Contribution à la détection de communautés chevauchantes pour l'analyse des réseaux transactionnels complexes

Safa El Ayeb

Normandie Université

UNIVERSITÉ CAEN NORMANDIE

THÈSE

Pour obtenir le diplôme de doctorat

Spécialité **INFORMATIQUE**

Préparée au sein de l'**Université de Caen Normandie**

Contribution à la détection de communautés chevauchantes pour l'analyse des réseaux transactionnels complexes

Présentée et soutenue par

SAFA EL AYEB

Thèse soutenue le 23/05/2023

devant le jury composé de :

M. HAMAMACHE KHEDDOUCI	Professeur des universités - UNIVERSITE LYON 1 CLAUDE BERNARD	Rapporteur du jury
MME PASCALE KUNTZ	Professeur des universités - Polytechnique Nantes	Rapporteur du jury
MME CÉCILE BOTHOREL	Maître de conférences - Ecole nationale supérieure Mines-Télcom Atlantique Bretagne (IMT)	Membre du jury
M. LUC BRUN	Professeur des universités - ENSICAEN	Membre du jury
M. BAPTISTE HEMERY	Docteur - ORANGE LABS CAEN	Membre du jury
M. JEAN-LOUP GUILLAUME	Professeur des universités - UNIVERSITE LA ROCHELLE	Président du jury
MME ESTELLE CHERRIER	Maître de conférences - ENSICAEN	Directeur de thèse
M. CHRISTOPHE CHARRIER	Maître de conférences - Université de Caen Normandie	Co-directeur de thèse

Thèse dirigée par **ESTELLE CHERRIER** (GREYC ALGORITHMIQUE) et **CHRISTOPHE CHARRIER** (GREYC ALGORITHMIQUE)

MIIS

GREYC
Laboratoire de recherche en sciences du numérique

REMERCIEMENTS

J'aimerais profiter de ces quelques lignes pour exprimer ma profonde gratitude envers toutes les personnes qui ont contribué à la réalisation de cette thèse. Le chemin parcouru jusqu'à ce stade n'aurait pas été possible sans le soutien, l'encouragement et l'apport précieux de nombreuses personnes exceptionnelles.

Je tiens à exprimer ma reconnaissance envers mes directeurs et encadrants de thèse Baptiste, Estelle, Christophe et Fabrice, pour leur confiance en moi et pour leur présence tout au long de cette aventure. Votre soutien constant a été une véritable source de motivation, m'aidant à rester déterminée même face aux moments les plus difficiles. Vos conseils avisés, vos disponibilités et votre expertise m'ont permis d'évoluer et de repousser mes limites. Je suis reconnaissante d'avoir eu la chance de travailler avec vous, et de tous les souvenirs précieux que nous avons partagés.

Je souhaite remercier Orange pour avoir financé ces trois années de thèse. Mon expérience à Orange était l'opportunité de faire de très belles rencontres au sein de différentes équipes. Je tiens également à exprimer ma gratitude envers les personnes avec qui je n'ai pas directement collaboré, mais avec qui j'ai partagé d'innombrables déjeuners, pauses café et discussions chaleureuses.

Je remercie également mes collègues du laboratoire GREYC, les membres de l'équipe SAFE, les services administratifs et ceux de l'école doctorale, ainsi que les docto-

rants avec qui j'ai partagé mon bureau, je suis reconnaissante d'avoir eu l'opportunité de travailler et d'apprendre à vos côtés, et de toutes les informations que vous m'avez généreusement fournies, en particulier pour la préparation de ma soutenance.

Je remercie les membres de mon comité de suivi individuel de thèse, Bruno Cremilleux et Mohammed Haddad pour leur conseils éclairés.

Je remercie les membres de mon jury, Hamamache Kheddouci et Pascale Kuntz d'avoir accepté de rapporter minutieusement cette thèse, ainsi que Cécile Bothorel, Luc Brun, et Jean-Loup Guillaume d'avoir pris le temps de l'examiner.

Je tiens à exprimer ma reconnaissance inégale envers mes parents, *Lotfi et Noura*, les mots n'exprimeront pas ma gratitude pour votre présence et votre guidance durant tout le chemin qui a mené à cette thèse. La confiance que vous avez placée en moi est la fondation sur laquelle j'ai pu construire mes accomplissements d'aujourd'hui. Merci pour tout ce que vous avez apporté à ma vie. Je remercie mon mari, *Achref*, pour son soutien inconditionnel et sa contribution indéniable pour la réalisation de cette thèse. Ta confiance et tes encouragements ont été les piliers sur lesquels je me suis appuyé tout au long de ce parcours et m'ont donné la force de continuer là où j'en avais le plus besoin. Ta présence est une source de joie et de soutien qui m'inspire chaque jour. Enfin, je remercie mes frères, *Mohamed et Malek*, mes beaux-parents, *Mansour et Moufida*, mes cousines, et tous mes amis pour leur support constant et leur croyance en mes capacités. Je suis reconnaissante de vous avoir à mes côtés.

Je remercie toutes les personnes qui ont contribué de près ou de loin à la réalisation de cette thèse. Votre collaboration a été d'une valeur inestimable.

Merci du fond du cœur.

RÉSUMÉ

L'analyse des réseaux sociaux est fondée sur l'étude des interactions sociales pour la compréhension des comportements individuels et collectifs au sein des systèmes complexes. Les réseaux sociaux peuvent être représentés sous forme de graphes, qui sont des structures de données mathématiques, pour les modéliser et étudier leurs propriétés. Une des nombreuses problématiques liées à l'analyse des réseaux sociaux concerne la détection de communautés qui vise à identifier des groupes fortement connectés. Cette thèse est motivée par l'étude de la détection des communautés sur des données de transactions issues du service financier Orange Money. Ces transactions sont modélisées par un multigraphe où les nœuds représentent les utilisateurs du service, et les liens représentent leurs échanges. Au cours de cette thèse, on s'intéresse à la détection de communautés chevauchantes. Ce type de communautés reflète bien la réalité en associant chaque individu à plusieurs communautés à la fois. Différents algorithmes ont été testés à cet effet. Compte tenu de la nature sensible des données, nos tests ont été effectués sur des données synthétiques inspirées des utilisations réelles du service financier. Les graphes ainsi étudiés sont très volumineux et peuvent comporter plusieurs millions de nœuds et d'arêtes. C'est dans ce sens que la première contribution de la thèse porte sur la comparaison de différentes méthodes de réduction d'un multigraphe. L'objectif de ces méthodes est de transformer le multigraphe en graphe simple pondéré afin de faciliter les phases de manipulation, de stockage, et d'analyse. Une étude comparative a permis de déterminer si la réduction du multigraphe impacte la qualité des communautés ob-

tenues. Dans la deuxième contribution, nous proposons quatre nouvelles métriques d'évaluation extrinsèques pour la détection des communautés chevauchantes. Ces métriques sont comparées aux métriques de l'état de l'art afin d'estimer leur efficacité.

ABSTRACT

Social network analysis is based on the study of social interactions to understand individual and collective behaviors in complex systems. Social networks can be represented as graphs, which are mathematical data structures, to model them and study their properties. One of the many problems related to the analysis of social networks concerns the detection of communities, which aims at identifying strongly related groups. This thesis is motivated by the study of community detection on transaction data from the Orange Money financial service. These transactions are modeled by a multigraph where the nodes represent the users of the service, and the links represent their exchanges. In this thesis, we are interested in the detection of overlapping communities. This type of community reflects reality by associating each individual to several communities simultaneously. Different algorithms have been tested for this purpose. Given the sensitive nature of the data, our tests were performed on synthetic data inspired by the real uses of the financial service. The graphs studied are very large and can have several million nodes and edges. It is in this sense that the first contribution of the thesis concerns the comparison of different methods for reducing a multigraph. The objective of these methods is to transform the multigraph into a simple weighted graph in order to facilitate the manipulation, storage and analysis phases. A comparative study has allowed us to determine if the reduction of the multigraph impacts the quality of the obtained communities. In the second contribution, we propose four new extrinsic evaluation metrics for the detection of overlapping communities. These metrics are compared

to the state-of-the-art metrics in order to estimate their effectiveness.

Table des matières

Liste des tableaux

Liste des figures

Liste d'abréviations

Introduction

Les réseaux sociaux sont omniprésents dans les systèmes qui nous entourent et se manifestent sous différentes formes dans notre quotidien, à travers les médias sociaux, les moteurs de recherche et la plupart des nouvelles technologies. Les réseaux sociaux sont des structures complexes de relations sociales entre des individus, des groupes ou des organisations. Au cours des dernières années, l'analyse des réseaux sociaux a suscité un intérêt considérable en raison de son potentiel à traiter de nombreuses études de cas du monde réel. Cette approche vise à explorer à la fois les structures sociales de base et l'influence des éléments caractérisant les différentes connexions. Ainsi, l'analyse des réseaux sociaux permet de mettre en évidence ces différentes formes de relations et de les étudier en profondeur pour mieux comprendre les dynamiques sociales à l'œuvre au sein de ces réseaux.

Les réseaux sociaux peuvent prendre différentes formes, telles que les réseaux professionnels, les réseaux familiaux, les réseaux d'amis, les réseaux de connaissances, etc. Au sein du réseau, la notion de relation peut prendre plusieurs formes, et peut être variée et complexe, allant de relations personnelles telles que l'amitié et la confiance, à des relations plus impersonnelles telles que l'intérêt commun ou la similarité d'opinions. Ces relations peuvent varier en termes de force et de durée. Ces réseaux permettent également de comprendre comment l'information et les ressources sont partagées, comment les décisions sont prises et comment les individus et les groupes interagissent les uns avec les autres dans des applications telles que l'analyse de l'opinion publique, la recommandation de contenu, la détection de

fraude, etc. Les réseaux de la vie réelle sont généralement complexes de par leur nature, en raison de la multitude de connexions et de relations qui s'y établissent entre les différents acteurs. Parmi les caractéristiques principales des réseaux complexes, on peut citer leur grande taille, leur connectivité élevée, leur hétérogénéité et leur capacité à présenter des propriétés émergentes qui ne peuvent être déduites uniquement de l'analyse de leurs éléments individuels. C'est pourquoi les réseaux constituent un outil exceptionnel pour l'analyse des systèmes complexes d'objets en interaction et la compréhension des mécanismes sous-jacents lors de l'étude de divers phénomènes, tels que la diffusion d'informations, la propagation des maladies, la coopération et le partage de ressources, la formation de communautés et la prise de décisions collectives, etc.

Objectifs de la thèse

Dans le cadre de cette thèse CIFRE proposée par *Orange* en collaboration avec l'équipe *SAFE* (Security, Architecture, Forensics, biomEtrics) du laboratoire *GREYC* (Groupe de Recherche en Informatique, Image, Automatique et Instrumentation), nous proposons d'analyser le réseau des transactions du service de paiement mobile « Orange Money », en utilisant des techniques d'analyse des réseaux sociaux, notamment la détection des communautés chevauchantes. L'analyse des réseaux sociaux a déjà été utilisée avec succès sur des données bancaires et mobiles, dans des études socio-économiques afin de découvrir les habitudes d'achat des clients, détecter la fraude, prédire l'évolution de l'exploitation du service par les clients (Centellegher et al., 2018), etc. La nature des transactions financières exige un certain niveau de confiance entre les parties impliquées. Ainsi, avant même d'envoyer une transaction, il existe un lien social, sous une forme ou une autre, entre les individus. Cette thèse vise à exploiter les données de transaction des services de paiement mobile d'Orange Money pour retracer et identifier ces liens sociaux préexistants.

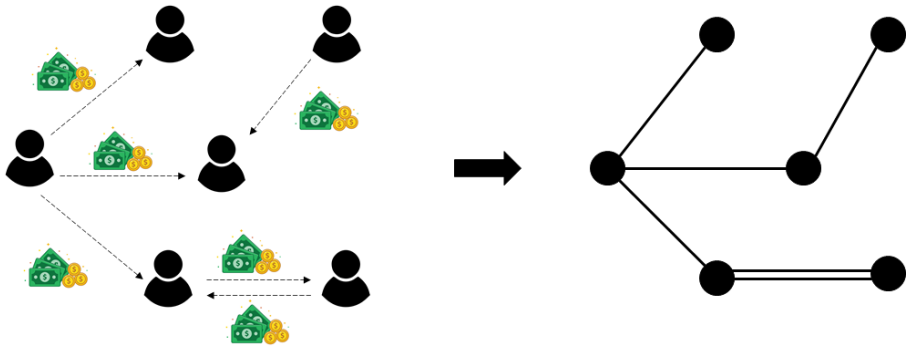


FIGURE 1 – Illustration de la transformation d'un réseau social de transactions en graphe. Sur ce graphe, les nœuds représentent les utilisateurs et les arcs représentent les échanges financiers.

Bien que variées, les principales applications des réseaux sociaux reposent sur l'identification des nœuds les plus importants dans un réseau, la compréhension de la structure et la dynamique du réseau, la prédiction de liens, etc. Dans ce travail, nous nous intéressons particulièrement à la détection de communautés. Les réseaux réels se décomposent en modules densément connectés, appelés communautés, ayant des liens peu nombreux entre eux. Les communautés correspondent à des entités comportementales ou fonctionnelles au sein du réseau. Elles présentent des groupes d'individus partageant des propriétés communes ou jouant des rôles similaires dans le réseau (Fortunato, 2010). L'importance de l'étude des communautés réside dans leur présence dans différents types d'organisations sociales, tels que les cercles familiaux, les groupes de travail et d'amitié, les villes et les villages, ainsi que dans les groupes virtuels tels que les communautés en ligne. C'est pourquoi la détection des communautés est devenue un domaine de recherche fondamental et très pertinent dans le domaine de l'analyse des réseaux sociaux.

Les communautés peuvent avoir des applications concrètes dans des disciplines variées telles que l'identification des clients ayant des intérêts similaires dans des réseaux des relations d'achat par exemple, le repérage des communautés d'utilisateurs de médias sociaux tels que *Twitter* basées sur des groupes d'âge, de localisation géo-

graphique, etc., l'analyse des communautés dans les réseaux professionnels tels que *LinkedIn* pour comprendre les relations entre les membres et les opportunités de carrière, l'étude des communautés dans les réseaux de collaborations scientifiques pour identifier les partenaires de recherche potentiels et les tendances dans le domaine, etc. Ces exemples ne représentent qu'une infime partie des nombreuses applications de l'analyse des réseaux sociaux, qui continuent à être explorées et développées dans de nombreux domaines de recherche. Dans le cadre de la détection de communautés, les résultats peuvent prendre diverses formes, notamment des communautés disjointes, qui ne partagent aucun nœud, des communautés chevauchantes, qui partagent certains nœuds, des communautés floues ou encore des communautés dynamiques. Notre étude se concentre principalement sur la recherche de communautés chevauchantes, car celles-ci reflètent mieux la réalité où un individu peut appartenir simultanément à plusieurs communautés qui se chevauchent.

Les communautés chevauchantes permettent de capturer cette complexité des liens entre les individus et de mieux modéliser la structure du réseau de transactions. En utilisant des techniques de détection de communautés chevauchantes, il est possible d'identifier des groupes d'individus qui sont liés par des transactions financières, même s'ils appartiennent à plusieurs communautés concurremment. La découverte de communautés au sein de ce réseau de transactions permettra d'avoir une vision plus complète de sa structure et de mieux comprendre les interactions entre les individus. En combinant la détection de communautés et l'utilisation de métriques pertinentes et efficaces nous pourrions avoir une bonne compréhension des utilisations du service Orange Money à travers différentes applications telles que le suivi des flux monétaires, l'identification des transactions suspectes entre des communautés éloignées pour détecter la fraude, la détermination des ponts qui jouent le rôle d'intermédiaire entre les communautés et qui sont importants pour la propagation de nouveaux services ou même pour l'adoption du service lui-même, etc.

L'un des plus grands défis liés à la détection de communautés dans les réseaux

sociaux est la capacité d'évaluer les résultats générés. L'évaluation est un véritable problème pour les réseaux réels, qui ne fournissent que peu de données. Les mesures d'évaluation dans ce domaine peuvent être utilisées soit pour évaluer les performances d'un algorithme de détection de communautés, soit pour comparer les différents algorithmes appliqués aux même réseaux. Pour évaluer et comparer les algorithmes de détection de communautés, la littérature a accordé beaucoup d'attention aux mesures d'évaluation qui permettent de déterminer la qualité de la partition sur la base de mesures topologiques, ou de la comparer à une référence appelée « vérité terrain ».

C'est en ce sens que notre étude a donné lieu aux contributions principales suivantes :

i L'étude de l'application de détection de communautés chevauchantes sur le graphe transactionnel d'Orange Money.

ii La proposition de nouvelles métriques d'évaluation des résultats de détection de communautés chevauchantes permettant de pallier les défauts des métriques standards de l'état de l'art.

iii L'exploration de l'effet de la réduction d'un multigraphe en se basant sur différentes formules de calcul de poids sur le résultat de la détection de communautés.

Organisation du manuscrit

Le manuscrit est organisé en six chapitres en vue d'introduire le contexte, l'état de l'art et les contributions de la thèse. Il est articulé de la manière suivante :

— Le **Chapitre 1** positionne le contexte de la thèse, présente les services de paiement mobile, et en particulier le service Orange Money. Il introduit également les réseaux sociaux et leurs applications. Les données Orange Money sont présentées, ainsi que les motivations qui sous-tendent l'utilisation des

réseaux sociaux pour analyser ces données.

- Le **Chapitre 2** pose les concepts de base liés aux réseaux et aux graphes. Des définitions de base seront présentées. Les données synthétiques exploitées ainsi que leur pertinence dans le cadre de notre thèse y sont présentées.

— Le **Chapitre 3** étudie le concept de détection de communautés. Il présente les algorithmes les plus utilisés dans l'état de l'art et justifie pourquoi les communautés chevauchantes sont les plus adéquates pour nos travaux. Finalement, une comparaison des différents algorithmes de détection de communautés chevauchantes est effectuée.

— Le **Chapitre 4** présente les métriques d'évaluation de l'état de l'art dans le contexte de la détection de communautés chevauchantes. Les nouvelles métriques d'évaluation extrinsèques proposées pour évaluer les résultats de la détection de communautés chevauchantes sont présentées dans ce chapitre.

— Le **Chapitre 5** aborde la problématique de la réduction du multigraphe et propose différentes méthodes proposées pour résoudre ce problème. Une analyse de l'effet de la réduction sur la détection de communautés chevauchantes est réalisée.

— Le **Chapitre 6** conclut ce manuscrit et donne plusieurs perspectives à cette thèse. Nous proposons des pistes pour des recherches futures.

Chapitre 1

Contexte

Les travaux de la thèse se placent dans le contexte d'analyse des données Orange Money qui offre des services bancaires sur mobile. Nous présenterons dans ce chapitre les services de paiement mobile en général, et le service Orange Money en particulier. Nous introduirons également les données recueillies par ce service et étudiées dans cette thèse.

Sommaire

1.1 Introduction

Le progrès constant des technologies a favorisé la création d'outils et des mécanismes innovants qui ont eu un impact significatif sur de nombreux domaines, modifiant leurs processus de travail et renforçant leurs activités. Dans le domaine financier,

ceci a suscité la création de nouvelles solutions de paiement, capables de changer les comportements des utilisateurs. Ainsi, les cartes bancaires et les paiements numériques ont offert des alternatives aux moyens de paiement conventionnels, tels que les chèques et les espèces. Cette révolution a également permis à des institutions non bancaires telles que PayPal, Amazon Pay, Google Pay, etc. de se joindre à l'offre des moyens de paiement. Des entreprises de télécommunications ont également rejoint le marché des services financiers. En effet, le raccordement des télécommunications et du domaine bancaire a donné lieu à trois nouvelles offres : le mobile bancaire, le paiement mobile, et le commerce mobile (Chaix, 2013). Le mobile bancaire regroupe les services fournis par des institutions financières et accessibles par téléphone tels que les virements bancaires, les suivis et gestions des comptes personnels, etc. Le paiement mobile correspond aux paiements réalisés à partir d'un téléphone mobile. Il réfère à tout paiement dont l'initiation, l'activation ou la confirmation est effectuée à travers un téléphone mobile. Le commerce mobile est une extension du commerce électronique impliquant l'utilisation d'un téléphone mobile et effectué sur un réseau de télécommunications. Ceci implique la réservation d'un ticket de cinéma ou de musée, l'accès à des hôtels, les stationnements, etc. Ces services permettent aux utilisateurs de gérer leur argent avec leur téléphone portable à travers des technologies simples telles que les USSD (Unstructured Supplementary Service Data), les SMS (Short Message Service), les applications mobiles, etc.

Dans le cadre de cette thèse, nous nous intéressons particulièrement à l'étude des données fournies par le service Orange Money : un service de paiement mobile proposé par l'entreprise Orange en Afrique. Les données fournies par ce service peuvent être transformées sous la forme d'un réseau social et analysées par le biais des techniques adéquates.

Dans ce premier chapitre, nous allons présenter l'évolution des services de paiement mobile en Afrique dans un premier temps, et du service Orange Money dans un deuxième temps. Nous allons par la suite nous pencher sur l'analyse des réseaux

sociaux, et expliquer les motivations derrière le choix de cette approche pour étudier les données Orange Money. Finalement, nous allons introduire les données Orange Money étudiées, et expliquer leur transformation en réseau.

1.2 Le paiement mobile

Le paiement mobile est l'exemple d'une innovation qui a particulièrement bien réussi sur le continent africain, non pas en raison de l'originalité de la technologie proposée, mais surtout parce que le service répondait à un vrai problème au niveau local, celui de la sous-bancarisation (Rennard, 2020) parallalèment à la très large diffusion des mobiles au sein de la population.

1.2.1 L'essor du paiement mobile en Afrique

Le paiement mobile ou le m-paiement se rapporte à toute transaction ayant une valeur monétaire entre deux parties, faite par le moyen d'un téléphone portable ou une autre forme de technologie mobile capable de traiter en toute sécurité une opération financière sur un réseau sans fil (Ondrus and Pigneur, 2005). Les transactions peuvent inclure des transferts d'argent entre personnes, des paiements de factures, des achats en ligne, et plus encore.

Selon Diniz *et al.* (Diniz et al., 2011), le raccordement du marché des télécommunications avec le marché bancaire avait deux grandes motivations : d'un côté, favoriser l'inclusion financière pour des populations peu bancarisées (afin de devenir capable d'épargner, d'investir, etc.), et d'un autre côté diversifier les revenues des opérateurs téléphoniques en proposant des offres variées à une proportion plus large de la population tout en limitant les investissements en infrastructure (Chaix and Torre, 2015).

En effet, de nombreuses personnes en Afrique, en particulier dans les zones rurales, n'ont pas accès à un compte bancaire traditionnel. Aujourd'hui encore, l'accès à

un compte bancaire reste difficile pour une grande partie de la population des pays africains. Les facteurs contribuant à limiter le taux de bancarisation de ces régions sont divers : les conditions strictes d'ouverture de compte, les frais récurrents élevés, la rareté des agences bancaires, etc.

Ainsi, les solutions de paiement mobile répondent à un besoin : pallier le manque de canaux formels de transactions financières et offrir des services financiers basés sur un compte prépayé en utilisant exclusivement des espèces. Ils ont permis aux personnes n'ayant pas accès aux comptes bancaires conventionnels de participer à l'économie, et d'accéder à des opérations simples, telles que l'envoi ou la réception de l'argent d'une région à une autre, l'achat sur des sites de commerce électronique, le paiement des factures, etc.

Les services de m-paiement sont également utiles pour les personnes qui ont accès à un compte bancaire, mais qui préfèrent utiliser des méthodes de paiement plus pratiques et rapides. Ainsi, grâce à la dématérialisation de la monnaie et des transactions, les services de paiement mobile ont fondamentalement transformé le secteur des services financiers.

Par conséquent, dans les pays à faible niveau de bancarisation, le paiement mobile est non seulement sans concurrent, mais il devient un vecteur essentiel de la bancarisation. La figure 1.1 en est un bon exemple où on voit l'évolution du nombre de comptes bancaires par rapport au nombre de comptes de paiement mobile au Kenya entre les années 2007 et 2016. Sur cette figure, le nombre de comptes de monnaie mobile dépasse le nombre de comptes bancaires classiques entre les années 2010 et 2015, ce qui reflète le succès de ce service.

Les services de paiement mobile en Afrique ont commencé à se développer dans les années 2000 avec l'adoption de la technologie de téléphonie mobile. Ils ont rapidement gagné en popularité en raison de leur facilité d'utilisation, nécessitant seulement l'acquisition d'un téléphone mobile et étant basés sur des outils auxquels

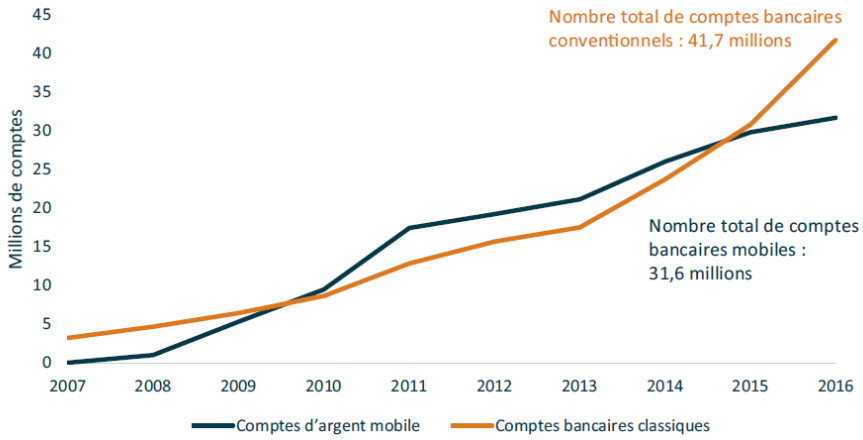


FIGURE 1.1 – Nombre total de comptes bancaires conventionnels et de comptes bancaires par téléphonie mobile au Kenya 2006-2015.

les gens sont déjà accoutumés tels que les SMS, ou les codes USSD.

Aujourd'hui, 70% du marché global du paiement mobile se passe en Afrique (184 millions d'utilisateurs actifs) (GSMA, 2021). En 2007, le succès du mobile bancaire a été immédiat. Près d'un quart de la population utilisait le service en 2012. Selon le Groupe de la Banque Africaine de Développement, environ 41% de la population africaine utilise actuellement des services de paiement mobile. La figure 1.2 permet de visualiser l'évolution du pourcentage du nombre d'adultes possédant un compte de paiement mobile en Afrique subsaharienne entre les années 2014 et 2017.

Les avantages des services de paiement mobile sont nombreux. Ils permettent aux utilisateurs de faire des transactions financières de manière simple et rapide, sans avoir à se rendre dans un guichet bancaire ou un point de vente. Ils offrent également une certaine sécurité, car les transactions sont effectuées en ligne et protégées par des mesures de sécurité telles que des codes PIN et des authentifications à deux facteurs. Cependant, il y a également quelques inconvénients : ils peuvent être coûteux pour les utilisateurs qui doivent payer des frais pour chaque transaction, ils peuvent également être limités par la couverture réseau ce qui signifie que les utilisateurs peuvent ne pas être en mesure de les utiliser dans certaines régions ou dans des

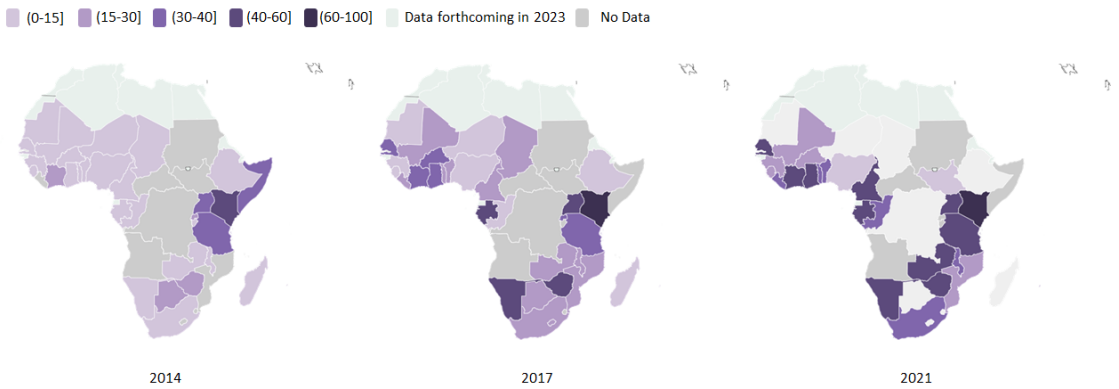


FIGURE 1.2 – Comptes de paiement mobile en Afrique subsaharienne sur la période 2014-2017 (Demirguc-Kunt et al., 2018).

zones rurales.

Avec la crise sanitaire, les banques centrales, les gouvernements et les autorités sanitaires ont largement incité la population à se détourner de l'utilisation de l'argent liquide et à se tourner massivement vers des solutions de paiement mobile jugées plus sécurisées. En 2020, la valeur des transactions du paiement mobile a atteint plus de deux milliards de dollars par jour, et devrait dépasser les trois milliards de dollars par jour d'ici la fin 2022 (GSMA, 2021). Aujourd'hui, il y a plusieurs fournisseurs de services de paiement mobile en Afrique : M-Pesa depuis 2007, Orange Money depuis 2008, MTN Mobile depuis 2010, Airtel Money depuis 2010, Ecocach depuis 2011, Wave Mobile Money depuis 2019, et encore plein d'autres. Le travail de cette thèse est fondé sur l'analyse des données relatives au service Orange Money.

1.2.2 Le service Orange Money

Orange a commencé à offrir ses services de paiement mobile « Orange Money » en Côte d'Ivoire en 2008. Depuis, le service s'est beaucoup développé et est devenu porteur de confiance. Orange Money connaît aujourd'hui une expansion sur 17 pays africains comme le montre la carte de la figure 1.3. Le service compte plus de soixante-dix millions de comptes. Actuellement, le flux hebdomadaire des

transactions représente un montant supérieur à un milliard de dollars.

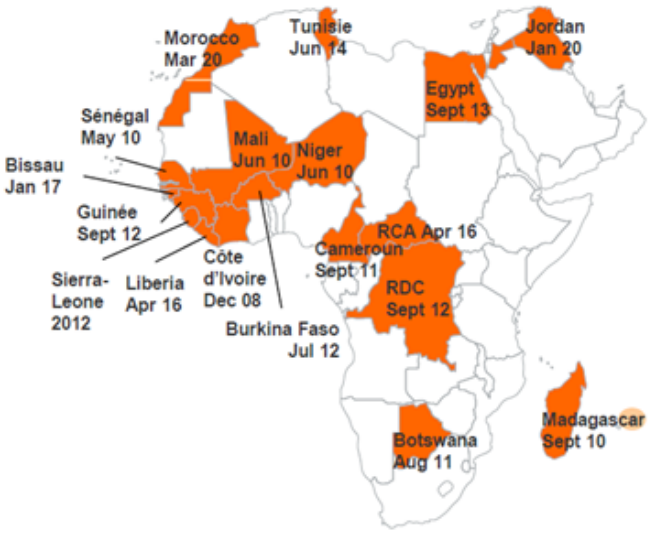


FIGURE 1.3 – Carte des pays intégrant le service Orange Money.

Comme tout service de paiement mobile, les transactions Orange Money forment un réseau de clients qui s'envoient de l'argent pour différents motifs. Concrètement, le système d'Orange Money s'est élaboré autour d'un réseau de distributeurs (plus de 500 000 sur le continent africain) répartis sur l'ensemble du territoire. Ces distributeurs ou points de vente rechargent sur le téléphone un solde Orange Money grâce à la somme en espèces que lui remettent les clients ou inversement. Ce solde peut, par la suite, être utilisé à travers les différentes prestations qu'offre Orange Money.

À l'origine, la proposition des services financiers d'Orange Money était simple : un encaissement en espèces (ou *cash-in*) pour alimenter le solde client, des retraits (ou *cash-out*) pour le retrait d'argent liquide en échange d'une diminution du solde client, des transferts pour le déplacement d'argent d'un compte à un autre, des paiements marchands utilisés pour acquérir des biens ou des services, et des achats de minutes de communication. Ces services étaient accessibles depuis des téléphones basiques en USSD, sans même avoir besoin de connexion internet. Aujourd'hui,

l'offre des services s'est largement enrichie avec des services proches des services bancaires tels que le crédit, l'épargne, le *bank to wallet*, le paiement web, les transferts internationaux, le paiement de factures, etc. Orange Money intervient actuellement dans la collecte de taxes, le versement de bourses aux étudiants, le paiement de l'électricité, etc.

Comme on peut le constater, les clients du service Orange Money sont des personnes qui ont une relation sociale préexistante et se connaissent généralement dans la vie réelle. Ce type de système de transactions financières repose sur la confiance et les liens personnels entre les parties concernées. L'utilisation des paiements mobiles permet à ces personnes de transférer des fonds facilement et en toute sécurité, sans avoir à passer par une institution bancaire traditionnelle. Les clients du service peuvent être des familles qui s'envoient de l'argent entre elles, un employeur qui paye le salaire de ses salariés, des amis qui collectent des fonds pour un événement, ou un marchand qui se fait payer par le paiement mobile. Il est donc légitime de présumer l'existence d'un lien social établi entre les clients du service Orange Money.

Notre objectif à travers cette thèse est de vérifier si, en étudiant uniquement les données de transactions, on arrive à retrouver ce lien social préexistant. Pour cette raison, nous proposons dans nos travaux une analyse des données de transactions Orange Money en les représentant sous forme de réseau social. Dans la section suivante, nous allons définir la notion de réseau social, et expliquer les raisons derrière le choix de cette structure pour représenter les données Orange Money.

1.3 L'analyse par réseaux sociaux

Comme abordé précédemment, les travaux de cette thèse ont pour but d'étudier et d'analyser les données de transactions d'Orange Money, en s'appuyant sur une approche d'analyse de réseaux sociaux. L'analyse des réseaux sociaux est une approche unique ayant pour objectif principal d'explorer les structures sociales en mettant

en avant les liens et les relations sociales au-delà des caractéristiques individuelles de ces entités. Elle représente à la fois un canal pour la définition des concepts sociaux importants, une alternative théorique à l'hypothèse d'indépendance des acteurs, ainsi qu'un cadre pour tester les théories sur les relations sociales structurées (Wasserman and Faust, 1994). Ainsi, l'analyse des réseaux sociaux vise à expliquer les phénomènes collectifs sociaux émergents, à comprendre les propriétés sociales à travers les interactions, et à déterminer comment celles-ci influencent les caractéristiques collectives et individuelles observées.

Au sens des sciences humaines, **un réseau social** est formé par un ensemble d'acteurs sociaux tels que des personnes, des groupes, ou des organisations et par leurs relations mutuelles, représentés sous forme de sommets et de liens. Ces relations peuvent se nouer dans divers cercles sociaux tels que la famille, le voisinage, l'affiliation au même groupe de loisirs, la sphère amicale ou encore le milieu professionnel. Elles peuvent également présenter des canaux de transfert d'informations, de biens, d'argent, etc. L'étude de ces relations représente l'essence de l'analyse des réseaux sociaux. La figure 1.4 présente l'exemple d'un réseau social.

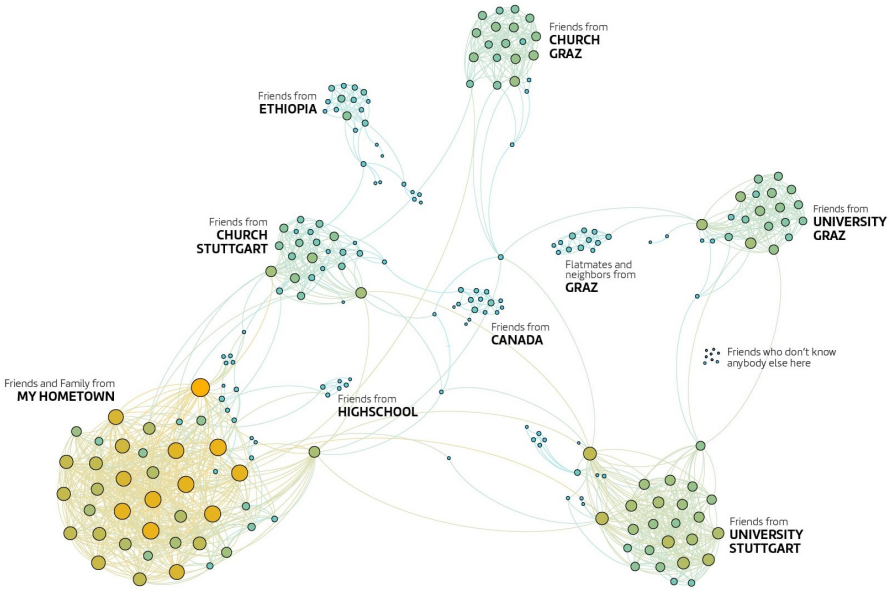


FIGURE 1.4 – Exemple d'un réseau social (Röhm, 2014).

À ses débuts, la science des réseaux sociaux était la filière des sciences sociales centrée sur l'étude des interactions entre différents acteurs. Bien que Barnes soit le premier à avoir utilisé le terme « réseau social » en 1954 dans la référence (Barnes, 1954), la science des réseaux a connu ses débuts avec Moreno dès 1934. Celui-ci, dans le livre (Moreno, 1934) a établi les bases de la sociométrie et s'est intéressé à l'effet des relations et des interactions sociales sur les actions des individus et leurs comportements et psychologies. À partir de 1960, l'analyse des réseaux sociaux connaît une expansion avec les travaux de Harisson White et ses étudiants à l'université de Harvard. Le sociologue et son groupe ont généralisé en effet l'étude de ces structures sociales par des applications variées (Wasserman and Faust, 1994) telles que l'analyse des conflits dans les systèmes d'entreprises, l'étude des amitiés, l'analyse de l'appartenance aux communautés scientifiques, l'étude de la mobilité et de progression dans les organisations, l'exploration de la force des liens sociaux, etc.

Depuis, plusieurs recherches des réseaux du monde réel ont permis de mettre en évidence des caractéristiques intéressantes telles que les six degrés de séparations, le phénomène du petit monde, la distribution en loi de puissance des degrés, l'autosimilarité dans les réseaux complexes, etc. (Milgram, 1967; Watts and Strogatz, 1998; Song et al., 2005). Les dernières décennies ont été marquées par une avancée impressionnante dans le domaine de l'analyse des réseaux sociaux. Ceci a généré des recherches approfondies dans le domaine avec les travaux de Wasserman et Faust (Wasserman and Faust, 1994), Newman et Clauset (Newman and Girvan, 2004b; Clauset et al., 2004), Aggrawal (Aggarwal and Lu, 2010), Fortunato (Fortunato, 2010), Barabási (Barabási and Pósfai, 2016), et beaucoup encore.

Par voie de conséquence, les réseaux sociaux ont été employés pour étudier les changements de comportements sociaux, politiques ou économiques d'individus et ont trouvé des applications dans les domaines de la sociologie, la biologie, la santé, la politique, la prise de décision, la diffusion et l'adoption de nouvelles technologies,

les études de marché, etc. Aujourd'hui, l'analyse des réseaux sociaux est devenue un outil utilisé par les sciences *dures* dans le domaine scientifique et les travaux de recherches se sont accrus, car les mesures initialement destinées à l'analyse des réseaux sociaux ont pu être généralisées pour d'autres types de réseaux. (Barabási and Pósfai, 2016) soutiennent même que pour pouvoir comprendre des systèmes de plus en complexes, il est indispensable d'appréhender les réseaux qui les constituent.

Avec l'évolution des techniques de cartographie, la capacité de stockage des ordinateurs, la science des graphes, et la science des données, ce domaine a connu une réelle révolution. Les réseaux sont maintenant au centre de nouvelles technologies allant des moteurs de recherche tels que Google, aux plateformes de médias sociaux tels que Facebook, Twitter, LinkedIn, etc.

1.3.1 Exemples de réseaux sociaux

La figure 1.5 et la figure 1.6 présentent deux exemples de réseaux sociaux largement étudiés dans l'état de l'art. Le premier réseau représente le réseau du *Club de karaté de Zachary* (Zachary, 1977). Ce réseau a été créé en 1977 en se basant sur des données réelles, afin d'étudier les relations entre les membres d'un club de karaté et les communautés qui se sont formées lorsque le club a été divisé en deux factions. Il est devenu célèbre pour avoir été largement utilisé comme un cas d'étude pour la détection de communautés. Le réseau est composé de 34 noeuds représentant les membres du club, et les liens entre les noeuds représentent les relations sociales entre les membres.

Le deuxième réseau est le réseau du *football universitaire américain* (Park and Newman, 2005). Ce réseau présente le calendrier des matchs de la saison régulière de football universitaire de l'automne 2000. Sur ce réseau, les sommets représentent les universités participant au tournoi et les arêtes représentent les matchs joués. Le graphe contient 115 nœuds et 613 arêtes. Le réseau de football universitaire américain est largement utilisé dans l'état de l'art, notamment pour comparer les

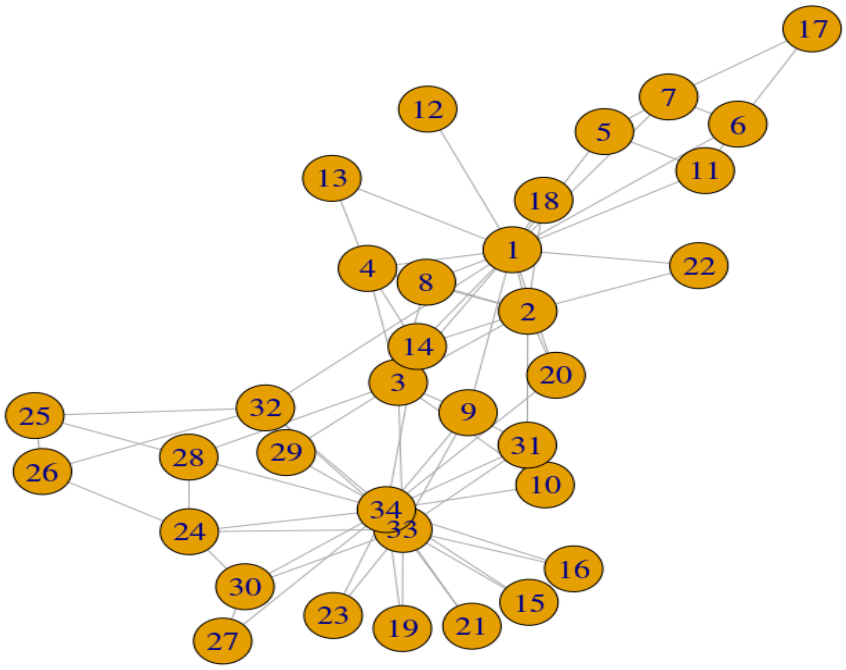


FIGURE 1.5 – Présentation du réseau social du club de karaté de Zachary.

algorithmes de détection de communautés.

1.3.2 Applications

Bien que diverses, les applications sur les réseaux sociaux peuvent être focalisées sur trois axes principaux (Aggarwal et al., 2018). Ces axes sont :

- L'identification des acteurs centraux dans le réseau en utilisant des mesures statistiques telles que la centralité ;
- La prédiction et l'analyse des liens en explorant la propagation d'influence ou des flux entre les individus ;
- La détection de communautés qui se base sur la recherche de groupes denses à l'intérieur du réseau.

La notion de réseau social apparaît comme une approche adaptée permettant l'analyse dans le contexte de nos travaux. L'analyse des transactions d'ORange Money en

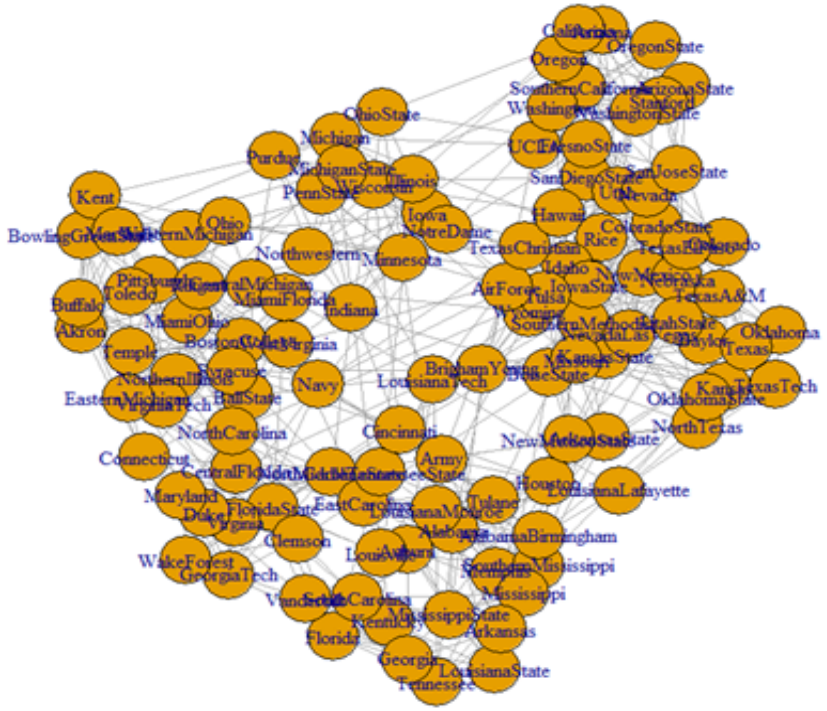


FIGURE 1.6 – Illustration du réseau social du *football universitaire américain.*

se basant sur les réseaux permettra une bonne compréhension des comportements individuels et collectifs des clients du système financier, tout en se basant sur les interactions entre eux. Les données de transactions peuvent alors prendre la forme d'un réseau social , où les sommets représentent les utilisateurs du système financier, et les liens correspondent aux interactions financières, telles qu'un paiement ou un virement.

L'analyse des réseaux sociaux offre une approche complète pour l'étude des données relatives au service de paiement mobile et permet de mieux comprendre les relations sociales et économiques sous-jacentes qui façonnent l'utilisation et l'impact de ce service. Elle présente ainsi un outil puissant permettant de cartographier les connexions et les flux d'argent au sein d'un réseau, ce qui peut fournir des informations précieuses sur la manière dont l'argent est échangé et dont il circule entre les différents utilisateurs.

En outre, l'analyse des réseaux sociaux peut être utilisée pour identifier les acteurs clés au sein du réseau, comprendre comment le pouvoir et l'influence opèrent au sein du réseau et comment cette dynamique affecte le fonctionnement global du réseau, identifier des modèles et des tendances dans les données, telles que la fréquence des transactions, les montants communs des transactions et l'impact d'événements spécifiques sur le réseau, découvrir les modèles de circulation des flux financiers dans le réseau, etc. Dès lors, une bonne compréhension des relations entre les utilisateurs du système financier devrait permettre d'ajuster et d'améliorer les services proposés.

1.4 Réseau des données de transactions Orange Money

Avec l'expansion des services de paiement mobiles, les données de transactions d'Orange Money représentent un volume très important, de l'ordre de plusieurs dizaines de giga-octets de données quotidiennes. Cependant, ces données sont encore relativement peu valorisées. L'analyse de ces données doit nous permettre de mieux comprendre l'utilisation du service Orange Money par ses clients et ainsi permettre de l'adapter, le personnaliser et l'enrichir de nouvelles fonctionnalités.

L'analyse des réseaux sociaux a déjà été utilisée avec succès sur des données bancaires et de transfert d'argent pour réaliser des études socio-économiques (Centellegher et al., 2018), découvrir les habitudes d'achat des clients (Di Clemente et al., 2018), lutter contre la fraude (El Ayeb et al., 2020), etc.

1.4.1 Description des données Orange Money

Dans le cadre des travaux de cette thèse, nous disposons des données de transactions du service Orange Money. Ces données comportent un ensemble de transactions effectuées sur une période donnée. Chaque transaction dans nos données représente un échange d'argent mobile entre deux utilisateurs du service. Une transaction dis-

pose d'un nombre d'attributs tels que les identifiants des utilisateurs respectifs après anonymisation, le montant de la transaction, la date de la transaction, etc. Chaque ligne de cette base de données est spécifique à une transaction effectuée. Le tableau 1.1 présente un exemple simplifié de la base de données des transactions d'Orange Money.

TABLEAU 1.1 – Extrait des données de transactions.

Sender user id	Sender category	Receiver user id	Receiver category	Transaction amount	Service type	Transaction Date-time
user 1	customer	user 2	customer	1500	PEER TO PEER	11/01/2022 11 :00
user 2	customer	user 1	customer	3000	PEER TO PEER	26/02/2022 16 :15
user 5	retailer	user 3	customer	1500	CASH-IN	08/03/2022 18 :06
user 3	customer	user 7	customer	500	PEER TO PEER	21/03/2022 10 :20
user 3	customer	user 2	customer	500	PEER TO PEER	15/04/2022 15 :55
user 4	retailer	user 3	customer	700	CASH-IN	04/06/2022 11 :30
user 3	customer	user 6	merchant	1000	MERCHPAY	18/07/2022 9 :00
user 2	customer	user 3	customer	5000	PEER TO PEER	04/09/2022 16 :30
user 3	customer	user 2	customer	500	PEER TO PEER	06/10/2022 15 :13
user 7	customer	user 3	customer	800	PEER TO PEER	23/10/2022 12 :00
user 3	customer	user 5	retailer	1500	CASH-OUT	07/11/2022 14 :15
user 5	retailer	user 3	customer	1500	CASH-IN	04/12/2022 11 :01
user 3	customer	user 2	customer	500	PEER TO PEER	29/12/2022 17 :45

La première étape de nos travaux consiste à transformer les données tabulaires présentées par 1.1 sous forme de réseau. Ceci permettra de représenter les données sous forme de sommets et de liens, plutôt que sous la forme de lignes et de colonnes. Cela est particulièrement utile lorsque les données ont des relations élaborées et dépendantes les unes des autres. Cette transformation permet également de visualiser ces liens complexes de manière intuitive sous forme de graphes. On peut visualiser la représentation du réseau de transactions étudié sur la figure 1.7. Chaque ligne du tableau est convertie en lien permettant de joindre les deux sommets (ou clients), définis par leurs identifiants uniques, qui ont effectué le transfert d'argent. Ce lien

porte des informations telles que le type de la transaction, sa date de transmission, et le montant impliqué.

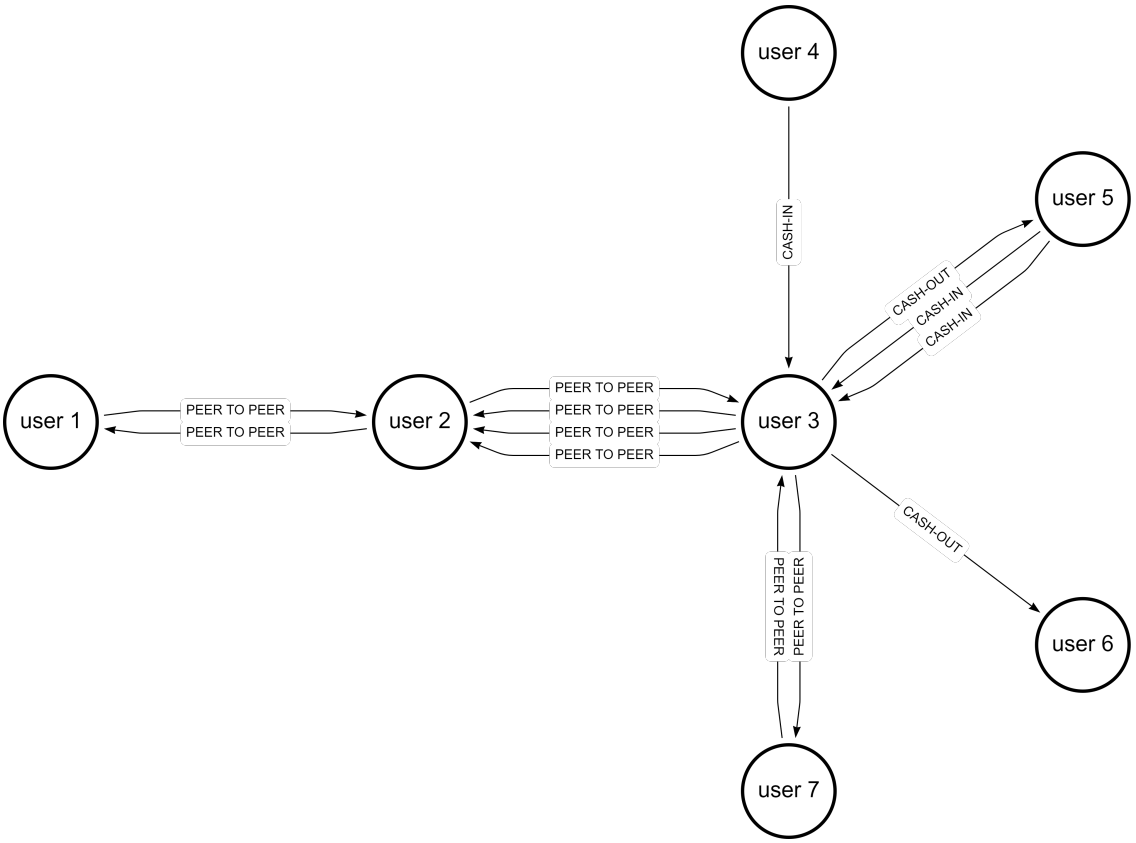


FIGURE 1.7 – Extrait d'un exemple du réseau de transactions.

Dans la suite des travaux de la thèse, nous nous intéressons particulièrement aux transactions « peer to peer » ou transferts d'argent entre les abonnés. Cependant, toutes les applications et contributions de la thèse sont applicables aux autres types de transactions. Notre choix est principalement orienté par le nombre considérable de ces transactions et le fait qu'elles sont les plus représentatives des interactions sociales par rapport au reste.

1.5 Conclusion

Parce que les réseaux présentent un type de données spécifique, leur traitement nécessite également des méthodes, des mesures, et des tests spécifiques (Wasserman and Faust, 1994). Dans ce chapitre, nous avons présenté le service de paiement mobile et discuté de son importance croissante dans la part des transactions actuelles. Nous avons également présenté le service Orange Money, dont les données de transactions sont au centre de nos travaux. Nous avons par la suite présenté l'analyse des réseaux sociaux et la motivation derrière la transformation des données de transactions en réseau social. Nous avons vu comment les données de transactions peuvent être utilisées pour créer des graphes qui reflètent les relations entre les différents utilisateurs et comment cela peut permettre de découvrir des perspectives cachées dans les données.

Cette analyse des données de transactions par le moyen des réseaux sociaux est un outil puissant pour mieux comprendre les besoins des utilisateurs et améliorer les services financiers. L'analyse des réseaux repose sur une multitude de méthodes pour étudier ces structures.

Dans les prochains chapitres, nous allons explorer spécifiquement l'application de la détection de communautés sur le réseau de transactions. Pour ce faire, nous allons poursuivre nos travaux en nous concentrant sur les propriétés des réseaux et des graphes, en abordant notamment les applications de détection de communautés sur le réseau social de transactions Orange Money, et en étudiant comment adapter ces applications aux particularités des données en notre possession.

Chapitre 2

Généralités sur les graphes et réseaux

Les réseaux sociaux constituent un type spécifique de réseaux dont les composantes principales sont les acteurs sociaux (personnes, groupes, entreprises, etc.) ainsi que leurs interactions mutuelles. Fondée sur l'importance des relations, l'analyse des réseaux sociaux a trouvé dans la théorie des graphes une base pour l'étude de ses propriétés. Nous présentons dans ce chapitre les définitions de base et les propriétés générales liées aux réseaux et aux graphes.

Sommaire

2.1 Introduction

Définis comme un ensemble d'éléments connectés par leurs interactions, les réseaux représentent depuis quelques années un outil fondamental pour la description de plusieurs phénomènes sociaux, économiques, physiques, biologiques, etc. Initialement motivée par des observations du monde réel, l'analyse des réseaux a permis de comprendre les similarités entre les mécanismes sous-jacents des systèmes complexes, et de quantifier leurs structures et leurs dynamiques.

Ainsi, l'exploration des réseaux sociaux a trouvé dans la théorie des graphes les aspects mathématiques indispensables permettant de décrire à la fois les objets et leurs relations (Iñiguez et al., 2020). Les graphes offrent une représentation adéquate et des concepts théoriques facilitant l'étude de ces structures. En effet, le recours aux graphes pour résoudre des problèmes complexes remonte au XVIIIe siècle avec l'énigme des sept ponts de Königsberg dans la capitale de la Prusse orientale. En 1736, cette énigme a été résolue par Leonhard Euler en posant les bases de la théorie des graphes (Euler, 1741). Le développement de la théorie des graphes a ainsi favorisé l'essor de la science des réseaux sociaux. Initialement motivée par des observations du monde réel, elle a permis de comprendre les similarités entre les mécanismes sous-jacents des systèmes complexes, et de quantifier leurs structures et leurs dynamiques.

Ce chapitre vise à présenter les propriétés des réseaux et des graphes. Les principaux points théoriques à connaître pour une meilleure compréhension du contenu de cette thèse sont présentés dans ce chapitre. Nous allons également présenter les données synthétiques que nous allons utiliser pour le reste de nos travaux.

2.2 Propriétés des réseaux

Au cours de cette thèse, notre travail s'est focalisé sur l'étude des réseaux sociaux, plus précisément le réseau social de transactions décrivant les échanges monétaires

des clients du service Orange Money. Cependant, les réseaux peuvent avoir différentes formes et tailles. Ils ne se limitent pas à la sphère sociale, ils peuvent également traiter différents types d'interactions : les réseaux de protéines et de gènes en biologie, les réseaux neuronaux en médecine, les réseaux de transports tels que les réseaux ferroviaires et les réseaux de routes reliant des régions urbaines, les réseaux de capteurs électriques, etc.

Le choix des éléments de tout réseau dépend du type de données et devrait fournir la compréhension la plus appropriée et la plus utile du phénomène à traiter. Selon le type des composants du réseau, on peut retrouver des réseaux unimodaux (simples) lorsque tous les éléments sont de même type ou des réseaux multimodaux (complexes) pouvant inclure plusieurs ensembles d'éléments (Hawe et al., 2004).

Dans la littérature, la conception des réseaux repose en général sur trois familles de variables (Wasserman and Faust, 1994) :

— Les *variables structurelles* permettent de caractériser de façon globale les liens entre les paires d'éléments en interaction. Elles sont calculées sur l'ensemble du réseau. Elles permettent d'évaluer la force d'un lien d'amitié, une co-citation dans un article ou une route entre deux régions (Fuhrer and Cucchi, 2012), etc.

— Les *variables de composition* mesurent des caractéristiques individuelles des éléments du réseau. Elles servent à caractériser leurs degrés d'influence et d'importance (Kamau et al., 2018). Ces variables sont définies au niveau des éléments pour indiquer des propriétés telles que le type, l'âge, la localisation géographique, le nombre d'habitants, etc.

— Les *variables d'affiliation* sont destinées à caractériser l'appartenance des éléments du réseau à certains groupes définis. Ces variables sont calculées non pas au niveau individuel, mais au niveau de groupes. Ces groupes peuvent présenter des individus appartenant à la même organisation, participant aux

mêmes évènements, ou ayant des habitudes similaires. Dans le réseau de transactions Orange Money par exemple, ces groupes peuvent être des familles, des amis, un marchand et l'ensemble de sa clientèle, etc.

Dans le contexte de l'étude des réseaux sociaux, (Lazega, 2007) a ajouté un quatrième type de variable : les *variables de comportement*. Il s'agit du comportement potentiellement influencé par la position des acteurs et de leurs voisins dans le réseau. Cette variable est plus particulièrement exploitée dans le contexte d'étude d'évolution de comportements ou d'adoption d'un service ou une technologie.

Durant les dernières années, la recherche sur les réseaux a été marquée par un mouvement substantiel où l'accent s'est déplacé de l'analyse de petits réseaux et de leurs composantes individuelles vers l'étude des propriétés statistiques à grande échelle des réseaux complexes.

Définition 1 *« Un réseau complexe est un réseau composé de multiples entités en interaction dont le comportement collectif entraîne l'émergence de propriétés ne pouvant être déduites des propriétés individuelles de ses éléments » (Society, 2022).*

Des exemples de systèmes complexes comprennent les fourmilières, les systèmes économiques, le climat, les systèmes nerveux, les êtres humains, etc.

L'homologie de structure entre les systèmes réels de réseaux et les graphes a permis aux chercheurs de se baser sur la théorie des graphes afin de comprendre certains systèmes naturels et même de prédire certains comportements. Dans la littérature, la séparation entre réseaux et graphes est assez subtile. Dans cette thèse, nous allons nous baser sur les appellations *graphes* pour désigner les objets mathématiques abstraits et *réseaux* pour désigner des systèmes réels. Dans la partie suivante, nous allons présenter des définitions préliminaires reliées aux graphes ainsi que leurs propriétés.

2.3 Les graphes

Le succès de l'analyse des réseaux tient principalement à l'exploitation de puissants outils mathématiques tels que les graphes. Les graphes offrent une description adéquate des réseaux, exhibant les acteurs en tant que « sommets », et leurs relations, en tant qu'« arêtes ».

2.3.1 Définitions et notations

Définition 2 *Un graphe est une structure mathématique formée par un ensemble fini de points qu'on appelle sommets ou nœuds et de liaisons qu'on appelle arêtes ou arcs. Il est souvent noté sous la forme d'un couple $\mathcal{G} = (\mathcal{V}, E)$, où $\mathcal{V}$ présente l'ensemble des sommets (Vertices) et E l'ensemble des arêtes (Edges).*

Un graphe est défini par son ordre $|\mathcal{V}|$, exprimant le nombre de ses nœuds, et par sa taille $|E|$, exprimant le nombre de ses arêtes. Une arête est un couple $(v_i; v_j)$ reliant les sommets v_i et v_j.

Un sous-graphe $\mathcal{G}' = (\mathcal{V}', E')$ de $\mathcal{G} = (\mathcal{V}, E)$ est un graphe tel que $\mathcal{V}' \subseteq V$ et $E' \subseteq E$.

Un graphe *partiel* de $\mathcal{G}$ est un graphe ayant le même nombre de sommets que $\mathcal{G}$, mais avec moins d'arêtes c'est-à-dire $\mathcal{V}' = \mathcal{V}$ et $E' \subset E$.

Définition 3 *Un graphe est dit complet lorsque tous ses couples de sommets sont liés par des arêtes.*

Le nombre d'arêtes d'un graphe complet d'ordre $|\mathcal{V}|$ est égal à $\frac{|\mathcal{V}|(|\mathcal{V}|-1)}{2}$.

Deux sommets v_i et v_j d'un graphe sont dits *voisins* ou *adjacents*, s'ils sont reliés par une arête, c'est-à-dire s'il existe une arête $e \in E$ telle que $e = (v_i; v_j)$. Deux arêtes sont adjacents s'ils ont une extrémité en commun.

Définition 4 *On appelle distance entre deux sommets le nombre minimal d'arêtes les liant. La distance la plus courte mesurée en nombre de liens entre deux sommets*

est nommée distance géodésique d_{geo}.

Le *voisinage* d'un nœud v_i est noté $N(v_i)$, avec $N(v_i) = \{v_j \in \mathcal{V}; (v_i; v_j) \in E\}$. On note $deg(v_i)$ le degré d'un sommet. Il est égal au cardinal du voisinage du nœud v_i :

$$deg(v_i) = |N(v_i)|; \tag{2.1}$$

Le *degré moyen* est la moyenne des degrés de tous les nœuds d'un graphe. Dans un souci de concision, plus de définition reliées aux graphes peuvent être retrouvées dans le Glossaire 5.5.

2.3.2 Types de graphes

Selon sa structure et ses composants, un graphe peut être de type simple, orienté, pondéré, biparti, multigraphe ou multicouche, etc. Dans ce qui suit, nous allons présenter quelques types de graphes rencontrés dans l'état de l'art.

- Graphe simple : un graphe simple est un graphe ayant un lien simple symétrique (sans orientation et sans attribut) entre les paires de nœuds qui le construisent. Les graphes simples sont largement étudiés car ils permettent de décrire de manière claire et concise une multitude de systèmes réels où les éléments sont reliés entre eux tels que la participation à un évènement donné, l'achat d'un produit, un réseau des voies ferrées, etc.

Un graphe simple est défini par $\mathcal{G} = (\mathcal{V}, E)$, où E représente l'ensemble des arêtes symétriques vérifiant : $(v_i; v_j) = (v_j; v_i); \forall v_i, v_j \in \mathcal{V}$.

- Graphe orienté : un graphe orienté est un graphe où les arêtes ont un sens : du nœud émetteur au nœud récepteur. Ce sens traduit l'orientation de circulation de l'échange d'information, ou de ressources entre les couples de nœuds. Les graphes orientés peuvent décrire le réseau de citations scientifiques, des réseaux d'échanges d'appels téléphoniques, des réseaux de transport de colis postaux, etc.

Un graphe orienté est souvent noté par $\mathcal{G} = (\mathcal{V}, E)$, où E représente l'ensemble d'arêtes orientées vérifiant : $(v_i; v_j) \neq (v_j; v_i); \forall v_i, v_j \in \mathcal{V}$.

Pour les graphes orientés, on peut distinguer entre le demi-degré intérieur $deg(v_i)^{in}$ ou nombre d'arêtes arrivant à un sommet, et le demi-degré extérieur $deg(v_i)^{out}$ ou nombre d'arêtes partant d'un sommet.

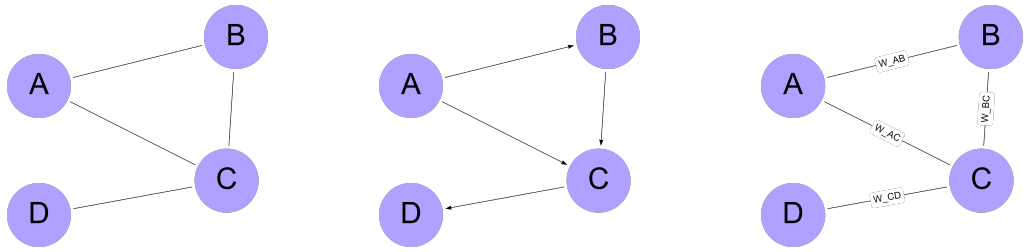
On a alors :

$$deg(v_i) = deg(v_i)^{in} + deg(v_i)^{out} \tag{2.2}$$

- Graphe pondéré : un graphe pondéré est un graphe qui attribue à chaque arête un poids positif : les poids reflètent l'intensité du lien entre chaque couple de sommets et peuvent représenter toute propriété des liens en relation avec le problème étudié. Ainsi les poids peuvent représenter le coût d'achat d'un produit donné, le nombre de visites d'un patient pour son médecin, le temps nécessaire pour se déplacer entre deux villes, etc.

Un graphe pondéré est défini par $\mathcal{G} = (\mathcal{V}, E, W)$ où W est une fonction qui à chaque arête associe un réel positif $W : E \longrightarrow \mathbb{R}^+$.

La figure 2.1 présente les différents types de graphes susmentionnés.



(a) Graphe simple (b) Graphe orienté (c) Graphe pondéré

FIGURE 2.1 – Différents types de graphes.

- Multigraphe : à l'inverse du graphe simple, un multigraphe est un graphe qui autorise des liens multiples (aussi appelés liens parallèles) entre les nœuds. En d'autres termes, dans un multigraphe, deux sommets peuvent être connectés par plus d'un

lien. Généralement, l'analyse de ces graphes implique une simplification par remplacement des liens multiples en liens simples ayant plusieurs attributs, afin de permettre l'application des méthodes conventionnelles (Ducruet, 2012). Un multigraphe est défini par $\mathcal{G} = (\mathcal{V}, E, \alpha)$ où α représente le cœfficient multiplicatif pour chaque arête $(v_i; v_j)$ indiquant le nombre de fois qu'elle est présente. La figure 2.2 présente un exemple d'un multigraphe.

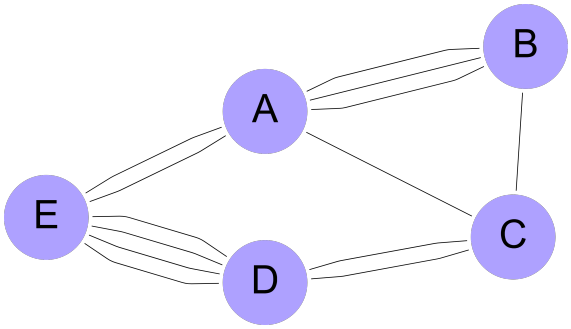


FIGURE 2.2 – Exemple d'un multigraphe

- Graphe multicouche : Un graphe multicouche ou multiplex est un graphe composé d'un ensemble de nœuds de même type, reliés par différents types de relations. Chaque couche contient le même ensemble de nœuds, mais correspond à un type de relation distinct. Par exemple, dans le cas de réseaux bibliographiques, on peut définir un multiplex où les nœuds sont les auteurs et chaque couche correspond à une relation différente : co-publication, co-citation, co-cités, co-participation à une conférence, etc. (Kanawati, 2013). Les notations utilisées pour des graphes simples doivent être étendues pour permettre de représenter les structures qui ont des couches en plus des nœuds et des liens. D'une manière formelle, un graphe multiplex structuré en n couches peut être défini par : $\mathcal{G} = (\mathcal{V}, E_1, ..., E_n)$.

La figure 2.3 présente un exemple d'un graphe multicouche.

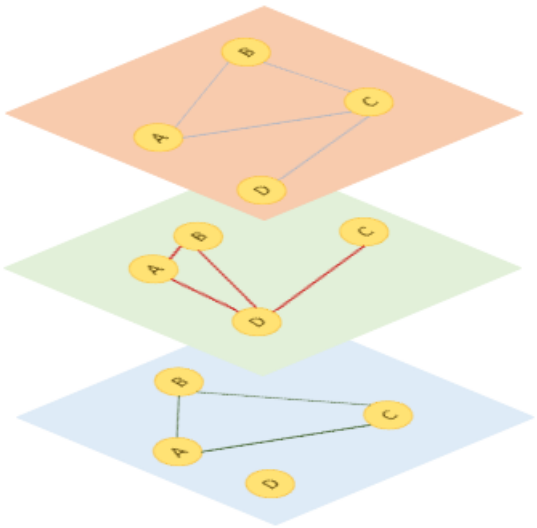


FIGURE 2.3 – Exemple d'un graphe multicouche.

2.3.3 Modes de représentation

Il existe différentes méthodes de représentation des graphes. Parmi les méthodes les plus utilisées on retrouve : la représentation graphique, la liste d'adjacence, la matrice d'adjacence, et la liste des arêtes. Pour illustrer ces méthodes, nous allons utiliser le graphe simple, non orienté et non pondéré de la figure 2.4. Ce graphe est composé de cinq nœuds (A,B,C,D,E) et cinq arêtes.

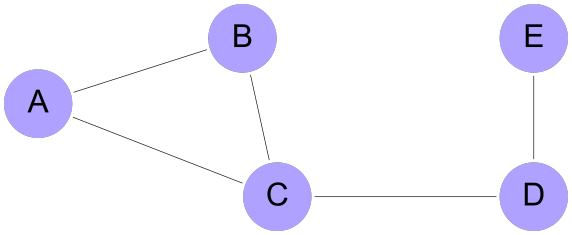


FIGURE 2.4 – Exemple d'un graphe simple.

2.3.3.1 La liste d'adjacence

Une liste d'adjacence recense tous les voisins de chaque nœud du graphe sous forme de liste. Le graphe de la figure 2.4 a donc la liste d'adjacence : {(A : B,C), (B : A,C), (C : A,B,D), (D : C,E), (E : D)}. Les listes d'adjacence permettent l'exploration

facile des voisins de chaque nœud. Généralement, les listes d'adjacence ne sont pas ordonnées.

2.3.3.2 La matrice d'adjacence

La matrice d'adjacence A d'un graphe ayant $|\mathcal{V}|$ nœuds est une matrice de taille $|\mathcal{V}| \times |\mathcal{V}|$. Ses éléments vérifient :

$$\begin{cases} A_{v_i v_j} = 1 & \text{si } (v_i,\, v_j) \in \text{E}; \\ A_{v_i v_j} = 0 & \text{sinon.} \end{cases}$$

Pour un graphe pondéré, les éléments de la matrice d'adjacence prennent les valeurs des poids des arêtes :

$$\begin{cases} A_{v_i v_j} = w_{v_i v_j} & \text{si } (v_i; v_j) \in E; \\ A_{v_i v_j} = 0 & \text{sinon.} \end{cases}$$

Pour un graphe non orienté, la matrice d'adjacence est une matrice symétrique c'est-à-dire $A_{v_i v_j} = A_{v_i v_j}$ pour tous les nœuds v_i et v_j. Pour le graphe de la figure 2.4, la matrice d'adjacence est égale à :

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Cette représentation matricielle permet d'acquérir des informations sur la topologie du graphe, et les relations entre les nœuds, et de calculer des indicateurs locaux et globaux du graphe. Dans le cas d'un multigraphe, les valeurs de la matrice d'adjacence correspondent au nombre d'arêtes entre les paires de nœuds.

2.3.3.3 La liste d'arêtes

Comme l'indique son nom, la liste d'arêtes correspond à la liste de l'ensemble des arêtes du graphe. Chaque élément de cette liste contient les deux nœuds formant les extrémités de l'arête correspondante. Si le graphe est orienté, chaque élément de la liste correspond à un couple de nœuds source et destination. S'il est pondéré, chaque élément contiendra une troisième information, qui est le poids de l'arête entre les nœuds. Le graphe exemple aura la liste d'arêtes suivante : {(A,B), (A,C), (B,C), (C,D), (D,E)}.

Il est à noter que les complexités des méthodes de représentation citées varient selon l'opération effectuée. En général, les listes d'adjacence et les listes d'arêtes nécessitent moins d'espace de stockage. De ce fait, ces deux structures sont souvent utilisées pour la représentation des graphes peu denses. En contrepartie, les matrices d'adjacence sont plus utilisées pour les manipulations mathématiques.

2.4 Mesures sur les nœuds

Dans cette partie, nous allons présenter les mesures permettant d'évaluer les caractéristiques individuelles des nœuds. Les mesures sur les nœuds permettent de quantifier et de caractériser l'importance d'un nœud par rapport aux autres nœuds dans le graphe.

2.4.0.1 Le cœfficient de partitionnement

Le *cœfficient de partitionnement* CC (Clustering Coefficient) (Watts and Strogatz, 1998) mesure le degré du regroupement des nœuds d'un graphe. Il permet de caractériser le degré de connectivité du voisinage d'un nœud. Plus précisément, il permet de calculer la probabilité que deux nœuds soient connectés, lorsqu'ils ont un nœud voisin en commun. Pour un nœud $v_i \in \mathcal{G}$, de degré $deg(v_i)$, et un nombre de liens l_{v_i} reliant ses voisins, le cœfficient de partitionnement est défini par :

$$CC(v_i) = \frac{2l_{v_i}}{deg(v_i)(deg(v_i) - 1)} \tag{2.3}$$

Par définition, le cœfficient de partitionnement varie entre 0 et 1. Plus un voisinage est relié, plus le cœfficient est proche de 1. Il existe une forme globale du cœfficient de partitionnement, prenant en considération l'ensemble du graphe. Il représente la moyenne des cœfficients de partitionnement individuels. Un cœfficient de partitionnement global égal à 1 correspond à un graphe complet. Le cœfficient de partitionnement global est défini par :

$$< C >= \frac{1}{|\mathcal{V}|} \sum_{v_i \in \mathcal{V}} CC(v_i) \tag{2.4}$$

2.4.0.2 Les centralités

La centralité permet d'évaluer la place qu'occupe un nœud dans le graphe. Elle peut être calculée de plusieurs manières. Le degré, le vecteur propre, la proximité et les plus courts chemins sont différentes mesures de la notoriété d'un nœud (Wasserman and Faust, 1994). Bien qu'il existe une analogie conceptuelle considérable entre ces notions, elles ont chacune leurs particularités. Une approche permettant de distinguer les différentes mesures consiste à considérer les caractéristiques des nœuds qui possèdent des valeurs de centralité élevées et leur influence sur le reste du graphe dans chacun des cas (Valente et al., 2008). Dans ce qui suit, nous allons présenter les centralités les plus connues de l'état de l'art.

- La centralité de degré (degree centrality) (Freeman, 1979) est souvent employée dans les contextes où on cherche le nœud populaire au sein du graphe. Cette centralité reflète la capacité de chaque nœud à établir des relations avec les autres. Elle est donnée par la somme des liens (entrant, sortant ou les deux à la fois) dans lesquelles un nœud est engagé. Pour un nœud donné, la centralité de degré est donnée par l'équation :

$$Centralité\ de\ degré\ (v_i) = \frac{deg(v_i)}{|\mathcal{V}| - 1} \tag{2.5}$$

La majorité des applications mettant en œuvre la centralité de degré repose sur la recherche des nœuds influents dans des réseaux bibliométriques (Kretschmer and Kretschmer, 2007), des réseaux de média sociaux tels que Twitter ou Facebook (Rachman et al., 2013; Yustiawan et al., 2015), etc.

- La centralité des vecteurs propres (eigenvector centrality) (Bonacich, 1972) est une généralisation de la centralité de degré permettant d'évaluer l'incidence des nœuds en se basant sur l'influence relationnelle. Cette centralité repose sur l'idée que l'importance d'un nœud est définie par l'importance de ses voisins. Ainsi, un score de centralité de valeur propre important implique que le nœud est relié à des voisins qui ont aussi des scores élevés. La centralité de valeurs propres d'un nœud v_i est donnée par :

$$Centralité\ de\ valeurs\ propres\ (v_i) = \frac{1}{\lambda} \sum_{v_j \in N(v_i)} C_{vp}(v_j) \tag{2.6}$$

où $\lambda \neq 0$ est la plus grande valeur propre de la matrice d'adjacence A, et $C_{v_p} = [C_{vp}(v_1), C_{vp}(v_2), \ldots, C_{vp}(v_n)]^T$ est le vecteur propre associé à λ.

Les sociologues ont été les pionniers à utiliser des versions de la centralité du vecteur propre pour étudier les relations dans des réseaux sociaux (Bonacich and Lloyd, 2001). Des travaux ont été également réalisés pour appliquer la centralité des vecteurs propres en biologie pour l'étude des protéines et des gènes (Parisutham and Rethnasamy, 2021).

- La centralité d'intermédiarité (betweenness centrality) (Freeman, 1977) est la centralité la plus fréquemment mise en œuvre. Elle permet de déterminer les « nœuds relais » ou ayant une influence sur le flux de l'information circulant dans le graphe. Ces nœuds forment généralement des points de passages importants permettant de relier rapidement deux sommets du graphe. La centralité d'intermédiarité repose sur le calcul du nombre de fois où un nœud se trouve sur les plus courts

chemins reliant toutes les autres paires de nœuds. Un niveau élevé de centralité d'intermédiarité n'est pas forcément corrélé avec un degré important du sommet : un nœud avec un faible degré faisant le lien entre deux groupes de sommets aura une centralité d'intermédiarité élevée. La centralité d'intermédiarité est donnée par :

$$\textit{Centralité d'intermédiarité}\,(v_i) = \sum_{v_j \neq v_i, v_k \neq v_i, v_j \neq v_k} \frac{\sigma_{v_j v_k}(v_i)}{\sigma_{v_j v_k}} \tag{2.7}$$

où $\sigma_{v_j v_k}(v_i)$ est le nombre des plus courts chemins entre deux sommets quelconques v_j, et $v_k \in \mathcal{V}$, passant par v_i.

La centralité d'intermédiarité peut également être appliquée aux arêtes pour mesurer le nombre des plus courts chemins entre deux nœuds, passant par une arête spécifique. Malgré sa large utilisation, la centralité d'intermédiarité présente l'inconvénient d'être coûteuse en termes de mémoire et de temps de calcul. Pour ceci, de nombreux articles ont considéré l'optimisation de cette métrique, notamment par des heuristiques (Puzis et al., 2012).

- La centralité de proximité (closeness centrality) (Beauchamp, 1965) permet d'identifier les nœuds capables de propager de manière efficace de l'information au sein du graphe. Plus un nœud est central, plus il est étroitement lié aux autres nœuds du graphe, et interagit facilement avec eux. La centralité de proximité mesure le nombre de nœuds par lesquels un nœud donné doit passer pour entrer en contact avec les autres nœuds. Pour un nœud donné, elle est égale à l'inverse de la somme des distances géodésiques le séparant du reste du graphe.

$$\textit{Centralité de proximité}\,(v_i) = \frac{1}{\sum_{\substack{v_j \in \mathcal{V} \\ v_j \neq v_i}} d_{geo}(v_i; v_j)} \tag{2.8}$$

Etant donné que la centralité de proximité permet d'identifier les individus susceptibles d'acquérir un contrôle au sein du réseau, elle a servi dans diverses études telles que l'étude des réseaux d'organisations pour détecter les groupes frauduleux ou criminels (Krebs, 2002). On lui trouve également des applications dans le do-

maine d'analyse des documents, plus spécifiquement l'extraction des phrases clés (Boudin, 2013).

- La centralité PageRank (PR) a été initialement introduite dans le contexte de requête de pages web par les concepteurs du moteur de recherche *Google*. PageRank a été employé afin de créer un classement de qualité pour chaque page web, permettant à ce moteur de recherche de fournir des résultats de forte précision (Page et al., 1999). Ainsi, la centralité PageRank évalue l'importance de chaque nœud en se basant sur le nombre de liaisons lui arrivant et l'importance des nœuds sources correspondants. La formule de la centralité PageRank est donnée par (2.9) :

$$PR(v_i) = (1-\alpha) + \alpha \times \sum_{v_k \neq v_i} \frac{PR(v_k)}{s(v_k)} \tag{2.9}$$

où $v_1, \cdots, v_k$ correspondent aux nœuds pointant vers v_i, $s(v_i)$ correspond aux nombres de liens sortants de v_i, et α est le facteur de *Dampington* variant entre 0 et 1 (en général $\alpha = 0.85$). Bien que la notion du PageRank ait été originairement conçue pour l'étude des pages web, d'autres utilisations ont suivi permettant de créer des alternatives adaptées aux différents domaines étudiés telles que le *GeneRank* en biologie (Morrison et al., 2005), et *TURank* et *TwitterRank* dans le cadre de l'analyse du réseau de *Twitter* (Yamaguchi et al., 2010; Weng et al., 2010).

2.5 Les données synthétiques

La recherche dans le domaine de l'analyse des réseaux sociaux a été fortement conditionnée par la disponibilité des données d'interaction à étudier. Depuis ses débuts, elle a été tributaire de la collecte de l'information à travers des questionnaires, des entretiens ou des observations directes. Aujourd'hui, des réseaux réels sont accessibles à travers des interfaces de programmation d'application (API), en payant des entreprises pour l'accès aux données spécifiques, sur des plateformes

telles que SNAP [1] et Networkrepository [2]. Cependant, même si aujourd'hui il y a plus de données réelles disponibles, leur complexité rend leur étude quelque peu difficile. (Humski et al., 2018) attribuent la complexité de l'étude des données réelles à la difficulté d'accès surtout pour des raisons de confidentialité, et la nature des données qui sont limitées, aléatoires, et souvent insuffisantes pour la validation de certaines approches ou métriques à cause de leur incertitude et leur distributions déséquilibrées.

Pour nos travaux basés sur l'étude des données d'Orange Money, nous sommes confrontés à deux défis majeurs concernant les données réelles : le premier est le problème de confidentialité, tandis que le second est le manque de vérité terrain :

— Par nature, les réseaux sociaux renferment des données à caractère personnel à des degrés variables. Une donnée à caractère personnel est « toute information se rapportant à une personne physique identifiée ou identifiable ». Les données d'Orange Money à titre d'exemple renferment toutes les opérations effectuées par un utilisateur, ses achats, sa balance (le montant d'argent disponible sur le compte à un moment donné), etc. Outre l'aspect éthique et la politique interne stricte des organisations pour protéger la confidentialité des utilisateurs, l'accès à ce type de données répond à des contraintes réglementaires en raison de leur nature sensible. Plus particulièrement, depuis 2018, la General Data Protection Regulation (RGPD) (General Data Protection Regulation) [3] contrôle le traitement des données personnelles sur le territoire de l'Union Européenne.

— La vérité terrain : Une vérité terrain aussi appelée « gold standard » désigne les données et les faits concrets observés sur le « terrain ». En d'autres termes, la vérité terrain se réfère à une forme de connaissance construite à partir de

1. http ://snap.stanford.edu/data/index.html
2. https ://networkrepository.com/networks.php
3. https ://gdpr-info.eu/

l'expérience directe et l'observation empirique de la réalité des réseaux étudiés. Pour les réseaux réels, cette vérité terrain n'est pas toujours disponible. Dans ce cas, la pratique standard consiste à traiter les caractéristiques disponibles des nœuds, également appelées méta-données, pour générer la vérité terrain. Considérons l'exemple du réseau du *club de karaté de Zachary* largement utilisé dans l'état de l'art (Zachary, 1977). La vérité terrain de ce réseau est basée sur des méta-données décrivant la tendance politique de chacun de ses membres. Cependant, la corrélation entre ces méta-données et la vérité terrain est susceptible d'être discutée (Peel et al., 2017). Par conséquent, la vérité terrain des réseaux réels ne peut pas être considérée en tant que connaissance absolue ou définitive, car elle est souvent influencée par les perceptions et les biais de celui qui la construit et des méta-données observées.

Compte tenu de ces aspects et de la sensibilité des données étudiées, pour les travaux de cette thèse nous avons décidé de travailler sur un graphe de transactions synthétiques. Les données simulées sont « des données autosuffisantes ayant pour but d'avoir des propriétés statistiques similaires à celles des données originales » (Lopez-Rojas et al., 2016). Elles ne présentent plus de problèmes de sécurité ou de confidentialité et devraient émuler le comportement des clients du service. Pour les réseaux synthétiques, la vérité terrain est également fournie manuellement en fonction du processus de génération du réseau (Chakraborty et al., 2018). Cela signifie que les données de vérité terrain sont fournis lors de la création du réseau synthétique et correspondent à des caractéristiques fixées au préalable.

La simulation est également utile pour évaluer les performances des algorithmes testés, avant de les appliquer à des données réelles. Dans l'état de l'art, plusieurs travaux ont proposé des générateurs de données synthétiques dans différents contextes. Certains de ces modèles se basent sur des caractéristiques démographiques ou géographiques des individus étudiés, d'autres considèrent la dynamique structurelle du réseau, et d'autres enfin associent divers attributs aux éléments du réseau et cal-

culent la probabilité de l'existence d'une liaison en fonction de la similarité de ces attributs. Des travaux ont été également réalisés dans le domaine des transactions financières (Lopez-Rojas et al., 2013, 2016).

On rappelle que les travaux de cette thèse se basent sur la détection de communautés à partir des transactions Orange Money. Dans ce but, notre équipe de recherche chez Orange a développé un générateur de transactions synthétiques basé sur les usages réels du service Orange Money. L'objectif de ce générateur est de fournir une vérité terrain pour l'application de la détection de communautés. Notre objectif est de disposer d'un grand ensemble de données simulées avec des dizaines de millions d'utilisateurs et des centaines de millions de transactions. Ce générateur nous a permis de créer des réseaux de transactions nous permettant de d'effectuer les différents tests nécessaires avant l'application sur les données réelles. Comme le montre la figure 2.5 ce générateur se base sur deux étapes : la génération des communautés et la génération des transactions.

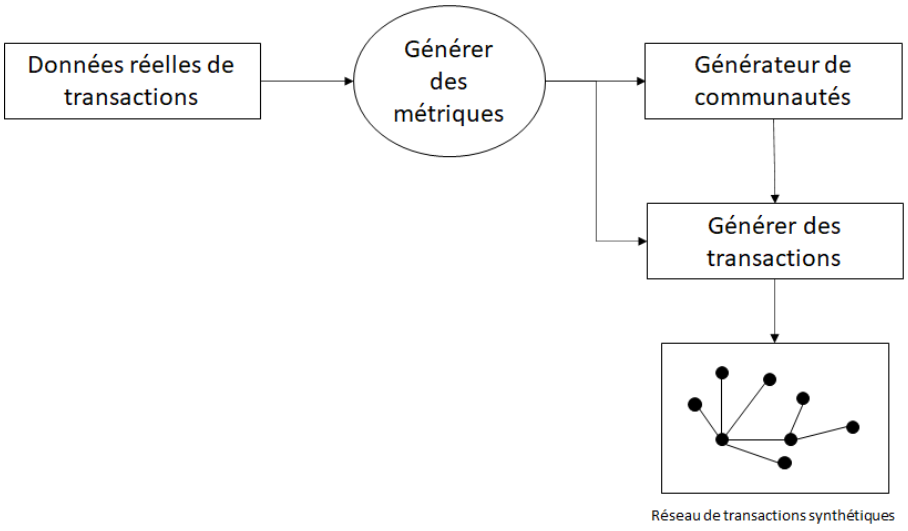


FIGURE 2.5 – Les étapes du simulateur des données de transactions.

Le générateur de communautés : La génération des communautés consiste à générer des communautés aléatoires en se basant sur des paramètres d'entrée. Ces paramètres incluent le nombre de nœuds requis, la taille moyenne et maximale des

communautés, le modèle désigné, le type de communautés générées (chevauchantes ou disjointes), etc. Un modèle décrit un type prédéfini de communautés représentant un concept social connu telles que les familles ou les tontines (une association de personnes qui se connaissent qui se partagent leur épargne pour financer ensemble des projets personnels ou collectifs [1]). Les communautés ainsi générées doivent refléter la réalité de l'utilisation habituelle du service Orange Money. Les nœuds représentent les différentes entités Orange Money : des personnes, des marchands, et des distributeurs, et les liens représentent les relations entre ces différentes entités. Les communautés ainsi construites serviront de base pour le générateur de transactions.

Le générateur de transactions : L'objectif de ce simulateur est de générer des transactions simulant les échanges entre les utilisateurs Orange Money. Comme pour l'outil de génération de communautés, les transactions doivent être représentatives des comportements habituels du service financier. Ceci inclut les répartitions entre les différents types de transactions selon les différents types d'utilisateurs, les répartitions des utilisations selon les jours et les heures de la semaine conformément aux règles de seuils et de balance, etc. Ces critères sont configurables par extraction de données issues des transactions réelles. Le générateur de transactions prend en entrée les communautés issues du simulateur de communautés, ainsi qu'un ensemble de paramètres tels que le nombre maximal de transactions et le nombre moyen de transactions par semaine, la date de début, la configuration des contraintes de répartition, etc. Afin de simuler la réalité, des singularités telles que des liens aléatoires ou imprévus ont été ajoutées au modèle pour refléter des occurrences imprévisibles dans les données.

2.6 Conclusion

Dans ce chapitre, nous avons présenté les propriétés mathématiques des réseaux et des graphes, en mettant en avant leurs liens étroits. Nous avons vu comment les

1. http ://www.tontine8.com/

graphes peuvent être utilisés pour modéliser les relations entre les différents éléments d'un système sous forme de nœuds et d'arêtes. Bien que la différence entre ces deux concepts soit assez subtile, dans ce manuscrit nous allons utiliser les termes réseau et graphe de manière indifférenciée pour désigner le réseau des transactions d'Orange Money. Au cours de ce chapitre, nous avons également présenté des métriques sur les nœuds, qui permettent de quantifier et de caractériser l'importance d'un nœud dans un graphe. Dans les chapitres suivants, nous allons nous concentrer sur la détection de communautés dans les graphes transactionnels. Nous examinerons les différentes méthodes de détection de communautés pouvant être utilisées afin d'extraire des informations utiles sur les relations entre les éléments.

Chapitre 3

La détection des communautés chevauchantes

La détection de communautés fait partie des principales applications de l'analyse des réseaux. Elle vise à mettre en évidence des groupes densément liés. Dans ce chapitre, nous allons donner des définitions de la notion de communautés, ainsi que les différents types rencontrés. Nous allons également examiner les algorithmes de détections de communautés les plus couramment utilisés dans la littérature, et qui seront utilisés pour les travaux à venir.

Sommaire

3.1 Introduction

Les réseaux offrent un outil exceptionnel pour l'analyse des systèmes complexes d'objets en interaction. En observant les réseaux du monde réel, on peut voir qu'ils se décomposent naturellement en modules densément connectés, appelés communautés. La détection de communautés dans les réseaux a gagné en importance ces dernières années en raison de son utilité pour comprendre les relations et les structures des systèmes les plus complexes. Elle permet de découvrir des groupes de nœuds qui ont des liens plus forts entre eux que vers l'extérieur du groupe. Les communautés peuvent fournir des informations importantes sur les relations et les caractéristiques des nœuds dans un réseau. Dans les réseaux sociaux à titre d'exemple, les communautés peuvent révéler des groupes de personnes qui ont des intérêts semblables. Dans les réseaux de transport, les communautés peuvent révéler des groupes de villes ou de régions qui sont connectées par des liens de transport importants. Ainsi, les communautés dans les réseaux peuvent être utilisées pour mieux comprendre la structure d'un système et peuvent fournir des informations utiles pour la prise de décision dans divers domaines d'application. La détection de communautés est donc utilisée dans de nombreux domaines

Cependant, malgré les progrès réalisés dans ce domaine, la détection des communautés reste une tâche complexe en raison de l'absence de définition universelle de ce qu'est une communauté. Chaque approche existante définit sa propre notion d'une communauté liée au contexte étudié (Peel et al., 2017; Rossetti et al., 2016). Ceci a permis d'avoir des approches diversifiés du problème, résultant en une multitude d'approches et d'algorithmes visant à percevoir ces structures dans toutes leurs formes : disjointes, chevauchantes, dynamiques, etc. Chacune de ces méthodes a ses propres avantages et inconvénients en fonction de l'application et des données.

Au cours de ce chapitre, nous allons présenter un état de l'art des travaux effectués dans le domaine de détection de communautés aussi bien disjointes que chevau-

chantes. Nous allons commencer par donner quelques définitions générales de la structure communautaire. Par la suite, nous allons aborder les approches de détection de communautés disjointes et chevauchantes les plus notables dans l'état de l'art, tout en étant attentifs à leurs complexités, étant donné la taille des données dont nous disposons.

3.2 La détection de communautés

Le problème de détection de communautés représente un des axes fondamentaux de l'analyse des réseaux et des graphes. Afin de bien percevoir le problème de détection de communautés, une étape fondamentale consiste à comprendre ce qu'est une communauté.

3.2.1 Qu'est-ce qu'une communauté ?

Les réseaux sociaux en tant que systèmes complexes par nature tendent à avoir une organisation non aléatoire. En effet, les observations des premières études effectuées dans les travaux de sociologie ont mis en évidence l'existence de certaines disparités et certains modèles spéciaux (White, 1961). Parmi l'une de ces caractéristiques des réseaux du monde réel, on retrouve les structures communautaires.

La formation des communautés dans les réseaux peut être due à différents facteurs. L'un d'entre eux est l'**homophilie**, qui se réfère au phénomène par lequel les éléments ont tendance à interagir et à se lier avec des éléments similaires à eux. Par exemple, les personnes qui ont les mêmes opinions politiques peuvent s'associer pour former une communauté et les protéines qui ont des fonctions similaires peuvent s'assembler pour former des complexes protéiques. Une autre raison qui peut contribuer à la formation des communautés est l'**influence**. Il s'agit de l'impact que les éléments du réseau ont les uns sur les autres lorsqu'ils interagissent. Par exemple, les individus peuvent être influencés par les opinions, les comportements

et les idées des personnes avec lesquelles ils interagissent pour adopter des idées similaires et les protéines qui interagissent peuvent se modifier mutuellement pour adopter une fonction similaire.

De manière intuitive, on peut définir une communauté comme un groupe d'individus ayant considérablement plus de relations entre eux par rapport aux individus des autres groupes. Ces communautés représentent des entités partageant certaines propriétés ou jouant des rôles similaires et permettent d'étudier les caractéristiques des réseaux (Chakraborty et al., 2017).

Bien que cette thématique ait été largement étudiée au cours de la dernière décennie, il n'existe encore aucune définition ou formule universelle d'une communauté. Dans un graphe, un consensus stipule qu'une communauté représente « *un groupe de sommets (un sous-graphe) plus densément reliés entre eux qu'avec le reste des nœuds du graphe* » (Mittal and Bhatia, 2020). Selon (Newman, 2006), une communauté est définie en tant que « *groupe de nœuds reliés avec une densité d'arêtes supérieure à la moyenne* ». Pour cette raison, le problème de détection de communauté est souvent assimilé à un problème de partitionnement de graphe (Schaub et al., 2017). La détection de communautés peut servir à différents motifs :

- Comprendre la structure d'un graphe : En identifiant les communautés au sein d'un graphe, on peut mieux comprendre la structure globale du graphe et la façon dont il est organisé. Cela peut être utile pour identifier des modèles et des tendances dans les données.
- Identifier les groupes ayant des caractéristiques communes : Dans de nombreux cas, les nœuds d'une communauté possèdent des caractéristiques communes. En identifiant ces communautés, on peut étudier ces caractéristiques et comprendre comment elles affectent la structure globale du graphe.
- Améliorer la précision des prédictions : Dans certains cas, les caractéristiques d'un nœud au sein d'une communauté sont plus prévisibles en fonction des caractéristiques des autres nœuds de sa communauté. En identifiant les com-

munautés, on peut améliorer la précision des prédictions faites sur les nœuds individuels en considérant les caractéristiques de la communauté dans son ensemble.

Dans l'ensemble, la détection des communautés est une application importante pour l'étude des graphes car elle permet aux chercheurs de comprendre et d'analyser plus en détail la structure et les caractéristiques du graphe. En fonction du contexte, les communautés correspondent à des entités comportementales ou fonctionnelles au sein du réseau. En considérant les réseaux sociaux, par exemple, les communautés peuvent référer à des groupes d'amis, des familles, des voisins géographiques, des collègues de travail, etc. En biologie, les communautés de gènes dans les réseaux de protéines sont étudiées afin d'estimer le pronostic des patients (Wu et al., 2011). Dans le contexte du commerce électronique, les communautés illustrent des groupes ayant des intérêts communs pertinents dans les stratégies de marketing ciblé. Dans le contexte des réseaux sociaux mobiles, les communautés chevauchantes ont été étudiées dans le réseau des appels téléphoniques et des textes échangés entre les utilisateurs (Kim and Kim, 2014).

Ainsi, chaque contexte nécessite une définition précise de la communauté, et examine le problème de manière différente. Pour ceci, il existe plusieurs formulations pour caractériser le concept de communauté. Certains problèmes s'intéressent à une caractérisation globale des communautés dans le réseau, d'autres se concentrent sur une définition sur le plan local.

Définitions globales :

Les définitions globales des communautés étudient la structure globale du réseau. Une définition triviale des communautés repose sur la similarité entre les nœuds. Les distances entre chaque couple de nœuds sont calculées, et ceux qui sont les plus similaires sont regroupés ensemble. Une autre définition globale examine le nombre de liens entre les communautés (cut-size), ou le cut-size normalisé (Shi and Malik,

2000).

Définitions locales :

Les communautés sont des parties du graphe faiblement connectées avec le reste des nœuds (Fortunato, 2010). Elles peuvent être considérées comme étant des entités indépendantes. Les définitions locales des communautés s'intéressent à ces groupes et à leurs voisinages immédiats. On peut trouver plusieurs définitions locales des communautés, telles que les cliques, n-cliques, n-clubs, n-clans, k-cores, etc. (Wasserman and Faust, 1994).

Dans la littérature, plusieurs types de communautés existent. Elles peuvent dépendre du contexte ou des données traitées. (Mittal and Bhatia, 2020) définissent trois principaux types de communautés :

- Les communautés dynamiques : Ce sont des communautés qui changent avec le temps. Elles peuvent s'élargir ou rétrécir ;

- Les communautés disjointes : Ce sont des communautés qui ne possèdent pas de nœud en commun ;

- Les communautés chevauchantes : Ces communautés partagent un ou plusieurs individus.

Dans le reste de ce chapitre, nous allons considérer uniquement les communautés disjointes et chevauchantes, car elles sont les plus étudiées dans l'état de l'art, et les plus adaptées à notre problématique. Dans ce qui suit, nous allons formuler les problèmes de la détection des communautés disjointes et chevauchantes, et nous allons proposer quelques exemples de travaux associés à ces structures.

3.2.2 Définition du problème

3.2.2.1 Les communautés disjointes

Considérons un réseau présenté par un graphe $\mathcal{G} = (\mathcal{V}, E)$, l'objectif d'un algorithme de détection de communautés est d'établir une partition $P_{dis} = \{C_1, C_2, \ldots, C_n\}$ de nœuds de sorte à satisfaire les conditions étudiées plus tôt, c'est-à-dire qu'il doit y avoir beaucoup plus d'arêtes à l'intérieur de ces communautés que d'arêtes entre les communautés. Dans ce contexte, chaque nœud est attribué à une communauté unique. Les communautés disjointes vérifient :

$$\forall i \neq j, C_i \cap C_j = \varnothing$$

3.2.2.2 Les communautés chevauchantes

Définition 5 *Une structure de communautés chevauchantes P_{chev} peut être définie comme une couverture de V en n communautés $P_{chev} = \{C_1, C_2, \ldots, C_n\}$ où un nœud v_i peut participer à une ou plusieurs communautés C_k.*

Dans les prochaines sections, nous allons présenter des approches de détection de communautés disjointes et chevauchantes présentes dans l'état de l'art. Cependant, il n'existe pas une taxonomie unique pour la classification des méthodes de détection de communautés. Plusieurs catégorisations se présentent basées sur différentes fonctions de qualité, attributs des nœuds et des arêtes, parcours de graphe, etc. A titre d'exemple, dans (Fortunato, 2010), les méthodes de détection de communautés sont regroupées en méthodes de segmentation traditionnelles, méthodes divisives, méthodes basées sur la modularité, méthodes spectrales, méthodes dynamiques, et méthodes basées sur l'inférence statistique. En revanche, d'après (Bohlin et al., 2014), les méthodes de détection sont associées à trois modèles : le modèle nul, le modèle de blocs, et le modèle de flux. Selon (Schaub et al., 2017), les différentes approches sont synthétisées en quatre catégories : approches basées sur la coupe (*cut-based*), approches basées sur la densité interne, approches basées sur l'équi-

valence stochastique (stochastic equivalent), et approches dynamiques. Finalement, dans (Ahajjam and Badir, 2022) une classification plus détaillée des méthodes de détection a été établie en : méthodes basées sur la marche aléatoire (*random walks*), méthodes basées sur le partitionnement du graphe, méthodes hiérarchiques, méthodes basées sur la segmentation partitionnelle, méthodes basées sur la classification spectrale, méthodes basées sur la modularité, et méthodes exploitant les nœuds centraux. D'un autre côté, (Xie et al., 2013) regroupe les méthodes de détection de communautés en méthodes de percolation de cliques, méthodes de partitionnement de liens, méthodes d'expansion et d'optimisations locales, méthodes de détection approximatives (*fuzzy detection*) et méthodes basées sur les agents et méthodes dynamiques.

Comme on peut le constater à partir des différentes classifications présentées, il n'existe pas une manière unique pour présenter les méthodes de détection de communautés. Dans ce qui suit, nous allons adopter une classification basée sur une synthèse des différentes taxonomies rencontrées dans l'état de l'art.

3.3 Approches pour la détection de communautés disjointes

Les approches de détection de communautés disjointes peuvent être regroupées selon leurs démarches en approches séparatives, approches d'optimisation de modularité, approches de propagation de label, et approches dynamiques. Il est évident que notre liste de méthodes n'est pas exhaustive et ne peut contenir toutes les méthodes existantes pour la détection de communautés. Néanmoins, nous allons présenter les méthodes les plus connues, et les plus utilisées en pratique.

3.3.1 Approches séparatives

Les approches séparatives de détection de communautés sont des méthodes qui se basent sur la définition d'une fonction de coût pour mesurer la qualité de la séparation des communautés, et qui par la suite utilisent des algorithmes d'optimisation pour minimiser cette fonction de coût et trouver une partition des nœuds qui optimise la qualité de la séparation des communautés.

Girvan and Newman, 2002

(Girvan and Newman, 2002) ont proposé une méthode divisive GN pour la détection des communautés basée sur la suppression progressive des arêtes du réseau. Cette méthode exploite la centralité d'intermédiarité qui exprime l'influence des arêtes sur le passage du flux d'information dans le réseau. En effet, si un réseau contient des groupes denses, ces derniers sont liés par un nombre limité d'arêtes ayant des centralités d'intermédiarité élevées. En enlevant ces arêtes, les communautés seront séparées. Cette structure est apparente sur la figure 3.1. Pour tester la qualité des divisions obtenues, les auteurs ont introduit une nouvelle notion, à savoir la modularité. La modularité représente un modèle efficace permettant de quantifier la qualité d'une communauté en se basant sur la différence du nombre d'arêtes intra-communautaires observées et le nombre d'arêtes intra-communautaires attendues dans un réseau aléatoire. Cette métrique sera étudiée davantage dans le chapitre suivant.

En 2016, (Moon et al., 2016) ont adapté la méthode *GN* en utilisant un algorithme de parallélisation basé sur le modèle MapReduce [1]. Le calcul parallèle a permis à l'algorithme de découvrir des communautés dans de grands réseaux.

1. MapReduce est un paradigme de programmation qui permet une extensibilité massive sur des centaines ou des milliers de serveurs (https ://www.ibm.com/fr-fr/topics/mapreduce).

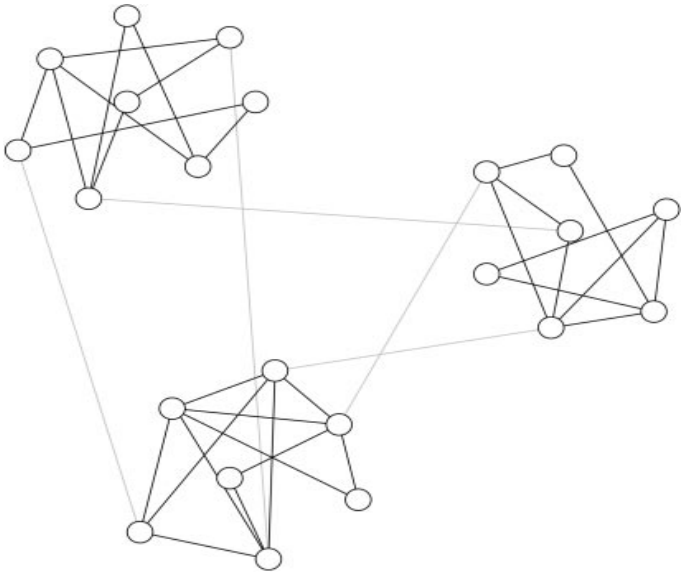


FIGURE 3.1 – Représentation d'une structure communautaire par la méthode *GN* (Girvan and Newman, 2002).

3.3.2 Approches d'optimisation de la modularité

Dans (Newman and Girvan, 2004b), il a été démontré que de meilleures partitions de réseau sont corrélées à une valeur de modularité élevée. Les méthodes de maximisation de la modularité sont basées sur cette théorie pour construire des communautés. Cette classe de méthodes a été largement utilisée dans l'état de l'art grâce à la versatilité de la modularité (Wadhwa and Bhatia, 2012).

Clauset-Newman-Moore, 2004

Le concept principal de la méthode de (Clauset et al., 2004) CNM repose sur une approche hiérarchique agglomérative, visant à joindre progressivement les nœuds afin d'optimiser le gain de modularité ΔQ. Le gain de la modularité proposée est donné par :

$$\Delta Q_{c_i c_j} = \frac{1}{2|E|} - \frac{deg(v_i)deg(v_j)}{4|E|^2}, \tag{3.1}$$

L'algorithme commence avec n nœuds représentant n communautés individuelles, et se termine lorsque tous les sommets sont regroupés dans la même communauté. Plutôt que de mettre à jour la matrice d'adjacence et par la suite calculer le gain de modularité, les auteurs ont mis en place une structure où les valeurs de ce gain sont calculées directement. Les auteurs reprochaient à la méthode *GN* sa consommation excessive de temps de calcul et de mémoire en stockant les valeurs nulles de la matrice d'adjacence. Cependant, un des inconvénients de la méthode CNM est sa tendance à diviser le réseau en deux grandes communautés.

Blondel *et al.*, 2008

La méthode (*Louvain*) de (Blondel et al., 2008) repose également sur une agrégation itérative des communautés afin de réaliser une maximisation de la modularité. Elle est divisée en deux phases. Au cours de la première phase, les communautés sont formés en attribuant chaque nœud à l'une des communautés de ses voisins qui engendre le gain de modularité (3.1) positif maximal. Contrairement à la méthode CNM, un nœud peut être revisité plusieurs fois et assigné à une nouvelle communauté si cela engendre un gain positif de la modularité.

À la deuxième phase, l'algorithme construit un graphe dont les sommets sont les communautés trouvées à la phase précédente, et les arêtes sont les liens entre ces communautés. Les étapes de la première phase sont ainsi appliquées sur le nouveau graphe. Ces deux phases sont répétées, jusqu'à ce qu'aucune augmentation de modularité ne soit possible. Le résultat est alors donné par les dernières communautés trouvées comme le montre la figure 3.2. La méthode *Louvain* fait parti des meilleures méthodes de détection de communautés disjointes en terme de complexité.

3.3.3 Approche de propagation de label

Les méthodes de propagation de label sont une famille de méthodes qui reposent sur le principe de propagation de l'information à travers les nœuds d'un réseau. Elles

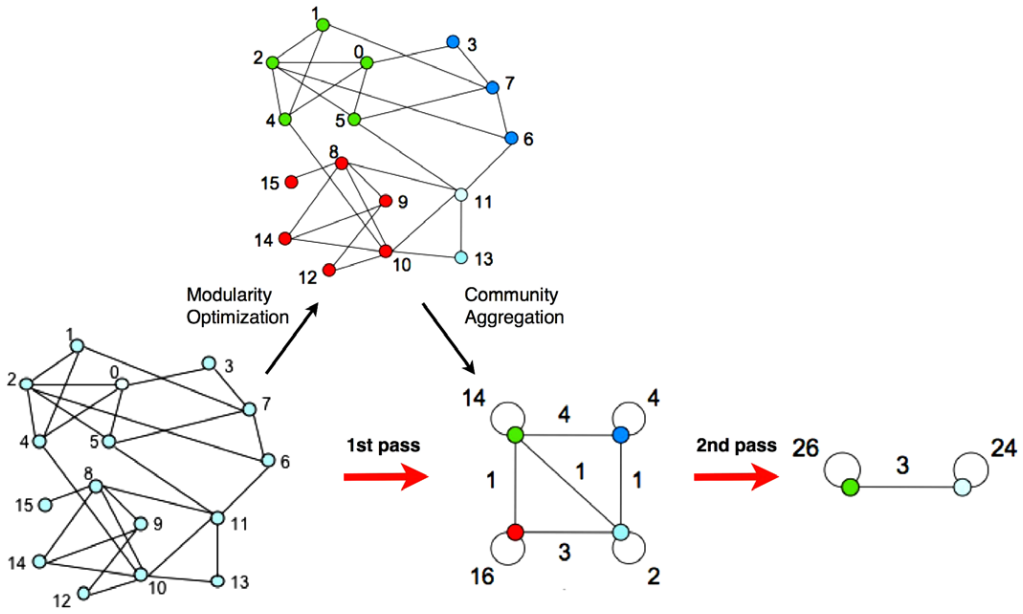


FIGURE 3.2 – Étapes de la méthode de détection de communautés *Louvain* (Blondel et al., 2008).

utilisent un ensemble de labels ou d'étiquettes pour marquer les nœuds et propagent ces étiquettes en utilisant des règles définies. Les méthodes de propagation de label ont pour objectif de regrouper des nœuds qui ont des étiquettes similaires. En raison de la nature dense des liens à l'intérieur d'une communauté, et leur faiblesse à l'extérieur, il est admis qu'un message émis par un nœud et retransmis par ses voisins est plus susceptible de rester dans la communauté du nœud source que de se diffuser aux autres communautés (Kanawati, 2013). Un des avantages des méthodes de propagation de label est le temps de calcul par rapport aux autres méthodes. Ces méthodes ont aussi prouvé leur capacité à gérer des réseaux réels de grande dimension.

Raghavan *et al.*, 2007

La méthode de propagation de label détaillée dans la référence (Raghavan et al., 2007) se base sur un algorithme itératif. À chaque itération, un nœud envoie un label à son voisinage direct et reçoit en contrepartie celui des nœuds qui l'entourent.

Chaque nœud détermine le label majoritaire qu'il adopte pour l'itération suivante en utilisant un principe de vote. Notons l_{v_i} le label d'un nœud v_i, et $N^l(v_i)$ le voisinage de v_i portant le label l, l'assignation du label est donnée par la formule suivant :

$$l_{v_i} = (\arg\max)_l |N^l(v_i)|$$

Un état d'équilibre est atteint lorsque chaque nœud a son label égal à celui de tous ses voisins. À la fin de l'algorithme, les nœuds ayant le même label forment les communautés.

Zong-Wen *et al.*, 2014

Les méthodes de propagation de label présentent l'inconvénient d'être instables. Ceci s'explique par le choix aléatoire du label que prend un nœud si plusieurs labels majoritaires se trouvent dans son voisinage. Pour ceci, beaucoup de travaux ont cherché à contourner le choix aléatoire du label. (Zong-Wen et al., 2014) ont mis en place un algorithme de propagation de label guidé par consensus, qu'ils ont appelé *Label Propagation Algorithm with Coverage Weighting (LPAcw)*. La méthode proposée commence par l'application de l'algorithme de propagation de label de manière répétitive afin d'obtenir plusieurs ensembles de communautés. Un nouveau graphe pondéré est alors généré, où les arêtes ayant des poids plus importants représentent les paires de nœuds fréquemment placés dans la même communauté. L'assignation des nouveaux labels est par la suite effectuée, en considérant les poids des arêtes du graphe.

3.3.4 Approches de processus dynamique

Les approches dynamiques pour la détection de communautés se concentrent sur la manière dont les communautés évoluent dans le temps. Ces approches reposent sur l'idée de surveiller en permanence le réseau au fur et à mesure de ses changements et de suivre l'évolution des communautés.

Pons and Latapy, 2005

La méthode *Walktrap* proposée par (Pons and Latapy, 2005) est une méthode de détection de communautés basée sur le principe des marches aléatoires (*random walks*). La marche aléatoire est déterminée par les valeurs de la matrice de transition P construite comme suit : à chaque étape, la probabilité de se déplacer d'un nœud v_i vers un nœud v_j d'une marche de longueur t est donnée par :

$$P^t_{v_i v_j} = \frac{A_{v_i v_j}}{deg(v_i)}$$

où A représente la matrice d'adjacence. Afin de construire les communautés, les auteurs ont proposé une nouvelle distance entre les nœuds donnée par par :

$$r_{v_i v_j} = \sqrt{\sum_{v_k \in \mathcal{V}} \frac{(P^t_{v_i v_k} - P^t_{v_j v_k})^2}{deg(v_k)}}$$

La distance est faible lorsque les nœuds appartiennent à la même communauté, elle est importante dans le cas contraire. Ainsi, en commençant par n communautés comportant un nœud unique, à chaque itération de l'algorithme, une nouvelle partition est construite selon une approche agglomérative en calculant les distances entre les nœuds. Au bout de $n-1$ itérations, l'algorithme génère une séquence de communautés encastrées comme le montre le dendrogramme de la figure 3.3. La partition ayant la modularité la plus importante est finalement sélectionnée.

3.3.5 Tableau récapitulatif

Dans ce qui précède, nous avons présenté un tour d'horizon des principales méthodes de détection de communautés disjointes en les regroupant selon les différentes approches proposées. Certaines se basent sur la propagation de label, d'autres sur l'optimisation de la modularité, et elles ont chacune leurs avantages et inconvénients. Nous proposons le tableau 3.1 permettant de récapituler les méthodes citées

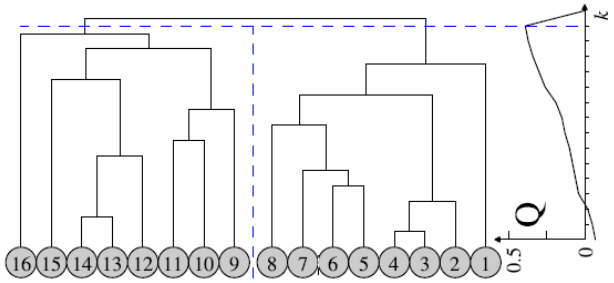


FIGURE 3.3 – Représentation d'un arbre hiérarchique des communautés généré par la méthode *Walktrap* (Pons and Latapy, 2005).

et leurs complexités temporelles estimées pour un graphe peu dense ($m \approx n$) où n est le nombre de nœuds et m le nombre d'arêtes.

TABLEAU 3.1 – Tableau récapitulatif des méthodes de détections de communautés de l'état de l'art.

Approches	Référence	Méthode	Complexité
Approches séparatives	(Girvan and Newman, 2002)	GN	$\mathcal{O}(nm^3)$
Approches d'optimisation de la modularité	(Clauset et al., 2004)	CNM	$\mathcal{O}(nlogn^2)$
	(Blondel et al., 2008)	louvain	$\mathcal{O}(m)$
Approches de propagation de label	(Raghavan et al., 2007)	label propagation	$\mathcal{O}(m)$
	(Zong-Wen et al., 2014)	*LPAcw*	
Approches de processus dynamique	(Pons and Latapy, 2005)	walktrap	$\mathcal{O}(n^3)$

3.4 Approches de détection de communautés chevauchantes

Nous avons choisi de présenter les méthodes de communautés chevauchantes phares de l'état de l'art selon la classification en approches de percolation de cliques, approches centrées nœuds, approches de propagation de label, et approches basées sur l'optimisation globale et locale. Il est à noter que plusieurs des méthodes chevau-

chantes sont basées sur des améliorations et des adaptations des méthodes conçues pour les communautés disjointes. On rappelle que dans le cadre de la détection des communautés chevauchantes les nœuds sont capables d'appartenir à plusieurs communautés concurremment.

3.4.1 Approches de percolation de cliques

La méthode de percolation de cliques (CPM) consiste à former des cliques ou des sous-ensembles complets de sommets de manière répétée, en commençant par des cliques de tailles minimales, jusqu'à ce qu'une seule grande clique qui traverse le réseau entier soit formée.

Dans ce contexte, les auteurs de (Palla et al., 2005) fondent leur méthode *CFinder* sur le principe qu'une communauté est composée de plusieurs sous-graphes complets qui se partagent un grand nombre de leurs nœuds. Ils définissent une communauté comme étant l'union de toutes les k-cliques adjacentes (c'est à dire qui se partagent $k-1$ membres). La première étape de ce processus est donc l'identification de toutes les cliques de taille k (où k est préalablement défini). Par la suite, un nouveau graphe est construit de telle sorte que ses sommets représentent les cliques précédemment identifiées. Dans ce graphe, deux k-cliques sont connectées si elles sont adjacentes. Finalement, les communautés finales sont formées par les ensembles des cliques connectées. Empiriquement, de petites valeurs de k (comprises entre 3 et 6) ont donné de bons résultats (Palla et al., 2005; Lancichinetti and Fortunato, 2009). Néanmoins, ces valeurs ne sont pas performantes avec les grands réseaux.

(Farkas et al., 2007) ont proposé une variation *CPMw* de la méthode de percolation de cliques pour les graphes pondérés en définissant un seuil d'intensité pour les cliques. L'intensité d'une clique est définie comme étant la moyenne géométrique des poids de ses liens. Pour cette méthode, seules les k-cliques ayant une intensité supérieure à un seuil fixé seront incluses dans les communautés formées.

3.4.2 Approches centrées sur les nœuds

Les approches centrées sur les nœuds pour la détection de communautés se basent sur des caractéristiques individuelles des nœuds pour identifier les communautés.

Chen *et al.*, 2010

La méthode *Wcommunity* (Chen et al., 2010) est une méthode de détection de communautés chevauchantes qui se base sur la notion de force des nœuds et prend en entrée un graphe pondéré. La force d'un nœud est une mesure de sa connectivité dans un réseau pondéré. Plus précisément, c'est la somme des poids des arêtes qui sont connectées à ce nœud. Cela signifie que les nœuds ayant une force élevée sont ceux qui sont connectés à de nombreux autres nœuds avec des liens forts (poids élevés).

La force d'un nœud v_i est donnée par :

$$k_{v_i} = \sum_{v_j \in \mathcal{V}} w_{v_i v_j}. \tag{3.2}$$

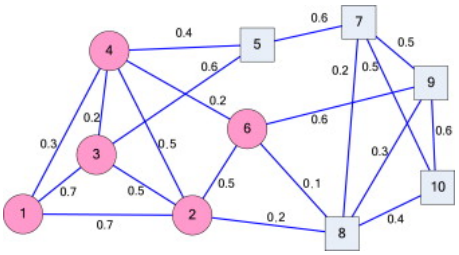
où $w_{v_i v_j}$ représente le poids de l'arête liant les deux nœuds v_i et v_j. Dans le cas où les nœuds v_i et v_j ne sont pas connectés : $w_{v_i v_j} = 0$.

La méthode *Wcommunity* est composée de deux étapes qui vont se répéter jusqu'à la découverte de toutes les communautés. Ces étapes consistent à : trouver des communautés initiales et les étendre. Pour la recherche des communautés initiales, les forces des nœuds sont calculées par l'équation (3.2). L'ensemble des nœuds ayant la valeur maximale de force et leurs voisins constituent la communauté initiale. Pour tous ces nœuds, leurs degrés d'appartenance doivent être supérieurs au seuil fixé (0.5), dans le cas contraire le nœud est retiré. Le degré d'appartenance d'un nœud v_i à une communauté C est défini par :

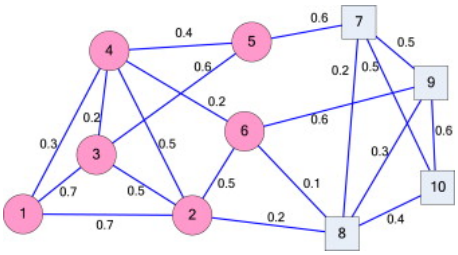
$$B^C(v_i) = \frac{\sum_{v_j \in C} w_{v_i v_j}}{k_{v_i}}. \tag{3.3}$$

Une fois la communauté initiale définie, la deuxième étape consiste à parcourir tous

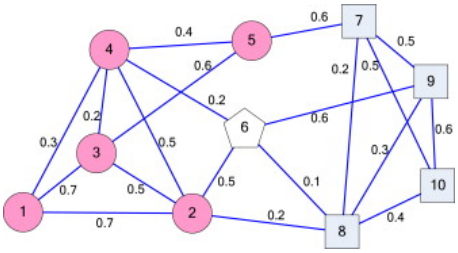
les voisins de la communauté initiale et les joindre à cette dernière si leurs degrés d'appartenance sont supérieurs au seuil et si leurs ajouts augmentent la modularité. Ce processus à deux étapes (recherche de communauté initiale et expansion) est ainsi répété jusqu'à ce que toutes les communautés du graphe soient trouvées comme le montre la figure 3.4.



(a) Définition de la communauté initiale (1,2,3,4,6).



(b) Expansion de la communauté initiale (ajout du nœud 5).



(c) Le résultat final de l'algorithme.

FIGURE 3.4 – Représentation des étapes de détection de communautés d'un graphe de 10 nœuds par l'algorithme *Wcommunity* (Chen et al., 2010).

Rossetti *et al.*, 2020

L'algorithme *Angel* (Rossetti, 2020) est une version améliorée en temps de calcul et en complexité d'un algorithme précédemment proposé appelé *Demon* (Coscia et al., 2012). Cette méthode utilise une approche descendante et centrée sur les nœuds pour détecter les communautés chevauchantes. L'algorithme se déroule en deux étapes : la construction des communautés locales, et l'aggrégation de ces dernières afin d'obtenir l'ensemble des communautés globales. La construction des communautés locales est effectuée en créant des sous-réseaux appelés réseaux « *EgoMinusEgo* ». En ef-

fet, pour chaque nœud du graphe, le réseau « *EgoMinusEgo* » est formé par son réseau de voisinage (Ego Network) incluant l'ensemble de ses voisins et leurs liens duquel on retire le nœud même. Un algorithme de propagation de label (Raghavan et al., 2007) est par la suite appliqué sur ce sous-graphe permettant de générer un ensemble de communautés locales. Le nœud est retiré de son réseau de voisinage afin de ne pas biaiser le processus de détection de communautés en raison de sa liaison à tous les nœuds de ce sous-graphe. Ce processus de recherche de communautés locales est appliqué à l'ensemble des nœuds du réseau. La deuxième étape de l'algorithme *Angel* consiste à fusionner les communautés locales pour former des communautés plus grandes en fonction de leur similarité. La similarité entre deux communautés est mesurée en fonction de la fraction de nœuds qu'elles partagent, et les communautés sont fusionnées si leur similarité dépasse un certain seuil défini.

La méthode *Angel* présente l'avantage d'une complexité généralement plus faible en comparaison avec les autres méthodes de détection de communautés chevauchantes. Elle est également déterministe, et peut être parallélisable.

3.4.3 Approches par propagation de label

Les approches de détection de communautés par propagation de label sont basées sur la propagation d'informations ou d'étiquettes sur les nœuds d'un graphe en utilisant des méthodes itératives. Ces approches peuvent identifier des communautés en analysant les étiquettes des nœuds et leur propagation à travers leurs relations.

Gregory, 2010

L'algorithme de propagation de recouvrement de communautés (Copra), (Gregory, 2010) est une variante de l'algorithme de Raghavan *et al.* (3.3.3) où les nœuds peuvent posséder plusieurs identifiants de communautés. L'algorithme *COPRA* est conçu pour détecter les communautés qui se chevauchent en utilisant un processus en deux étapes : la propagation d'étiquettes et l'agrégation des communautés sur

la base d'un vote entre voisins. Au départ, chaque nœud du graphe est étiqueté avec une valeur unique. Cette étiquette est sous la forme (C, b_t) où C représente l'identifiant de la communauté et b_t représente le coefficient d'appartenance. A une itération t de l'algorithme, pour un nœud v_i, ce coefficient est donné par :

$$b_t(C, v_i) = \frac{\sum_{v_j \in N(v_i)} b_{t-1}(C, v_j)}{|N(v_i)|} \tag{3.4}$$

De manière récursive, l'étape de propagation consiste à mettre à jour les étiquettes des nœuds en les remplaçant par l'identifiant de la communauté la plus répandue entre ses voisins. La convergence de l'algorithme est atteinte lorsque le nœud a la majorité des étiquettes de ses voisins. Les communautés sont formées par les nœuds partageant les mêmes étiquettes comme on peut le voir sur l'exemple de la figure 3.5. L'algorithme est relativement rapide et il est capable de détecter les communautés qui se chevauchent avec une grande précision.

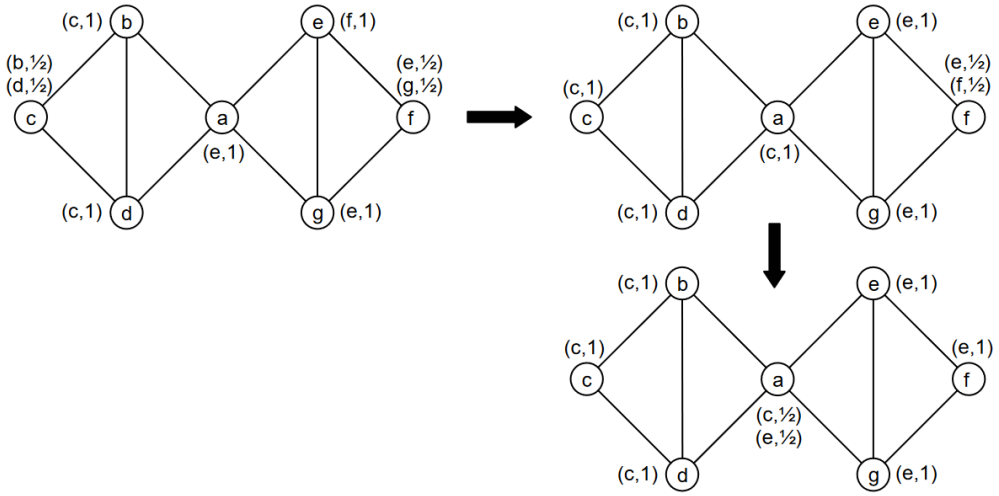


FIGURE 3.5 – Illustration sur un exemple de la méthode *Copra* (Gregory, 2010).

Xie *et al.*, 2011

L'algorithme Speaker-Listener Label Propagation (Speaker-listener Label Propagation Algorithm (SLPA)) (Xie et al., 2011) fait partie des algorithmes les plus connus et utilisés pour la détection de communautés chevauchantes. Il s'agit d'une

approche de propagation de label qui se base sur des règles pour déterminer comment les étiquettes sont propagées d'un nœud à l'autre. Contrairement à la méthode LPA (Raghavan et al., 2007), ici le nœud est autorisé à posséder plus qu'un label puisqu'il peut appartenir à plusieurs communautés à la fois. Initialement, chaque noeud commence avec une étiquette unique qui contient son identifiant. Au cours de chaque itération de l'algorithme, *SLPA* utilise une approche de dialogue pour la propagation d'étiquettes, où chaque nœud du réseau joue alternativement le rôle de récepteur d'étiquettes, ou *listener*, et d'émetteur d'étiquettes, ou *speaker*. Le nœud récepteur d'étiquettes reçoit un des labels de ses voisins selon des règles d'émission et de réception. Ce processus est répété jusqu'à ce que les étiquettes des nœuds cessent de changer ou atteignent un état d'équilibre, ou le nombre maximum prédéfini d'itérations T est atteint. Une étape de post-traitement a alors lieu permettant d'affiner les résultats en supprimant les communautés imbriquées.

Sedighpour and Bagheri, 2018

L'algorithme PASLPA (Parallel Advanced Speaker-Listener Label Propagation Algorithm) est basé sur une implémentation parallèle de l'algorithme *SLPA* (Xie et al., 2011) proposé par (Sedighpour and Bagheri, 2018). Cet algorithme a été conçu pour pallier les problèmes de scalabilité en étant capable de fonctionner sur des plateformes de traitement distribuées telles que *Spark*. Cela permet de traiter des réseaux de grandes tailles efficacement en utilisant la puissance de calcul distribuée.

Comme il s'agit d'une variante de l'algorithme *SLPA*, *PASLPA* fonctionne également de manière asynchrone : les nœuds sont mis à jour de manière non simultanée. En utilisant le système de fichiers distribués *Hadoop*[1], le graphe est divisé en plusieurs parties. L'algorithme *SLPA* est alors exécuté sur chaque partition du graphe en parallèle. Les résultats obtenus pour chaque partition sont stockés sur le système de fichiers distribués *Hadoop*. Finalement, les résultats obtenus pour chaque

1. https ://hadoop.apache.org/

partition sont fusionnés pour obtenir les communautés finales. Ces étapes de partitionnement, d'exécution et de fusion sont répétées plusieurs fois afin d'améliorer la qualité de la détection de communauté. Au cours de l'étape de post-traitement, seuls les labels ayant une probabilité d'occurrence supérieure à un seuil fixé seront retenus.

3.4.4 Approches basées sur l'optimisation globale et locale

Les approches basées sur l'optimisation globale et locale cherchent respectivement à maximiser ou minimiser une mesure de qualité globale pour la partition du réseau en communautés et à optimiser localement la qualité de la partition en maximisant ou minimisant des critères locaux à chaque nœud.

Rosvall *et al.*, 2009

La méthode proposée dans (Rosvall et al., 2009), appelée *Infomap*, fait partie des méthodes les plus connues des approches basées sur l'optimisation globale. La méthode *Infomap* utilise la théorie de l'information pour décomposer les réseaux en modules ou communautés en utilisant un processus de compression d'information pour identifier les communautés dans les réseaux à travers un processus dynamique, à savoir la marche aléatoire. Elle partitionne le réseau en communautés correspondant à des modules qui peuvent être compressés de manière optimale. En effet, regrouper les nœuds en communautés ayant des liens intra-communauté forts et des liens inter-communauté faibles devrait compresser la description de la marche aléatoire parcourant ces nœuds. Les communautés retenues sont celles qui permettent d'optimiser une fonction de qualité analogue à la modularité : la *map equation*[1]. Cette équation joue un rôle clé dans la méthode *Infomap* en permettant de quantifier la longueur de la description de la marche aléatoire.

Infomap est capable de reconnaître les réseaux aléatoires (sans communautés). Elle

1. https ://www.mapequation.org/

est également conçue pour être efficace pour les réseaux de grande taille.

Lancichinetti and Fortunato, 2011

La méthode d'optimisation locale des statistiques d'ordre (OSLOM) (Lancichinetti et al., 2011) est une méthode qui se base sur l'optimisation locale. Elle est élaborée pour détecter les communautés statistiquement significatives dans les réseaux complexes tout en prenant en compte la direction et le poids des arcs, ainsi que l'aspect hiérarchique des communautés. Les auteurs définissent la signification statistique comme la probabilité de trouver une communauté similaire dans un modèle nul ne possédant aucune structure communautaire.

La méthode *OSLOM* se déroule en trois étapes : l'initialisation, l'agrégation et le raffinement. Commençant par une division aléatoire des nœuds en communautés, à chaque itération, l'algorithme agrège les nœuds en communautés en fonction de leurs liens avec d'autres nœuds. Les nœuds sont associés à une communauté s'ils ont un nombre élevé de liens avec d'autres nœuds de cette communauté. Ce processus continue jusqu'à ce qu'aucune amélioration ne puisse être apportée. En raison de sa nature stochastique, ce processus est répété plusieurs fois pour assurer la stabilité des communautés. Les nœuds partageant les mêmes labels de communautés sont groupés, et de multiples niveaux hiérarchiques sont obtenus.

Hollocou *et al.*, 2016

L'algorithme Walkscan (Hollocou et al., 2016) est basé sur une modification de l'algorithme PageRank. Il utilise des marches aléatoires pour explorer le graphe localement, à partir de chaque nœud, et détecter des sous-graphes denses qui correspondent à des communautés locales. Plus précisément, chaque marche aléatoire part d'un nœud initial et suit une probabilité de transition qui est basée sur la similarité des nœuds dans le voisinage local. À chaque étape de la marche aléatoire, un nœud est choisi en fonction de sa similarité avec le nœud actuel, puis la probabilité de transition est mise à jour en fonction du nœud choisi. Cette procédure est

répétée pour chaque marche aléatoire jusqu'à ce qu'un certain critère d'arrêt soit atteint. L'idée derrière ce processus est le fait que les nœuds qui sont visités par la marche aléatoire en même temps sont plus susceptibles d'appartenir à la même communauté. Les sous-graphes locaux sont fusionnés en utilisant une heuristique de fusion qui prend en compte à la fois la similarité entre les sous-graphes et leur connectivité globale dans le réseau.

3.4.5 Approches de factorisation matricielle non négative

Yang and Leskovec, 2013

L'algorithme *Bigclam* (Cluster Affiliation Model for Big Networks) (Yang and Leskovec, 2013) est un algorithme de détection de communautés basé sur une factorisation matricielle non négative (NMF). Il est basé sur un modèle génératif probabiliste qui appréhende la structure du réseau à partir d'un modèle d'affiliation du graphe (AGM) représenté par un graphe biparti d'affiliation où chaque nœud est relié à la communauté à laquelle il appartient. F est a matrice non négative qui contient l'ensemble des poids des arêtes d'affiliations pour tous les nœuds du graphe.

Étant donné un graphe $\mathcal{G}$, l'algorithme *Bigclam* recherche le meilleur ajustement entre F et la matrice d'adjacence de $\mathcal{G}$. Il découvre ainsi les communautés en combinant les méthodes de factorisation de matrices non négatives avec la descente de gradient stochastique par blocs. L'algorithme *Bigclam* a permis de réaliser des améliorations en termes de scalabilité, et de la qualité des communautés détectées.

3.4.6 Tableau récapitulatif

Dans ce qui précède, nous avons présenté quelques méthodes de détection de communautés chevauchantes. Nous proposons le tableau 3.2 permettant de récapituler les méthodes citées ainsi que leurs complexités temporelles estimées pour un graphe peu dense ($m \approx n$) où n est le nombre de nœuds et m le nombre d'arêtes.

TABLEAU 3.2 – Tableau récapitulatif des méthodes de détections de communautés de l'état de l'art.

Approches	Référence	Méthode	Complexité
Percolation de clique	(Palla et al., 2005)	CPM	
	(Farkas et al., 2007)	CPMw	
Approche de propagation de label	(Xie et al., 2013)	SLPA	$\mathcal{O}(Tn)$
	(Gregory, 2007)	Copra	$\mathcal{O}(Tn)$
	(Sedighpour and Bagheri, 2018)	PASLPA	
Approches centrées nœuds	(Chen et al., 2010)	Wcommunity	$\mathcal{O}(n^2)$
	(Rossetti, 2020)	Angel	
Approches d'expansion et d'optimisation local	(Rosvall and Bergstrom, 2007)	Infomap	
	(Lancichinetti et al., 2011)	Oslom	$\mathcal{O}(n)$
	(Hollocou et al., 2016)	Walkscan	
Approches de factorisation matricielle non négative	(Yang and Leskovec, 2013)	Bigclam	$\mathcal{O}(n^2)$

Dans le reste de ce manuscrit, nous allons considérer uniquement la détection de communautés chevauchantes. En effet, la réalité de l'utilisation du service Orange Money révèle que les individus peuvent appartenir à plusieurs communautés, telles que leur communauté familiale, leur communauté d'amis, leur communauté d'employeurs, leur communauté de cité, etc. Ces communautés ne sont pas disjointes, mais se chevauchent souvent, et il est donc important de les prendre en compte dans l'analyse.

3.5 Application

Afin de comparer les performances des différentes méthodes de détection de communautés chevauchantes présentées sur le graphe des données synthétiques, nous avons appliqué ces méthodes et évalué leur efficacité selon plusieurs critères. Nous avons ainsi mesuré le temps d'exécution, la taille et le nombre de communautés générées,

le degré interne moyen, ainsi que la fraction moyenne des arêtes pointant en dehors de la communauté elle-même, ODF moyen (Average Overlapping Density Fluctuations) (cf. Section 5.5). Ces critères ont été choisis pour leur pertinence pour la comparaison des performances des algorithmes. Dans cette section, nous présentons les résultats de cette évaluation pour chaque méthode citées précedemment.

Les données synthétiques ont été générées avec le simulateur présenté dans le Chapitre 2. Le tableau 3.3 présente les caractéristiques du multigraphe généré.

TABLEAU 3.3 – Tableau descriptif du graphe de données de transactions synthétiques.

	nœuds	arêtes	densité	diamètre	CC
Graphe de transactions	10 000	319 330	0,006 3	40	0,368

Pour ces tests, nous avons utilisé les algorithmes suivants : CPM, Slpa, Angel, Bigclam, Wcommunity, Walkscan, Copra, et Oslom. Ces algorithmes ont été implémentés en utilisant les configurations par défaut. Les résultats obtenus sont présentés dans le tableau 3.4.

TABLEAU 3.4 – Comparaison des algorithmes de détection de communautés chevauchantes de l'état de l'art.

Algorithme	*CPM*	*Slpa*	*Angel*	*Bigclam*	*Wcommunity*	*Walkscan*	*Copra*	*Oslom*
Temps d'exécution (s)	1, 013	5905, 95	1067, 11	72, 04	80, 16	1, 59	0, 253	283, 15
Nombre de communautés	917	930	1	1	1904	1	1	353
Taille moyenne	5	12, 19	9972	10000	6, 62	10000	10000	31, 71
Degré interne moyen	15, 2	22, 47	63, 67	63, 86	8, 03	63, 86	63, 86	52, 25
ODF moyen	63, 99	35, 57	0, 18	0	62, 82	0	0	21, 11

Le tableau de résultats 3.4 met en évidence les différences entre les performances des différents algorithmes de détection de communautés. Bien que certains de ces algorithmes soient rapides en temps d'exécution, ils ont tendance à échouer dans

la détection de communautés denses et à regrouper tous les nœuds dans une seule communauté. Nous pouvons expliquer cela par le fait que le graphe de transactions possède une faible densité, ce qui en fait un graphe spécial présentant des caractéristiques distinctes.

Le degré interne moyen mesure la moyenne des degrés internes des nœuds dans une communauté donnée alors que l'ODF moyen quantifie la fraction des liens qui pointent vers des nœuds en dehors des communautés. Plus l'ODF est faible, plus les communautés sont densément connectées et cohérentes. Ainsi, une bonne partition devrait avoir des communautés avec un degré interne moyen élevé et un ODF moyen faible, ce qui signifie que les communautés sont à la fois denses et bien séparées les unes des autres. En prenant en compte le graphe de transactions Orange Money à titre d'exemple, nous observons que, bien que l'algorithme basé sur la propagation de label (Slpa) donne de bons résultats, il peut être excessivement coûteux en termes de mémoire et de temps de calcul. Pour ceci, des améliorations telles que *PASLPA* pouvant être parallélisables permettraient de contourner ces difficultés.

Il convient de souligner que tous les algorithmes de détection de communautés ont généré des communautés chevauchantes. Ces résultats nous permettront d'orienter notre choix d'algorithmes pour les tests des prochains chapitres.

3.6 Discussion

Les travaux de l'état de l'art sont unanimes sur le fait qu'il n'y a pas une seule définition de communauté valable pour tous les contextes. Chaque méthode de détection conçoit son propre mécanisme pour diviser le réseau, ce qui amène à retrouver des partitions complètement distinctes en nombre, en taille et en densité. De surcroît, les méthodes de détection de communautés se basent chacune sur des critères supplémentaires différents tels que le degré des nœuds, la comparaison entre la cohésion interne et externe, etc. (Fortunato, 2010). Un autre inconvénient observé avec les

méthodes de détection de communautés c'est que la majorité de ces méthodes sont non déterministes : pour le même graphe en entrée, l'application de ces méthodes peut générer différentes partitions pour différentes exécutions. Ceci peut être expliqué par le parcours non unique de l'ensemble des nœuds et d'arêtes, ou par le choix aléatoire que la méthode est amenée à prendre lorsque plusieurs alternatives sont possibles.

Le réseau de transactions Orange Money est un graphe qui contient des millions de nœud et des dizaines voire des centaines de millions d'arcs. Pour étudier ces réseaux de grande taille, nous avons besoin d'algorithmes de détection de communautés adaptés, capables de gérer efficacement les contraintes de mémoire et de temps liées à ces grandes tailles de données. Ces algorithmes doivent être capables de détecter des communautés à différentes échelles, d'utiliser des techniques de parallélisation pour accélérer les calculs et d'être adaptés pour être utilisés sur des plateformes distribuées.

Il est donc important de noter que certaines des méthodes de détection de communautés que nous avons présentées précédemment peuvent ne pas être adaptées pour analyser des graphes de grande taille. Ces méthodes peuvent ne pas être efficaces en termes de temps de calcul ou d'occupation mémoire, ou peuvent ne pas être suffisamment scalables pour gérer les contraintes liées à la taille des données.

Il est donc important de choisir les méthodes adaptées pour chaque cas d'utilisation tels que dans le cas d'un graphe pondéré, ou un multigraphe par exemple. Il faut également être conscient que certains algorithmes peuvent être bons pour certaines propriétés du graphe mais pas pour d'autres, et qu'il est important de bien comprendre les avantages et les limites de chaque algorithme avant de les utiliser pour des analyses spécifiques.

Il convient également de noter que la plupart des algorithmes de détection de communautés sont paramétrables, ce qui signifie qu'ils disposent de plusieurs paramètres

qui peuvent être ajustés pour obtenir des résultats optimaux. Ainsi, il est important de bien paramétrer ces algorithmes pour garantir la qualité des partitions de communautés et leur adaptation au contexte spécifique du graphe étudié. Un mauvais réglage des paramètres peut conduire à des partitions de communautés de faible qualité ou à des temps de calcul trop longs. En conséquence, l'ajustement précis des paramètres des algorithmes est une étape cruciale dans la recherche de communautés dans les graphes.

3.7 Conclusion

La détection des communautés suscite de plus en plus d'intérêt dans la communauté des sciences des réseaux. Cependant, il n'existe pas de définition quantitative de la communauté qui soit explicitement mise en œuvre unanimement dans tous les algorithmes. On a vu au cours de ce chapitre qu'il existe un nombre important de méthodes de recherche de ces sous-groupes denses, et encore plus de propositions d'algorithmes. Il est donc problématique de distinguer les différences topologiques des structures communautaires obtenues à travers différentes méthodes, même si les concepts associés sont perceptibles en théorie. Par voie de conséquence, de nombreux travaux ont été publiés ces dernières années pour évaluer les résultats générés par différentes méthodes aussi bien en proposant des métriques qualitatives, ou des métriques permettant de les comparer à une référence. Dans le prochain chapitre, nous allons explorer quelques-unes de ces métriques. Nous allons discuter leurs avantages et leurs inconvénients et présenter les motivations qui sous-tendent la proposition de nouvelles métriques pour surmonter certains de ces inconvénients.

Chapitre 4

Métriques d'évaluation pour la détection de communautés chevauchantes

Au cours de ce chapitre nous allons présenter les métriques d'évaluation permettant de déterminer la qualité d'une communauté ou d'une partition de communautés données. On distingue deux types fondamentaux de ces métriques : les métriques intrinsèques et les métriques extrinsèques. La contribution dans ce chapitre vise à compenser les critiques adressées aux métriques standard en proposant de nouvelles métriques pour l'évaluation de la détection de communautés chevauchantes.

Sommaire

4.1 Introduction

L'une des applications les plus importantes des réseaux sociaux et des réseaux en général repose sur la recherche des communautés. La présence de ces groupes denses a été observée dans divers réseaux du monde réel. De ce fait, la détection de communautés dans les réseaux a suscité un grand intérêt au cours de la dernière décennie (Chakraborty et al., 2017; Kim and Kim, 2014). Bien que la communauté ne soit pas un concept défini avec précision, un consensus général implique qu'une communauté représente *un groupe de sommets densément connectés, partageant certaines propriétés ou jouant des rôles similaires au sein du réseau.*

En fonction des caractéristiques du réseau, le résultat de la détection des communautés peut conduire à des communautés disjointes, des communautés qui se chevauchent, des communautés dynamiques, etc. Au niveau de cette thèse, nous nous intéressons particulièrement aux communautés chevauchantes. Pour évaluer et comparer les algorithmes de détection de communautés, les travaux de l'état de l'art ont proposé une multitude de métriques d'évaluation (Emmons et al., 2016; Chakraborty et al., 2017). Les métriques d'évaluation peuvent être soit des métriques de qualité qui évaluent la qualité structurelle des communautés, soit des métriques de récupération d'information qui comparent le résultat à un *gold standard*, également appelé *vérité terrain*. Les algorithmes de détection de communautés sont généralement classés en fonction de leur capacité à récupérer le plus grand nombre de communautés existantes, sur la base de la vérité terrain. Malgré le grand nombre de mesures d'évaluation existantes, très peu sont applicables aux communautés qui se chevauchent. De plus, il est important de disposer d'une métrique simple et facile à interpréter pour les algorithmes de détection de communautés.

Au cours de ce chapitre, nous allons d'abord examiner les métriques d'évaluation des communautés chevauchantes, et nous allons par la suite proposer quatre nouvelles métriques de récupération d'information. Ces métriques appartiennent à la

classe des métriques de récupération d'information. Chacune des métriques proposées considère un aspect spécifique du réseau et est conçue pour fournir une explication claire. Notre objectif derrière cette contribution est de surmonter les inconvénients des métriques de récupération d'information standard, à savoir la difficulté d'interprétation des résultats.

4.2 Evaluation de la détection de communautés chevauchantes

L'un des plus grands défis liés à la détection de communautés dans les réseaux sociaux est la capacité à évaluer les résultats générés. L'évaluation est un véritable problème pour les réseaux réels, où seules quelques données sont fournies. Les mesures d'évaluation dans ce domaine peuvent être utilisées soit pour évaluer la performance d'un algorithme de détection de communautés, soit pour comparer les performances de différents algorithmes appliqués au même ensemble de nœuds.

Les métriques d'évaluation peuvent être formellement classées en deux catégories : les métriques intrinsèques et extrinsèques (Chakraborty et al., 2017).

4.2.1 Métriques d'évaluation intrinsèques

Les métriques intrinsèques, ou de qualité, évaluent les propriétés structurelles des communautés identifiées. Elles évaluent dans quelle mesure les éléments de chaque communauté sont similaires et comment ils diffèrent des éléments d'autres communautés, en fonction d'une métrique. Bien que de nombreuses métriques intrinsèques aient été proposées pour évaluer les communautés, peu d'entre elles sont appropriées pour les communautés chevauchantes. Les métriques de qualité intrinsèques les plus courantes disponibles pour les communautés qui se chevauchent sont la centralité Hub, le degré interne moyen, la conductance (Almeida et al., 2011), et la modularité (Newman and Girvan, 2004b).

4.2.1.1 La centralité Hub

La structure interne d'une communauté peut se présenter sous différentes formes. Alors que certaines communautés s'organisent autour d'un ou de plusieurs hubs (nœuds reliés à un grand nombre de nœuds de la même communauté), d'autres maintiennent une distribution uniforme des liens entre ces nœuds. Ceci peut être évalué par la mesure de la centralité hub ou la Hub Dominance (Lancichinetti et al., 2010). Cette métrique est conçue pour identifier le niveau d'organisation centrale autour des nœuds bien connectés. Pour une communauté, plus cette métrique est élevée, plus il est probable qu'elle ait une structure de type hub. Le calcul de la centralité hub pour une communauté C_i est donné par l'équation (4.1)

$$Centralité\ hub(C_i) = \frac{\max_{v_j \in C_i} deg_{int}(v_j)}{n_i - 1} \tag{4.1}$$

où n_i correspond au nombre de nœuds de la communauté C_i, et deg_{int} correspond au degré interne du nœud v_i dans la communauté C_i. Le numérateur correspond au degré le plus élevé de la communauté Ci, si l'on ne considère que les liens internes. La mesure varie entre 0 et 1. Un score élevé de centralité hub correspond à une forte structure centralisée. Une valeur proche de 0 correspond au pire cas où tous les nœuds sont isolés. Une valeur proche de 1 indique qu'il existe au moins un nœud relié à tous les autres.

4.2.1.2 Le degré interne moyen

Le degré interne moyen des nœuds d'une communauté est une mesure de la cohésion interne de cette communauté. Il est calculé en prenant la moyenne des degrés des nœuds qui appartiennent à la communauté. Le degré d'un nœud est le nombre de liens qu'il possède avec les autres nœuds de la communauté. Un degré interne moyen élevé des nœuds d'une communauté indique que les nœuds qui appartiennent à la même communauté sont fortement connectés entre eux. Un degré interne moyen

élevé est souhaitable pour les communautés car cela signifie qu'il y a une bonne cohésion interne dans la communauté et que les nœuds qui appartiennent à la communauté sont similaires.

Pour une communauté, une valeur élevée du degré interne moyen indique une forte cohésion interne, mais cela ne signifie pas nécessairement qu'il y a plus de nœuds dans la communauté. Il est possible qu'une communauté de taille modeste ait un degré interne élevé si les nœuds qui la composent sont très connectés entre eux, tandis qu'une communauté de taille importante peut avoir un degré interne faible si les nœuds qui la composent sont peu connectés entre eux. Le degré interne moyen des nœuds d'une communauté est donc un indicateur de la cohésion interne de cette communauté, indépendamment de sa taille.

4.2.1.3 La conductance

La conductance est une métrique qui permet d'évaluer la qualité de la détection d'une communauté (Shi and Malik, 2000). Elle mesure la proportion de liens qui relient les sommets d'une communauté à des sommets en dehors de celle-ci. Plus précisément, la conductance d'une communauté est le rapport entre le nombre d'arêtes qui pointent à l'extérieur de la communauté et le minimum entre le nombre d'arêtes avec une extrémité dans la communauté ou le nombre d'arêtes qui n'ont pas d'extrémité dans la communauté. Pour une communauté C_i, elle est donnée par la formule :

$$Conductance(C_i) = \frac{|E^{in}_{C_i}|}{2|E^{in}_{C_i}| + |E^{out}_{C_i}|}$$

où $|E^{in}_{C_i}|$ mesure le nombre total d'arêtes complètement internes à la communauté C_i, et $|E^{out}_{C_i}|$ mesure le nombre d'arêtes pointant à l'extérieur de la communauté.

(Yang and Leskovec, 2013) ont montré que la conductance est une bonne métrique pour évaluer les communautés des graphes du monde réel. Dans le contexte de la détection de communautés, une partition avec une faible conductance est considérée

comme une partition de qualité car cela signifie qu'il y a peu de liens reliant les sommets d'une communauté à des sommets en dehors de celle-ci, donc il y a une forte densité de liens entre les sommets d'une même communauté. En d'autres termes, la plupart de ses bords sont internes à la communauté. La conductance du graphe est la moyenne de la conductance de chaque communauté.

4.2.1.4 La modularité

La modularité est une mesure permettant d'évaluer la qualité d'un partitionnement donné. Elle a été introduite par (Newman and Girvan, 2004a) et a été fortement utilisée dans des algorithmes de détection de communauté depuis. Cette mesure permet de déceler des relations denses dans un graphe par rapport à un graphe aléatoire.

En considérant une partition $P = \{C_1, \ldots, C_n, \ldots, C_p\}$, d'un graphe $\mathcal{G} = (\mathcal{V}, E)$, la modularité est donnée par :

$$Q(P) = \frac{1}{2m} \sum_{v_i, v_j \in \mathcal{V}} (A_{v_i v_j} - (\frac{deg(v_i) deg(v_j)}{2m}) \delta(c_{v_i}, c_{v_j}) \tag{4.2}$$

où m est le nombre d'arêtes du graphe, $A_{v_i v_j}$ la valeur dans la matrice d'adjacence entre les sommets v_i et v_j , $deg(v_i)$ le degré du nœud v_i, c_{v_i} l'identifiant de la communauté à laquelle appartient le nœud v_i, c_{v_j} l'identifiant de la communauté auquel appartient le nœud v_j, et $\delta(c_{v_i}, c_{v_j})$ la fonction de *Kronecker* qui vaut 1 si les deux nœuds appartiennent à la même communauté et 0 sinon.

(Nepusz et al., 2008) ont étendu la définition de la modularité dans le contexte des communautés chevauchantes. Dans cette nouvelle formulation, la fonction de *Kronecker* de l'équation 4.3 est remplacée par $s_{v_i v_j} = \sum_{C \in P} \alpha_{v_i C} \alpha_{v_j C}$ où $\alpha_{v_i C}$ indique le degré d'appartenance du nœud v_i à la communauté C. Les degrés d'appartenance vérifient : $0 \leq \alpha_{v_i C} \leq 1$ et $\sum_{C \in P} \alpha_{v_i C} = 1 \ \forall v_i \in \mathcal{V}, \forall C \in P$. La formule de la

modularité chevauchante est alors donnée par :

$$Q_{ov}(P) = \frac{1}{2m} \sum_{v_i, v_j \in \mathcal{V}} [A_{v_i v_j} - (\frac{deg(v_i)deg(v_j)}{2m})] s_{v_i v_j} \tag{4.3}$$

La valeur de la modularité varie entre -1 et 1. Une valeur égale à -1 signifie qu'il n'y a pas de lien entre les nœuds des communautés d'une partition. Par opposition, une valeur égale à 1 révèle une forte structure communautaire. Une valeur nulle de la modularité sous-entend une structure complètement aléatoire.

4.2.2 Métriques d'évaluation extrinsèques

Les mesures de qualité extrinsèque évaluent plutôt la façon dont les communautés résultantes sont proches de la vérité terrain. Ces métriques sont également appelées métriques de récupération de l'information car elles mesurent la capacité des algorithmes à récupérer l'information à partir de la vérité terrain. Dans le cas des réseaux synthétiques, comme nous l'avons vu dans le chapitre 2, les communautés de base sont fournies manuellement lors du processus de génération du réseau (Chakraborty et al., 2018). Cependant, pour les réseaux réels, cette vérité terrain n'est pas toujours disponible. Nous considérons que, lorsque la vérité terrain n'est pas disponible, aucune évaluation ne peut être réalisée sur la performance d'un algorithme de détection de communautés chevauchantes, et aucune comparaison entre ces algorithmes ne peut être effectuée. Par conséquent, dans ce chapitre, nous ne considérons que des réseaux synthétiques où la vérité terrain est automatiquement disponible, ou des réseaux réels où la vérité terrain est connue.

Dans l'état de l'art, les mesures de récupération d'information les plus populaires sont l'information mutuelle normalisée (ONMI) (Danon et al., 2005), l'indice de Rand (IR) (Hubert and Arabie, 1985) et le score F1 (Yang and Leskovec, 2013). Ces mesures de récupération d'information comparent deux ensembles de communautés (pas nécessairement le même nombre) sur la base de différents critères.

Malgré le grand nombre de métriques d'évaluation de la détection de communautés existantes, la plupart des articles se concentrent sur les communautés disjointes. Mais ces dernières années, une attention croissante a été portée à l'évaluation de la détection des communautés qui se chevauchent (Goldberg et al., 2010), (Lutov et al., 2019).

L'évaluation des communautés qui se chevauchent est plus difficile en raison des nœuds qui appartiennent à plus d'une communauté. Une adaptation des métriques les plus connues telles que l'information mutuelle normalisée et le Rand Index aux cas de chevauchement a été proposée, même si leur efficacité est encore discutable. Dans ce qui suit, nous allons présenter les métriques les plus employées pour l'évaluation des communautés chevauchantes.

4.2.2.1 ONMI

L'ONMI (Overlapping Normalized Mutual Information) est un ajustement de l'Information Mutuelle Normalisée (NMI) pour les communautés qui se chevauchent proposée par (Lancichinetti et al., 2009). Basée sur la théorie de l'information, la métrique NMI est devenue l'une des métriques les plus populaires lorsqu'il s'agit d'évaluer la pertinence des communautés (Danon et al., 2005). Pour nos travaux nous utilisons la formulation de ONMI :

$$ONMI(R, G) = \frac{I(R : G)}{max(H(R), H(G))}; \tag{4.4}$$

où $I(R : G)$ représente l'information mutuelle entre le résultat et la vérité terrain, et H représente l'entropie d'une partition.

L'un des inconvénients du ONMI, relevé par (Zhang, 2015), est l'effet de taille finie qui implique que le score ONMI moyen glisse vers le haut avec le nombre de communautés prédites, quel que soit le nombre de communautés de base. La valeur de ONMI varie entre 0 et 1, où 1 correspond à des partitions identiques, et 0 à des

partitions complètement distinctes.

4.2.2.2 Indice Omega

L'indice Omega (Collins and Dent, 1988) est l'adaptation aux communautés chevauchantes de l'indice Rand ajusté (Adjusted Rand Index (ARI)) introduit par Hubert et Arabie (Hubert and Arabie, 1985). L'ARI ne prend en compte que les partitions disjointes. À l'origine, l'indice de Rand (Rand Index (RI)) (Rand, 1971) est basé sur l'accord entre toutes les paires de nœuds du graphe : une paire de nœuds est dite en accord si elle est assignée aux mêmes communautés. L'ARI est ensuite amélioré à partir de l'IR. Il considère l'accord observé et l'accord attendu entre les partitions : un accord observé est la fraction de paires de nœuds classées de la même façon dans les deux partitions.

L'expression de l'indice Oméga de deux partitions C_i et C_j est alors donnée par :

$$Omega(C_i, C_j) = \frac{observed(C_i, C_j) - expected(C_i, C_j)}{1 - expected(C_i, C_j)} \tag{4.5}$$

où le numérateur représente l'accord observé (observed) ajusté par l'accord attendu (expected), et le dénominateur est l'accord maximal possible ajusté par l'accord attendu. Le score le plus élevé possible de 1 indique que deux communautés correspondent parfaitement sur la façon dont chaque paire de nœuds est partitionnée, il est proche de 0 dans le cas d'un partitionnement aléatoire. Les valeurs de l'indice Oméga ne sont pas affectées par le nombre de communautés (contrairement à NMI). Cependant, sa complexité de calcul est élevée.

4.2.2.3 Score F1 moyen

Le score F1 est une mesure de précision et de rappel utilisée en statistiques et en apprentissage automatique pour évaluer la performance d'un modèle de classification binaire. Plus précisément, le score F1 est la moyenne harmonique de la précision et

du rappel. La précision mesure la proportion de vrais positifs parmi les exemples classés positifs, tandis que le rappel mesure la proportion de vrais positifs qui ont été correctement identifiés. Le score F1 combine ces deux mesures pour donner une mesure unique de la performance du modèle. Il est particulièrement utile dans les cas où les classes sont déséquilibrées, c'est-à-dire lorsque l'une des classes est beaucoup plus fréquente que l'autre. Le score F1 est compris entre 0 et 1, où 1 indique une performance parfaite et 0 indique une performance très faible. En général, un score F1 élevé indique une meilleure performance du modèle. Le score F1 est souvent utilisé en conjonction avec d'autres mesures de performance pour évaluer la qualité d'un modèle de classification. Le score F1 moyen est la moyenne des scores F1 de la meilleure correspondance entre la communauté de vérité terrain et chaque communauté détectée, et des scores F1 de la meilleure correspondance entre la communauté détectée et chaque communauté de vérité terrain. Le score F1 moyen est donné par :

$$F1_{average}(C, C') = \frac{1}{2}(\frac{1}{|C|}\sum_{C_i \in C} F1(C_i, C'_{g(i)}) + \frac{1}{|C'|}\sum_{C'_i \in C'} F1(C_{g'(i)}, C'_i)) \tag{4.6}$$

où $g(i) = \underset{j}{argmax}\, F1(C_i, C'_j)$, $g'(i) = \underset{j}{argmax}\, F1(C_j, C'_i)$, et $F1$ est la moyenne harmonique de Précision et Rappel. Pour calculer le score moyen de $F1$, nous devons déterminer quelle communauté détectée C' correspond à quelle communauté de vérité terrain C. L'un des inconvénients du score F1 est qu'il accorde une importance égale à la précision et au rappel. Il est également calculé comme une moyenne des scores F1 des paires de communautés, ce qui peut entraîner un écart-type élevé.

4.3 Contribution

La première contribution des travaux de la thèse repose sur la proposition d'un ensemble de métriques pour l'évaluation de la détection de communautés chevauchantes, plus précisément des métriques de validation basées sur la vérité terrain.

Le besoin d'une nouvelle métrique vient de l'inadéquation observée avec les métriques de l'état de l'art. En effet, lors de nos tests, nous avons été confrontés au nombre limité de métriques d'évaluation adaptées aux communautés chevauchantes. Nous rappelons que l'évaluation des communautés chevauchantes est plus difficile que celle des communautés disjointes en raison de l'appartenance des nœuds à plus d'une communauté.

Il a été prouvé que les métriques adaptées (initialement conçues pour les communautés disjointes), comme ONMI, donnaient des résultats moins précis que les métriques standard sous-jacentes (Emmons et al., 2016). Un autre défaut observé avec les métriques disponibles est qu'elles comparent les partitions de manière globale : si cette approche peut conduire à de bons résultats, elle passe généralement à côté des informations concernant les similarités et les dissimilarités entre les partitions réelles et détectées.

Lorsque nous avons essayé de comparer différents résultats de détection de communautés chevauchantes à la vérité terrain par le biais des des métriques disponibles, nous avons rencontré un véritable défi, à savoir que les valeurs obtenues étaient toutes faibles et non concluantes. Le tableau 4.1 ci-dessous illustre une sélection des résultats obtenus à partir des métriques ONMI, F1 et indice omega, lors de l'application algorithmes *Slpa*, *Oslom* et *Wcommunauté* au graphe de transactions synthétiques.

Les résultats obtenus ont rendu impossible la distinction entre les partitions données ou le choix de la meilleure. Comme on peut le constater, ces valeurs sont toutes très faibles, inférieures à 0,1 même. Ceci rend difficile la formulation de conclusions ou d'observations pertinentes à partir de celles-ci.

Sur la base des critiques ci-dessus, nous proposons de définir un ensemble de métriques de validation basées sur la vérité terrain. Les mesures proposées devraient prendre en compte la structure des communautés obtenues et les comparer avec

TABLEAU 4.1 – Résultats obtenus avec des résultats de détection de communautés chevauchantes.

Algorithme	R^1	R^2	R^3	R^4	R^5
ONMI	$0,0905$	$0,0663$	$0,0555$	$0,0755$	$0,0662$
Score F1	$0,0014$	$0,0021$	$0,0013$	$0,0023$	$0,0025$
Indice Omega	$0,027$	$0,0155$	$0,0132$	$0,0288$	$0,0258$

la vérité terrain. Par conséquent, nos mesures combinent les caractéristiques des mesures basées sur le chevauchement et des mesures basées sur la structure. Les métriques doivent évaluer la correspondance entre le résultat d'un algorithme de détection de communautés et une vérité terrain en explorant les similitudes ou les différences des communautés.

Nous proposons quatre mesures différentes pour comparer les communautés qui se chevauchent. Plus précisément, nous présentons des métriques qui visent à comparer la similarité d'un résultat de détection de communautés qui se chevauchent, à une vérité terrain donnée (El Ayeb et al., 2022b). Ces métriques sont le *taux d'inclusion*, le *taux de couverture*, l'*écart de chevauchement*, et l'*écart de distribution*, qui seront détaillées par la suite.

Chaque métrique proposée considère un aspect structurel des communautés acquises par l'algorithme de détection des communautés. En conséquence, nous considérons qu'un bon résultat doit avoir de bons scores pour les quatre métriques. En fait, les quatre métriques doivent être considérées simultanément pour comprendre pleinement les résultats, car certaines métriques peuvent donner de bons scores pour certains mauvais résultats. La combinaison des quatre métriques permettra de prévenir les cas où de bons scores masqueraient des résultats insatisfaisants. Considérer les métriques conjointement vise plus spécifiquement à pénaliser les grosses commu-

nautés englobantes et les petites communautés fractionnées obtenues par certains algorithmes.

Dans le contexte de l'évaluation extrinsèque de la détection de communautés chevauchantes, on estime qu'une bonne métrique doit être capable de mesurer avec précision la similitude entre les communautés détectées et les communautés de vérité terrain, tout en étant robuste aux variations dans les données et aux erreurs de groupement en prenant en compte à la fois la précision et la couverture des communautés détectées, ainsi que la quantité de chevauchements présents dans le graphe. Elle doit également être facile à interpréter et à comparer entre différentes méthodes de détection de communautés.

Dans ce qui suit, considérons un graphe $\mathcal{G} = (\mathcal{V}, E)$, pour lequel il existe deux partitions de communautés : une vérité terrain $G = G_1, G_2, \ldots, G_n$ de taille n et un résultat de détection de communautés $R = R_1, R_2, \ldots, R_m$ de taille m. Nous supposons qu'une communauté ne contient pas de nœuds en double. Essentiellement, ces deux groupes n'ont pas le même nombre de communautés (n peut être différent de m) mais contiennent le même ensemble de nœuds. Les communautés de R et de G se chevauchent également.

Pour toutes les métriques, les deux ensembles de communautés $G = G_1, G_2, \ldots, G_n$, et $R = R_1, R_2, \ldots, R_m$ constituent les entrées, et la sortie est une mesure $d(R, G)$. Chaque mesure proposée doit également remplir certaines propriétés de base :

— Positivité : La métrique doit être positive $d(R, G) \geq 0$.

— Intervalle de définition : Nous avons choisi de définir le domaine des métriques comme étant $[0, 1]$ i.e. $d(R, G) \in [0, 1]$, où 0 signifie que les partitions sont complètement différentes et 1 qu'elles sont identiques.

— $d(G, G) = d(R, R) = 1$.

4.3.1 Nouvelle métrique 1 : Taux d'inclusion

Une façon simple de définir la similarité entre deux ensembles de communautés R et G est de considérer à quel point R représente G, en tenant compte de l'inclusion des communautés dans les deux groupes. Pour définir le taux d'inclusion, et le taux de couverture, nous devons commencer par rappeler les définitions de la *précision* et du *rappel*.

En considérant une communauté R_i de la partition résultat R et une communauté G_j de la partition de vérité terrain G, la *précision* définit le nombre de nœuds correctement classés sur le volume du résultat. D'autre part, le *rappel* définit le nombre de nœuds correctement segmentés sur le volume de la vérité terrain.

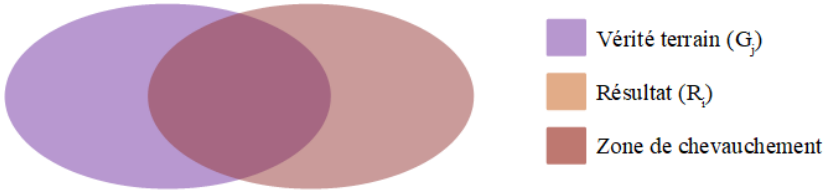


FIGURE 4.1 – Précision et rappel

Si l'on considère les partitions du résultat et de la vérité terrain de la figure 4.1, la précision et le rappel sont donnés par les équations suivantes :

$$précision = \frac{|R_i \bigcap G_j|}{|R_i|},\ rappel = \frac{|R_i \bigcap G_j|}{|G_j|}, \forall (i,j) \in [|1,m|] \times [|1,n|].$$

La précision ne tient pas compte des erreurs de sous-segmentation, tandis que la sur-segmentation n'est pas reflétée dans le rappel (Yeghiazaryan and Voiculescu, 2018). Par conséquent, nous prenons en considération la combinaison de ces deux mesures.

Une fois les concepts de base de la précision et du rappel définis, nous pouvons maintenant introduire la première métrique proposée, à savoir le taux d'inclusion.

Le taux d'inclusion est une métrique qui vise à mesurer l'inclusion des communautés de résultats dans les communautés de base. Il est essentiel de disposer d'une métrique de récupération de l'information qui évalue la similarité d'un résultat de détection de communautés avec la vérité terrain. L'idée de base de cette métrique était le besoin d'une mesure qui estime le degré de représentativité des communautés de résultats par rapport aux communautés de base. Pour une communauté de résultats donnée R_i, le taux d'inclusion individuel est donné par le taux de précision maximum.

$$\begin{aligned} taux\ d'inclusion(R_i, G) &= \max_j(précision(R_i, G_j)) \\ &= \max_j(\frac{R_i \bigcap G_j}{|R_i|}) \end{aligned} \tag{4.7}$$

Le taux d'inclusion global est alors défini par le rapport d'une somme pondérée des taux d'inclusion individuels divisée par la somme des tailles des communautés de résultats. Le poids que nous avons choisi est la taille des communautés de résultats.

$$\begin{aligned} Taux\ d'inclusion(R, G) &= \frac{\sum_i taux\ d'inclusion(R_i) \times |R_i|}{\sum_i |R_i|} \\ &= \frac{\sum_i \max_j(R_i \bigcap G_j)}{\sum_i |R_i|} \end{aligned} \tag{4.8}$$

Idéalement, un score proche de 1 indique une correspondance exacte entre le résultat et la vérité terrain, tandis qu'un score de 0 reflète une divergence complète entre les deux.

4.3.2 Nouvelle métrique 2 : Taux de couverture

Alors que le taux d'inclusion considère la similarité du point de vue des communautés de résultats, le taux de couverture la considère sous l'angle de la vérité terrain. Notre objectif est d'identifier deux métriques complémentaires. Par analogie avec

le taux d'inclusion basé sur le maximum de la précision, le taux de couverture est une fonction du rappel maximum.

Pour une communauté de vérité terrain donnée G_j, le taux de couverture individuel est donné par le taux de rappel maximal.

$$\begin{aligned} taux\,de\,couverture(G_j, R) &= \max_i(rappel(R_i, G_j)) \\ &= \max_i(\frac{R_i \bigcap G_j}{|G_j|}) \end{aligned} \tag{4.9}$$

Pour obtenir le taux de couverture global, nous choisissons un poids égal à la taille des communautés de la vérité terrain.

$$\begin{aligned} Taux\,de\,couverture(R, G) &= \frac{\sum_j taux\,de\,couverture(G_j) \times |G_j|}{\sum_j |G_j|} \\ &= \frac{\sum_j \max_i(R_i \bigcap G_j)}{\sum_i |G_j|} \end{aligned} \tag{4.10}$$

Le score est minimal (proche de 0) lorsque le résultat diffère complètement de la vérité terrain, et maximal (égale à 1) lorsque le résultat est en accord parfait avec la vérité terrain.

4.3.3 Nouvelle métrique 3 : L'écart de chevauchement

L'écart de chevauchement est une métrique extrinsèque qui se base sur la comparaison du nombre de nœuds chevauchants dans la partition résultat et dans la vérité terrain. L'étape initiale du calcul de cette métrique consiste à calculer les taux de chevauchement entre les communautés résultats. Pour une paire de communautés, le taux de chevauchement est calculé en divisant le nombre de nœuds communs qu'elles partagent par la plus petite taille entre les deux communautés. Le taux de chevauchement pour un couple de communautés de résultats, par exemple, est donné par l'équation suivante 4.11 :

$$taux\ de\ chevauchement(R_i, R_j) = \frac{|R_i \bigcap R_j|}{min(|R_i|, |R_j|)} \tag{4.11}$$

Remarque : le taux de chevauchement est symétrique.

De cette manière, le taux de chevauchement d'une partition peut être vu comme une propriété qui représente la moyenne des taux de chevauchement des paires de communautés qu'elle contient. Il est donnée par :

$$taux\ de\ chevauchement(R) = \frac{1}{|R|} \sum_{\substack{R_i, R_j \\ R_i \neq R_j}} taux\ de\ chevauchement(R_i, R_j) \tag{4.12}$$

En ce qui concerne le taux de chevauchement, il n'existe pas de score idéal, car celui-ci dépend des caractéristiques des données analysées. Cependant, vu que l'écart de chevauchement est une mesure de récupération d'information, on souhaite que cette propriété soit identique pour R et G et idéalement le taux de chevauchement de R soit proche de celui de G. D'où la définition de l'écart de chevauchement :

$$\begin{aligned} &\Delta\, chevauchement(R, G) = \\ &1 - |taux\ de\ chevauchement(R) - taux\ de\ chevauchement(G)| \end{aligned} \tag{4.13}$$

Bien qu'il soit intuitif de s'attendre à ce que l'écart soit proche de zéro lorsque les deux partitions sont identiques, nous avons délibérément choisi de définir des métriques dont les valeurs sont proches de 1 lorsqu'elles sont maximales. Par conséquent, nous avons utilisé une formule de normalisation pour ajuster les résultats de manière à refléter cette convention de mesure. Ainsi, une mesure d'un écart de chevauchement proche de 1 indique que les taux de chevauchement du résultat et de la vérité terrain sont similaires et un score proche de 0 indique que ces taux sont très différents.

4.3.4 Nouvelle métrique 4 : L'écart de distribution

La distribution d'un nœud est égale au nombre de communautés auxquelles il appartient dans une partition. Pour une partition P donnée et un nœud v_i, elle est notée par $n_P(v_i)$.

L'écart de distribution compare donc le nombre de communautés auxquelles chaque nœud appartient dans les communautés de résultats R et dans la vérité terrain G. Pour chaque nœud v_i, l'écart de distribution est donné par :

$$\delta\, distribution(v_i) = |n_G(v_i) - n_R(v_i)| \tag{4.14}$$

L'écart de distribution de deux partitions du même graphe $\mathcal{G} = (\mathcal{V}, E)$ est donné par la moyenne des écarts de distribution individuels de tous les nœuds. Nous utilisons une fonction exponentielle pour ramener les valeurs dans l'intervalle $[0, 1]$.

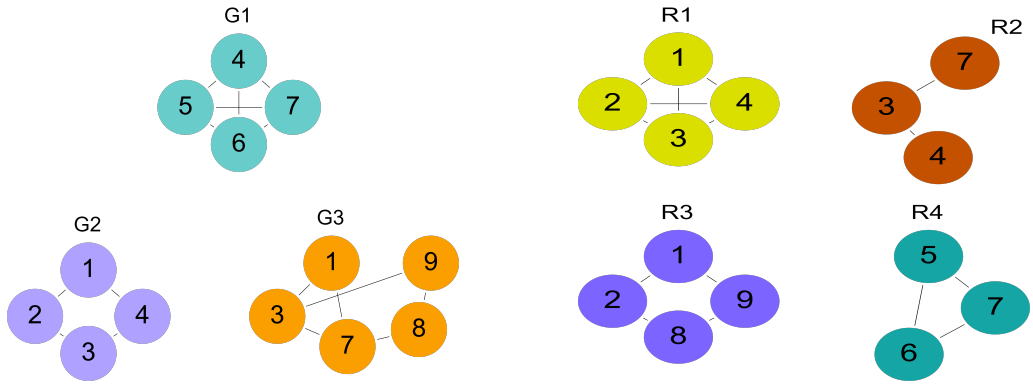
$$\Delta\, distribution(R, G) = \exp\left(-\frac{\sum_i \delta\, distribution(v_i)}{|\mathcal{V}|}\right) \tag{4.15}$$

Ce score est proche de 0 lorsque les deux partitions sont complètement divergentes, et de 1 lorsque les deux sont concordantes.

4.3.5 Illustration

Nous allons tout d'abord évaluer les métriques sur des données synthétiques de petite taille construites manuellement. Ceci nous permettra d'analyser leur validité et les ajuster en conséquence. Dans un premier temps, nous avons construit manuellement une vérité terrain de neuf nœuds. Nous avons ensuite établi un bon résultat potentiel qui aurait pu être fourni par un algorithme de détection de communautés. Considérons la vérité terrain (G) et la partition (R) données par la figure 4.2. À gauche (la vérité terrain), les neuf nœuds sont répartis en trois communautés, à savoir G_1, G_2 et G_3. À droite (le résultat), les mêmes nœuds sont répartis en quatre

communautés, R_1 à R_4. Les partitions de ces deux communautés représentent des structures similaires, mais avec quelques variations mineures. Les deux partitions présentent également des chevauchements.



(a) Les communautés de la vérité terrain. (b) Les communautés du résultat.

FIGURE 4.2 – Illustration d'une vérité de terrain synthétique et d'un résultat potentiel.

Les deux partitions ont approximativement la même structure, et un nombre proche de communautés (G a trois communautés et R en a quatre). Afin d'évaluer la similarité entre le résultat et la vérité terrain, nous allons appliquer les quatre métriques proposées. Par exemple, en considérant la vérité terrain G, la partition R aurait les valeurs suivantes pour les métriques proposées :

— $Taux\ d'inclusion(R,G) = \dfrac{4+2+3+3}{4+4+3+3} = \dfrac{12}{14} = 0,857$

— $Taux\ de\ couverture(R,G) = \dfrac{4+3+3}{4+4+5} = \dfrac{10}{13} = 0,769$

— $\Delta\ chevauchement(R,G) = 1 - |\dfrac{\frac{2}{3}+\frac{2}{4}+0+0+\frac{1}{3}+0}{6} - \dfrac{\frac{1}{4}+\frac{2}{4}+\frac{1}{4}}{3}|$
$= 0,917$

— $\Delta\ distribution(R,G) = \exp(-\dfrac{0+1+0+0+0+0+0+0+0}{9}) = 0,895$

Sur la base des valeurs élevées du taux d'inclusion et du taux de couverture ci-dessus, nous pouvons conclure que la partition R est bien incluse dans la vérité terrain, mais aussi que la vérité terrain est bien représentée dans cette partition avec un taux de couverture élevé. L'écart de distribution élevé indique que les nœuds sont disposés de manière similaire dans les deux partitions. L'écart de chevauchement est aussi élevé. En résumé, la partition R est un bon résultat que nous pourrions obtenir d'un algorithme de détection de communautés.

Dans les sections suivantes, nous allons générer des partitions synthétiques de résultats. Lors de la création des communautés synthétiques, notre objectif est de mettre l'accent sur des exemples extrêmes pour tester les performances de la métrique. Chaque partition synthétique a une caractéristique qui impacte un aspect spécifique tel que l'inclusion, le chevauchement ou le nombre de communautés. Nous allons tester le comportement des métriques en conséquence.

4.4 Expérimentations

Afin d'évaluer l'efficacité et de tester les limites pratiques des métriques proposées, nous allons les appliquer sur de petites données synthétiques manuelles avec différents résultats potentiels, et sur un réseau classique de la littérature : le réseau de football universitaire américain (Park and Newman, 2005).

4.4.1 Application sur des données synthétiques

Dans cette section, nous allons étudier les partitions synthétiques qui sont le résultat d'une sous-segmentation et d'une sur-segmentation. La vérité terrain utilisée pendant ces tests est la même vérité terrain G créée dans la section précédente.

4.4.1.1 Sous-segmentation

La sous-segmentation consiste à regrouper plusieurs communautés de la vérité terrain en une seule. Cela implique d'avoir moins de communautés. Ces communautés sont généralement étendues. Afin de tester les nouvelles métriques proposées dans le cas d'une sous-segmentation, nous établissons quatre partitions synthétiques à partir du même ensemble de nœuds présenté précédemment (figure 4.2). Pour commencer, nous avons considéré quatre partitions tests R^1, R^2, R^3 et R^4. La première partition choisie prend la forme d'une énorme communauté unique qui englobe tous les nœuds. Ce type de résultat a été rencontré lors de nos tests ultérieurs. La deuxième partition choisie est un ensemble de neuf ($N = 9$) communautés de taille $N - 1$. Dans chaque communauté, un seul nœud est manquant. Pour les troisième et quatrième partitions, nous avons opté pour une solution qui divise le réseau en deux parties, respectivement sans chevauchement et avec un nœud de chevauchement. L'objectif est d'observer si le chevauchement peut affecter les résultats dans ce cas. Les partitions et leurs tailles sont données dans le tableau 4.2.

TABLEAU 4.2 – Communautés correspondant à une sous-segmentation.

	G	R^1	R^2	R^3	R^4
Communautés	1 2 3 4 4 5 6 7 1 3 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 1 2 3 4 5 6 7 9 1 2 3 4 5 6 7 9 1 2 3 4 5 6 8 9 1 2 3 4 5 7 8 9 1 2 3 4 6 7 8 9 1 2 3 5 6 7 8 9 1 2 4 5 6 7 8 9 1 3 4 5 6 7 8 9 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 5 6 7 8 9

Dans le tableau 4.3, nous présentons les différents résultats de ces partitions pour le taux d'inclusion, le taux de couverture, l'écart de chevauchement, l'écart de dis-

tribution, ainsi que l'ONMI, le score F1 et l'indice Oméga (page 79).

TABLEAU 4.3 – Évaluation des métriques standard et des nouvelles métriques sur les 4 partitions résultats.

	R^1	R^2	R^3	R^4
Taux d'inclusion	$0,55$	$0,55$	$0,77$	$0,7$
Taux de couverture	1	1	$0,76$	$0,76$
Écarts de chevauchement	$0,66$	$0,45$	$0,66$	$0,86$
Écart de distribution	$0,64$	$0,0014$	$0,64$	$0,57$
ONMI	0	$0,15$	$0,36$	$0,23$
Score F1	$0,71$	$0,52$	$0,75$	$0,74$
Indice Oméga	0	0	$0,25$	$0,12$

Nous observons que les valeurs des taux de couverture sont plutôt élevées. Les deux premières partitions ont des taux de couverture de 1, et R^3 et R^4 ont des valeurs de 0,76. Cela signifie que pour les cas de sous-segmentation, en raison de leur grande taille, les communautés de résultats auraient tendance à couvrir au maximum les communautés de vérité terrain. Cela explique donc les taux de couverture élevés. Le taux d'inclusion, quant à lui, varie de 0,55 à 0,77. Il est au plus bas lorsqu'il s'agit de très grandes communautés, mais augmente avec R^3 et R^4. Pour cette paire de partitions, nous remarquons une augmentation de l'écart de chevauchement et une diminution de l'écart de distribution (puisque le nœud 5 est représenté une fois dans G, mais deux fois dans R^4). Maintenant, si nous considérons les valeurs des métriques standard de récupération d'information considérées, nous pouvons voir que ONMI et l'indice Omega capturent une certaine différence entre les différentes partitions. Cependant, cela n'est pas aussi évident pour le score F1. Trois des quatre scores F1 sont supérieurs à 0,71, même si le résultat est mauvais par rapport à la vérité terrain.

Dans un second temps, nous avons élargi l'échelle des tests en utilisant des jeux de

données plus grands contenant 18 partitions résultats. Les figures 4.3 et 4.4 présentent les résultats des nouvelles métriques proposées et des métriques de l'état de l'art pour dix-neuf partitions présentant des caractéristiques de sous-segmentation.

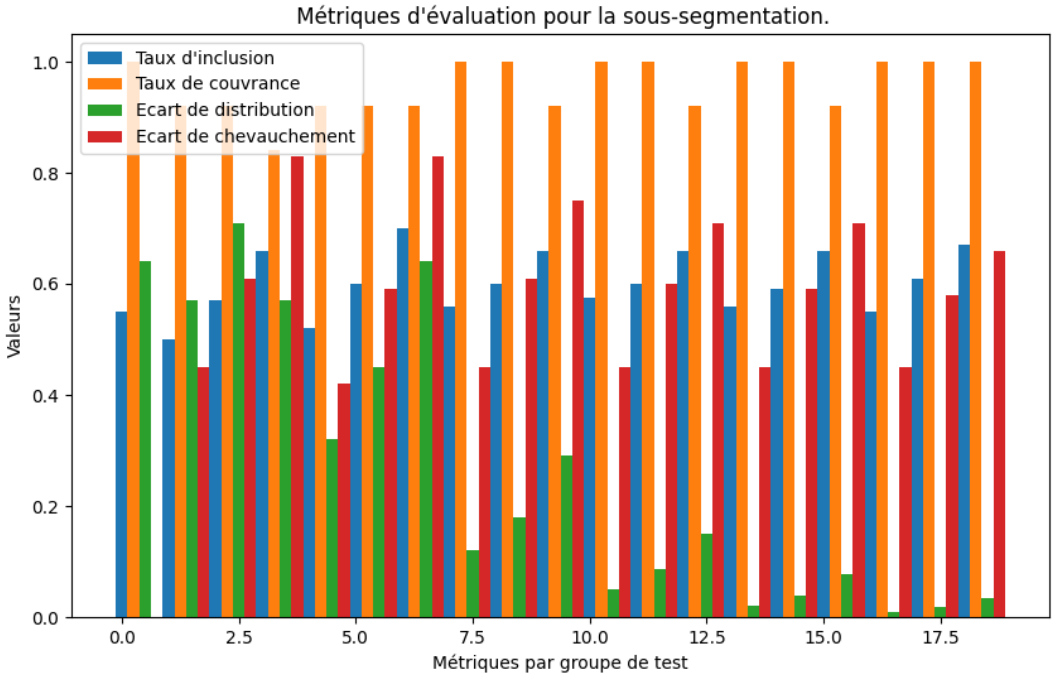


FIGURE 4.3 – Variation des nouvelles métriques pour les résultats de sous-segmentation.

Les résultats obtenus sont en accord avec les tests précédents. Pour toutes les partitions, les taux de couverture montrent une tendance régulière et gardent des valeurs élevées par comparaison aux taux d'inclusion. Les métriques de l'état de l'art en contrepartie présentent une forte fluctuation. Ainsi, les résultats des tests effectués sur les données de sous-segmentation valident que nos métriques permettent de mieux évaluer la qualité des communautés obtenues, en comparaison avec les métriques utilisées habituellement pour mesurer la qualité des communautés basées sur la vérité terrain.

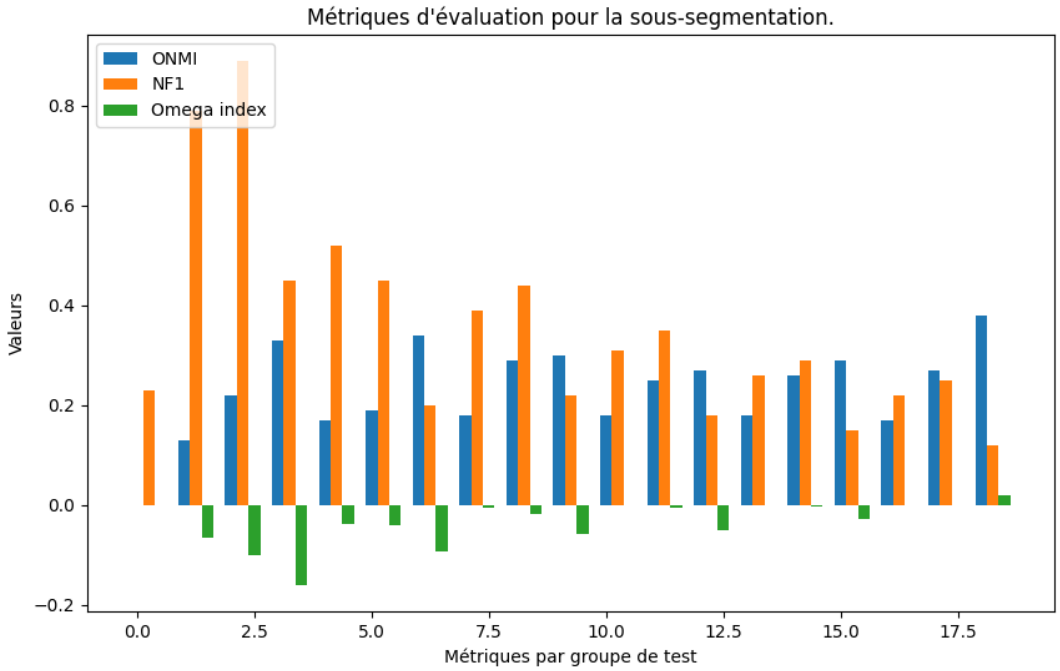


FIGURE 4.4 – Variation des métriques standard pour les résultats de sous-segmentation.

4.4.1.2 Sur-segmentation

Cette deuxième section examine un deuxième cas de biais spécial qui est la sur-segmentation. On parle de sur-segmentation lorsqu'il y a un nombre de communautés plus élevé que prévu. Ces communautés sont généralement de petites tailles. À cette fin, nous testons les métriques proposées sur trois partitions synthétiques. La première partition est un cas extrême où le résultat est neuf communautés uniques. Nous supposons que l'algorithme de détection des communautés ne parvient pas à découvrir la structure des communautés dans le réseau et affecte chaque nœud à sa communauté individuelle. Pour les deuxième et troisième partitions, nous supposons que l'algorithme de détection de communautés réussit à identifier une communauté de base, mais divise le reste du réseau respectivement en communautés unitaires et en communautés de paires. Les éléments des différentes partitions sont donnés dans le tableau suivant 4.4 :

TABLEAU 4.4 – Communautés correspondant à une sur-segmentation.

	G	R^1	R^2	R^3
Communautés	1 2 3 4 4 5 6 7 1 3 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 5 6 7 8 9	1 2 3 4 4 5 6 7 8 9

Les résultats du taux d'inclusion, du taux de couverture, de l'écart de chevauchement et de l'écart de distribution, ainsi que le ONMI, le score F1 et l'indice Oméga pour ces partitions sont donnés dans le tableau suivant 4.5.

TABLEAU 4.5 – Métriques standard et nouvelles dans le cas d'une sur-segmentation.

	R^1	R^2	R^3
Taux d'inclusion	1	1	1
Taux de couverture	0,23	0,53	0,61
Écarts de chevauchement	0,66	0,66	0,75
Écart de distribution	0,64	0,64	0,71
ONMI	0,75	0,4	0,51
Score F1	0,37	0,45	0,64
Indice Omega	0	0,20	0,34

Le tableau 4.5 montre une comparaison des différentes métriques pour les partitions de sur-segmentation. Pour ces partitions, nous observons des valeurs maximales du taux d'inclusion. Ce résultat confirme que le taux d'inclusion fournit des informa-

tions sur la façon dont les communautés de résultats sont contenues dans la vérité terrain. En particulier pour les cas de sur-segmentation où les communautés sont plutôt de petite taille, elles ont plus de chances d'être incluses dans une des communautés de base. En revanche, nous constatons que les taux de couverture sont plus faibles que ceux observés avec la sous-segmentation. Quant aux écarts de distribution et de chevauchement, les valeurs varient respectivement entre 0,64 et 0,71 et entre 0,66 et 0,75. Ces valeurs élevées montrent que même avec de petites communautés, l'occurrence des nœuds dans les communautés et le nombre de nœuds qui se chevauchent sont comparables à la vérité terrain. En termes de score F1 et d'indice Oméga, nous remarquons une augmentation de la métrique de R^1 à R^3 à mesure que la taille des communautés de résultats augmente. Les valeurs du ONMI ne sont pas concluantes. La valeur la plus élevée de 0,75 est trouvée avec les communautés unitaires, ce qui est censé être notre pire résultat.

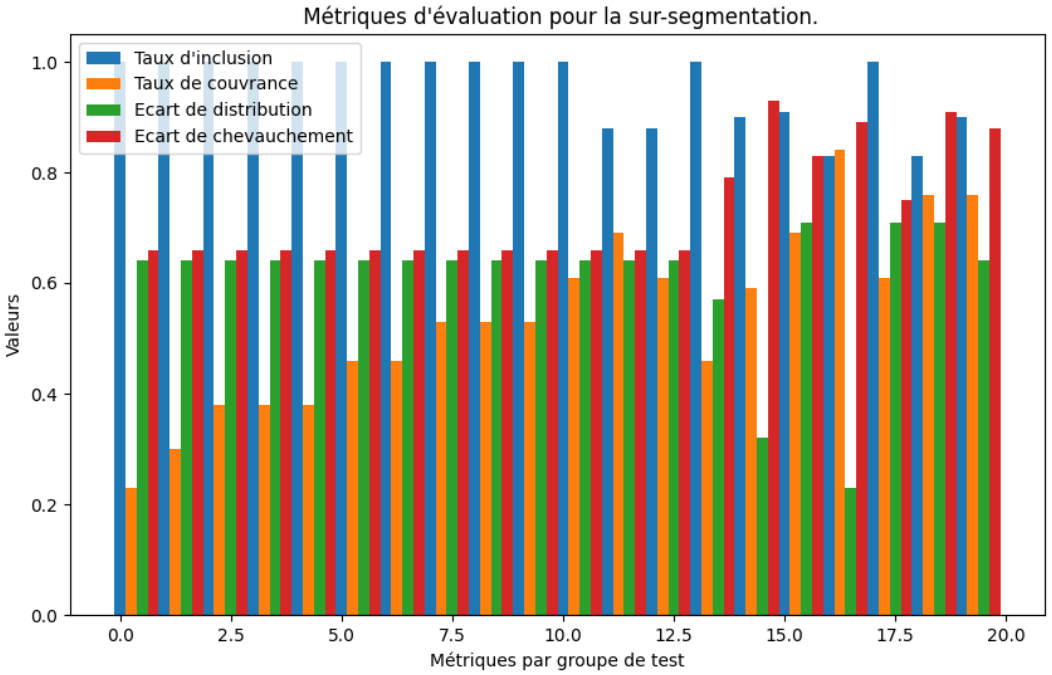


FIGURE 4.5 – Variation des nouvelles métriques pour les résultats de sur-segmentation.

Afin de généraliser les résultats obtenus avec un nombre restreint de données, nous avons généralisé les tests pour un ensemble de vingt partitions caractérisant une sur-segmentation. Les figures 4.5 et 4.6 présentent les résultats des nouvelles métriques proposées et des métriques de l'état de l'art ces partitions.

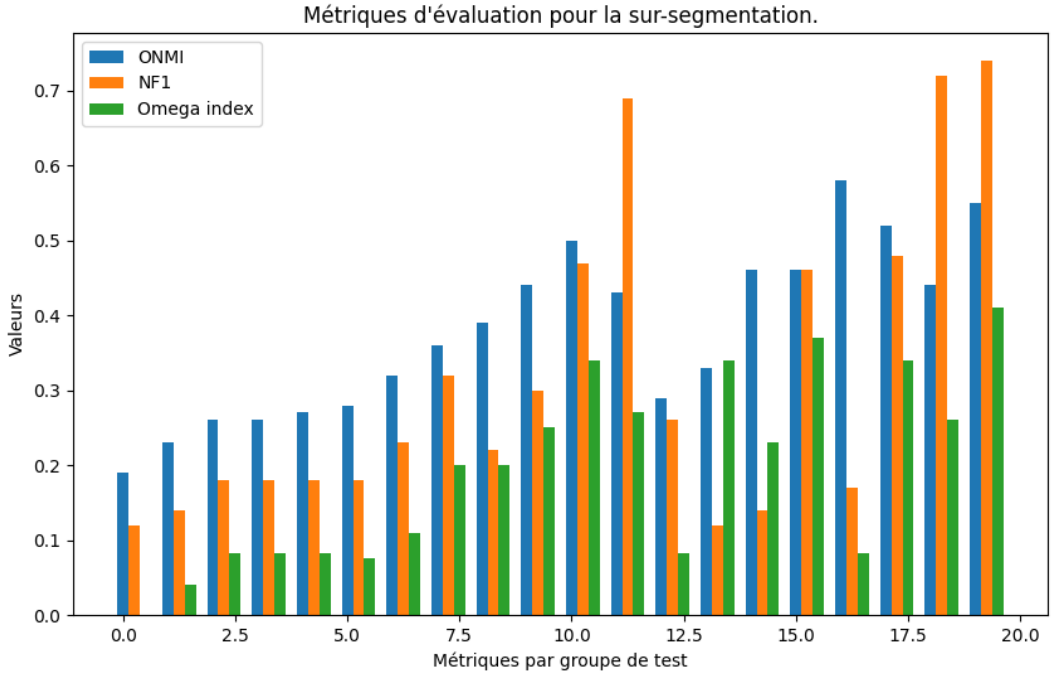


FIGURE 4.6 – Variation des métriques standard pour les résultats de sur-segmentation.

Les résultats confirment les observations faites lors des tests précédents où, contrairement au cas de sous-segmentation, les taux d'inclusion sont très élevés, voire atteignent la valeur maximale, et les taux de couverture sont moins importants. En revanche, les métriques standard présentent des variations importantes et aucune tendance stable n'est observée.

Les tests effectués sur les données de sur-segmentation ont validé que nos métriques fournissent une meilleure comparaison des communautés obtenues avec la vérité terrain, en particulier par rapport aux métriques standard.

4.4.1.3 Altérations successives

Dans cette section, nous étudions des partitions de résultats synthétiques qui correspondent presque parfaitement à la vérité terrain, mais avec quelques variations à différents degrés. En considérant la vérité terrain $G = G1, G2, G3$ de la figure 4.2 comme point de départ, nous allons générer douze partitions résultats en effectuant des altérations successives introduisant chacune un changement dans la structure des communautés. Ces modifications sont principalement de trois types : modification de l'appartenance des nœuds (remplacement de nœuds d'une communauté à une autre, ajout ou suppression de nœuds dans les communautés), modification du nombre de communautés (ajout ou suppression d'une communauté), fusion ou séparation de communautés. Contrairement à la sous-segmentation et la sur-segmentation présentées précédemment, ces résultats pourraient être un résultat « acceptable » d'un algorithme de détection de communautés chevauchantes. Chacune des partitions a un nombre différent de communautés et un niveau différent de chevauchement. Dans le tableau 4.6, quelques-unes de ces partitions.

TABLEAU 4.6 – Exemples de communautés obtenues par des altérations successives.

	G	R^1	R^2	R^3	R^4
Communautés	1 2 3 4	1 2 3 4	1 2 3 4 9	1 2 3 6 9	2 3 6 9
	4 5 6 7	3 4 7	3 4 7	3 6 7	3 6 7
	1 3 7 8 9	1 2 8 9	1 2 8	1 2 8	1 2 8
	5 6 7	2 5 6 8 9	2 5 6 8 9	2 5 4 8 9	2 5 4 8 9

Nous présentons dans les figures 4.7 et 4.8 les différents résultats de ces partitions pour le taux d'inclusion, le taux de couverture, l'écart de chevauchement, l'écart de distribution, ainsi que le ONMI, le score F1, et l'indice Oméga. À première vue, on remarque que plus on modifie le résultat initial G, plus les taux d'inclusion et de couverture se dégradent. En fait, ces deux métriques suivent généralement

le même schéma. Ce résultat confirme le fait que moins les communautés sont représentatives de la vérité terrain, plus les taux d'inclusion et de couverture sont mauvais. L'écart de chevauchement suit ces mêmes variations. En ce qui concerne les écarts de distribution, on remarque que cette tendance n'est pas aussi visible. En examinant les métriques de récupération de l'information standard, nous observons que le ONMI, le score F1 et l'indice Oméga suivent de manière générale la même progression que les taux d'inclusion et de couverture.

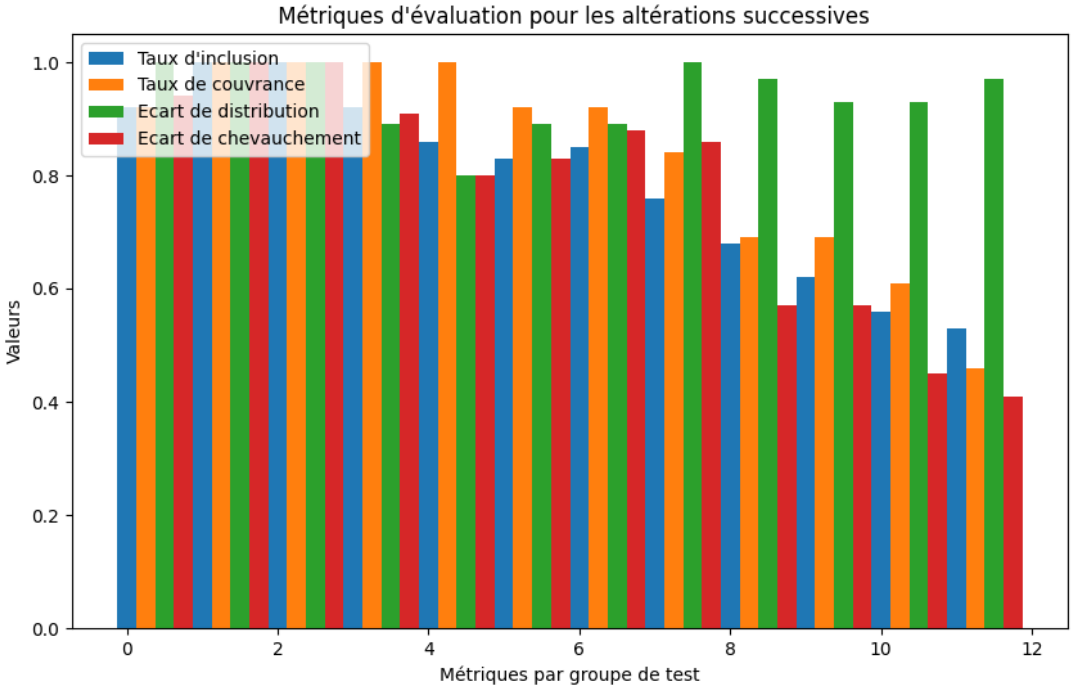


FIGURE 4.7 – Variation des Taux d'inclusion, taux de couverture, écarts de distribution et écarts de chevauchement pour les résultats d'altérations successives.

En résumé, nous avons testé les métriques que nous avons proposées sur différents jeux de données synthétiques présentant des caractéristiques de sous-segmentation, de sur-segmentation, et d'altérations successives de la vérité terrain. À travers les résultats obtenus, nous avons pu démontrer que les métriques proposées varient en fonction de la topologie des partitions, confirmant que nos métriques sont plus réactives aux changements de structure des communautés. Ces variations n'étaient

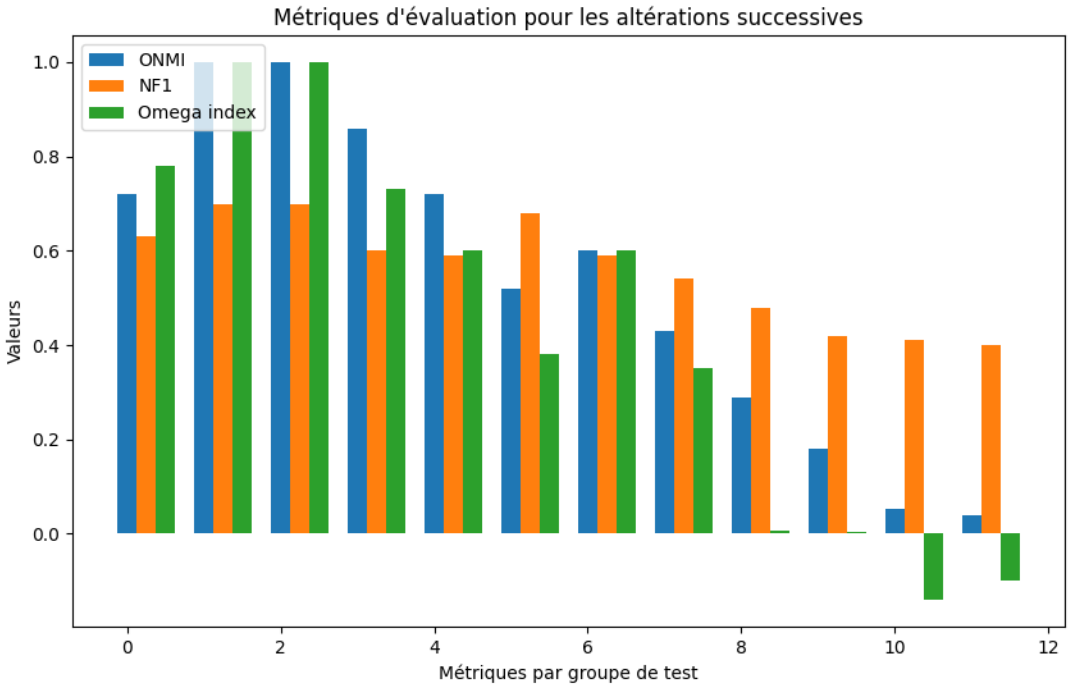


FIGURE 4.8 – Variation des ONMI, scores F1 et indices Oméga pour les résultats d'altérations successives.

pas aussi visibles avec les métriques standard où nous avons eu de bonnes et de mauvaises valeurs pour chacun des cas considérés. Ces nouvelles métriques vont permettre d'améliorer et de faciliter le processus d'évaluation des résultats d'une détection de communautés chevauchantes en résolvant le problème d'interprétation par rapport aux métriques standard. Dans la prochaine partie, nous allons évaluer les résultats de la détection de communautés appliquées sur un réseau réel : celui du football universitaire américain (Park and Newman, 2005).

4.4.2 Application sur le réseau de football universitaire américain

Nous avons également testé nos métriques sur le réseau de football universitaire américain (Park and Newman, 2005). Ce réseau présente le calendrier des matchs

de football universitaire pour la saison régulière de l'automne 2000. Les sommets représentent les universités et les arêtes les matchs de la saison régulière. Il contient 115 nœuds et 613 arêtes et peut être divisé en 12 communautés selon les conférences athlétiques. Chaque communauté comprend 8 à 12 équipes. Le réseau de football universitaire américain est largement utilisé dans la littérature, notamment pour comparer les algorithmes de détection de communautés ou les mesures d'évaluation grâce à la disponibilité d'une vérité terrain.

Dans notre étude, nous avons appliqué quatre algorithmes de détection de communautés qui se chevauchent sur ce réseau afin d'obtenir différentes partitions. Ces algorithmes sont Angel (Rossetti, 2020), BigClam (Yang and Leskovec, 2013), Wcommunity (Chen et al., 2010), et Walkscan (Hollocou et al., 2016) (cf. Section 3.4). Ces algorithmes prennent des paramètres en entrée. Nous avons donc fait varier chacun de ces paramètres afin de tester l'effet des variations sur les partitions générées et par conséquent sur les métriques proposées. Enfin, nous avons calculé les métriques proposées et les métriques classiques de récupération d'information telles que l'information mutuelle normalisée, le score F1 et l'indice Oméga. Les résultats de ces métriques sont donnés dans le tableau 4.7.

Comme les algorithmes ont des paramètres à définir, nous présentons trois résultats de trois partitions détectées distinctes R^1, R^2, et R^3 afin de considérer la déviation des métriques. Il est à noter que notre objectif n'est pas de comparer les performances des algorithmes introduits, mais de considérer les métriques mises en œuvre sur les différents résultats qu'ils génèrent. Pour les différents algorithmes, le taux d'inclusion varie de 0,17 à la valeur maximale 1.

Les résultats obtenus à travers ces tests nous permettent de constater que les variations des nouvelles métriques sont dans l'ensemble en adéquation avec les métriques de l'état de l'art. La variation des paramètres des algorithmes de détection de communautés génère des partitions différentes, qui ont des comportements variables face aux nouvelles métriques. D'une part, pour les algorithmes Bigclam, Wcommunity et

TABLEAU 4.7 – Différentes métriques correspondant aux données "Football club américain".

Algorithme	*Angel*			*Bigclam*			*Wcommunity*			*Walkscan*		
	R^1	R^2	R^3	R^1	R^2	R^3	R^1	R^2	R^3	R^1	R^2	R^3
Taux d'inclusion	0,72	0,80	0,9	0,17	0,2	0,31	0,17	0,96	1	0,17	0,26	0,31
Taux de couverture	0,058	0,61	0,46	1	0,93	0,57	1	0,53	0,22	1	0,51	0,23
Écarts de chevauchement	0,58	0,63	0,69	0,58	0,58	0,58	0,58	0,58	0,58	0,58	0,58	0,58
Écart de distribution	0,21	0,44	0,13	0,53	0,53	0,53	0,53	0,52	0,53	0,53	0,30	0,26
ONMI	0,019	0,37	0,31	0	0,012	0,037	0	0,40	0,13	0	0,021	0,0082
Score F1	0,55	0,55	0,37	0,3	0,28	0,37	0,3	0,18	0,12	0,3	0,39	0,42
Indice Omega	0,027	0,45	0,30	0	0,025	0,091	0	0,44	0,17	0	0,085	0,013

Walkcan, la première partition R^1 révèle une structure de sous-segmentation avec un taux de couvertures maximal (égal à 1) et un taux d'inclusion faible. D'autre part, les troisièmes partitions R^3 produites par les algorithmes *Angel* et *Wcommunity* manifestent une structure de sur-segmentation avec des taux d'inclusion maximaux et des taux de couverture plus faibles.

En revanche, les variations des métriques standard en fonction des différentes partitions ne sont pas aussi évidentes. En effet, c'est le ONMI qui présente des performances meilleures que les scores F1 et l'indice Oméga, qui tous deux produisent des résultats très similaires pour les différentes partitions.

En définitive, les résultats obtenus appuient les critiques (Jebabli et al., 2015) selon lesquelles peu d'interprétation ou d'information sur la structure des partitions peuvent être déduite à partir des métriques de récupération d'information standard, en raison de leur manque de capacité à fournir une analyse détaillée de ces structures. En contrepartie, les métriques proposées corrigent ce problème en fournissant une compréhension des structures et les topologies des partitions. Il est ainsi possible

de différencier une partition d'une autre de manière plus précise, et sélectionner par exemple les paramètres pour ajuster un algorithme conséquemment.

4.5 Discussion et Conclusion

En résumé, au cours de ce chapitre nous avons proposé des métriques de récupération d'information dans le cadre de l'évaluation des résultats obtenus par une détection de communautés chevauchantes. Nous avons commencé par présenter les métriques intrinsèques et extrinsèques existantes pour définir par la suite quatre nouvelles métriques extrinsèques : le taux d'inclusion, le taux de couverture, l'écart de distribution, et l'écart de chevauchement. Le but de cette contribution est de surmonter les limites des métriques d'évaluation existantes et fournir des métriques qui permettent d'avoir une lecture plus claire et détaillée des résultats des différents algorithmes de détection de communautés obtenus. Les nombreuses expériences ont mis en évidence que les métriques proposées varient en fonction de la topologie des partitions, tout en résolvant le problème d'interprétation par rapport aux métriques standard. Toutefois, les métriques ayant été validées sur des données de faible dimension, la prochaine étape consiste à les employer dans le contexte de l'évaluation de la détection de communautés appliquée au multigraphe de transactions. Cela nous permettra de tester l'efficacité de nos métriques sur des données de dimensions plus grandes et plus complexes, et d'ajuster les algorithmes en fonction des résultats attendus.

Chapitre 5

Méthodes de réduction du multigraphe

Les multigraphes représentent des types spécifiques de graphes où les nœuds peuvent être reliés par plus d'une arête. Dans ce chapitre nous proposons d'appliquer différentes méthodes de réduction du multigraphe. Nous allons étudier l'effet de ces réductions sur le processus de détection de communautés. Pour ce faire, nous allons comparer les résultats obtenus par les graphes réduits par rapport au multigraphe initial en utilisant les différentes métriques d'évaluation intrinsèques et extrinsèques introduites au chapitre précédent.

Sommaire

5.1 Introduction

Un multigraphe est un type spécifique de graphe où il est possible d'avoir des arêtes multiples entre les nœuds. En d'autres termes, contrairement à un graphe simple, dans un multigraphe les couples de nœuds peuvent êtres liés par plusieurs liens à la fois. Les multigraphes sont souvent utilisés pour modéliser les réseaux complexes, où une seule arête peut ne pas être suffisante pour représenter la nature complète de la relation. Dans la vie réelle, la plupart des réseaux sociaux exhibent des structures de multigraphes. À titre d'exemple, en considérant les transactions Orange Money, chaque client peut réaliser à maintes reprises différents types de transactions avec d'autres utilisateurs. Pour ceci, la structure de multigraphe semble adéquate pour représenter les interactions des utilisateurs sur le graphe de transactions. Cependant, bien que l'exploitation des multigraphes puisse être utile dans divers contextes et applications, ils sont généralement plus complexes et plus difficiles à étudier que les graphes simples. Dans ce chapitre, notre objectif est d'appliquer des méthodes de réduction du multigraphe et de tester leur incidence sur le processus de détection de communautés. Pour nos travaux, une réduction du multigraphe fait référence à la substitution des liens multiples entre chaque couple de nœuds par un seul lien. Nous allons commencer par présenter la problématique de la réduction du multigraphe et définir six méthodes de réduction. Par la suite, nous allons étudier les effets de la réduction sur les résultats de détection de communautés en appliquant des algorithmes adéquats sur le multigraphe ainsi que sur les versions réduites et en comparant les résultats obtenus à travers des métriques intrinsèques et extrinsèques.

5.2 Du multigraphe au graphe pondéré

À mesure que la complexité des données augmente, la définition classique d'un graphe simple devient insuffisante à la représentation de la sémantique complexe des réseaux du monde réel. Plus précisément, dans ces réseaux, il est possible d'observer

de multiples relations entre l'ensemble des acteurs. Ces relations peuvent décrire des interactions identiques récurrentes dans le temps, ou décrire différents types d'interactions (Papalexakis et al., 2013; Higaki et al., 2020). Les graphes ayant plusieurs arêtes reliant les nœuds sont appelés multigraphes.

Les multigraphes sont de plus en plus étudiés dans divers domaines, car ils fournissent une représentation plus complète et précise à de nombreux systèmes du monde réel. Il existe de nombreuses applications des multigraphes pour l'analyse des réseaux. Parmis ces applications on peux citer : l'étude de différents types de relations et comment chacune influence la structure globale du réseau, la modélisation de la diffusion de l'information ou de l'influence, l'étude de la contribution des différentes relations pour le rôle des nœuds intermédiaires, l'identification des communautés, etc.

Ainsi, l'utilisation des multigraphes pour représenter les relations entre les entités des réseaux permet d'avoir un graphe plus riche en information. Cependant, l'étude des multigraphes présente plusieurs inconvénients par rapport aux graphes simples :

— Défis algorithmiques : de nombreux algorithmes conçus pour fonctionner sur des graphes simples ne sont pas applicables sur des multigraphes, ou doivent être adaptés afin de gérer les arêtes multiples.

— Plus d'espace de stockage requis : les multigraphes nécessitent généralement plus d'espace de stockage que les graphes simples.

— Plus difficiles à visualiser : il peut être plus difficile de représenter visuellement un multigraphe, en particulier lorsqu'il existe un nombre important d'arêtes entre les mêmes sommets.

5.2.1 Définition du problème

Malgré leur capacité à représenter naturellement les réseaux réels, les multigraphes n'ont pas été beaucoup étudiés dans les travaux de recherche. D'autant plus que la

majorité des travaux concernant ce type de structure se concentrent sur les graphes avec des arêtes de types différents, ou ce que l'on appelle les réseaux multi-couches (page 29). Pour l'analyse des multigraphes portant le même type d'arêtes, une démarche générale observée revient à considérer une version simple de ce graphe. Les graphes simples peuvent être dérivés des multigraphes, en réduisant les arêtes multiples entre deux nœuds en une arête simple, c'est ce qu'on appelle le processus de *réduction*.

Peu d'applications de réduction de multigraphes sont présentes dans l'état de l'art (Qi et al., 2011; Acosta-Mendoza et al., 2015; Rutkowski et al., 2021). En ce qui concerne nos travaux, notre intérêt principal réside dans la réduction du multigraphe en un graphe pondéré. Les graphes pondérés ont été étudiés dans différents contextes, car de nombreux réseaux du monde réel sont intrinsèquement pondérés. Leurs arêtes ayant des poids différents décrivent des liens et des flux plus ou moins forts entre les nœuds. Comme tous les types de graphes, les graphes pondérés ont attiré l'attention pour différentes applications, en particulier pour la détection de communautés (Kim and Kim, 2014; Duan et al., 2009; Majmudar and Vavasis, 2020).

A notre connaissance aucune étude n'a été effectuée pour vérifier comment ce processus affecte les opérations menées par la suite. Dans la suite des travaux, nous considérons le multigraphe de transactions Orange Money, où un seul type de transaction sera considéré, soit les transferts d'argent. Notre objectif dans ce chapitre est de simplifier la structure du graphe transactionnel en réduisant le nombre de ses arêtes par différentes méthodes de réduction et d'étudier l'effet de ces réductions sur la détection de communautés tout en espérant de préserver les propriétés du graphe et la force des liens entre les nœuds.

Si la réduction du multigraphe peut entraîner une perte d'informations et de détails, elle permet aussi d'éliminer le bruit qui peut être présent dans le graphe. Nous aspirons à ce que la transition d'un multigraphe vers un graphe pondéré permettra

de limiter cette perte d'information, notamment celles utiles aux algorithmes de détection de communautés, en concevant des poids sur les nouvelles arêtes à partir des attributs des arêtes éliminées. Dans ce qui suit, nous allons présenter des méthodes de réductions basées sur des méthodes de calculs de poids différentes.

5.2.2 Contribution : Les méthodes de réduction proposées

Nous fondons les travaux de ce chapitre sur l'hypothèse que la réduction du multigraphe ne devra pas affecter considérablement le processus de détection des communautés. Nous allons tenter de valider cette hypothèse en nous appuyant sur différentes méthodes de réduction (El Ayeb et al., 2022a). Nous avons choisi cinq méthodes pour transformer le multigraphe en un graphe orienté pondéré. Pour chaque méthode, le poids de l'arête sera calculé différemment. Dans le cadre des données d'Orange Money, les arêtes du multigraphe possèdent divers attributs liés au contexte tel que le montant et la date de la transaction. Par souci de simplicité, nous ne prenons pas en compte les autres attributs. Les poids sont ainsi calculés à partir des attributs de ces arêtes. Les méthodes que nous présentons sont nommées suivant les fonctions de pondération : *simple*, *occurrence*, *somme*, *moyenne*, *moyenne mensuelle* et *score temporel*. Pour les cinq méthodes, nous procédons comme suit : toutes les arêtes dirigées d'un nœud émetteur à un nœud récepteur sont remplacées par une arête dirigée pondérée. Les poids sur les arêtes fourniront des informations sur la relation entre deux nœuds.

Dans ce chapitre, nous considérons le graphe orienté $\mathcal{G} = (\mathcal{V}, E)$. Soit v_i et $v_j \in \mathcal{V}$ deux nœuds de $\mathcal{G}$. $\mu(v_i, v_j)$ désigne le nombre d'arêtes joignant v_i et v_j dans $\mathcal{G}$. La réduction du multigraphe consiste à combiner les arêtes entre chaque couple de nœuds v_i, v_j en une seule arête qui porte le poids $w(v_i; v_j)$. Nous allons également conserver l'orientation du graphe. Les poids ne sont donc pas symétriques, et nous avons :

$$w(v_i; v_j) \neq w(v_j, v_i)$$

5.2.2.1 La méthode simple

La méthode simple consiste à remplacer les arêtes multiples en un arc simple portant un poids égal à 1. Ce graphe nous servira de référence par rapport aux autres graphes pondérés et représentera une étape intermédiaire entre le multigraphe et les graphes pondérés.

$$w_{simple}(v_i; v_j) = 1; \quad \forall v_i, v_j \in \mathcal{V} \tag{5.1}$$

FIGURE 5.1 – Illustration de la méthode de réduction « Simple ».

5.2.2.2 La méthode de l'occurrence

La méthode de réduction *occurrence* se base le calcul du nombre d'arêtes entre chaque couple de nœuds. Le poids « occurrence » est donné par :

$$w_{occurrence}(v_i; v_j) = \mu(v_i; v_j); \quad \forall v_i, v_j \in \mathcal{V} \tag{5.2}$$

FIGURE 5.2 – Illustration de la méthode de réduction « Occurence ».

5.2.2.3 La méthode de la somme des montants

Pour la méthode de réduction *somme des montants*, nous prenons en compte l'attribut « montant de la transaction ». Pour ceci, le poids de l'arête réduite correspond

à la somme des montants des transactions échangées entre le nœud émetteur v_i et le nœud récepteur v_j sur la période étudiée. Le poids « somme » est donné par :

$$w(v_i, v_j) = \sum_{k=1}^{T} montant\ de\ transaction(v_i, v_j); \quad \forall v_i, v_j \in \mathcal{V} \tag{5.3}$$

où T représente le nombre de transactions entre v_i et v_j.

FIGURE 5.3 – Illustration de la méthode de réduction « Somme des montants ».

5.2.2.4 La méthode de la moyenne des montants

La méthode de réduction *moyenne des montants* repose sur le calcul de la moyenne des montants des transactions échangées au cours de la période étudiée. Le poids « moyenne » est donné par :

$$w_{moyenne}(v_i, v_j) = \frac{\sum_{k=1}^{T} montant\ de\ transaction(v_i, v_j)}{\mu(v_i; v_j)}; \quad \forall v_i, v_j \in \mathcal{V} \tag{5.4}$$

FIGURE 5.4 – Illustration de la méthode de réduction « Moyenne des montants ».

5.2.2.5 La méthode de la moyenne mensuelle des montants

Pour cette méthode de réduction, nous intégrons un deuxième attribut des arêtes, soit « la date de la transaction ». Ainsi, le poids d'une arête réduite est égal à la moyenne mensuelle des transactions échangées entre chaque couple de nœuds sur

la période étudiée. Une moyenne mensuelle est la moyenne des montants transmis entre deux clients sur la période de leur échange. La moyenne et la moyenne mensuelle sont toutes deux pertinentes pour comprendre le modèle de l'échange entre les nœuds. Alors que la moyenne mesure la tendance centrale de cet échange, la moyenne mensuelle peut être utile pour comprendre sa tendance temporelle. Le poids « moyenne mensuelle » est donné par :

$$w_{moyennemensuelle}(v_i, v_j) = \frac{\sum_{k=1}^{T} montant\ de\ transaction(v_i, v_j)}{|nombre\ de\ mois|}; \quad \forall v_i, v_j \in \mathcal{V} \quad (5.5)$$

FIGURE 5.5 – Illustration de la méthode de réduction « Moyenne mensuelle des montants ».

5.2.2.6 La méthode du score temporel

La dernière méthode de réduction *score temporel* est basée sur le calcul d'un score qui tient compte des montants des transactions ainsi que la date de chacune. Le calcul du score est basé sur une fonction décroissante de puissance deux, et un facteur τ. L'idée de départ de ce score est qu'une transaction perd de son importance avec le temps. Après un temps donné τ, le score de la transaction baisse à la moitié de sa valeur initiale. Le poids « score temporel » est donné par :

$$w_{scoretemporel}(v_i, v_j) = \sum_{k=1}^{T} montant\ de\ transaction(v_i, v_j) * 2^{\frac{-(t-t_f)}{\tau}}; \quad \forall v_i, v_j \in \mathcal{V} \quad (5.6)$$

Ce score est calculé sur la base de toutes les transactions échangées avec un montant $montant\,de\,transaction(v_i, v_j)$, où t est la date de chaque transaction, t_f est la date de la dernière transaction échangée. On a fixé τ à 30 jours.

Cette méthode de réduction est particulièrement avantageuse dans le cas où on veut mettre à jour le graphe en lui ajoutant de nouvelles transactions en temps réel ou de manière régulière (hebdomadaire, mensuelle, annuelle, etc.). Dans ce cas, il suffit dans un premier temps de créer et stocker le graphe initial. À chaque fois que de nouvelles transactions sont ajoutées, les poids des arêtes seront mis à jour en calculant simplement les nouveaux poids. Cette méthode de calcul permet de traiter rapidement de grands volumes de données, ce qui lui confère un avantage en termes de rapidité et d'efficacité.

Comme présenté dans cette section, les poids proposés sont reliés au contexte et à la nature des données Orange Money et sont basés sur trois variables qui sont le nombre, le montant et la date des transactions échangées. Les scores de réductions obtenus à l'aide des différentes méthodes sont relativement similaires. Notre contribution ne réside pas tant dans la proposition des scores, qui est une tâche bien connue, mais plutôt dans notre approche objective de leur comparaison, ainsi que dans notre étude approfondie de l'impact de la réduction sur le processus de détection de communautés. La prochaine section de notre analyse portera sur la comparaison des effets de ces différentes méthodes de réduction.

5.3 Protocole expérimental

Dans cette partie, nous allons appliquer des algorithmes de détection de communautés chevauchantes sur le multigraphe de transactions, ainsi que sur les six méthodes réduites. Nous allons par la suite comparer les résultats obtenus par différentes approches afin de déterminer quelle méthode est la plus efficace. L'évaluation sera effectuée par le biais de métriques extrinsèques et intrinsèques. Les métriques ex-

trinsèques utilisées sont celles détaillées au chapitre précédent : le taux d'inclusion, le taux de couverture, l'écart de chevauchement, le taux de distribution (cf. Section 4.3), le ONMI, et le score F1. Les métriques intrinsèques employées sont le degré interne moyen, la conductance, et la modularité.

Nous allons également comparer les temps d'exécution des méthodes de détection de communautés sur le multigraphe et sur les graphes pondérés, et en considérant séparément le processus de réduction et celui de détection de communautés. Cette séparation repose sur le fait que la réduction sera effectuée une seule fois, et le graphe sera stocké dans sa forme réduite par la suite.

Dans ce qui suit, nous allons commencer par la présentation des données employées pour construire le multigraphe de transactions. Nous allons introduire les algorithmes mis en oeuvre, et nous allons étudier les résultats obtenus.

5.3.1 Les données de transactions

Dans le cadre des tests de ce chapitre, nous allons utiliser le multigraphe généré à travers le simulateur de transactions et employé pour les tests du Chapitre 3. Il contient 10 000 nœuds et plus de 300 000 transactions sur une période de douze mois. La vérité terrain comprend 3 626 communautés ayant une taille moyenne de 10 de nœuds par communauté.

5.3.2 Les algorithmes

Pour effectuer la détection de communautés sur les graphes décrits plus tôt, nous proposons d'appliquer des méthodes de détection de communautés chevauchantes. Il est important également de choisir un algorithme capable de traiter efficacement un multigraphe. Au départ, nous avions sélectionné trois algorithmes pour nos tests, à savoir *Slpa*, *Oslom* et *Wcommunity*. Lors des tests sur différents graphes réduits, nous avons constaté que l'algorithme *Oslom* produisait des valeurs nulles pour les

métriques d'évaluation extrinsèques standard. C'est pourquoi nous avons opté pour les algorithmes *Slpa* et *Wcommunity* en modifiant ses paramètres par défaut (cf. Section 3.4). En ajustant les paramètres de l'algorithme *wcommunity*, nous avons pu obtenir des résultats améliorés en termes de qualité de communautés détectées. Ces algorithmes sont capables de traiter les multigraphes, aussi bien que les graphes pondérés et permettent de générer des communautés chevauchantes.

5.3.3 Les métriques d'évaluation

Pour l'évaluation des résultats de détection de communautés de chacun des algorithmes sur les différents graphes, nous allons avoir recours à un ensemble de métriques intrinsèques et extrinsèques. Il est possible de retrouver ces définitions dans le chapitre précédent, où elles ont été présentées de manière détaillée.

Les métriques intrinsèques que nous utiliserons sont : le degré interne moyen, la conductance et la modularité. En ce qui concerne les métriques extrinsèques, nous allons recourir aux métriques extrinsèques que nous avons proposées dans le cadre des travaux de cette thèse : le taux d'inclusion, le taux de couverture, l'écart de chevauchement, et l'écart de distribution, ainsi que des métriques de l'état de l'art : ONMI et le score F1. Chacune des quatre métriques que nous avons proposées prend en compte un aspect structurel des communautés acquises par l'algorithme de détection des communautés. En conséquence, nous considérons qu'une bonne partition devrait avoir de bons scores pour les quatre métriques.

Afin de permettre une comparaison avec les métriques standard, nous avons développé un score moyen qui repose sur les taux d'inclusion et de couverture, tels que décrits par l'équation suivante :

$$ScoreMoyen = 2 \times \frac{Taux\ d'inclusion * Taux\ de\ couverture}{Taux\ d'inclusion + Taux\ de\ couverture} \tag{5.7}$$

5.4 Les résultats

Les résultats obtenus par l'application des algorithmes *Slpa* et *Wcommunity* sur le multigraphe ainsi que sur les graphes réduits sont présentés respectivement dans les tableaux 5.1 et 5.2. Ces tableaux regroupent en colonne les types de graphes examinés soit : le multigraphe, le graphe simple, le graphe de l'occurrence, le graphe de la somme, le graphe de la moyenne, le graphe de la moyenne mensuelle, et le graphe du score temporel. Sur les lignes, on retrouve les métriques extrinsèques : le taux d'inclusion, le taux de couverture, l'écart de chevauchement, l'écart de distribution, le score moyen de ces quatre métriques, le ONMI, et le score F1, et les métriques intrinsèques : le degré interne moyen, la conductance, et la modularité.

TABLEAU 5.1 – Métriques extrinsèques et intrinsèques des résultats de l'algorithme *Slpa*.

	Slpa						
	Multigraph	Simple	Occurrence	Somme	Moyenne	Moyenne mensuelle	Score temporel
Taux d'inclusion	**0,837**	0,706	0,789	0,785	0,748	0,773	0,783
Taux de couverture	0,506	**0,675**	0,535	0,519	0,606	0,551	0,447
Écarts de distribution	0,074	0,073	0,077	0,077	0,075	0,078	**0,082**
Écarts de chevauchement	0,99	0,99	0,99	0,99	0,99	0,99	0,99
Score moyen	0,630	**0,690**	0,637	0,624	0,669	0,643	0,569
ONMI	0,230	**0,278**	0,188	0,172	0,222	0,190	0,125
Score F1	**0,266**	0,235	0,246	0,245	0,245	0,244	0,233
Degré interne moyen	29,16	3,65	4,01	3,81	4,79	3,99	2,99
Conductance	**0,321**	0,460	0,521	0,54	0,42	0,52	0,62
Modularité	0,0109	**0,0952**	0,0407	0,0318	0,0504	0,0381	0,0187

À travers les résultats des deux tableaux, on a pu comparer les résultats de détection de communautés sur le multigraphe ainsi que sur les graphes réduits proposés. La

TABLEAU 5.2 – Métriques extrinsèques et intrinsèques des résultats de l'algorithme *Wcommunity*.

	Wcommunity						
	Multigraph	Simple	Occurrence	Somme	Moyenne	Moyenne mensuelle	Score temporel
Taux d'inclusion	**0,815**	0,803	0,804	0,791	0,780	0,776	0,737
Taux de couverture	0,576	0,597	0,585	0,577	**0,613**	0,602	0,592
Écarts de distribution	0,076	0,072	0,079	0,081	0,075	0,077	**0,084**
Écarts de chevauchement	0,998	0,998	0,999	0,999	0,998	0,998	0,999
Score moyen	0,674	0,684	0,677	0,667	**0,686**	0,678	0,656
ONMI	**0,250**	0,202	0,213	0,215	0,223	0,212	0,213
Score F1	**0,260**	0,245	0,249	0,244	0,244	0,245	0,232
Degré interne moyen	27,6	1,47	3,32	3,51	3,74	3,72	4,62
Conductance	**0,44**	0,78	0,65	0,63	0,62	0,61	0,51
Modularité	0,0165	0,048	0,049	0,049	0,058	0,053	**0,079**

première réflexion qu'on peut faire à partir des résultats des métriques extrinsèques est que les écarts observés entre le multigraphe et les graphes réduits sont très faibles. Si on considère le score moyen par exemple, on peut constater qu'il varie entre les valeurs 0,569 et 0,690 pour l'algorithme *Slpa* et entre 0,656 et 0,686 pour l'algorithme *Wcommunity* pour tout graphe confondu. Le score ONMI varie entre 0,125 et 0,278 pour l'algorithme *Slpa*, et entre 0,202 et 0,250 pour l'algorithme *Wcommunity*. Finalement, le score F1 varie entre 0,233 et 0,266 pour l'algorithme *Slpa* et entre 0,232 et 0,260 pour l'algorithme *Wcommunity*.

En examinant les résultats des deux algorithmes de manière plus approfondie, on peut constater que les taux d'inclusion sont assez élevés pour toutes les partitions. Les taux de couverture sont moins notables, mais restent assez élevés avec des valeurs au-dessus de 0,447. Nous avons comparé les tailles moyennes des communautés

de la vérité terrain et des partitions résultats. Les résultats sont illustrés par la figure 5.6. La taille moyenne correspond à la moyenne des tailles des différentes communautés dans chaque partition. Comme les courbes le montrent, la taille moyenne des communautés de la vérité terrain est légèrement supérieure à celle des partitions générées par les algorithmes de détection de communautés.

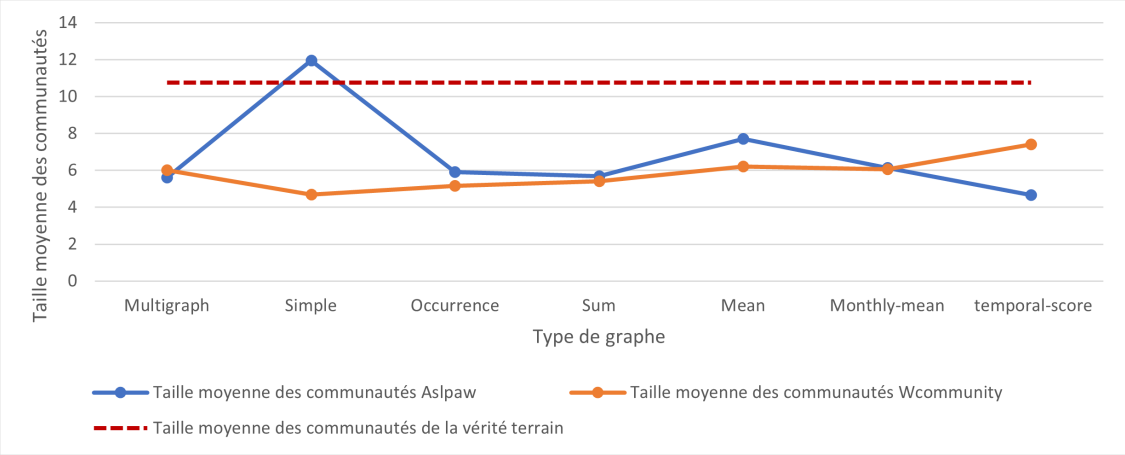


FIGURE 5.6 – Variation de la taille moyenne des communautés pour les résultats de détection de communautés et la vérité terrain.

Les taux de chevauchement des résultats sont très similaires aux taux de chevauchement de la vérité terrain, même si les nombres de nœuds chevauchants dans les deux partitions sont significativement différents comme le montrent les courbes de la figure 5.7. En effet, alors que le nombre des nœuds chevauchants des résultats varie entre 1 123 et 2 608 pour *Slpa*, et 783 et 2 532 pour *Wcommunity*, le nombre de nœuds chevauchant dans les communautés de la vérité terrain est égal à 9 429. Ce qui explique les valeurs élevées des valeurs des écarts de chevauchement est le fait que ce score est normalisé par la taille des communautés qui permet d'éliminer cette disparité.

D'un autre côté, nous pouvons constater que les taux de distribution sont significativement faibles pour toutes les partitions (inférieurs à 0,1). En tenant compte de cette observation, nous avons comparé le nombre de communautés dans les parti-

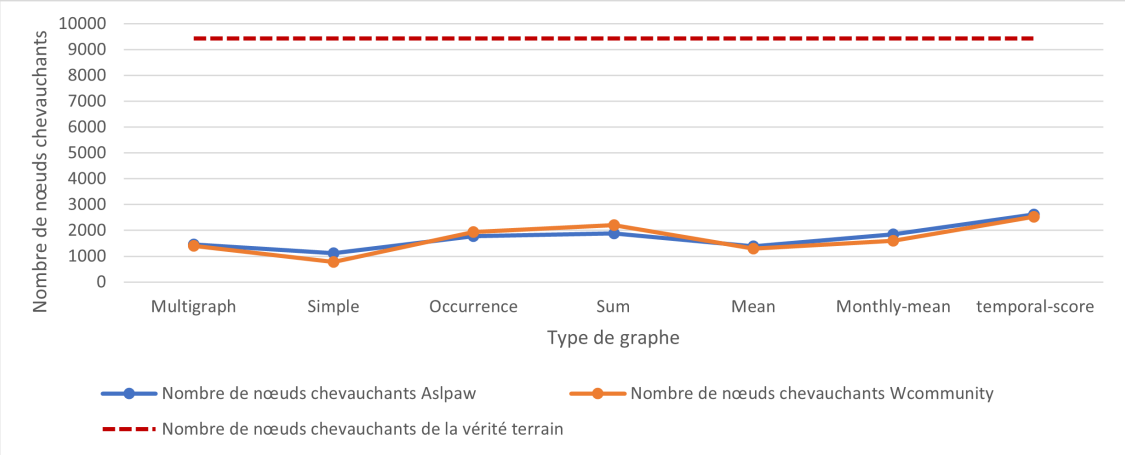


FIGURE 5.7 – Variation du nombre de nœuds chevauchant pour les résultats de détection de communautés et la vérité terrain.

tions résultats par rapport à la vérité terrain comme le montre la figure 5.8. Nous pouvons voir sur cette figure que le nombre de communautés de la vérité terrain (3 626) est remarquablement supérieur à celui des partitions résultats (entre 1 471 et 2 649 pour *Slpa* et entre 1 728 et 2 333 pour *Wcommunity*). Étant basée sur une comparaison des nombres de communautés auxquelles un nœud appartient dans les résultats et la vérité terrain, on suppose donc que les faibles valeurs des taux de distributions sont liées d'une part à la disparité des nombres de communautés entre la vérité terrain et les partitions résultats, et à la disparité du nombre de nœuds chevauchants de l'autre. Plus spécifiquement, les nœuds peuvent appartenir à un plus grand nombre de communautés dans la vérité terrain que dans les partitions résultantes, ce qui explique les taux de distribution significativement plus faibles.

En examinant dans un second temps les métriques intrinsèques, il est possible de constater que la conductance est maximale pour le multigraphe en tenant compte des deux algorithmes, et que les valeurs de la modularité sont généralement faibles avec des valeurs maximales observées avec le graphe simple pour le premier algorithme et le graphe du score temporel pour le deuxième. En ce qui concerne le degré interne moyen, les résultats sont nettement dissociés. Les degrés internes moyens du

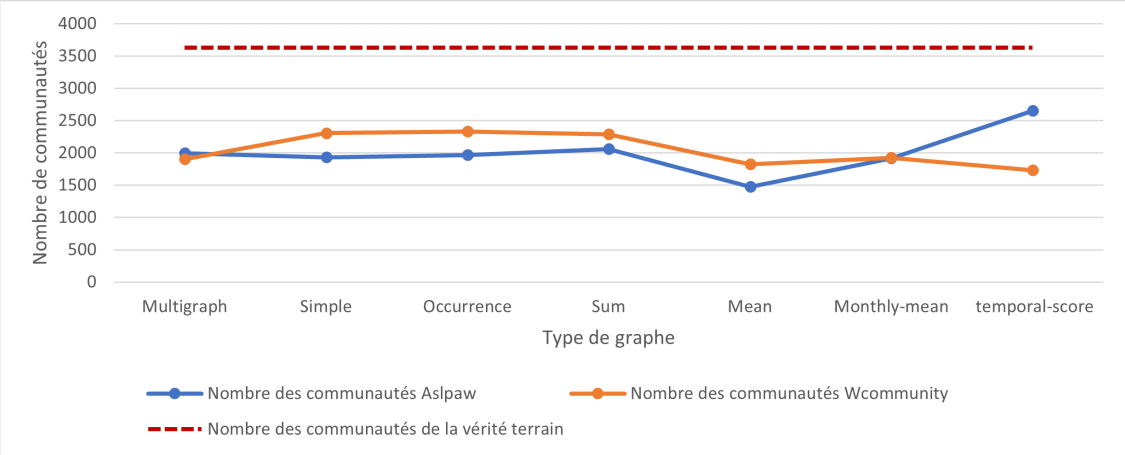


FIGURE 5.8 – Variation du nombre de communautés pour les résultats de détection de communautés et la vérité terrain.

multigraphe sont de l'ordre de 29,16 et 27,6 alors que les degrés internes moyens des graphes réduits varient entre 1,47 et 4,79 comme le montre la figure 5.9. Sachant que le degré interne moyen de la vérité terrain est de l'ordre de 3,25, il est trivial d'observer cette disproportion vu qu'au sein du multigraphe les nœuds sont reliés par plus d'arêtes par rapport aux graphes réduits.

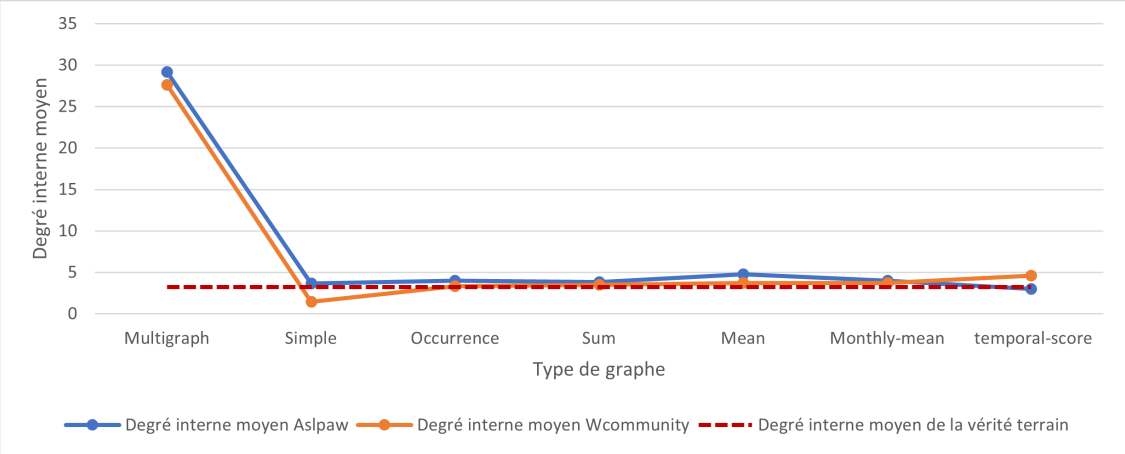


FIGURE 5.9 – Variation du degré interne moyen pour les résultats de détection de communautés et la vérité terrain.

Les résultats recueillis à travers nos tests ont révélé que l'hypothèse que nous avions

posée était bien fondée. Notre analyse a montré que les résultats obtenus sont très similaires, indiquant que le processus de la réduction du multigraphe en graphes simples pondérés n'a pas eu d'impact significatif sur l'application de la détection des communautés chevauchantes. Les métriques extrinsèques sont très comparables, et les modularités des communautés des graphes réduits sont meilleures que celles du multigraphe initial. Ceci prouve qu'afin de réduire la complexité des calculs et l'espace de stockage, réduire le multigraphe initial semble être une bonne solution. Plus particulièrement dans le contexte où la détection de communautés sera effectuée de manière régulière (dans un contexte d'application en temps réel par exemple), le multigraphe peut être réduit et stocké, et l'algorithme sera appliqué sur la version réduite. Notre étude a permis également de mettre en évidence les limites des méthodes de détection de communautés chevauchantes actuellement utilisées. À travers nos tests, nous avons remarqué que la variation de la méthode du calcul du poids du graphe réduit n'avait pas un grand impact sur la détection de communautés. D'autant plus, nous avons constaté que le graphe simple avait de bons voire les meilleurs résultats que les graphes pondérés. Ceci met en évidence les limitations des algorithmes testés, et permet d'envisager des pistes pour des améliorations futures pour des algorithmes de détection de communautés qui ont la capacité de mieux gérer les poids des graphes.

5.5 Conclusion et Discussion

Ce chapitre a apporté des éléments de réponse sur l'effet de la réduction du multigraphe des données de transactions. Dans notre contexte d'étude, une réduction fait référence à la transformation d'un multigraphe à un graphe simple pondéré. Pour ceci, nous avons mis en place différentes méthodes de calcul du poids des arêtes d'un graphe, et avons comparé les résultats de la détection de communautés sur les graphes réduits par rapport au multigraphe initial. Notre analyse a permis de mettre en évidence que la réduction du multigraphe n'a pas eu d'impact significatif

sur les résultats finaux. Pour simplifier le processus de détection de communautés, il est donc possible de réduire le multigraphe sans qu'il y ait une perte d'information significative sur les résultats générés pour la détection de communautés. Nos résultats montrent également que le choix du poids à calculer n'est pas déterminant. Les performances des différentes méthodes de réductions sont très comparables, avec un léger avantage pour le graphe simple. Le choix de ce poids dépendra des contextes spécifiques des applications, ce qui laisse une marge de liberté pour les futurs cas d'études.

Conclusion et perspectives

Nous nous sommes intéressés au cours de cette thèse à l'étude des données de transactions du service financier Orange Money en le modélisant sous forme d'un réseau social. Plus spécifiquement, nous avons concentré nos recherches sur l'application de la détection de communautés chevauchantes sur ce réseau social de transactions dans le but de récupérer les liens sociaux qui existent entre les utilisateurs. La détection de communautés chevauchantes a permis de réduire la complexité du réseau en identifiant des sous-groupes d'utilisateurs ayant des interactions denses. Cette approche a été employée afin d'explorer les propriétés structurelles du réseau et pour obtenir une meilleure compréhension des dynamiques sociales sous-jacentes dans le réseau étudié.

Dans ce manuscrit, nous avons commencé par présenter le cadre de l'étude et posé la problématique de la thèse. Dans le premier chapitre, nous avons décrit le contexte de notre recherche. Nous avons présenté les services de paiement mobile et le service Orange Money en particulier. Nous avons également présenté l'analyse des réseaux sociaux et la motivation derrière la transformation des données de transactions en réseau social afin de créer des graphes qui reflètent les relations entre les différents utilisateurs et comment ceci peut permettre de découvrir des connaissances cachées dans les données.

Le deuxième chapitre s'est penché sur les propriétés mathématiques des réseaux et

des graphes. Nous avons vu les différents types de graphes permettant de modéliser les relations entre les différents éléments d'un système sous forme de nœuds et d'arêtes. Nous avons également présenté des métriques sur les nœuds, qui permettent de quantifier et de caractériser l'importance d'un nœud dans un graphe. Par la suite, nous avons abordé les données synthétiques ainsi que les générateurs de communautés et de transactions nous permettant de respecter la confidentialité des données et d'avoir une vérité terrain nécessaire pour nos tests futurs.

Le troisième chapitre a été consacré à la détection de communautés chevauchantes. Nous avons commencé par définir la notion de communauté de manière locale et globale. Un état de l'art sur les algorithmes de détection de communautés chevauchantes a été détaillé. Ces méthodes permettent d'extraire des informations utiles sur les relations entre les éléments d'un graphe. Ensuite, une première comparaison a été établie afin d'orienter notre sélection d'algorithmes pour les prochaines étapes de notre recherche.

Dans le quatrième chapitre, nous nous sommes focalisé sur les métriques d'évaluation dans le contexte de la détection de communautés chevauchantes. Nous avons présenté les métriques intrinsèques et extrinsèques de l'état de l'art, et de nouvelles métriques extrinsèques ont été proposées. Ces nouvelles métriques ont été validées et comparées avec les métriques existantes de l'état de l'art sur des données synthétiques et des données réelles.

Finalement, le cinquième chapitre a examiné la problématique de réduction d'un multigraphe. Ce type de graphe particulier a été étudié, et différentes méthodes de réduction ont été présentées et appliquées sur le multigraphe de transactions Orange Money. Les résultats ont été par la suite comparés par le biais des métriques de l'état de l'art ainsi que des nouvelles métriques.

Première contribution : les métriques d'évaluation

La première contribution des travaux de cette thèse concerne l'évaluation de la détection de communautés chevauchantes. L'évaluation fait partie des principaux défis liés à la détection de communautés dans les réseaux sociaux. Les mesures d'évaluation permettent soit d'évaluer les performances d'un algorithme de détection de communautés, soit de comparer les performances de différents algorithmes appliqués au même ensemble de nœuds. L'évaluation des communautés qui se chevauchent est plus difficile en raison de l'appartenance des nœuds à plus d'une communauté. Pour ceci, nous avons trouvé que les métriques de l'état de l'art ne parviennent pas à évaluer de manière suffisante la structure communautaire. Dans ce contexte, nous avons proposé quatre métriques d'évaluation extrinsèques : le *taux d'inclusion*, le *taux de couverture*, le *taux de chevauchement*, et *l'écart de distribution*. Les métriques extrinsèques sont des métriques qui comparent les résultats générés par la détection de communautés avec une vérité terrain. Les résultats obtenus renforcent les limites des métriques de récupération d'information standard qui fournissent peu d'informations sur la structure des partitions. En contrepartie, les métriques proposées dans cette thèse offrent une solution à ce problème en permettant une analyse détaillée de la structure et de la topologie des partitions. Il est donc possible de distinguer de manière plus précise une partition d'une autre et de sélectionner les paramètres adéquats pour ajuster les algorithmes en conséquence. Les analyses ont également révélé que les métriques proposées sont plus adaptatives aux changements de structure, tout en résolvant le problème d'interprétation de compréhension par rapport aux métriques de l'état de l'art.

Deuxième contribution : les méthodes de réduction du multigraphe

La deuxième contribution des travaux de la thèse sont centrées sur la réduction du multigraphe de transactions. On appelle réduction d'un multigraphe le processus

de simplification de ce graphe en réduisant le nombre des arêtes multiples entre les nœuds. En rapport avec cette question, nous avons proposé différentes méthodes de réductions qui se basent sur la transformation du multigraphe en graphe simple en remplaçant les arêtes multiples par des arêtes pondérées. Chaque méthode proposée repose sur une fonction différente de pondération des arêtes et permet de créer un graphe pondéré et orienté. Après avoir appliqué des algorithmes de détection de communautés chevauchantes sur le multigraphe initial ainsi que sur les versions réduites, les différents résultats ont été étudiés. La comparaison basée sur des métriques intrinsèques et extrinsèques a montré que les communautés des graphes pondérés sont en général proches de celles du multigraphe et dans certains cas sont structurellement meilleures en nous basant sur la modularité. A l'issue des résultats obtenus, nous avons constaté que la réduction du multigraphe n'a pas eu d'effet notable sur les résultats finaux de la détection de communautés chevauchantes. Ainsi, il est envisageable de simplifier le processus de détection en réduisant le multigraphe sans sacrifier une quantité significative d'informations liées aux résultats générés.

Perspectives et orientations futures

Cette thèse valide la première étape d'un projet visant à analyser les données socio-transactionnelles du service de paiement mobile *Orange Money*. Durant nos travaux, nous avons exploré l'applicabilité et l'efficacité des techniques d'analyse de réseaux sociaux, en particulier la détection de communautés chevauchantes, afin de mieux comprendre ces données et d'obtenir des informations précieuses pour des applications telles que le suivi des flux financiers et la détection de fraudes. Les travaux de cette thèse se poursuivront par une deuxième thèse dont l'objectif est de concevoir des algorithmes permettant d'identifier des groupes d'intérêts (famille, amis, clientèle, tontine, etc.) dans le réseau social des transactions d'Orange Money à travers des approches de Machine Learning, notamment le Graph Machine Learning. Elle visera également à explorer le lien entre l'appartenance à un groupe et le comporte-

ment d'un utilisateur, ainsi que l'analyse et la prédiction de l'évolution temporelle des différents groupes identifiés.

Au cours de cette thèse, nous avons cherché à contribuer à la détection de communautés chevauchantes dans les réseaux complexes. Cependant, il reste de nombreux aspects qui nécessitent une exploration future approfondie :

- Jusqu'à ce jour, il n'y a pas de consensus dans l'état de l'art sur la manière dont le problème de la détection des communautés devrait être considéré ou sur la définition même d'une communauté. Durant nos différentes phases de tests, nous avons constaté que les algorithmes *Slpa*, *Oslom* et *Wcommunity* en changeant ces paramètres par défaut fournissent les meilleures partitions de communautés en termes de qualité et de performance. Ces algorithmes ont réussi à détecter des communautés denses et bien séparées malgré la faible densité du graphe de transactions, ce qui en fait des choix pertinents pour la suite de notre étude. Ainsi, chaque algorithme se base sur sa propre définition de ce qu'est une communauté lors de la recherche de telles structures. Cette limitation a été spécialement observée lors de la comparaison des méthodes de réduction dans le Chapitre 5, où il a été constaté que le graphe simple présentait une meilleure performance globale que certains graphes pondérés. Ainsi, une piste d'amélioration consiste à proposer un algorithme de détection de communautés qui prend davantage en considération les poids des arêtes.
- Dans le Chapitre 4, nous avons présenté quatre métriques d'évaluation extrinsèques pour la détection de communautés. Une piste d'amélioration consisterait à corréler ces métriques extrinsèques avec les métriques intrinsèques, afin d'avoir une vue plus détaillée de la topologie des communautés détectées. Une autre possibilité serait de combiner ces quatre métriques en une seule, sous forme d'une somme pondérée, ou en faisant varier les poids en fonction des cas et de l'aspect que nous souhaitons mettre en avant. Ce type de métrique combinée pourrait fournir une vision globale et complète de la qualité de la

détection de communautés.

- Le cas d'usage « détection de fraude » : une piste de recherche future serait d'explorer l'utilisation de la détection de communautés chevauchantes dans le contexte de la détection de fraudes. Pour cela, on doit supposer qu'un fraudeur n'appartient à aucune communauté, contrairement aux utilisateurs légitimes qui appartiennent à plusieurs communautés du graphe et ont des voisins qui appartiennent à des communautés en commun.

  Pour identifier les transactions frauduleuses effectuées par les nouveaux utilisateurs du service Orange Money, nous avons élaboré un protocole expérimental suivant en nous basant sur les données de transactions ainsi que sur les données d'appels (Call Detail Record (CDR)) :

  1. Générer des données : Données de transaction et données de CDR labellisées sécurisée/frauduleuse sur une période de 15 mois incluant trois mois de données sans fraudes. Injecter des échanges frauduleux et de nouveaux arrivants sur la période des 12 mois restants.
  2. Créer un graphe multicouche pondéré orienté de transactions et de CDR de la période des trois mois sans fraude : le poids des arcs correspond à un score temporel défini.
  3. Effectuer une détection de communautés chevauchantes.
  4. Établir une distance de sécurité sur la base des transactions sécurisées : étudier la distribution des distances communautaires paire à paire des voisins de chaque nœud sur la durée des trois mois et le nombre de communautés par nœud.
  5. Ajouter une semaine de transactions incluant des nouveaux nœuds : nouveaux arrivants et des fraudeurs.
  6. Calculer la distance entre les voisins des nœuds ayant de nouvelles transactions.
  7. Décider si les nœuds sont gardés dans le graphe ou pas.

8. Mettre à jour le graphe.
9. Réeffectuer une détection de communautés / ajuster les communautés.
10. Calculer l'efficacité du processus de filtrage (précision/rappel).

En se basant sur une distance entre les communautés plutôt qu'une distance entre les individus, il serait possible de prédire plus rapidement la légitimité d'un nouvel utilisateur du service Orange Money en calculant la distance communautaire entre ses voisins. Cette approche pourrait améliorer l'efficacité de la détection de fraudes en réduisant le temps de traitement et en fournissant une mesure plus précise de la probabilité de fraude.

Enfin, nous concluons ce manuscrit par le fait que les travaux de cette thèse seront intégrés à une bibliothèque contenant des outils d'analyse de graphes pour Orange.

Publications

Conférences Internationales

El Ayeb, S., Hemery, B., Jeanne, F., and Cherrier, E., "Community detection for mobile money fraud detection." in *Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, Paris, France, December 2020, pp. 1–6

El Ayeb, S., Hemery, B., Jeanne, F., Cherrier, E., and Charrier, C., "Evaluation metrics for overlapping community detection," in *2022 IEEE 47th Conference on Local Computer Networks (LCN)*, 2022

El Ayeb, S., Hemery, B., Jeanne, F., Charrier, C., and Cherrier, E. P., "Multigraph transformation for community detection applied to financial services," in *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, November 2022

Conférence nationale

El Ayeb, S., Hemery, B., Jeanne, F., and Cherrier, E., "Multigraph transformation for community detection applied to financial services." in *French Regional Conference on Complex Systems FRCCS 2023*, Le Havre, France, May 2023

Annexes

GLOSSAIRE

Nous proposons un glossaire permettant de définir certaines notions, principalement liées aux graphes.

Dans ce cadre, nous considérons un graphe $\mathcal{G} = (\mathcal{V}, E)$ où $\mathcal{V}$ représente l'ensemble des sommets et E l'ensemble d'arêtes.

L'excentricité : est la distance maximale existant entre un sommet et les autres sommets du graphe.

Le diamètre : d'un graphe est donné par la valeur de l'excentricité maximale.

La densité : d'un graphe est le rapport entre le nombre d'arêtes existantes et le nombre possible d'arêtes. La densité varie entre 0 et 1. Un graphe complet a une densité égale à 1.

Une chaîne : de longueur k est une suite finie de k liens consécutifs reliant deux nœuds donnés. Une chaîne est *élémentaire* lorsque chaque nœud y apparaît au plus une fois. Une chaîne est *simple* lorsqu'elle passe par un lien au plus une fois. Dans le cas d'un graphe orienté, on parle de chemins.

Un chemin eulérien : est un chemin qui passe par chaque lien exactement une seule fois.

Un chemin hamiltonien : est un chemin qui passe par chaque nœud exactement une seule fois.

Un cycle : (ou circuit) est une chaîne (ou chemin) élémentaire ayant un point d'arrivée identique au point d'arrivée.

Une clique : est un sous-graphe complet de $\mathcal{G}$.

Un graphe connexe : est un graphe ayant une chaîne ou un chemin entre tous ses sommets.

L'ODF (Overlapping Density Fluctuations) : est une mesure utilisée pour évaluer la qualité des partitions de communautés chevauchantes dans les graphes. Plus précisément, l'ODF est définie comme la moyenne de la fraction des liens d'un nœud qui pointent vers des nœuds en dehors de sa propre communauté. Elle mesure la densité de chevauchement dans les communautés. Plus l'ODF d'un nœud est faible, plus il est densément connecté à d'autres nœuds de sa propre communauté plutôt qu'à des nœuds en dehors de cette communauté. Pour une communauté C, l'ODF est donnée par :

$$\frac{1}{|C|}\sum_{v_i \in C}\frac{|(v_i, v_j) \in E : v_j \notin C}{deg(v_i)} \tag{5.8}$$

L'ODF moyen est alors donnée par la moyenne des scores des différentes communautés.

Graphe signé : un graphe signé est un type spécial des graphes pondérés où les arcs peuvent avoir deux valeurs possibles : un signe positif ou un signe négatif. Un arc positif traduit généralement une qualité positive telle que l'attirance ou la confiance. Par analogie, un arc négatif décrira une opposition telle que le rejet ou la méfiance entre les nœuds . Les graphes signés sont utilisés pour la modélisation des réseaux sociaux d'amitié par exemple, ou des réseaux de relations diplomatiques entre les pays.

Un graphe signé est défini par $\mathcal{G} = (\mathcal{V}, E, \sigma)$ où $\sigma : E \to +/-$ est une fonction de mapping qui à tout arc $e \in E$ associe un signe positif $(+)$ ou négatif $(-)$.

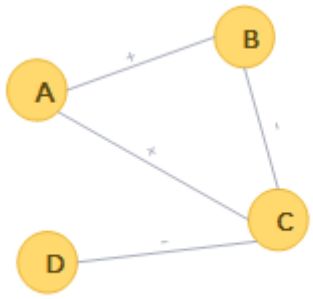


FIGURE 5.10 – Graphe signé.

- Graphe biparti : un graphe biparti est un graphe dont les nœuds peuvent être divisés en deux ensembles disjoints. Les arcs d'un graphe biparti relient des nœuds appartenant à deux ensembles différents. De ce fait, deux nœuds du même ensemble ne peuvent pas être reliés comme le montre la figure 5.11. Grâce à leurs propriétés, les graphes bipartis ont été exploités dans divers domaines afin de modéliser des systèmes ayant deux groupes indépendants qui interagissent.

Un graphe biparti est défini par $\mathcal{G} = (\mathcal{V}, E)$ avec $\mathcal{V} = A \cup B$ où A et B sont deux ensembles disjoints et $(v_i, v_j) \in E$ avec $v_i \in A$ and $v_j \in B$. La figure 5.11 présente un graphe biparti.

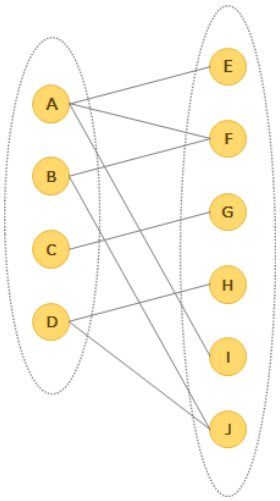


FIGURE 5.11 – Graphe biparti.

Bibliographie

Acosta-Mendoza, N., Carrasco-Ochoa, J. A., Martínez-Trinidad, J. F., Gago-Alonso, A., and Medina-Pagola, J. E., "A new method based on graph transformation for fas mining in multi-graph collections," in *Pattern Recognition*. Springer International Publishing, 2015, pp. 13–22.

Aggarwal, C. C. and Lu, E. Y.-E., "Graph mining techniques for networking applications : A review." *Network Science for Military Coalition Operations : Information Exchange and Interaction*, pp. 42–62, 2010.

Aggarwal, K., Kapoor, K., and Srivastava, J., "Data Mining Techniques for Social Networks Analysis," in *Encyclopedia of Social Network Analysis and Mining*. New York, NY : Springer, 2018, pp. 557–567.

Ahajjam, S. and Badir, H., "Community detection in social networks." in *Principles of Social Networking : The New Horizon and Emerging Challenges*. Springer Singapore, 2022, pp. 91–107.

Almeida, H., Guedes, D., Meira, W., and Zaki, M. J., "Is there a best quality metric for graph clusters ?" in *Machine Learning and Knowledge Discovery in Databases*, Michelangelo, C., Jaakko, H., Ljupčo, T., Celine, V., and Sašo, D., Eds., vol. 10535. Skopje, Macedonia : Springer, 2011, pp. 44–59.

Barabási, A.-L. and Pósfai, M., "Network Science." Cambridge University Press, 2016, chapter 4, pp. 1–57.

Barnes, J. A., "Class and committees in a norwegian island parish." *Human Relations*, vol. 7, no. 1, pp. 39–58, 1954.

Beauchamp, M. A., "An improved index of centrality." *Behavioral science*, vol. 10, no. 2, pp. 161–163, 1965.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E., "Fast unfolding of communities in large networks." *Journal of Statistical Mechanics : Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.

Bohlin, L., Edler, D., Lancichinetti, A., and Rosvall, M., "Community detection and visualization of networks with the map equation framework." in *Measuring Scholarly Impact : Methods and Practice*, 2014.

Bonacich, P., "Factoring and weighting approaches to status scores and clique identification." *Journal of mathematical sociology*, vol. 2, no. 1, pp. 113–120, 1972.

Bonacich, P. and Lloyd, P., "Eigenvector-like measures of centrality for asymmetric relations." *Social Networks*, vol. 23, no. 3, pp. 191–201, 2001.

Boudin, F., "A comparison of centrality measures for graph-based keyphrase extraction." in *International joint conference on natural language processing (IJCNLP)*, Nagoya, Japan, October 2013, pp. 834–838.

Centellegher, S., Miritello, G., Villatoro, D., Parameshwar, D., Lepri, B., and Oliver, N., "Mobile money : Understanding and predicting its adoption and use in a developing economy." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 4, pp. 1–18, 2018.

Chaix, L., "Le paiement mobile : perspectives économiques, modèles d'affaires

et enjeux concurrentiels," Ph.D. dissertation, 2013. [Online]. Available : https://theses.hal.science/tel-00983937

Chaix, L. and Torre, D., "Le double rôle du paiement mobile dans les pays en développement." *Revue économique*, vol. 66, no. 4, pp. 703–727, 2015.

Chakraborty, T., Dalmia, A., Mukherjee, A., and Ganguly, N., "Metrics for community analysis : A survey." *ACM Computing Surveys*, vol. 50, no. 4, pp. 54 :1–54 :37, 2017.

Chakraborty, T., Cui, Z., and Park, N., "Metadata vs. ground-truth : A myth behind the evolution of community detection methods." in *Companion Proceedings of the The Web Conference (WWW '18)*. Republic and Canton of Geneva, Switzerland : International World Wide Web Conferences Steering Committee, 2018, pp. 45–46.

Chen, D., Shang, M., Lv, Z., and Fu, Y., "Detecting overlapping communities of weighted networks via a local algorithm," *Physica A : Statistical Mechanics and its Applications*, vol. 389, no. 19, pp. 4177–4187, 2010.

Clauset, A., Newman, M. E. J., and Moore, C., "Finding community structure in very large networks." *Physical Review E*, vol. 70, no. 6, p. 066111, 2004.

Collins, L. M. and Dent, C. W., "Omega : A general formulation of the rand index of cluster recovery suitable for non-disjoint solutions." *Multivariate Behavioral Research*, vol. 23, no. 2, pp. 231–242, 1988.

Coscia, M., Rossetti, G., Giannotti, F., and Pedreschi, D., "DEMON : A local-first discovery method for overlapping communities," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12*, Beijing, China, August 2012, p. 615–623.

Danon, L., Díaz-Guilera, A., Duch, J., and Arenas, A., "Comparing community structure identification." *Journal of Statistical Mechanics : Theory and Experiment*, vol. 2005, no. 9, p. P09008, 2005.

Demirguc-Kunt, A., Klapper, L., Singer, D., Ansar, S., and Hess, J., *The Global Findex Database 2017 Measuring Financial Inclusion and the Fintech Revolution.* The World Bank, 2018. [Online]. Available : https://books.google.fr/books?id=lYxaDwAAQBAJ

Di Clemente, R., Luengo-Oroz, M., Travizano, M., Xu, S., Vaitla, B., and González, M. C., "Sequences of purchases in credit card data reveal lifestyles in urban populations." *Nature Communications*, vol. 9, no. 1, 2018.

Diniz, E., Porto de Albuquerque, J., and Cernev, A., "Mobile money and payment : A literature review based on academic and practitioner - oriented publications (2001 - 2011)," in *Proceedings of SIG GlobDev Fourth Annual Workshop*, Shanghai, China, December 2011.

Duan, D., Li, Y., Jin, Y., and Lu, Z., "Community mining on dynamic weighted directed graphs." in *Proceedings of the 1st ACM international workshop on Complex networks meet information & knowledge management (CNIKM '09)*, ser. CNIKM '09. Association for Computing Machinery, November 2009, pp. 11–18.

Ducruet, C., "Multigraphes, multiplexes, et réseaux couplés." CNRS, UMR Géographie-cités, Technical Report, 2012. [Online]. Available : https://shs.hal.science/halshs-00746129

El Ayeb, S., Hemery, B., Jeanne, F., and Cherrier, E., "Community detection for mobile money fraud detection." in *Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, Paris, France, December 2020, pp. 1–6.

El Ayeb, S., Hemery, B., Jeanne, F., Charrier, C., and Cherrier, E. P., "Multigraph transformation for community detection applied to financial services," in *2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, November 2022.

El Ayeb, S., Hemery, B., Jeanne, F., Cherrier, E., and Charrier, C., "Evaluation metrics for overlapping community detection," in *2022 IEEE 47th Conference on Local Computer Networks (LCN)*, 2022, pp. 355–358.

Emmons, S., Kobourov, S., Gallant, M., and Börner, K., "Analysis of network clustering algorithms and cluster quality metrics at scale." *PLoS ONE*, vol. 11, no. 7, p. e0159161, 2016.

Euler, L., "Solutio problematis ad geometriam situs pertinentis," vol. 53, pp. 128–140, 1741.

Farkas, I., Ábel, D., Palla, G., and Vicsek, T., "Weighted network modules," *New Journal of Physics*, vol. 9, no. 6, pp. 180–180, 2007.

Fortunato, S., "Community detection in graphs." *Physics Reports*, vol. 486, no. 3, pp. 75–174, 2010.

Freeman, L. C., "A set of measures of centrality based on betweenness." *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.

——, "Centrality in social networks conceptual clarification." *Social networks*, vol. 1, no. 3, pp. 215–239, 1979.

Fuhrer, C. and Cucchi, A., "Relations between social capital and use of ict : A social network analysis approach." *International Journal of Technology and Human Interaction*, vol. 8, no. 2, pp. 15–42, 2012.

Girvan, M. and Newman, M. E. J., "Community structure in social and biological networks." *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

Goldberg, M. K., Hayvanovych, M., and Magdon-Ismail, M., "Measuring similarity between sets of overlapping clusters." in *IEEE Second International Conference*

on Social Computing (SocialCom), Minneapolis, Minnesota, August 2010, pp. 303–308.

Gregory, S., "An algorithm to find overlapping community structure in networks." in *11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, Warsaw, Poland, September 2007, pp. 91–102.

——, "Finding overlapping communities in networks by label propagation." *New Journal of Physics*, vol. 12, no. 10, p. 103018, 2010.

GSMA, "State of the industry report on mobile money 2021," 2021, https://www.gsma.com/sotir/, Last accessed on 12-07-2022.

Hawe, P., Webster, C., and Shiell, A., "A glossary of terms for navigating the field of social network analysis." *Journal of Epidemiology & Community Health*, vol. 58, no. 12, pp. 971–975, 2004.

Higaki, A., Uetani, T., Ikeda, S., and Yamaguchi, O., "Co-authorship network analysis in cardiovascular research utilizing machine learning (2009–2019)," *International Journal of Medical Informatics*, vol. 143, p. 104274, 2020.

Hollocou, A., Bonald, T., and Lelarge, M., "Improving pagerank for local community detection." 2016.

Hubert, L. and Arabie, P., "Comparing partitions." *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.

Humski, L., Pintar, D., and Vranić, M., "Analysis of facebook interaction as basis for synthetic expanded social graph generation," *IEEE Access*, vol. 7, pp. 6622–6636, 2018.

Iñiguez, G., Battiston, F., and Karsai, M., "Bridging the gap between graphs and networks," *Communications Physics*, vol. 3, no. 1, pp. 1–5, 2020.

Jebabli, M., Cherifi, H., Cherifi, C., and Hamouda, A., "Overlapping community detection versus ground-truth in amazon co-purchasing network," in *11th international conference on signal-image technology & internet-based systems (SITIS)*. IEEE, 2015, pp. 328–336.

Kamau, N., Margret, W., and Hillary, B., "Structural analysis of social networks revealed by small holder banana farmers in muranga county, kenya." *Journal of Agricultural Science and Food Research*, vol. 9, no. 2, p. 6, 2018.

Kanawati, R., "Détection de communautés dans les grands graphes d'interactions (multiplexes) : état de l'art," Laboratoire d'Informatique de Paris-Nord (LIPN), Technical Report, 2013. [Online]. Available : https://hal.science/hal-00881668

Kim, P. and Kim, S., "A detection of overlapping community in mobile social network." in *Proceedings of the 29th Annual ACM Symposium on Applied Computing (SAC '14)*, New York, USA, 2014, pp. 175–179.

Krebs, V. E., "Mapping networks of terrorist cells." *Connections*, vol. 24, no. 3, pp. 43–52, 2002.

Kretschmer, H. and Kretschmer, T., "A new centrality measure for social network analysis applicable to bibliometric and webometric data." *COLLNET Journal of Scientometrics and Information Management*, vol. 1, no. 1, pp. 1–7, 2007.

Lancichinetti, A. and Fortunato, S., "Community detection algorithms : A comparative analysis." *Physical Review E*, vol. 80, no. 5, p. 056117, 2009.

Lancichinetti, A., Fortunato, S., and Kertesz, J., "Detecting the overlapping and hierarchical community structure of complex networks." *New Journal of Physics*, vol. 11, no. 3, p. 033015, 2009.

Lancichinetti, A., Kivelä, M., Saramäki, J., and Fortunato, S., "Characterizing the Community Structure of Complex Networks," *PLOS ONE*, vol. 5, no. 8, p. e11976, 2010.

Lancichinetti, A., Radicchi, F., Ramasco, J. J., and Fortunato, S., "Finding statistically significant communities in networks." *PloS one*, vol. 6, no. 4, p. e18961, 2011.

Lazega, E., *Réseaux sociaux et structures relationnelles.* Presses Universitaires de France, 2007.

Lopez-Rojas, E., Elmir, A., and Axelsson, S., "Paysim : A financial mobile money simulator for fraud detection." in *28th European Modeling and Simulation Symposium (EMSS)*, Larnaca, Chypre, September 2016, pp. 249–255.

Lopez-Rojas, E. A., Gorton, D., and Axelsson, S., "Retsim : A shoestore agent-based simulation for fraud detection." in *25th European Modeling and Simulation Symposium (EMSS)*, Athens, Greece, September 2013, pp. 25–34.

Lutov, A., Khayati, M., and Cudré-Mauroux, P., "Accuracy evaluation of overlapping and multi-resolution clustering algorithms on large datasets." in *The 6th IEEE International Conference on Big Data and Smart Computing (BigComp)*, Kyoto, Japan, February 2019, pp. 1–8.

Majmudar, J. and Vavasis, S., "Provable overlapping community detection in weighted graphs." *Advances in Neural Information Processing Systems*, vol. 33, no. 1, pp. 19 028–19 038, 2020.

Milgram, S., "The small world problem." *Psychology today*, vol. 2, no. 1, pp. 60–67, 1967.

Mittal, R. and Bhatia, M. P. S., "Classification and comparative evaluation of community detection algorithms." *Archives of Computational Methods in Engineering*, vol. 28, p. 1417–1428, 2020.

Moon, S., Lee, J.-G., Kang, M., Choy, M., and Lee, J.-w., "Parallel community detection on large graphs with MapReduce and GraphChi," *Data & Knowledge Engineering*, vol. 104, pp. 17–31, 2016.

Moreno, J. L., *Who shall survive ? : A new approach to the problem of human interrelations*. Nervous and Mental Disease Publishing Co, 1934.

Morrison, J. L., Breitling, R., Higham, D. J., and Gilbert, D. R., "Generank : using search engine technology for the analysis of microarray experiments." *BMC bioinformatics*, vol. 6, no. 1, pp. 1–14, 2005.

Nepusz, T., Petróczi, A., Négyessy, L., and Bazsó, F., "Fuzzy communities and the concept of bridgeness in complex networks." *Physical Review E*, vol. 77, no. 1, p. 016107, 2008.

Newman, M. E. J., "Finding community structure in networks using the eigenvectors of matrices." *Physical Review E*, vol. 74, no. 3, p. 036104, 2006.

Newman, M. E. J. and Girvan, M., "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2, p. 026113, 2004.

——, "Finding and evaluating community structure in networks." *Physical Review E*, vol. 69, no. 2, p. 026113, 2004.

Ondrus, J. and Pigneur, Y., "A disruption analysis in the mobile payment market." in *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, Big Island, Hawaii, 2005, pp. 84c–84c.

Page, L., Brin, S., Motwani, R., and Winograd, T., "The pagerank citation ranking : Bringing order to the web." Stanford InfoLab, Technical Report, 1999.

Palla, G., Derényi, I., Farkas, I., and Vicsek, T., "Uncovering the overlapping community structure of complex networks in nature and society," *Nature Research Journals*, vol. 435, no. 7043, pp. 814–818, 2005.

Papalexakis, E. E., Akoglu, L., and Ience, D., "Do more views of a graph help ? community detection and clustering in multi-graphs." in *Proceedings of the 16th*

International Conference on Information Fusion, Istanbul, Turkey, July 2013, pp. 899–905.

Parisutham, N. and Rethnasamy, N., "Eigenvector centrality based algorithm for finding a maximal common connected vertex induced molecular substructure of two chemical graphs." *Journal of Molecular Structure*, vol. 1244, p. 130980, 2021.

Park, J. and Newman, M. E. J., "A network-based ranking system for us college football." *Journal of Statistical Mechanics : Theory and Experiment*, vol. 2005, no. 10, p. P10014, 2005.

Peel, L., Larremore, D. B., and Clauset, A., "The ground truth about metadata and community detection in networks." *Science Advances*, vol. 3, no. 5, p. e1602548, 2017.

Pons, P. and Latapy, M., "Computing communities in large networks using random walks." in *Computer and Information Sciences (ISCIS 2005)*, vol. 3733. Springer, 2005.

Puzis, R., Zilberman, P., Elovici, Y., Dolev, S., and Brandes, U., "Heuristics for speeding up betweenness centrality computation." in *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing (PASSAT)*, Amsterdam, Netherlands, Sept. 2012, pp. 302–311.

Qi, X., Wu, Q., Zhang, Y., Fuller, E., and Zhang, C.-Q., "A novel model for DNA sequence similarity analysis based on graph theory." *Evolutionary Bioinformatics*, vol. 7, p. EBO.S7364, 2011.

Rachman, Z. A., Maharani, W., and Adiwijaya, "The analysis and implementation of degree centrality in weighted graph in social network analysis." in *2013 International Conference of Information and Communication Technology (ICoICT)*, Bandung, Indonesia, March 2013, pp. 72–76.

Raghavan, U. N., Albert, R., and Kumara, S., "Near linear time algorithm to detect community structures in large-scale networks." *Physical Review E*, vol. 76, no. 3, p. 036106, 2007.

Rand, W. M., "Objective criteria for the evaluation of clustering methods." *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.

Rennard, M., "Les innovations africaines, sources d'inspiration pour l'innovation européenne d'orange." *Le journal de l'école de Paris du management*, vol. 143, no. 3, pp. 17–23, 2020.

Rossetti, G., "Exorcising the demon : Angel, efficient node-centric community discovery." in *The 9th International Conference on Complex Networks and Their Applications (VIII). COMPLEX NETWORKS 2019*, ser. Studies in Computational Intelligence, Cherifi, H., Gaito, S., Mendes, J. F., Moro, E., and Rocha, L. M., Eds., vol. 881. Springer, Cham, 2020, pp. 152–163.

Rossetti, G., Pappalardo, L., and Rinzivillo, S., "A novel approach to evaluate community detection algorithms on ground truth," in *Complex Networks VII : Proceedings of the 7th Workshop on Complex Networks (CompleNet)*, ser. Studies in Computational Intelligence. Springer International Publishing, Cham., 2016, pp. 133–144.

Rosvall, M., Axelsson, D., and Bergstrom, C. T., "The map equation." *The European Physical Journal Special Topics*, vol. 178, no. 1, pp. 13–23, 2009.

Rosvall, M. and Bergstrom, C. T., "An information-theoretic framework for resolving community structure in complex networks." *Proceedings of the national academy of sciences*, vol. 104, no. 18, pp. 7327–7331, 2007.

Rutkowski, E., Sargant, J., Houghten, S., and Brown, J. A., "Evaluation of communities from exploratory evolutionary compression of weighted graphs." in *IEEE Congress on Evolutionary Computation (CEC)*, June 2021, pp. 434–441.

Röhm, J., "Social network vizualization using facebook and gephi." 2014, https://www.jroehm.com/2014/10/29/social-network-vizualiation/, Last accessed on 12-07-2022.

Schaub, M. T., Delvenne, J.-C., Rosvall, M., and Lambiotte, R., "The many facets of community detection in complex networks." *Applied Network Science*, vol. 2, no. 1, p. 4, 2017.

Sedighpour, N. and Bagheri, A., "Paslpa - overlapping community detection in massive real networks using apache spark." in *9th International Symposium on Telecommunications (IST)*, Tehran, Iran, December 2018, pp. 233–240.

Shi, J. and Malik, J., "Normalized cuts and image segmentation." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

Society, C. S., "What are complex systems ?" 2022, https://cssociety.org/about-us/what-are-cs, Last accessed on 16-10-2022.

Song, C., Havlin, S., and Makse, H. A., "Self-similarity of complex networks." *Nature*, vol. 433, no. 7024, pp. 392–395, 2005.

Valente, T. W., Coronges, K., Lakon, C., and Costenbader, E., "How correlated are network centrality measures ?" *Connections (Toronto, Ont.)*, vol. 28, no. 1, p. 16, 2008.

Wadhwa, P. and Bhatia, M. P. S., "Social networks analysis : trends, techniques and future prospects." in *4th International Conference on Advances in Recent Technologies in Communication and Computing (ARTCom2012)*, Bangalore, India, October 2012, pp. 1–6.

Wasserman, S. and Faust, K., *Social Network Analysis : Methods and Applications*. Cambridge University Press, 1994.

Watts, D. J. and Strogatz, S. H., "Collective dynamics of 'small-world' networks." *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

Weng, J., Lim, E.-P., Jiang, J., and He, Q., "Twitterrank : Finding topic-sensitive influential twitterers." in *Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM '10)*. New York, NY : Association for Computing Machinery, 2010, pp. 261–270.

White, H., "Management conflict and sociometric structure." *American Journal of Sociology*, vol. 67, no. 2, pp. 185–199, 1961.

Wu, K., Taki, Y., Sato, K., Sassa, Y., Inoue, K., Goto, R., Okada, K., Kawashima, R., He, Y., Evans, A. C., and Fukuda, H., "The overlapping community structure of structural brain network in young healthy individuals." *PLOS ONE*, vol. 6, no. 5, p. e19608, 2011.

Xie, J., Szymanski, B. K., and Liu, X., "Slpa : Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process." in *2011 IEEE 11th International Conference on Data Mining Workshops*, December 2011, pp. 344–349.

Xie, J., Kelley, S., and Szymanski, B. K., "Overlapping community detection in networks : The state-of-the-art and comparative study." *ACM Computing Surveys*, vol. 45, no. 4, pp. 43 :1–43 :35, 2013.

Yamaguchi, Y., Takahashi, T., Amagasa, T., and Kitagawa, H., "Turank : Twitter user ranking based on user-tweet graph analysis." in *Web Information Systems Engineering – WISE*, Hong Kong, China, December 2010, pp. 240–253.

Yang, J. and Leskovec, J., "Overlapping community detection at scale : a nonnegative matrix factorization approach." in *Proceedings of the sixth ACM international conference on Web search and data mining, WSDM '13*. Association for Computing Machinery, 2013, pp. 587–596.

Yeghiazaryan, V. and Voiculescu, I., "Family of boundary overlap metrics for the evaluation of medical image segmentation." *Journal of Medical Imaging*, vol. 5, no. 1, p. 015006, 2018.

Yustiawan, Y., Maharani, W., and Gozali, A. A., "Degree centrality for social network with opsahl method," *Procedia Computer Science*, vol. 59, pp. 419–426, 2015.

Zachary, W. W., "An information flow model for conflict and fission in small groups," *Journal of anthropological research*, vol. 33, no. 4, pp. 452–473, 1977.

Zhang, P., "Evaluating accuracy of community detection using the relative normalized mutual information." *Journal of Statistical Mechanics : Theory and Experiment*, vol. 2015, no. 11, p. 11006, 2015.

Zong-Wen, L., Jian-Ping, L., Fan, Y., and Petropulu, A., "Detecting community structure using label propagation with consensus weight in complex network." *Chinese Physics B*, vol. 23, no. 9, p. 098902, 2014.

Contribution à la détection de communautés chevauchantes pour l'analyse des réseaux transactionnels complexes.

L'analyse des réseaux sociaux est fondée sur l'étude des interactions sociales pour la compréhension des comportements individuels et collectifs au sein des systèmes complexes. Les réseaux sociaux peuvent être représentés sous forme de graphes, qui sont des structures de données mathématiques, pour les modéliser et étudier leurs propriétés. Une des nombreuses problématiques liées à l'analyse des réseaux sociaux concerne la détection de communautés qui vise à identifier des groupes fortement connectés. Cette thèse est motivée par l'étude de la détection des communautés sur des données de transactions issues du service financier Orange Money. Ces transactions sont modélisées par un multigraphe où les nœuds représentent les utilisateurs du service, et les liens représentent leurs échanges. Au cours de cette thèse, on s'intéresse à la détection de communautés chevauchantes. Ce type de communautés reflète bien la réalité en associant chaque individu à plusieurs communautés à la fois.

Contribution to the detection of overlapping communities for the analysis of complex transactional networks.

Social network analysis is based on the study of social interactions to understand individual and collective behaviors in complex systems. Social networks can be represented as graphs, which are mathematical data structures, to model them and study their properties. One of the many problems related to the analysis of social networks concerns the detection of communities, which aims at identifying strongly related groups. This thesis is motivated by the study of community detection on transaction data from the Orange Money financial service. These transactions are modeled by a multigraph where the nodes represent the users of the service, and the links represent their exchanges. In this thesis, we are interested in the detection of overlapping communities. This type of community reflects reality by associating each individual to several communities simultaneously.

Mots clés : Analyse des réseaux sociaux, détection de communautés, multigraphe, métriques d'évaluation, réseaux complexes.

Normandie Univ., UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen, France