calculons une mesure qui évalue la force de connexion de chaque nœud avec ses voisins immédiats.

Généralement, les nœuds normaux ont tendance à appartenir à des structures denses qui apparaissent dans des sous-espaces de dimensions. Par contre, les nœuds atypiques ont tendance à se connecter aléatoirement aux nœuds du réseau à travers toutes les dimensions. Par conséquent, à l'encontre des anomalies qui se caracté-

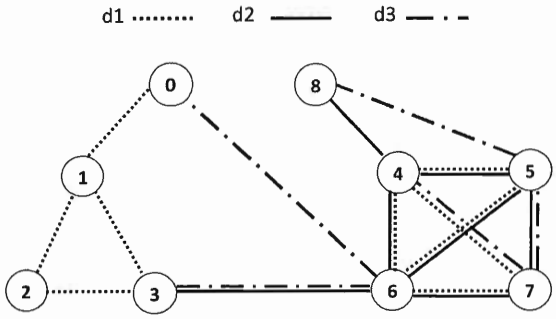


Figure 2.1: Un exemple d'un réseau de 9 noeuds et 3 dimensions.

risent par une structure topologique irrégulière, les nœuds normaux qui font partie d'une même région dense, partagent un même patron d'interactions, c.-à-d., ils ont relativement un nombre élevé de voisins en commun. En suivant ce raisonnement, nous définissons la force de connexion f entre deux nœuds u et v comme suit :

$$f(v,u) = \frac{\mid N(u, S_{u,v}) \cap N(v, S_{u,v}) \mid}{\mid N(u, D) \cap N(v, D) \mid} \tag{2.1}$$

D est l'ensemble de toutes les dimensions du réseau, alors que $S_{u,v} \subseteq D$ est l'ensemble des dimensions qui connectent u et v uniquement.

Dans la fonction définie par la formule 2.1, $N(u, S_{u,v})$ dénote l'ensemble des voisins de u dans le sous-espace de dimensions S_{uv}, c.-à-d. les nœuds auxquels le nœud u se connecte directement dans $S_{u,v}$. $N(u, D)$ dénote l'ensemble des voisins de u dans toutes les dimensions D du réseau. Afin d'illustrer ces notations et ces concepts, nous considérons le réseau multidimensionnel de la figure 2.1. Ce réseau contient 9 nœuds interconnectés suivant 3 dimensions : $D = \{d_1, d_2, d_3\}$. Examinons l'interaction entre les nœuds 4 et 5 qui se connectent à travers les dimensions d_1 et d_2. Dans ce cas, le sous-espace de dimensions qui connectent les nœuds 4 et 5 est défini par $S_{4,5} = \{d_1, d_2\}$, les voisins du nœud 4 dans ce sous-espace sont $N(4, S_{4,5}) = \{5, 6, 7, 8\}$ et les voisins de 5 dans le même sous-espace

de dimensions sont $N(5, S_{4,5}) = \{4, 6, 7\}$. Les voisins du nœud 4 sur toutes les dimensions D sont définis par $N(4, D) = \{5, 6, 7, 8\}$ et les voisins de 5 dans D sont $N(5, D) = \{4, 6, 7, 8\}$. Dans ce qui suit, nous expliquons le raisonnement derrière la définition de la fonction $f(u, v)$ telle que décrite par l'équation (2.1).

Le numérateur de l'équation 2.1, à savoir le terme : $\mid N(u, S_{u,v}) \cap N(v, S_{u,v}) \mid$, correspond au nombre de voisins en commun entre les nœuds u et v dans le sous-espace de dimensions $S_{u,v}$, alors que le dénominateur, à savoir le terme : $\mid N(u, D) \cap N(v, D) \mid$, retourne le nombre de voisins en commun entre les deux nœuds, u et v, dans tout l'espace de dimensions D. La force de connexion entre deux nœuds u et v se calcule par le rapport des termes $\mid N(u, S_{u,v}) \cap N(v, S_{u,v}) \mid$ et $\mid N(u, D) \cap N(v, D) \mid$. Spécifiquement, ce rapport, tel que défini par la fonction $f(u, v)$, calcule le ratio du nombre de voisins en commun des nœuds u et v dans les dimensions qui les connectent exclusivement $(S_{u,v})$ sur le nombre des voisins en commun dans tout l'espace de dimensions D.

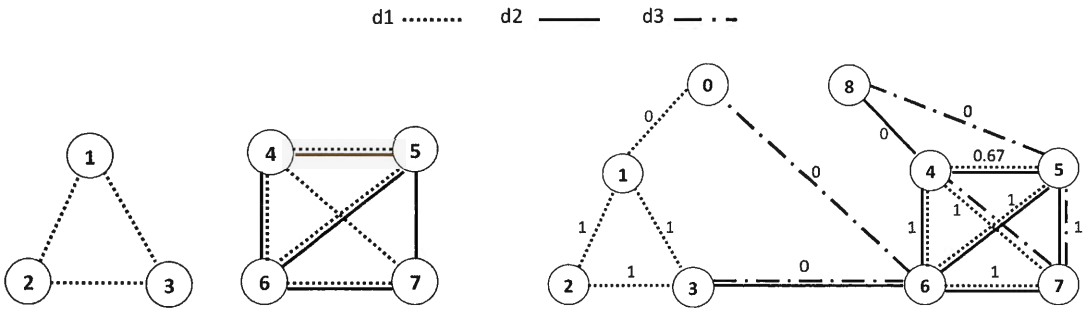
De cette manière, $f(u, v)$ prend en considération deux facteurs pour évaluer la connexion entre deux nœuds u et v. D'une part, la fonction f évalue la force de connexion à travers les dimensions qu'utilisent deux nœuds pour se connecter (c.-à-d., $S_{u,v}$) par rapport à l'ensemble de dimension (c.-à-d., D). D'une autre part, la fonction f compare la structure topologique des nœuds en analysant leur voisinage direct. Ainsi, la fonction $f(u, v)$ révèle la fermeté de connexion entre les pairs de nœuds et leur tendance à former des groupes de nœuds denses en considérant la pertinence de connexion en termes de nombre de voisins partagés à travers les dimensions du réseau.

Les valeurs de la fonction $f(u, v)$ sont comprises dans l'intervalle [0,1]. Une valeur élevée de $f(u, v)$ indique que les nœuds u et v partagent un grand nombre de voisins dans le sous-espace de dimensions $S_{u,v} \subseteq D$. Une valeur faible de $f(u, v)$

suggère que les nœuds u et v disposent de connexions éparses et aléatoires puisqu'ils n'ont pas assez de voisins en commun pour appartenir à un même groupe dense de nœuds. Une valeur nulle de $f(u, v)$ indique que les nœuds u et v n'ont aucun voisin en commun. Globalement, $f(u, v)$ fournit une mesure relative de la force de connexion entre les nœuds d'un graphe multidimensionnel et facilite, par conséquent, la détection des nœuds atypiques qui n'appartiennent pas à une structure dense.

Pour des fins d'illustration, considérons encore une fois l'exemple de la figure 2.1. À partir de ce réseau multidimensionnel, deux structures de groupe denses peuvent être identifiées. Le premier groupe C_1 est défini par les nœuds 1, 2 et 3 qui sont étroitement connectés dans la dimension d_1. Le deuxième groupe C_2 est formé par les nœuds 4, 5, 6 et 7 fermement reliés dans les dimensions d_1 et d_2 (voir figure 2.2a). Les nœuds 0 et 8, cependant, n'appartiennent à aucune région dense identifiable dans le réseau puisqu'ils disposent d'un patron d'interaction irrégulier par rapport aux autres nœuds du réseau. Il convient de noter qu'en inspectant soigneusement le réseau, nous nous apercevons que les connexions des nœuds dans la dimension d_3 sont éparses. Par conséquent, la dimension d_3 est une dimension non pertinente puisqu'elle ne renferme aucune structure significative comme c'est le cas avec les dimensions d_1 et d_2 qui sont des dimensions pertinentes.

Les valeurs $f(u, v)$ qui reflètent la force de connexion entre les nœuds de l'exemple de la figure 2.1 sont montrées par la figure 2.2b. Toutes les pairs de nœuds u et v qui appartiennent à une même région dense et partagent des voisins dans le sous-espace de dimensions qui les connectent ($S_{u,v}$) se caractérisent par des valeurs $f(u, v)$ élevées. Les pairs de nœuds qui partagent peu de voisins ou bien qui n'ont pas de voisins en commun reçoivent des valeurs $f(u, v)$ faibles. Notons que, même si le réseau incorpore une dimension non pertinente d_3, les valeurs retournées par la fonction $f(u, v)$ permettent de discriminer les noeuds faiblement connectés de



(a) Les communautés identifiées. (b) Les valeurs $f(u, v)$ estimées.

Figure 2.2: Analyse de l'exemple 2.1.

ceux qui appartiennent à une communauté.

Les nœuds atypiques sont faiblement connectés aux nœuds du réseau à travers toutes les dimensions. Par conséquent, les valeurs $f(u, v)$ associées à ces nœuds sont faibles et s'approchent de 0. Donc, pour avoir un degré d'anomalie d'un nœud u, nous calculons la moyenne des valeurs $f(u, v)$ de ses voisins directs à travers toutes les dimensions D du réseau. Précisément, nous spécifions un score d'anomalie $AS(u)$ d'un nœud u par la formule suivante :

$$AS(u) = \frac{\sum_{v \in N(u,D)} f(u, v)}{\mid N(u, D) \mid} \tag{2.2}$$

Comme montré dans l'équation 2.2, $AS(u)$ calcule la moyenne des valeurs $f(u, v)$ des voisins d'un nœud u dans toutes les dimensions D. De cette manière, les nœuds qui ont des voisins en commun reçoivent un score $AS(u)$ élevé comparativement aux nœuds aléatoirement connectés qui ne partagent pas suffisamment de voisins avec le reste des nœuds du réseau. Ainsi, les nœuds avec des connexions atypiques auront des scores faibles par rapport aux nœuds densément connectés.

Pour illustrer, reprenons, encore une fois, l'exemple de la figure 2.2b. Les nœuds

0 et 8 reçoivent les scores d'anomalie les plus faibles : $AS(0) = AS(8) = 0$. Les autres nœuds du réseau, étant membres d'une structure dense, reçoivent des scores élevés comparativement aux nœuds 0 et 8. Dans l'ensemble, $AS(u)$ fournit une mesure relative qui facilite la distinction entre les nœuds normaux et les nœuds qui correspondent à des anomalies.

2.3 Phase 2 : Identification automatique des anomalies

Dans cette section, nous expliquons la démarche que nous avons élaborée pour détecter automatiquement les anomalies en nous basant sur les scores $AS(u)$ calculés dans la première phase. Avant de donner des détails sur notre méthode de détection, nous rappelons que les nœuds atypiques se caractérisent par des scores relativement faibles comparativement aux nœuds qui appartiennent à des régions denses. Pour identifier les anomalies, une solution naïve consiste à trier les scores $AS(u)$ selon un ordre ascendant et sélectionner, de suite, les k nœuds ayant les scores les plus faibles comme des anomalies. Le problème majeur de cette solution simple réside dans le choix subjectif de la valeur de k. En d'autres termes, cette solution est sensible à la sélection du paramètre k. Une valeur de k non appropriée affecte d'une façon négative les résultats de la détection. De plus, la spécification d'une valeur k au préalable est davantage problématique et non pratique puisqu'il est difficile de préciser une valeur qui convient à tous les ensembles de données.

Une autre alternative consiste à fixer un seuil de séparation entre les scores $AS(u)$ faibles et élevés. Les nœuds qui ont un score en dessous du seuil s'identifient comme des anomalies. Cependant, l'identification des anomalies sera dépendante du seuil fixé et, par conséquent, sera étroitement affectée si un mauvais seuil est utilisé. De plus, il est difficile de trouver un seuil qui convient à tous les réseaux. Effectivement, pour qu'il soit optimal, un seuil doit être fixé pour chaque réseau d'une manière complètement empirique. Avec une telle démarche informelle, il

est impossible d'être objectif ou cohérent. Afin de résoudre ces problèmes, nous développons dans ce qui suit une approche probabiliste, qui s'appuie sur des considérations théoriques bien établies, pour la détection automatique des anomalies et ce sans avoir à préciser le nombre d'anomalies à identifier ni un seuil empirique du score $AS(u)$ au préalable.

Pour identifier les anomalies, nous nous intéressons aux nœuds ayant les scores $AS(u)$ les plus faibles. Pour ce faire, nous proposons une approche probabiliste qui modélise la distribution statistique des scores $AS(u)$. À partir de cette modélisation, il est possible de partitionner les scores en plusieurs composants. Le composant disposant des scores les plus faibles correspond au composant des nœuds atypiques qu'on cherche. Afin d'avoir une modélisation rigoureuse et précise de la distribution des scores $AS(u)$, nous utilisons un modèle de mélange. Un modèle de mélange permet de décrire la distribution des données au moyen d'une formulation mathématique qui se base sur le fait que les données se subdivisent en des sous-populations homogènes qu'on appelle composants. Plus précisément, nous utilisons le modèle de mélange de la loi beta. La distribution beta admet une grande variété de formes et, par conséquent, permet de modéliser de nombreuses situations avec des données à support fini. Contrairement à d'autres distributions statistiques, la distribution beta permet de modéliser des modes multiples, symétriques et asymétriques (Ma et Leijon, 2009), (Bouguila *et al.*, 2006), (Ji *et al.*, 2005).

L'adaptabilité et la flexibilité de la distribution beta à modéliser des situations complexes et variables sont des caractéristiques uniques dont d'autres distributions ne disposent pas. Par exemple, la distribution gaussienne permet d'avoir une modélisation des données sous une forme de distribution symétrique uniquement. Cependant, dans plusieurs applications, les données ne prennent pas toujours des formes symétriques. Comme mentionné dans (Boutemedjet *et al.*, 2010), la dis-

tribution gaussienne peut mener à une modélisation moins précise des données. Effectivement, à cause de sa restriction à une forme symétrique, la gaussienne risque de surestimer le nombre de composants et de faire un partitionnement erroné des données. En revanche, à l'encontre de plusieurs autres distributions univariées, le modèle de mélange beta se caractérise par une maniabilité qui permet de modéliser une variété de formes. En effet, la distribution de la loi beta peut être : en forme de U, en forme de L, en forme de J, strictement convexe, une droite, strictement concave, symétrique ou non (Ma et Leijon, 2009), (Bouguila *et al.*, 2006), (Ji *et al.*, 2005). De cette manière, la distribution gaussienne et les autres distributions comme la loi Gamma, la loi exponentielle ainsi que les distributions uniformes sont des cas particuliers de la distribution beta. La grande flexibilité de la distribution beta permet une modélisation adéquate des scores d'anomalie, ce qui conduit, par conséquent, à une détection précise des nœuds ayant un patron d'interaction irrégulier (c.-à-d., les anomalies).

2.3.1 Définition du modèle de la distribution beta

Selon la théorie des probabilités et en statistiques, la loi beta est une famille de lois de probabilités continues qui se définit dans l'intervalle [0,1]. Dans notre cas, pour modéliser les scores $AS(u)$, nous procédons, d'abord, à la normalisation des scores entre 0 et 1 sans changer leurs propriétés statistiques. De ce fait, soit ω_i, $(i = 1, \ldots, n)$, les scores normalisés, tel que n est le nombre total des nœuds dans le réseau. Formellement, les scores normalisés $\{\omega_i\}$ suivent une distribution de la forme :

$$F(\omega) = \sum_{l=1}^{p} \alpha_l B_l(\omega, x_l, y_l) \tag{2.3}$$

où $B_l(.)$ indique la densité du $l$ième beta composant, p est le nombre de com-

posants, x_l et y_l $(x_l, y_l > 0)$ sont les paramètres de forme du composant l, α_l $(l = 1, \ldots, p)$ sont les coefficients de mélange avec $\alpha_l > 0$ pour $l = 1, \ldots, p$ et $\sum_{l=1}^{p} \alpha_l = 1$.

Chaque composant de la distribution beta se caractérise par une densité spécifique paramétrée par les deux paramètres de forme x_l et y_l. Formellement, la fonction de densité du $l$ième composant est formulée par :

$$B_l(\omega, x_l, y_l) = \frac{\Gamma(x_l + y_l)}{\Gamma(x_l)\Gamma(y_l)} \omega^{x_l - 1}(1 - \omega)^{y_l - 1} \tag{2.4}$$

tel que $\Gamma(.)$ est la fonction Gamma définie par $\Gamma(\lambda) = \int_0^\infty t^{\lambda - 1} \exp(-t)dt;\ t > 0$.

2.3.2 Estimation des paramètres d'un composant

Pour estimer les paramètres x_l et y_l d'un composant beta, une approche commune consiste à utiliser la technique du maximum de vraisemblance (Ma et Leijon, 2009). L'estimateur du maximum de vraisemblance est un estimateur statistique utilisé pour inférer les paramètres de la distribution de probabilité d'un échantillon donné. Dans notre cas, la fonction de vraisemblance du $l$ième composant est définie comme suit :

$$L_{B_l}(x_l, y_l) = \prod_{\omega \in B_l} B_l(\omega, x_l, y_l) = \left(\frac{\Gamma(x_l + y_l)}{\Gamma(x_l)\Gamma(y_l)}\right)^{n_l} \prod_{i=1}^{n_l} (\omega_i)^{x_l - 1} \prod_{i=1}^{n_l} (1 - \omega_i)^{y_l - 1} \tag{2.5}$$

avec n_l est la taille du $l$ième composant. Pour des raisons de facilité de calcul, le log-vraisemblance (le logarithme de la fonction de vraisemblance) est plus souvent utilisé pour déterminer les valeurs des paramètres x_l et y_l. Explicitement, on l'exprime par :

$$\begin{aligned}\log(L_{B_l}(x_l, y_l)) &= n_l \log(\Gamma(x_l + y_l)) - n_l \log(\Gamma(x_l)) \\ &\quad - n_l \log(\Gamma(y_l)) + (x_l - 1)\sum_{i=1}^{n_l} \log(\omega_i) \\ &\quad + (y_l - 1)\sum_{i=1}^{n_l} \log(1 - \omega_i) \end{aligned} \tag{2.6}$$

L'estimation des paramètres x_l et y_l qui maximisent la fonction de vraisemblance revient à considérer l'égalité à zéro du différentiel de $\log(L_{B_l}(x_l, y_l))$ par rapport aux variables x_l et y_l comme formulé par ce qui suit :

$$\frac{\partial}{\partial x_l}\log(L_{B_l}(x_l, y_l)) = \frac{n_l\Gamma'(x_l + y_l)}{\Gamma(x_l + y_l)} - \frac{n_l\Gamma'(x_l)}{\Gamma(x_l)} + \sum_{i=1}^{n_l}\log(\omega_i) = 0 \tag{2.7}$$

et

$$\frac{\partial}{\partial y_l}\log(L_{B_l}(x_l, y_l)) = \frac{n_l\Gamma'(x_l + y_l)}{\Gamma(x_l + y_l)} - \frac{n_l\Gamma'(y_l)}{\Gamma(\beta_l)} + \sum_{i=1}^{n_l}\log(1 - \omega_i) = 0 \tag{2.8}$$

Le système d'équations (2.7) et (2.8) n'admet pas de solution analytique précise. Touefois, les paramètres $\hat{x_l}$ et $\hat{y_l}$ peuvent être estimés itérativement par la résolution du système d'équations (2.7) et (2.8) en utilisant les méthodes de calcul itératif comme la méthode de Newton-Raphson (Jennrich et Robinson, 1969). En effet, la méthode de Newton-Raphson est un algorithme itératif qui sert à trouver une solution numérique aux équations en question. Précisément, nous estimons le vecteur de paramètres $\hat{\theta}_l = (\hat{x_l}, \hat{y_l})^T$ itérativement par

$$\hat{\theta}_l^{(I+1)} = \hat{\theta}_l^{(I)} - h_l^T H_l^{-1} \tag{2.9}$$

tel que I est un index d'itération, h_l^T est le vecteur de la première dérivée et H_l^{-1} est la deuxième dérivée du logarithme de la fonction de vraisemblance du $l$ième composant. Le vecteur h_l est défini par :

$$h_l = \begin{pmatrix} h_l^1 \\ h_l^2 \end{pmatrix} \tag{2.10}$$

avec

$$\begin{aligned} h_l^1 &= n_l\Big[\psi(x_l + y_l) - \psi(x_l)\Big] + \sum_{i=1}^{n_l} \log(\omega_i) \\ h_l^2 &= n_l\Big[\psi(x_l + y_l) - \psi(y_l)\Big] + \sum_{i=1}^{n_l} \log(1 - \omega_i) \end{aligned} \tag{2.11}$$

tel que $\psi(.)$ est la fonction digamma définie par $\psi(\lambda) = \frac{\Gamma'(\lambda)}{\Gamma(\lambda)}$.

La matrice H_l est définie par :

$$H_l = \begin{pmatrix} \frac{\partial h_l^1}{\partial x_l} & \frac{\partial h_l^1}{\partial y_l} \\ \frac{\partial h_l^2}{\partial x_l} & \frac{\partial h_l^2}{\partial y_l} \end{pmatrix} \tag{2.12}$$

avec

$$\begin{aligned} \frac{\partial h_l^1}{\partial x_l} &= n_l[\psi(x_l + y_l) - \psi'(x_l)], \\ \frac{\partial h_l^1}{\partial y_l} &= \frac{\partial h_l^2}{\partial x_l} = n_l[\psi'(x_l + y_l)], \\ \frac{\partial h_l^2}{\partial y_l} &= n_l[\psi'(x_l + y_l) - \psi'(x_l)] \end{aligned} \tag{2.13}$$

$\psi'(.)$ est la fonction trigamma donné par $\psi'(\lambda) = \frac{\Gamma''(\lambda)}{\Gamma(\lambda)} - [\frac{\Gamma'(\lambda)}{\Gamma(\lambda)}]^2$.

L'algorithme Newton-Raphson converge à l'itération où l'estimation des paramètres x_l et y_l change par une valeur inférieure à une petite valeur positive ϵ qui permet d'atteindre les valeurs $\hat{x}_l$ et $\hat{y}_l$. Il convient de noter que, dans notre implémentation, nous utilisons les estimateurs de moments de la distribution beta pour définir les valeurs initiales de $\hat{\theta}_l^{(0)}$ dans (2.9) (Bain et Engelhardt, 1987). Dans cette technique, la moyenne et la variance attendues sont égales respectivement à la moyenne et à la variance de l'échantillon. Plus précisément, les estimateurs des moments utilisés sont :

$$\hat{x_l}^{(0)} = \bar{X}_l\Big[\frac{\bar{X}_l(1-\bar{X}_l)}{S_l^2} - 1\Big],$$
$$\hat{y_l}^{(0)} = (1-\bar{X}_l)\Big[\frac{\bar{X}_l(1-\bar{X}_l)}{S_l^2} - 1\Big] \tag{2.14}$$

où $\bar{X}_l$ est la moyenne de l'échantillon et S_l^2 est la variance du $l$ième composant.

2.3.3 L'algorithme EM pour la distribution beta

Afin de trouver les paramètres du maximum de vraisemblance de la distribution beta, le calcul itératif de la solution de l'équation de vraisemblance peut se faire par l'application de l'algorithme Espérance-Maximisation (*Expectation-Maximization*), souvent abrégé EM (Dempster *et al.*, 1977). D'une itération à une autre, l'algorithme EM permet d'assurer que les valeurs prises par la vraisemblance augmentent d'une façon monotone jusqu'à la convergence. Cela se fait en deux étapes successives : une étape d'évaluation de l'espérance (E), où nous calculons l'espérance de la vraisemblance en tenant compte des dernières variables observées, et une étape de maximisation (M), où nous estimons le maximum de vraisemblance des paramètres en maximisant la vraisemblance trouvée à l'étape E.

Ce problème d'estimation de modèles de mélange dans le cadre de l'algorithme EM implique la formulation du problème en un problème de données manquantes. Ainsi, nous enrichissons les données par l'introduction d'une variable indicatrice z_{il}, $(i = 1, \ldots, n)$, $(l = 1, \ldots, p)$ pour chaque score d'anomalie normalisé ω_i. Ici, il convient de noter que z_{il} est une variable cachée qui indique quelle composante du mélange a permis de générer chacune des observations. La variable indicatrice z_{il} prend une valeur binaire qui renseigne sur l'appartenance d'un score normalisé ω_i à un composant l tel que précisé par (2.15).

$$z_{il} = \begin{cases} 1 & si\ \omega_i\ appartient\ au\ composant\ l \\ 0 & sinon \end{cases} \tag{2.15}$$

Ainsi, l'ensemble des données est défini par les ensembles de valeurs $\{z_{il}\}$ et $\{\omega_i\}$. De ce fait, la fonction de vraisemblance peut, désormais, être réécrite comme :

$$L_{B_l}(\Theta, \alpha, z) = \prod_{i=1}^{n} \prod_{l=1}^{p} (\alpha_l B_l(\omega_i, x_l, y_l))^{z_{il}} \tag{2.16}$$

La version logarithmique de la fonction de vraisemblance est redéfinie par :

$$log(L_{B_l}(\Theta, \alpha, z)) = \sum_{i=1}^{n} \sum_{l=1}^{p} z_{il} log(\alpha_l B_l(\omega_i, x_l, y_l)) \tag{2.17}$$

tel que

- $\Theta = \{x_1, y_1, \ldots, x_p, y_p\}$ sont les paramètres inconnus du modèle de mélange.
- $z = \{z_1, \ldots, z_n\}$, avec $z = (z_{i1}, \ldots, z_{ip})^T$ est le vecteur des variables indicatrices z_{il}.
- $\alpha = (\alpha_1, \ldots, \alpha_p)$ sont les coefficients de mélange qui représentent les proportions d'objets dans chaque composant.

Selon cette perspective, l'algorithme EM peut être appliqué en considérant les variables z_{il} comme des données manquantes. Pour estimer les paramètres du modèle, chaque donnée manquante est remplacée par son espérance :

$$\hat{z_{il}}^{(t)} = \frac{\hat{\alpha_l}^{(t)} B_l(\omega_i, \hat{x_l}^{(t)}, \hat{y_l}^{(t)})}{\sum_{j=1}^{p} \hat{\alpha_j}^{(t)} B_l(\omega_i, \hat{x_j}^{(t)}, \hat{y_j}^{(t)})} \tag{2.18}$$

Algorithme 1 : L'algorithme EM

Entrée : $\{\omega_i\}$

Résultat : $\hat{\Theta}, \hat{\alpha}, \hat{z}$

début

répéter

Étape E :

- Calculer $\hat{z_{il}}^{(t)}$ en utilisant l'équation (2.18).

Étape M :

- Calculer les coefficients de mélange de chaque composant de sorte que $\hat{\alpha_l}^{(t+1)} = \frac{\sum_{i=1}^{n} \hat{z_{il}}^{(t)}}{n}$

- Estimer $\hat{\theta}^{(t+1)}$ en utilisant (2.9), (2.11) et (2.13).

jusqu'à *Convergence :* $|log(L_{B_l}(\Theta, \alpha, z))^{(t+1)} - log(L_{B_l}(\Theta, \alpha, z))^{(t)}| \simeq 0$;

fin

où t dénote l'indice de l'itération courante. L'algorithme EM itère les deux étapes d'éspérance (E) et maximisation(M) pour produire une séquence de valeurs $\{\hat{\theta}\}^{(t)}$, $(t = 0, 1, 2, \ldots)$. L'algorithme s'arrête lorsque le changement du logarithme de fonction de vraisemblance donné par l'équation (2.17) devient négligeable.

Il convient de noter que l'algorithme EM nécessite une initialisation des paramètres de chaque composant. Toutefois, l'algorithme est sensible à l'initialisation (Figueiredo et Jain, 2002). Par conséquent, il est plus approprié d'initialiser les paramètres en utilisant un algorithme de partitionnement non supervisé (*clustering*). À cette fin, nous implémentons l'algorithme K-means pour partitionner les scores $\{\omega_i\}_{i=1,\ldots,n}$ en p composants (Hartigan et Wong, 1979). Après le partitionnement, nous estimons les paramètres initiaux $x_l^{(0)}$ et $y_l^{(0)}$ de chaque composant.

2.3.4 Estimation du nombre de composants p

L'utilisation de la distribution beta permet d'avoir une modélisation flexible des données qui décrit adéquatement la distribution des scores d'anomalies. Pour construire le modèle, il faut estimer également le nombre de composants p et les paramètres associés à chaque composant. Une façon d'aborder le problème consiste à varier le nombre de composants p de 1 à p_max et à calculer à chaque itération des métriques de performance pour identifier un nombre optimal de composants. Pour ce faire, nous implémentons un processus à deux étapes. Dans un premier lieu, nous calculons les paramètres qui maximisent la fonction de vraisemblance en variant p de 1 à p_max. Dans un deuxième lieu, nous calculons un critère de performance pour les différents paramètres estimés et nous sélectionnons la valeur p qui revoit la valeur optimale du critère de performance.

Plusieurs critères informationnels ont été proposés pour aider à estimer le nombre de composants dans les données (Smyth, 2000). Dans notre méthode, nous utilisons le critère d'information bayésien BIC (*Bayesian Information Criterion*) qui a été initialement proposé par Schwarz (Schwarz *et al.*, 1978). Le critère BIC est illustré par la formule suivante :

$$BIC(p) = -2L_p + Nb_p \log(N) \tag{2.19}$$

tel que L_p est le logarithme de la vraisemblance du modèle estimé et Nb_p est le nombre de paramètres estimés. Le nombre optimal de composants p est le nombre de composants qui minimise $BIC(p)$. La procédure d'estimation du nombre de composants du modèle de mélange est donnée par l'Algorithme 2.

Algorithme 2 : Estimation du nombre optimal de composants p

Entrée : $\{\omega_i\}$, p_max

Résultat : Le nombre optimal de composants p

début

  pour $p = 1$ **à** p_max **faire**

    si $p==1$ **alors**

      Estimer $\hat{x}$ et $\hat{y}$ en utilisant la méthode de Newton-Raphson basée sur (2.9);

      Calculer la valeur $BIC(p)$ en utilisant (2.19);

    sinon

      Appliquer K-means pour initialiser les paramètres de l'algorithme EM;

      Appliquer l'Algorithme 1 pour estimer les paramètres de la distribution $\hat{x}_l$ et $\hat{y}_l$ $(l = 1, \ldots, p)$;

      Calculer la valeur de $BIC(p)$ en utilisant (2.19) ;

    fin

  fin

  Sélectionner le nombre de composants $\hat{p}$, tel que $\hat{p} = \arg_{min} BIC(p)$;

fin

Une fois le nombre optimal de composants est identifié, il est possible d'utiliser les résultats de l'algorithme EM pour dériver une classification afin d'affilier chaque score ω_i au composant approprié du modèle de mélange. L'affiliation d'un ω_i à un composant se fait selon le maximum à postériori de la variable indicatrice $\hat{z_{il}}$. Pour identifier les anomalies, nous nous intéressons à identifier le composant beta qui correspond aux valeurs les plus faibles de ω_i. Par conséquent, les nœuds associés à ces scores sont les nœuds atypiques (les anomalies) qu'on cherche. L'Algorithme 3 présente toutes les étapes suivies pour identifier les nœuds anomalies.

Algorithme 3 : Identification automatique des anomalies

Entrée : Un réseau multidimensionnel G

Résultat : Les anomalies AN

début

Pour chaque noeud u dans G estimer $AS(u)$ en utilisant (2.2);

Estimer $\{\omega_i\}_{i=1,\ldots,n}$ en normalisant les valeurs $AS(u)$ entre 0 et 1;

Estimer la fonction de densité de probabilité des scores d'anomalies normalisés pour différentes valeurs p où $p = 1, \ldots, p\_max$ en utilisant l'Algorithme 2;

Sélectionner le modèle de mélange avec le nombre optimal de composants qui minimise BIC;

Utiliser les resultats de EM pour classifier les ω_i dans le composant correspondant;

Selectionner le composant beta qui correspond aux valeurs ω_i les plus petites;

Identifier les nœuds de G associés aux ω_i qui appartiennent au composant sélectionné et les sauvegarder dans AN;

Retourner AN;

fin

2.4 Conclusion

Dans ce chapitre, nous avons présenté une nouvelle approche de détection d'anomalies dans les réseaux multidimensionnels. L'approche proposée procède en deux phases. Dans la première phase, nous élaborons une fonction qui relate la force de connexion entre les pairs de nœuds d'un réseau multidimensionnel afin de calculer un score d'anomalie pour chaque nœud. La fonction permet d'analyser un réseau multidimensionnel en tant que tel, sans considérer une agrégation de dimensions ou une analyse indépendante de chaque dimension. Dans la deuxième phase, nous développons une méthode probabiliste qui se base sur la distribution beta. Le but est de modéliser la distribution des scores d'anomalie pour identifier

automatiquement les nœuds atypiques. En utilisant cette approche, on est en mesure d'identifier avec précision les anomalies dans les réseaux multidimensionnels même en présence de dimensions non pertinentes. Les résultats des expérimentations dans le chapitre suivant corroborent nos propos.

CHAPITRE III

ÉVALUATION DE L'APPROCHE PROPOSÉE

Ce chapitre présente une évaluation empirique de la performance de notre approche sur une variété de réseaux synthétiques et réels. À noter que, dans la littérature courante, il n'y a aucune méthode de détection d'anomalies dans les réseaux multidimensionnels à laquelle notre approche peut être comparée. De plus, tel que discuté dans l'introduction, nous ne pouvons pas considérer une méthode de détection d'anomalies dédiée aux réseaux monodimensionnels pour les fins de comparaison puisque ces méthodes ne sont pas appropriées au cas de notre étude. Dans ce contexte, nous élaborons un ensemble d'expérimentations pour évaluer de façon objective l'applicabilité de notre approche. Nous commençons d'abord par décrire le cadre expérimental et analyser les résultats de notre approche sur des réseaux synthétiques. Par la suite, nous présentons une description détaillée des données réelles testées et nous interprétons la pertinence des résultats obtenus.

3.1 Expérimentations sur des réseaux synthétiques

3.1.1 Mécanisme de génération

Pour démontrer l'efficacité de notre approche, nous menons, de prime abord, des expérimentations sur des réseaux artificiellement générés. Pour ce faire, nous adoptons le modèle de génération de réseaux synthétiques basé sur le modèle de parti-

tion plantée, défini dans (Condon et Karp, 2001). Ce modèle de génération offre un environnement contrôlable qui permet de simuler des configurations diverses de réseaux multidimensionnels, et ce en variant un certain nombre de paramètres. Plus précisément, nous spécifions le nombre de nœuds n du réseau. Nous précisons également le nombre de structures denses ainsi que l'intervalle de leurs densités interne et externe définies respectivement par $[\vartheta_{int_{min}}, \vartheta_{int_{max}}]$ et $[\vartheta_{ext_{min}}, \vartheta_{ext_{max}}]$.

De plus, nous fixons le nombre total de dimensions tout en indiquant le nombre de dimensions pertinentes et non pertinentes. Ici, il convient de souligner que chaque dimension, qu'elle soit pertinente ou non pertinente, dispose d'un bruit de fond. Le bruit de fond est un ensemble de connexions aléatoires et éparses entre l'ensemble des nœuds du réseau. La densité du bruit de fond est sélectionnée aléatoirement à partir d'une marge de valeur $[\kappa_{min}, \kappa_{max}]$. Il est à noter que les structures denses sont plantées sur le bruit de fond des dimensions pertinentes uniquement. De ce fait, une structure dense peut apparaître dans une ou plusieurs dimensions pertinentes. Lors du processus de génération, le nombre de dimensions pertinentes ainsi que la densité de connexion interne et externe d'une structure dense particulière sont sélectionnés aléatoirement à partir des marges de valeurs introduites en paramètres. Par ailleurs, les anomalies correspondent aux nœuds du bruit de fond qui ne font partie d'aucune structure dense sur toutes les dimensions du réseau.

Le réseau synthétique de la figure 0.2 du chapitre Introduction est un exemple produit par le générateur qui vient d'être décrit. Cet exemple est simple et a été fourni à titre illustratif, dans l'intention de montrer que les interactions entre les nœuds diffèrent en fonction des dimensions et en fonction de l'appartenance aux régions denses. Toutefois, les réseaux générés pour l'évaluation sont plus complexes et relatent des situations diverses.

Tableau 3.1: Description des réseaux synthétiques.

	#Nœuds	#Normaux #Anomalies	#Dimensions	#Pertinentes #Non_Pertinentes
Réseau_1	500	475 25	4	3 1
Réseau_2	1 000	950 50	5	3 2
Réseau_3	3 000	2 850 150	6	3 3
Réseau_4	5 000	4 750 250	8	3 5
Réseau_5	10 000	9 500 500	12	3 9

Le tableau 3.1 résume les caractéristiques des cinq réseaux synthétiques que nous évaluons dans ce mémoire parmi plusieurs autres réseaux artificiels testés. Le processus de génération a été fait en variant différents paramètres. Autrement dit, chaque réseau synthétique dispose d'une configuration particulière. Précisément, chacun des réseaux artificiellement générés se distingue par un nombre spécifique de nœuds normaux et d'anomalies ainsi qu'un nombre spécifique de dimensions pertinentes et non pertinentes. Comme indiqué par le tableau 3.1, nous avons varié la taille des réseaux de 500 à 10 000 nœuds ainsi que le nombre d'anomalies de 25 à 500 nœuds. Nous avons également varié la dimensionnalité des réseaux de 4 à 12 tout en fixant le nombre de dimensions pertinentes à 3 et en variant le nombre de dimensions non pertinentes de 1 à 9. Par ce fait, nous examinons l'adaptabilité de notre méthode dans des réseaux avec un nombre d'anomalies et un nombre de dimensions non pertinentes élevés. Par le fait même, nous évaluons

la robustesse et la capacité de notre approche à identifier les anomalies dans des situations diverses.

3.1.2 Critères d'évaluation

Pour évaluer la performance de notre approche, nous avons considéré un ensemble de critères d'évaluation. Les critères que nous utilisons représentent des métriques utilisées pour l'évaluation de la qualité des résultats obtenus à partir de réseaux où la nature des nœuds est connue à l'avance. Il convient de noter que, dans le cas des réseaux synthétiques, à l'encontre des réseaux réels, la nature des nœuds est connue *à priori*. De ce fait, dans les cinq réseaux synthétiques générés, nous distinguons les nœuds normaux et les anomalies. Ce partitionnement n'est pas considéré au moment de l'application de notre approche, mais on l'utilise, par la suite, comme référence (*ground truth*) pour mesurer la conformité des résultats. Dans l'évaluation de notre approche sur les réseaux artificiels, nous considérons les indicateurs de performance suivants :

- Le coefficient de corrélation de Matthews (MCC) (Matthews, 1975) est une mesure utilisée pour évaluer la correspondance entre le partitionnement identifé et le partitionnement de référence. Cette mesure est utilisée lorsque la distribution des données entre les différentes catégories d'objets est disproportionnelle. Les valeurs de cet indicateur sont comprises entre -1 et +1 et se calculent comme suit :

$$MCC = \frac{VP \times VN - FP \times FN}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}} \tag{3.1}$$

  tel que

  - VP : le nombre de vrais positifs, c-à-d, le nombre des anomalies correctement identifiées comme anomalies.

- VN : le nombre de vrais négatifs, c-à-d, le nombre des nœuds normaux correctement identifiés comme normaux.
- FP : le nombre de faux positifs, c-à-d, le nombre de nœuds normaux identifiés comme des anomalies.
- FN : le nombre de faux négatifs, c-à-d, le nombre d'anomalies identifiées comme des nœuds normaux.

La mesure MCC prend en considération les vrais et faux positifs et négatifs. Des valeurs proches de +1 indiquent une bonne identification de la nature des nœuds. En revanche, des valeurs qui s'approchent de -1 témoignent une détection aléatoire.

- Le taux de vrais positifs (TVP) mesure la proportion des nœuds anomalies qui ont été correctement identifiés comme des anomalies.

$$TVP = \frac{VP}{VP + FN} \tag{3.2}$$

La valeur TVP est comprise entre 0 et 1. Si la valeur du TVP est proche de 1, cela signifie que peu de nœuds normaux ont été classifiés comme anomalies et que la classification peut être considérée comme "précise". De ce fait, une valeur TVP élevée suggère un taux de détection correct élevé.

- Le taux de faux positifs (TFP) correspond à la proportion des nœuds normaux qui ont été incorrectement identifiés comme des anomalies.

$$TFP = \frac{FP}{FP + VN} \tag{3.3}$$

Tout comme le TVP, le TFP prend des valeurs dans l'intervalle [0,1]. Une valeur proche de 0 indique que la détection des nœuds normaux a été faite avec précision. Une valeur proche de 1 indique une mauvaise identification des nœuds normaux.

- La F-mesure est la moyenne harmonique entre la précision et le rappel de la classe des anomalies. Cet indicateur est formulé par :

$$F - mesure = \frac{2 \times Precision \times Rappel}{Precision + Rappel} \tag{3.4}$$

tel que :

$$Precision = \frac{VP}{VP + FP} \tag{3.5}$$

et

$$Rappel = \frac{VP}{VP + FN} \tag{3.6}$$

Les valeurs de cette métrique varient entre 0 et 1. Plus la valeur de F-mesure s'approche de 1, mieux la détection se conforme au partitionnement de référence.

3.1.3 Résultats et discussion

Avant de discuter les résultats obtenus pour les réseaux synthétiques, il convient de préciser que nous avons fixé la valeur de p_max à 5. Comme discuté dans le chapitre précédent, nous varions le nombre de composants de 1 à p_max. Nous sélectionnons, ensuite, le nombre optimal de composants qui minimise le critère BIC. Ici, le lecteur doit être avisé que la valeur de p_max n'est pas restreinte à 5 et que toute autre valeur entière peut être utilisée. Toutefois, après plusieurs expérimentations sur différents réseaux, nous avons découvert que, dans la majorité des cas, le nombre optimal de composants du modèle de mélange beta varie de 2 à 3. Par ailleurs, les anomalies sont discernées par le composant beta ayant les scores d'anomalies les plus faibles. Les nœuds associés à ces scores sont identifiés comme des anomalies.

Le tableau 3.2 illustre les résultats d'évaluation de notre approche sur les 5 réseaux synthétiques décrits dans le tableau 3.1. Comme on peut le voir à partir du

Tableau 3.2: Les résultats d'évaluation des réseaux synthétiques.

	MCC	TVP	TFP	F-mesure
Réseau_1	0.9602	96.00%	0.00%	0.9795
Réseau_2	0.9044	90.00%	0.40%	0.9000
Réseau_3	0.9681	96.00%	0.10%	0.9696
Réseau_4	0.9593	93.20%	0.04%	0.9608
Réseau_5	0.9968	99.80%	0.03%	0.9960

tableau 3.2, notre approche montre une bonne performance lorsqu'on augmente le nombre d'anomalies et le nombre de dimensions non pertinentes. Effectivement, la moyenne des critères d'évaluation MCC et F-meausre sur les cinq réseaux synthétiques testés sont, respectivement, de l'ordre de 0.9577 et 0.9611. Ces résultats démontrent la conformité de la détection par rapport au *ground truth* utilisé. D'autant plus, notre méthode retourne une moyenne égale à 0.114% de TFP, c.-à-d., la proportion des nœuds normaux qui ont été prédits comme anomalies. Ici, il convient de souligner que 0.4% est le taux de faux positifs le plus élevé produit par notre approche sur le Réseau_2. Ces chiffres démontrent clairement que peu de nœuds normaux ont été incorrectement identifiés comme des anomalies.

Dans l'ensemble de ces expérimentations, quelques anomalies ont été identifiées comme des nœuds normaux. Un taux moyen de 95% de TVP a été rapporté par notre approche sur l'ensemble des cinq réseaux testés. Cela veut dire que, en moyenne, 5% des nœuds anomalies ont été incorrectement identifiés comme des nœuds normaux. Ce partitionnement n'est pas nécessairement incorrect. En se basant sur le fait que les nœuds atypiques sont générés aléatoirement tout au long des différentes dimensions du réseau, il est probable que les anomalies se placent dans des régions denses. Dans ces circonstances, certains nœuds atypiques peuvent

possiblement recevoir des scores d'anomalie, $AS(u)$, élevés et seront considérés, par conséquent, comme des nœuds normaux. En grande partie, les résultats présentés par le tableau 3.2 suggèrent que l'approche proposée performe bien dans différents réseaux avec un nombre variable d'anomalies. De plus, l'identification des anomalies demeure cohérente même dans le cas de réseaux caractérisés par un nombre de dimensions non pertinentes elevé.

3.2 Expérimentations sur des réseaux réels

Dans cette section, nous évaluons l'efficacité de notre approche sur cinq réseaux réels : (1) deux réseaux de collaboration en recherche scientifique DBLP, (2) un réseau agrégé de médias sociaux Friendfeed, (3) un réseau d'interactions de protéines de la mouche à fruits Drosophila Melanogaster et (4) un réseau de transport aérien européen. Contrairement aux réseaux synthétiques, nous n'avons aucune connaissance *à priori* sur le partitionnement des nœuds, c.-à-d., une connaissance au préalable de l'appartenance des nœuds à une catégorie spécifique, à savoir, anomalie ou nœud normal, est manquante. De ce fait, nous ne pouvons pas utiliser des métriques supervisées qui se basent sur l'existence d'un partitionnement de référence comme on a fait dans les réseaux synthétiques. Dans ce contexte, nous avons adopté une approche objective qui consiste à l'interprétation des interactions des nœuds par une investigation manuelle et une visualisation graphique des matrices d'adjacence des réseaux. Dans ce qui suit, nous présentons une description de l'ensemble des réseaux réels testés et nous analysons les résultats obtenus par notre approche.

3.2.1 Description des réseaux réels

- **Les réseaux de collaboration scientifique DBLP**

  Nous proposons l'analyse de deux réseaux multidimensionnels qui ont été extraits du réseau de collaboration scientifique en ligne DBLP. Nous les carac-

térisons par DBLP1 et DBLP2. Les réseaux DBLP1 et DBLP2 incorporent, respectivement, 1 230 et 3 090 nœuds (auteurs). Dans les deux réseaux, les nœuds s'interconnectent via trois dimensions d'analyse. Dans la première dimension, une connexion existe entre deux auteurs s'il y a des citations entre eux. Dans la deuxième dimension, deux auteurs se connectent s'ils ont écrit un article ensemble. La troisième dimension représente les connexions entre les auteurs qui ont publié des articles ayant en commun au moins trois mots clés dans le résumé ou le titre (Papalexakis *et al.*, 2013).

- **Le réseau agrégé de médias sociaux Friendfeed**

  Ce réseau est extrait de l'agrégateur de médias sociaux Friendfeed (Celli *et al.*, 2010). Dans ce système, les utilisateurs peuvent publier des messages et peuvent aussi les commenter comme dans Facebook ou autre média social. Les utilisateurs ont la possibilité d'associer leur compte Friendfeed à d'autres systèmes comme YouTube et Twitter. Dans notre cas, le réseau étudié est un réseau tridimensionnel qui modélise l'interaction de 6 407 utilisateurs. Ces utilisateurs sont, à la base, des utilisateurs de la plateforme Friendfeed, et ils associent leur(s) compte(s) Youtube et/ou Twitter à leur compte Friendfeed. Dans ce réseau, l'analyse se fait selon trois types de dimensions : (1) Friendfeed, (2) Twitter et (3) YouTube. Ces dimensions relatent les connexions que peut avoir un utilisateur avec les autres membres du réseau sur trois systèmes différents.

- **Le réseau d'interactions de protéines de Drosophila Melanogaster**

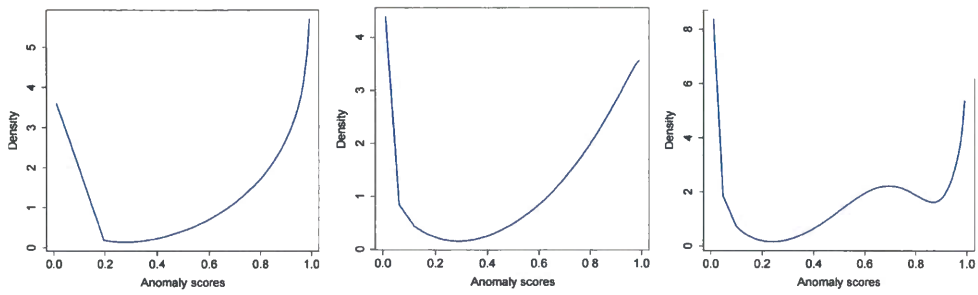
  Ce réseau représente l'interaction de 8 215 protéines (nœuds) de la mouche à fruits Drosophila Melanogaster (De Domenico *et al.*, 2015), (Stark *et al.*, 2006). L'interaction entre les protèines est définie par sept types de dimensions (d_1 : Interaction directe, d_2 : Interaction génétique suppressive, d_3 : Interaction génétique additive, d_4 : Association physique, d_5 : Colocalisa-

tion, d_6 : Association et d_7 : Interaction génétique synthétique).

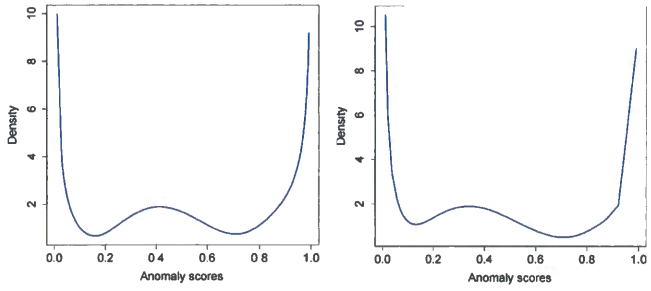
- **Le réseau de transport aérien européen**
  Ce réseau modélise l'interaction de 450 aéroports européens (nœuds) à travers 37 dimensions. Chaque dimension correspond à une compagnie aérienne (Cardillo *et al.*, 2013). Dans ce réseau, les connexions représentent des vols offerts par différentes compagnies aériennes qui relient les aéroports de départ et de destination. Particulièrement, ce réseau a la distinction d'avoir un nombre élevé de dimensions comparativement aux autres réseaux réels et plusieurs parmi ses dimensions sont des dimensions avec des connexions éparses.

3.2.2 Résultats et discussions

Dans cette section, nous discutons les résultats d'application de notre approche sur les réseaux réels précédemment décrits. Pour chaque réseau, nous estimons les scores d'anomalie des nœuds. Nous modélisons, ensuite, la distribution de ces scores selon notre modèle probabiliste qui exploite la distribution beta. Tout comme dans les données synthétiques, nous fixons la valeur de p_max à 5 et nous sélectionnons, après, le nombre optimal de composants qui minimise le critère BIC. Suite à plusieurs tests, nous trouvons que le nombre optimal de composants beta varie entre 2 et 3. Cela est clairement visible à partir des courbes de densité des scores d'anomalies des cinq réseaux réels testés (voir figure 3.1). Les courbes de densité nous permettent de constater la grande flexibilité du modèle beta et son adaptabilité à modéliser des distributions de données ayant des formes diverses. Dans chaque courbe de la figure 3.1, le premier composant (proche de zéro) représente les valeurs des scores d'anomalies les plus faibles. Par conséquent, les nœuds associés aux scores qui sont groupés dans ce composant sont identifiés comme des anomalies.



(a) DBLP1. (b) DBLP2. (c) Friendfeed.



(d) Drosophila. (e) Air transportation.

Figure 3.1: La densié de probabilité des scores estimés pour les 5 réseaux réels.

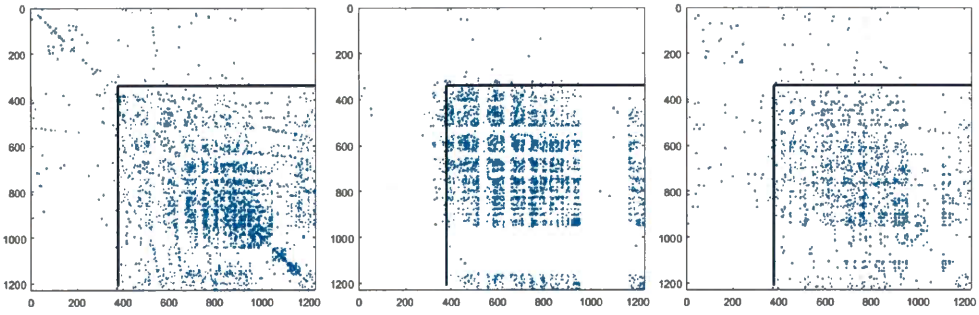
3.2.2.1 Résultats des réseaux DBLP1 et DBLP2

Dans cette section, nous détaillons les résultats obtenus pour les réseaux de collaboration scientifique DBLP1 et DBLP2. Pour le réseau DBLP1, 371 d'un ensemble de 1 230 nœuds ont été identifiés comme des anomalies en ayant des scores en dessous de 0,22. Une majorité de ces anomalies ont reçu des scores nuls ou avoisinant zéro. Néanmoins, la plupart des nœuds normaux ont eu un score en dessus de 0,8. Similairement, dans le réseau DBLP2 qui contient 3 090 nœuds, 750 nœuds ont été identifiés comme des nœuds atypiques avec une marge de score dans l'intervalle [0 ;0,25]. Tel est le cas pour DBLP1, la plupart des nœuds normaux dans DBLP2 ont acquis un score d'anomalies en dessus de 0,8, tandis que le plus grand nombre d'anomalies a obtenu un score avoisinant zéro. À travers les courbes de

densité des réseaux DBLP1 et DBLP2 illustrées par les figures 3.1a et 3.1b, nous constatons la présence de deux composantes. Le premier composant, dans les deux courbes, représente l'ensemble des valeurs faibles des scores d'anomalies, à savoir, les valeurs des scores associés aux anomalies identifiées. Le deuxième composant représente les valeurs élevées des scores qui correspondent aux valeurs associées aux nœuds normaux détectés.

Les réseaux DBLP1 et DBLP2 modélisent la collaboration entre auteurs. Par ce fait, les nœuds qui ont été identifiés comme des anomalies représentent des auteurs qui n'ont pas suffisamment d'interaction avec les autres membres du réseau sur les trois dimensions d'analyse. En effet, dans notre investigation, nous avons constaté que ces membres n'ont pas une contribution significative comparativement aux auteurs associés aux nœuds normaux identifiés. Pour fournir une illustration visuelle du résultat obtenu, nous montrons les figures 3.2 et 3.3 qui illustrent, respectivement, les matrices d'adjacence des réseaux DBLP1 et DBLP2. Il convient de souligner que, dans cette représentation, les nœuds sont placés selon leurs scores d'anomalies dans un ordre ascendant.

Dans ces figures, chaque dimension contient un bloc dense de nœuds (délimité par une ligne noire solide) entouré par une région éparse de nœuds qui correspondent aux nœuds atypiques. Ces nœuds atypiques modélisent les auteurs qui ont des connexions irrégulières sur les trois dimensions d'analyse. De ce fait, ils sont des auteurs qui n'ont pas cité ou coécrit assez d'articles. De plus, il est possible que leurs publications n'aient pas en commun un minimum de trois mots clés dans le titre ou le résumé avec les publications des autres auteurs du réseau (DBLP1 et DBLP2).

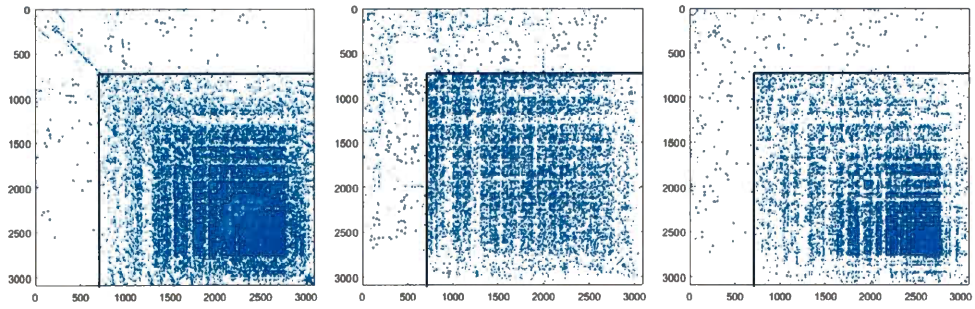


(a) d_1 : Citation commune.

(b) d_2 : Co-réalisation d'un article.

(c) d_3 : Mots clés en commun.

Figure 3.2: Les matrices d'adjacence de DBLP1.



(a) d_1 : Citation commune.

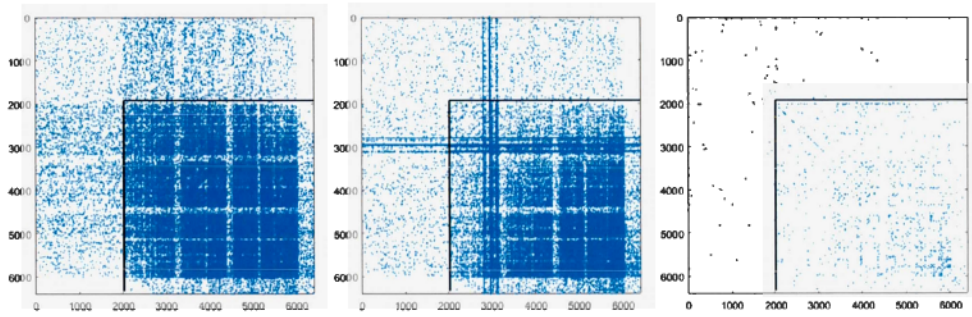
(b) d_2 : Co-réalisation d'un article.

(c) d_3 : Mots clés en commun.

Figure 3.3: Les matrices d'adjacence de DBLP2.

3.2.2.2 Résultats du réseau Friendfeed

En ce qui concerne le réseau social Friendfeed, 2 175 nœuds d'un ensemble de 6 407 ont été sélectionnés comme des nœuds ayant des connexions atypiques. Tout comme les réseaux DBLP, nous montrons dans la figure 3.4 les matrices d'adjacences des trois dimensions du réseau de sorte que les nœuds soient triés suivant un ordre ascendant de leurs scores d'anomalie. Ayant les scores $AS(u)$ les



(a) FriendFeed. (b) Twitter. (c) YouTube.

Figure 3.4: Les matrices d'adjacence de Friendfeed.

plus faibles, les anomalies sont placées en haut de chaque matrice. Tels qu'illustrés par la figure 3.4, les nœuds atypiques sont connectés d'une façon éparse sur les trois dimensions du réseau. Par contre, les nœuds normaux sont étroitement connectés, particulièrement, sur les dimensions Friendfeed et Twitter. Cela se manifeste par les blocs denses des figures 3.4a et 3.4b. Ici, il est approprié de rappeler que les utilisateurs de Friendfeed peuvent utliser leurs comptes Friendfeed pour voir le contenu de leurs comptes Twitter et Youtube s'ils sont associés. L'échantillon de données que nous avons utilisé suggère que les participants sont plus actifs dans Friendfeed et Twitter que sur Youtube.

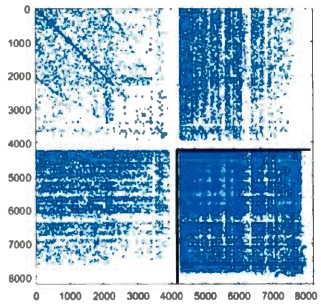
3.2.2.3 Résultats du réseau d'interactions de protéines de Drosophila Melanogaster

Dans le réseau d'interactions de protéines de Drosophila Melanogaster, 4 357 nœuds ont été détectés comme des nœuds avec des connexions atypiques. Notons que 97% des anomalies identifiées ont reçu un score nul parmi lesquels il y a plusieurs nœuds qui sont isolés, c.-à-d., des nœuds qui ne se connectent à aucun nœud sur toutes les dimensions du réseau. Pour évaluer le résultat obtenu, nous montrons, d'abord, les matrices d'adjacence des sept dimensions du réseau (voir

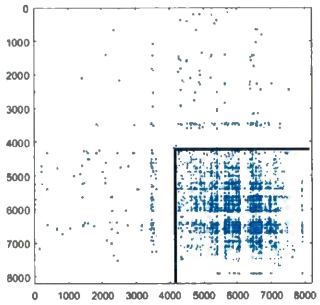
figure 3.5). Nous rappelons que ces matrices d'adjacence ont été construites selon un ordre ascendant des scores d'anomalies. De ce fait, nous nous attendons à ce que les anomalies se placent en premier et se suivent après par les nœuds normaux. Ici, il convient de noter que le réseau drosophila incorpore trois dimensions non pertinentes, à savoir, d_5, d_6 et d_7. Comme montré par les figures 3.5e, 3.5f et 3.5g, ces dimensions ne relatent aucune structure de nœuds significative. Sur le reste des dimensions qui sont des dimensions pertinentes (c.-à-d. d_1, d_2, d_3 et d_4), nous délimitons les nœuds normaux identifiés par un trait noir solide.

Sur la première dimension qui représente l'interaction directe entre les protéines (voir la figure 3.5a), nous constatons la présence de blocs qui paraissent denses en dehors de la zone délimitée des nœuds normaux identifiés. Lors de notre investigation de la structure topologique du réseau, nous avons découvert que ces blocs ne représentent pas concrètement des groupes denses de nœuds. Plus précisément, il y a des nœuds qui forment des chaines isolées composées de deux à quatre nœuds et d'autres nœuds qui se connectent sous forme d'étoile. Dans une telle situation, il est attendu que ces nœuds faiblement connectés reçoivent des scores $AS(u)$ infimes et se sélectionnent, par la suite, comme des anomalies comparativement aux autres nœuds densément connectés qui reçoivent des scores $AS(u)$ élevés.

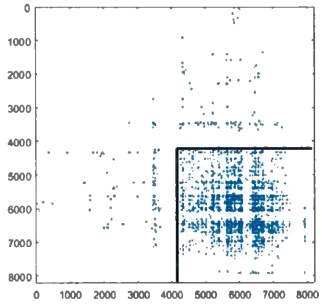
Nous rappelons que l'objet de ce mémoire est l'identification des nœuds qui sont aléatoirement connectés au reste des nœuds et qui n'appartiennent à aucune région dense à travers toutes les dimensions du réseau. De ce fait, dans le cas du réseau d'interaction de protéines Drosophila, nous ne fournissons pas une analyse détaillée de l'aspect biologique des nœuds anomalies identifiés. En effet, notre approche identifie les anomalies (les protéines) d'un point de vue topologique uniquement. D'autant plus, même si les nœuds identifiés sont considérés atypiques selon la définition d'anomalie adoptée dans ce mémoire, ils peuvent caractériser certains phénomènes biologiques qui peuvent être, éventuellement, évalués par des experts



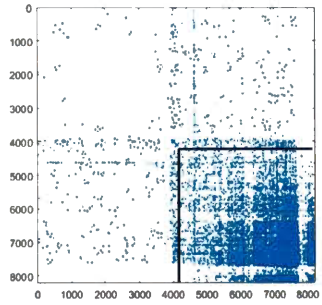
(a) d_1 : Interaction directe.



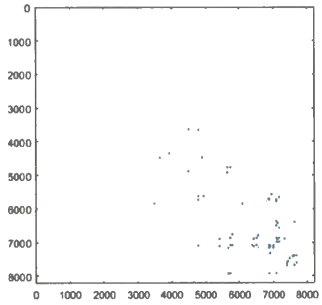
(b) d_2 : Interaction génétique suppressive.



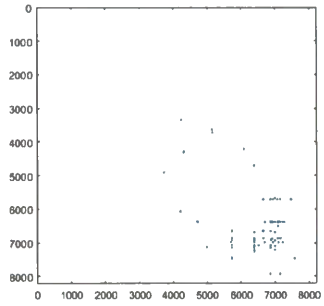
(c) d_3 : Interaction génétique additive.



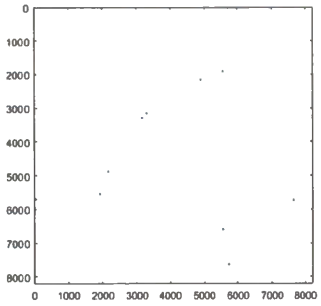
(d) d_4 : Association physique.



(e) d_5 : Colocalisation.



(f) d_6 : Association.



(g) d_7 : Interaction génétique synthétique.

Figure 3.5: Les matrices d'adjacence du réseau Drosophila.

du domaine en utilisant des connaissances biologiques existantes ou de nouvelles hypothèses.

3.2.2.4 Résultats du réseau de transport aérien européen

Pour le réseau de transport aérien européen, les 198 nœuds atypiques identifiés sont principalement des aéroports secondaires qui n'ont pas un trafic de vols entrants et sortants élevés comparativement aux aéroports principaux qui se caractérisent par plusieurs vols offerts par différentes compagnies aériennes. Les liens (vols) que détiennent les aéroports secondaires (identifiés comme des nœuds atypiques) sont des vols épars qui ne montrent aucun patron d'interaction compact avec les aéroports principaux qui se sont caractérisés par des scores élevés. Prenons l'exemple de quelques aéroports français modélisés dans l'ensemble de données testées. Les nœuds qui correspondent aux aéroports Clermont-Ferrand Auvergne (CFE), Limoges-Bellegarde (LIG) et Béziers-Cap d'Agde (BZR) ont été marqués comme anomalies. Selon les statistiques de l'année 2016 de l'union des aéroports français [1] qui ordonnent les aéroports selon leur trafic, les trois aéroports susmentionnés, à savoir, CFE, LIG et BZR, sont classés parmi les aéroports ayant un trafic très bas. Ces aéroports sont moins occupés et moins fréquentés comparativement aux aéroports classés au premier rang comme Paris Charles-de-Gaulle (CDG), Paris Orly (ORY) et Nice-Côte d'Azur (NCE). Les nœuds associés à ces aéroports sont identifiés comme des nœuds normaux par notre approche.

Les matrices d'adjacence du réseau de transport aérien européen n'ont pas été présentées dans ce mémoire en raison du nombre élevé de dimensions (37 dimensions). Toutefois, les résultats illustrés par les figures 3.2, 3.3, 3.4 et 3.5 sont représentatifs du comportement général de notre approche. Ces figures ont une caractéristique

1. voir `http://www.aeroport.fr/`

commune qui réside dans le fait que les nœuds atypiques ont des connexions aléatoires et éparses sur toutes les dimensions du réseau, contrairement aux nœuds normaux qui ont tendance à être fermement connectés dans un sous-ensemble ou sur toutes les dimensions du réseau, ce qui est le cas avec le réseau de transport.

3.3 Conclusion

Dans ce chapitre, nous avons évalué les performances de l'approche que nous proposons pour la détection des anomalies dans les réseaux multidimensionnels. L'ensemble des expérimentations sur les réseaux synthétiques suggèrent que notre approche identifie avec précision les anomalies même en présence d'un nombre élevé de dimensions non pertinentes. La détection des anomalies se fait automatiquement et aucune intervention humaine n'est requise. Ce fait appuie l'adaptabilité de notre approche au cas des données réelles où la catégorie des nœuds n'est pas connue au préalable.

CONCLUSION

Dans ce mémoire, nous avons proposé une approche de détection des anomalies dans les réseaux multidimensionnels. Dans notre conception, une anomalie est un nœud qui dispose de connexions éparses et qui n'appartient à aucune région dense à travers toutes les dimensions d'un réseau multidimensionnel. Pour identifier ce type d'anomalie, nous avons élaboré une approche en deux étapes.

En premier, nous avons examiné la structure topologique du graphe pour assigner un score d'anomalie à chaque nœud du réseau. Les scores d'anomalies sont calculés à partir d'une fonction, nouvellement suggérée, qui permet d'attribuer des scores faibles aux anomalies et des scores élevés aux nœuds normaux. Cette fonction a la particularité de considérer le réseau multidimensionnel en tant que tel sans avoir à faire une agrégation ou une analyse indépendante des dimensions.

Ensuite, pour identifier automatiquement les anomalies, nous avons utilisé le modèle de mélange beta. Ce modèle probabiliste permet de subdiviser les scores d'anomalies estimés en des sous-populations homogènes (composants). Le composant ayant les scores d'anomalies les plus faibles correspond au composant des nœuds atypiques. Ainsi, la détection des anomalies se fait d'une manière complètement systématique, sans intervention humaine pour fixer un seuil empirique de détection ou pour préciser un nombre k d'anomalies au préalable.

Les expérimentations sur des réseaux synthétiques suggèrent que notre approche identifie les anomalies avec précision et retourne des résultats consistants même en présence de dimensions non pertinentes. Tout de même, les résultats sur les données réelles suggèrent que l'approche proposée peut être un outil efficace et

pratique dans plusieurs applications réelles. À titre d'exemple, notre méthode peut être appliquée comme une étape de prétraitement de données qui précède l'application des algorithmes de partitionnement pour les réseaux multidimensionnels. De cette manière, il sera, désormais, possible d'éliminer les *outliers* qui entravent le processus de partitionnement.

Pour conclure, suite à la qualité des résultats obtenus, nous pensons que ce travail présente un moyen efficace qui peut être appliqué dans différents contextes pratiques. Dans la suite de nos recherches, nous explorerons différentes pistes pour étendre ce travail. Une des possibilités à envisager est l'identification simultanée des communautés et des *outliers* dans un sous espace de dimensions d'un réseau multidimensionnel. Une autre possibilité est l'étude du dynamisme du réseau, et ce en analysant le comportement des nœuds au cours du temps pour évaluer leur normalité. De plus à son applicabilité aux réseaux multidimensionnels, ce travail peut être adapté à d'autres types de réseaux tels que les réseaux dynamiques. Un réseau dynamique peut être pensé comme un réseau multidimensionnel tel que chaque dimension représente une vue sur les données à un temps t bien défini. De plus, nous pourrons, éventuellement, élargir ce travail par l'étude des anomalies dans les réseaux qui sont à la fois multidimensionnels et hétérogènes, c.-à-d., les réseaux qui incorporent plusieurs types de nœuds et de liens.

RÉFÉRENCES

Aggarwal, C. C. (2005). On Abnormality Detection in Spuriously Populated Data Streams. Dans *Proc. 5th International Conference on Data Mining*, 80–91. SIAM.

Aggarwal, C. C. (2017). Outlier Detection in Graphs and Networks. In *Outlier analysis* 369–397. Springer.

Akoglu, L., Chandy, R. et Faloutsos, C. (2013). Opinion Fraud Detection in Online Reviews by Network Effects. Dans *Proc. 7th International Conference on Weblogs and Social Media.*

Akoglu, L., McGlohon, M. et Faloutsos, C. (2010). Oddball : Spotting Anomalies in Weighted Graphs. *Advances in Knowledge Discovery and Data Mining*, 410–421.

Akoglu, L., Tong, H. et Koutra, D. (2015). Graph Based Anomaly Detection and Description : a Survey. *Data Mining and Knowledge Discovery*, 626–688.

Angiulli, F. et Pizzuti, C. (2002). Fast Outlier Detection in High Dimensional Spaces. Dans *Proc. European Conference on Principles of Data Mining and Knowledge Discovery*, 15–27. Springer.

Bain, L. J. et Engelhardt, M. (1987). *Introduction to Probability and Mathematical Statistics.* Brooks/Cole.

Battiston, F., Nicosia, V. et Latora, V. (2014). Structural Measures for Multiplex Networks. *Physical Review E.*

Bishop, C. M. (1994). Novelty Detection and Neural Network Validation. *IEE Proceedings-Vision, Image and Signal processing*, *141*(4), 217–222.

Boccaletti, S., Bianconi, G., Criado, R., Del Genio, C. I., Gómez-Gardenes, J., Romance, M., Sendina-Nadal, I., Wang, Z. et Zanin, M. (2014). The Structure and Dynamics of Multilayer Networks. *Physics Reports*, *544*(1), 1–122.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. et Hwang, D.-U. (2006). Complex Networks : Structure and Dynamics. *Physics reports*, *424*(4), 175–308.

Bouguessa, M., Missaoui, R. et Talbi, M. (2014). A Novel Approach for Detecting Community Structure in Networks. Dans *Tools with Artificial Intelligence (ICTAI), 2014 IEEE 26th International Conference on*, 469–477. IEEE.

Bouguila, N., Ziou, D. et Monga, E. (2006). Practical Bayesian Estimation of a Finite Beta Mixture Through Gibbs Sampling and its Applications. *Statistics and Computing*, *16*(2), 215–225.

Boutemedjet, S., Ziou, D. et Bouguila, N. (2010). Model-Based Subspace Clustering of Non-Gaussian Data. *Neurocomputing*, *73*(10), 1730–1739.

Breunig, M. M., Kriegel, H.-P., Ng, R. T. et Sander, J. (2000). Lof : Identifying Density-Based Local Outliers. Dans *ACM sigmod record*, volume 29, 93–104. ACM.

Cardillo, A., Gómez-Gardenes, J., Zanin, M., Romance, M., Papo, D., del Pozo, F. et Boccaletti, S. (2013). Emergence of Network Features from Multiplexity. *Scientific Reports*, *3*.

Celli, F., Di Lascio, F. M. L., Magnani, M., Pacelli, B. et Rossi, L. (2010). Social Network Data and Practices : The Case of Friendfeed. Dans *Proc. International Conference on Social Computing, Behavioral Modeling, and Prediction*, 346–353.

Chandola, V., Banerjee, A. et Kumar, V. (2009). Anomaly Detection : A Survey. *ACM Computing Surveys (CSUR)*, *41*(3).

Chandola, V., Banerjee, A. et Kumar, V. (2012). Anomaly Detection for Discrete Sequences : A Survey. *IEEE Transactions on Knowledge and Data Engineering*, *24*(5), 823–839.

Chouchane, A. et Bouguessa, M. (2017). Identifying Anomalous Nodes in Multidimensional Networks. *IEEE on Data Science and Advanced Analytics*.

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., Nicosia, V., Arenas, A. et Latora, V. (2015). Structural Reducibility of Multilayer Networks. *Nature Communications*, *6*.

Dempster, A. P., Laird, N. M. et Rubin, D. B. (1977). Maximum Likelihood From Incomplete Data Via The EM Algorithm. *Journal of the royal statistical society.*, 1–38.

Ding, Q., Katenka, N., Barford, P., Kolaczyk, E. et Crovella, M. (2012). Intrusion

as (anti) Social Communication : Characterization and Detection. Dans *Proc. 18th International Conference on Knowledge Discovery and Data Mining*, 886–894. ACM.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. *et al.* (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Dans *Kdd*, volume 96, 226–231.

Fathaliani, F. et Bouguessa, M. (2015). A Model-Based Approach for Identifying Spammers in Social Networks. Dans *Data Science and Advanced Analytics*, 1–9. IEEE.

Fawcett, T. et Provost, F. (1999). Activity Monitoring : Noticing Interesting Changes in Behavior. Dans *Proc. 5th International Conference on Knowledge Discovery and Data Mining*, 53–62. ACM.

Figueiredo, M. A. T. et Jain, A. K. (2002). Unsupervised Learning of Finite Mixture Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(3), 381–396.

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.

Hackett, A., Cellai, D., Gómez, S., Arenas, A. et Gleeson, J. P. (2016). Bond Percolation on Multiplex Networks. *Physical Review X*, *6*(2).

Hartigan, J. A. et Wong, M. A. (1979). Algorithm as 136 : A K-means Clustering Algorithm. *Journal of the Royal Statistical Society.*, *28*(1), 100–108.

Hawkins, D. M. (1980). *Identification of Outliers.*, volume 11. Springer.

Huang, J., Sun, H., Han, J., Deng, H., Sun, Y. et Liu, Y. (2010). Shrink : A Structural Clustering Algorithm for Detecting Hierarchical Communities in Networks. Dans *Proc. 19th ACM International Conference on Information and Knowledge Management*, 219–228.

Huang, J., Sun, H., Song, Q., Deng, H. et Han, J. (2013). Revealing Density-Based Clustering Structure from The Core-Connected Tree of a Network. *IEEE Transactions on Knowledge and Data Engineering*, *25*(8), 1876–1889.

Jennrich, R. I. et Robinson, S. M. (1969). A Newton-Raphson Algorithm for Maximum Likelihood Factor Analysis. *Psychometrika*, *34*(1), 111–123.

Ji, Y., Wu, C., Liu, P., Wang, J. et Coombes, K. R. (2005). Applications of Beta-Mixture Models in Bioinformatics. *Bioinformatics*, *21*(9), 2118–2122.

Kriegel, H.-P., Kröger, P., Schubert, E. et Zimek, A. (2009). Loop : Local Outlier Probabilities. Dans *Proceedings of the 18th ACM conference on Information and knowledge management*, 1649–1652. ACM.

Krügel, C., Toth, T. et Kirda, E. (2002). Service Specific Anomaly Detection for Network Intrusion Detection. Dans *Proceedings of the 2002 ACM symposium on Applied computing*, 201–208. ACM.

Li, X. et Han, J. (2007). Mining Approximate Top-k Subspace Anomalies in Multi-Dimensional Time-Series Data. Dans *Proc. 33rd international conference on Very large data bases*, 447–458. VLDB Endowment.

Ma, Z. et Leijon, A. (2009). Beta Mixture Models and the Application to Image Classification. Dans *16th IEEE International Conference on Image Processing*, 2045–2048.

Matthews, B. W. (1975). Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, *405*(2), 442–451.

Papalexakis, E. E., Akoglu, L. et Ience, D. (2013). Do More Views of a Graph Help ? Community Detection and Clustering in Multi-Graphs. Dans *Proc. 16th IEEE International Conference on Information Fusion*, 899–905.

Ramaswamy, S., Rastogi, R. et Shim, K. (2000). Efficient Algorithms for Mining Outliers from Large Data Sets. Dans *ACM Sigmod Record*, volume 29, 427–438. ACM.

Savage, D., Zhang, X., Yu, X., Chou, P. et Wang, Q. (2014). Anomaly Detection in Online Social Networks. *Social Networks*, *39*, 62–70.

Schwarz, G. *et al.* (1978). Estimating The Dimension of a Model. *The annals of statistics*, *6*(2), 461–464.

Smyth, P. (2000). Model Selection for Probabilistic Clustering Using Cross-Validated Likelihood. *Statistics and Computing*, *10*(1), 63–72.

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.

Tan, P.-N., Steinbach, M. et Kumar, V. (2006). *Introduction to Data Mining*. Pearson Education India.

Tang, J., Chen, Z., Fu, A. et Cheung, D. (2002). Enhancing Effectiveness of Outlier Detections for Low Density Patterns. *Advances in Knowledge Discovery and*

Data Mining, 535–548.

Xu, X., Yuruk, N., Feng, Z. et Schweiger, T. A. (2007). Scan : A Structural Clustering Algorithm for Networks. Dans *Proc. 13th ACM International Conference on Knowledge Discovery and Data Mining*, 824–833.

Yamanishi, K., Takeuchi, J.-I., Williams, G. et Milne, P. (2004). On-line Unsupervised Outlier Detection Using Finite Mixtures with Discounting Learning Algorithms. *Data Mining and Knowledge Discovery*, *8*(3), 275–300.

Yeung, D.-Y. et Chow, C. (2002). Parzen-Window Network Intrusion Detectors. Dans *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 4, 385–388. IEEE.

Zimek, A., Schubert, E. et Kriegel, H.-P. (2012). A Survey on Unsupervised Outlier Detection in High-Dimensional Numerical Data. *Statistical Analysis and Data Mining : The ASA Data Science Journal*, *5*(5), 363–387.