MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE

UNIVERSITE MOULOUD MAMMERI, TIZI-OUZOU

FACULTE DES SCIENCES

DEPARTEMENT DE MATHEMATIQUES

MEMOIRE DE MASTER 2

SPECIALITE: MATHEMATIQUE

OPTION: RECHERCHE OPERATIONNELLE ET AIDE A LA DECISION

Présenté par:

KHELOUFI NADJIA et LESLOUS YAMINA

Sujet:

Outils de graphes pour l'analyse des réseaux sociaux

Devant le jury d'examen composé de:

M. Mohamed Ouanes;	M. de conférences A;	U.M.M.T.O ;	Président
M. Oukacha Brahim;	M. de conférences A;	U.M.M.T.O ;	Rapporteur
M. Kasdi Kamel;	M. assistant A ;	U.M.M.T.O ;	Examinateur
M^{me}. Louadj Kahina;	M. de conférences B ;	U.M.M.T.O ;	Examinatrice

Soutenu le: 11/10/2012

Remerciements

Je tiens à exprimer nos plus vifs remerciements et toute notre gratitude à :

– M. Oukacha Brahim pour avoir accepté de nous proposer le sujet de ce mémoire et qui a bien su nous guider, On tiens à lui exprimer notre plus profond respect pour sa patience et pour tous les efforts qu'il a consentis et tout le temps précieux qu'il a consacré dans la réalisation de ce travail.

– Tous les membres de jury pour avoir accepté d'examiner et de juger ce travail.

Je tiens également à remercier tous ceux qui nous ont encouragé, motivé et aidé tout au long de l'année.

Table des matières

Introduction générale

L'analyse des réseaux sociaux est l'un des domaines des sciences sociales les plus proches des mathématiques en ce sens qu'il utilise des mathématiques et qu'il suscite des travaux de la part de mathématiciens ou de statisticiens.

Les réseaux sociaux sont utilisés principalement pour décrire les interactions entre les entités sociales. Ils sont encore utilisés pour modéliser d'autres types d'interaction comme les liens entre un ensemble de pages Web ou bien les citations bibliographiques dans une corpus de documents. Le cadre mathématique des réseaux est bien approprié pour décrire plusieurs systèmes composés d'un grand nombre d'entités qui interagissent entre elles. Chaque entité est représentée par un nœud du réseau et chaque interaction par un lien entre deux nœuds. Il est donc possible de modéliser ces réseaux par des graphes.

Pour la plupart de ces réseaux, la difficulté provient principalement du grand nombre d'entités, ainsi que de la façon dont elles sont interconnectées. Une approche naturelle pour simplifier de tels systèmes consiste donc à réduire leur taille. Cette simplification n'est pas faite aléatoirement, mais de telle façon à ce que les nœuds de la même composante aient plus de liens entre eux qu'avec les autres composantes. Ces groupes de nœuds ou composantes sont appelés communautés d'intérêt. La connaissance communautés des réseaux nous aide à bien comprendre leurs fonctionnements et comportements, et à appréhender les performances de ces systèmes.

L'étude de structures de communautés a de nombreuses applications dans divers disciplines. En sociologie, les réseaux de connaissances, de travail, ou encore d'amitié, sont caractérisés par des groupes d'individus fortement connectés entre eux, représentant les communautés d'intérêt. En informatique, les réseaux physiques et logiques d'ordinateurs, les graphes du web contiennent des communautés. En biologie, l'importante application est les réseaux de neurones et les réseaux métaboliques.
Le problème traité en vue de l'élaboration de notre projet de mémoire de fin d'étude porte sur l'utilisation d'outils de graphes pour l'analyse des réseaux sociaux .
En résumé, pour présenter notre notre travail, nous l'entamerons par un premier chapitre dont lequel nous donnerons les définitions et la terminologie utilisé par la suite.

Dans le deuxième chapitre, Nous aborderons les réseaux et leur modélisation en graphes; nous donnerons des exemples de réseaux sociaux ainsi que les différentes méthodes de représentation. Nous parlerons aussi de partitionnement de graphes, des méthodes de partitionnement existantes.

Le troisième chapitre chapitre nous présenterons la formalisation du problème. Nous définirons la communauté, les méthodes de détection de communautés seront classées selon leur principe de détection. Nous parlerons de : partitionnement de graphes, d'approches séparatives, d'approches agglomératives, d'approches basées sur les motifs. Une fois les communautés sont formées, nous nous intéressons à l'étude et l'analyse des structures de graphes émergés.

Dans le quatrième chapitre on a abordé réseau social du web qui est parmi les réseaux sociaux les plus étudiés en ce moment.
Dans le dernier chapitre on a essayé de programmer un algorithme de détection de communauté qui optimise la modularité qui l'algorithme basé sur l'optimisation spectrale.

Et nous terminerons par une conclusion oú nous allons mentionner les questions qui vont se poser et serons d'excellents thèmes de recherche dans ce domaine.

Chapitre 1

DEFINITIONS ET CONCEPTS DE BASE

1.1 Introduction

La théorie des graphes est née en 1736 quand Euler démontra qu'il était impossible de traverser chacun des sept ponts de la ville russe de königsberg (aujourd'hui Kaliningrad) une fois exactement et de revenir au point de départ. Les ponts enjambent les bras de la *Pregel* qui coulent de part et d'autre de l'île de $Kneiphof$.
Dans la figure suivante, les nœuds représentent les rives. La théorie des graphes constitue un domaine des mathématiques qui, historiquement, s'est aussi développé au sein de disciplines diverses telles que la chimie (modélisation de structures), la biologie, les sciences sociales (modélisation des relations) ou en vue d'applications industrielles (problème du voyageur de commerce). Elle constitue l'un des instruments les plus courants et les plus efficaces pour résoudre des problèmes discrets posées en Recherche Opérationnelle (R.O.). De manière générale, un graphe permet de représenter simplement la structure, les connexions, les cheminements possibles d'un ensemble complexe comprenant un grand nombre de situations, en exprimant les relations, les dépendances entre ses éléments (réseau de communication, réseaux ferroviaire ou routier, arbre généalogique, diagramme de succession de tâches en gestion de projet, ...). En plus de son existence purement mathématique, le graphe est aussi une structure de données puissante pour l'informatique (les réseaux sociaux de web).

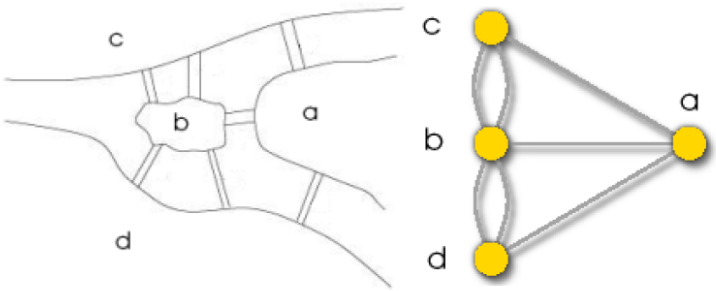


FIG. 1.1 – *Les sept ponts de Königsberg*

1.2 Concepts fondamentaux de la théorie des graphes

1. **Sommet**
   Un sommet est l'unité de base d'un réseau, il en représente une ressource. Dans un

réseau social on parle d'acteur. Le terme nœud est également utilisé pour désigner un sommet.

2. **Arête**
   Une arête est une connexion entre deux sommets. On parle également d'arc ou de lien.

1. **Graphe**
   Un graphe permet de décrire un ensemble d'objet et leurs relation, c.à.d, le lien entre les objets.
   Les objets sont les **sommets** ou encore **nœuds** du graphe.

2. **Graphe non orienté**
   Un graphe non orienté est la donnée d'un couple (X, U), où X est un ensemble fini de sommets et U un ensemble d'arêtes. Ce graphe sera noté $G = (X, U)$ et sera sous entendu comme graphe non orienté.

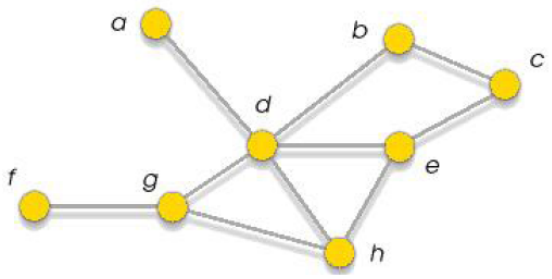


FIG. 1.2 – *Un graphe non orienté*

3. **Graphe orienté**
   $G = (X, E)$ est un graphe dont toute les arêtes sont orientées. Les éléments de E sont appelés arcs de G.

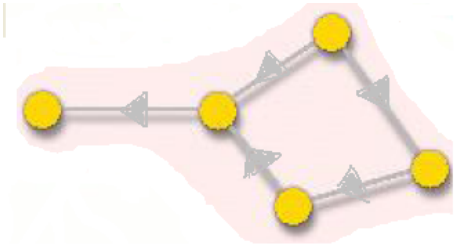


FIG. 1.3 – *Un graphe orienté*

4. **Graphe valué**
   Un graphe valué est un graphe orienté $G = (X, U)$ muni d'une application

$$d : \begin{cases} U \longrightarrow & \mathbb{R} \\ u \longmapsto & d(u), \end{cases}$$

   Où $d(u)$ est le coût, longueur ou distance de l'arc u.
5. **Ordre du graphe**
   On appelle ordre du graphe G=(X,U) le nombre de sommets du graphe.
   L'ordre de G est donc le cardinal de X et noté |X|.
6. **La distance ou l'écart**
   La distance d(x,y) entre deux sommets x et y dans un graphe est la longueur du chemin $x - y$ le plus court qui les relie, si elle existe, sinon la distance est
   d(x,y) $= \infty$.
7. **Le diamètre**
   Le diamètre est la distance maximale entre deux sommets.
8. **Arcs adjacents**
   Deux arcs sont adjacents s'ils ont une extrémité commune. De plus, deux arcs sont dits consécutifs si l'extrémité initiale de l'un est l'extrémité terminale de l'autre.
9. **Successeurs et prédécesseurs**
   Soit x un sommet du graphe G. On dit que le sommet y est un successeur de x s'il existe un arc ayant son extrémité initiale en x et son extrémité terminale en y. L'ensemble des successeurs de x est noté : $\Gamma^+(x)$.
   De même, le sommet y est un prédécesseur de x, s'il existe un arc ayant son extrémité initiale (I) en x et son extrémité terminale (T) en y. L'ensemble des prédécesseurs de x est noté: $\Gamma^-(x)$.

10. **Sommets voisins**
    Pour un arc $e = (x, y)$, l'élément x est l'extrémité initiale (I) de e, y son extrémité terminale (T), les sommets x et y sont dits alors voisins.
11. **Incident**
    Un sommet et une arête sont incident, lorsque le sommet est une extrémité de l'arête.

12. **Sommet (nœud) isolé**
    Un sommet qui n'a aucun voisin est un sommet isolé.
13. **Degré**
    Le degré d'un sommet est le nombre de ses arêtes adjacentes.

14. **Chaîne, chemin**
    Une chaîne de longueur k est une séquence de sommets distincts tels que x_{i+1} est un successeur de x_i pour tout indice i allant de 0 à k. Les sommets x_0 et x_k sont appelés les extrémités de la chaîne.
    Un chemin d'un graphe orienté G est une chaîne où deux arcs consécutifs sont dans le même sens.

15. **Cycle, Circuit**
    Un cycle est une chaîne dont les deux extrémités coïncident.
    Un circuit est un chemin dont les deux extrémités coïncident.

16. **Longueur d'un chemin**
    La longueur d'un chemin C noté $d(C)$ est la somme des longueurs des arcs qui le compose.

17. **Plus court chemin, plus long chemin**
    Un chemin C° joignant x à y est dit de longueur minimale (ou maximale) si il minimise (ou maximise) $d(C)$, pour tous les chemins C joignant x à y.

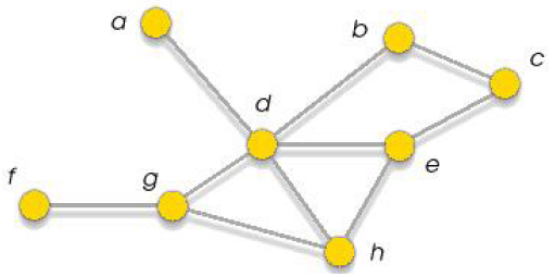
18. **Distribution des degrés d'un graphe**
    La distribution des degrés d'un graphe G est l'association de chaque entier k au nombre de nœuds de G ayant un degré égale à k.

19. **Sous graphe**
    Un sous graphe $G^{'} = (X^{'}, U^{'})$ de $G = (X, U)$ est un graphe dont l'ensemble des sommets est un sous-ensemble de X, et dont l'ensemble des arêtes $U^{'}$ est un sous-ensemble de U tel que toute arête de $U^{'}$ joint deux sommets de $X^{'}$.

20. **Arbre**
    Un arbre est un graphe connexe sans cycle.

21. **Multigraphe**
    Un multigraphe est un graphe qui peut contenir des boucle et plus d'une arête entre ses sommets.

22. **Graphe biparti**
    Un graphe $\mathbf{G} = (X,E)$ est biparti si l'ensemble des sommets peut être partitionné en deux sous-ensembles disjoints



22.

FIG. 1.4 – *Un graphe connexe*

Il peut y avoir plusieurs composantes connexes dans un graphe. Cependant, la composante connexe est le plus grand sous-graphe que l'on peut obtenir. Par conséquent, la composante connexe d'un graphe connexe est ce même graphe. Voici en rouge, bleu, vert les trois composantes connexes du graphe :

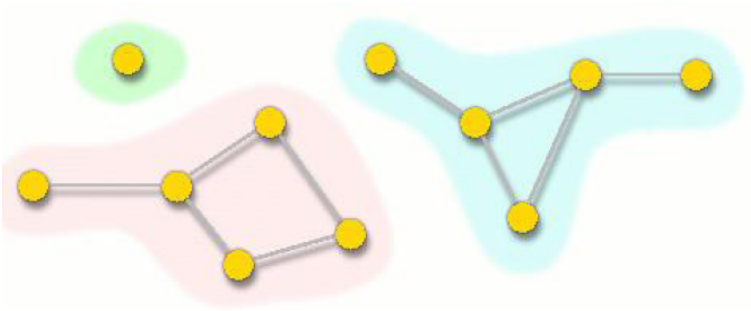


FIG. 1.5 – *Composantes connexes*

k-connexité

Un graphe est dit k-connexe si k est le nombre minimum de nœuds à retirer pour que le graphe obtenu soit non-connexe.

Définition 1.1. Matrice d'adjacence

Considérons un graphe G =(X, E) comportant n sommets. La matrice d'adjacence de G est égale à la matrice A = (a_{ij}) de dimension n ×n telle que:

$$a_{ij} = \begin{cases} 1 & \text{si (i, j)} \in \text{E;} \\ 0 & \text{sinon.} \end{cases}$$

Un graphe orienté quelconque a une matrice d'adjacence quelconque, alors qu'un graphe non orienté possède une matrice d'adjacence symétrique. La matrice d'adjacence possède quelques propriétés qui peuvent être exploitées. Considérons un graphe G et sa matrice d'adjacence associée A :

- la somme des éléments de la $i^{ème}$ ligne de A est égale au degré sortant $d^-(x_i)$ du sommet x_i de G.
- la somme des éléments de la $j^{ème}$ colonne de A est égale au degré entrant $d^+(x_j)$ du sommet x_j de G.

Définition 1.2. Matrice d'incidence

Considérons un graphe orienté sans boucle G = (X, E) comportant n sommets $x_1 \ldots, x_n$ et m arêtes $e_1, \ldots, e_m$. On appelle matrice d'incidence (aux arcs) de G la matrice M = (m_{ij}) de dimension n×m telle que :

$$m_{ij} = \begin{cases} 1, & \text{si } x_i \text{ estl'extremité initiale de } e_j; \\ -1, & \text{si } x_i \text{ estl'extremité terminale de } e_j; \\ 0, & \text{si } x_i \text{ n'est pas une extremité de } e_j. \end{cases}$$

Pour un graphe non orienté sans boucle, la matrice d'incidence (aux arêtes) est définie par:

$$\begin{cases} 1, & \text{si } x_i \text{ est une extremité de } e_j; \\ 0, & \text{sinon.} \end{cases}$$

Définition 1.3. La matrice diagonale
La matrice diagonale d'un graphe G est une matrice où tous les éléments, à part la diagonale, sont nuls. Les éléments de la diagonale sont le nombre de voisins de chaque nœud du graphe G.

1.3 Domaines d'application de la théorie des graphes

Les graphes (et par conséquent la théorie des graphes) sont utilisés dans de nombreux domaines. On peut donner quelques exemples :

- Les réseaux de communication : réseaux de routes représentés par une carte routière, réseaux de chemin de fer, de téléphone, de relais de télévision, réseaux électriques, réseaux des informations dans une organisation, etc...
- La gestion de la production : graphes potentiels-étapes plus connu sous le nom de graphes $PERT$ ["Programme *Evaluation and Research Task*" ou "Programme *Evaluation Review* Technique"]
- L'étude des circuits électriques : $Kirchhof$, qui a étudié les réseaux électriques, peut être considéré comme un des précurseurs de cette théorie ;
- la chimie, la sociologie et l'économie .

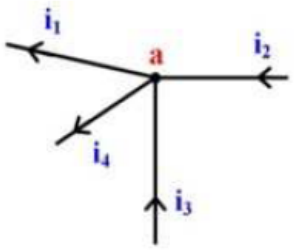


FIG. 1.6 – *Illustration de la loi des noeuds de Kirchhoff* ($i_1 + i_4 = i_2 + i_3$).

1.4 Définitions des notions utilisées dans les réseaux sociaux

Définition 1.4. Densité
La densité pour un graphe est la proportion d'arêtes dans le graphe. Si n est la taille du graphe G et m le nombre de liens qui existent entre ses nœuds, alors :

$$Densite(G) = 2m/(n(n-1)) \tag{1.1}$$

Définition 1.5. Distance moyenne
Moyenne des plus courts chemins pour toutes les paires de nœuds du graphe.

Définition 1.6. Centralité sommets/arétes
Somme des relations dans laquelle un acteur est engagé. Moins un acteur est central, plus il est dépendant d'un ou quelques membres pour établir des relations au sein du réseau, c.à.d. la capacité de chaque membre d'établir des relations avec les autres parties. Indépendance de l'acteur du fait de la multiplicité des relations qu'il entretient.

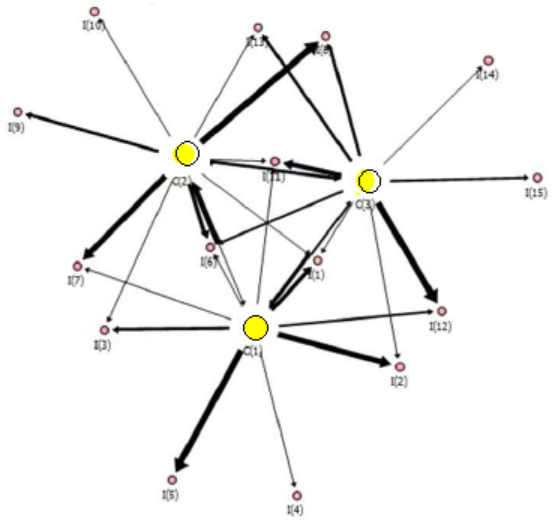


FIG. 1.7 – *Exemple de centralité d'intermédiarité de sommet [Freeman* 1977*].*

Définition 1.7. Centralité de proximité
Nombre d'individus par lequel l'acteur doit passer pour entrer en contact avec les autres acteurs du système.

Définition 1.8. Trou structural
Se réfère à l'absence de lien entre deux nœuds. Les trous structuraux peuvent être stratégiquement remplis en connectant un ou plusieurs nœuds ensembles. Ils peuvent aussi être entretenus afin de maximiser la rentabilité d'un réseau. Selon le concept de capital social; si un nœud relie deux autres nœuds ensembles sans qu'ils ne soient autrement liés entre eux, il est possible pour ce nœud de contrôler leurs communications.

Définition 1.9. Sommets influents/pouvoir/préstige
Degré d'importance d'un nœud (individu) dans un réseau social.

Définition 1.10. Coefficient de clustering
– Le coefficient de *clustering* d'un nœud est la probabilité que deux voisins d'un nœud soient aussi voisins entre eux.
– Le coefficient de *clustering* d'un graphe est la moyenne des coefficients de clustering de tous ses nœuds.

Définition 1.11. Cohésion dans un groupe
Dans le contexte des graphes, on qualifie un groupe de cohésif s'il existe plus de liens entre ses membres que vers l'extérieur. Des auteurs suggèrent que la k-connexité puisse être une estimation de cette cohésion structurelle du réseau social, car elle mesure à quel point un groupe dépend de ses constituants pour conserver son caractère de groupe. La connexité simple peut être insuffisante dans ce but. En effet, considérons deux types de groupes: un oú l'autorité est centrée sur un leader et l'autre oú elle est dispersée entre les membres; l'un et l'autre type sont connexes et de même densité, mais alors que la suppression du nœud central déconnecte la structure du premier, il n'y a pas d'acteur jouant un rôle équivalent dans l'autre structure par exemple un cycle.

Définition 1.12. Distribution des degrés en loi de puissance
La distribution des degrés d'un réseau d'interactions suit une loi de puissance de type : f(k) $\sim k^{-\gamma}$; où, k désigne le degré, γ est l'exposant de la loi de puissance. En pratique, la valeur de γ est comprise entre 2 et 3. Une distribution des degrés en loi de puissance signifie qu'il existe beaucoup de sommets de faibles degrés et très peu de sommets de forts degrés. γ représente la vitesse de décroissance de la courbe des degrés. Plus γ est grand et plus la probabilité d'obtenir des sommets de forts degrés est petite.

Définition 1.13. Un algorithme de clustering hiérarchique
L'algorithme part d'une structure dans laquelle chaque sommet est identifié comme une petite communauté. On itére alors les étapes suivantes : on calcule des distances entre communautés et on fusionne les deux les plus proches en une nouvelle communauté. Le nombre est réduit d'un à chaque étape, et le processus s'arrête lorsqu'il n'y a plus qu'une seule communauté correspondant au graphe entier. On obtient ainsi une structure hiérarchique de communautés qui peut être représentée sous une forme arborescente appelée dendrogramme : les feuilles sont les sommets du graphe tandis que les nœuds représentent les communautés crées et sont reliés en fonction des fusions de ces dernières. La racine de la structure correspond au graphe entier. Il existe plusieurs façons de définir la distance entre deux communautés. La plus simple (single linkage) considère que la distance entre deux communautés est la distance minimale entre deux sommets de celles-ci. A l'opposé, on peut considérer la distance maximale (complete linkage).

Définition 1.14. Les six degrés de séparation
Les six degrés de séparation est une théorie établie par le hongrois Frigyes Karinthy en 1929 qui évoque la possibilité que toute personne sur le globe peut être reliée à n'importe quelle autre, au travers d'une chaîne de relations individuelles comprenant au plus cinq autres maillons.
Cette théorie est reprise en 1967 par Stanley Milgram à travers l'étude du petit monde.

Définition 1.15. L'effet du petit monde
L'effet du petit monde également connu sous le nom « paradoxe de Milgram »(car ses résultats semblent contraires à l'intuition) est l'hypothèse que chacun puisse être relié à n'importe quel autre individu par une courte chaîne de relations sociales.

Définition 1.16. L'expérience de Milgram (le petit monde)
Elle trouve leur origine dans une expérimentation menée aux États-Unis par Stanley Mil-

gram et ses collègues. En résumé, à des personnes tirées au hasard, on demande d'acheminer un paquet par la poste à un même individu cible qu'elles ne connaissent pas. On leur donne quelques renseignements généraux sur cet individu : le collège où il a fait ses études, son lieu de résidence, sa profession, etc. Bien sûr, on ne leur procure ni son nom ni son adresse. Puisqu'elles ne peuvent pas déterminer de qui il s'agit, on leur demande d'envoyer le paquet à une personne de leur réseau dont elles pensent qu'elle sera la plus susceptible de l'acheminer vers l'individu cible. Par exemple, cet individu étant agent de change, si elles connaissent un agent de change, elles lui enverront le paquet. À nouveau, la personne qui le reçoit est chargée de la même mission, et ce jusqu'à ce que le paquet finisse par parvenir à quelqu'un qui se trouve être en mesure d'atteindre le destinataire final. En moyenne, il a fallu 5,2 intermédiaires pour réaliser cet objectif. C'est finalement fort peu à l'échelle d'une société comme les États-Unis.

1.5 Notions sur la complexité

En général, on mesure l'efficacité d'un algorithme par une expression mathématique C(n). Celle ci exprime le nombre d'opérations élémentaires indispensables à son exécution en fonction de la taille des données en entrée, tout en considérant le pire des cas. C'est le nombre maximum d'étapes de calculs nécessaires en fonction de n pour aboutir à une solution optimale. Cette expression mathématique s'appelle (complexité de l'algorithme). Si cette complexité est C(n), on dit que cet algorithme est en O(C(n)).
Un algorithme est dit polynômial ou efficace si le nombre d'opérations nécessaires pour résoudre un problème avec celui ci est borné par une fonction polynômiale d'un paramètre caractérisant la taille du problème.

1. **La classe NP**
   Un problème appartient à la classe NP si on peut déterminer sa solution en un temps polynômial. On dit qu'elle regroupe les problèmes faciles (la classe P) et les problèmes difficiles (la classe NP-Complet et les problèmes ouverts).
2. **La classe P**
   Elle regroupe les problèmes les plus faciles de la classe NP. Ce sont les problémes que l'on peut résoudre en temps polynômial, c'est-à-dire qu'il existe un algorithme pour résoudre le problème dont le temps d'exécution est de la forme O(n^k) pour toutes les entrées, où k est une constante fixée et n est la taille de l'entrée.
3. **La classe NP-Complet**
   Elle regroupe les problèmes les plus difficiles de la classe NP. car bien qu'ils travaillent depuis des années sur les problèmes NP-complets, aucun n'a trouvé d'algorithme polynômial pour les résoudre. Si l'on est capable de résoudre un seul problème NP-complet en temps polynômial, alors par définition des réductions polynômiaux dans NP, on pourra résoudre tous les problèmes de la classe NP en temps polynômial, et donc P = NP.

Conclusion

La branche de la sociologie qui étudie les réseaux sociaux individuels ou collectifs a utilisé la théorie des graphes pour l'analyse de ces réseaux pour plusieurs raisons : premièrement, elle fournit un vocabulaire qui peut être employé pour désigner beaucoup de propriétés structurelle des relations sociales; deuxièmement, elle offre des formulations mathématiques et des idées avec lesquels beaucoup de propriétés relationnelles peuvent être quantitativement évaluées et mesurées et dernièrement, elle permet de présenter les réseaux sociaux comme des modèles de jeux et de relations.

Chapitre 2

RESEAUX SOCIAUX

2.1 Introduction

Un réseau social est un ensemble de relations entre un ensemble d'acteurs. Cet ensemble peut être organisé (une entreprise, par exemple) ou non (comme un réseau d'amis) et ces relations peuvent être de nature fort diverse (pouvoir, échanges de cadeaux, conseil, etc.), spécialisées ou non, symétriques ou non. Les acteurs sont le plus souvent des individus, mais il peut aussi s'agir de ménages, d'associations, etc. L'essentiel est que l'objet d'étude soit bien la relation entre éléments, autrement dit l'interaction ou l'action réciproque entre ces éléments. Des recherches pionnières ont été menées sur ces questions tant par des sociologues, comme Georg Simmel ou Jacob Moreno (1934), que par des ethnologues comme Radcliffe-Brown, Firth, Barnes (1954) ou Bott (1971). Elles sont à l'origine de l'important développement de l'analyse des réseaux sociaux auquel on assiste depuis le début des années 1970. Selon cette perspective, les réseaux ne sont pas un mode d'organisation sociale particulier et leur analyse n'est pas une fin en soi. L'étude des graphes des relations n'est pas davantage conçue comme un simple outil technique venant s'ajouter à la panoplie déjà bien fournie du sociologue. L'analyse de réseaux sociaux est au contraire ici le moyen d'élucider des structures sociales et de s'interroger sur leurs rôles.

2.1.1 Quelques exemples de réseaux d'interactions sociales

Les réseaux d'acteurs

Le graphe des acteurs est un graphe dont les sommets sont des acteurs et deux acteurs sont reliés s'ils ont joué ensemble dans un film. En 2000, la taille du graphe d'acteurs étudié était de 449913 sommets.

Les réseaux de connaissances

Le graphe de connaissance est un graphe dont les sommets sont des personnes et deux personnes sont liées si elles se connaissent. Bien entendu, une telle relation ne peut être définie de façon très formelle et le graphe ne peut être totalement construit.

Les réseaux d'appels téléphoniques

Le graphe des appels téléphoniques est un graphe orienté dont les sommets représentent des numéros de téléphones et un arc signale qu'un numéro a au moins une fois appelé un autre.

2.2 Les caractéristiques d'un réseau social

La principale caractéristique est l'effet du petit monde issu de la célèbre expérience de [Milgram 1967], Ainsi toute personne dans un réseau social est connectée à toute autre personne par un chemin de courte distance. Le plus court chemin entre deux sommets dans

un réseau social de taille n est de l'ordre de log(n). Ainsi lorsque la taille du réseau augmente, la longueur des plus courts chemins n'augmente que très peu. De plus les membres de ce réseau possèdent la faculté de trouver facilement ces plus courts chemins. Une autre caractéristique est issue de la tendance de l'homme à se socialiser en groupe ce qui donne aux réseaux sociaux une forte tendance au clustering et une structure en communautés. Si un sommet A est connecté à un sommet B et que ce sommet B est connecté à un sommet C, alors A et C ont une forte probabilité d'être également connectés, on parle aussi de transitivité. La troisième caractéristique est la distribution des degrés, par exemple si cette distribution suit une loi de puissance, à savoir que plus on considère un degré élevé, plus le nombre de sommets qui ont ce degré dans un même réseau est faible. Si on prend par exemple le célèbre réseau d'amis du club de karaté de Zachary en 1977, représenté par un graphe non orienté, non pondéré et non étiqueté(Figure 2.1). Ce club a été scindé en deux clubs, les membres du premier sont représentés par des sommets ronds et blancs, les membres du deuxième sont représentés par des sommets carrés et grisés.

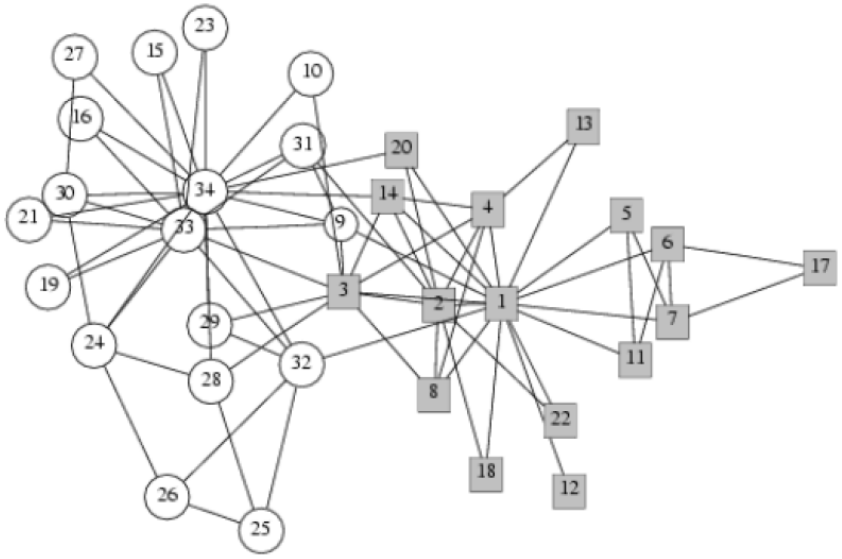


FIG. 2.1 – *Le club de karaté de Zachary s'est divisé en deux clubs, les membres du premier club sont représentés par des ronds blancs et les membres du second par des carrés grisés.*

La figure suivante montre la répartition des degrés dans le réseau social du club de karaté du club de Zachary.

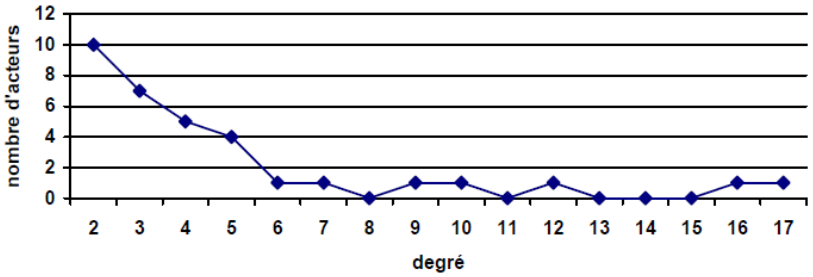


FIG. 2.2 – *répartition des degrés du club de karaté de Zachary.*

2.3 L'analyse des réseaux sociaux

l'analyse des réseaux consiste à examiner la structure et la configuration de ces relations et de chercher à identifier leurs causes et leurs conséquences. L'analyse des réseaux existe depuis longtemps. On peut citer par exemple Durkheim qui, en 1897, disait que les suicides des individus apparaissaient lorsqu'ils étaient dépourvus de liens sociaux qui les empêchaient de commettre le suicide, mais La première personne à avoir représenté un réseau social est Jacob Levy Moreno au début des années 1930. Son objectif étant de visualiser graphiquement un réseau social, il a représenté les personnes par des points et une relation entre deux personnes par des flèches. Cette représentation est depuis désignée par le terme sociogramme, mais on parlait également de toiles en raison de leur aspect en toile d'araignée. Cette forme de visualisation, aussi peu innovante qu'elle puisse paraître de nos jours, fut un premier outil d'identification rapide des caractéristiques d'un réseau social. Moreno a ainsi introduit le concept d'étoile pour désigner les personnes ayant le plus de relations dans un réseau social, en référence à l'étoile formée par un point et ses connections. Les mathématiciens ont rapidement fait le rapprochement entre les représentations sociogrammes et la théorie des graphes au sens mathématique.
Nous avons ainsi tous les concepts sociologiques pour définir un réseau social : les individus, leurs liens (contacts), leurs affinités et l'environnement les entourant. D'un point de vue technologique, le réseau définit un ensemble d'équipements interconnectés qui servent à acheminer un flux d'informations. Il existe le réseau informatique que l'on retrouve dans les entreprises par exemple, le réseau téléphonique ou encore le réseau des réseaux : Internet.
La figure suivante présente un sociogramme utilisé par Moreno pour représenter la position d'un leader isolé (individu A). L'individu A est susceptible d'influencer 50 personnes, virtuellement prédisposées en sa faveur via l'individu B.

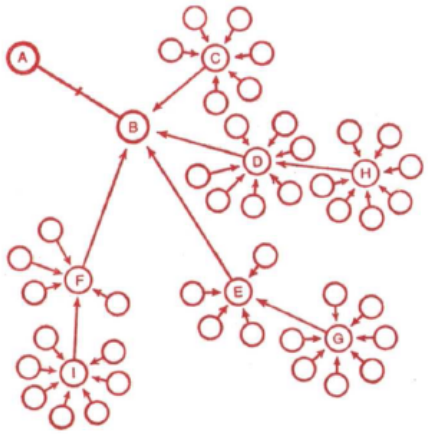


FIG. 2.3 – *Sociogramme de leader isolé [Moreno 1934].*

2.4 Les précautions à prendre pour une meilleur analyse

L'analyse des réseaux sociaux a adopté dans ses techniques de représentation un certain nombre de précautions qui correspondent à des questionnements d'autant plus importants qu'ils touchent à la validité même de la méthode. Nous citerons :

2.4.1 Définir la frontière du réseau à analyser

La conception des frontières du réseau soulève un problème méthodologique important. En effet, la délimitation de la frontière d'un réseau a une influence directe sur la validité externe des résultats. Elle nécessite que le chercheur ait une vision claire de l'objet à étudier. Le terme de réseaux complets utilisés en analyse des réseaux sociaux ne correspond à aucune réalité. Les réseaux sont dits complets par convention. Il faut, en effet, délimiter l'objet de recherche et étudier un réseau fini. Ce dernier doit être constitué par des individus entretenant entre eux des relations plus denses qu'avec l'extérieur. Néanmoins, un réseau n'est jamais « fini » en soi et de ce fait ne peut être identifié à un acteur collectif. Il s'imbrique à d'autres réseaux, s'agrège ou disparaît au sein de méta-réseaux. Ainsi, l'analyse structurale ne peut uniquement fonctionner sur la description d'un objet sans contour. Elle a besoin au préalable de se reposer sur un contexte d'étude et de s'interroger sur les acteurs individuels (profils, types de ressources échangées$\cdots$).

2.4.2 L'échantillonnage

Tous les individus n'ont pas le même statut. Certains sont plus centraux que d'autres. D'autres sont des individus ponts qui relient des sous-groupes entre eux. Le fait de ne pas sélectionner ce type d'individus peut fausser radicalement la perception. Donc il faut faire attention a choisir un échantillon représentatif.

2.5 La représentation d'un réseau social

On peut représenter un réseau social par deux façons. Les premières visualisations de réseaux sociaux ont été introduites par Moreno, il s'agissait alors de représentations nœuds-liens. Les matrices d'adjacence ont fait leur apparition par la suite.

1. **Représentation nœuds liens**
   Dans cette représentation, un nœud représente un acteur et un lien représente une relation du réseau. Il s'agit de la représentation la plus courante des reseaux. Le trés grand avantage des diagrammes nœud-lien est leur intuitivitè (la grande majorité des lecteurs peuvent les comprendre). En revanche, qu'ils soient dessinés manuellement (figure 2.4) ou générés automatiquement (figure 2.5), leur lisibilité dépend totalement du placement des nœuds dans le plan.

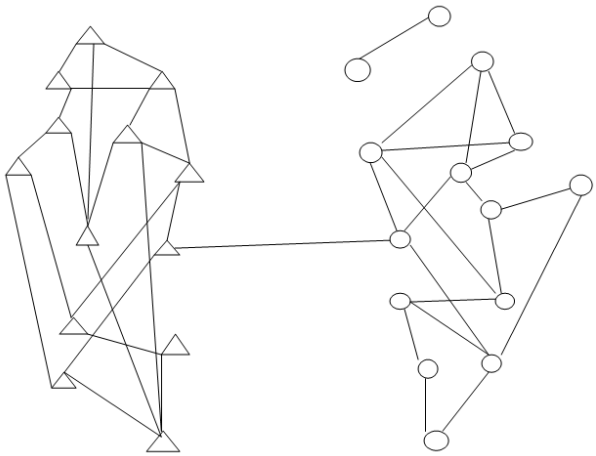


FIG. 2.4 – *Réseau d'amitiés entre garçons (triangles) et filles (cercles)*

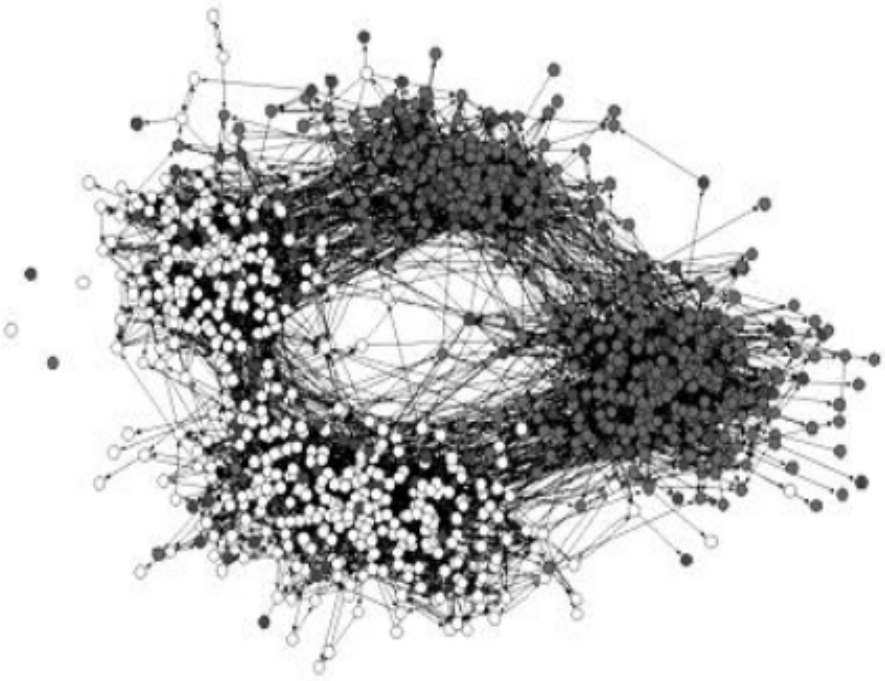
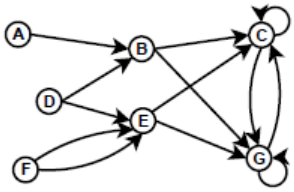


FIG. 2.5 – *Réseau d'amitiés entre lycéens*

2. **Représentation matricielle**
   Une matrice d'adjacence représente chaque sommet d'un réseau à la fois comme une ligne et comme une colonne. Si deux sommets sont connectés, la case correspondant a l'intersection de la ligne et de la colonne est marquée. Traditionnellement, on utilise une valeur numérique (0 marquant l'absence de connexion, 1 marquant la presence).

Et voici un exemple représenté par les deux méthodes:



	A	B	C	D	E	F	G	∑
A	0	1	0	0	0	0	0	1
B	0	0	1	0	0	0	1	2
C	0	0	1	0	0	0	1	2
D	0	1	0	0	1	0	0	2
E	0	0	1	0	0	0	1	2
F	0	0	0	0	2	0	0	2
G	0	0	1	0	0	0	1	2
∑	0	2	4	0	3	0	4	13

FIG. 2.6 – *Un exemple représenté par les deux méthodes*

Le tableau suivant liste les principaux avantages et inconvénients des représentations matricielles par rapport aux diagrammes nœud-lien.

pour	contre
Absence de superpositions des Nœuds (ce qui permet de pouvoir toujours lire les étiquettes porté par les nœuds).	Taille de l'espace visuel requis à un niveau de détail équivalent plus important que le diagramme nœud –lien.
Absence de croisement des liens (ce qui permet de toujours d'identifier la source et destination des connexions).	Difficulté à effectuer des tâches de suivi de chemins.
Facilité avec laquelle il est possible d'identifier l'absence de connexions.	Manque de familiarité, les matrices paraissent moins intuitives que les diagrammes

FIG. 2.7 – *Avantages et inconvénients des représentations matricielles par rapport aux diagrammes nœud-lien.*

2.6 Modélisation des réseaux

Un réseau peut être modélisé par un graphe G = (X,E), où X est l'ensemble des nœuds du graphe G (représentant les individus du réseau), et E l'ensemble des arêtes entre les nœuds du graphe G (représentant les liens entre les individus du réseau). Pour une meilleure analyse et compréhension du réseau, ce dernier est généralement décomposé en sous-groupes de nœuds. La décomposition n'est pas faite aléatoirement, mais de telle façon à ce que les nœuds de la même composante aient plus de liens entre eux qu'avec les autres composantes. Ces groupes de nœuds (composantes) sont appelés communautés d'intérêt. Une communauté peut être définie, d'une manière structurelle, comme un ensemble de nœuds qui ont plus de liens entre eux qu'avec les autres groupes, et d'une manière sémantique, comme un ensemble de nœuds partageant les mêmes centres d'intérêt. Dans ce qui suit, nous allons essayer de donner quelques présentations de différents modèles de réseaux d'interactions.

1. **Modèle aléatoire d'Erdös et Rényi**
   Entre 1950 et 1960, P. Erdös et A. Rényi (ER) ont étudié et proposé les premiers modèles de réseaux d'interactions appelés les graphes aléatoires. Le principe de ce

modèle consiste à générer un graphe à n sommets et m arêtes. Les arêtes entre les sommets sont choisies d'une manière aléatoire, noté $G_{n,p}$ où n est la taille du graphe et p la probabilité avec laquelle le lien entre deux sommets est établi. Ils sont des graphes statiques : le nombre de sommets est fixe et ne change plus au cours du temps. Or, il existe des réseaux d'interactions tel que le Web, où le nombre de pages Web et le nombre de liens augmentent chaque jour. En d'autres termes, la modélisation d'Erdös et Rényi ne prend pas en compte la dynamique des réseaux.

Propriétés des graphes aléatoires

Le coefficient de clustering d'un graphe aléatoire est donné par le nombre de liens entre les voisins d'un nœuds divisé par le nombre maximal des liens possibles. Du moment que la probabilité p est uniforme pour tout le réseau, alors le coefficient de clustering moyen C est simplement donné par p. Étant donné que tous les nœuds dans un graphe aléatoire ont pratiquement un degré proche du degré moyen du graphe, alors le coefficient de clustering décroît avec la taille du graphe n.

Par ailleurs, dans ce type de réseaux, la distribution des degrés des nœuds suivent une loi de Poissson, qui indique que la plupart des nœuds ont approximativement le même nombre de liens.

2. **Modèle petits mondes ("Small world")**

   Un graphe petit monde (small world) est caractérisé par une distance moyenne faible et un fort coefficient de clustering. Il existe plusieurs modèles générant des graphes avec une distance moyenne faible et d'autres générant des graphes avec un fort coefficient de clustering, mais il n'existe que très peu de modèles regroupant les deux propriétés. Watts et Strogatz proposent une méthode pour générer des graphes petits mondes. Partant d'un anneau régulier à n sommets où chaque sommet est relié à ses 2k plus proches voisins (k voisins de chaque côté). Le coefficient de clustering d'un sommet x de l'anneau régulier est assez important : $\mathrm{C(x)} = \frac{3(k-2)}{4(k-1)}$ et la distance moyenne, dans un anneau régulier, est elle aussi très élevée. L'idée de Watts et Strogatz est de modifier suffisamment l'anneau régulier en déplaçant les arêtes afin de diminuer la distance entre les sommets. Concrètement, pour chaque sommet et pour chaque arête, selon une probabilité p, l'arête est redirigée vers un autre sommet choisi d'une manière aléatoire et uniforme. Avec une probabilité $1 - p$, l'arête est gardée.

3. **Modélisation des graphes sans echelle**

   Les travaux de Watts et Strogatz ont attiré l'attention sur les graphes de terrains. L'appellation graphe de terrain, proposée initialement par Bruno Gaume, n'est pas universellement acceptée en France et l'on peut aussi rencontrer par exemple les termes réseau d'interactions ou réseau complexe. Ces appellations font référence aux mêmes objets qui possèdent la dénomination anglo-saxonne bien reconnue de complex network. Ils se caractérisent par le fait que des interactions simples et facilement compréhensibles entre sommets à l'échelle microscopique produisent des propriétés et des comportements macroscopiques difficilement interprétables. Du point de vue de l'algorithmique ou de la théorie des graphes, l'originalité du sujet tient au fait que les graphes, outre leur grandes tailles, sont obtenus sur la base de données réelles. On

a cherché a mieux les caractériser encore. Babarabasi et al. (1999) ont ainsi montré qu'ils font partie d'une autre classe très intéressante de graphes, les graphes sans échelle. Cela signifie que la répartition des degrés des sommets suit une loi de puissance : la probabilité P(k) qu'un sommet du graphe considèré aie k voisins décroît en suivant une loi de puissance P(k)= $k^{-\lambda}$, o$\grave{u}$ λ est une constante caractéristique du graphe, alors que dans le cas des graphes aléatoires, c'est une loi de Poisson qui est à l'œuvre. La structure sans échelle se traduit donc par la présence d'un très grand nombre de sommets de faible degré et d'un nombre faible mais non négligeable de sommets de très haut degré. Ceci donne aux graphes sans échelle une structure qui peut être vue comme hiérarchique : localement, des sommets de très haut degré sont reliés a des sommets de moins haut degré, eux-mêmes reliés a des sommets de degré encore moindre, et ainsi de suite jusqu'a la masse des sommets de très faible degré.

2.7 Conclusion

Les réseaux sociaux et leurs analyses sont devenus un véritable challenge dans différents domaines de recherche. En effet, plusieurs recherches se sont intéressées à l'étude des types de relations entre les entités et leur impact sur le réseau considéré. Les graphes représentent les propriétés topologiques essentielles en traitant le réseau comme une collection des noeuds et des arêtes. Pour la plupart de ces réseaux, la complexité provient principalement du grand nombre d'entités ainsi que de la façon avec laquelle elles sont interconnectées. La détection de communautés dans un réseau est un moyen couramment utilisé pour simplifier ces réseaux étant donné que le nombre de clusters est en général bien plus petit que le nombre de nœuds dans le réseau.
Dans le chapitre qui va suivre on va essayer d'introduire et de définir ce moyen.

Chapitre 3

LES COMMUNAUTES D'INTERET

3.1 Introduction

Dans la littérature, on ne trouve pas une définition bien claire des communautés. La plupart des chercheurs dans ce domaine se restreignent à donner des définitions plutôt structurelle et sémantique. Par contre, aucune définition formelle, en accord avec l'ensemble des chercheurs, n'est donnée à ce jour. La notion de communautés est généralement décrite comme étant des sous structures du graphe avec des liens denses entre les membres de la communauté et peu de liens avec les autres communautés comme illustré par la figure suivante :

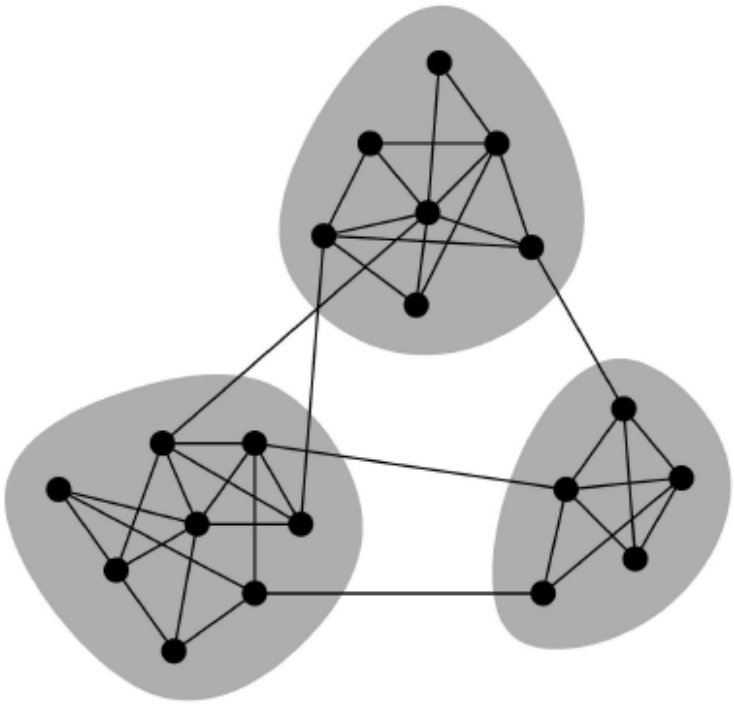


FIG. 3.1 – *Exemple de structures de communautés dans un graphe.*

3.2 L'intérêt de détection des communautés

Notons que les communautés peuvent avoir des interprétations différentes suivant le type de réseau considéré. Ainsi dans les réseaux sociaux elles correspondent à des ensembles d'individus possédant des points communs et dont les liens sociaux sont naturellement plus forts. De maniére totalement différente, dans les réseaux métaboliques les communautés correspondent à des fonctions biologiques de la cellule. Pour les réseaux d'information, elles peuvent correspondre à des thématiques. Par exemple les pages Web traitant d'un même sujet se référencent mutuellement et la détection de communautés dans le graphe du Web

est une piste envisagée pour améliorer les moteurs de recherche. De maniére générale, la détection de communautés est un outil important pour la compréhension des structures et des fonctionnements des grands graphes. En effet, il est souvent impossible d'appréhender la structure d'un graphe en ne connaissant que ses propriétés locales, c'est à dire des propriétés à l'échelle des sommets et de leurs voisins directs. Les communautés permettent de donner un point de vue macroscopique sur la structure des graphes. En effet la grande taille des graphes considérés est une limite majeure en terme de complexité des algorithmes utilisables. Dans certain cas, utiliser les communautés pour diviser le graphe permet d'effectuer des calculs séparés moins coûteux sur chaque communauté. Ce procédé permet d'envisager des gains de complexité pour les algorithmes. La détection de communautés peut aussi être utilisée pour la visualisation des grands graphes de terrain, on peut par exemple envisager des visualisations multi-échelles des communautés permettant d'apprécier les structures du graphe à différentes échelles.

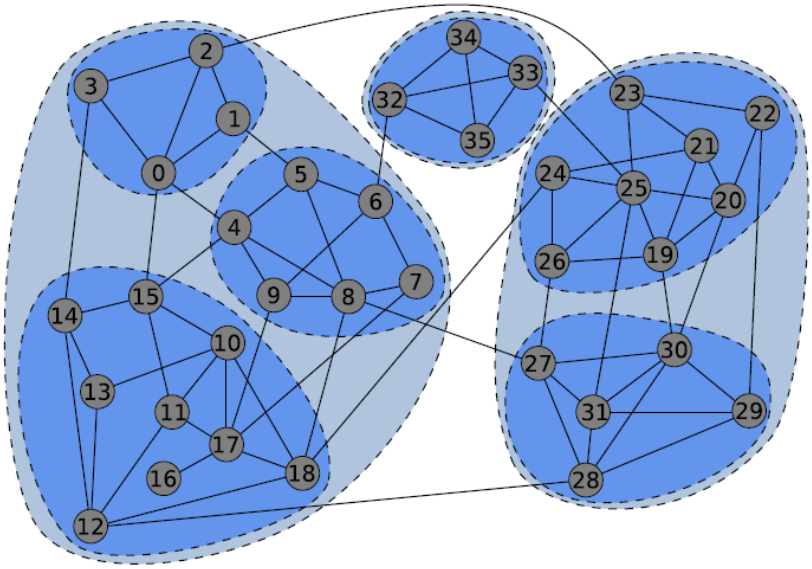


FIG. 3.2 – *Exemple de structures de communautés dans un graphe : deux partitions en communautés correspondant à deux échelles différentes.*

3.3 Fonction de qualité

Pour un même réseau, les approches de détection de communautés peuvent donner des partitionnements en communautés différents. Pour mesurer la qualité d'un tel ou tel partitionnement, des fonctions de qualité ont été définies. Une fonction de qualité est une mesure qui permet de donner un "score" à un partitionnement. Dans le problème de détection de communautés, la fonction de qualité la plus connue et la plus utilisée est la

modularité de Newman.
La modularité de Newman est basée sur le principe suivant: Pour un graphe G = (X,E) et une partition P = (C_1,C_2,...,C_k) des sommets de X, la modularité est formellement définie par :

$$Q(P) = (1/2w)\sum_{i=1}^{N}\sum_{j=1}^{N}(w_{ij} - ((w_i^{out}w_j^{in}))/2w)\delta(C_i,C_j) \tag{3.1}$$

Où, w_{ij} est un élément de la matrice d'adjacence du graphe G, $w_i^{out} = \sum_j wij$ est la somme des poids des liens sortants du nœud i et $w_j^{in} = \sum_i w_{ij}$ la somme des poids des liens entrants au nœud j, $2w = \sum_{ij} w_{ij}$ est la somme totale des poids des liens du graphe G. C_i est l'indice de la communauté à laquelle le nœud i appartient, et $\delta(x,y)$ le symbole de Kronecker qui est égal à 1 si $x = y$, 0 sinon. Pour simplifier, nous donnons la formule de modularité pour les graphes simples (non orientés et non pondérés). Elle est donnée par :

$$Q(P) = \sum_{c\in p}(e_c - a_c^2) \tag{3.2}$$

Où e_c représente la proportion de liens à l'intérieur de C et a_c la proportion de liens connectés à C, par rapport au nombre total de liens.
Cette valeur est toujours inférieure à 1 et permet de comparer les qualités de deux partitions pour un même graphe et, en conséquence, permet de comparer les performances d'algorithmes de détection de communautés. Si cette mesure de qualité n'est pas la seule existante et a plusieurs lacunes, elle reste néanmoins la plus fiable et la plus utilisée. L'inconvénient majeur de cette fonction est que son optimisation est un problème NP-complet et nécessite donc d'utiliser des méthodes approchées dés lors que l'on souhaite l'utiliser sur des graphes de grande taille.

3.4 Méthodes de détection des communautés

La plupart des méthodes de détection de communautés se basent sur l'intuition qu'une structure communautaire est par nature hiérarchique, c'est-à-dire qu'une communauté est elle même composée de sous communautés qui sont à leur tour composées de sous-sous-communautés. Nous allons citer ici les principales méthodes qui ont été proposées à ce jour. D'une manière générale, nous pouvons classer toutes les méthodes, approches et algorithmes en quatre grandes familles d'approches.
Il s'agit de :

1. **Approches par partitionnement de graphes**
   Le but du partitionnement de graphe est de grouper les sommets d'un graphe en un nombre prédéterminé de parties (et de tailles elles aussi prédéterminées) tout en

minimisant le nombre d'arêtes tombant entre les différents groupes. Cette approche ne convient pas totalement à la détection de communautés car elle a l'inconvénient de requérir une connaissance préalable du nombre de communautés recherchées ainsi que de leurs tailles. Nous citons les deux méthodes générales ayant eu le plus de succès.

(a) **Algorithme de la bissection spectrale**
Le principe de la méthode consiste à calculer le vecteur propre correspondant à la plus petite valeur propre non nulle de la matrice Laplacienne du graphe. La matrice Laplacienne L d'un graphe G est calculée comme suit $L = D - A$, où A est la matrice d'adjacence du graphe G et D la matrice diagonale. Suivant les signes des nœuds dans le vecteur propre, les nœuds du graphe sont séparés en deux groupes (les nœuds de signe + dans un groupe, et ceux ayant un signe − dans un autre). La complexité de cette méthode est de l'ordre de $O(n^3)$. et on obtient avec cet algorithme de bons résultats lorsque le graphe possède effectivement deux grandes communautés de tailles similaires, ce qui n'est qu'un cas particulier dans l'optique de détection de communautés.

(b) **La méthode de Kernighan et Lin**
L'idée principale de la méthode KL est de diviser le graphe en un nombre g connu de groupes, tout en minimisant le nombre des arêtes entre les groupes. Au début, le graphe est divisé aléatoirement en g groupes. Les nœuds seront déplacés d'un groupe à un autre de telle façon à minimiser le nombre d'arêtes entre les groupes. Tous les nœuds doivent être déplacés exactement une seule fois. A chaque étape, on déplace le nœud qui donne la meilleure amélioration (le nœud qui minimise le plus les arêtes inter-groupes). La complexité du pire cas est en $O(n^3)$.

2. **Approches séparatives**
Ces Approches construisent l'arbre de manière inverse. L'arbre est construit à partir du graphe entier, en retirant itérativement les arêtes par poids décroissant. Les arêtes sont retirées une à une, et à chaque étape les composantes connexes du graphe obtenu sont identifiées à des communautés. Le processus est répété jusqu'au retrait de toutes les arêtes. Les méthodes existantes diffèrent par la façon de choisir les arêtes à retirer.

(a) **Les approches de Radicchi et Al. et d'Auber et Al. basées sur le clustering d'arrêtes**
La détection des arêtes intercommunautaires est ici basée sur le fait que de telles arêtes sont dans des zones peu clustérisées. Radicchi et al proposent un coefficient de clustering (d'ordre g) d'arêtes. Il est défini comme étant le nombre de cycles de longueur « g » passant par l'arête, divisé par le nombre total de tels cycles possibles (étant donné les degrés des extrémités de l'arête). Cet algorithme retire donc à chaque étape l'arête de plus faible clustering (d'un ordre donné, 3 ou 4 en pratique). Chaque suppression d'arête ne demande alors qu'une mise à jour locale des coefficients de clustering. La complexité totale est en $O(m^2)$. L'approche d'Auber et al. utilise un clustering d'arêtes d'ordre 3.

(b) **L'algorithme de Fortunato et Al basé sur la centralité d'information**
Le principe de cet algorithme est le même que toutes les approches séparatives. La

métrique utilisée par Fortunato et al. pour identifier les arêtes inter-communautés est la centralité d'information. La centralité d'information d'une arête est la diminution de la capacité du réseau due à la suppression de cette arête. Pour toute arête, on calcule la capacité du réseau puis la diminution due à la suppression de cette arête, et on affecte cette diminution au score de l'arête. Le calcul de la capacité du réseau est coûteux en temps. La complexité de cet algorithme est de $O(m^3.n)$.

(c) **L'intermédiarité de Newman et Girvan**
Neuwman et Girvan ont défini une mesure pour détecter les arêtes inter-communautés à supprimer en premier pour trouver les communautés. La mesure proposée est l'intermédiarité (en Anglais, betweeness). L'intermédiarité est une mesure qui donne des valeurs fortes aux arêtes qui ont une extrémité dans une communautés i et une autre extrémité dans une autre communauté j, et des valeurs faibles aux arêtes d'une même communauté. L'intermédiarité d'une arête est calculée comme le nombre des plus courts chemins du graphe qui passent par cette arête. On calcule l'intermédiarité de chaque arête du graphe, et on trie les arêtes dans l'ordre décroissant du score de l'intermédiarité. A chaque étape, l'arête ayant la plus forte intermédiarité sera supprimée. Les composantes connexes issues de la suppression des arêtes représenteront les communautés recherchées. L'inconvénient majeur de cette approche est le coût du calcul de l'intermédiarité des arêtes, pour chaque suppression d'une arête, le score de toutes les arêtes va être affecté, du coup, il sera recalculé, ce qu'est très coûteux. La complexité de l'algorithme est d'ordre de $O(m^2.n)$.

3. **Approches agglomératives**
Le principe global de ces approches consiste à regrouper les noeuds itérativement en communautés. Au début, les nœuds sont considérés comme si chacun d'eux constituait une communauté à part, c.à.d. il y a autant de communautés que de nœuds. Les communautés sont réunies deux à deux jusqu'à avoir une grande communauté représentant l'ensemble des nœuds du graphe. Il faut souligner qu'à chaque étape de regroupement de deux communautés, une métrique (fonction de qualité) est calculée et le partitionnement ayant la plus haute valeur de la métrique considérée représente le meilleur partitionnement du graphe en communautés.
Ci-dessous, nous présenterons les algorithmes les plus connus dans la littérature appartenant à la classe des approches agglomératives :

(a) **L'algorithme d'optimisation de la modularité proposé par Newman et amélioré par Clauset et al**
L'algorithme démarre avec un ensemble de nœuds isolés (tous les nœuds du graphe) considérés chacun comme une communauté. Les communautés sont jointes deux à deux jusqu'à avoir une grande communauté représentant le graphe entier. Newman introduit la notion de modularité; il s'agit de la fonction Q mesurant la qualité d'une partition du graphe en communautés . Afin de maximiser cette

quantité l'algorithme glouton proposé fusionne à chaque étape les communautés permettant d'avoir la plus grande augmentation de la modularité. Pour améliorer les performances de l'algorithme, seules les communautés ayant une arête entre elles peuvent être fusionnées à chaque étape. Chaque fusion se fait en O(n) et la mise à jour des valeurs des variations de Q (pour chaque nouvelle fusion possible) peut être effectuée facilement en O(m) . La complexité globale est donc O((m + n)n). Cette méthode est très rapide et permet de traiter de très grands graphes. La qualité des partitions obtenues est cependant moins bonne qu'avec des méthodes plus coûteuses. La complexité de cette méthode a été améliorée par Clauset en utilisant une structure de données adaptée.

(b) **Un algorithme efficace et sémantique pour la détection de communautés**
Les propriétés spectrales de la matrice Laplacienne du graphe, $G = (X,U)$ sont utilisées pour détecter des communautés. En effet les coordonnées x et y des vecteurs propres correspondant aux plus petites valeurs propres non nulles sont corrélées lorsque les sommets x et y sont dans la même communauté. Une distance entre sommets est alors calculée a partir de ces vecteurs propres, cette distance étant ensuite utilisée dans un algorithme de clustering hierarchique. Le nombre de vecteurs propres a considérer est a priori inconnu. Plusieurs calculs sont successivement effectués en prenant en compte différents nombres de vecteurs propres, et le meilleur résultat est retenu. Les performances de l'algorithme sont limitées par les calculs des vecteurs propres qui se fait en $O(n^3)$ pour une matrice creuse. Une amélioration de cette approche a été proposée et utilise une version normalisée de la matrice Laplacienne.

4. **Approches à base de motif**
Dans de nombreux réseaux, les communautés issues en regardant seulement le simple lien entre les nœuds ne sont pas très représentatives et ne nous fournissent pas assez de connaissance sur le réseau. Hormis le simple lien, les communautés peuvent être détectées en se basant sur un motif. Plusieurs réseaux (réseaux biologiques, réseau du web, etc.) contiennent de petits graphes qui ont de grandes fréquences d'apparition. Ces petits graphes, généralement à trois, quatre ou cinq nœuds, sont appelés motifs. Un motif est dit bien représenté dans un réseau donné, si la fréquence d'apparition du motif dans ce réseau est plus grande que celle dans un réseau aléatoire ayant les mêmes caractéristiques. Parmi les motifs possibles, un des plus simples est le triangle ; qui représente l'unité de base de la transitivité dans un graphe. Ce motif est bien représenté dans de nombreux réseaux réels.

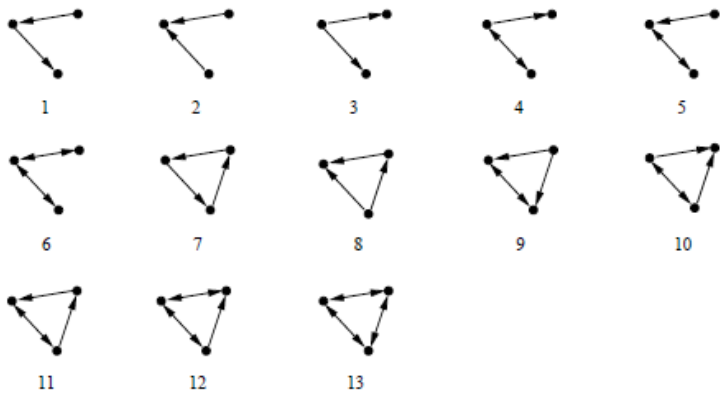


FIG. 3.3 – *Liste de tous les motifs possibles à trois nœuds.*

Et comme algorithme se basant sur cette approche on peut citer:

(a) **CFinder**
Cette méthode est basée sur une recherche de motifs locaux. Une communauté est définie comme une chaîne de k-cliques adjacentes. Une k-clique est un sous-ensemble de k sommets tous adjacents les uns aux autres, et deux k- cliques sont adjacentes si elles partagent k -1 sommets. L'avantage immédiat d'une telle approche est la détection de communautés avec recouvrement, un sommet pouvant appartenir à plusieurs k-cliques non forcément adjacentes. Une limite de cet algorithme est qu'il nécessite un paramétrage : la valeur de k (la taille des communautés à considérer). Généralement on choisi la valeur k= 4 acceptée comme la plus efficace.

3.5 Analyser les communautés

Une fois que les communautés sont formées, nous nous intéressons à l'étude et l'analyse des structures de graphes émergés, à la manière dont les nœuds sont interconnectés entre eux. En effet, plus les nœuds sont connectés plus la relation sociale entre eux est forte, et vice versa. Pour ce faire, nous nous basons sur quelques propriétés de la théorie des graphes qui sont :

(a) **Le degré moyen des communautés (deg-moy)**
Le degré moyen d une communauté C de n nœuds est calculé de la façon suivante :

$$degmoy(C) = \sum_{i}^{n} degré(i)/n \tag{3.3}$$

où, degré(i) est le nombre de voisins du nœud i dans la communauté C.

(b) **Le degré maximum (deg-max)**
Il représente la valeur maximale de tous les degrés des nœud d'une même communauté.

(c) **Le nombre de degré** 1 **(nb-deg**1**)**
C'est le nombre de nœuds de degré 1, c. à. d. les nœuds feuilles dans le graphe. Du point de vue "communauté", il s'agit des nœuds qui ne sont connectés qu'à une seule communauté.

(d) **La densité des communautés (d-comm)**
La densité moyenne des communautés d'un partitionnement s'obtient par le calcul de la densité de chaque communauté à part (considéré comme un graphe). Soit C une des communautés formées, et soient n et m le nombre de nœuds dans C et le nombre d'arêtes dans C, respectivement. La densité de C est calculée comme suit :

$$d - comm(C) = m/(n.(n-1)/2) \tag{3.4}$$

où, $n.(n-1)/2$ est le nombre maximum d'arêtes que peut avoir un graphe de n nœuds (nombre d'arêtes d'un graphe complet de n nœuds).

(e) **Coefficient de Clustering (CC)**
Le CC est la proportion du nombre de triangles dans un graphe donné.
Il est calculé comme la moyenne des CC de tous les nœuds du graphe et il est donné par l'équation :

$$(CC(i) = 2.k/degre(i).(degre(i)-1) \tag{3.5}$$

Où k représente le nombre d'arêtes entre les voisins du nœud i.

(f) **La Distribution de degrés (dis-deg)**
Elle représente l'énumération du nombre de noeuds pour chaque degré. La distribution de degrés est représentée sur un graphique où l'axe des abscisses (x) contient les degrés et l'axe des ordonnées (y) contient le nombre de nœuds pour chaque degré.

Dans notre étude des structures de communautés, nous notons la relation célébrité-fan entre les nœuds. Il existe trois types d'interactions (voir la figure suivante) :

- L'interaction célébrité-fan : combien de célébrités y-a-t-il dans la communauté que les fans connaissent et avec lesquelles ils interagissent ?
- L'interaction célébrité-célébrité : les célébrités connaissent-elles d'autres célébrités dans la communauté ? Combien ?
- L'interaction fan-fan : les fans connaissent-ils d'autres fans dans la communauté ? Combien ?

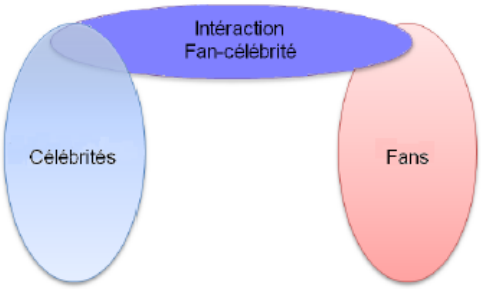


FIG. 3.4 – *Les interactions dans une communauté.*

3.6 Corrélation entre réseaux initiaux et structures de communautés

Selon la valeur de la densité, nous obtenons des structures de communautés différentes. En effet, pour une **faible densité**, les nœuds d'une même communauté n'ont pas tendance à se connecter entre eux et se contentent de se connecter presque à un même nœud.
Cela se traduit par l'apparition de plusieurs nœuds de degré 1 et peu de nœuds de degré élevé. En théorie des graphes, les structures qui répondent à cette particularité sont des structures de type :

- **Chemin / Cycle**
  Le nombre de nœuds de degré 1 de cette classe est très petit (0 *ou* $\approx 2/$ n) et le degré maximum deg-max est aussi très petit ($\approx 2/(n-1)$). Dans cette classe, on ne trouve pas de fans ou de célébrités évidente dans la communauté. Tous les nœuds jouent le même rôle. L'interaction est très faible étant donné que tous les nœuds ne connaissent que peu de membres de la communauté.
- **Étoile**
  Le nombre de nœuds de degré 1 de cette classe est très grand et le degré maximum est aussi très grand. Il est claire que le centre de l'étoile est la seule célébrité de la communauté. Tous les autres nœuds (les feuilles) sont des fans. Dans cette classe, il n'existe pas d'interaction entre les fans.
- **Bistar**
  Le nombre de nœuds de degré 1 de cette classe est très grand (≈ 1) et le degré maximum est moyen ($\approx 1/2$). Les deux centres du bistar sont les célébrités de la communauté. Tous les autres nœuds sont les fans. Aussi dans cette classe, comme la précédente, il n'existe que peu d'interaction entre les fans.
- **Spider**
  Le nombre de nœuds de degré 1 de cette classe est moyen ($\approx 1/2$) et le degré

maximum est aussi moyen ($\approx 1/2$). Il est évident que les feuilles du spider sont les fans de la communauté et la racine la célébrité. Les autres nœuds de la communauté ont un rôle d'intermédiarité entre les feuilles et la racine. L'interaction entre ces nœuds est très faible quant à l'interaction entre fans est inexistante.

- **Caterpillar**
  Le nombre de nœuds de degré 1 de cette classe est moyen (entre 1/2 et des valeurs proches de 1) et le degré maximum est petit ($O(log(n)/(n-1$))). Dans ce type de structures, il y a des célébrités qui n'interagissent pas beaucoup entre elles. Chaque célébrité a un groupe de fans avec lesquels elle interagit. Un fan connaît seulement une célébrité et n'interagit pas avec les autres fans.
- **Arbre**
  Le nombre de nœuds de degré 1 de cette classe est petit ou moyen (en général, entre $O(log(n)/n$) et 1/2 , cependant, dans certain cas particuliers ce nombre peut être plus grand mais sans atteindre des valeurs proches de (1) et le degré maximum est petit ($O(log(n)/n-1$)).
  Les interactions dans ce type de structure sont un mélange de celles des graphes Spider et graphes Caterpillar.
- **Roue**
  Le nombre de nœuds de degré 1 dans cette classe est très petit (≈ 0) et le degré maximum est très grand (≈ 1). Le centre du Roue est l'unique célébrité de la communauté. Tous les autres nœuds sont des fans. Dans cette classe, il existe une faible interaction entre les fans.

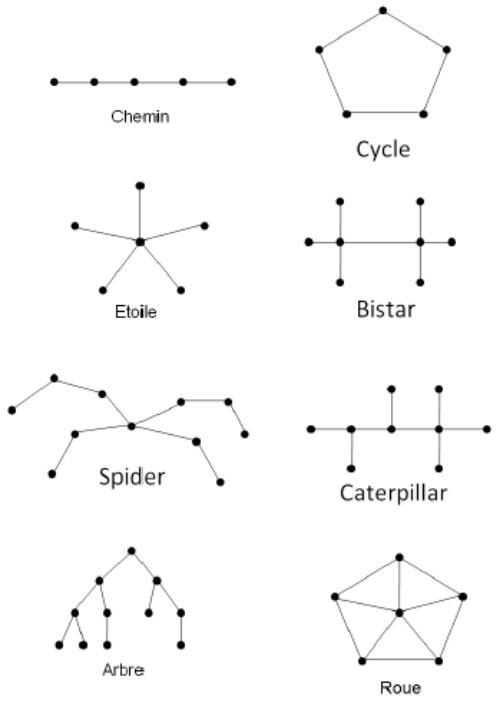


FIG. 3.5 – *Les classes de graphes de densité faible.*

Dans ce même esprit, nous pouvons observer que les types des structures de graphes de **densité moyenne** regroupent un grande nombre de nœuds de degré 1 et également un grand nombre de degré élevé. En effet, un graphe ne peut avoir une densité moyenne et un grand nombre de nœuds de degré 1, ou très peu de degré maximum. Donc, les combinaisons possibles sont celles s'apparentant à des structures de graphes de types :

- **Eventail**
  Un graphe Eventail $F_{n,m}$ est la jointure d'un graphe vide de n nœuds et un chemin de m nœuds (chaque nœuds du graphe vide est connecté à tous les nœuds du chemin). Le nombre de nœuds de degré 1 de cette classe est très petit (0) et le degré maximum peut être moyen ou grand (moyen si n et m sont proche, grand sinon). Nous considérons deux différents cas de degré maximum. Si le degré maximum est grand, alors la partition avec le plus petit nombre de nœuds contiendra les célébrités de la communauté. Sinon, les rôles ne sont pas évidents à distinguer. Toutefois, dans tous les cas, les interactions à l'intérieur de la communauté sont importantes puisque chaque nœud sait et interagit avec de nombreux autres nœuds.
- **Sun**
  Le graphe Sun est obtenu en ajoutant un nœud voisin à tous les nœuds d'un graphe complet (clique). Le nombre de nœuds de degré 1 de cette classe est moyen ($\approx 1/2$) et le degré maximum est aussi moyen ($\approx 1/2$). Les feuilles de ce

graphe sont les fans de la communauté et les nœuds du sous-graphe complet sont les célébrités. Chaque célébrité interagit avec presque toutes les autres célébrités, alors que les fans n'interagissent qu'avec une seule célébrité (ou au plus avec peu de célébrités). Dans cette classe, il n'y a presque pas d'interaction entre fans.

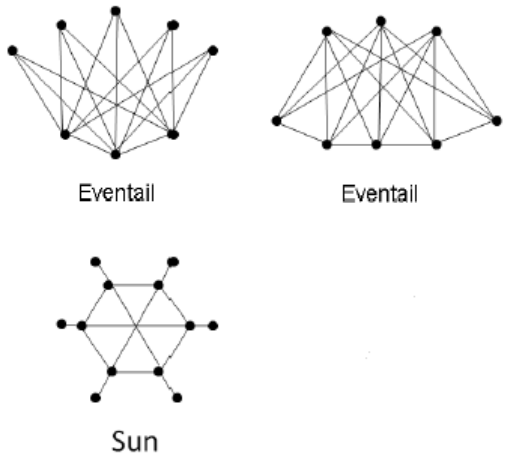


FIG. 3.6 – *Les classes de graphes de densité moyenne.*

Pour ce qui est des structures de communautés ayant une **densité moyenne forte**, l'unique possible combinaison est celle quand le nombre de nœuds de degré 1 est petit et le degré maximum est grand. Les graphes de cette classe sont les graphes complets (les cliques).

– **Clique**
  Le nombre de nœuds de degré 1 de cette classe est nul (= 0) et le degré maximum est très grand (= 1). Chaque nœud connaît et interagit avec tous les autres nœuds de la communauté. La communauté est alors fortement connectée.

FIG. 3.7 – *Les classes de graphes de densité forte.*

3.7 Conclusion

Dans ce chapitre, nous avons introduit les concepts de communautés dans les réseaux. Nous avons donné les définitions des communautés d'intérêt et cité les apports et importances de ces dernières dans la compréhension du fonctionnement des réseaux sociaux, on a cité les méthodes et les approche les plus répondu dans le domaine de la détection des communautés et on a classifié ces communautés selon des propriétés qu'on a cité précédemment. Dans le chapitre suivant, nous allons étudier le graphe social web, implémentér un algorithme basé sur l'optimisation spectrale sous MATLAB et étudier les résultats.

Chapitre 4

RESEAUX SOCIAUX ET WEB

4.1 Introduction

L'internet a été conçu au départ pour offrir un service simple, à savoir connecter tous les ordinateurs du monde, de la façon la plus économique possible. Mais il n'a pas été conçu pour un type d'applications particulier. Ceci dit, il a bien été utilisé jusqu'à ces derniéres années par un petit nombre d'applications spécifiques (courrier électronique, transfert de fichiers). Ces applications sont extrêmement utiles, et elles représentent d'ailleurs toujours la majorité des données échangées sur l'Internet mais le service qui a rendu l'Internet populaire auprés du grand public est le "World Wide Web" (toile d'araignée mondiale, abrégée en web) qui a commencé à se répandre en 1993. Le web repose sur trois idées principales qui sont la navigation par hypertexte, le support du multimédia, et l'intégration des services préexistants. Avec l'arrivé de web 2.0 en 2004, internet a subit beaucoup de changement. Grâce à la révolution technologique, l'internet se voit maintenant proposer des sons, vidéos et images en temps réel mais aussi d'avantage l'interactivité avec les sites internet. Par ailleurs le web 2.0 a amené l'évolution des réseaux sociaux. Les réseaux sociaux sont qualifiées de $\ll$ la grande affaire du moment $\gg$. Ils sont de plus en plus utilisés et surtout pour un public trés large.On note que les adolescents et les étudiants ont été les utilisateurs précurseurs de se type de sites. Ces sites sont maintenant largement répandus, afin de toucher toute les catégories de la population.

4.2 Problématique

Dans les sites Web collaboratifs actuels, un effort de saisie important est demandé aux utilisateurs afin d'identifier la communauté à laquelle ils appartiennent (description du profil personnel, du réseau social, etc.).
Cependant, ces sites exigent de chaque utilisateur une description explicite de son réseau social ou de son profil. De plus, seules les communautés ainsi explicitées sont identifiées. Or un grand nombre de communautés d'utilisateurs existent de façon implicite dans de nombreux domaines. Par exemple, tout site de musique généraliste rassemble une communauté d'utilisateurs ayant des goûts musicaux variés. Mais cette communauté est en fait composée de sous-communautés potentiellement disjointes, toutes liées à la musique (la communauté des amateurs de musique pop, de musique rock, etc). Découvrir et identifier précisément ces communautés implicites est un gain pour de nombreux acteurs : le propriétaire du site, les régies publicitaires en ligne et surtout, les utilisateurs du système.Dans ce memoire, nous proposons une méthode de détection de communautés basée sur les actions des utilisateurs.

4.3 Web

Le web est représenté par des milliards de pages écrites par des personnes indépendantes. Ces pages sont interconnectées par des liens hypertextes que les internautes utilisent pour se déplacer d'une page à l'autre. Chaque personne créant une page Web choisit librement les pages vers lesqulles pointer et peut changer d'avis au cours du temps. sans qu'il y ait un contrôle global. Lorsque l'on veut étudier le web, on se retrouve donc face à un immense ensemble d'éléments interconnéctés pour lesquels aucune autorité centrale n'existe. Dans ce cas, obtenir à un moment donné une vue d'ensemble des pages disponibles et de leurs interconnexions par les liens hypertextes devient trés difficile. Face à une telle situations les spécialistes du domaine ont fait appel à la théorie des graphes afin de connaître la structure globale du web. En effet le web peut être modélisé sous forme d'un graphe où les sommets représentent les pages web et le arcs représentent les liens hypertextes reliant ces pages.

4.4 L'historique du Web

Le Word Wide Web (www) a été mis en place par Tim-Berners Lee qui est considéré comme le père fondateur du web. Au milieu des années 1990, Internet fait son apparition au grand public en version 1.0 via des pages statiques codés en HTML. Il s'agit de sites non interactifs principalement destinés à la recherche d'informations : encyclopédies, etc. Au début des années 2000, le web évolué et est devenu dynamique. Il s'agit de sa version 1.5. Il est maintenant possible de consulter du contenu dynamique en ligne, via des bases de données : boutique en ligne, etc. En 2004, DaleDougherty utilise le terme Web 2.0 qui sera vite repris par Tim OReailly, spécialiste du World Wide Web. Le web 2.0 se caractérise par la prise de pouvoir des internautes sur internet, grâce notamment aux réseaux sociaux. Plus qu'un bouleversement technologique, l'apparition du web 2.0 prend une véritable dimension sociologique puisqu'il replace le consommateur à la source de l'information. Le web "classique" était composé de pages reliées entre elles, alors que le Web 2.0 est constitué d'un ensemble d'utilisateurs reliés entre eux. Tous les nouveaux services se développent avec une philosophie particulière : ne pas "maltraiter" l'internaute et prendre en compte ses besoins. Pour ce faire, les sites demandent trés peu d'informations lors de l'inscription (afin de ne pas rebuter l'internaute), développent des services interactifs généralement gratuits (l'objectif est de donner le maximum de services gratuitement). L'internaute est donc enfin devenu le centre d'intérêt d'internet. Depuis quelques mois, on entend de plus en plus parler du Web 3.0, c'est-à-dire un web encore plus humain, encore plus intelligent.

4.5 Le réseau social web

Lorsque l'on veut étudier le Web, on se retrouve donc face à un immense ensemble d'éléments interconnectés pour lesquels aucune autorité centrale n'existe. Dans ce cas, obtenir à un moment donné une vue d'ensemble des pages disponibles et de leurs interconnexions par les liens hypertextes devient très difficile. Face à une telle situation, les spécialistes du domaine ont fait appel à la théorie des graphes afin de connaître la structure globale du web. En effet, le web peut être modélisé sous forme d'un graphe où les sommets représentent les pages web et les arcs représentent les liens hypertextuels reliant ces pages. D'une façon explicite le graphe du web, noté G , est défini comme suit : G est un graphe simple orienté non pondéré $G = (X,U)$possédant $X =$ n sommets et $U =$ m arcs tel que :
X= { x_i / x_i est une page web}
$U = \{x_i x_j \in X \times X;\ x_i\ pointe\ vers\ x_j\}$

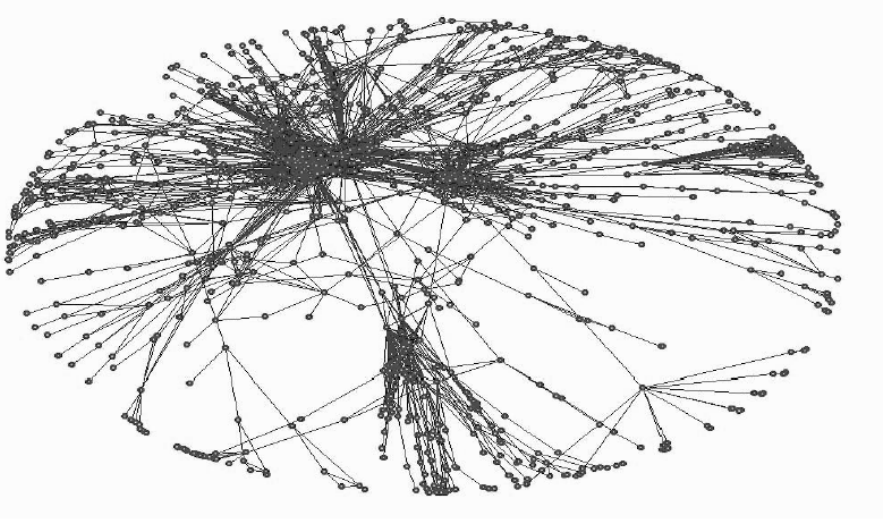


FIG. 4.1 – *Sous graphe du web*

Définition 4.1. Les blogs Les blogs sont de mini sites web, similaires à de petits territoires personnels qui permettent d'échanger des informations sur ses passions, ses goûts, sa vie privée... et par extension de revendiquer son identité. Faciles d'accès et à la création ils sont recensés au nombre de 156 millions en 2011. Leur longévité est cependant limité de part la difficulté à offrir du contenu toujours nouveau, leur pertinence et donc leur capacité à être attractifs ainsi qu'à la grande concurrence face à laquelle ils doivent faire face.

4.6 Quelques exemples de réseaux sociaux web

4.6.1 MySpace

MySpace est bien évidemment un réseau social. Il entre dans la catégorie des réseaux implicites ceux qui n'ont pas été créés pour être des réseaux sociaux mais qui, de par leur évolution en sont devenus.
Selon Wikipédia, « MySpace est une plate forme qui met gratuitement à disposition de ses membres enregistrés un espace web personnalisé, permettant d'y faire un blog, d'y envoyer ses photos/vidéos/sons et d'y remplir diverses informations personnelles». Le site possède aussi un système de messagerie qui permet de communiquer entre membres et de se construire son « réseau d'amis ». Il s'agit d'une définition très courte mais qui couvre bien le domaine d'activité de MySpace.
Selon une étude de ComScore Media Metrix16, MySpace générerait 51,4 millions de visiteurs uniques sur le mois de mai 2006.

4.6.2 Twitter

Twitter a été créé en 2006 dans le but de permettre à ses utilisateurs de partager facilement de courts messages textuels appelés tweets. Le système ayant été initialement conçu pour partager les tweets via SMS, une limite de 140 caractères a été fixé a ses messages courts . Bien que le système soit massivement utilisé via le web et via des applications développées pour ordinateurs; cette contrainte de 140 caractères n'a jamais été levée. La croissance de se service est actuellement important, en mois d'avril 2010 Twitter comptait plus de 6 millions utilisateurs enregistrés, 300000 nouveaux compts par jour.

4.7 Les différents types de réseaux sociaux web

Les réseaux sociaux peuvent être divisés en plusieurs catégories. Selon les différents avis, plusieurs classifications sont proposées, on peut citer la suivante qui est basé sur le fait que les réseaux sociaux fournissent des outils qui facilitent le processus de mise en relation au tour d'un centre d'intérêt commun :

(a) **Les réseaux plate-forme de partage**
D'un côté les réseaux sociaux de type plate-forme, mettent à disposition des contenus sans création de profil, sans qu'il y ait besoin d'avoir une appartenance à la communauté. Cependant, si l'internaute souhaite faire partie de cette communauté, en diffusant ses propres vidéos par exemple, il doit obligatoirement se créer une fiche d'identité (profil). Ainsi, l'internaute est reconnu en tant que membre de la dite communauté. Une fois membre, il est possible de rechercher des profils, de diffuser et de partager des contenus. Par contre, il n'y a pas de mise en relation à proprement parler. Les membres du réseau ne peuvent interagir entre eux qu'en

réagissant, sous forme de commentaires, à des contenus publiés. Ainsi les réseaux de type plate-forme sont essentiellement orientés vers le partage et la diffusion de contenu (souvent des vidéos ou des photos). Exemple : youtube...

(b) **Les réseaux personnels (généralistes ou thématiques)**
De l'autre côté, nous avons les réseaux sociaux personnels (généralistes ou thématiques) qui se rapprochent réellement plus de l'univers social et partage du contenu. La lecture du contenu est toujours libre d'accès, comme pour les réseaux de type plate-forme. Cependant, la création d'un profil est nettement plus avancée. Ce profil, véritable carte d'identité, est essentiel pour entrer dans la communauté mais surtout pour y participer et créer des liens avec les autres membres. En général, ce profil est présenté par un espace personnel (page web) visible par tous les internautes. Il s'agit d'un espace réservé aux membres où ils ont la possibilité d'y mettre tout ce qu'ils souhaitent : textes, histoires, journal intime, photos de vacances ou encore vidéos. La mise en relation des membres se fait de manière très simple : soit par un lien vers l'autre profil que l'internaute insère manuellement soit en invitant l'autre membre à se joindre à son cercle d'amis. Les réseaux sociaux personnels sont plus orientés vers la diffusion d'informations que vers la relation entre membres. Ils sont le plus souvent orientés vers un centre d'intérêt (lecture,cinéma...) avec le but de faire partager ses passions avec le reste de la communauté.Exemple : Friendster...
Les réseaux personnels et thématique sont les mêmes que les généralistes, mais ils sont orientés autour d'une thématique : les voitures, la cuisine ...

(c) **Les réseaux professionnels**
Les réseaux sociaux professionnels sont les plus avancés d'un point de vue des fonctionnalités proposées pour la gestion de sa communauté. La création de son profil est primordiale pour pouvoir profiter de tous les services associés aux réseaux professionnels. L'objectif de ce type de sites est clairement de se construire un réseau, le plus grand et pertinent possible. Que ce soit dans le cadre d'une recherche d'emploi, pour trouver des capacités de financement ou encore des opportunités de partenariat, les réseaux sociaux peuvent s'avérer bien utiles pour ce genre de demande. De ce fait, la création de sa fiche personnelle (son curriculum vitae ici) est vitale. Si vous voulez être trouvé ou trouver du monde, il faut que cette fiche soit la plus complète possible tout en définissant bien ses objectifs professionnels. En fait, l'utilisation d'un réseau social en ligne doit se faire à l'image d'un véritable réseau. D'un point de vue de la mise en relation entre membres, les réseaux professionnels sont certainement les plus aboutis tout comme la recherche de profil. Elles aboutissent souvent sur les profils recherchés et en quelques clics, avec un message de demande de mise en contact, l'invitation est envoyée par courrier électronique. Par la suite, la personne invitée a le choix de refuser ou d'accepter la mise en relation. Si elle l'accepte, elle sera en contact direct avec la

personne et aura accès à toutes ses informations professionnelles mais également, et surtout, verra le réseau du membre ainsi que son degré de proximité avec ses autres contacts. La création de son réseau n'est pas plus compliquée que cela et permet d'être rapidement en relation avec « le monde entier ».
Exemple : 6nergies. . . .

4.8 Découverte des communautés Web

Dès les premiers travaux sur la reconnaissance des communautés sur le Web (par exemple Gibson et al. (1998)), le lien hypertexte est utilisé comme base de raisonnement. L'apport majeur en la matière est l'algorithme HITS de Kleinberg (1998), définissant les notions d'autorités et de hubs, structurant une communauté autour d'un sujet donné. Imafuji et Kitsuregawa (2002) concluent à l'appartenance d'une page à une communauté si cette page est plus majoritairement référencée depuis l'intérieur de la communauté que depuis son extérieur. Ils utilisent un algorithme de flot maximum afin d'isoler les noeuds faisant partie d'une même communauté, en se basant sur l'algorithme proposé par Flake et al. (2000). Dourisboure et al. (2007) identifient au sein d'un graphe du Web les communautés comme autant de sous-graphes denses et bipartis au sein de ce graphe. Le graphe biparti représente d'une part les centres d'intérêt de la communauté et d'autre part ceux qui citent la communauté (les hubs). Cette méthode permet de mettre en évidence les éventuels partages des mêmes centres d'intérêt par plusieurs communautés d'acteurs, ou au contraire le partage de mêmes acteurs par plusieurs centres d'intérêt des communautés. Ces approches fournissent une analyse avancée des liaisons entre les différentes pages structurant une communauté thématique, mais ne permettent pas en revanche de rapprocher des utilisateurs de par leurs intérêts ou activités : le partage de lien hypertexte n'étant plus nécessairement la base de l'activité communautaire dans les échanges sociaux du Web collaboratif (évaluation de contenu par l'utilisateur ...).

4.9 Communauté dans un réseau social web

Depuis 1999, les communautés font partie des éléments d'observation locaux auxquels on prête beaucoup d'attention. En réalité, un premier problème en relation directe avec la structure locale du graphe du Web consiste à définir une communauté. Un second problème important consiste à proposer des méthodes pour les détecter automatiquement. L'existence dans le graphe du web de zones plus densément connectées que d'autres constitue une des caractéristiques non triviales de ce graphe. Ces zones sont appelées communautés (par analogie avec les réseaux sociaux) et correspondent intuitivement à des groupes de sommets plus fortement connectés entre eux qu'avec les autres sommets, comme illustré par la figure suivante :

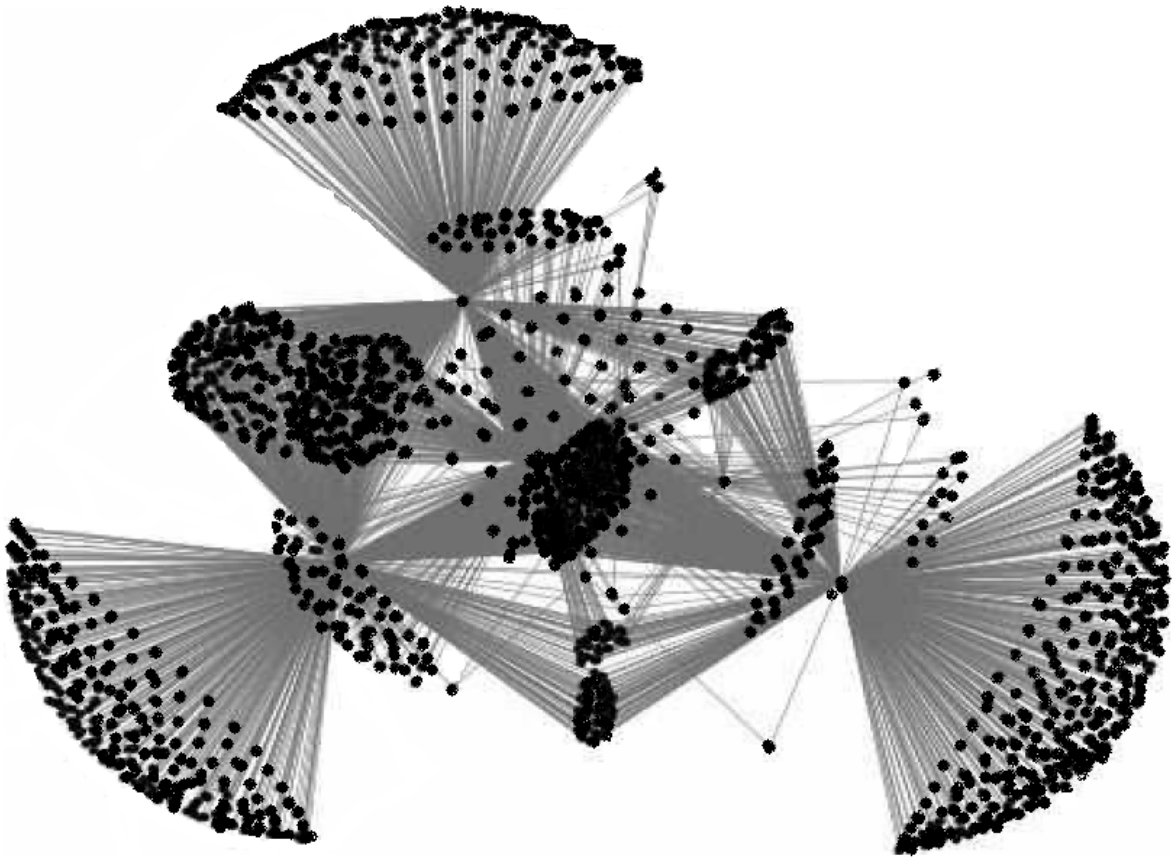


FIG. 4.2 – *Les zones denses du graphe du web*

4.10 conclusion

Web est le support d'une grande variété de services utilisant des techniques informatiques différentes. Il est conçu de façon telle que plusieurs communautés ayant des intérêts divers (recherche, éducation, loisir, commerce.) peuvent coexister sur le même réseau.
Dans le chapitre suivant on va essayer de programmer un algorithme qui va partitionner ces communautés d'une façon optimale.

Chapitre 5

L'APPLICATION

5.1 L'application

Le problème de détection de communautés a été transformé en un problème d'optimisation de modularité. Comme nous l'avons mentionné précédemment, la modularité est une fonction de qualité de partitionnement du réseau en communautés. Les bonnes valeurs de la modularité correspondent à de bonnes communautés dans le réseaux. Plusieurs travaux ont été proposés pour détecter les communautés en optimisant la modularité. Nous présentons ici l'un des plus importants d'entre eux: **Un algorithme basé sur l'optimisation spectrale**.
Newman et al. proposent un algorithme de détection de communautés en optimisant la fonction de modularité. Les auteurs définissent une matrice de modularité calculée à partir de la matrice d'adjacence du graphe. La matrice de modularité est inspirée de la définition de la modularité, et son optimisation est faite en utilisant les valeurs et vecteurs propres de cette matrice. La matrice de modularité est calculée comme suit. Soit A_{ij} la matrice d'adjacence du graphe G de m arêtes. La matrice de modularité B_{ij} est donnée par :

$$B_{ij} = A_{ij} - (k_i k_j / 2m) \tag{5.1}$$

Où, k_i est le degré du nœud i. L'algorithme proposé calcule le vecteur propre de la plus petite valeur propre positive. Un premier partitionnement du graphe en deux composantes est fait suivant les signes du vecteur propre (les nœuds de signe + dans un groupe et ceux de signe − dans un autre). Le processus est répété pour les deux sous-graphes obtenus jusqu'à ce que les composantes ne seront plus divisibles (pas de valeur propre positive de la matrice de modularité des sous-graphes obtenus). La complexité de cet algorithme est de $O(n^2 logn)$.

5.1.1 Le programme sous MATLAB

```
disp('Donner le nombre de pages web n')
 n=input('n')
  disp('Donner
le nombre des arétes')
 m=input('m')
%La matrice d'adjacence
for i=1:n
    for j=1:n
        A(i,j)=input('A(i,j) =');
    end;
end;
  A

%les degrés des i
for i=1:n
```

```
    d(i)=0;
for j=1:n
        d(i)=d(i)+A(i,j);
    end;
   d(i);
end;
%Les degres de j
for j=1:n
    k(j)=0;
    for i=1:n
        k(j)=k(j)+A(i,j);
    end;
    k(j);
end;
%La matrice de modularité
for i=1:n
    for j=1:n
        B(i,j)=A(i,j)-d(i)*k(j)/(2*m);
    end;
end; B
%Les valeurs propres
C=eig(B);
%La plus petite valeur propre positive
for i=1:n
    if (C(i)>0)
        valp=C(i);
        break;
   end;
end; valp
 C
  disp('la plus petite valeur propre est valp')
   for t=i+1:n
    if (C(t)>0)&&(C(t)<=valp)
        valp=C(t);end; end; valp
[V,D]=eig(B)
 D;

for j=1:n
 for i=1:n
         D(i,j);
         valp;

         if(D(i,j)==valp)
```

```
            j;
         break;
         end;
        end;

     if (D(i,j)==valp)
         j;
         break;
     end;
 end;
j
          for i=1:n
     V(i,j);
 end;
 disp('Le vecteur propre de la plus petite valeur propre est')
 V(:,j)
 for i=1:n
     F=V(:,j);
 end;
 F
 disp(' les valeurs + dans COMp et les - dans COMn')
for h=1:n
    if (F(h)>0)
        COMP=h;
        COMp(h)=COMP;
    elseif(F(h)<0)
        COMN=h;
        COMn(h)=COMN;
         end;
  end;
     COMn

   for h=1:length(COMn)
      if(COMn(h)==0)
          COMp
       break;
      else
          break;
      end;
   end;
   %la matrice des communautés
   VE=zeros(n,n);
   VEC=zeros(n,n);
```

```
   for i=1:n
       for j=1:length(COMn)
      VE=COMn(j);
      VEC(1,j)=VE;
       end;
     for j=1:length(COMp)
      VE=COMp(j);
      VEC(2,j)=VE;
 end;
 end;
disp('VEC est la matrice des communautés chaque ligne est une
communauté')
 VEC
 K=2
 disp('f est le nombre de ligne de VEC')
 f=2
  while(K<n)
for h=1:2*f
 if(h>n)
 break;
 else
 VEC(h,:)

  for i=1:length(VEC(h,:))
     for j=1:length(VEC(h,:))
         if(VEC(h,j)==0)||(VEC(h,i)==0)
             A(i,j)=0;
         elseif (i==j)
     A(i,j)=0;
         else
     A(i,j)=1;
         end;
         end;
     end;
 A
 if A==zeros(n,n)
 end;

 disp('les degrés des i')
for i=1:n
    d(i)=0;
for j=1:n
        d(i)=d(i)+A(i,j);
```

```
    end;
   d(i);
end; disp('Les degres de j') for j=1:n
    k(j)=0;
    for i=1:n
        k(j)=k(j)+A(i,j);
    end;
    k(j);
end; disp('La matrice de modularité') for i=1:n
    for j=1:n
        B(i,j)=A(i,j)-d(i)*k(j)/(2*m);
    end;
end; B;
 disp('Les valeurs propres')
  C=eig(B)
  disp('La plus petite valeur propre positive')
   for i=1:n
    if (C(i)>0)
        valp=C(i);
        break;
    else
   end;
end; valp
 disp('la plus petite valeur propre est valp')
 for t=i+1:n
    if (C(i)~=0)&&(C(t)>0)&&(C(t)<valp)
        valp=C(t);
    else
 end;
end; valp
[V,D]=eig(B);
 D

 for j=1:n
     for i=1:n
         if D(i,j)==valp
             j
         break;
         end;
 end;

     if D(i,j)==valp
         disp('le j de la plus PETIE')
```

```
        j
        break;
 end;
 end;
j

         for i=1:n
     V(i,j);
 end;
 disp('Le vecteur propre de la plus petite valeur propre est')
 V(:,j)
 for i=1:n
     F=V(:,j);
 end;
 F

 disp(' les valeurs + dans COMp et les - dans COMn')
for i=1:n
    if (F(i)>0)
 COMP=i;
        COMp(i)=COMP;
    elseif(F(i)<0)

        COMN=i;
        COMn(i)=COMN;
    else
         end;
  end;
  COMn
    for i=1:length(COMn)
      if(COMn(i)==0)
          COMp
       break;
      else
          disp('y a qu une communauté')
          break;
      end;
   end;
   disp('la matrice des communautés')
   VE=zeros(n,n);
   for i=1:n
       for j=1:length(COMn)
      VE=COMn(j);
```

```
      VEC(n,j)=VE;
       end;
     for j=1:length(COMp)
      VE=COMp(j);
      VEC(n-1,j)=VE;
    end;
    n==n-1
 end;
 end;
 VEC
 end;
 K=K+1
 end;
VEC
```

Et pour l'exemple d'exécution on peut prendre la matrice d'adjacence symétrique suivante:

```
A =

     1     0     0     0     1     0     0
     0     1     0     1     0     1     0
     0     0     0     1     1     1     0
     0     1     1     1     0     1     0
     1     0     1     0     1     1     1
     0     1     1     1     1     1     0
     0     0     0     0     1     0     1
```

FIG. 5.1 – *La matrice d'adjacence*

Et après l'exécution du programme on a eu le résultat suivant :

```
VEC =

     0     2     0     0     5     6     0
     1     0     3     4     0     0     7
     0     0     0     0     0     0     0
     0     0     0     0     0     0     0
     0     0     0     0     0     0     0
     0     2     0     0     5     6     0
     0     0     0     0     0     0     0
```

FIG. 5.2 – *La matrice des communautés*

Oú chaque ligne représente une communauté.

5.2 Conclusion

On a remarqué que l'algorithme nous aide a définir des communautés d'une façon optimale : il devise premièrement le graphe en deux communautés, ensuite il vérifie la possibilité de deviser ces sous communautés chacune en deux sous-sous communautés et on remarque que la solutions optimale est d'avoir seulement deux communautés.

5.3 Conclusion générale

Nous nous sommes intéressés dans ce mémoire à l'étude de structures de communautés dans les réseaux. L'étude de structures de communautés dans les réseaux est d'une importance capitale pour la compréhension des fonctionnalités et comportements des réseaux. Dans cette problématique de recherche, les travaux se focalisent sur deux parties : détection de communautés et analyse de communautés. La détection de communautés consiste à identifier les groupes de nœuds dans un graphe constituant les communautés. La difficulté réside dans le fait que ni le nombre ni la taille des communautés ne sont connus à priori. L'analyse de communautés, quant à elle, consiste à faire sortir des propriétés des communautés détectées, les analyser et ensuite essayer de comprendre les fonctionnalités et le comportement du réseau étudié. Nous avons modélisé les réseaux sociaux par les graphes et nous les caractérisons par un ensemble de paramètres.

Les recherches dans ce domaine sont focalisés beaucoup plus sur le web qui est devenu une source d'information très importante, il donne libre accès à une masse d'informations très variées sur tous les domaines. Ce qui explique sa taille qui ne cesse d'augmenter, dépassant le 1 billion de pages web. La recherche d'information devient de plus en plus difficile, pour remédier à ce problème, notre projet a été de résoudre le problème de recherche d'information par une des alternatives existantes qui est le partitionnement du web en communautés. Avant de prendre notre décision sur le choix des méthodes de résolution, nous avons comparé les méthodes existantes dans la littérature. Les deux critères ayant influencé notre choix, sont : la qualité des résultats ainsi que le temps d'exécution.

A travers notre étude, on a étudié les approches de détection de communautés et les propriétés utilisées pour leurs analyse , Cependant, plusieurs interrogations concernant les communautés ont été posées. Parmi ces interrogations, nous pouvons citer, leurs construction dans le temps, leur robustesse face à des perturbations ou bruits introduits sur ces structures (par exemple : disparition d'un lien entre 2 nœuds,$\cdots$). L'aspect de la robustesse des réseaux constitue un challenge très important pour comprendre leur fonctionnement, le comportement des entités les constituant et surtout pour comprendre les interactions qui peuvent se produire entre elles, permettant l'émergence de certains comportements qui n'étaient pas du tout prévisibles au préalable. Actuellement, les études de la robustesse des réseaux qui existent dans la littérature traitent cet aspect du point de vue purement structurel, c.à.d. toutes les perturbations sont appliquées soit sur les nœuds, soit sur les arêtes du graphe.

Pour ce qui est de notre étude, nous nous sommes intéressés a l'implémentation d'un algorithme basé sur l'optimisation spectrale sous l'outil de développement **MATLAB** 7 en utilisant les matrice d'adjacence et la seul condition pour cette matrice et qu'elle soit symétrique c.à.d. le graphe engendré est non orienté. .

Bibliographie

[1] A. H. Dekker. B. D. Colber. *Network Robustness and Graph Topology*. Defence Science and Technology Organisation (DSTO), 2004.

[2] A. Mislove, M. Marcont, K. P. Gummadi, P. Druschel, B. Bhattacharjee. *Measurement and Analysis of Online Social Networks*. Max Planck Institute for Software Systems, Rice University, University of Maryland.

[3] B. Karrer, E. Levina, M. E. J. Newman. *Robustness of community structure in networks*. Université de Michigan, 2008.

[4] B. Serrour. *Détection et Analyse de Communautés dans les Réseaux*. Université Claude Bernard Lyon 1, 2012.

[5] C. Thovex. *Réseaux de Compétences : de l'Analyse des Réseaux Sociaux à l'Analyse Prédictive de Connaissances*. Université de Nantes, 2012.

[6] D. B. Horn, T. A. Finholt, J. P. Birnholtz, D. Motwani, S. Jayaraman. *Six degrees Of Jonathan Grudin: A social network analysis Of The evolution and impact Of CSCW research*. Université de Michigan.

[7] E. Birmelé. *Etude structurelle des réseaux : modéles aléatoires, motifs et cycles*. l'Université d'Evry-Val d'Essonne, 2011.

[8] E. Griechisch, Andràs Pluhàr. *Community Detection by using the Extended Modularity*. Acta Cybernetica, 2010.

[9] E. Navarro, R. Cazabet. *Détection de communautés, étude comparative sur graphes réels*. Université de Toulouse, 2010.

[10] F. Filliettaz, M. Gregori. *Un enjeu pour l'enseignement Comprendre les réseaux sociaux numériques*. DIP Genève sous licence Creative Commons, 2011.

[11] F. Rossi et N. Villa-Vialaneix. *Représentation d'un grand réseau à partir d'une classification hiérarchique de ses sommets*. Journal de la Société Française de Statistique, 2011.

[12] G. Erétéo. *Analyse des réseaux sociaux et web sémantique: un état de l'art*. Agence nationale de la recherche, 2009.

[13] H. Hacid. *Le Web Social 2010*. Université Lyon 2,2010.

[14] L. Saglietto. *Quelques points de repéres deans l'étude des réseaux par la théorie des graphes*. Networks and communication studies,2006.

[15] M. Gilli. *Numerical methods in finance*. Université de Geneva et Institut suisse de finance, 2008.

[16] M. Giraud. *Les réseaux sociaux peuvent-t- il devenir un nouvel outils marketing pour l'entreprise*, 2009.

[17] M. Laris-eisti. *Introduction à l'analyse des réseaux sociaux*. Université d'Evry Val d'Essonne, 2009.

[18] M. Mercanti-Guérin. *Analyse des reseaux sociaux et communautes en ligne : quelles applications en marketing* .

[19] M. Teixeira. *L'émergence de réseaux sociaux sur le Web comme nouveaux outils de marketing*. Université d'Ottawa, 2009.

[20] N. Henry, J.D. Fekete. *Représentations visuelles alternatives pour les réseaux sociaux*. Lavoisier,2008.

[21] O. Le Deuff. *Le succès du web 2.0 : histoire, techniques et controverse*. Université Rennes 2, 2007.

[22] P. Pons. *Détection de communautés dans les grands graphes de terrain*. Université Paris 7, 2007.

[23] P. Torloting. *Enjeux et perspectives des réseaux sociaux*. Institut superieur du commerce de paris, 2006.

[24] S. Goyal .A. Vigier. *Robust Networks*. Université de Cambridge, 2010.

[25] S. Lemmouchi. *Etude de la Robustesse des Graphes Sociaux Emergents*. Université Claude Bernard Lyon 1 (UCBL), 2012.

[26] W. Y. Yang,W. Cao, T. S. Chung, J. Morris. *Applied numerical methods using Matlab*. Université de Chung-Ang, Université de Chung Chung-Ang Pennsylvania State, The Université d'Auckland, 2005.

[27] Y. Upadrashta. *Emerging social networks in peer-to-peer systems*. University of Saskatchewan.