الجمـــهوريـــة الجـــزائريـــة الديـمقراطيـــــــة الشـعبيـــــة

République Algérienne Démocratique ET Populaire

وزارة التعـلـيـم الـعـالـــي و الـبـحـث الـعـلـمـي

Ministère de l'enseignement supérieur et de la Recherche Scientifique

جـامـعـة عمّـــار ثـلـيـــجـــي بالأغـــواط

Université Amar Telidji de Laghouat

كـلـيـــــة الـعـلـــــوم

Faculté des sciences

Département de mathématique et informatique

Mémoire de Master

Domaine : Mathématique et informatique (MI)

Filière : Informatique

Option : Système d'Iformation et de Décision

PAR : **GUENDOUZ AYOUB ABDESSAMAD**

Thème

Étude comparative de méthodes de détection de communautés dans les réseaux sociaux en ligne

Soutenu publiquement devant le jury composé de :

Mr L.CHELLAMA	Université de Laghouat	Président
Mr A.LAKHDARI	Université de Laghouat	Examinateur
Mlle S.BENKOUIDER	Université de Laghouat	Examinatrice
Mr Y.OUINTEN	Université de Laghouat	Encadreur
Mr M.BOUAKKAZ	Université de Laghouat	Co-encadreur

ANNÉE UNIVERSITAIRE 2015/2016

Dédicaces

“ Je dédie ce mémoire

à

mes parents

mes frères et sœurs

et toute ma famille

et

mes amis ”

Remerciements

Je remercie tout d'abord ALLAH, le tout-puissant de m'avoir donné la force et la patience pour accomplir ce travail.

Mes remerciements s'adressent également à toutes les personnes qui ont contribué de près ou de loin avec leurs conseils ou avec leurs encouragements à l'accomplissement de ce travail.

Je tiens à exprimer ma sincère reconnaissance et remerciements à Mr. Yousef OUINTEN, professeur à l'université amar telidji de laghouat d'avoir accepté d'encadrer et de diriger mes travaux. Mes remerciement vont aussi à mon co-encadreur Mr. Mustapha BOUAKKAZ qui n'a pas cessé de m'aider et de m'encourager pour l'accomplissement de ce mémoire.

Je remercie tous le personnel de l'université d'Amar TELIDJI de Laghouat, l'université qui m'a accueilli bras ouvert, mes remerciements vont particulièrement aux enseignants et administrateurs du département de mathématique et informatique

Enfin, j'exprime mes vifs remerciements à toute ma famille et spécialement à mes parents.

Résumé

L'analyse de réseaux sociaux est un outil qui s'impose dans de nombreux domaines de la science. Un de ces outils spécifiques à l'analyse de réseaux sociaux est la détection de communautés. De nombreux algorithmes de détection de communautés ont été développés dans ce domaine.

L'objectif de ce travail est de faire une étude comparative entre deux algorithmes de détection de communautés dans les réseaux sociaux en ligne, en détaillant leurs principes.

L'algorithme optimal de détection de communautés, de manière à ce qu'il satisfasse efficacement le problème de la minimisation des liens inter-communautés, et la maximisation des liens intracommunautés se pose comme problématique dans ce travail.

Dans le contexte de ce travail, nous avons choisi deux algorithmes pour appliquer les expérimentations. Ces algorithmes sont : l'algorithme CPM la méthode de clique percolation (Clique Percolation Method), et l'algorithme de BOUAKKAZ (Diamant).

Nous avons comparé et interprété les résultats de ces algorithmes par les mesures d'évaluer de performance communément utilisées : le rappel, la précision, la F-mesure et la modularité.

Mots-clés : Analyse de réseaux sociaux, détection de communautés, CPM (la méthode de clique percolation), BOUAKKAZ (Diamant), les mesures de performance.

Abstract

The social network analysis is a tool that is required in many areas of science. One such specific tools for social network analysis is community detection. Many of community detection algorithms have been developed in this area.

The objective of this work is to make a comparative study between two communities detection algorithms in the online social networks detailing their principles.

The optimal algorithm for community detection, so that it effectively meets the problem of minimizing inter-community links, and maximizing intracommunautés links arises as an issue in this work.

In the context of this work, we chose two algorithms to implement the experiments. These algorithms are : the algorithm CPM (Clique Percolation Method) and Bouakkaz algorithm (Diamond).

We compared and interpreted the results of these algorithms by measures to evaluate performance commonly used : the recall, precision, F-measure and modularity.

Keywords : social network analysis, community detection, CPM (clique percolation method), Bouakkaz (Diamond), performance measures.

Table des matières

Table des figures

Page

Liste des tableaux

Page

Introduction

Parmi les différents types de médias sociaux sont les réseaux sociaux qui sont devenus une grande partie de notre vie sociale qui apparaissent avec la prospérité de l'internet et du web 2.0.

Ils permettent aux personnes de se connecter entre eux en offrant des options et des services très attirants qui facilitent leurs utilisations, à cause de sa concurrence entre eux pour satisfaire les besoins de ces millions d'utilisateurs

Après la grande évolution de tous ces réseaux sociaux, les chercheurs se sont intéressé à ce type de réseau en focalisant leurs analyses pour explorer et détecter les différentes communautés qui sont sous-jacentes à ces réseaux. C'est le sujet qui est au cœur de ce mémoire, où nous nous intéressons à la détection de communautés dans les réseaux sociaux.

La détection de communautés dans les réseaux sociaux en ligne est un sujet récent et vaste pour l'ensemble de la théorie des réseaux sociaux qui base sur la théorie des graphes. Nous nous intéressons à déterminer l'appartenance de chaque acteur.

Dans le cadre de notre travail, nous avons organisé ce mémoire en quatre chapitres.

Dans le premier chapitre, nous présentons des généralités sur les réseaux sociaux en définissant les médias sociaux, les réseaux et les réseaux sociaux et de présenter certains services et applications des réseaux sociaux.

Le deuxième chapitre est dédié à l'état de l'art, nous commençons par les concepts de l'analyse des réseaux sociaux et leurs tâches et quelques notions de la théorie des graphes en consacrant le reste du chapitre aux détections des communautés et leur problématique.

Dans le troisième chapitre, nous expliquons les deux algorithmes utilisés dans l'étude comparative en détail en donnant des exemples et en expliquant les mesures de performance utilisées.

Dans le quatrième chapitre, nous commençons par présenter l'environnement de travail, ensuite nous décrivons les algorithmes choisis et leurs mesures de performance. À la fin nous avons interprété et comparé les résultats trouvés après l'exécution de chaque algorithme de cette étude.

Nous concluons le mémoire par une conclusion générale.

Chapitre 1

Généralités sur les réseaux sociaux en ligne

Avec la prospérité de l'internet, et du web 2.0, de nombreux réseaux sociaux et sites de médias sociaux apparaissent. Ils sont en compétition pour présenter les meilleurs services et offres afin d'attirer plus d'utilisateurs. Les gens peuvent facilement se connecter les uns aux autres dans l'espace cybernétique. De nos jours, les réseaux sociaux sont largement utilisés. Certains utilisateurs ont restreint l'utilisation d'internet aux blogs et médias sociaux. Il y a ceux qui vérifient leur SRS (Service de réseautage social) dès la première heure le matin, et avant de se coucher. Les SRS sont devenus une addiction. Il existe plusieurs services de réseautage tels que Facebook, Twitter, YouTube ; cependant, la liste n'est pas limitée à ces trois sites.

1.1 Médias sociaux

Les médias sociaux sont les collectifs de canaux de communication en ligne et technologies dédiées pour créer des communautés en ligne, l'interaction, la publication, la discussion, la messagerie, le partage de contenu et de collaboration. Les sites web et applications dédiées à des forums, microblogging, réseaux sociaux, et les wikis sont parmi les différents types de médias sociaux [CCS+10].

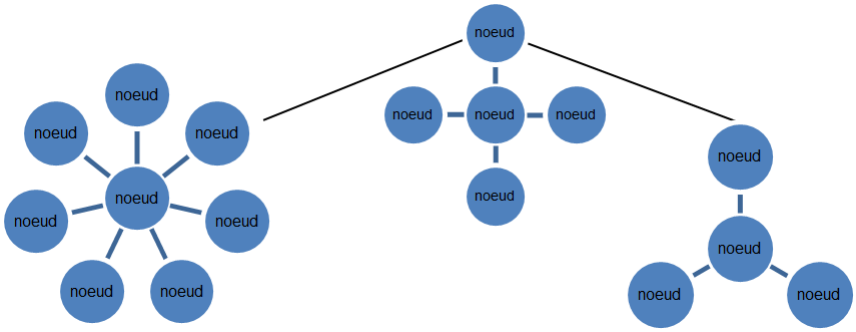

FIGURE 1.1: Un réseau

1.2 Un réseau (un terme polysémique)

Historiquement et étymologiquement, le réseau est un filet, un tissu, ou un entrelacement de fils et les figures de sa généalogie montrent que cette référence originelle est persistante jusqu'à nos jours [Zam12]. Et dans le contexte de notre mémoire le réseau est un ensemble de nœuds reliés entre eux par des liens [Bac14], figure 1.1.

1.3 Un réseau social

Un réseau social est un ensemble d'identités sociales, telles que des individus, ou encore des organisations, reliés entre eux par des liens créés lors d'interactions sociales [WF94]. Deux termes sont importants dans cette définition :

(i) les individus ou organisations qui sont généralement caractérisés par des propriétés (ou attributs) qui constituent pour nous les informations de leurs profils.

(ii) Les liens entre les individus ou organisations qui constituent l'élément majeur donnant un sens au réseau social (interactions entre amis, collègues, contacts téléphoniques, ...etc).

1.4 Les services et les applications des réseaux sociaux en ligne

A ce jour, il y a des centaines de sites de réseaux sociaux, avec potentialités technologiques diverses, en soutenant un large éventail d'intérêts et de pratiques.

Bien que leurs caractéristiques technologiques clés sont assez uniformes, les cultures qui ont émergé autour de sites de réseaux sociaux sont variés.

Il existe plusieurs types de réseaux sociaux qui permettent de satisfaire les différents besoins des utilisateurs [Via15]

1.4.1 Les réseaux sociaux populaire

1.4.1.1 Facebook

Facebook [1] est le deuxième site le plus visité selon Alexa [2] . Il a été fondé par Mark Zuckerberg en 2004 qui a commencé à Harvard et diffusé à d'autres universités avant d'être pour tout le monde, il permet l'échange de vidéos, images, fichiers et textes. L'échange peut être via des groupes postes, des messages privés ou des postes de murs. Il permet de partager les liens, de créer des pages et des groupes, de publier des applications, des publicités et des jeux. Il est disponible en 70 langues. Il y a 1,55 milliard d'utilisateurs de facebook tandis que jamais dans l'historique de l'humanité un média n'avait réussi à toucher plus d'un milliard d'utilisateurs [Cav16], dont 1,09 milliard d'utilisateurs actifs quotidiens en moyenne en Mars 2016 qui passent en moyenne 6h45 par mois, le nombre moyen de photos publiées sont 300 millions de photos par jour ; et 500 millions de likes quotidiens. Facebook propose également des solutions efficaces pour aider les entreprises à contacter leurs clients [Via15].

FIGURE 1.2: Logo facebook

1. www.facebook.com
2. Alexa : société fondée en 1996 et l'une des groupes d'AMAZON, son site web fourni les statistiques et les classements des sites web selon leurs nombre de visiteurs.

1.4.1.2 Twitter

Twitter [3] est l'un des réseaux sociaux les plus utilisés, il permet l'échange facile des informations et des photos. La longueur maximale d'un message sur twitter est limitée à 140 caractères. Le message peut être un tweet ou un message privé. L'échange de messages est permis si l'émetteur suit le récepteur et le récepteur suit l'émetteur. Twitter a été lancé en 2006 par Jack Dorsey, il a plus de 320 millions d'utilisateurs actifs par mois et 1 milliards de Tweets sont envoyés par mois, il est disponible en 35 langues. Contrairement à Facebook, centré sur le réseau d'amis proches, Twitter permet de suivre librement n'importe quel utilisateur (ami, marque, personnalité), tandis que certains utilisateurs, l'utilisent quasiment comme un chat public pour interagir avec un noyau dur d'amis : entreprises, marques, médias, journalistes et créateurs s'en servent également comme d'un outil de promotion efficace.

FIGURE 1.3: Logo Twitter

1.4.1.3 Google Plus (G+)

Google+ [4] est un service à mi-chemin entre Facebook et Twitter lancée 2011 et concurrence Facebook en se positionnant derrière : deuxième plus grand réseau social au monde. Au lancement, Google + est accessible sur invitation avant d'être ouvert au grand public. Ces invitations, envoyées aux personnalités du monde informatiques, ont ajouté un aspect de privilège et ont suivi la communication de Gmail. Il permet en effet de communiquer avec vos amis et votre famille en limitant la visibilité de vos messages et photos à un groupe défini de personnes (grâce aux « cercles »). Pour autant, des utilisateurs pourront vous suivre sans que vous ayez besoin de les accepter en tant qu'amis au préalable. Les pages entreprises permettent aux marques de communiquer vers leurs clients. En bref, ce n'est pas réellement Google + qui est utilisé mais toutes les autres fonctionnalités de Google en parallèle.

3. www.twitter.com
4. www.plus.google.com

FIGURE 1.4: Logo Google Plus

1.4.1.4 Tumblr

Tumblr [5] est une plateforme de microblogging et de réseautage. Il a été fondé par David Karp en 2007. Etymologie du terme : de l'anglais « a-tumble » qui signifie pêle-mêle ou encore « tumble » [KKL12], il contient plus de 225 millions de blogs et plus de 104 milliards de publications de types : images, vidéos, audios, citations, ou liens. Tumblr est une plateforme permettant de publier des textes, citations, liens, photos, sons et vidéos de manière ultra-simple sans passer par la création fastidieuse d'un blog, mais avec des possibilités intéressantes de customisation au niveau du design de votre page.

FIGURE 1.5: Logo Tumblr

5. www.tumblr.com

1.4.2 Les réseaux sociaux professionnels

1.4.2.1 Linkedin

Linkedin [6] est le plus populaire des réseaux sociaux dédiés aux professionnels créé en 2003 par Reid Hoffman et Allen Blue et trois autres entrepreneurs qui permettant de valoriser ses compétences et de se connecter avec son réseau : amis, collègues, partenaires, clients, avec 300 millions d'utilisateurs au monde. Il aide les personnes à rechercher un emploi, adresser son CV aux entreprises présentés sur le réseau. Les utilisateurs de LinkedIn habituellement affiliés à leur réseau de travail utilisent le site pour maintenir une liste des contacts de personnes qu'ils connaissent et avoir une confiance au sein de leur ligne de travail. Ce réseau de contacts est utilisé pour maintenir la communication, l'information commerciale, et se référer les uns aux autres.

FIGURE 1.6: Logo Linkedin

1.4.2.2 Viadeo

Viadeo [7] a été créé en 2004 par Dan Serfaty et Thierry Lunati. C'est d'abord une plateforme destinée à faciliter les interactions entre les membres du club d'entrepreneurs qu'ils ont fondé : Agregator. La première version de Viadeo se nomme Viaduc [Pal15].

Après une première levée de fonds pour 5 millions d'euros en 2006, le groupe atteint 1 millions de membres en 2007 et Viaduc change de nom pour devenir Viadeo.

Principaux concurrents : LinkedIn, Xing et le petit nouveau Amplement.

Dix ans plus tard Viadeo compte 65 millions de membres et des filiales en Chine, Russie, Afrique du Nord. Il est le numéro 1 en Chine avec 25 millions de membres

6. www.linkedin.com
7. www.viadeo.com

et en France, avec 9 millions d'utilisateurs. Il est disponible en huit langues : arabe, français, anglais, espagnol, italien, portugais, russe et allemand.

Viadeo est un réseau très implanté régionalement en France et tourné dès l'origine vers les petites et moyennes entreprises et les TPE (Très Petite Entreprise), ce qui ne l'empêche pas de compter les plus grandes sociétés internationales parmi ses membres. Il permet à l'utilisateur de réseauter, créer et étendre leur réseau professionnel, construire leur visibilité numérique, trouver des partenaires, des opportunités et des clients, contactez les décideurs, suivre des entreprises et trouver un emploi.

FIGURE 1.7: Logo Viadeo

1.4.3 Les réseaux sociaux vidéo

1.4.3.1 YouTube

YouTube [8] est tout simplement la plus grande plateforme pour regarder et partager les vidéos en ligne. Il a été crée en 2005 par Steve Chen, Chad Hurley et Jawed Karim, qui étaient des employés de PayPal [9], rachetée par Google en 2006, le premier vidéo publié on YouTube intitulé 'Me at the zoo' montre fondateur Jawed Karim à san diego zoo.

8. www.youtube.com
9. PayPal est un service de paiement en ligne qui permet de payer des achats, de recevoir des paiements, ou d'envoyer et de recevoir de l'argent

FIGURE 1.8: Logo YouTube

1.4.3.2 Dailymotion

Dailymotion [10] est un site français pour partager et regarder les vidéos, a été crée en 2005 par Benjamin Bejbaum et Olivier Poitrey ; est un concurrent de YouTube. Il attire 300 millions d'utilisateurs qui regardent 3,5 milliards de vidéos sur son player chaque mois, à travers le monde. Il offre des fonctionnalités telles que la diffusion dans la TV et dans certains cinémas en paris.

FIGURE 1.9: Logo Dailymotion

1.4.4 Les applications de messagerie sur mobile

1.4.4.1 Facebook Messenger

Facebook Messenger [11] est l'application mobile de Facebook dédiée au chat et à l'envoi de messages gratuits. Facebook Messenger permet même d'appeler gratuitement les amis entre eux depuis le téléphone.

10. www.dailymotion.com
11. www.messenger.com

FIGURE 1.10: Logo Messenger

1.4.4.2 Viber

Viber [12] est permettant d'écrire des textos gratuitement auprès de tous les amis équipés de l'application. Il est disponible dans un version desktop. Il utilise le numéro de téléphone pour identifier les utilisateurs.

FIGURE 1.11: Logo Viber

1.5 L'addiction aux réseaux sociaux

De nos jours, il est difficile pour un internaute de passer à côté des réseaux sociaux. Selon une étude américaine, de nombreux utilisateurs, notamment les plus jeunes, développeraient une réelle addiction aux réseaux sociaux.[Pai16] Une vidéo de Maitre Chat postée sur YouTube revient sur ce phénomène.

Peut-on être accro aux réseaux sociaux ?

A en croire, une récente étude américaine et une vidéo française qui l'illustre, la réponse est oui, l'addiction aux réseaux sociaux est une réalité. Guetter les commen-

12. www.viber.com

taires de ses amis sur Facebook, ajouter une photo sur son site Instagram, publier un tweet, etc., les raisons de flâner sur tous ces réseaux sociaux sont nombreuses.

Il est en outre, aujourd'hui, facile de rester connecter en permanence avec son smartphone. Cette vidéo révèle de 48% des personnes âgées de 18 à 34 ans consultent leur compte Facebook dès le réveil. Les internautes français passent, eux, en moyenne 1h45 par jour sur les réseaux sociaux. Bien que les consultations de ces pages web soient généralement courtes, elles sont répétées jusqu'à 14 fois par jour, un mécanisme d'addiction qui est comparable à celle du tabac par ses effets sur le cerveau.

Une dépendance qui génère du stress et de l'anxiété

Ainsi, les interactions des internautes sur les différents réseaux sociaux entrainent des réactions positives dans leurs cerveaux. Cette dépendance est encore accentuée par la réception de notifications qui sont considérées comme une récompense et motivent l'utilisateur à retourner sur ses sites préférés, un véritable cercle vicieux.

Un tel comportement n'est pas sans conséquence. En effet, l'abus de réseaux sociaux aurait un effet négatif sur l'humeur et le sommeil et peut entrainer du stress, de l'anxiété et même provoquer une dépression. Ces risques diminuent par le sevrage, alors il serait certainement profitable de savoir déconnecter de temps en temps.

Conclusion

Dans ce chapitre, nous avons présenté quelques concepts et définitions liés aux réseaux sociaux, avec quelques dates sur leur historique, et les différents types de ces réseaux et leurs services et applications. On termine le chapitre, par l'impact de l'addiction de ces réseaux sur notre vie quotidienne.

Chapitre 2

État de l'art

Dans ce chapitre, nous allons présenter une brève revue de littérature sur la détection de communautés. Comme il existe de nombreuses approches proposées, nous allons retenir celles ayant le plus d'intérêts de la part de la communauté scientifique. Ces approches illustrent aussi la diversité de méthodologies, et donnent une vue d'ensemble de techniques proposées, selon leurs principes méthodologiques. Mais nous allons commencer d'expliquer c'est quoi l'analyse des réseaux sociaux et pourquoi l'utiliser ; et d'expliquer leurs tâches.

2.1 Analyse des réseaux sociaux

L'analyse des réseaux sociaux est menée dans le domaine des sciences sociales depuis les années 1930 [BD07]. Cette analyse vise d'une part, à identifier les structures sociales distinctes dans les réseaux, et d'autre part, à expliquer le comportement des individus au sein de ces structures sociales, au moyen des études ethnographiques, de modèles mathématiques (théorie des graphes) ou d'éléments de la sociométrie. L'accessibilité de plus en plus grandissante des données sociales des utilisateurs avec l'explosion du Web 2.0 a ouvert la voie à des expérimentations sociales ou automatisées beaucoup plus importantes [BD07]. L'analyse de réseaux sociaux "réels" fait partie du domaine d'étude des " grands graphes réels", des "réseaux complexes" ou encore les "grands graphes de terrain". La discipline a pris son essor dans les années 90 avec la généralisation d'Internet, l'analyse de la topologie de l'Internet, la mise en place de réseaux pairs à pairs ou encore l'accès à des données de réseautage.

L'analyse de réseaux sociaux s'appuie sur les acquis de la théorie des graphes [Sco88] pour formaliser le réseau social comme un ensemble de nœuds et de liens, où chaque nœud modélise un acteur et chaque lien une relation entre deux acteurs. Une valeur peut être affectée à un lien, et représentera alors la force de celui-ci. Elle peut servir à représenter l'importance d'une relation, que ce soit en comptant simplement le nombre d'occurrences de cette relation, ou en prenant en compte d'autres processus de pondération (qualité de l'interaction, système d'évaluation, ...etc).[NED15]

2.2 Préliminaires sur les graphes

La théorie des graphes est devenue un domaine très important pour l'analyse de réseaux sociaux. Donc, il convient d'introduire au préalable certaines notions et définitions utiles de la théorie des graphes [For10].

Graphe : Un graphe non-orienté G = (V,E) est composé d'un ensemble V de nœuds et d'un ensemble E de paires (non ordonnées) de nœuds nommés arêtes (ou liens).

Nous adoptons les notations suivantes : n représente le nombre de nœuds (n = |V|) et m le nombre d'arêtes (m = |E|), le graphe est dit d'ordre n et de taille m. Les arêtes de graphe peuvent être pondérées grâce à une fonction de poids w permettant de modéliser plus finement les interactions entre nœuds, nous obtenons ainsi un graphe pondéré G = (V,E,w)

Graphe connexe : Un graphe G = (V,E) est connexe si, quels que soient les nœuds u et v de V, il existe un chemin de u vers v.

Graphe complet : Un graphe non orienté G = (V,E) est dit complet si quelque soit la paire (u,v) de V, il existe un arête reliant les deux nœuds u et v.

Sous-graphe : Un sous-graphe d'un graphe G est un graphe constitué de certains nœuds de G et de toutes les arêtes qui les relient.

Clique : Une clique de G est un sous-graphe complet de G. On parle de k-clique pour désigner un graphe complet de k nœuds.

n-clique : Un n-clique est un sous-graphe maximale de telle sorte que la distance entre chaque paire de nœuds n'est pas supérieur à n.

n-clan : Un n-clan est un n-clique, dont le diamètre ne dépasse pas n.

n-club : Un n-club, est un sous-graphe maximal de diamètre n.

k-plex : Un k-plex est un sous-graphe maximal dans lequel chaque nœud est adjacent à tous les autres nœuds du graphe, sauf au plus k d'entre eux

k-core (k-noyau) : est un sous-graphe maximal dans lequel chaque nœud est adjacent à au moins k autres nœuds du sous-graphe

LS-sets (LS-ensembles) : est un sous-graphe de telle sorte que le degré interne de chaque nœud est supérieure à son degré externe.

Degré d'un nœud : Dans un graphe non orienté, le degré d'un nœuds v $\in$ V est le nombre d'arêtes auxquelles ce nœuds appartient.

Voisinage : Le voisinage d'un nœud correspond à l'ensemble de tous ses nœuds adjacents. Autrement dit, l'ensemble des voisins d'un nœuds v.

Distance : La distance entre deux nœuds est la longueur de la plus courte chaîne entre ces nœuds ; elle est aussi appelée distance géodésique.

Diamètre : Le diamètre d'un graphe est la plus grande distance entre deux nœuds de ce graphe.

Densité d'un graphe : La densité D d'un graphe, est définie par le rapport du nombre d'arêtes du graphe sur le nombre d'arêtes d'un graphe complet ayant n nœuds

Matrice d'adjacence : Une matrice d'adjacence A d'un graphe G d'ordre n est une représentation matricielle exactement équivalente au graphe. Cette matrice (n×n) est binaire, $a_{ij} = 1$ s'il existe un lien entre les nœuds n_i et n_j, sinon $a_{ij} = 0$.

2.3 Les grands graphes de terrain (Les réseaux complexes)

On rencontre diverses qualifications pour les graphes (ou réseaux) dans le domaine de la détection de communautés. L'appellation "grands graphes de terrain" vient du fait que ces graphes sont de grandes tailles et obtenus de manière empirique en s'appuyant sur des données réelles. Elle fait référence aux mêmes objets dénommés "Complex network" ou "real-world graphs". Dans ces graphes, est la présence d'une forte densité locale, et d'une faible densité globale du graphe. Cette propriété fondamentale, traduit la capacité des nœuds à se regrouper en clusters ou groupes.

La plupart des graphes de terrain ont en communs des propriétés non-triviales [WS98] :

- Loi de puissance : les distributions des degrés de graphes de terrain sont la plupart du temps hétérogènes. En effet, nous rencontrons beaucoup de nœuds avec un faible degré. Quelques nœuds avec un très fort degré, qui jouent nécessairement des rôles particuliers par rapport aux autres nœuds, et tous les degrés intermédiaires sont aussi observables ;
- Le degré moyen des nœuds est relativement faible par rapport à la taille du graphe. Mais la densité globale est faible contrairement à la densité locale qui est élevée ;
- Les graphes de terrain ont quasiment toujours une composante qui contient la très grande majorité des nœuds. Elle est appelée composante géante ;
- La distance moyenne dans la composante géante est petite par rapport au nombre de nœuds ;
- Le coefficient de clustering est élevé, qu'il quantifie comment les voisins d'un noeud dans un graphe sont bien connecté.

2.4 Les graphes aléatoires

En mathématiques, un graphe aléatoire est un graphe qui est généré par un processus aléatoire. Le premier modèle de graphes aléatoires a été introduit par Paul Erdös et Alfréd Rényi [ER59].

Sommairement, un graphe aléatoire de taille n est un graphe de n nœuds dont nous avons choisi aléatoirement les arêtes, en fixant la probabilité d'avoir une arête entre les paires de nœuds (probabilité identique pour chaque paire de nœuds).

Les graphes aléatoires sont utilisés pour évaluer la complexité en moyenne d'algorithmes, utilisant les graphes, ou encore pour modéliser de vrais réseaux. Ces graphes ont des distributions de degrés homogènes, et un faible coefficient de clustering. Or, les réseaux du monde réel comme le web, l'internet, la collaboration des auteurs, etc, se détournent considérablement de ce modèle aléatoire. De ce fait, ils ne sont pas un bon modèle pour les graphes de terrain. Un domaine de recherche essentiel se consacre d'ailleurs à construire des modèles de graphes permettant d'imiter les propriétés des graphes de terrain [BA99, GL06].

2.5 Les techniques d'analyse des réseaux sociaux

L'analyse des réseaux sociaux comporte une variété de tâches, nous énumérons quelques-uns qui sont parmi les plus pertinentes pour ce domaine [TL10b] :

2.5.1 Analyse de centralité

L'analyse de centralité vise à identifier les acteurs «les plus importants» dans un réseau social. La centralité est une mesure de calibrer l'importance d'un acteur. Cela aide à comprendre l'influence sociale, et de la puissance dans un réseau.

2.5.2 Modélisation du réseau

La modélisation du réseau tente de simuler le réseau du monde réel par l'intermédiaire des mécanismes simples. Tels que les modèles présentés dans des réseaux complexes à grande échelle, peuvent être capturés.

2.5.3 Analyse de l'influence

2.5.3.1 Diffusion de l'information

La diffusion de l'information étudie comment l'information se propage dans un réseau. la diffusion de l'information facilite également la compréhension de la dynamique culturelle.

2.5.3.2 Le marketing viral

La modélisation du processus de diffusion de l'information, en collaboration avec l'analyse de la centralité et de communautés, peut aider à atteindre plus marketing viral rentable. C'est, seul un petit groupe d'utilisateurs sont sélectionnés pour la commercialisation. Heureusement, leur adoption peut influencer les autres membres du réseau, de sorte que l'avantage est maximisé.

2.5.4 Prédiction des liens

Prendre une capture d'un réseau social à l'instant t, permet à prédire de nouvelles interactions entre les membres qui ne sont jamais interagis avant, un exemple illustré dans la figure 2.1.

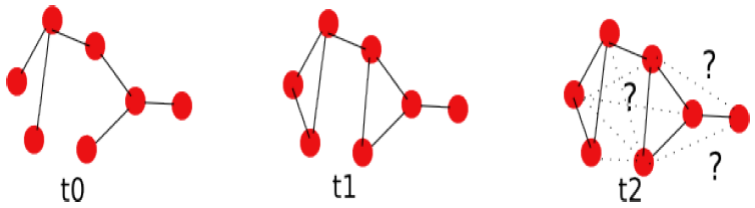

FIGURE 2.1: Prédiction des liens

2.5.5 Détection de communautés

Cette technique identifie les communautés à travers l'étude des structures de réseau et leur topologie ; où leurs acteurs dans un réseau social en ligne forment des groupes.

2.6 Détection de communautés

Au cours des dernières années, et en particulier après la publication de l'article de Girvan et Newman [GN02], la détection de communautés devenue un domaine de recherche très important, et des centaines des algorithmes sont publier, dont l'identification de la structure de communauté a suscité l'intérêt d'un grand nombre de chercheurs, dans diverses disciplines. Notamment, avec l'entrée en jeu des physiciens et des mathématiciens. Depuis, une panoplie de techniques a été proposée, et de nouvelles méthodes continuent régulièrement d'apparaître. Ceci étant, nous pouvons dire que l'identification de la structure de communauté est un domaine vaste, et interdisciplinaire auquel contribuent plusieurs communautés scientifiques.

2.6.1 Problématique

La problématique de détection de communautés dans les réseaux sociaux est de trouver le moyen optimal de détection de communautés, dont le nombre de commu-

nautés que l'on cherche à obtenir, ne peut être connu à l'avance. On peut le définir de la manière suivante : pour un réseau donné, comment le décomposer en un nombre inconnu de communautés de manière à ce qu'elles satisfassent efficacement le problème de la minimisation des liens inter-communautés, et la maximisation des liens intra-communautés.

2.6.2 Définitions de communauté

En raison de la diversité et travaux connexes au domaine de réseaux sociaux, il n'y a pas de définition unique acceptée de la communauté [PKVS12]. En réalité, la définition dépend souvent du type du réseau d'interaction considéré (réseaux sociaux, réseaux biologiques, ...etc).

La communauté peut définir comme la suite : une définition sémantique et l'autre structurelle [NED15], (Figure 2.2).

Définition sémantique : une communauté est un ensemble de nœuds qui partagent les mêmes centres d'intérêt ou ayant le même profil.

Définition structurelle : une communauté est un ensemble de nœuds fortement liés entre eux et faiblement liés avec les autres nœuds du graphe [GN02].

De plus, on peut décrire la communauté comme explicite ou implicite [ZDP14] :

Communautés explicites : sont créés à la suite de la décision humaine, pour acquérir les membres sur la base des intérêts communs. Exemple de ces communautés sont Facebook, LinkedIn, Twitter and Flickr Groups.

Communautés implicites : sont défini par les intérêts des utilisateurs, et les connexions (implicites) entre les utilisateurs ne sont pas explicitement créés par les utilisateurs eux-mêmes, mais purement évoluera en fonction de leurs intérêts tel qu'illustré par leur comportement en ligne. Les communautés implicites sont définies en se référant à la structure du réseau. La notion la plus établie de la communauté a l'intérieur d'un réseau est basé sur le principe que certains ensembles de noeuds sont plus densément connectés les uns aux autres que le reste du réseau. Selon que

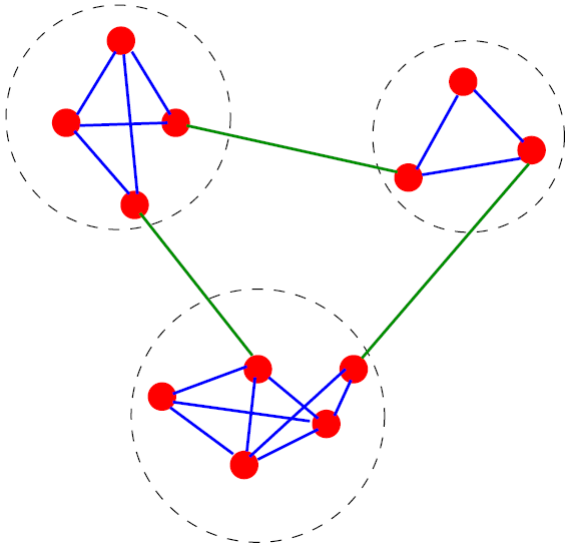

FIGURE 2.2: Un réseau constitué de 3 communautés. [For10]

cette propriété de noeuds est considéré localement (sur un sous-ensemble connecté de noeuds) ou globalement (sur le réseau entier).

2.6.3 Attributs des communautés

Plusieurs attributs qui peuvent distinguer la structure de communauté, parmi ceux-ci :

Communautés chevauchante

Où se trouve des noeuds appartiennent aux plusieurs communautés au même temps. Figure 2.3, les noeuds de couleur rouge sont des noeuds chevauchent entre les communautés

Communautés disjointe

Elle considère que chaque noeud appartient à une seule communauté seulement, figure 2.4.

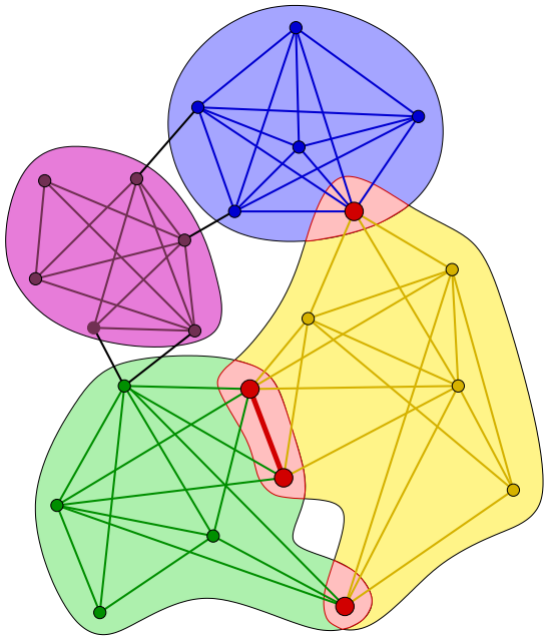

FIGURE 2.3: Communautés chevauchante [PDFV05]

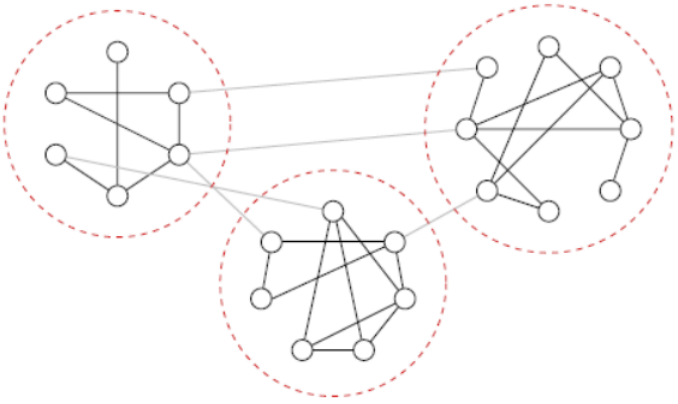

FIGURE 2.4: Communautés disjointes [NG04]

2.6.4 Domaines d'application

En plus des réseaux sociaux, la détection de communautés concerne également de nombreux autres types de réseaux : réseaux biologiques (réseaux métaboliques entre gènes et protéines, réseaux de neurones, ...etc), réseaux d'infrastructures (réseaux de transports, réseaux de distribution électriques, réseau de l'internet ...etc), les réseaux d'information (réseau internet de pages Web), réseaux linguistiques (réseaux de synonymies, réseaux sémantiques, ...etc), ...etc. Cette multitude de domaines d'application suscite le développement de très nombreux méthodes et algorithmes génériques de détection de communautés pouvant être exploités dans chaque domaine.

2.7 Les méthodes de détection de communautés

La diversité des méthodes qui sont apparues dans la littérature pour détecter les communautés est encore plus importante. Il existe des méthodes qui se basent sur le contenu textuel de noeuds dans le réseau, et autres méthodes sont basées sur la structure du graphe. Pour cela nous allons résumer les classes les plus importantes, de telles méthodes en manière méthodologique selon [PKVS12]

Selon le principe méthodologique sous-jacente, ainsi que la définition adoptée de la communauté, nous considérons cinq grandes catégories de détection de communautés et le regroupement de graphe : (A) la découverte de sous-graphe cohésive, (B) le regroupement de noeuds, (C) optimisation de la qualité de la communauté, (D) de division, et (E) à base du modèle (Figure 2.5).

2.7.1 La découverte de sous-graphe cohésive

Les méthodes de cette classe supposent une spécification des propriétés structurales qu'un sous-graphe du réseau doit satisfaire pour être considéré comme une communauté. Une fois qu'une telle structure du sous-graphe est spécifié, la méthode impliquent l'énumération de ces structures dans le réseau en cours d'étude. Comme exemple de structure cohésive, il y a cliques, n-cliques, k-cores, LS-set et lambda set, en outre une méthode telle que CPM (Clique Percolation Method) [PDFV05], et SCAN [XYFS07] sont aussi des méthodes de la même classe.

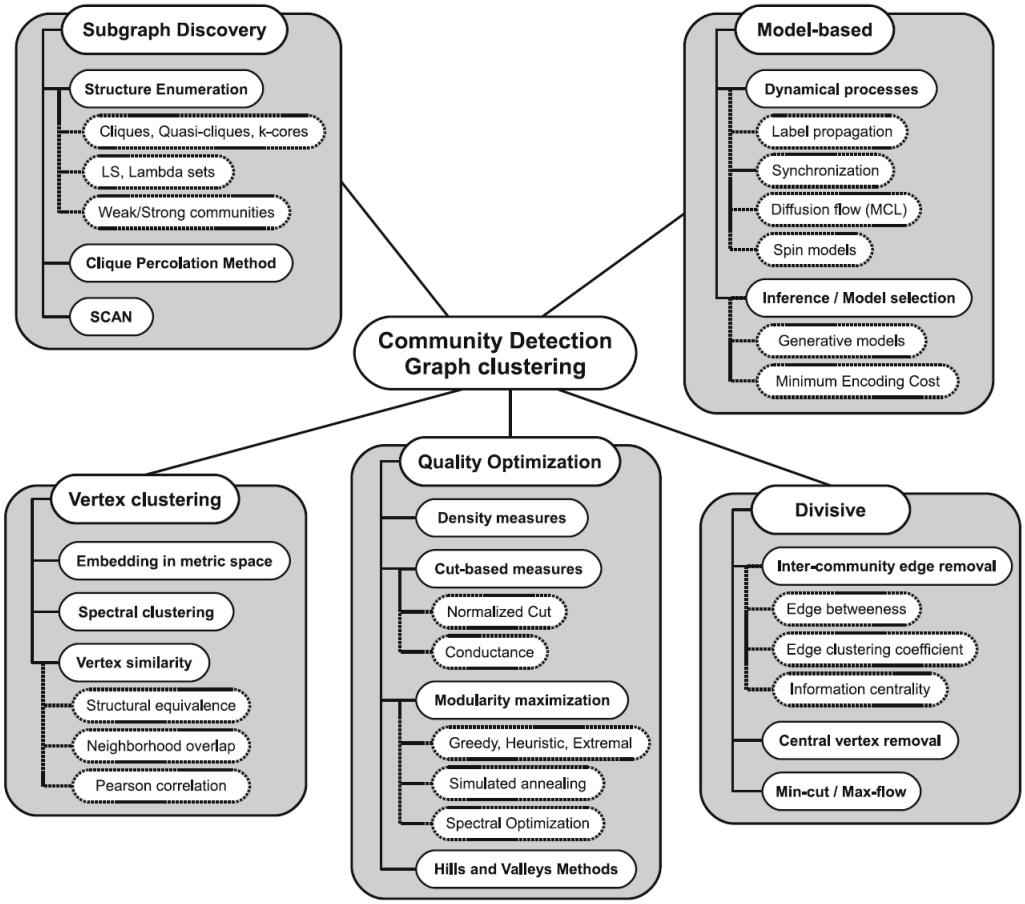

FIGURE 2.5: Classification des méthodes de détection des communautés [PKVS12]

2.7.2 Le regroupement des noeuds

De telles techniques proviennent de la recherche traditionnelle de regroupement de données. Un moyen typique de la coulée d'un problème de regroupement des noeuds, à celui qui peut être résolu par des méthodes de regroupement de données classiques (Comme k-means et le regroupement hiérarchique (agglomératif)) est en intégrant les noeuds du graphe dans un espace vectoriel, où les distances entre les paires des noeuds peuvent être calculées.

Une autre méthode populaire consiste à utiliser le spectre du graphe pour les noeuds du graphe de la cartographie des points dans un espace de faible dimension, où la structure de cluster est plus profonde [DM04, Lux06]. D'autres mesures vertex de similarité telles que l'équivalence structurelle [BBA75] et le chevauchement des

voisins a été utilisé pour calculer la similarité entre les noeuds du graphe [WF94]. Enfin, une méthode remarquable, appelée Walktrap [PL05] qui faite des marches aléatoire basé sur la similarité entre les noueds qui sont des communautés en première étape, pour obtenir finalement à une seule communauté.

2.7.3 Optimisation de la qualité de la communauté

Il y a un très grand nombre de méthodes qui sont fondées sur la base de l'optimisation de certaine mesure à base de graphe, de la qualité de la communauté. Les mesures : la densité du sous-graphe et à base du coupe, telle que la coupe normalisée [SM00] et la conductance [Kan04], étaient parmi les premières à être utilisées pour quantifier la qualité de certaines divisions du réseau en communauté. Une nouvelle vague de recherche a été stimulée par la mesure de la modularité. Les schémas de maximisation de modularité approximatives abondent en littérature, telle que l'algorithme heuristique itératif de louvain le plus populaire [BGLL08]. Autres méthodes appartiennent aussi à ces classes, visant à l'optimisation des mesures locales de qualité de la communauté, tel que la modularité locale et la modularité du sous-graphe [Cla05, LWP06]

2.7.4 Méthodes divisives

Les méthodes de cette classe reposent sur l'identification des éléments de réseau (arêtes et noeuds) qui sont positionnés entre les communautés, qui sont consistant à scinder un réseau en plusieurs communautés en éliminant itérativement les liens entre les nœuds. Ces types commencent par une seule communauté au démarrage, qui est le réseaux en le divisant, en appliquant la mesure de centralité d'intermédiarité jusqu'à avoir n communautés à un seul nœud, l'algorithme Edge Betweenness [NG04] est un exemple de ce type de méthode ; autre exemple, c'est la méthode (min-cut/max-flow) [FLG00, IKN05] adopter une perspective de division différente : ils essaient d'identifier les coupes de graphes (ensembles d'arêtes qui séparent le graphe en pièces), qui ont une taille minimale.

2.7.5 Méthodes à base du modèle

Les Méthodes à base du modèle sont des méthodes, soit qu'ils sont considérés comme un processus dynamique, ce qui révèle ses communautés, ou qu'ils sont considérés comme un modèle sous-jacent de la nature statistique, qui peut générer la division du réseau en communautés. Exemple de processus dynamique, c'est l'algorithme propagation de labels [RAK07] qui se base sur le principe que chaque nœud change de communauté en fonction de la communauté à laquelle appartiennent ses voisins. Un nœud fait partie de la communauté qui contient le plus grand nombre de nœuds voisins. De plus, la détection de la communauté peut être jeté comme un problème d'inférence statistique [Has06], en supposant un certain modèle probabiliste sous-jacente, comme le modèle de partition planté, qui génère la structure de la communauté et l'estimation des paramètres de ce modèle.

2.7.6 Comparaison des algorithmes

Le tableau 2.1 présente une comparaison de certains algorithmes, de ces méthodes de détection de communautés selon la complexité temporelle [PKVS12, For10].

On remarque que l'algorithme 'Louvain' et 'La propagation de labels' ont une meilleur complexité O(n) que les autres algorithmes.

Méthode	Complexité
La découverte de sous-graphe cohésive	
CPM : Clique Percolation Method [PDFV05]	O(exp(n))
Le regroupement des noeuds	
Walktrap [PL05]	O(n^2log n)
Optimisation de la qualité de la communauté	
Louvain [BGLL08]	O(n)
Méthodes divisives	
Min-cut/Max-flow [IKN05]	O(n^3log n)
Méthodes à base du modèle	
La propagation de labels [RAK07]	O(n)

TABLE 2.1: Complexité de certains algorithmes de méthodes de détection de communautés

Conclusion

Dans ce chapitre, nous avons expliqué les tâches d'analyse des réseaux sociaux, et quelques notions des théories des graphes, avant d'introduire dans le domaine de détection de communautés, en donnant par la suite une définition des communautés et leurs attributs et domaines d'application. Enfin les méthodes de détection de communautés existantes dans la littérature et la complexité de certains algorithmes de ces méthodes.

Chapitre 3

Les algorithmes d'étude comparative

Dans ce chapitre, nous allons expliquer les algorithmes de notre étude comparative avec des exemples détaillés, et d'expliquer les mesures choisis pour évaluer la performances des algorithmes ; nous avons choisi deux algorithmes pour la détection de communautés, la méthode de clique percolation CPM (Clique Percolation Method) [PDFV05], et l'algorithme de BOUAKKAZ (Diamant) [Bou16].

3.1 La méthode de clique percolation (CPM : Clique Percolation Method)

La méthode de clique percolation par Palla et al. est basée sur l'hypothèse selon laquelle une communauté se compose d'un ensemble de sous-graphes complets (cliques) chevauchants, et détecte les communautés par la recherche de cliques adjacentes. La CPM construit donc des communautés de k-cliques, qui correspondent aux graphes complets de k noeuds. Elle commence donc par identifier toutes les cliques de taille k dans un réseau. Une fois qu'elles ont été identifiées, un nouveau graphe est construit de telle sorte que chaque noeud représente une de ces k-cliques. Deux nœuds sont alors connectés si les k cliques les représentant partagent k-1 noeuds. Des composants connectés dans le nouveau graphe déterminent quelles cliques composent les communautés. Puisque un noeud peut être dans plusieurs k-cliques en même temps, le chevauchement entre les communautés est possible. La figure 3.1 illustre un exemple de l'algorithme CPM avec k = 3. CPM est adaptée aux réseaux

denses. Empiriquement, les petites valeurs de k (précisément entre 3 et 6) se sont révélées intéressantes [PDFV05, LFK09]. Nous avons choisi l'algorithme de cpm, car il est l'un des meilleurs algorithme de détection de communautés chevauches.

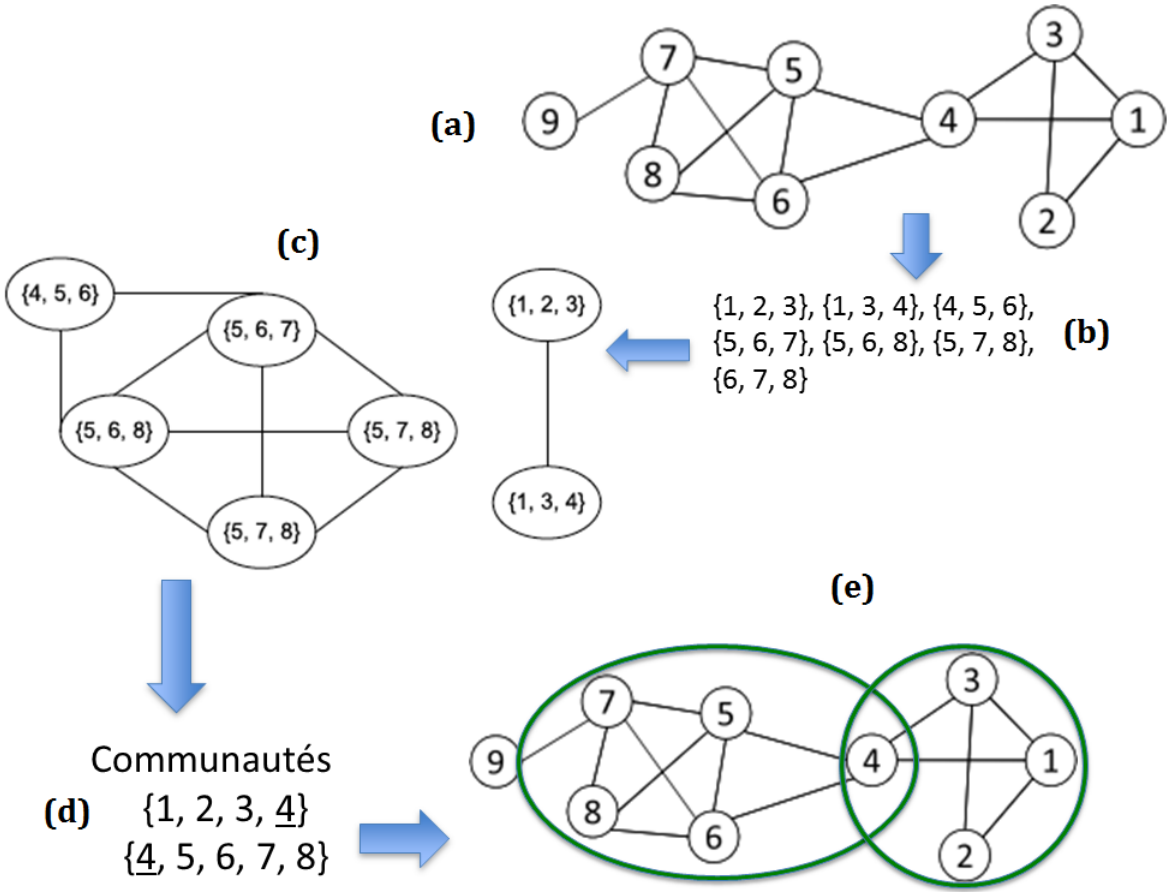

FIGURE 3.1: Exemple de l'algorithme CPM avec k = 3 [TL10a]

Exemple

Dans cet exemple Figure 3.1, nous allons expliquer les étapes à suivre pour détecter les communautés dans cet graphe par la CPM :

(a). Représenter le réseau par un graphe (non orienté, non valué), à partir d'une matrice d'adjacence.

(b). Extraire de tous les cliques de taille k = 3.

(c). Construire le graphe de cliques à partir des cliques qui sont déjà extraites dans l'étape (b), où chaque noeud représente une de ces k-cliques. On construit des composants connectés par une série de k-cliques adjacentes qui atteint l'un de l'autre.

(d). Chaque composant connecté représente une communauté ; dans cet exemple, les cliques {1,2,3} et {1,3,4} partagent k-1 noeuds qui est égale à 2, qui sont les noeuds {1,3} ; ainsi, les deux cliques se réunissent dans une seule communauté qui est {1,2,3,4} ; et ainsi de suite pour le reste de cliques.

(e). Visualiser les communautés sur le graphe, qui sont des communautés chevauchantes entre elles, dont le noeud 4 est un noeud chevauche qui appartient aux deux communautés au même temps, alors que le noeud 9 n'appartient à aucune communauté, car il n'est pas relié avec des noeuds qui permettent de composer un clique de taille 3 parce que leur degré est égale à 1.

3.2 Algorithme de BOUAKKAZ (Diamant)

L'algorithme de BOUAKKAZ (Diamant) a été proposé par [Bou16], c'est un algorithme basé sur l'approche de fonction TAG de mesures textuelles [BLO14]. Cette nouvelle fonction TAG (Textual Aggregation by Graph) utilisée pour l'agrégation textuelle. Cette approche consiste à extraire à partir d'un ensemble de termes celui les plus représentatifs du corpus en utilisant un graphe. Il prend en entrée tous les termes extraits de corpus. Pour extraire les motifs fréquents cette approche passe par 3 principales étapes :

1. L'extraction des mots-clés avec leurs fréquences : l'ensemble des mots est obtenu en analysant les documents du corpus, en éliminant les mots vides de sens, et en sélectionnant les mots les plus signifiants. Il existe plusieurs méthodes pour avoir les mots significatifs, cette approche a utilisé la fréquence comme mesure de pertinence d'un mot, et considère le mot de fréquence supérieure ou égale à 30 % comme un mot significatif.
2. La construction de la matrice d'affinité et le graphe d'affinité correspond : pour construire la matrice d'affinité, cette approche calcule la matrice des fréquences où les colonnes sont les termes et les lignes sont les documents textuels du corpus. Après, la matrice d'affinité est calculée à partir de cette

dernière matrice. La matrice d'affinité entre mots résultants est une matrice carré où les lignes et les colonnes représentent les mots du corpus. Ensuite le graphe d'affinité est construit à partir de cette matrice.

3. La construction du cycle à partir du graphe d'affinité et la sélection des mots-clés les plus représentatifs.

L'approche de fonction TAG a donné de très bons résultats et ne nécessite en plus ni connaissances externes ni la spécification du nombre des mots représentatifs du corpus à l'avance.

Exemple

La figure 3.2 illustre les étapes nécessaires pour détecter les communautés par l'algorithme de BOUAKKAZ (Diamant) :

(a). Le corpus utilisé dans l'article de [Bou16] contient des articles scientifiques, mais peut contenir des mots, ensemble de mots (titre) ou des paragraphes.

(b). En utilisant la fonction TAG, on extrait les fréquences des mots-clés et remplir la matrice de fréquence.

(c). On calcule l'affinité entre les mots-clés selon la formule 3.1 pour créer le graphe d'affinité à partir de matrice d'affinité, où TF_{kj} est la fréquence d'occurrence de terme t_j dans le document d_i.

$$f(x) = \begin{cases} \sum_k \quad TF_{kj} \quad si \quad i = j \\ \sum_k \quad (TF_{ki} * TF_{kj}) \quad sinon \end{cases} \tag{3.1}$$

(d). Création des cycles selon le poids de chaque colonne dans la matrice d'affinité, en commençant de manière aléatoire par choisir une colonne et prenant la plus grande valeur ; la ligne de cette valeur c'est la colonne qui sera choisie ensuite ; après la création de tous les cycles, on choisit le meilleur cycle qui a la plus grande valeur.

(e). L'agrégat de diamant qui forme les communautés, Figure 3.3.

(f). Les liens sémantiques sont des liens entre les individus dans chaque communauté.

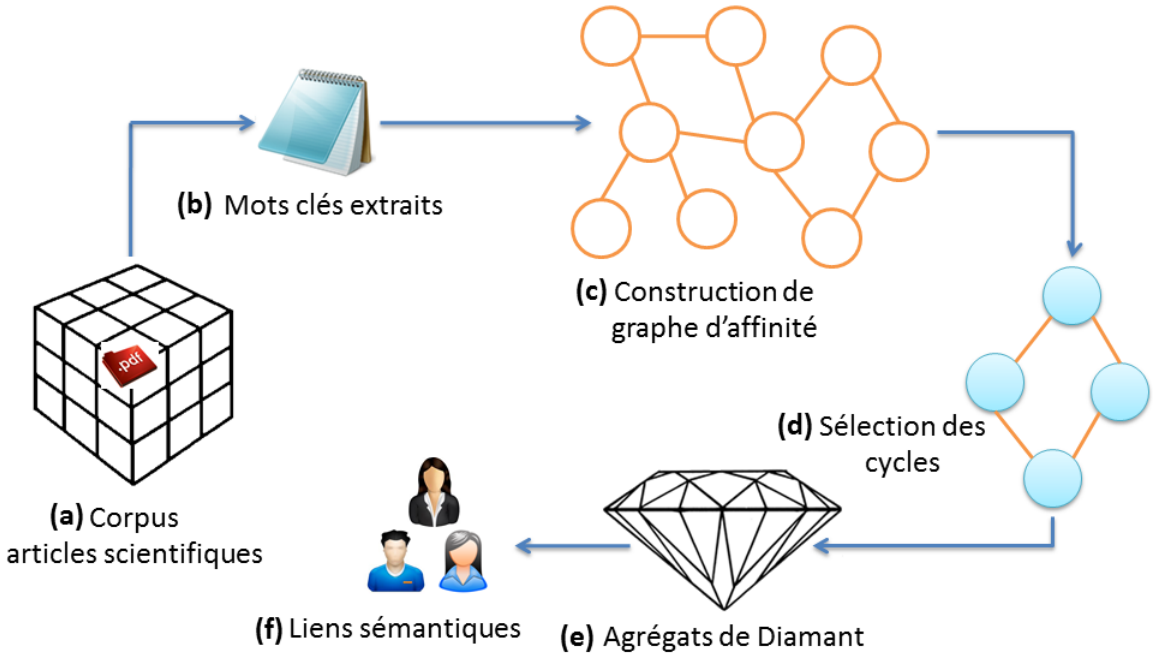

FIGURE 3.2: Algorithme de BOUAKKAZ (Diamant) [Bou16]

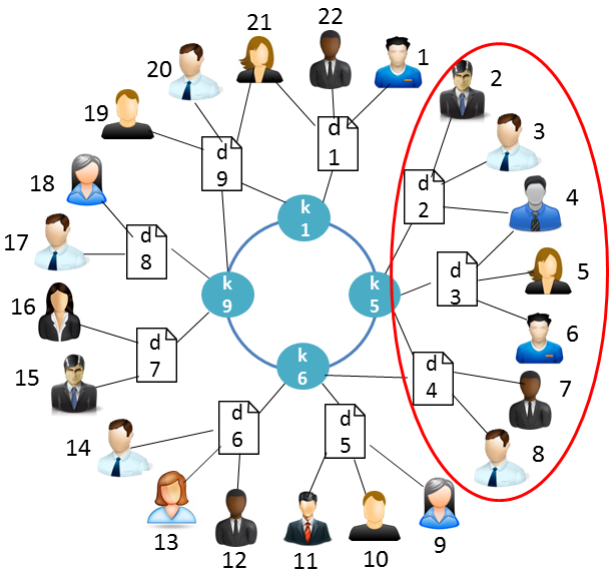

FIGURE 3.3: Agrégats de Diamant (communautés finales) [Bou16]

3.3 Mesures de performance

Dans notre contexte de travail, nous avons choisi comme mesure de performance pour évaluer les algorithmes d'étude comparative que nous avons choisis, la mesure Rappel, Précision, ainsi que la moyenne harmonique entre ces deux, caractérisée par la F-mesure [RH07], nous l'avons choisie, car elle est la plus utilisée ; et une autre mesure plus populaire et plus utilisée aussi, qui est appelée la modularité [NG04].

Pour utiliser la F-mesure., il nécessite d'un ensemble initial de communautés qui est $\{C^{'} = C^{'}_1, C^{'}_2, C^{'}_3, C^{'}_4 ... C^{'}_{k'}\}$, et d'un ensemble de communautés détectées par un des algorithmes à comparer $\{C = C_1, C_2, C_3, C_4 ... C_k\}$, où $k^{'}$ est le nombre de communautés initiales, et k est le nombre de communautés détectées par l'un des algorithmes ; et un élément a_{ij} qui représente le nombre de nœuds se trouvant à la fois dans la communauté $C^{'}_i$ et C_j (c'est à dire $C^{'}_i \cap C_j$), où i varie $de\, 1 \,à\, k^{'}$ et j varie $de\, 1 \,à\, k$, il faut calculer le rappel et la précision avant de calculer la F-mesure.

3.3.1 Le rappel

Dans notre contexte, le rappel est calculé comme la portion des nœuds de communautés initiales $C^{'}_i$ qui sont présents dans l'une des communautés C_j détectées par un des algorithmes à comparer, en mesurant ainsi la façon dont la communauté C_j détectée est complétée par rapport à la communauté initial $C^{'}_i$. Nous calculons le rappel par la formule 3.2. Le rappel de réseau entier, se calcule par la moyenne.

$$Rappel(C^{'}_i, C_j) = \frac{a_{ij}}{|C^{'}_i|} \tag{3.2}$$

3.3.2 La précision

La précision est calculée comme étant la portion de la communauté C_j détectée, qui est un membre de la communauté initiale $C^{'}_i$, en mesurant ainsi la façon homogène de la communauté C_j détectée par rapport à la communauté initiale $C^{'}_i$. Nous pouvons calculer ce rapport à l'aide de la formule 3.3. La précision de réseau entier, se calcule par la moyenne.

$$Précision(C^{'}_i, C_j) = \frac{a_{ij}}{|C_j|} \tag{3.3}$$

3.3.3 La F-mesure

La F_{mesure} fournit une vision équilibrée utile pour les algorithmes de détections des communautés, qui combine entre le rappel et la précision en effectuant une moyenne harmonique entre ces deux mesures. On commence de calculer la $F_{mesure(Locale)}$ à l'aide de la formule 3.4 qui retenue pour l'évaluation entre deux communautés $C_i^{'}$ et C_j.

$$F_{mesure(Locale)}(C_i^{'}, C_j) = \frac{2 * Rappel * Précision}{Rappel + Précision} \tag{3.4}$$

La $F_{mesure(Globale)}$ de réseau entier, sera enfin calculée en appliquant la formule 3.5, où n est la somme de tous les éléments a_{ij}.

$$F_{mesure(Globale)}(C^{'}, C) = \sum_{C_i^{'}} \frac{|C_i^{'}|}{n} \max_{C_j}(F_{mesure(Locale)}(C_i^{'}, C_j)) \tag{3.5}$$

3.3.4 La modularité

La modularité Q est une mesure de qualité qui a été introduite par Newman & Girvan [NG04], qui mesure la solidité de la structure de la communauté détectée par l'un des algorithmes. La meilleure structure de communauté est celle qui maximise la modularité, elle indique si le nombre d'arêtes inter-communautés est faible, alors que le nombre d'arêtes intra-communautés est élevé. La modularité est comprise entre -1 et 1, cette mesure est définie par la formule 3.6, où k est le nombre total des communautés, l_{C_i} est le nombre d'arêtes d'une communauté C_i, d_{C_i} est la somme de degrés de tous les nœuds de la communauté C_i et L est le nombre d'arêtes du réseau entier.

$$Q = \sum_{i=1}^{k} \left[\frac{l_{C_i}}{L} - \left(\frac{d_{C_i}}{2L} \right)^2 \right] \tag{3.6}$$

Conclusion

Dans ce chapitre, nous avons vu les deux algorithmes que nous avons choisis pour notre étude comparative, en expliquant leurs étapes avec des exemples, et les mesures de performances qu'ils utilisent pour évaluer les algorithmes.

Dans le prochain chapitre, nous allons présenter notre outil programmé, et discuterons les résultats d'expérimentation.

Chapitre 4

Implémentation et Expérimentations

L'objectif de ce chapitre est de présenter notre outil que nous avons développé pour comparer les deux algorithmes choisis, et de discuter les résultats d'expérimentation, pour les évaluer par les mesures de performance qu'on a expliquées dans le chapitre précédent.

4.1 Environnement de travail

L'outil de cette étude comparative a été implémenté en langage java, ce langage intègre les concepts les plus intéressants des technologies informatiques récentes dans une plateforme de développement riche et homogène. L'approche objet de ce langage, sa portabilité et sa gratuité, le placent parmi les outils de programmation les plus efficaces. Depuis la version 1.4.2, Java dispose d'outils modernes d'installation et de mise à jour. Il est maintenant possible de télécharger le JDK (Java Development Kit) ou le JRE (Java Runtime Environment) sur leur site officiel [1].

Nous avons installé la version JDK1.8 dans un ordinateur ayant un processeur Intel ® CORE i5, une RAM de 8Go et disque dur de 500 GO avec le système d'exploitation windows 7 ultimate service pack 1 de type 64 bits.

L'outil a été écrit dans l'éditeur de code NetBeans IDE 8.0, C'est un éditeur parmi les IDE (Integrated Development Environment) Java. Il simplifie grandement l'édition et la gestion d'un programme. Ils intègrent les fonctionnalités suivantes :

1. www.oracle.com/technetwork/java/javase/downloads/index.html

- Éditeur de textes avec mise en couleur des mots-clés Java, des commentaires.
- Complétion automatique (menus contextuels proposant la liste des méthodes d'un objet).
- Génération automatique des dossiers nécessaires à l'organisation d'un programme et des paquetages des classes.
- Intégration des commandes Java et de leurs options dans des menus et des boîtes de dialogue appropriés.
- Débogueur pour corriger les erreurs.

Nous avons visualisé notre réseau social avec SocNetV [2] (Social Network Visualizer), c'est un projet open-source pour construire, un outil multi-plateforme flexible et facile à utiliser pour l'analyse et la visualisation des réseaux sociaux, en ciblant principalement les chercheurs sociaux. L'application offre une interface graphique facile.

SocNetV nous permet de construire des réseaux sociaux en quelques clics ou d'importer le réseau à partir d'un fichier de différents formats (graphml, GraphViz, matrice d'adjacence, Pajek, csv, ...etc) et de les modifier en fonction de nos besoins.

4.2 Algorithmes et explications

Afin de comparer les résultats des deux algorithmes d'étude comparative de détection de communautés, il nécessite d'appliquer les deux algorithmes sur les mêmes données et dans le même contexte en utilisant les mêmes mesures de performances ; pour cela, nous avons choisi d'utiliser le même réseau social pour notre travail.

Les différentes étapes d'implémentation de ces algorithmes sont montrés dans la figure 4.1

2. www.socnetv.sourceforge.net

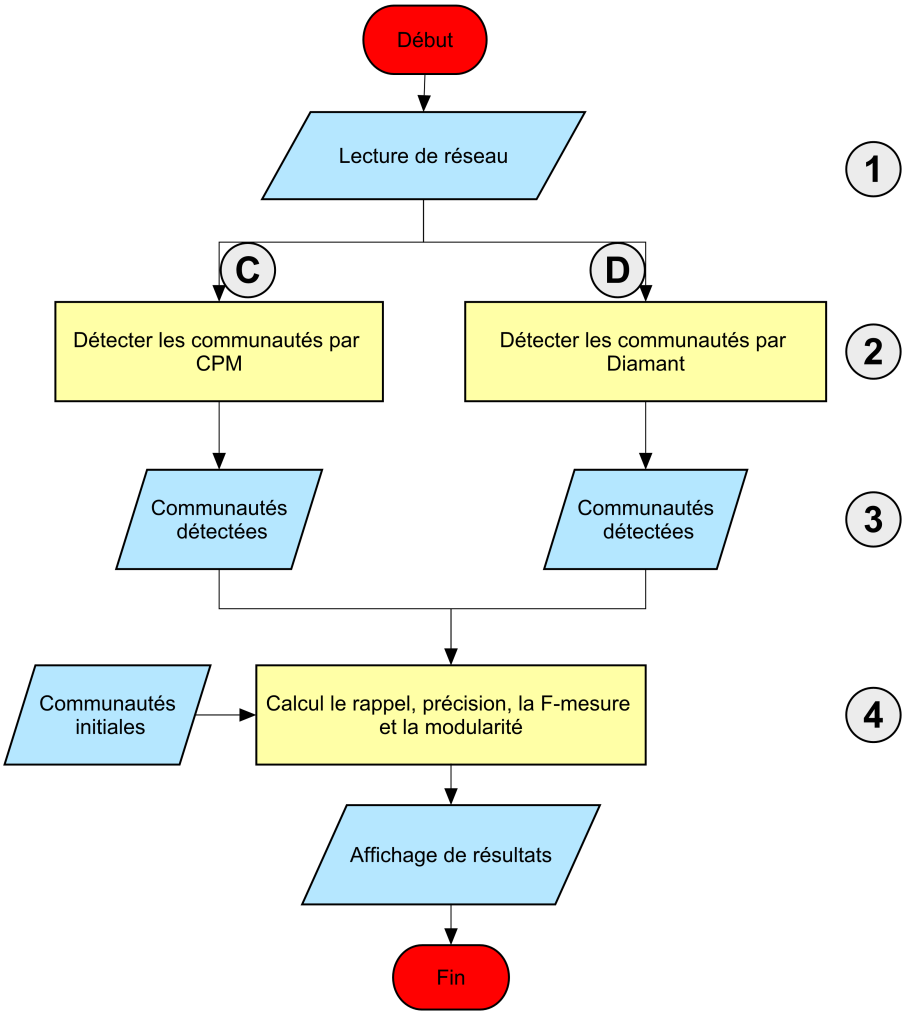

FIGURE 4.1: L'organigramme global de l'implémentation

L'étape (1) : c'est une opération d'entrée/sortie d'une étape très importante, dans laquelle le programme lit le fichier texte qui a stocké le réseau social sous forme d'une matrice symétrique, ce qu'il facilite son interprétation.

L'étape (2) : cette étape d'implémentation est distinguée en deux étapes indépendantes, une étape (C) consiste à appliquer l'algorithme de CPM de détection de communautés et l'autre (D) consiste à appliquer l'algorithme de BOUAKKAZ (Diamant) ; comme elles sont montrées dans les figures 4.2 et 4.3 respectivement.

L'étape (3) : c'est l'opération de sortie qui présente les résultats d'étape (2) qui

sont les communautés détectées et deviendrons comme entrée pour l'étape (4).

L'étape (4) : cette étape consiste à calculer les mesures de performances : le rappel, la précision, la F-mesure et la modularité pour les deux algorithmes, ces mesures sont décrites en détail dans les figures 4.4 et 4.5 respectivement. Elle nécessite les communautés initiales comme entrée aussi.

Dans la suite nous allons voir les algorithmes et les procédures qui sont incluent dans chaque étape.

4.2.0.1 Détection de communautés par l'algorithme CPM

La figure 4.2 représente les étapes d'exécution de l'algorithme CPM, il commence la première étape par lire la taille de clique k pour extraire tous les cliques en deuxième étape, puis il construit la matrice d'adjacente de cliques à partir de matrice Tous_cliques[][] à condition que chaque deux nœuds (cliques) partagent k-1 noeuds, ainsi créer un lien entre tous pairs de noeuds qui confirment cette condition, puis il prend la matrice Cliques_liens[][] comme entrée pour la fonction Parcours_en_Largeur() qui trouve les composants connectés et ainsi les communautés détectées.

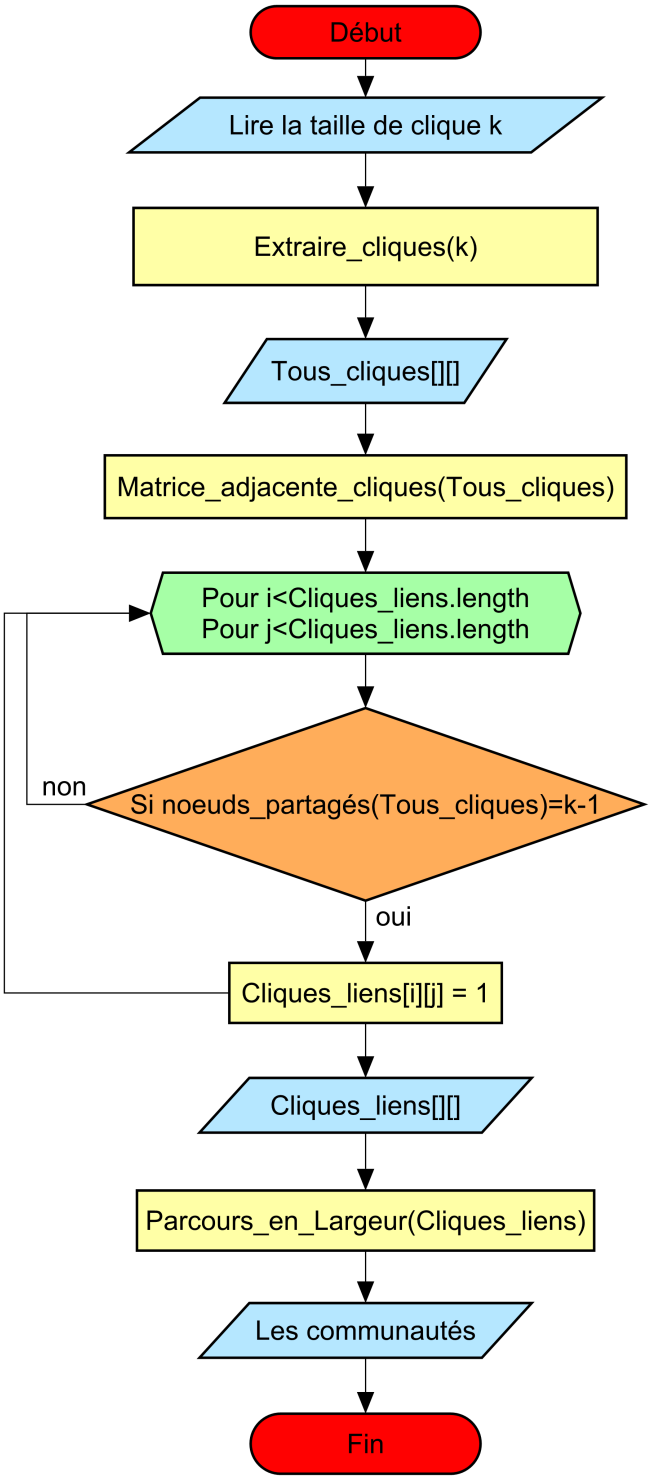

FIGURE 4.2: L'organigramme de l'algorithme CPM

4.2.0.2 Détection de communautés par l'algorithme de BOUAKKAZ (Diamant)

La figure 4.3 montre les étapes de l'algorithme de BOUAKKAZ (Diamant) en détail, il utilise les termes les plus représentatifs extraits à partir d'un corpus comme entrée de la fonction TAG() pour extraire les Mots-clés avec leurs fréquences afin de remplir la matrice de fréquence, ensuite il utilise cette matrice pour calculer l'affinité à l'aide de la formule 3.1 qu'on a déjà présenté dans le chapitre précédent pour créer le graphe d'affinité en utilisant la matrice résultante qui est Matrice_affinité[][], en utilisant cette dernière matrice, il construit les cycles, puis il choisit le meilleur, enfin il agrégat le diamant qui forme les communautés détectées.

4.2.0.3 Calcul de la F-mesure

On présente dans la figure 4.4 l'organigramme de la fonction de F-mesure qui monte la manière de calculer cette mesure. elle recherche pour chaque deux communautés pour trouver le nombre de noeuds a_{ij} qui appartient à ces communautés, et retient le nombre total n de tous ces éléments, puis elle calcule le rappel et la précision en utilisant la valeur a_{ij} pour les deux mesures et C_i', C_j respectivement pour les utiliser comme entrée pour calculer la $F_{mesure(Locale)}$ entre deux communautés seulement, puis elle suffit de choisir la $F_{mesure(Locale)}$ maximale obtenue avec cette dernière pour le calcul enfin la $F_{mesure(Globale)}$ pour le réseau entier avec la valeur n en entrée.

4.2.0.4 Calcul de la modularité

Cette fonction calcule la modularité comme présente dans la figure 4.5, le calcul de cette mesure est simple, c'est juste une application numérique de la formule 3.6.

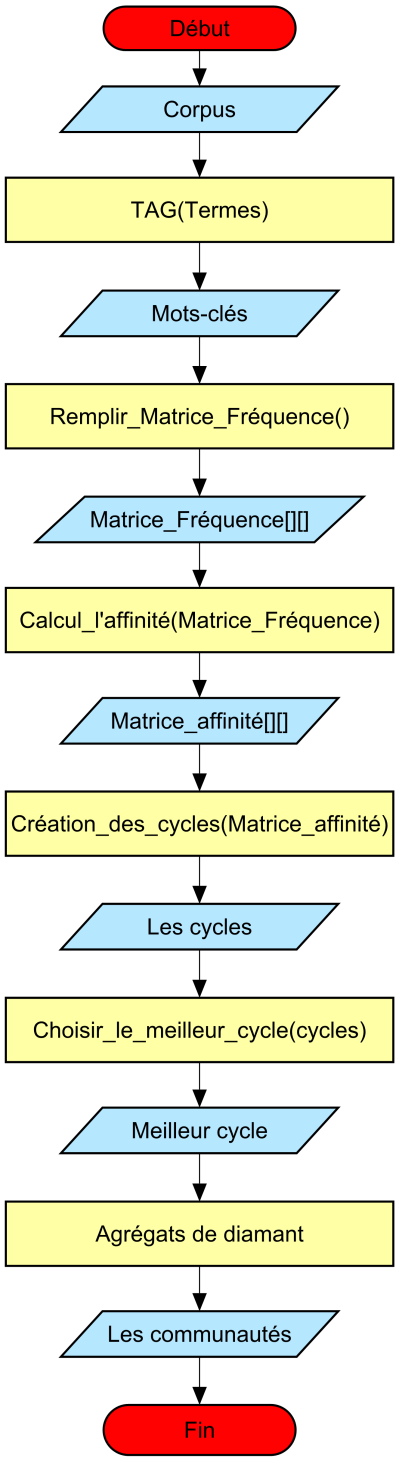

FIGURE 4.3: L'organigramme de l'algorithme de BOUAKKAZ (Diamant)

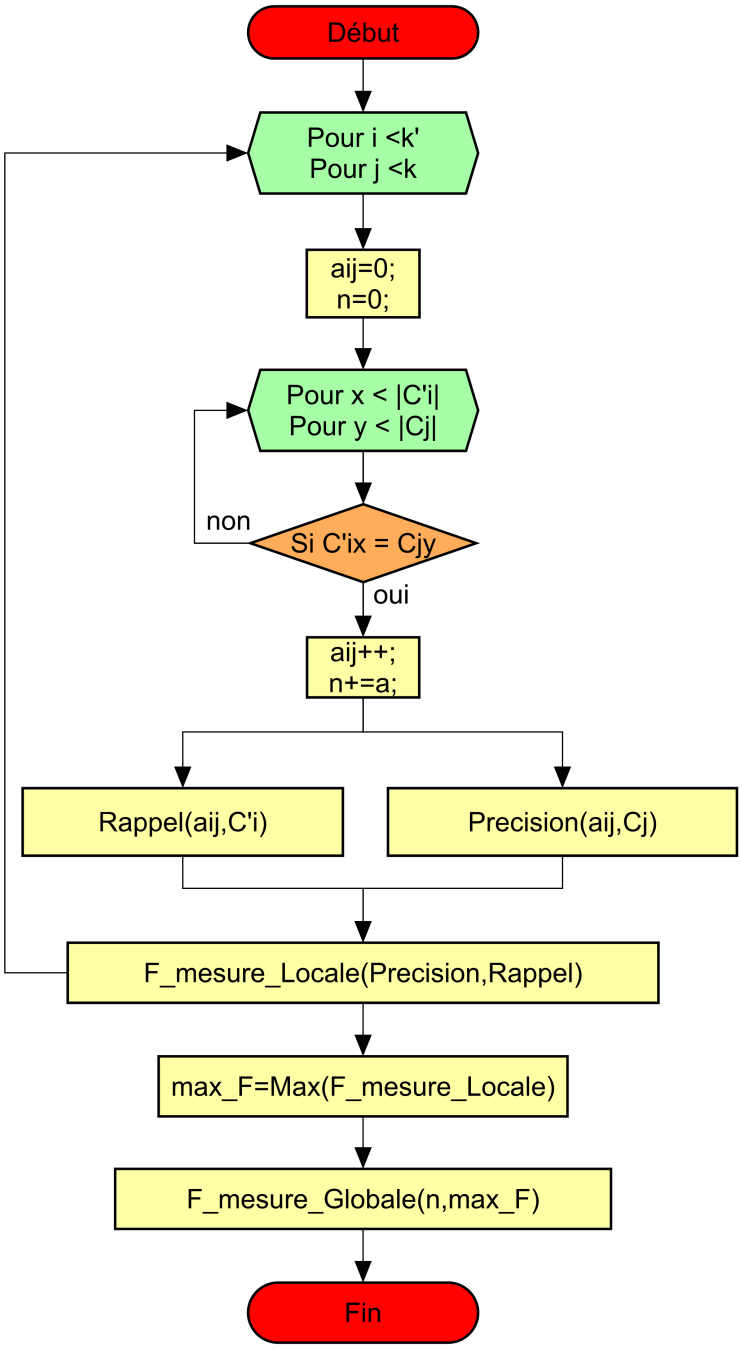

FIGURE 4.4: L'organigramme de la F-mesure

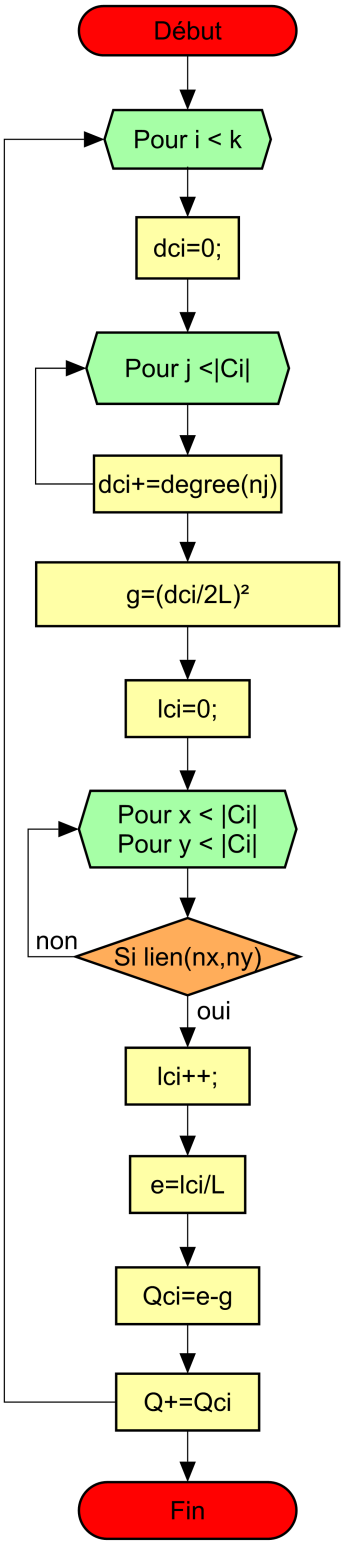

FIGURE 4.5: L'organigramme de la modularité

4.3 Expérimentations et résultats

Pour assurer une comparaison efficace entre les deux algorithmes d'étude comparative de détection de communautés, nous les avons testés sur un réseau réel du laboratoire ERIC (Entrepôts, Représentation et Ingénierie des Connaissances) de l'université de lyon, c'est pour ça, nous avons choisi de prendre une visualisation sur le réseau pour présenter les communautés initiales qui sont constituées à partir de membres d'équipe du laboratoire ERIC, avant d'appliquer les algorithmes de détection de communautés et une autre visualisation après l'application pour présenter les communautés finales détectées.

4.3.1 Présentation du laboratoire ERIC

Le laboratoire ERIC [3] a été créé en 1995, ERIC est une unité de recherche dont les établissements de tutelle sont l'Université Lumière Lyon 2 et l'Université Claude Bernard Lyon 1. ERIC fait également partie de l'Institut des Sciences de l'Homme et est rattaché à l'École Doctorale Infomaths.

Les recherches du laboratoire ERIC se situent dans les domaines de la **science des données** et de l'**informatique décisionnelle**. Elles visent à valoriser les grandes bases de données complexes, notamment dans les domaines des sciences humaines et sociales (SHS), et se situent dans les domaines suivants :

- Les **entrepôts de données** : intégration intelligente de données complexes, modélisation multidimensionnelle d'objets complexes, analyse en ligne personnalisée, sécurité du processus d'entreposage ;
- La **fouille de données** et la **décision** : apprentissage automatique, étude et fouille de graphes, analyse de données complexes, agrégation multicritère, fouille d'opinion, logiciels de fouille de données.

Il développe de nombreux partenariats académiques et industriels. Ses membres animent des formations en Licence, Master et Doctorat dans les universités Lyon 1 et Lyon 2. Ils coordonnent notamment une formation européenne d'excellence en Fouille de Données et Gestion des Connaissances, le Master Erasmus Mundus.

Le laboratoire est structuré en deux équipes de recherche :

Systèmes d'Information Décisionnels (SID) ;

3. eric.ish-lyon.cnrs.fr

Data Mining et Décision (DMD).

Les figures 4.6 et 4.7 représentent le réseau et les communautés initiales du laboratoire ERIC respectivement.

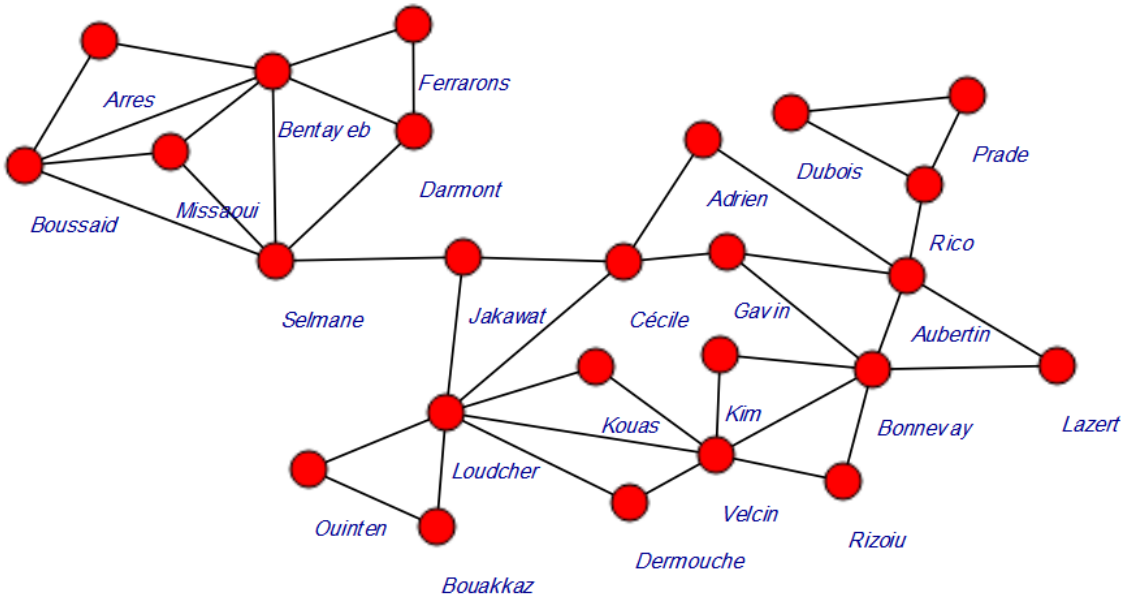

FIGURE 4.6: Le réseau du laboratoire ERIC

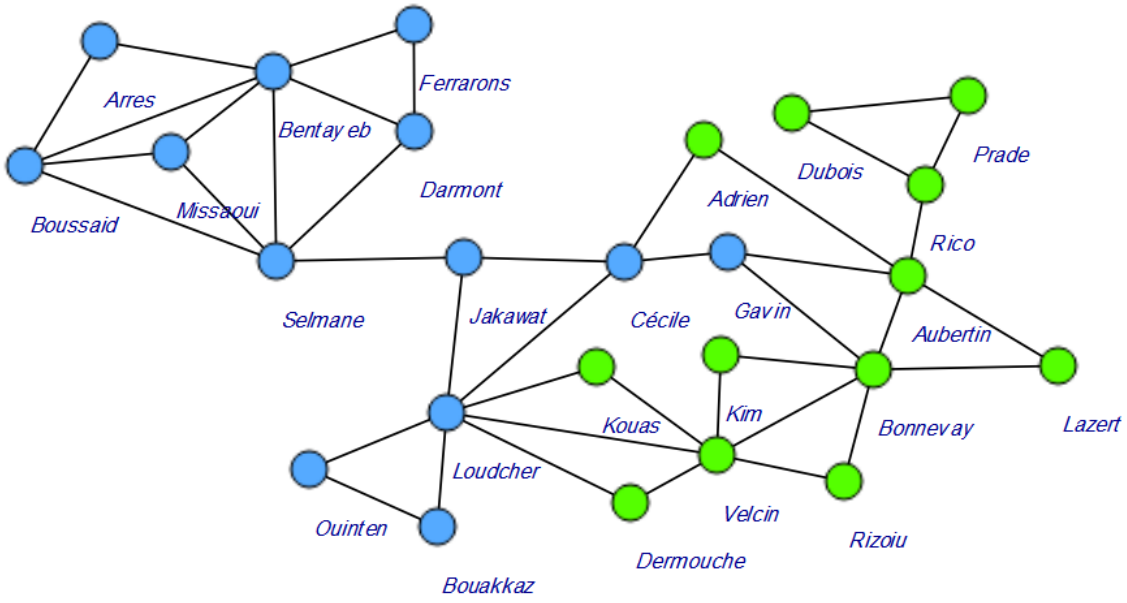

FIGURE 4.7: Les communautés initiales du laboratoire ERIC

Les figures 4.8 et 4.9 représentent les communautés détectées après l'exécution des algorithmes CPM et Diamant respectivement ; nc : nombre de communautés.

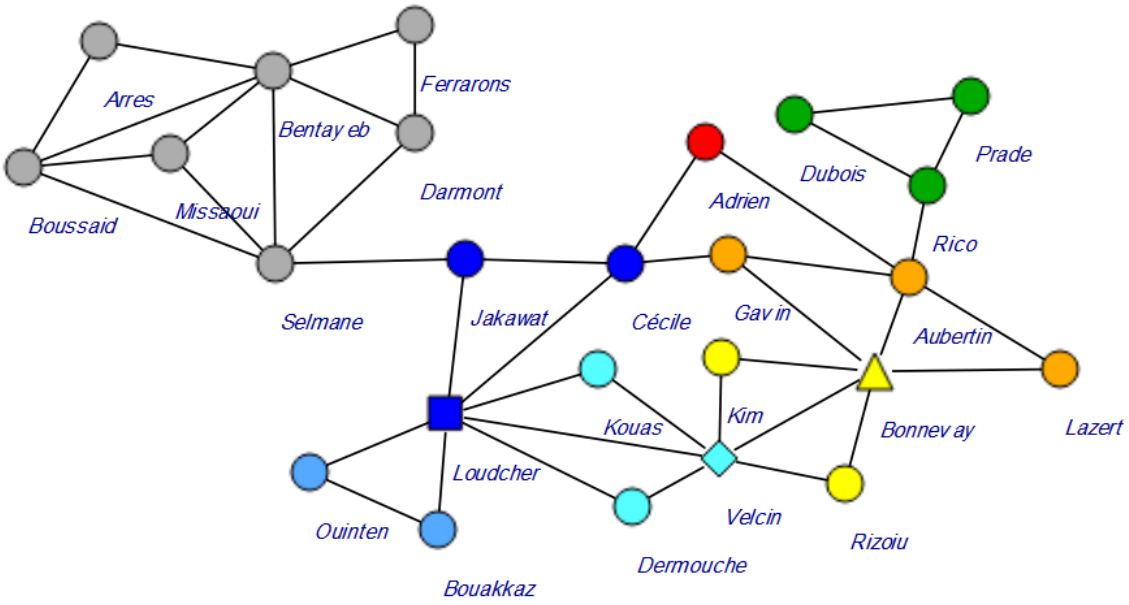

FIGURE 4.8: Résultat d'exécution de l'algorithme CPM, nc=7

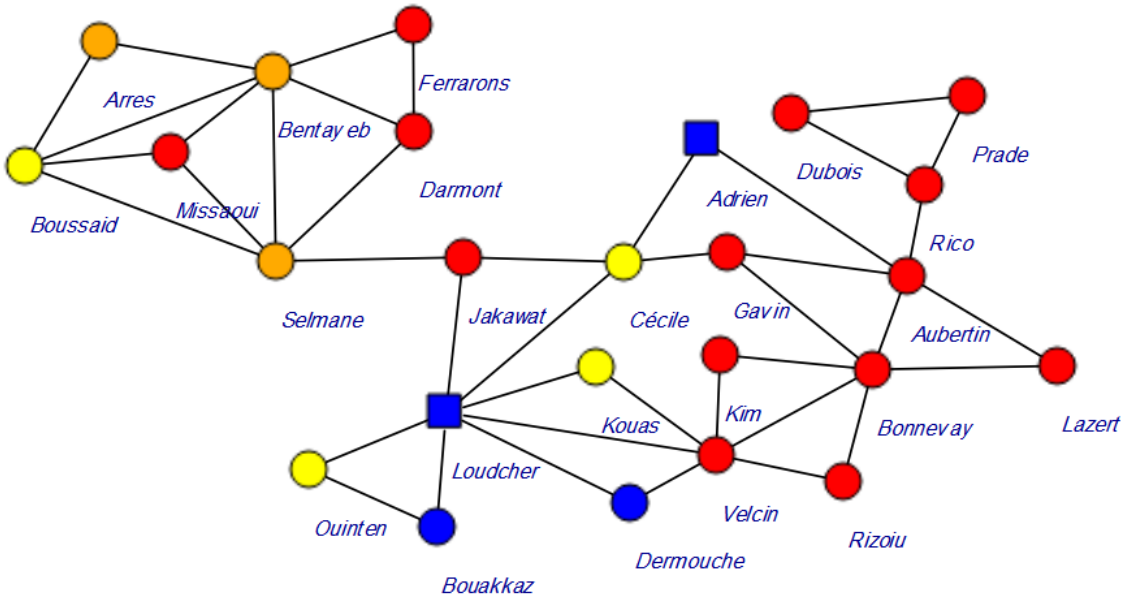

FIGURE 4.9: Résultat d'exécution de l'algorithme Diamant, nc=3

Le tableau 4.1 montre les différents résultats des mesures de performance après l'exécution de l'algorithme CPM avec k = 3 et l'algorithme Diamant.

		Algorithme	
		CPM	Diamant
Mesures	Rappel	24,86%	20,51%
	Précision	77,78%	60,00%
	F-mesure	62,50%	80,97%
	Modularité	61,15%	7,93%

TABLE 4.1: Les mesures de performance

La figure 4.10 est une représentation graphique du rappel, précisons, F-mesure et modularité des différents résultats du tableau précédent.

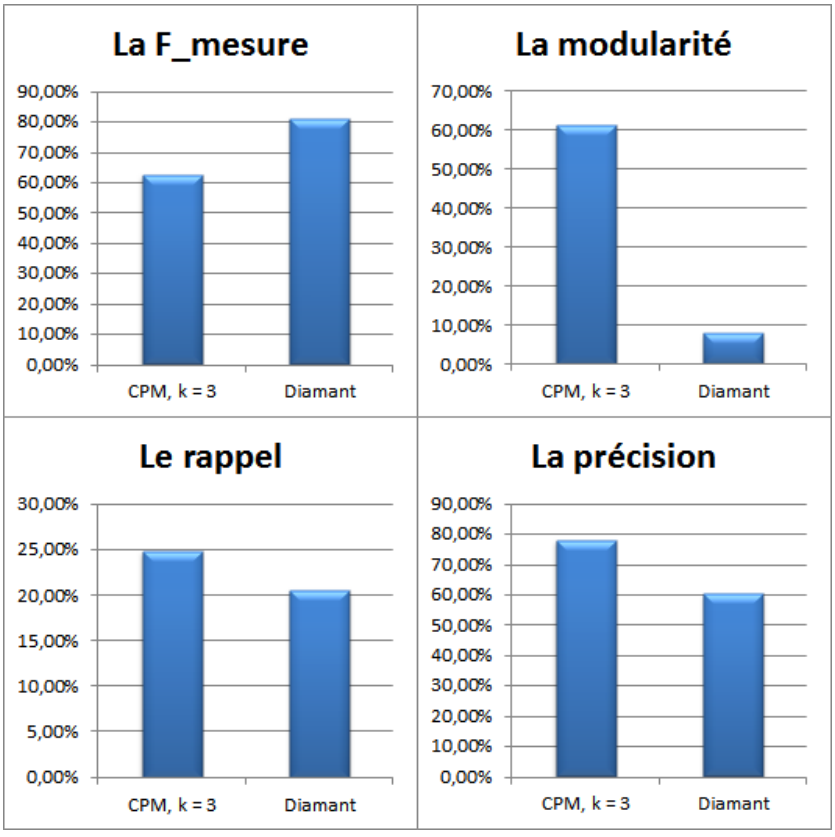

FIGURE 4.10: Représentation graphique du rappel, précisons, F-mesure et modularité

4.3.2 Comparaison et interprétation des résultats

Afin d'évaluer la performance de ces derniers algorithmes selon les mesures de performance (rappel, précisons, F-mesure et modularité) choisies dans ce cadre de travail. Nous comparons et nous interprétons les résultats obtenus et résumés dans le tableau 4.1 et représentés dans le graphe 4.10 selon les mesures qu'on a mentionnées précédemment.

4.3.2.1 Point de vue Rappel

Nous mentionnons qu'un rappel de valeur élevée signifie que les communautés sont plus pertinentes entre elles. On observe que le rappel est quasiment la même pour les deux algorithmes CPM et Diamant à 24,86% et 20,51% respectivement.

4.3.2.2 Point de vue Précision

Nous rappelons également qu'une précision de grande valeur est interprétée par une bonne qualité de détection. On remarque à travers le graphe de la précision que la performance de l'algorithme CPM est élevée à 77.78% par rapport à l'algorithme Diamant à 60%, cela exprime une détection correcte.

4.3.2.3 Point de vue F-mesure

Nous mentionnons qu'une valeur de F-mesure est proche de 1 réfère à une structure de communauté initiale proche et donc l'algorithme qui l'a détectée, a donné de bons résultats. En conséquence, l'algorithme Diamant a donné de très grande valeur de F-mesure à 80,97% par rapport à l'algorithme CPM à 62,50%.

4.3.2.4 Point de vue Modularité

Nous rappelons que la meilleure structure de communauté est celle qui maximise la modularité, pour cela, on remarque que l'algorithme CPM a donné des valeurs élevées de modularité à 61,15% par rapport à l'algorithme Diamant qui donne un résultat faible à 7,93%.

Enfin, nous pouvons déduire que la performance de l'algorithme Diamant qui est basé sur le contenue, en plus de la structure, est supérieur à celle de l'algorithme

CPM qui est basé seulement sur la structure, d'après la F-mesure. Par contre, ce dernier algorithme est supérieur à l'algorithme Diamant selon la modularité. La modularité faible produite par Diamant peut être expliquée par un petit nombre de liens générés. Avec un réseau plus dense la modularité pourrait augmenter.

Conclusion

Dans ce chapitre, nous avons exposé les résultats des expérimentations que nous avons réalisées. Les tests ont porté sur les deux algorithmes choisis en utilisant un réseau réel du laboratoire ERIC, dans le but de comparer et d'interpréter les résultats obtenus par les différentes mesures proposées dans cette étude.

Conclusion et perspectives

Pour atteindre les objectifs tracés pour cette étude comparative, nous avons décrit les différentes méthodes existantes de ce domaine dans l'état de l'art, afin de choisir deux algorithmes pour les utiliser dans ce travail ; ces deux algorithmes sont l'algorithme CPM la méthode de clique percolation (Clique Percolation Method), et l'algorithme de BOUAKKAZ (Diamant) ; les deux algorithmes ont été implémenté en langage JAVA.

Afin d'évaluer la performance de chaque algorithme choisi, nous avons appliqué notre outil programmé sur un réseau réel du laboratoire ERIC (Entrepôts, Représentation et Ingénierie des Connaissances) de l'université de lyon. Ainsi, nous avons visualisé ce réseau avec l'application SocNetV. Après la comparaison et l'interprétation des résultats obtenus par les différentes mesures proposées, nous conclurons que l'algorithme Diamant donne de meilleurs résultats avec la F-mesure en outre, il ne nécessite pas des paramètres à l'avance. Donc, nous pouvons le considérer comme un algorithme non-supervisé. D'autre part l'algorithme CPM donne de meilleurs résultats avec la modularité mais il faut paramétrer la valeur k qui est la taille de clique, pour démarrer l'exécution. Donc, il peut être considéré comme un algorithme supervisé.

Cette étude comparative, nous a permis également de découvrir le domaine d'analyse des réseaux sociaux, et en particulier la détection de communautés. En plus, nous a donné une expérience très importante avec la programmation orienté objet en langage JAVA.

Perspectives

Sur un plan plus général, et comme perspective pour ce mémoire, nous désirons trouver une nouvelle manière pour optimiser l'algorithme CPM dans le but d'éviter les paramétrages nécessaires dans l'exécution et par introduire la modularité dans la détection. Nous espérons que la proposition de cette optimisation permettra de l'effectuer comme un algorithme non-supervisé, afin de permettre une meilleure détection de communautés dans les réseaux sociaux.

Bibliographie

[BA99] Albert-Laszlo Barabasi and Reka Albert. Emergence of Scaling in Random Networks. *SCIENCE*, 286(October) :509–512, 1999.

[Bac14] Rémi Bachelet. *Réseaux sociaux (cours)*. 2014.

[BBA75] Ronald L Breiger, Scott A Boorman, and Phipps Arabie. An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. *Journal of Mathematical Psychology*, 12(3) :328–383, aug 1975.

[BD07] John G. Breslin and Stefan Decker. The future of social networks on the internet : The need for semantics. *IEEE Internet Computing*, 11(6) :86–90, 2007.

[BGLL08] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics : Theory and Experiment*, 10008(10) :6, 2008.

[BLO14] Mustapha Bouakkaz, Sabine Loudcher, and Youcef OUINTEN. Automatic Textual Aggregation Approach of Scientific Articles in OLAP context. In *Innovations in Information Technology (INNOVATIONS), 2014 10th International Conference on*, pages 30–35. IEEE, 2014.

[Bou16] Mustapha Bouakkaz. Algorithme de BOUAKKAZ (Diamant). *Communication personnelle*, 2016.

[Cav16] Frédéric Cavazza. Panorama des médias sociaux 2016 - FredCavazza, 2016. url : http ://www.fredcavazza.net/2016/04/21/panorama-des-medias-sociaux-2016/, consulté le 03/05/2016.

[CCS+10] Attias Cyril, Brayer Céline, Bruno Salvatore, Jacquot Caroline, Strul Richard, Thobellem Alexis, and Villalba Archana. Les médias sociaux. *IAB France*, 2010.

[Cla05] A Clauset. Finding local community structure in networks. *Physical Review E*, 72(2) :26132, 2005.

[DM04] Luca Donetti and Miguel A. Muñoz. Detecting network communities : a new systematic and efficient algorithm. *Journal of Statistical Mechanics : Theory and Experiment*, 10(2004) :8, 2004.

[ER59] P Erdös and a Rényi. On random graphs. *Publicationes Mathematicae*, 6 :290–297, 1959.

[FLG00] Gary William Flake, Steve Lawrence, and C Lee Giles. Efficient Identification of Web Communities. pages 150–160, 2000.

[For10] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5) :75–174, 2010.

[GL06] Jean-Loup Guillaume and Matthieu Latapy. Bipartite graphs as models of complex networks. *Physica A : Statistical Mechanics and its Applications*, 371(2) :795–813, nov 2006.

[GN02] M. Girvan and Mark E J Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12) :7821–7826, 2002.

[Has06] M. B. Hastings. Community detection as an inference problem. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 74(3) :6–9, 2006.

[IKN05] Hidehiko Ino, Mineichi Kudo, and Atsuyoshi Nakamura. Partitioning of Web graphs by community topology. *Proceedings of the 14th international conference on World Wide Web*, pages 661–669, 2005.

[Kan04] Ravi Kannan. On Clusterings : Good , Bad and Spectral. 51(3) :497–515, 2004.

[KKL12] Rania Kacimi, Ira Kokova, and Bénédicte Lelong. *Tumblr 2012 : guide d'utilisation pour un usage professionnel*. 2012.

[LFK09] Andrea Lancichinetti, Santo Fortunato, and Janos Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11 :1–18, 2009.

[Lux06] Ulrike Von Luxburg. A Tutorial on Spectral Clustering A Tutorial on Spectral Clustering. *Statistics and Computing*, 17(March) :395–416, 2006.

[LWP06] Feng Luo, James Wang, and Eric Promislow. Exploring Local Community Structures in Large Networks. In *2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)*, pages 233–239. IEEE, dec 2006.

[NED15] MOHAMED ABDELHAMID NEDIOUI. *FOUILLE ET APPRENTISSAGE AUTOMATIQUE DANS LES RESEAUX DYNAMIQUES*. Magister thesis, 2015.

[NG04] M. E J Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 69(2 2) :1–15, 2004.

[Pai16] David Pain. L'addiction aux réseaux sociaux est une réalité, la preuve en vidéo, 2016. url : http ://www.begeek.fr/laddiction-aux-reseaux-sociaux-realite-preuve-video-200733, consulté le 25/04/2016.

[Pal15] Bernard Pallardy. *Découvrir et utiliser Viadeo*. 2015.

[PDFV05] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043) :814–8, jun 2005.

[PKVS12] S Papadopoulos, Y Kompatsiaris, A Vakali, and P Spyridonos. Community detection in social media performance and application considerations. *Data Mining and Knowledge Discovery*, 24(3) :515–554, 2012.

[PL05] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, 10(2) :191–218, 2005.

[RAK07] Usha Nandini Raghavan, Reka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 76(3) :1–12, 2007.

[RH07] Andrew Rosenberg and Julia Hirschberg. V-measure : A conditional entropy-based external cluster evaluation measure. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1(June) :410–420, 2007.

[Sco88] John Scott. TREND REPORT SOCIAL NETWORK ANALYSIS. *Sociology*, 22(1) :109–127, 1988.

[SM00] J Shi and J Malik. Normalized Cuts and Image Segmentation. *Ieee Transactions on Pattern Analysis and Machine Intelligence*, 22(8) :888–905, 2000.

[TL10a] Lei Tang and Huan Liu. Community Detection and Mining in Social Media. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 2(1) :1–137, jan 2010.

[TL10b] Lei Tang and Huan Liu. GRAPH MINING APPLICATIONS TO SOCIAL NETWORK ANALYSIS. In *Managing and Mining Graph Data*, chapter 16, pages 487–513. C. Aggarwal,Haixun Wang, 2010.

[Via15] Rudy Viard. La Liste Complète des Réseaux Sociaux, 2015. url : http ://www.webmarketing-conseil.fr/liste-reseaux-sociaux/, consulté le 07/01/2016.

[WF94] Stanley Wasserman and Katherine Faust. *Social network analysis : Methods and applications*, volume 8. Cambridge university press, 1994.

[WS98] DJ J Watts and SH H Strogatz. Collective dynamics of'small-world'networks. *Nature*, 393(6684) :440–442, 1998.

[XYFS07] Xiaowei Xu, Nurcan Yuruk, Zhidan Feng, and Thomas a J Schweiger. SCAN : A Structural Clustering Algorithm for Networks. *13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 824–833, 2007.

[Zam12] Nisrine Zammar. *Réseaux Sociaux numériques : essai de catégorisation et cartographie des controverses.* PhD thesis, 2012.

[ZDP14] Wei Zhou, Wenjing Duan, and Selwyn Piramuthu. AC A Social Network Matrix for Implicit and Explicit. *Decision Support Systems*, 2014.