République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Centre Universitaire Belhadj Bouchaib -Ain Temouchent-

Institut des Sciences
Département des mathématiques et de l'informatique

Mémoire

Présenté pour l'obtention du Diplôme de Master en

Option : Réseaux et Ingénierie des données (RID)

Présenté par: M. GUERMOUCHE Amine

THÈME :

Détection et analyse de communautés dans les réseaux sociaux

Encadrant : Mme Khadidja MEGAGUE
Maitre Assistant "A" à C.U.B.B.A.T

Soutenu en Septembre 2018

Devant le Jury composé de :

Présidente : Mme BEDAD Fatima (M.A.A) C.U.B.B.A.T
Examinateur : M.MESSAOUDI Mohamed Amine (M.A.A) C.U.B.B.A.T
Encadrant : Mme Khadidja MEGAGUE (M.A.A) C.U.B.B.A.T

Résumé

Pour modéliser certains systèmes complexes, il est adéquat d'utiliser des structures mathématiques appelés graphes ou réseaux

La problématique posée par les graphes est de détecter les communautés. Le but étant d'en comprendre la structure en détectant une partition. De nombreux algorithmes ont été utilisés afin de résoudre ce problème.

Une des méthodes que nous avons utilisée est la percolation de cliques. Elle est basée sur la recherche d'un groupe de nœuds plus étroitement connectés les uns aux autres que les autres nœuds du réseau. Mais cet algorithme passe par plusieurs étapes ce qui lui rend très lent en exécution d'un nombre important de nœuds.

La deuxième méthode est l'algorithme de Louvain qui est actuellement un des meilleurs algorithmes en terme de complexité pour calculer des communautés sur de très grands graphes, Cette méthode a pour particularité d'implanter une méthode d'optimisation "gloutonne" locale de la modularité.

Enfin nous avons créé une heuristique. Elle simule de grands graphes réels, que nous avons obtenu de nombre aléatoires de nœuds et d'arêtes. Elle permet la détection et visualisation graphiques des communautés. Nous l'avons également testé sur des graphes réels, obtenus des réseaux sociaux.

Une étude comparative a été effectuée pour observer la qualité de partitionnement ainsi que le temps de détection des communautés par les algorithmes proposés.

Abstract

To model some complex systems, it is appropriate to use mathematical structures called graphs or networks.

The problem posed by graphs is to detect communities. The goal is to understand the structure by detecting a partition. Many algorithms have been used to solve this problem.

One of the methods we used is the percolation of cliques. It is based on finding a group of nodes more closely connected to each other than other nodes in the network. But this algorithm goes through several steps which makes it very slow in execution of a large number of nodes.

The second method is the Louvain algorithm, which is currently one of the best algorithms in terms of complexity for calculating communities on very large graphs. This method has the particularity of implementing a local "gluttonous" optimization method of modularity.

Finally we have created a heuristic. It simulates large real graphs, which we have obtained from random numbers of nodes and edges. It allows the detection and graphic visualization of communities. We also tested it on real graphs, obtained from social networks.

A comparative study was carried out to observe the quality of partitioning as well as the detection time of the communities by the proposed algorithms.

Table des matières

Table des figures

Liste des tableaux

Remerciements

Je tiens en premier lieu, à remercier, tout particulièrement mon encadreur Madame MEGAGUE Khadidja, pour la gentillesse et la patience qu'elle a manifestée à mon égard. Ainsi que de m'avoir guidé, dirigé et conseillé pour la réalisation de ce mémoire.

Je tiens également à remercier Mme BEDAD, pour m'avoir fait l'honneur de présider le Jury de soutenance de ce mémoire ainsi que M.MESSAOUDI pour sa participation en tant qu'examinateur ainsi que pour son aide apporté pendant mon cursus.

Je remercie également tous mes professeurs qui durant ma formation m'ont permit par leur savoir et leur pédagogie d'acquérir toutes les notions essentielles à mon perfectionnement.

Mes remerciements vont également à mon ami GADIRI Mohamed pour toute l'aide apportée, ma mère Tsouria BOUAYAD ALAM épouse TANDJAOUI, ma femme Sihem Née RAHALI, mes enfants Ibtissem, Lahcène, Ilyess et Karim pour leurs patience et encouragements.

Enfin, je remercie tous ceux que j'ai oublié de citer : tous ces amis proches et lointains.

Merci à toi lecteur de ce mémoire, de t'intéresser à mon travail de recherche. Tes critiques et réflexions lui donne un sens et une raison d'exister.

Chapitre 1

Introduction Générale

1.1 Contexte d'études

Internet est un réseau informatique mondial, il utilise l'ensemble des protocoles de la suite TCP/IP (Transmission Control Protocol / Internet Protocol), pour le transfert des Données (DoD Model)

Internet propose trois types de services fondamentaux : le courrier électronique (e-mail) ; le Web (les pages avec liens et contenus multimédia de ses sites Web) ; l'échange de fichiers par FTP (File Transfer Protocol). Actuellement, Le réseau Internet sert de plus en plus, aux communications téléphoniques et à la transmission de vidéos et d'audio en direct.

Le réseau du Web n'est pas le seul réseau d'importance, nous avons aussi pour autre exemple la Sociologie, la Chimie, etc.

Définition de la Théorie des graphes :

En mathématique et informatique, cette théorie étudie les graphes, lesquels sont des modèles abstraits de dessins de réseaux reliant des objets.

Ces modèles sont constitués par la donnée de « points », appelés nœuds ou sommets (en référence aux polyèdres), et de « liens » entre ces points ; ces liens sont souvent symétriques (les graphes sont alors dits non orientés) et sont appelés des arêtes.

Les algorithmes élaborés pour résoudre des problèmes concernant les objets de

cette théorie ont de nombreuses applications dans tous les domaines liés à la notion de réseau (réseau social, réseau informatique, télécommunications, etc.) et dans bien d'autres domaines (par exemple celui de la génétique) tant le concept de graphe, à peu près équivalent à celui de relation binaire (à ne pas confondre donc avec graphe d'une fonction), est général.

Par ailleurs, un réseau social est un ensemble d'acteurs et de relations que ces acteurs entretiennent

1.2 Problématique

Notre but en utilisant les graphes est de détecter les communautés sous jacentes. Il est important de comprendre la structure de ces graphes en détectant une partition ou un ensemble de nœuds ayant plus de liens entre eux qu'avec d'autres nœuds.

A cet effet, une bonne connaissance des propriétés des réseaux d'interactions est nécessaire pour prévoir leur évolution.

L'une de ces propriétés, communes des réseaux d'interaction, est la présence de zones très denses en liens, généralement appelées communautés.

Depuis plusieurs années, les réseaux sociaux sont une partie intégrante de nos vies et sont devenus des outils incontournables sur Internet quelle que soit la cible que l'on recherche.

Des dizaines de réseaux sociaux existent et se différencient plus au moins tous, certains afin de garder de la proximité entre amis, d'autres un peu plus professionnels, ou encore afin d'être à la pointe de l'actualité.

1.3 Objectif et Contribution

l'objectif de ce mémoire et de faire une étude sur la détection des communautés dans les différents type de réseaux sociaux en analysant et exploitant les propriétés de ces derniers afin de mieux comprendre leurs fonctionnements et comportements.

1.4 Organisation du manuscrit

Ce mémoire est structuré comme suit :

Chapitre 1 : introduction générale

Nous avons définit le contexte, les bases, outils et objectif de ce travail

Chapitre 2 : Réseaux

Nous avons utilisé pour ce travail les grands réseaux en général et les réseaux d'interactions en particulier

Chapitre 3 : Communautés

Notre but a été de comprendre comment s'organisent les communautés afin de mieux les détecter dans des graphes réels.

Chapitre 4 : Contribution et résultats

Notre contribution en dehors d'utiliser des algorithmes performant de la littérature a été de créer une heuristique performante en terme de temps de détection et de simulation de communautés.

Nous avons clos ce manuscrit par une conclusion générale suivie de quelques suggestions que nous avons jugés importantes.

1.5 Résultats

Nous avons simulé, par des Algorithmes issus de la littérature, de grands graphes réels, nous avons également testé des graphes obtenus des réseaux sociaux.

La diversité de ces algorithmes nous a poussé à créer une heuristique personnelle.

Les résultats obtenus mettent tous en évidence des communautés denses bien distinctes les unes des autres que nous avons visualisé par des outils graphiques d'une manière claire et visible.

Les temps d'exécution obtenus lors de ces tests mettent en évidence que notre heuristique a été parmi les plus performants comparativement aux algorithmes présents ce domaine.

Chapitre 2

Les réseaux

2.1 Introduction

En 1989 Tim Berners-Lee,informaticien Britannique de l'organisation Européenne pour la recherche nucléaire ou "CERN", a créé la première serveur (version) du web (World Wide Web) [52], représentée par un essai de quelques pages, qualifiée alors de vague mais prometteur, où il propose un projet hypertexte global permettant aux utilisateurs de travailler ensemble en alliant leurs connaissances. Depuis cette date il n'a cessé de croitre pour atteindre actuellement plusieurs milliards de pages.

Notre époque ("The connected Age)[48] explique la popularité actuelle de la notion de réseaux ainsi que le grand développement de l'Internet, à la vie quotidienne de l'être humain. Certains lie l'avènement des réseaux à l'adéquation de cette notion au monde actuel et notamment à l'économie [10].

Notre vie de tous les jours façonne nos grilles d'interprétation et d'analyse qui sont perméables à l'air du temps. Cette popularité est surtout congruente à la montée en puissance de l'Internet et des gigantesques opportunités offertes par ce dispositif en termes d'accès aux données et de collaborations scientifiques.

2.2 Réseaux d'interaction

On appelle réseau d'interactions, tout groupe d'entités qui interagissent entre eux de façon individuelle [3]

Nous pouvons dire que les réseaux d'interactions recouvrent ainsi des réseaux très différents tel que le réseau des routeurs de la Toile, le réseau des liens sociaux entre

individus, ou le réseau du métabolisme d'un être vivant via les réactions chimiques entre protéines.

Des études récentes sur les graphes réels ont montré qu'en modélisant ces réseaux, des propriétés communes sont observées malgré l'hétérogénéité de leurs origines (Sciences sociales , biochimie, Internet, Web ...etc.).

Il est important de connaitre les propriétés des réseaux d'interactions afin de prévoir leur évolution et de déterminer leurs capacités à résister à différents phénomènes.

2.2.1 Définition d'un graphe d'interaction

Dans de nombreux domaines, les graphes d'interaction, [3] représentent une modélisation. les acteurs sont les nœuds d'un graphe où une arête (arc) modélise une interaction entre les deux nœuds qu'elle relie.

Ce réseau est dynamique, des nœuds peuvent être ajoutés ou supprimés durant son évolution. Une interaction, dans un tel réseau, est définie selon ce que l'on cherche à modéliser (fig 1.1 :Exemple d'interaction entre protéines [20])

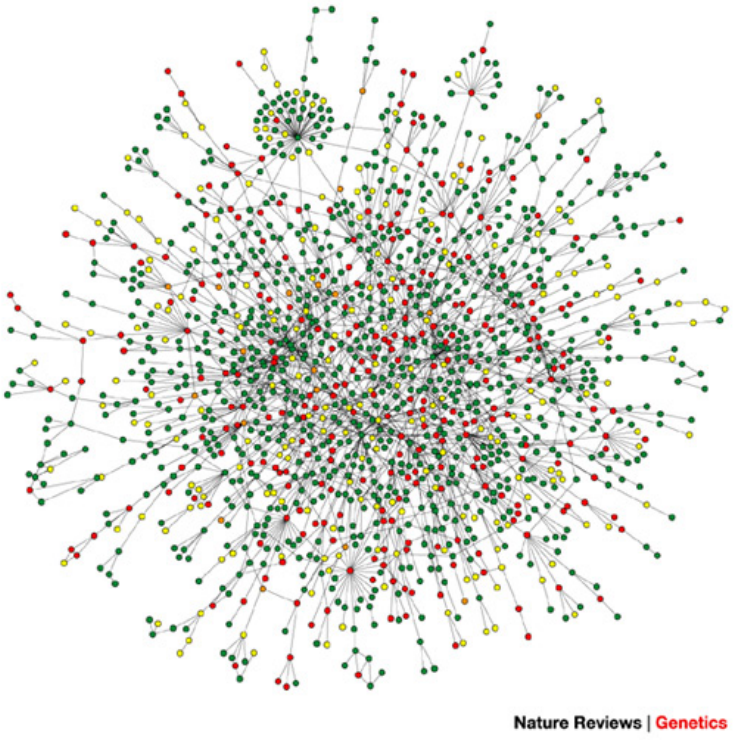


FIGURE 2.1 – Grand réseau d'interaction entre protéines

Parmi les réseaux d'interaction, nous avons :

Réseau Internet :

C'est un des réseaux les plus connu (appelé aussi les réseaux d'Internet)[24]

. Plusieurs applications utilisent ce réseau pour transmettre de l'information, tel que *www* "World-Wide-Web" ou encore p2p "pair-à-pair". Les instances de ces applications sont représentées par des nœuds et les relations qu'elles entretiennent entre elles, peuvent être vues comme des arêtes.

Réseaux sociaux :

Les graphes d'interactions sont utilisés couramment en sciences sociales pour modéliser des interactions entre individus [28]. Les sommets représentent alors les entités ou individus et les arêtes ou arcs une relation ou interaction entre eux.

2.2.2 Propriété des réseaux d'interaction

Vers la fin des années 1990, les chercheurs se sont confrontés à certains des grands réseaux observés sur le web. L'analyse de ces graphes réels a révélé des propriétés inattendues. Les graphes que l'on génère aléatoirement représentent un faible taux de "clustering".

Les grands graphes rencontrés en pratique sont en général :

1- Petit monde (small world)

2- Combinent une distance moyenne faible

3- Ont un fort(nombre important) de groupes de sommets fortement inter connectés (clusters) reliés entre eux par quelques liens.

Ils sont aussi sans échelle (scale free) : la distribution des degrés des sommets est hétérogène, elle suit une loi de puissance[3]. Un réseau de ce type est composé de quelques sommets dotés de nombreuses connections et de beaucoup de sommets faiblement connectés.

Distribution des degrés en loi de puissance

La distribution des degrés d'un réseau d'interactions suit une loi de puissance [2] de type : $P(k) = C * k^{-\lambda}$

Avec C est un paramètre qui dépend de la taille du réseau, k étant le degré et λ est l'exposant de la loi de puissance. Dans la pratique la valeur de λ est comprise entre 2 et 3 et ne dépend pas de la taille du graphe.

Cette distribution en loi de puissance induit qu'il existe beaucoup plus de sommets de faible degré et peu de sommets de fort degré. L'exposant représente la vitesse de décroissance de la courbe des degrés. Plus il est important, plus la probabilité d'obtenir des sommets de fort degré est petite. Les graphes en loi de puissances sont appelés graphes sans échelle (scale-free graphes).

Diamètre

Le diamètre d'un graphe [6] est le plus long des chemins parmi l'ensemble des plus courts chemins pour tout couple de nœuds dans le graphe. La distance moyenne est la moyenne des longueurs des plus courts chemins entre tous les couples de nœuds d'un graphe. Les réseaux d'interactions ont un petit diamètre et une distance moyenne faible de l'ordre de $\log(n)$ où n est la taille du graphe.

Coefficient de regroupement fort

Le coefficient de regroupement (d'agrégation, de clusterisation ou de clustering) est une mesure de la connectivité d'un graphe non orienté, introduite en 1998 par Watts et Strogatz [47] .

Soit $d(u)$ le degré du sommet u ; le coefficient de regroupement du sommet u noté $C(u)$ est défini par la formule suivante :

$C(u) = \frac{(2*nombredeconnexionentrelesvoisinsdeu)}{d(u)*(d(u)-1)}$

Coefficient d'agrégation

Le coefficient d'agrégation [47] C d'un graphe est égal à la moyenne des coefficients d'agrégation de l'ensemble de ses nœuds. Des études sur différents réseaux d'interactions ont montré que le coefficient d'agrégation de ces réseaux est élevé. pour exemple, dans un réseau social, les amis d'un même individu ont une grande probabilité d'être amis entre eux.

2.2.3 Modélisation des réseaux d'interaction

De nombreux scientifiques ont étudiés les grands graphes d'interaction [41] ce qui a permis de découvrir les principales propriétés de ces graphes et donc de les modéliser en se rapprochant des grands graphes réels . Les principaux modèles sont :

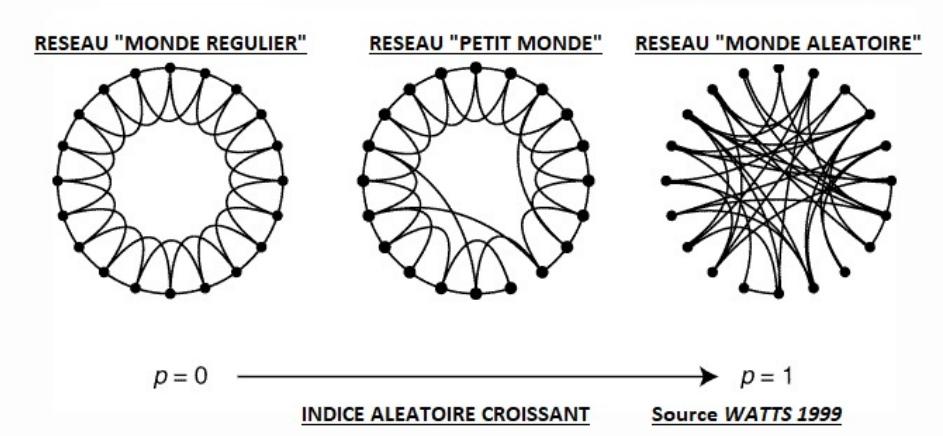


FIGURE 2.2 – Architecture type des réseaux [31]

1- Les Graphes aléatoires uniformes

A priori rien ne relie les différents graphes réalistes. Ce que pourrait expliquer que ces graphes ne sont que le résultat de connexions établies au hasard entre les nœuds.

L'introduction de la théorie des graphes aléatoires a été faite par Paul Erdös et Alfred Rényi, [16], juste après la découverte d'Edrös qui démontre que des méthodes probabilistes résolvent certains problèmes de la théorie des graphes.

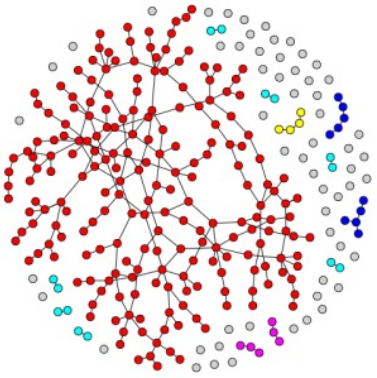


FIGURE 2.3 – Grand graphe aléatoire uniforme - Erdös-Rényi

Donc la seule méthode qui existait au départ pour la modélisation des grands réseaux d'interaction était par des graphes aléatoires Par défaut c'était la solution pour les réseaux réels.

2- Les graphes petit-monde

L'apparition du concept des « six degrés de séparation imaginé par le Hongrois Frigyes Karinthy en 1929 a permis la naissance des graphes petit monde (Small World). Cette théorie se base sur la possibilité que l'on peut trouvé un lien entre toute personne dans cette planète à travers une succession de relations individuelles comprenant au plus cinq maillons de plus.

En effet, le concept (ou notion) de small-world est issu des travaux bien connus de Stanley Milgram [29] dans les années 60. En deux mots, l'expérience consistait à demander à des habitants des états du Middle West Américain de faire parvenir une lettre à un destinataire de la Côte Ouest en utilisant comme intermédiaire des personnes qu'ils appelaient par leur prénom.

Milgram a eu la surprise de découvrir que la moyenne des chaînes parvenues au destinataire n'était que de (5,6). Cette expérience a donné lieu à un mythe devenu dans sa version populaire les six degrés de séparation : en clair, seules cinq personnes nous séparent de n'importe quelle autre personne dans le monde.

Par contre, actuellement, le statut de cette idée comme description de réseaux sociaux hétérogènes reste une question ouverte. Des études sont encore menées actuellement sur le « petit monde ».

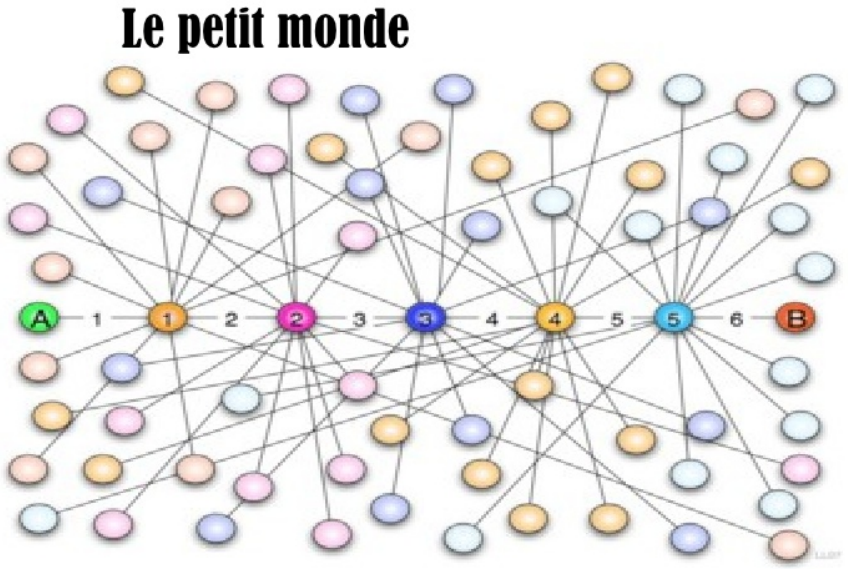


FIGURE 2.4 – Réseau Small World

3- Les réseaux scale-free

Depuis longtemps, les graphes aléatoires ont longtemps été utilisés pour modéliser les réseaux d'interactions. Récemment, des études sur les réseaux d'interactions ont révélées que la distribution des degrés de ces derniers suit une loi de puissance [2]. Alors que la distribution des degrés dans les graphes aléatoires suit une loi de Poisson[35]

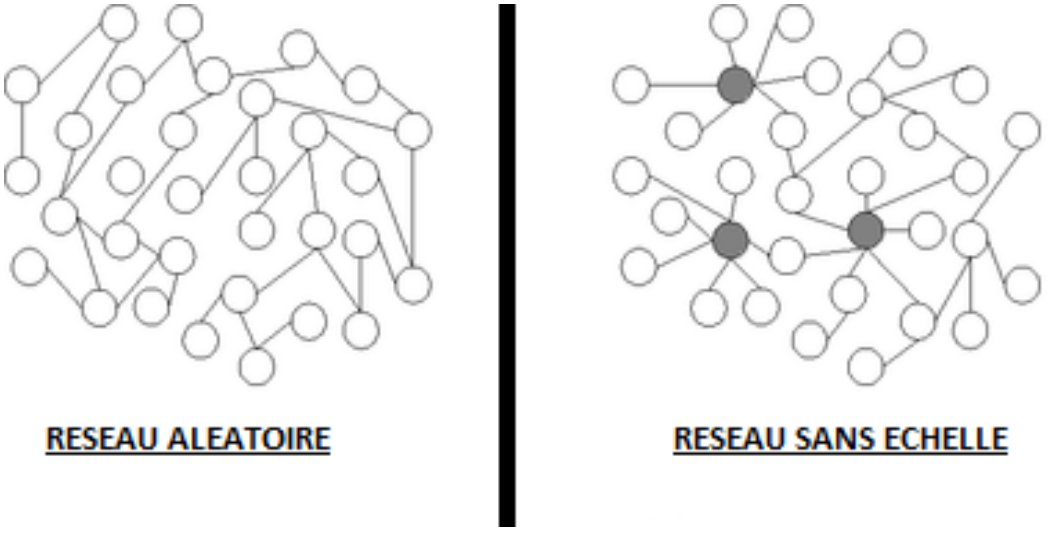


FIGURE 2.5 – Scall free

L'approche et la méthodologie sont radicalement différentes. on part de l'étude de très gros graphes dont on cherche à modéliser la structure et les règles de construction. [5].

Comme expliqué plus haut et contrairement à ce qui était communément admis, les graphes ne présentent pas une distribution gaussienne des degrés mais au contraire une distribution de type loi de puissance.

Les graphes empiriques étudiés mettent en évidence, non des liens aléatoires, mais des liens dûs à un processus d'attachement préférentiel (preferencial attachment).

L'exemple le plus probant est celui d'Internet : tout créateur d'un site sur un sujet précis crée une rubrique liens qui envoie de façon préférentielle vers les sites de référence concernant ce sujet, et la centralité de ces sites de référence tend à augmenter au cours du temps.

4- Graphe biparti :

En théorie des graphes, suivant le Théorème de Koening [30], on dit d'un graphe qu'il est :

a- Biparti :

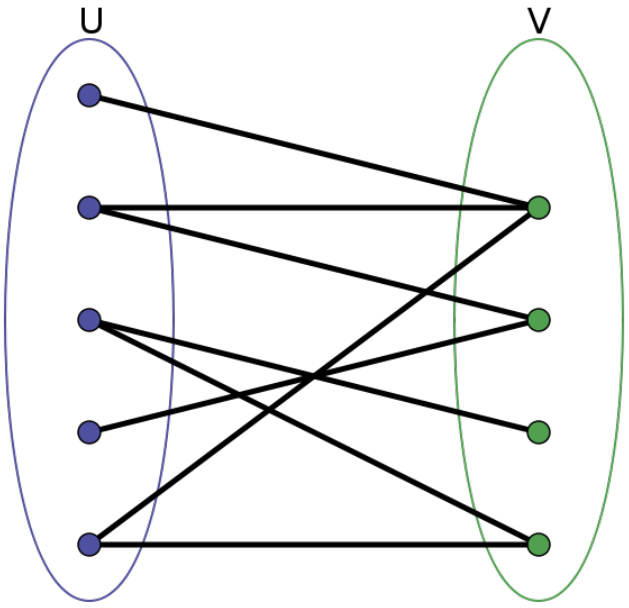
S'il existe une partition de son ensemble de sommets en deux sous-ensembles U et V telle que chaque arête ait une extrémité dans U et l'autre dans V.

Autrement dit, il n'y a aucune arête entre éléments de U et aucune arête entre éléments de V.

Cela nous permet d'étudier et de visualiser les relations entre deux ensembles distincts de sommets, d'où le terme synonyme de réseau 2 modes(2-mode network)

Un graphe biparti permet notamment de représenter une relation binaire.

FIGURE 2.6 – Graphe biparti simple



b- Biparti complet : (ou encore biclique)

S'il est biparti et contient le nombre maximal d'arêtes. En d'autres termes, il existe une partition de son ensemble de sommets en deux sous-ensembles $UetV$ telle que chaque sommet de U est relié à chaque sommet de V Si U est de cardinal m et V est de cardinal n le graphe biparti complet est noté K_{mn}

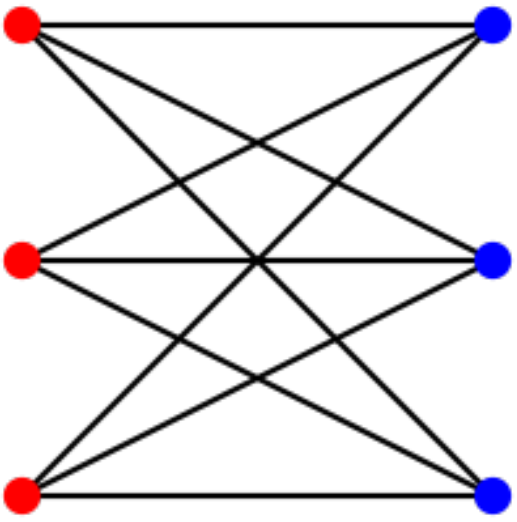


FIGURE 2.7 – Graphe biparti complet

2.3 Réseau social

2.3.1 Définition

Quotidiennement des liens de différentes natures sont tissés avec des personnes. que cela soit de type familial ou professionnel. Cela peut être des liens forts, lorsque la communication est intense. Ou lien faible lorsque nous n'aurons communiqué qu'une seule fois. Toutes ces relations constituent des réseaux sociaux. A l'origine, cela a fait l'objet d'études notamment en sciences sociales par des sociologues, des comportementalistes, des économistes, etc.

La définition de Wasserman d'un réseau social est un ensemble d'acteurs et de relations existant entre eux [46]. Pour exemple du réseau social des élèves d'une école qui ont déjà été dans la même classe. Les acteurs seront alors tous les élèves de l'école. La relation entre deux éléments sera alors "ont déjà été dans la même classe". L'analyse de ces liens peut prédire des caractéristiques des acteurs ou l'apparition de liens entre eux. On peut détecter des fortes connexions entre des groupes d'acteurs.

2.3.2 Analyse des réseaux sociaux

L'analyse des réseaux en sciences sociales se différencie des autres méthodes d'analyse de données, dû au fait qu'elle s'appuie sur l'étude des faits sociaux à partir de l'observation des interactions entre entités sociales. Elle suppose une formalisation relationnelle des données, les entités et leurs relations, dont la traduction sous forme de graphe ou de matrice permet la manipulation et la mesure [41]

C'est un domaine de recherche relativement bien identifié et structuré à l'heure actuelle. Un certain nombre de phénomènes sociaux ont été abordés ou ré-interrogés sous cet angle : l'identité, la dynamique des groupes, la diffusion des innovations, la mobilité sociale, l'amitié, les échanges.

Un large panorama de ces études est accessible dans les travaux actuels qui s'intéressent de plus en plus aux pratiques relationnelles en ligne et se concentrent sur l'étude de la dynamique et de la modélisation des réseaux [7]

Au cours de son évolution, l'analyse des réseaux sociaux a régulièrement fait appel au formalisme mathématique comme les théories des graphes et des ensembles

et le calcul matriciel. L'analyse structurale des réseaux sociaux, s'est développée en référence aux théories structuralistes, elle se focalise sur la découverte des propriétés des structures relationnelles constituées par les réseaux sociaux.

Elle se fonde en général sur l'étude de groupes constitués (institutions, groupes sociaux) ou de relations particulières dont l'étude permet d'expliquer le fonctionnement interne. La plus grande partie des mesures et indices proposés actuellement par les logiciels spécialisés est tirée de cette vision systémique des interactions sociales.

A cette première ligne de partage s'adjoint souvent une différence d'ambition scientifique : si certains présentent l'analyse des réseaux sociaux comme une théorie à part entière, d'autres au contraire n'y voient qu'une simple méthodologie ou une catégorie analytique.

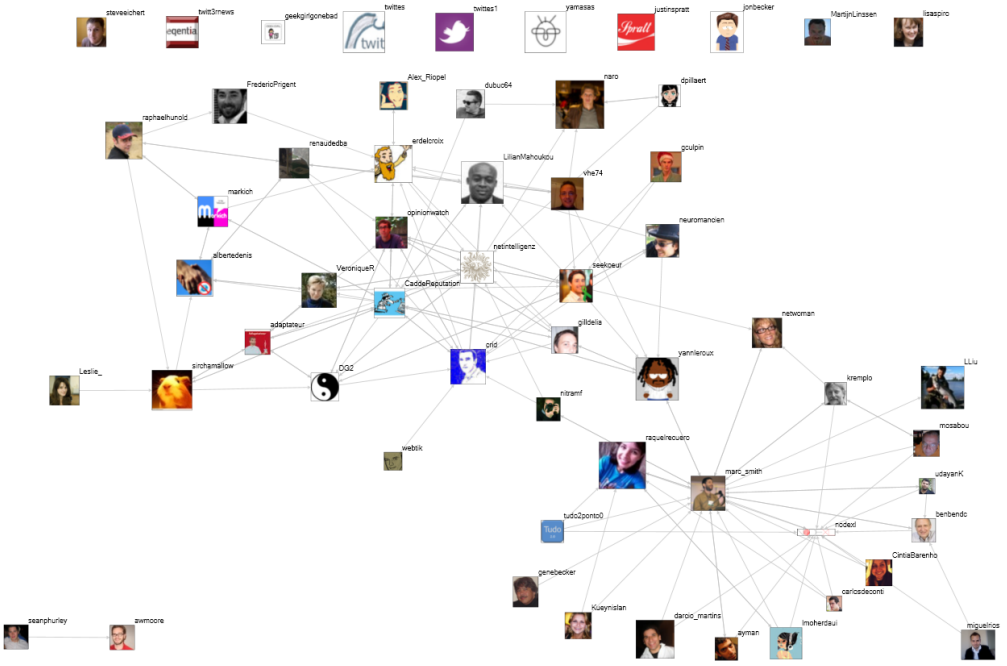


FIGURE 2.8 – Analyse réseaux sociaux

2.3.3 Réseau Social sur internet ou Réseautage Social

En informatique, l'avènement des réseaux sociaux en ligne a conduit à un regain d'intérêt pour leur analyse. Mais cet intérêt est dû non seulement à la disponibilité d'information portant sur les relations qui existent entre les acteurs, mais aussi sur les données permettant de décrire ou de caractériser ces derniers.

On dispose donc de réseaux où les acteurs ne sont pas seulement reliés entre eux mais ont également des informations attachées. Ces données attachées peuvent être des étiquettes, des vecteurs numériques, du contenu textuel, etc. Ces réseaux enrichis, désignés par le nom de réseaux d'information, peuvent être représentés par un graphe dont les sommets sont décrits par des attributs.

De tels réseaux ne sont que partiellement analysés par les méthodes traditionnelles. En effet, si la détection de communautés dans un graphe a fait l'objet de nombreuses recherches ayant abouti à plusieurs méthodes et algorithmes, celles-ci ne prennent pas en compte en général les attributs décrivant les sommets du graphe.

FIGURE 2.9 – "Réseautage" du Web

Par ailleurs la classification automatique, dont l'objet est de regrouper les éléments ayant les mêmes caractéristiques au regard d'une mesure de similarité, a également donné lieu à de nombreux travaux, mais les méthodes de classification automatique ne permettent pas de tirer parti, en plus des valeurs associées aux éléments, de leurs connexions dans un graphe.

C'est la raison pour laquelle des recherches récentes ont été consacrées à la détection de communautés exploitant les données relationnelles et leurs attributs. En effet, la prise en compte conjointe des deux types de données soulève des questions nouvelles liées à la façon de tirer le meilleur parti de l'ensemble des données.

Cette approche est notamment justifiée par le phénomène d'homophilie, qui traduit la tendance qu'ont les individus à se lier avec d'autres ayant des caractéristiques similaires.

2.3.4 Problématique actuelle

Pour information, en 2018, la planète compte plus de 7 milliards d'habitants, dont 4,2 milliards d'internautes (3,4 Milliards sur les réseaux sociaux) et le réseau social Facebook compte à lui seul plus de 2 milliards de comptes actifs (source : Hootsuite and We are Social, 3 ème trimestre 2018). Le nombre d'inscrits sur les réseaux sociaux sur le web augmente de deux cent millions par an.

L'augmentation du nombre de comptes entraine automatiquement une importante augmentation du volume de données produit. La problématique de la volumétrie des données est au cœur des challenges algorithmiques liés aux réseaux sociaux où la taille des réseaux sociaux complique leur analyse.

Pour pallier ce problème, plusieurs méthodes sont envisageables, l'une d'elles est de partitionner les réseaux sociaux en éléments plus petits.

Ce partitionnement permet alors de comprendre le comportement global du système en analysant localement les éléments qui le composent et leurs comportements.

2.4 Représentation des réseaux sociaux

Grâce à la démocratisation des technologies Internet, il devient plus facile de collecter de grandes quantités de données et d'extraire de grands réseaux sociaux afin de les étudier.

Si la collecte des données est grandement simplifiée, le problème de l'analyse se pose de manière aiguë. Plusieurs méthodes et outils existent pour analyser les réseaux sociaux. La majorité de ces systèmes sont basés sur des méthodes statistiques et fournissent un grand nombre de fonctionnalités d'analyse et de modélisation.

Dès les années 1930 sont apparues des représentations visuelles de réseaux sociaux. Par exemple, Jacob Moreno [14] a été un des pionniers à utiliser la représentation nœud-lien pour communiquer sur ses travaux (Figure 1.8). Dans cette représentation, un nœud représente un acteur et un lien représente une relation du réseau.

2.4.1 Analyse par les graphes

On considère le graphe non orienté (qui décrit une relation symétrique entre les sommets du graphe G) : $G = (V, E)$ où V est l'ensemble des sommets et $E \subseteq (V \star V)$ est l'ensemble des arêtes.

On notera N le nombre de sommets de G, avec $N = |V|$

Deux sommets $v, v^{'}$ sont adjacents si ils sont les extrémités d'une même arête du graphe, c'est à dire si $v, v^{'}, \in E$

Les arêtes décrivent les relations entre les sommets du graphe. Elles peuvent être valuées, c'est-à-dire qu'une valeur leur est attribuée. Une valuation forte indique alors une relation de forte intensité. On dit que le graphe est valué.

Les approches classiques pour l'analyse de réseaux sociaux considèrent ces derniers comme des graphe où ils peuvent être représentés à l'aide matrice dite d'adjacence A_{ij} qui indique les connexions entre les nœuds.

Graphe et Matrice correspondante

Sur la figure 1.10 ci-dessous, nous avons un graphe de sept (07) nœuds, huit (08) arêtes ainsi que la matrice d'adjacence correspondante

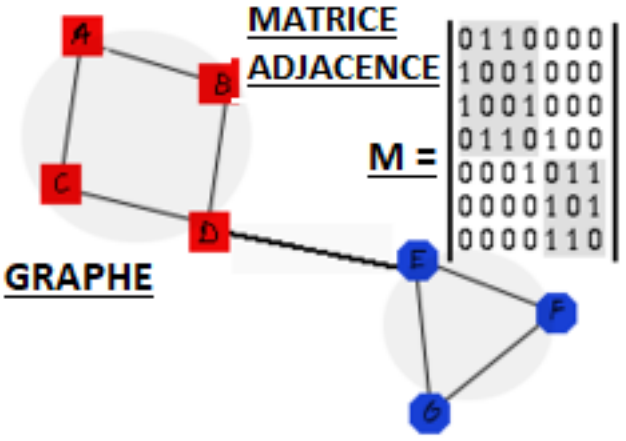


FIGURE 2.10 – Graphe et matrice d'adjacence correspondante

Les sept (07) nœuds sont vont de A à G et une arête a la valeur 1 si le lien existe entre deux (02) sommets (pour exemple il y a un lien entre A et B donc $(A, B) = 1$ et $(B, A) = 1$

Nous pouvons remarquer sur la matrice que les valeurs 1 sont fortement représentés dans les sous matrices AàD et EàG.

2.4.2 Notion clés des graphes

Définition de certaines variables

- Chemin entre 2 sommets :

  Est le nombre de sommets et arêtes qu'il faut parcourir pour aller de l'un à l'autre

- Plus court chemin :

  Nombre d'arêtes minimale qu'il faut pour parcourir la distance entre 2 sommets : C'est la distance géodésique (exemple δ_{IJ} est le plus court chemin entre les sommets I et J

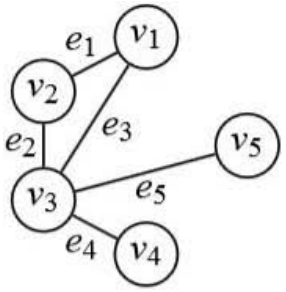


FIGURE 2.11 – Graphe connexe

  Exemple pour la fig 1.11, la distance géodésique entre v_1 et v_4 est :

  $\delta_{v1v4} = 2$ (e_3 plus e_4)

- Voisinage Ni d'un sommet :

  Ce sont les nœuds qui lui sont reliés. exemple pour le sommet v_3 on a $N_3 = 4$

- Graphe complet :

  Un graphe est dit complet si tous ses sommets sont adjacents deux à deux : $\forall(v, v^{'}), \in V \star V, (v, v^{'}) \subseteq E$

- Sous-graphe :

  $G^{'}$ est un sous graphe si $G^{'} = (V^{'}, E^{'})$ avec $V^{'} \subset V$ et $E^{'} \subset E$

  est composé des sommets de $V^{'}$ et des arêtes de E ayant leurs deux extrémités dans $V^{'}$.

- Clique

  $G^{'} = (V^{'}, E^{'})$ est un clique de G si tous les couples de sommets de V' sont reliés par une arête, c'est donc un sous-graphe complet de G.

- Composante connexe d'un graphe :

Un graphe non orienté $G = (V, E)$ est dit connexe si quels que soient les sommets u et v de V, il existe une chaîne reliant u à v

Un sous-graphe connexe maximal d'un graphe non orienté quelconque est une composante connexe de ce graphe.

- Graphe non orienté :

  un graphe est dit non orienté si $\forall(v, v') \in E(v', v) \in E$, c'est-à-dire si les arêtes sont faites de paires de sommets non ordonnées.

  Si les arêtes sont présentes sous forme de couples de sommets, avec une origine et une destination, alors le graphe est orienté.

- Diamètre d'un graphe :

  C'est la plus grande distance géodésique possible entre 2 sommets dans le graphe. exemple entre les sommets v_1 et v_4 de la figure 1.11, le diamètre est de 3.

- Densité d'un graphe :

  Rapport entre le nombre d'arêtes observées et le nombre maximal d'arêtes possibles. Densité = 0, tous les sommets sont isolés ; Densité = 1, tous les sommets sont reliés entre eux.

  un graphe est complet veut dire qu'il y a un lien entre chaque paire de sommets.

- Degré de centralité d'un sommet :

  En terme de voisinage est le nombre d'arêtes qui lui est associé (nombre de voisins adjacents).

  C'est le degré absolu, $C_D(v_i) = deg(v)$

  Pour le degré de centralité relatif,c'est le rapport entre $deg(v)$ et $deg(v') = C'_D(v_i)$ ou v' est le sommet ayant le degré maximum (ayant le plus grand nombre de voisins)

  Dans notre exemple (figure 1.11) . $C_D(v_2) = 2$ et $C'_D(v_2) = \frac{2}{4}$

2.5 Conclusion

Il est évident qu'actuellement, les réseaux sociaux prennent une grande place dans la vie quotidienne. En effet, ils possèdent de nombreux avantages, que cela soit sur le plan social et professionnel pour la majorité de ses utilisateurs. Il s'agit d'un phénomène de mode qui est de plus en plus sollicité par notre société.

A contrario, ils présentent des dangers non-négligeables, car il est facile d'être victime de vol d'identité, de cyber-harcèlement ou de cyberdépendance.

Ces dangers peuvent être quotidien. Nous pouvons dire que les réseaux sociaux possèdent des avantages mais aussi des inconvénients, c'est pour cela qu'il doit prendre en considération des risques qu'encoure un internaute en les utilisant.

Chapitre 3

Les communautés

3.1 Introduction

Les communautés sont variés et les définitions sont nombreuses. Par exemple dans un certain contexte, les membres d'une communauté doivent avoir des caractéristiques proches. Une communauté est en ensemble d'individu ayant plus de chances d'avoir d'avantages de relations internes qu'externes.

La détection de communautés dans un réseau complexe est une tâche qui se rapproche de la classification non-supervisée réalisée en fouille de données classique. Pour cette raison, l'évaluation des algorithmes accomplissant ce type de traitement s'est faite jusqu'ici exclusivement au moyen de mesures comparables à celles utilisées en fouille de données.

3.2 Détection de communautés

3.2.1 Communautés dans les réseaux sociaux

Parmi les principales méthodes de détection de communautés proposées dans la littérature, on peut citer celles qui optimisent une fonction de qualité pour évaluer la qualité d'une partition donnée [26], comme la modularité [32], la coupe ratio [42], la coupe min-max [15] ou la coupe normalisée [11]

- Modularité : La modularité est une mesure pour la qualité d'un partitionnement des nœuds d'un graphe ou réseau, en communautés . Elle est principalement utilisée en analyse des réseaux sociaux. Elle a été introduite par Newman [34].

C'est aussi une fonction d'optimisation pour certaines tâches de détection de communautés dans les graphes. Un bon partitionnement d'un graphe implique un nombre d'arêtes intra communautaires important et un nombre d'arêtes inter communautaires faible.

- Coupe : Une coupe correspond à la partition de l'ensemble de sommets V d'un graphe en deux ensembles disjoints S et T de sorte que l'ensemble des liens du graphe aient leurs extrémités dans chaque sous-ensemble de la partition [43].

  Par exemple on dit d'une coupe qu'elle est minimale si l'ensemble de liens entre les deux sous-ensembles est minimale.

les techniques hiérarchiques comme les algorithmes de division [18] , les méthodes spectrales [45] ou l'algorithme de Markov et ses extensions [39]. Ces techniques de partitionnement de graphes sont très utiles pour détecter des composantes fortement connectées dans un graphe.

L'objectif de la détection de communautés dans les graphes [27], ou dans les réseaux sociaux, est de créer une partition des sommets [40], en tenant compte des relations qui existent entre les sommets dans le graphe, de telle sorte que les communautés soient composées de sommets fortement connectés [32] et qu'elles soient peu reliées entre elles [9] Deux sommets classés ensemble doivent être plus liés entre eux, directement ou par l'intermédiaire d'autres sommets, que vis-à-vis de sommets placés dans d'autres classes.

Considérons un réseau social tel que Facebook ou Twitter [44] nous pouvons dire qu'il représente un ensemble d'acteurs sociaux, représentés par des individus ou des organisations. Les connexions entre eux représentant des interactions sociales. Ce réseau décrit une structure sociale dynamique par un ensemble de sommets et d'arêtes.

L'analyse des réseaux sociaux, basé principalement sur la théorie des graphes et une analyse sociologique, vise à étudier principalement : la détection de communautés

l'identification d'acteurs influents et l'étude et la prédiction de l'évolution des réseaux.

La détection de communautés consiste à former des groupes disjoints ou chevauchant

de sorte que les nœuds au sein d'un même groupe soient connectés d'une manière dense.

Dans le cas particulier de communautés disjointes, cela signifie aussi que les liens entre groupes sont faibles.

3.2.2 Détections des communautés

Soit $G = (V, E)$, graphe non orienté, notre objectif est de partitionner G en sous-graphes denses (communautés)

Le but de la détection de communauté est de trouver une partition $P = C_1, ..., C_k$ de k communautés de l'ensemble des nœuds de V. Chaque communauté C_i représente un sous-groupe de nœuds qui sont fortement connectés plus qu'ailleurs dans le graphe G. (Graphe non orienté)

Il convient de noter que la valeur de k est inconnue et doit être identifiée automatiquement.

Une communauté est formée par un ensemble d'individus qui interagissent souvent entre eux, ils s'agit donc de groupes d'individus qui ont tissés des liens forts.

Au sens du graphe, une communauté est constitué par un ensemble de nœuds qui sont fortement liés entre eux, et faiblement liés avec les nœuds situés en dehors de la communauté.

Les sommets d'une même communauté sont vus de la même manière par un individu à l'extérieur, tous les éléments d'une communauté se connaissent 2 à 2 et deux sommets d'une communauté sont interchangeables.

La détection de communautés a pour rôle de déterminer les groupes formés "implicitement"

En finalité, l'intérêt de la détection de communautés est vaste : identifier des profils types, effectuer des actions commerciales ciblées, ajuster des recommandations, identifier de principaux acteurs. Un réseau peut être partitionné (dans notre exemple de graphe(2.1) ci-dessous, il existe 04 parties)

3.2.3 Communautés disjointes ou chevauchantes

Dans notre exemple de la figure(2.1) , la communauté de couleur violette est disjointe, les trois autres se chevauchent, il y a des individus qui font partis des deux communautés.

Comme en clustering, l'appartenance à une communauté peut être "Crisp" (groupes disjoints) ou floue (groupes avec chevauchement)

A chaque itération, on scinde une classe en deux sous-classes disjointes suivant des principes similaires. Dans l'un ou l'autre cas, l'algorithme produit une hiérarchie de communautés, et l'on retient une partition à nombre de classes voulu, ou composée de classes qui maximisent la modularité. C'est un procédé très général qui peut s'appliquer à bien d'autres critères.

Les méthodes précédemment citées ont l'avantage d'être efficaces et de s'appliquer à de grands graphes. Mais elles produisent des partitions en classes disjointes, ce qui n'est pas toujours désiré ni justifié. En particulier dans les réseaux sociaux, où un individu peut appartenir à plusieurs groupes (de travail ou de relation). De même en Biologie dans lesquels on cherche à identifier des classes fonctionnelles. En effet, de nombreuses protéines ont plusieurs fonctions suivant les différents tissus.

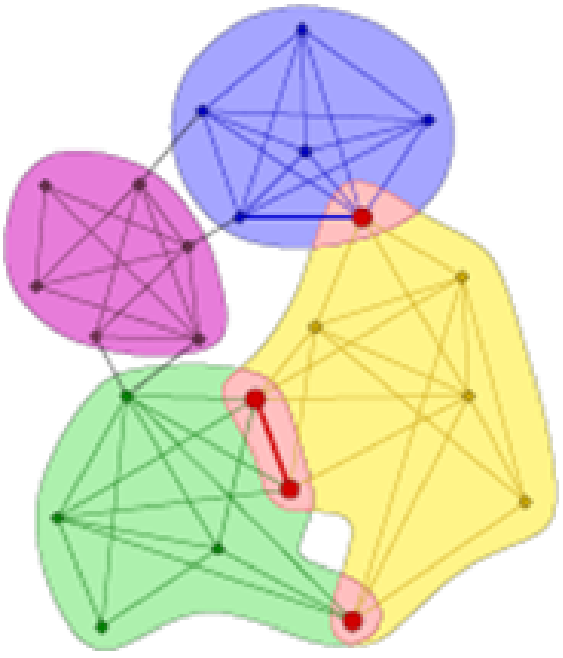


FIGURE 3.1 – communautés

Exemple de communautés sur graphes réels

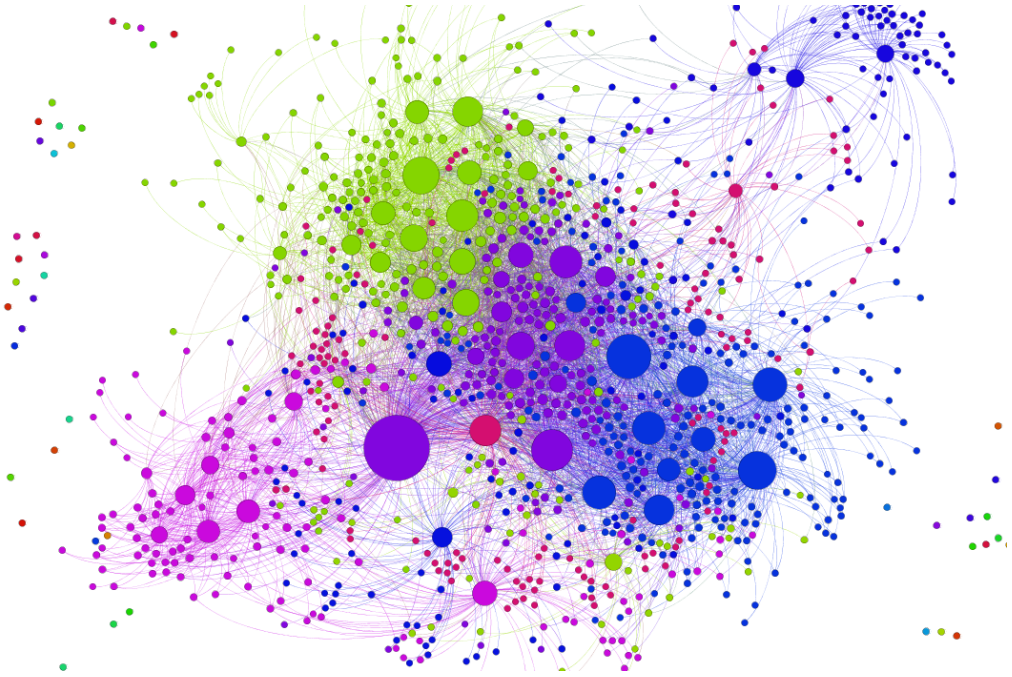


FIGURE 3.2 – Communautés sur graphes réels

Graphe représentant un réseau dense d'utilisateurs facebook, il contient des subdivisions internes appelées communautés. Il existe des méthodes qui permettent de mettre ces communautés en évidence, elles dépendent d'une comparaison de la densité des arêtes à l'intérieur d'un groupe avec la densité des arêtes qui relient ce groupe au reste du réseau.

3.2.4 Variables des réseaux sociaux

Les réseaux sociaux sont constitués de trois types de variables [46]

1- structurelle :

qui représente le graphe et les relations entre les nœuds

2- De composition (ou sémantique) :

pour décrire chaque nœud du réseau à partir de ses attributs (exemple, sexe, nationalité...etc)

3- d'affiliation :

destinée à décrire l'appartenance de chaque nœud à des groupes identifiés dans le réseau

Nous pouvons dire que l'analyse des réseaux sociaux est dédiée à la mesure de chacune de ces variables. Cette analyse permet d'identifier les nœuds importants ou des relations non triviales.

La détection de communautés constitue un type d'analyse mettant en œuvre les trois variables décrites précédemment, la principale idée est de trouver des partitions dont le nombre d'arêtes à l'intérieur des groupes est supérieure au nombre d'arêtes en dehors des groupes.

3.2.5 Filtrage et systèmes de recommandation

Le contenu disponible sur le Web est en très forte expansion, plus particulièrement le Web Social. Pour recevoir des publications de bonne qualité en évitant le superflu, nous devons utiliser un système de filtrage et de recommandation.

Le filtrage consiste à ne recevoir que du contenu des publications auxquels on a souscrit. tandis que la recommandation permet la découverte d'informations à travers la suggestion de nouveaux sujets à suivre.

Pour le calcul de recommandations et pour les réseaux sociaux en ligne, la détection de communautés peut recommander de nouveaux liens d'amitiés.

Dans le contexte de réseaux bibliographiques on peut penser à la recommandation de nouvelles collaborations scientifiques. Dans les applications réelles, les réseaux d'interactions sont des réseaux hétérogènes.

Par ailleurs, nombreuses sont les applications de la détection de communautés. C'est une étape obligatoire pour plusieurs opérations de traitement de grands graphes notamment pour la visualisation [4], la compression [22] et la parallélisation des calculs.

3.3 Algorithmes de détections

Il existe de nombreuses approches proposées, la communauté scientifique a retenu certaines comme les plus intéressantes.

Les Algorithmes de classification ou méthode de clustering [19] de type agglomératifs à l'exemple de Walktrap, ou divisifs à l'exemple de la mesure de la centralité d'intermédialité (Edge Betweness), algorithmes d'optimisation d'une fonction objective à l'exemple Fastgreedy [13] et les algorithmes à base de modèles à l'exemple propagation de Labels [38]

Ces approches illustrent aussi la diversité de méthodologies et donnent une vue d'ensemble des techniques proposées selon leurs principes méthodologiques.

Il s'agit de clarifier la relation entre le problème de la détection de communautés et le partitionnement de graphes.

En effet, le partitionnement de graphes consiste à regrouper les nœuds d'un graphe en un nombre généralement prédéterminé de sous-groupes homogènes en minimisant le nombre de liens entre les différents groupes alors que la détection de communautés effectuée la même opération avec ou sans exiger la connaissance à priori du nombre de communautés.

Nous présentons quelques algorithmes les plus connus de détection

3.3.1 Percolation de cliques

La méthode de percolation des cliques est une approche populaire pour analyser la structure communautaire des réseaux qui se chevauchent [17]. Elle est généralement

définie comme un groupe de nœuds plus étroitement connectés les uns aux autres que les autres nœuds du réseau.

Trois étapes principales caractérisent cet algorithme :

1- Il calcule l'ensemble de cliques de taille k dans le graphe cible G. k étant un paramètre de l'algorithme.

Les communautés sont alors construites à partir de k cliques qui correspondent à des sous graphes complets (tous les k nœuds sont reliés)

pour exemple un réseau k cliques avec $k = 3$ est un triangle.

Soit $C = c_1, ..., c_i$ l'ensemble des $k - cliques$

2- Il construit un graphe de cliques où chaque clique $c_i \in C$ est représentée par un nœud.

Deux nœuds c_i, c_j sont connectés par un lien si les deux cliques associées partagent $k - 1$ nœuds dans le graphe G.

3- Les communautés dans le graphe G sont alors les composantes connexes identifiées dans le graphe de cliques construit à l'étape 2.

Ce type d'algorithme fonctionne dans les graphes plutôt denses mais donne des performances limités dans les graphes de terrain.

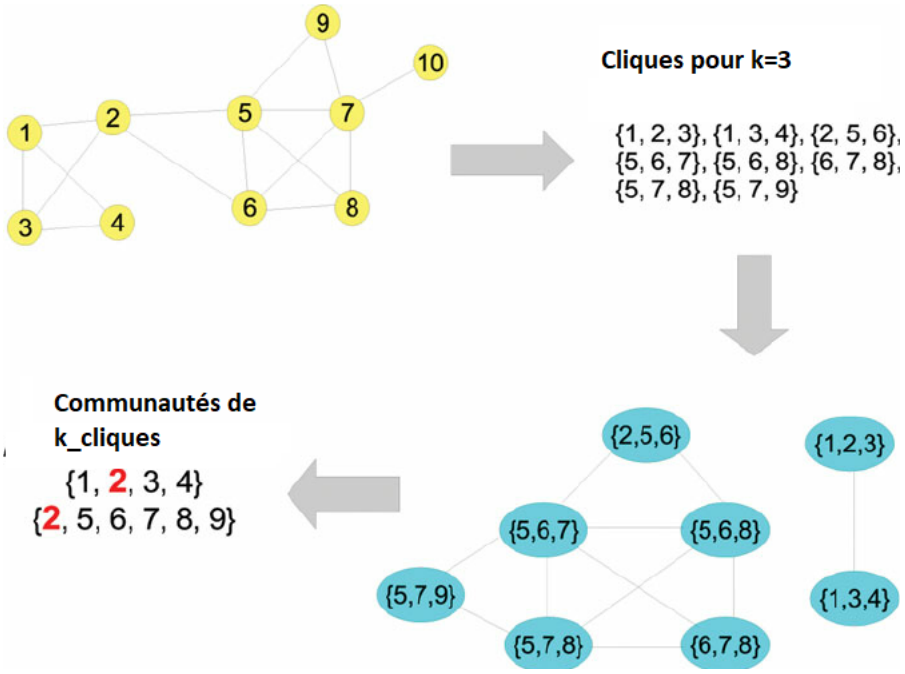


FIGURE 3.3 – Méthode de la percolation de clique

3.3.2 Approches divisives

L'approche divisive pour l'optimisation de la modularité, considère initialement le graphe entier comme une seule communauté.

Puis, itérativement on doit supprimer un lien du graphe pour optimiser la modularité.

- Un premier exemple est l'algorithme de Girvan-Newman [33] [21] où l'heuristique appliquée pour le choix du lien à supprimer tel que à chaque itération, on choisit le lien dont la maximalité de la centra-lité d'intermédiarité est prouvé.

  La centralité d'intermédiarité d'un lien e est donnée par la rapport du nombre de plus courts chemins passant par e et reliant n'importe quels couples de nœuds dans le graphe sur le nombre totale de plus courts chemins dans le graphe.

  Cette heuristique [33] repose sur le fait que les liens inter-communautaire auront forcément une centralité d'intermédiarité importante.

  La complexité de cet algorithme de l'ordre de $O(n^3)$est son principale inconvénient.

- Un autre exemple d'heuristique est celui proposé dans [37]

Dans cet algorithme le lien à supprimer à chaque itération est celui dont le coefficient de clustering est maximal.

Le coefficient de clustering d'un lien est défini d'une manière analogue au coefficient de clustering d'un nœud.

Celui ci est défini par le nombre de circuits qui passent par le lien sur le nombre de circuits possibles.

Le calcul de ce coefficient est local uniquement contrairement à la centralité d'intermédiarité.

La complexité de calcul de l'algorithme est de l'ordre de $O(n^2)$.

Pour cet algorithme, deux critères d'arrêt sont définis :

Le premier pour détecter de fortes communautés où le degré intra-communauté de chaque nœud est supérieur à son degré inter-communauté.

Le deuxième critère concerne la détection de communautés faibles et consiste à continuer l'itération tant que la somme des degrés intra-communauté des nœuds dans une communauté est supérieur à la somme de leurs degrés inter-communautés.

newpage

3.3.3 Algorithme de Louvain

Il est actuellement le meilleur algorithme en terme de complexité pour calculer des communautés sur de très grands graphes (réseaux complexes de graphe de terrain), il est en effet capable de traiter en moins de trois (03) heures des graphes ayant plus d'un milliard de sommets et d'arêtes.

Cette méthode [8] a pour particularité d'implanter une méthode d'optimisation "gloutonne" locale de la modularité.

Nous rappelons qu'en informatique, un algorithme glouton (greedy algorithm en anglais, parfois appelé aussi algorithme gourmand) est un algorithme qui suit le principe de faire, étape par étape, un choix optimum local. Dans certains cas cette approche permet d'arriver à un optimum global, mais dans le cas général c'est une heuristique.

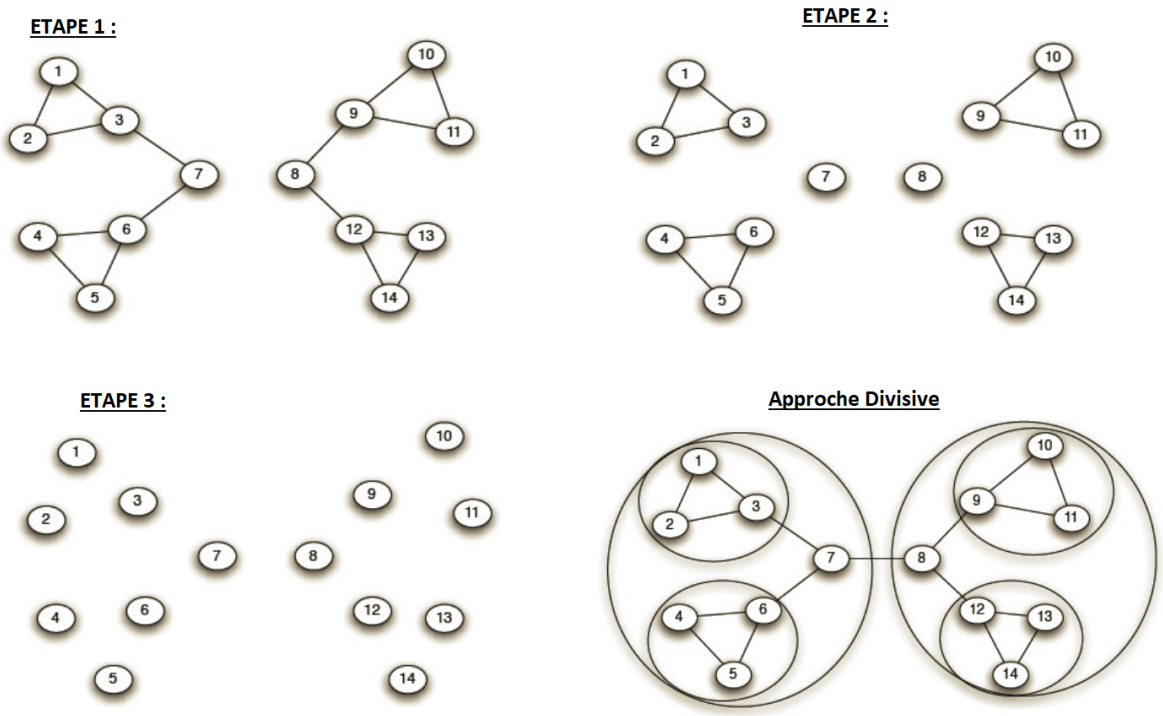


FIGURE 3.4 – Exemple d'approche divisive

A l'état initial chaque nœud est affecté à un communauté différente des autres. L'algorithme applique ensuite une itération de succession de deux phases :

- Phase d'affectation des nœuds :

  Pour chaque nœud x on évalue le gain de la modularité si on le déplace dans la communauté de ses voisins directs.

  On déplace x dans la communauté du voisin qui maximise le gain de la modularité.

  Si aucun gain n'est trouvé le nœud reste dans sa communauté.

- Phase de compression :

  On compresse le graphe obtenu en remplaçant chaque communauté par un seul nœud.

  Deux nœuds c_x, c_y dans le nouveau graphe sont liés par un lien s'il existe un lien entre un nœud de la communauté représentée par c_x et un nœud de la communauté représentée par c_y.

Le poids de lien entre deux communautés est égale à la somme des poids des liens reliant des nœuds de deux communautés.

La complexité théorique de l'algorithme n'a pas été étudiée, mais expérimentalement, cette complexité a été évaluée à $O(n \log n)$ ce qui fait de Louvain la plus rapide des méthodes pour l'identification de communautés.

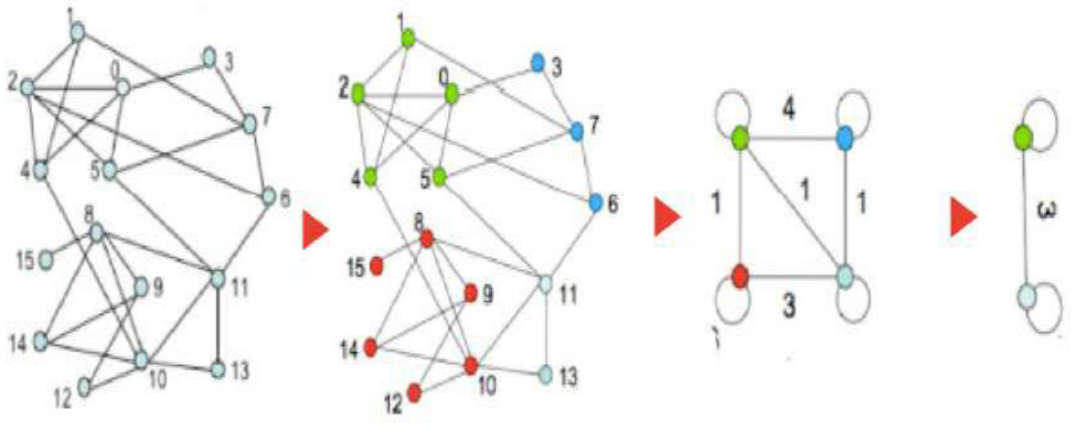


FIGURE 3.5 – Méthode de Louvain

3.3.4 Propagation de labels

Le premier algorithme qui utilisa l'idée de propagation de labels est proposé dans [38].

Basé sur un algorithme itératif où à chaque itération un nœud envoi à la fois son label à ses voisins directs, et reçoit aussi ceux de ses voisins. Chaque nœud détermine le label majoritaire qu'il adopte pour l'itération suivante.

Ce processus itératif donne un consensus sur un label précis pour chaque groupe de nœuds. Cette propagation de labels se fait en mode synchrone ou asynchrone.

Dans le mode synchrone, Il peut y avoir un problème de convergence lié à un échange infini de label entre deux nœuds. Ce problème a été évité dans le mode asynchrone. Une version semi-synchrone qui tente d'avoir les avantages des ces deux versions a été proposée par Bajec [1].

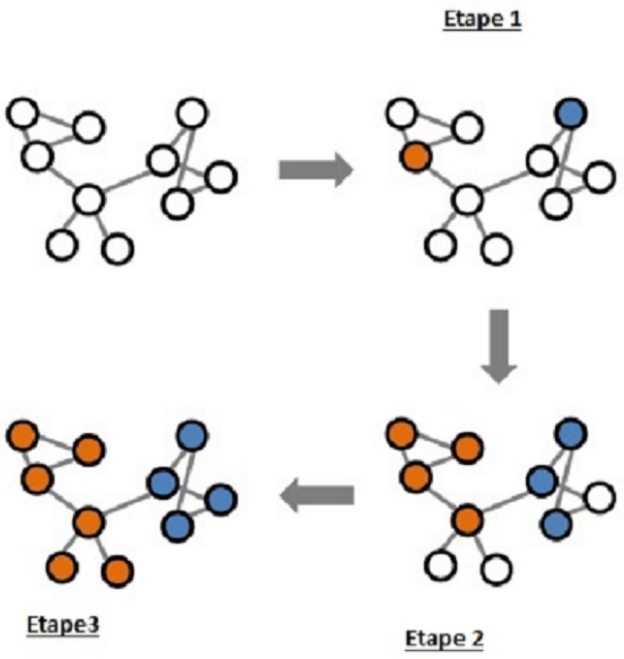


FIGURE 3.6 – Propagation des Labels

3.3.5 Walktrap

Cet algorithme proposé par Pons [36] a pour principe qu'une marche aléatoire partant d'un nœud a un taux de probabilités plus important à rester piégée pendant un certain temps dans la communauté du nœud de départ.

Si nous effectuons une marche aléatoire courte sur le graphe partant d'un sommet v.

La probabilité d'accéder à chacun des voisins de v en une étape est de $\frac{1}{\|\Gamma(v)\|}$

On peut donc calculer de la même manière, la probabilité de se trouver au sommet j en partant de i après avoir effectué aléatoirement k pas.

Cette probabilité définit une distance entre les paires de sommets du graphe dans laquelle deux sommets u et v sont proches à la condition que leur vecteurs de probabilité d'atteindre les autres sommets sont semblables.

nous calculons les probabilités pour toutes les paires de sommets, et notre algorithme les utilise pour partitionner le graphe par une méthode de clustering hiérarchique.

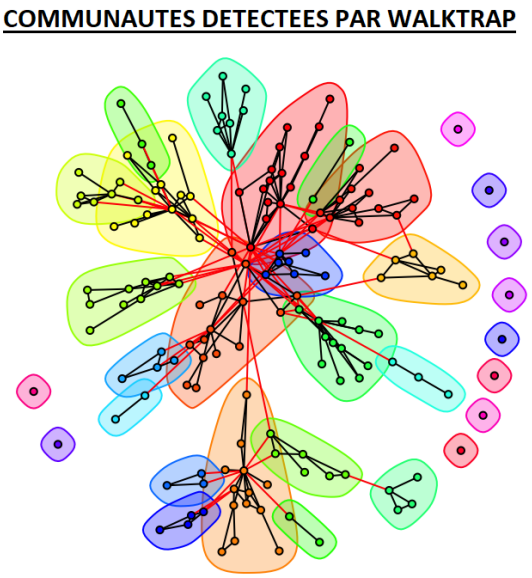


FIGURE 3.7 – Méthode de Walktrap

Commençant par n communautés ne contenant chacune qu'un seul sommet, l'algorithme cherche les deux communautés les plus proches, les fusionne, recalcule les distances, puis effectue une nouvelle fusion et ainsi de suite, jusqu'à n'obtenir qu'une seule communauté recouvrant tout le graphe.

3.3.6 Algorithme LICOD

Proposé dans [25], le principe est que les communautés se forment autour de nœuds spécifiques, appelés leaders. L'algorithme se résume en trois étapes :

1- Identifier de l'ensemble des Leaders L. Un nœud est déclaré leader si sa centralité est supérieure à la centralité de la plupart de ses voisins.

Différentes mesures de centralités peuvent être employées notamment la centralité de degré, la centralité d'intermédiarité et la centralité de proximité . Ici nous utilisons la centralité de proximité. Après, L est réduit en un ensemble C de communautés de Leaders.

Deux leaders sont regroupés s'ils ont un nombre de voisins communs élevé.

2- Chaque $x \in V$ forme un vecteur de préférence P_x^0 où les communautés identifiées dans C sont triées en ordre décroissant.

Le degré d'appartenance d'un nœud x à une communauté$c \in C$ est simplement donné par $min_{c,i}\, dist(x, ci)ci \in c.$(c peut être représentée par un ensemble de nœuds leaders c_i)

3- Chaque nœud aura son vecteur de préférence, on commence alors une phase d'intégration ou le vecteur de préférence d'un nœud est fusionné avec ceux de ses voisins directs.

Cela priorise la classe dominante dans l'ensemble des nœuds voisins.

Des algorithmes issue de la théorie de choix social sont utilisés dans cette phase de vote [12].

A la stabilisation, chaque nœud x est affecté aux communautés placées en tête de vecteur de préférence.

3.3.7 FastGreedy

Tout comme l'algorithme de Louvain, cette méthode est basée sur une optimisation gloutonne de la modularité [13]

En partant du partitionnement le plus fin (une communauté par nœud), les communautés sont rassemblées progressivement si cela permet d'augmenter la modularité.

Cette méthode peut être utilisée pour les réseaux de taille faible (une vingtaine ou trentaine de nœuds) pour avoir un temps d'exécution rapide.

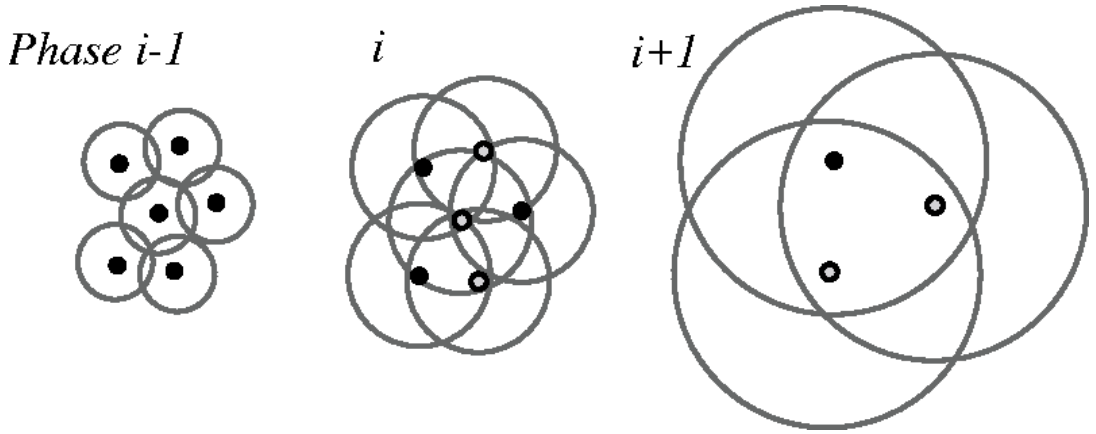


FIGURE 3.8 – Algorithme de Fastgreedy

3.4 Conclusion

Quels algorithmes choisir ?

En première intention, pour les approches de maximisation de modularité, il convient de préférer Louvain à Fastgreedy, en raison de la limite de résolution.

Louvain fait d'ailleurs partie des algorithmes de détection de communauté les plus rapides (plus que Walktrap par exemple).

Chapitre 4

Contribution et Résultats

4.1 Conception et Implémentation

4.1.1 Introduction

L'importante quantité de messages et d'actualités générés par tous les médias sociaux soulèvent de nombreux défis dans le traitement des données et informations à large échelle.

Toutes ces informations créent un besoin croissant pour des méthodes efficaces, capables de filtrer, d'analyser et de s'adapter au grand nombre de données par rapport au profil et besoins des utilisateurs.

Le travail que nous proposons considère la détection et l'analyse des communautés dans les réseaux sociaux.

Plus particulièrement nous nous intéressons à comparer les algorithmes les plus utilisés dans ce domaine. Nous avons également proposé une nouvelle approche.

Par ailleurs les solutions proposées détectent les communautés suivant des concepts différents.

4.1.2 Objectif

Notre projet a été de réaliser, dans un premier temps,plusieurs algorithmes de détection de communautés qui existent dans la littérature (Méthode Louvain et Percolation de cliques) ainsi qu'une heuristique personnelle qui crée et détecte les communautés.

Dans une deuxième phase, implémenter ces algorithmes avec plusieurs tests aléatoires de réseaux, puis de comparer les résultats vitesse d'exécution pour la détection de communautés.

Enfin, utiliser notre logiciel pour tester des réseaux sociaux réels et actifs afin de détecter leurs communautés.

Ce projet a été réaliser sous Netbeans en langage java , il est formé d'un package nommé detectionCommunautes.

Il comprend une classe principale "Pincipale.java" ainsi que de plusieurs autres classes :

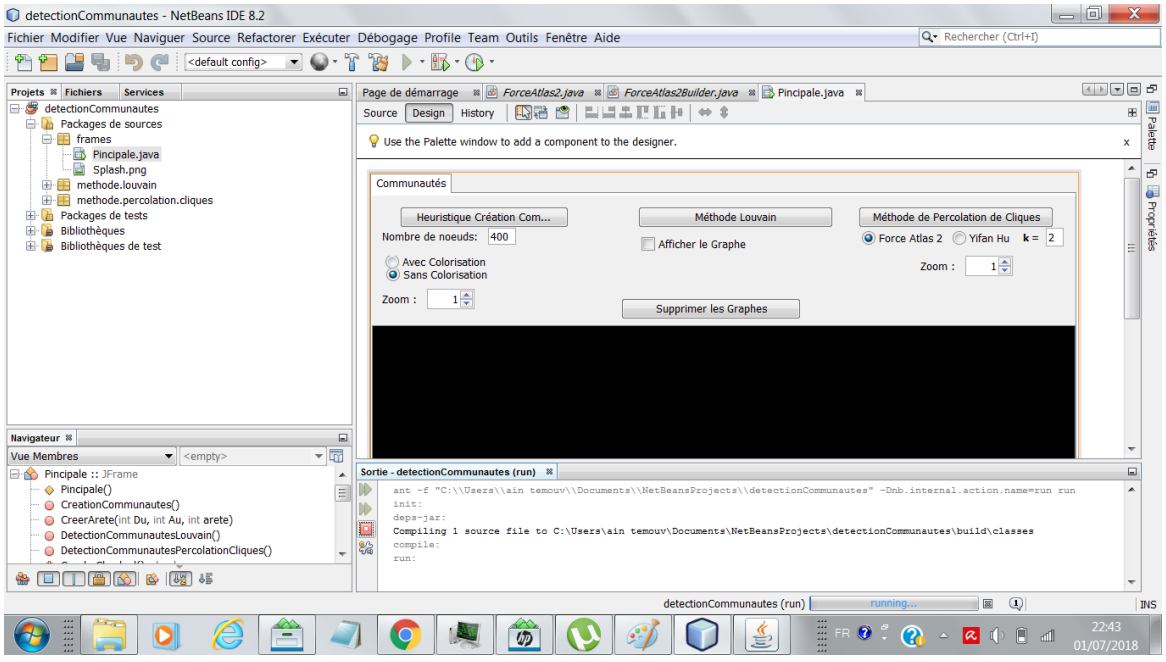


FIGURE 4.1 – Design "Pincipale"

4.1.3 Les Algorithmes programmés

- **Méthode Louvain** :

  Les classes utilisés dans notre programme sont :

  communautes.java et optimiseurModularite.java

  Nous rappelons que la méthode Louvain est un des meilleurs algorithmes en terme de capacité à traiter des graphes de très grandes complexité (plusieurs milliard de sommets et arêtes).

  Il est appelé également algorithme "gourmand". Il a pour principe de faire, étape par étape, un meilleur choix local.

  L'algorithme applique une itération successive de deux phases :

  une première phase d'affectation des nœuds, qui évalue le gain de la modularité lors de déplacement d'un nœud dans la communauté de ses voisins directs.

  On déplace ce nœud dans la communauté du voisin qui maximise le gain de la modularité

  Puis une phase compression :

  On compresse le graphe obtenu en remplaçant chaque communauté par un seul nœud.

  l'affichage du graphe ainsi obtenu se fait en graphstream, qui est bibliothèque Java pour l'édition et la manipulation des réseaux statiques et dynamiques.

- **Méthode Percolation de cliques** :

  Nous avons utilisé deux classes : cpm.java cpm.builder.java .

  Nous rappelons que trois étapes caractérisent cet algorithme

  Premièrement il faut calculer l'ensemble de cliques de taille k dans le graphe cible G.

  Deuxièmement construire un graphe de cliques où chaque clique est représentée par un nœud.

Le résultat étant que les communautés dans le graphe G sont alors les composantes connexes identifiées.

L'affichage du graphe se fait en gephi (02 méthodes sont proposées : Yifan Hu ou Force Atlas2)

- **Heuristique création et détection de communautés** :

La classe est pincipale.java

A partir du nombre de nœuds que définit préalablement l'utilisateur et dans une première étape, nous détectons aléatoirement (en utilisant des méthodes mathématiques appropriées) le nombre de nœuds d'une première communauté que nous créerons puis nous les relierons par un nombre aléatoire d'arêtes, tiré également d'équations mathématiques.

Une considération importante est que nous partant du principe que le graphe créé est connexe.

Par la suite, nous créerons une nouvelle communautés avec le nombre de nœuds et d'arêtes, en utilisant les mêmes méthodes mathématiques, totalement aléatoires.

Puis ainsi de suite et jusqu'à que le total de nœuds de toutes les communautés atteignent le nombre défini par l'utilisateur.

Un autre principe suivi de notre programme est que toutes les communautés sont reliés aléatoirement par un certain nombre de nœuds mais sans constituer une possibilité de communauté plus importante.

L'affichage du graphe a été réalisé par Graphstream, qui est une bibliothèque pour des réseaux dynamiques en Java, nous l'avons proposé en deux modes :

- Multicolore (pour bien visualiser chaque communauté)

- Mono couleur.

4.1.4 Détails du programme

En compilant la classe principale une fenêtre s'ouvre (Jframe : interface graphique) où toutes les commandes exécutant les programmes suivant le choix de l'utilisateur, sont présentes sous la forme de boutons à presser par l'utilisateur.

Dans notre classe principale (appelée "pincipale"), l'appel est fait en fonction du choix de l'utilisateur

1.Choix de la simulation des graphes par notre heuristique :

Nous définissons au départ le nombre de nœuds puis si nous cochons l'option "avec coloration" alors les communautés seront représentées par des couleurs différentes mais le temps d'exécution est légèrement plus important puisque l'assistant graphique graphstream fait la conception différenciée ; sinon si on coche l'option "sans coloration", on aura toujours l'assistant graphique qui mettra en évidence les communautés (nœuds d'une communauté plus proches les uns des autres et avec plus de liens qu'avec ceux d'une autre communauté), simplement il n y aura aucune différenciation de couleurs.

Au final les détails du temps d'exécution, le nombre de communautés avec chacune son nombre de nœuds et d'arêtes sera mis en évidence et un fichier d'extension csv qui représente la liste d'adjacence de ce graphe sera créé.

2.Choix de Louvain :

Sur l'interface, si l'utilisateur clique sur le bouton "Méthode Louvain", il doit choisir ensuite un fichier d'extension csv (pour MicroSoft Excel) ou un fichier texte

Ce fichier d'extension csv représente la liste d'adjacence d'un réseau simulé ou réel (extrait par exemple des réseaux sociaux). nous recherhons l'optimisation de la modularité

La classe optimiseur de modularité :

Le principe du partitionnement d'un réseau est de couvrir tous les nœuds.

La modularité est de définir cette qualité de partitionnement.

Sachant que l'on recherche des sous ensembles fortement denses en définissant la quantité d'arêtes si un sommet (nœud) est dans un sous ensemble C_i au lieu d'un autre C_j

3.Choix de la méthode Percolation de cliques :

Si l'utilisateur choisit "Percolation de cliques", il doit ensuite choisir une liste d'adjacence (fichier d'extension csv ou texte).

le corps du programme est comme suit :

On analyse la structure communautaire des réseaux qui se chevauchent.

Nous recherchons un groupe de nœuds plus étroitement connectés les uns aux autres que les autres nœuds du réseau.

Nous calculons l'ensemble de cliques de taille k dans le graphe cible G. k étant un paramètre de l'algorithme.

Les communautés sont alors détectées à partir de k cliques qui correspondent à des sous graphes complets (tous les k nœuds sont reliés)

Nous rappelons que Soit $C = c_1, ..., c_i$ l'ensemble des $k - cliques$

Deux nœuds c_i, c_j sont connectés par un lien si les deux cliques associées partagent $k - 1$ nœuds dans le graphe G.

Les communautés dans le graphe G sont alors les composantes connexes identifiées dans le graphe de cliques construit à l'étape 2.

Détails de la modélisation mathématique de notre Heuristique

Pour débuter notre programme de création et détection des communautés, l'utilisateur donne le choix du nombre de nœuds V du graphe.

Par la suite la matrice d'adjacence, du graphe, est mise à zéro. Cette matrice est de type $(v+1, v+1)$ pour éviter de définir la première ligne ou colonne par le chiffre "0"

La première étape importante est de simuler aléatoirement le nombre de nœuds de la première communauté.

Nous l'avons établi une équation entre deux variables qui representent les deux bornes minimale et maximale tirés du "nombre d'or" [23], qui est une variable irrationnel et qui a comme valeur 1,618133...

Nous calculons ensuite et toujours aléatoirement le nombre d'arêtes de la première communauté.

Nous définissons une équation entre le nombre d'or et deux variables qui sont :

La première variable est le nombre de d'arêtes qui garantit obligatoirement qu'un graphe soit connexe $\frac{(v-1)*(v-2)}{2} + 1$

La deuxième qui défini le nombre d'arêtes pour qu'un graphe soit connexe complet $\frac{v*(v-1)}{2}$

Ce qui veut dire que le nombre aléatoire d'arêtes calculé est entre ces deux contraintes.

Nous avons donc créée et définie la première communauté

Le nombre de nœuds de la première communauté étant connu, nous redéfinissons aléatoirement un nombre de nœuds pour une deuxième communauté (à la condition que la somme de tous les nœuds soit inférieur à celui fixé par l'utilisateur) puis nous recommençons le même processus pour les arêtes qui devront reliés ces nœuds et ainsi de suite jusqu'à l'atteinte des nœuds fixés par l'utilisateur.

Tant que le nombre de nœuds de départ n'est pas épuisé (par la déduction des nœuds de chacune des communautés créées), nous recommençons ces mêmes étapes.

Nous avons ajouté au rendu visuel de "graphstream" un nœud que nous avons nommé "Master" afin de bien visualiser et différencier les différentes communautés

Notre heuristique calcule simultanément le temps de création des communautés ainsi que le nombre de nœuds et d'arêtes de chacune d'elles.

Il complète la liste d'adjacence à partir des arêtes créées ainsi que le rendu graphique visuel de chaque communauté de manière distinct et bien visible(par graphstream) .

Au final un fichier d'extension "csv" est également créé, contenant la liste d'adjacence.

4.1.5 conclusion

Nous avons implémenté lors de ce projet deux algorithmes , la méthode de Louvain et la méthode de percolation de cliques.

Durant ce travail nous avons eu quelques difficultés : Temps d'exécution important, modification des données pour avoir une liste d'adjacence en format csv.

Nous avons également réussi à implémenté une heuristique qui donne de bons résultats en terme d'exécution en mode mono-couleur et des résultats acceptables en mode multicolore où chaque communauté est représenté par une couleur différente.

4.2 Programme et Résultats

4.2.1 Logiciels de programmation utilisés

Nous avons choisi de programmer les algorithmes de détection de communautés (Louvain, Percolation de clique et une méthode personnelle) dans les réseaux sociaux en utilisant le langage de programmation Java avec l'IDE libre Netbeans, nous rappelons que :

- Java :

  Est un langage de programmation et une plate-forme informatique qui a été créé par Sun Microsystems en 1995.

FIGURE 4.2 – Java

Il est évolué et très puissant permettant de presque tout faire.

Java est complètement orienté objet et un environnement d'exécution. Il permet de développer des applications d'une manière bien structurée, modulable, maintenable beaucoup plus facilement et efficacement. Cela augmente une fois de plus la productivité.

Il possède de nombreuses librairies tierces. les applications sont plus sûres et stables

Contrairement à la plupart des autres langages, Java met à la disposition du développeur une "API" (qui est une interface de programmation applicative : ensemble normalisé de classes, de méthodes ou de fonctions qui sert de façade par laquelle un logiciel offre des services à d'autres logiciels) très riche lui permettant de faire de très nombreuses choses. Java vous propose à peu près

tout ce dont vous avez besoin directement dans le "JDK" (ensemble de bibliothèques logicielles de base du langage de programmation Java)

- Netbeans :

Logiciel Open Source, développé par Sun en 2000.

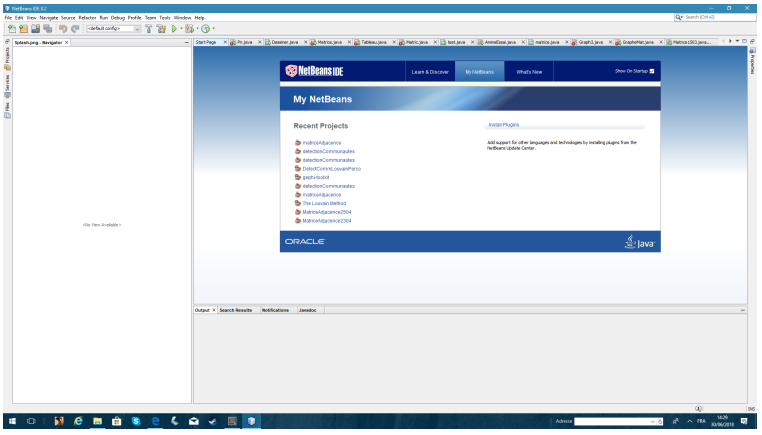


FIGURE 4.3 – Netbeans 8.2

C'est un IDE (Ou EDI, environnement de développement intégré) moderne (éditeur en couleurs, projets multi-langage, refactoring, éditeur graphique d'interfaces et de pages Web), Compilé en Java.

Nous avons choisi d'utiliser la dernière version 8.2

4.2.2 Outils de programmation utilisés

Outre le langage Java et l'EDI Netbeans, Nous avons utilisé 02 micro-ordinateurs, un ordinateur de bureau (Desktop) avec un processeur Intel(R) Core(TM) i3-2100 CPU 3.10 Ghz de 6 Go de RAM et une carte accélératrice NVIDIA de 4 Go, Windows 10 professionnel (64 bits) et un PC portable (LAPTOP) i5 de 4Go de RAM et une carte accélératrice de 4 Go.

Ces outils rapides et puissants nous ont permis de faire tourner des matrices importantes sous l'IDE.

4.2.3 Outils d'affichages graphiques des communautés

Nous avons utilisé les outils graphiques suivants, sous Netbeans :

Graphstream

Editer par "the GraphStream Team", il permet un affichage dynamique graphique des réseaux sous JAVA et Netbeans.[50]

Les nœuds et les arêtes sont en mouvement lors de la création de ces communautés ce qui donne un rendu intéressant.

Gephi

Ce logiciel a été développé depuis 2008 à l'université technologique de Compiègne (UTC), il peut être utilisé sous Netbeans. [51]

Ce logiciel offre néanmoins de vastes possibilités, en particulier en termes de visualisation graphique des réseaux.

Il a été pensé pour mettre en valeur les réseaux « petit monde » ou "Small World"

- **Force Atlas 2 pour Gephi**

  C'est un des algorithmes « maison » de Gephi, Tous les nœuds se repoussent entre eux, respectant le principe des aimants. Sachant que plus les nœuds sont éloignés, moins ils se repoussent.

  On déplace ces nœuds jusqu'à trouver un état stable.

- **Méthode Yifan Hu pour Gephi**

Afin de traiter de plus grandes bases de données, on pourra utiliser « Yifan Hu proportionnal », qui produit des travaux remarquables sur la visualisation de grands réseaux.

4.2.4 Algorithmes

Pour les algorithmes nous avons utilisé deux méthodes présentes dans la littérature ainsi qu'une heuristique personnelle

Louvain

Qui est actuellement un des meilleures algorithmes adaptatif pour la détection de communautés sur de très grands graphes

Percolation de cliques

Qui est approche populaire pour détecter des communautés à partir des $k-$ cliques

Heuristique personnelle

Elle nous permet à partir d'un nombre de nœuds qui est défini, au préalable par l'utilisateur de créer un nombre d'arêtes aléatoires avec la visualisation, par des graphes, du nombre de communautés ainsi créé.

4.2.5 Exemple de compilation

Pour faire un exemple, nous avons choisi un réseau de 521 nœuds les résultats obtenus sont :

1- Compilation de l'heuristique :

Elle a créé et mit en évidence la détection de **six (06) communautés.**

Le temps d'exécution, pour obtenir ce résultat est de **1.392 secondes**

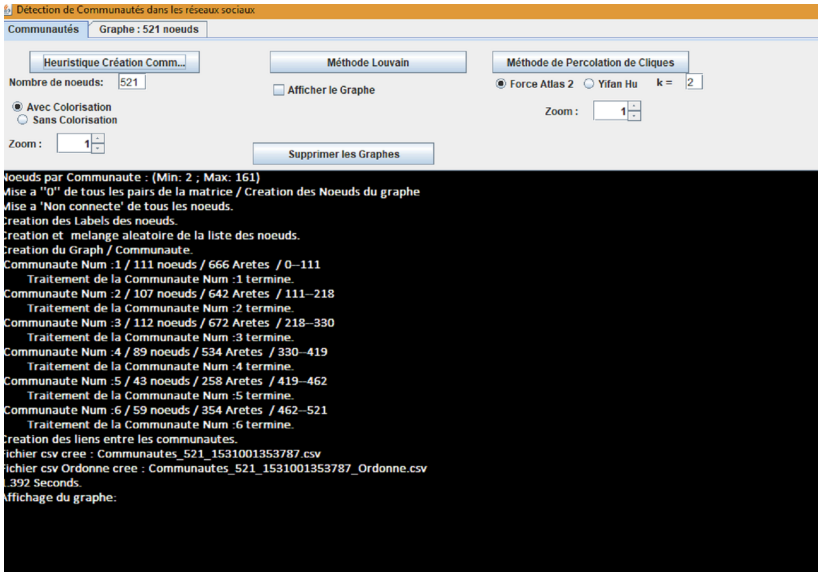


FIGURE 4.4 – Test pour 521 nœuds - Heuristique-

Nous avons également choisi l'option "couleurs" pour mettre en évidence, en couleurs différentes, toutes les communautés.

Nous distinguons bien, sur le graphe obtenu par Graphstream, six (06) communautés de couleurs différentes, cinq sur le pourtour et une au centre.

La communauté minimale a 43 nœuds, la maximale ayant 112 nœuds.

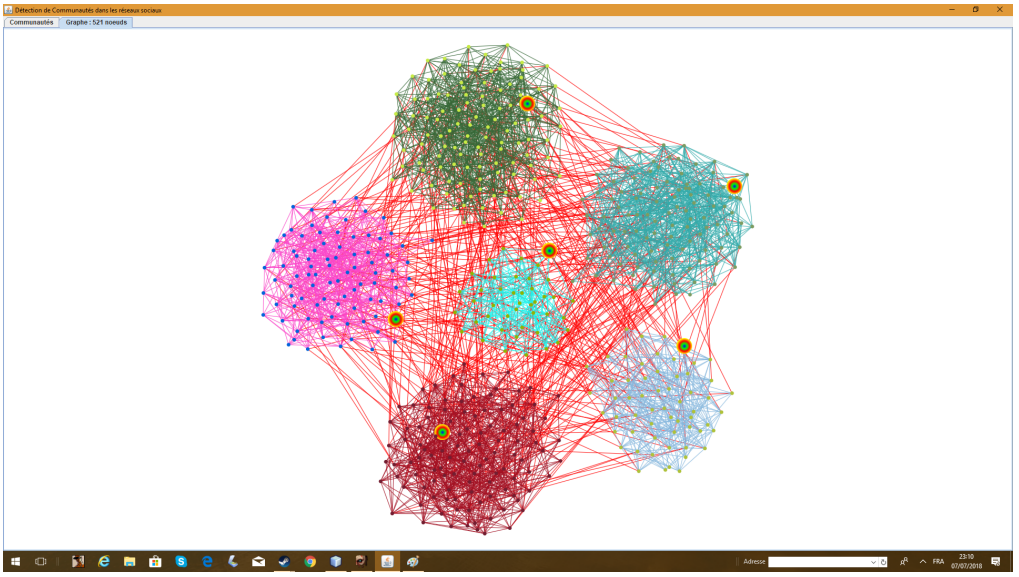


FIGURE 4.5 – Test pour 521 nœuds -Le graphe de l'heuristique-

2- Compilation de Louvain :

L'exécution nous détecte également **06 communautés**

Le temps d'exécution est de **0,211 secondes**

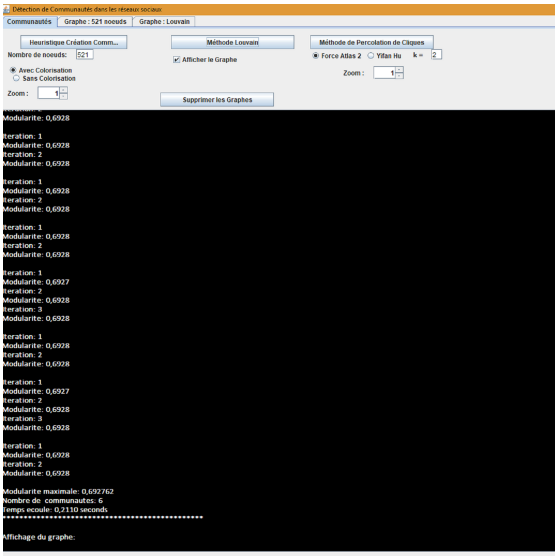


FIGURE 4.6 – Test pour 521 nœuds -Louvain-

Le graphe obtenu, en utilisant la méthode graphstream, met toujours, en évidence 06 communautés situées sur le pourtour.

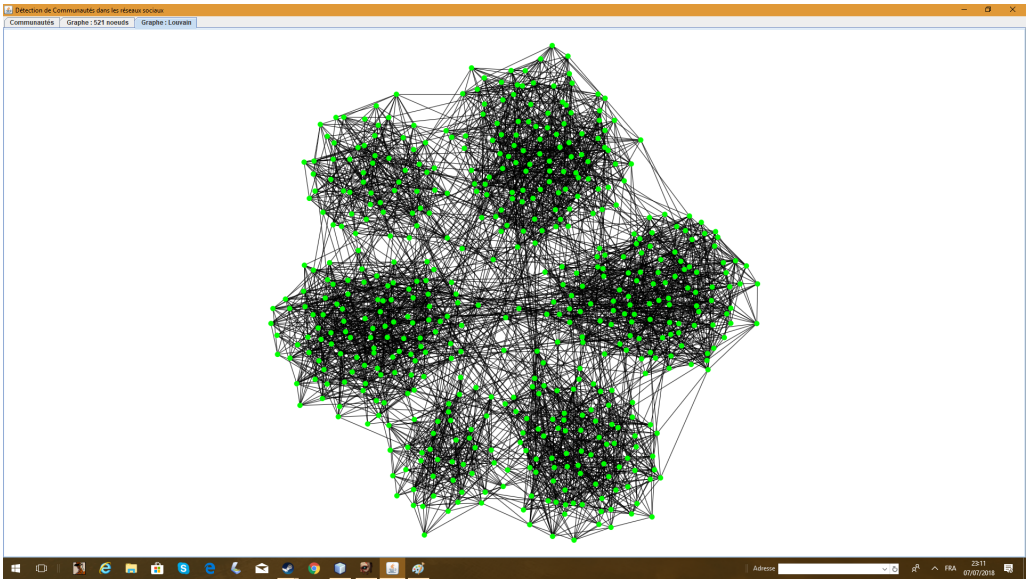


FIGURE 4.7 – Test pour 521 nœuds -Graphe de Louvain

3- Compilation de Clique Percolation Methode avec force Atlas2

Le temps d'exécution pour détecter 06 communautés est de 7,9650 secondes pour notre algorithme.

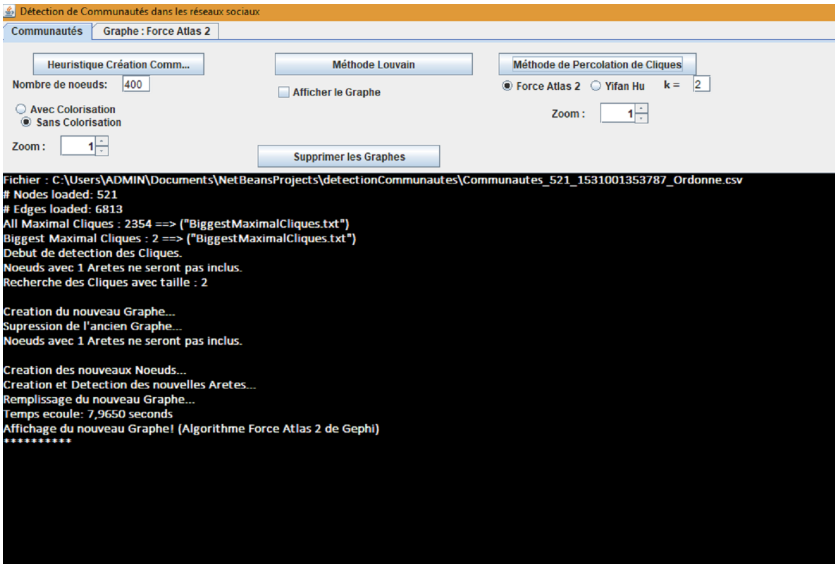


FIGURE 4.8 – Test pour 521 nœuds -Percolation de cliques

Le graphe obtenu pour CPM en utilisant Gephi (Force Atlas2)

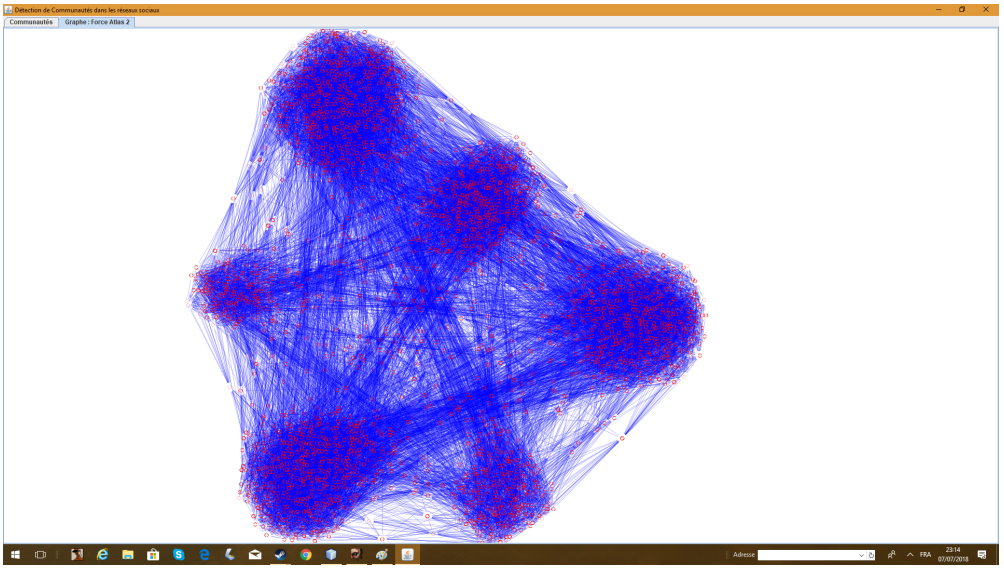


FIGURE 4.9 – Test pour 521 nœuds -Graphe de Percolation méthode Atlas 2 pour Gephi-

4- Compilation CPM avec méthode Yifan Hu

Le temps d'exécution obtenu est de 7,6080 secondes

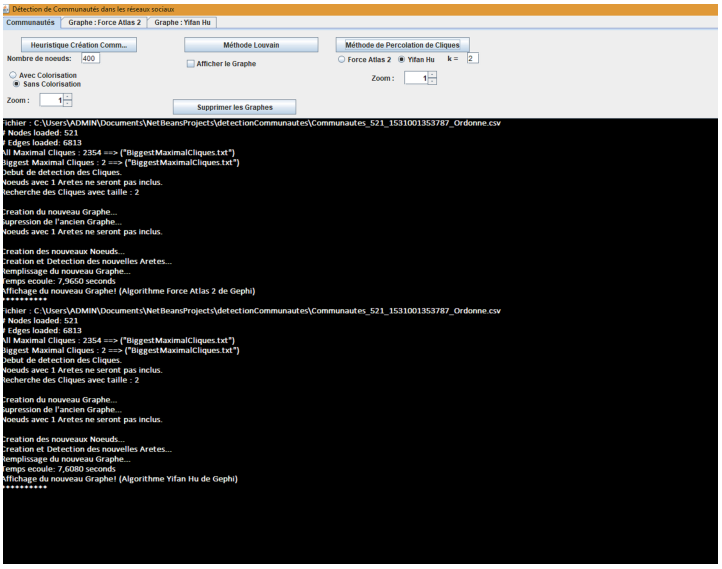


FIGURE 4.10 – Test pour 521 nœuds -Percolation méthode Yifan Hu sur Gephi

Le graphe obtenu, en utilisant la méthode Yifan Hu de Gephi

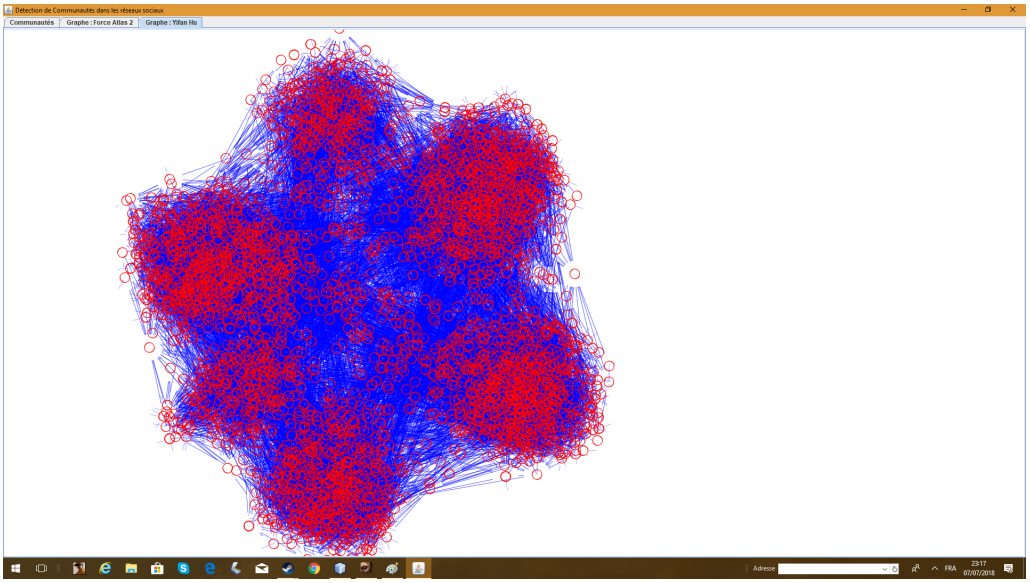


FIGURE 4.11 – Test pour 521 nœuds -Graphe de Yifan Hu

4.2.6 Comparatif de vitesse d'exécution

Pour tester notre programme, nous l'avons compilé avec un nombre de nœuds de départ (voir figure 4.1), nous pouvons constater qu'utiliser percolation pour un nombre assez important de nœuds génère un temps trop important car il réalise un parcours en profondeur de la matrice.

Concernant la distinction entre notre heuristique et Louvain, pour un nombre (Environ moins de 20.000 nœuds), Louvain donne des temps de détection plus rapide, au delà nous considérons que les deux sont assez proches. Mais nous devons prendre en considération que notre heuristique crée et forme les communautés.

N	Nbre de Noeuds	Nbre commu	Nbre Aretes	Louvain	Percolation	Heuristique
1	188	6	1099	0,053s	1,118s	0,078s
2	888	8	5882	0,091s	18,88s	0,375s
3	8880	8	61824	0,827s	Important	1,275s
4	12000	6	83753	3,494s	Important	3,588s
5	25000	9	174608	17,191s	Important	16,271s

TABLE 4.1 – Comparatif des trois programmes

4.2.7 Détection de communautés sur réseaux sociaux

Nous testons ensuite notre programme sur des cas de réseaux sociaux. Nous nous baserons sur des pages Web qui nous permettent d'avoir des matrices de réseau sociaux (exemple ci-dessous d'une page de l'université de Stanford [49]

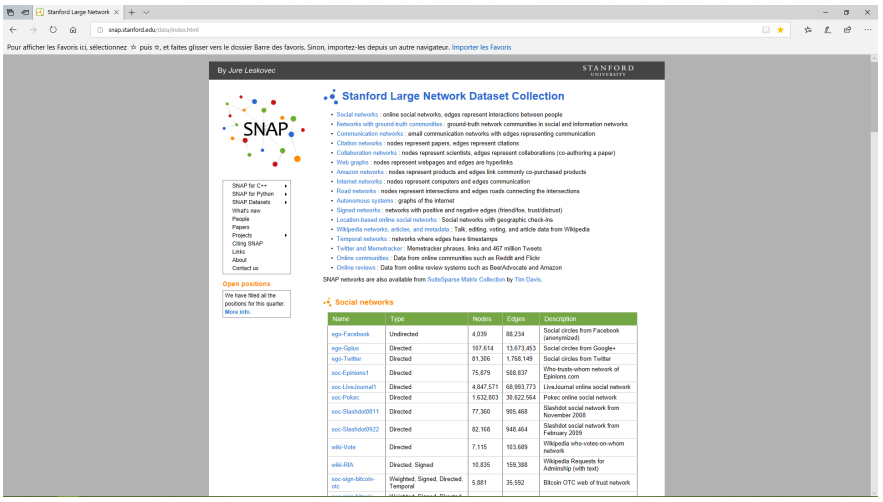


FIGURE 4.12 – Page Stanford Université

Après avoir transformés les données, du format texte de fichier de départ texte en format csv (Ms-Excel) pour la liste d'adjacence, nous le compilons avec la méthode Louvain

Test pour email-Eu-core :

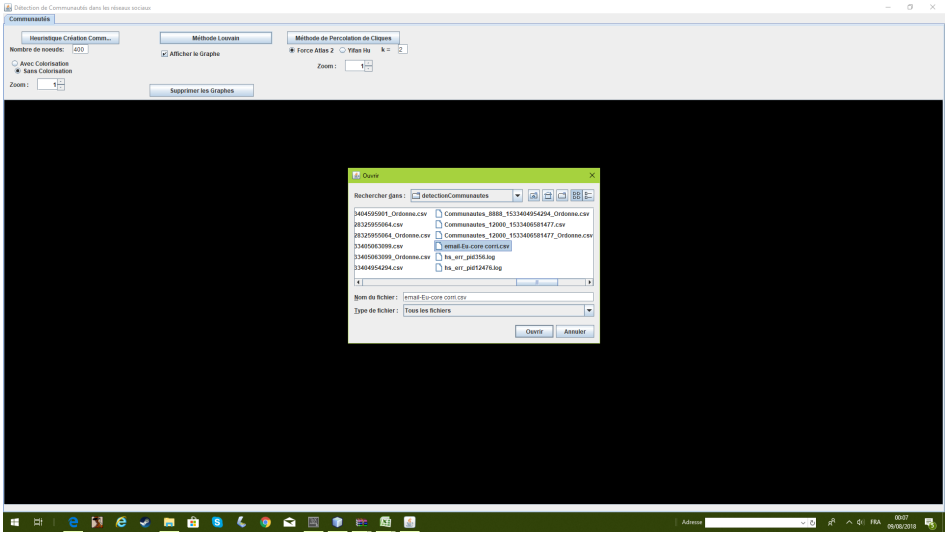


FIGURE 4.13 – Choix de Email EU Core

Sachant que nous avons réorganisé, ce fichier, pour plus de maléabilité, nous obtenons 48 communautés pour 1004 nœuds.

Le temps d'exécution est de **0,4380 secondes** et le résultat est :

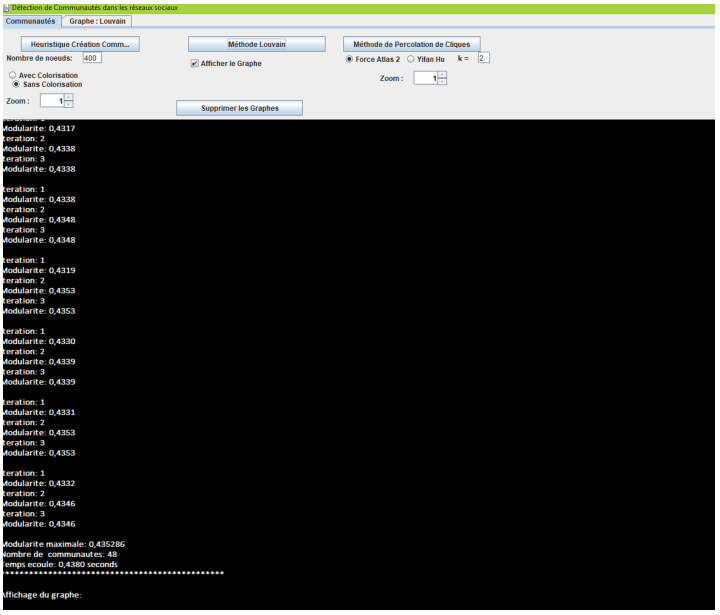


FIGURE 4.14 – Résultat pour Emai EU Core

Test pour wiki-vote, qui est issu de Wikipedia Network (Data et Metadata)

Pour une liste d'adjacence de 8297 nœuds et 71032 arêtes.

Le temps d'exécution est de **0,6410 secondes** et le résultat est : 3021 communautés

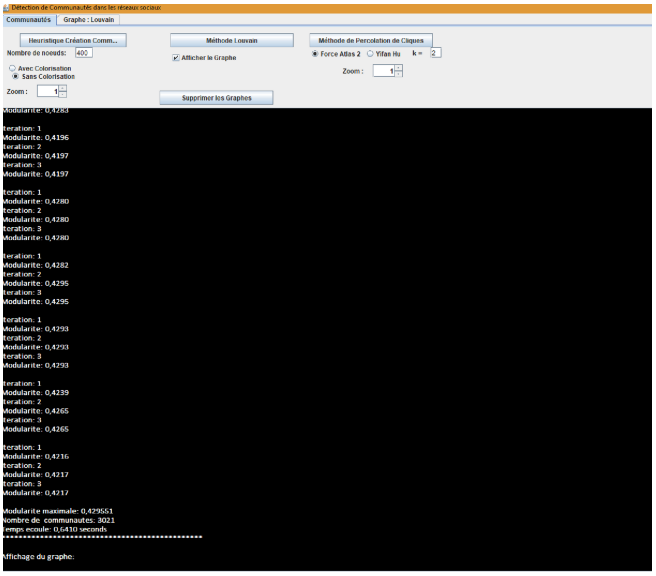


FIGURE 4.15 – Résultat pour Wiki Vote

Test pour ego facebook

Le nombre de nœuds est de 4038 et le nombre d'arêtes est de 88233

le nombre de communautés est de 15 pour un temps d'exécution de 0,6420 s

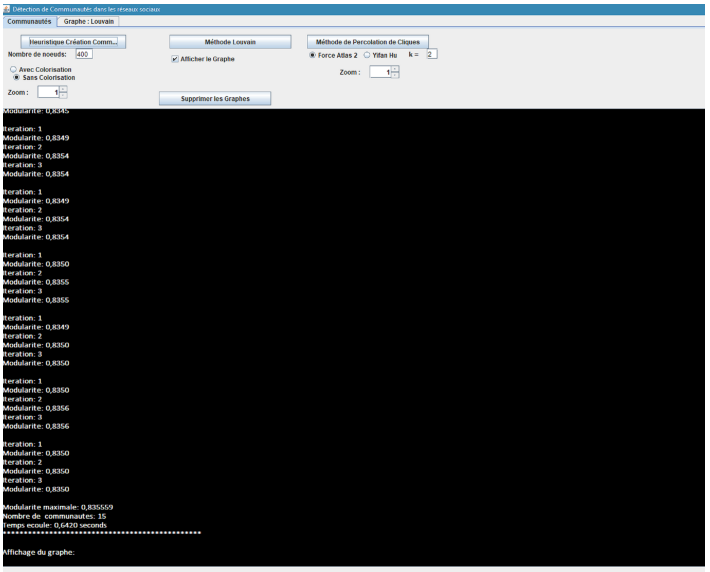


FIGURE 4.16 – Résultat pour Facebook

4.2.8 Conclusion

Nous avons testé les algorithmes présent dans la littérature (Louvain, Percolation de cliques). La diversité des algorithmes nous a poussé à créer une heuristique personnelle.

Nous avons recréé (simulé) des grands graphes réels, que nous avons obtenu de nombre aléatoires de nœuds. Nous avons également testé des graphes obtenus des réseaux sociaux.

Les résultats obtenus mettent tous en évidence des communautés denses bien distinctes les unes des autres que nous avons visualisé par "Graphstream" ou "Gephi" d'une manière claire et visible.

Les temps d'exécution obtenus lors de ces tests mettent en évidence que la méthode Louvain ainsi que pour notre heuristique ont été les plus performants dans ce domaine.

L'algorithme de percolation de cliques passe par plusieurs étapes ce qui lui rend très lent en exécution d'un nombre important de ne contrairement à Louvain qui est une méthode d'optimisation gloutonne indiqué pour les réseaux complexes. Mais ceci n'empêche pas qu'il est moins rapide en passant un certain seuil, puisque il décompose le graphe initiale en petit graphe ce qui le ralenti en passant à l'échelle.

Conclusion générale

Les activités interactives représentés par des graphes permettent d'appliquer les approches d'exploration, d'analyse ainsi que de fouille de réseaux complexes.

Le principal problème dans l'étude des réseaux complexes est celui de la détection des communautés qui sont représentés par des sous-graphes fortement denses mais faiblement connectés entre eux.

La littérature scientifique concernant la détection de communautés est très abondante. Il existe un nombre impressionnant d'approches différentes développés dans différentes disciplines.

Dans notre travail, nous sommes partis du principe que notre objectif est de faire une étude sur la détection des communautés dans les différents types de réseaux.

Pour ce faire, nous avons été amenés, dans une première étape, à analyser le fonctionnement et comportement de ces derniers puis nous nous sommes donnés les moyens afin de pouvoir exploiter leurs propriétés en exploitant des algorithmes présents dans la littérature , comme les approches rapides sur des très grands graphes, l'optimisation gloutonne de la modularité, ainsi qu'une heuristique personnelle qui simule des grands graphes en comptabilisant leurs communautés bien visibles et distinctes entre elles.

Perspective

Dans le future, il serait intéressant de modifier l'heuristique afin d'avoir une nouvelle possibilité de pouvoir détecter les communautés réelles des réseaux sociaux en s'appuyant sur une matrice d'adjacence ou liste d'adjacence.

Chapitre 5

Bibliographie

[1] Bajec, M. (2011). "Robust network community detection using balanced propagation". Nature. 40

[2] Barabasi, Albert-László ; Albert, Réka. (October 15, 1999)"Emergence of scaling in random networks" 12, 16

[3] Barabasi (2002) "Linked : The New Science of Networks". Perseus. 10, 11, 12

[4] Bastian, M., S. Heymann, et M. Jacomy (2009) "Gephi : An Open Source Software for Exploring and Manipulating Networks". Dans International AAAI Conference on Weblogs and Social Media, pp. 361–362. 35

[5] Beauquitte L, Ducruet C (2011) "Scale-free, small-world networks et géographie". In HAL 17

[6] J. C- Bermond et B. Bollobas (1981) "The Diameter of Graphs : a Survey, Congressus Numerantium", vol 32, 1981, p. 3-27 13

[7] Bidart C., Lavenu D. (2005) "Evolutions of Personal Networks and Life Events" 19

[8] Blondel, V. D., Guillaume, J.-l., and Lefebvre, E. (2008). "Fast unfolding of communities in large networks". pages 1–12. 38

[9] Brandes, U., M. Gaertler, et D. Wagner (2003) "Experiments on Graph Clustering Algorithms".11th European Symposium on Algorithms, pp. 568–579. Springer. 30

[10] Castells M. (1998)"La société en réseaux. Tome 1 : l'ère de l'information" . Paris, Fayard 10

[11] Chan, P. K., M. D. F. SCHLAG, et J. Y. Zien (1994) "Spectral K-way ratio-cut partitioning and clustering". IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 13(9), pp. 1088–1096. 29

[12] Chevaleyre et al. (2007) "Allocating goods on a graph to eliminate envy". 43

[13] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest et Clifford Stein (2001) " Introduction to Algorithms" MIT Press et McGraw-Hill, 2e éd. chap. 23 : Minimum Spanning Trees, p. 561-579 35, 44

[14] Daniel Derivry, « MORENO JACOB LEVY - (1892-1974)" Encyclopædia Universen " 23

[15] Ding, C., X. He, H. Zha, et M. Gu (2001) " A min-max cut algorithm for graph partitioning and data clusterin". Dans Proceedings IEEE International Conf on Data Mining, pp. 107–114. 29

[16] Erdős, P. ; Rényi, A. (1959 " On Random Graphs. I". Publicationes Mathematicae. 6 : 290–297. 14

[17] Evans, T S (2010)"Clique graphs and overlapping communities". Journal of Statistical Mechanics : Theory and Experiment. 2010 (12) : P12037. arXiv :1009.0638. Bibcode :2010JSMTE..12..037E. doi :10.1088/1742-5468/2010/12/P12037 35

[18] Flake, G., R. Tarjan, et K. Tsioutsiouliklis (2003) " Graph clustering and minimum cut trees". Internet Mathematics 1(4), pp. 385–408. 30

[19] Gabor J. Székely et Maria L. Rizzo (2005) " Hierarchical clustering via Joint Between-Within Distances : Extending Ward's Minimum Variance Method. " Journal of Classification, vol. 22 35

[20] Gavin et al. (2002) " protein identification Bioinformatic interpretation data" Nature, 415 : 11

[21] M. et M. Girvan (2004) " Finding and evaluating community structure in networks". Physical review E 69(2), pp. 1–16. 37

[22] Hernandez and Navarro, (2012) "Compressed representation of web and social networks via dense subgraphs". In Calderon-Benavides, L., Gonzalez-Caro, C. N., Chavez, E., and Ziviani, N., editors, SPIRE, volume 7608 of Lecture Notes in Computer Science, pages 264–276. 35

[23] Roger Herz-Fischler (1998) " A Mathematical History of the Golden Number", Dover 51

[24] Christian Huitema(1996) " et Dieu créa l'Internet..., Eyrolles", février 1996 12

[25] Kanawati, R. (2011). "Licod : Leaders identification for community detection in complex networks". In SocialCom/PASSAT, pages 577–582. IEEE. 42

[26] Kernighan, B. W. et S. Lin (1970) "An efficient heuristic procedure for partitioning graphs". Bell System Technical Journal 49(2), pp. 291–307 29

[27] Lancichinetti, A. et S. Fortunato (2009) "Community detection algorithms". a comparative analysis. Physical review E 80(5), pp. 056117. 30

[28] E. Mazzoni (2006) " Du simple tracement des interactions à l'évaluation des rôles et des fonctions des membres d'une communauté en réseau : une proposition dérivée de l'analyse des réseaux sociaux" ISDM – Information Sciences for Decision Making, 25, 2006, pp. 477-487 12

[29] S. Milgram et J. Travers (1969) "An Experimental Study of the Small World Problem".Sociometry, 32(4) :425–443 15

[30] M Wate Mizuno (2010)"The works of Konig Dénes (1884-1944) in the domain of mathematical recreations and his treatment of the recreational problems in his works of graph theory" (Thesis, University of Paris, Diderot (Paris 7) 17

[31] Cristopher Moore and M. E. J. Newman (2000) "Epidemics and percolation in small-world networks". Physical Review E, 61 :5678–5682, 2000 3, 14

[32] Newman, M. (2004)."Detecting community structure in networks". Physical Journal B-Condensed Matter and Complex Systems 38(2), pp. 321–330. 29, 30

[33] Newman, M. E. J. (2004a) " Coauthorship networks and patterns of scientific collaboration". Proceedings of the National Academy of Science of the United States (PNAS), 101 :5200–5205 37

[34] M.E.J Newman (2006) " Modularity and community structure in networks" USA, vol. 103 p. 8577–8582 29

[35] Siméon-Denis Poisson,(1837) "Recherches sur la probabilité des jugements en matière criminelle et en matière civile ; précédées des Règles générales du calcul des probabilités" 16

[36] Pons, P. and Latapy, M. (2006). "Computing communities in large networks using random walks". J. Graph Algorithms Appl., 10(2) :191–218. 41

[37] Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., and Parisi, D. (2004) "Defining and identifying communities in networks". In Proc. Natl. Acad. Sci. USA, pages 2658–2663. 37

[38] Raghavan, U. N., Albert, R., and Kumara, S. (2007). "Near linear time algorithm to detect community structures in large-scale networks". Physical Review E, 76 :1–12. 35, 40

[39] Satuluri, V. et S. Parthasarathy (2009) " Scalable graph clustering using stochastic flows : applications to community discovery". Dans Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 737–746. 30

[40] Schaeffer, S. (2007) "Graph clustering". Computer Science Review 1(1), pp. 27–64. 30

[41] Scott (1991) "Ils sont aussi appelés aussi réseaux hétérogènes, réseaux a deux dimensions (two-mode network), et pour la théorie des graphes, graphes bi-partis." 14, 19

[42] Shi, J. et J. Malik (2000) " Normalized cuts and image segmentation". Pattern Analysis and Machine Intelligence, IEEE Transactions on 22(8), pp. 888–905. 29

[43] L. Sunil Chandran et L. Shankar Ram (2002) " On the Number of Minimum Cuts in a Graph ", dans Computing and Combinatorics 30

[44] Taibi,M. (2013) " Une nouvelle approche de détection de communautés dans les réseaux sociaux ". Université du Québec en Ouatoutais 30

[45] Von Luxburg, U. (2007) " A Tutorial on Spectral Clustering". Statistics and Computing 17(4), pp. 1–32 30

[46] Wasserman, S. et K. Faust (1994b) "Social network analysis : Methods and applications" . Cambridge University Press 1994b. 19, 34

[47] Watts D. J. (1999 (réed. 2004)) "Small Worlds : The Dynamics of Networks Between Order and Randomness" . Princeton University Press, Princeton. 13

[48] Watts D. J. (2003)"Six Degrees : The Science of a Connected Age" . Norton , New York 10

[49] In page Web"http ://snap.stanford.edu/data/index.html" . 64

[50] In page Web "http ://graphstream-project.org" 56

[51] In page Web "https ://gephi.org" 56

[52] "La proposition de Tim Berners-Lee " in info.cern.ch. 10