REPUBLIQUE ALGERIENNE DEMOCRATIQUE ET POPULAIRE
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Mohamed Khider – BISKRA

Faculté des Sciences Exactes, et des Sciences de la Nature et de la Vie
Département d'Informatique

Mémoire en vue de l'obtention du diplôme de Magister en informatique

***Option*: Data Mining et Multimédia**

Intitulé :

FOUILLE ET APPRENTISSAGE AUTOMATIQUE DANS LES RESEAUX SOCIAUX DYNAMIQUES

Présenté par :

NEDIOUI MOHAMED ABDELHAMID

Devant le jury composé de :

Pr. Djedi Noureddine	**Professeur**	**President**	**Universite De Biskra**
Pr. Moussaoui Abdelouahab	**Professeur**	**Rapporteur**	**Universite De Setif**
Dr. Babahenini Mohamed Chaouki	**Maitre De Conferences A**	**Examinateur**	**Universite De Biskra**
Dr. Foudil Cherif	**Maitre De Conferences A**	**Examinateur**	**Universite De Biskra**

Année Universitaire 2014/2015

Dédicace

Je dédie ce modeste travail à :

L'âme de mon père

Ma femme et mes enfants

Toute la famille

Et tous mes amis

Remerciements

Tout d'abord, je remercie mon Dieux, qui m'a donné le tout pour pouvoir accomplir ce travail.

Je remercie tout particulièrement le Pr. MOUSSAOUI Abdelouahab qui a encadré ce mémoire. Il a toujours été disponible et a su me guider tout au long de mon mémoire tout en me laissant une grande liberté.

Je souhaite remercier vivement le Pr DJEDI Noureddine qui m'a fait l'honneur de présider le jury de soutenance.

Je remercie de plus les membres de mon jury, le Dr Mohamed Chaouki BABAHENINI et le Dr Foudil CHERIF de m'avoir fait l'honneur d'être rapporteurs de mémoire et surtout d'avoir accepté de juger mon travail dans des délais très courts malgré leurs emploi de temps plus que saturé.

Je tiens à adresser mes remerciements les plus sincères aux personnes qui m'ont accompagné au cours de ce mémoire.

Résumé

L’analyse de réseaux sociaux est un outil qui s’impose dans de nombreuses sciences. Un de ces outils spécifiques à l’analyse de réseaux sociaux est la détection de communautés. De nombreux algorithmes de détection de communautés ont été développés mais beaucoup ont une approche statique, c’est à dire ne considèrent pas que l’ordre d’apparition a une importance. De plus, ils posent le problème de la robustesse, car ces différents algorithmes proposent des résultats très différents.

L’objectif de ce travail est de proposer une nouvelle approche de détection de communautés qui serait stable, précise et efficace pour des réseaux sociaux avec des liens inter-communautés élevés. Pour cela, nous avons défini une nouvelle méthode qui fonctionne en deux phases. Durant la première phase, nous détectons tous les circuits afin de décomposer le réseau initial en petits groupes élémentaires. Dans la deuxième phase, nous proposons une procédure itérative ayant pour objectif l'identification des différentes communautés en fusionnant les différents sous graphes issus de la première phase via un principe de fusion utilisé dans les méthodes basées sur des cliques.

L’approche proposée est évaluée sur différents types de réseaux en variant le nombre de liens inter-communautés. La performance de l’approche proposée est comparée avec d'autres algorithmes de détection de communautés qui montre l’efficacité de notre approche.

Mots clés:
Classification dynamique, Fouille de données, Détection de communautés, Réseaux sociaux.

Abstract

The social networks analysis is a tool that is necessary in many sciences. One such specific tools for social network analysis is community detection. Many of community detection algorithms have been developed but many have a static approach, or do not consider the order of appearance in importance. De plus, they pose the problem of robustness, as these algorithms propose very different results.

The objective of this work is to propose a new community detection approach that would be stable, precise and effective social networks with high inter-community links. For this we have defined a new method that works in two phases. During the first phase, we detect all circuits to decompose the original network into small elementary groups. In the second phase, we propose an iterative procedure aimed identification of the different communities by merging the different sub graphs from the first phase through a principle used in fusion methods based on cliques.

The proposed approach is evaluated on different types of networks by varying the number of inter-community links. The performance of the comparative approach propose is with others algorithms communities detections which shows the effectiveness of our approach.

Keywords:
Dynamic classification, Data mining, Detection of community, Social networks.

ملخص

تحليل الشبكات الاجتماعية هي أداة ضرورية في كثير من العلوم .أحد هذه الأدوات لتحليل الشبكات الاجتماعية هو كشف للمجتمعات المحلية . وقد وضعت العديد من الطرق و البرامج للكشف عن المجتمعات لكن العديد منها له نهج ثابت، أي لا تنظر في ترتيب الظهور له الأهمية .بالإضافة إلى ذلك، فإنها تثير مشكلة متانة كما تقدم هذه البرامج نتائج مختلفة.

والهدف من هذا العمل هو اقتراح نهج جديد الكشف المجتمعات التي من شأنها أن تكون شبكات اجتماعية مستقرة ودقيقة وفعالة مع وصلات عالية بين المجتمع .لهذا قمنا بتحديد طريقة جديدة تعمل على مرحلتين .خلال المرحلة الأولى، كشف جميع الدوائر لتتحلل الشبكة الأصلية إلى مجموعات صغيرة ابتدائية . في المرحلة الثانية، فإننا نقترح إجراء تكراري بهدف تحديد المجتمعات المختلفة عن طريق دمج المجموعات الفرعية الناتجة عن المرحلة الأولى من خلال مبدأ المستخدمة في طرق الانصهار على أساس الزمر. يتم تقييم النهج المقترح على أنواع مختلفة من الشبكات من خلال تغيير عدد من الروابط بين المجتمع . تتم مقارنة أداء النهج المقترح مع المجتمعات الأخرى المكتشفة مع برامج اخرى و التي تظهر فعالية طريقتنا المقترحة.

كلمات مفتاحية :

تصنيف ديناميكي، التنقيب على المعطيات ، الكشف عن المجتمعات ، الشبكات الإجتماعية

Table des matières

Liste des figures

Liste des tableaux

Introduction générale

Les réseaux sociaux sont omniprésents depuis l'arrivée d'Internet. Ils permettent de représenter les interactions entre les différents individus d'un système, que ce soit un échange de photos entre amis, des mails ou des SMS, entre différents groupes d'individus.

Ces réseaux sont souvent modélisés par des graphes, une structure permettant l'encodage de données relationnelles. Quel que soit le domaine applicatif, cette modélisation sert à l'étude de la structure émergeant des entrelacements entre individus. Ainsi, les utilisateurs d'un réseau social auront tendance à former des groupes plus fortement connectés entre eux qu'avec le reste du réseau.

En général, la densité des liens entre les nœuds du réseau varie d'une zone à une autre, ce qui implique l'existence de groupes de nœuds fortement connectés entre eux mais faiblement reliés aux autres nœuds du réseau. Ces zones, appelées communautés, on peut les définir comme des ensembles de nœuds fortement liés entre eux, et plus faiblement liés avec le reste du réseau .

L'identification de communautés est un sujet important, puisqu'il peut être rencontré dans plusieurs domaines d'application et des situations du monde réel. L'identification des communautés nous permet également de déterminer le rôle de différents acteurs au sein des communautés et dans le réseau dans sa globalité.

Par ailleurs, beaucoup de travaux portant sur la définition puis la détection de ces communautés ont été effectués durant ces dernières années, Et nous en décrirons certaines dans le courant de la thèse.

L'objectif principal de cette thèse est de concevoir une approche de détection de communautés qui serait stable, précise et efficace même pour des réseaux sociaux . Pour cela, nous avons défini une nouvelle méthode qui fonctionne en deux phases. Durant la première phase, nous détectons tous les circuits afin de décomposer le réseau initial en petits groupes élémentaires. Dans la deuxième phase, nous proposons une procédure qui fusionne ces groupes élémentaires ce qui permet d'identifier les différentes communautés.

L'approche proposée est évaluée sur différents types de réseaux et leur performance est comparée à celle d'autres algorithmes de détection de communautés.

Organisation de la thèse

Les travaux de ce mémoire s'inscrivent dans ce contexte interdisciplinaire, en se concentrant sur la question algorithmique de détection de communautés, c'est -à-dire de groupes d'acteurs fortement liés entre eux et faiblement liés aux autres. Cette thèse est organisée comme suit :

- Chapitre 1 : *les réseaux sociaux.*

  Dans ce chapitre, nous donnerons quelques définitions et la terminologie. Nous aborderons les réseaux sociaux et leur modélisation en graphes, la définition structurelle et sémantique des communautés et nous donnerons des exemples de réseaux sociaux ainsi que leurs communautés d'intérêt correspondantes.

- Chapitre 2 : *Un état de l'art.*

  Dans ce chapitre, nous présentons l'état de l'art la détection de communautés, en couvrant la plupart de ses aspects, tout en nous consacrant plus particulièrement aux avancées les plus récentes du domaine, qui concernent directement la problématique de cette thèse.

- Chapitre 3 : *Méthode proposée*.

  Dans ce chapitre, nous présentons une nouvelle méthode de détection de communautés qui ne nécessite pas la connaissance à priori du nombre de communautés.

- Chapitre 4 *: Évaluation et Expérimentation.*

  Ce Chapitre présente les résultats expérimentaux de la méthode proposés sur des réseaux réels.

- Conclusion :

  Nous terminerons notre thèse par une conclusion, où nous rappellerons nos apports et contributions.

Chapitre 1

Les réseaux sociaux

Sommaire

1.1 Introduction

Les réseaux sociaux sur Internet sont devenus, depuis 2004, un fait qui ne cesse de croître avec les années. Ils permettent aux différents utilisateurs d'interagir en communauté et de se regrouper selon des critères qui leur sont importants. Ces réseaux sont de différents types. Certains sont connus de tous (Facebook, Twitter, Youtube) et comptent des millions de membres. D'autres exploitent des niches moins connues et peuvent passer relativement inaperçus ou rester confidentiels, tels les réseaux d'entreprise.

Enfin, certains des échanges peuvent aussi être assimilés à des réseaux sociaux : c'est le cas des mails et des SMS, qui définissent des échanges entre différents groupes d'individus.

Tous ces réseaux sociaux amassent de très nombreuses données : les amis, les messages, les images, la fréquence d'utilisation... tous ces échanges et informations sont soigneusement enregistrés. Dés lors se pose le problème de l'exploitation de cette masse d'informations.

Les réseaux sociaux sont représentés par des graphes non orientés avec des relations non orientées. Les graphes pondérés sont adaptés aux réseaux sociaux qui contiennent différents niveaux d'intensité dans les relations. Les graphes multipartis sont adaptés pour des réseaux sociaux incluant différents types de ressources manipulées par les acteurs et qui sont le support des interactions.

Dans la première partie, une définition des réseaux sociaux sera abordée, puis dans la seconde partie il sera question de l'analyse des réseaux sociaux. Enfin on va étudier ces propriétés.

1.2 Origines des réseaux sociaux

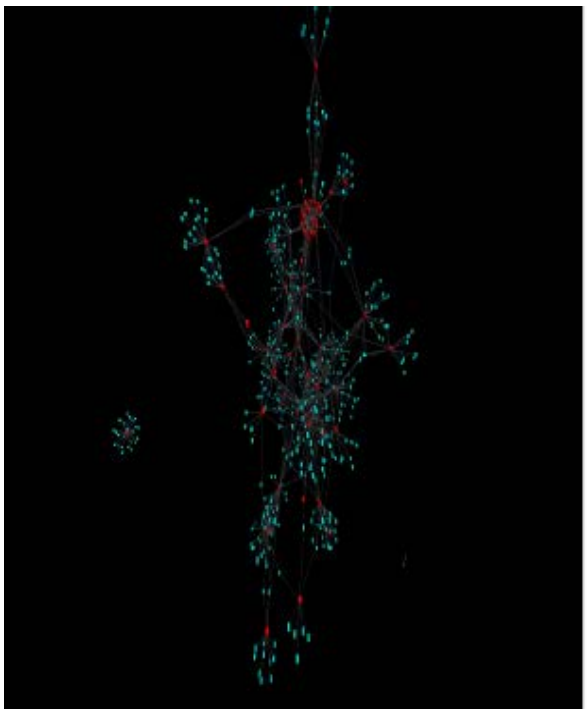
La première personne à avoir représenté un réseau social est ***Jacob Levy Moreno*** au début des années 1930 [Moeno,1933]. Son objectif étant de visualiser graphiquement un réseau social, il a représenté les personnes par des points et une relation entre deux personnes par des flèches.

Au milieu du 20ième siècle, ***Cartwright*** et ***Harary*** sont les premiers à avoir appliqué la théorie des graphes dans l'analyse des réseaux sociaux. Le graphe est devenu par la suite la représentation adoptée par toutes les sciences manipulant l'analyse des réseaux sociaux, dont la sociologie, les mathématiques et l'informatique.

1.2.1 Définition d'un réseau social :

Aujourd'hui, un réseau social est défini comme « une structure définie par des relations entre des individus ». Concrètement, c'est l'ensemble des individus avec qui une personne est en contact. Il s'agit également de liens entre des personnes : les habitants d'un quartier, une famille...

Le but des réseaux sociaux sur Internet est tout d'abord de rencontrer des personnes qui ont des intérêts communs, garder le contact avec ces personnes et enfin de reprendre contact avec des personnes perdues de vue ainsi que de maintenir le lien avec des personnes distantes. Le succès de ces réseaux sociaux sur Internet est essentiellement dû à la rapidité et à la simplicité des échanges. De plus, il répond au besoin d'appartenance de ceux-ci, en étant sur un réseau social, un utilisateur appartient à une communauté, avec des « amis » (Facebook), des « followers » (Twitter). Ainsi, les réseaux sociaux sur Internet se sont multipliés et développés en créant un réel phénomène et en répondant à un besoin humain.



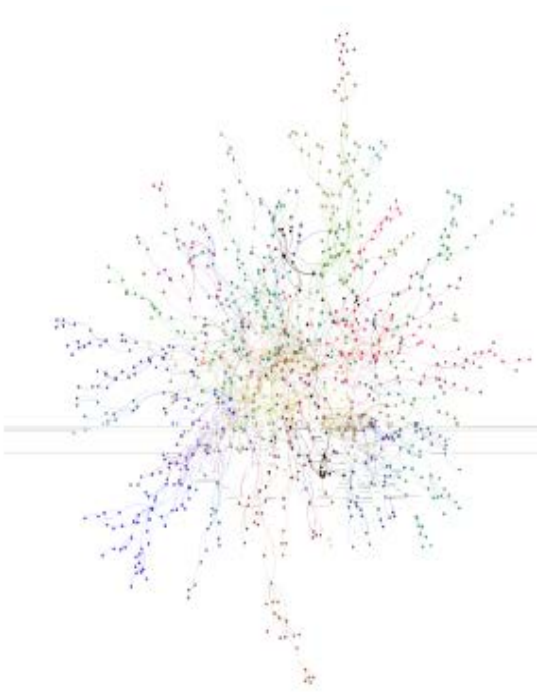


Figure 1.1: Réseau d'achats sur un site de e-commerce [Kanawati,2014]	***Figure 1.2:*** Réseau de collaborations scientifiques extrait de DBLP [Kanawati,2014]

1.2.2 Exemple de réseaux sociaux:

Dans le monde , il existe énormément de réseaux sociaux. On va donner plusieurs exemples de réseaux sociaux les plus connus dans cette section, qu'on va l'utiliser prochainement dans le chapitre d'évaluation.

1.2.2.1 Club de karaté du Zachary [Zachary,1977]

Le réseau de Zachary est un réseau social des membres d'un club de karaté de l'université San Francisco aux Etats Unis. Le club de karaté compte 78 membres. Zachary a fait une étude sur les membres du club qui ont des relations d'amitié en dehors du club. Parmi les 78 membres du club, seuls 34 ont des relations d'amitié. Le réseau d'amitié issu de ce club est représenté dans la figure 1.3 (a). Il est constitué de 34 nœuds représentant les membres du club et 78 liens représentant les amitiés entre les membres.

1.2.2.2 Le graphe de Pages Web [Givan et Newman,2004]

Ce graphe est un réseau social d'indexation de pages web. Les nœuds de ce graphe représentent les pages web et les arcs représentent les hyperliens d'une page à une autre.

Dans ces graphes, les liens sont orientés (généralement, si une page *A* indexe une autre page *B*, le contraire n'est pas systématiquement vrai - la page *B* peut ne pas indexer la page *A*). Le groupe de pages qui ont beaucoup de liens entre elles, peut être une communauté de pages web qui traitent un même sujet. Rassembler les pages du même sujet en communautés, peut faciliter et accélérer le processus d'indexation et de réponse aux requêtes.

1.2.2.3 Les livres de politique Américaine [Newman,2013]

Ce réseau est collecté par V. Krebs et contient des livres sur la politique américaine vendus par le libraire en ligne Amazon.com. Il contient 405 classifiés manuellement en trois groupes notamment "libérales", "neutres", ou "conservateurs". Ces classifications ont été affectées séparément par Mark Newman d'après les descriptions et les commentaires sur les livres sur Amazon. Les 441 liens entre les livres représentent l'achat de livres par des acheteurs .

1.2.2.4 Le graphe du Football Américain [KREBS,2008]

C'est est un réseau de relations entre les équipes du football américain de première division entre des universités aux USA. Les nœuds du réseau représentent les équipes (étiquetes par les noms des collèges) et les liens représentent les matchs entre équipes (un lien est mis entre deux équipes si elles se rencontrent durant la saison). Les équipes sont divisées en conférences de 8 à 12 équipes chacune. Il existe plus de matchs entre les équipes de la même conférence qu'avec les autres. Une équipe joue en moyenne sept matchs dans sa conférence et quatre avec des équipes d'autres conférences (une équipe joue autour de 75% de ses matchs dans sa conférence). Les communautés réelles de ce réseau sont les groupes d'équipes (nœuds) de mêmes conférences.

1.3 Analyse des réseaux sociaux

L'analyse de réseaux sociaux s'appuie sur les acquis de la théorie des graphes [John,1988] pour formaliser le réseau social comme un ensemble de nœuds et de liens où chaque nœud modélise un acteur et chaque lien une relation entre deux acteurs. Une valeur peut être affectée à un lien et représentera alors la force de celui-ci. Elle peut servir à représenter l'importance d'une relation, que ce soit en comptant simplement le nombre d'occurrences de cette relation, ou en prenant en compte d'autres processus de pondération (qualité de l'interaction, système d'évaluation, ...).

A ce titre, l'analyse de réseaux sociaux dispose d'outils mathématiques issus directement de la théorie des graphes mais aussi d'outils et de techniques qui lui sont propres. On retrouvera donc dans l'analyse de réseaux sociaux des terminologies de la théorie des graphes telles que le degré, la force, ou le poids d'un lien.

Les travaux de [Zachary,1977], suivant les relations existantes entre les 34 membres d'un club de karaté sur une période de 3 ans. On trouve notamment dans cet ouvrage un exemple de modélisation en graphe d'un réseau social, les individus étant représentés par des nœuds, les liens entre ces individus étant représentés par des arêtes, pouvant avoir des attributs spécifiques.

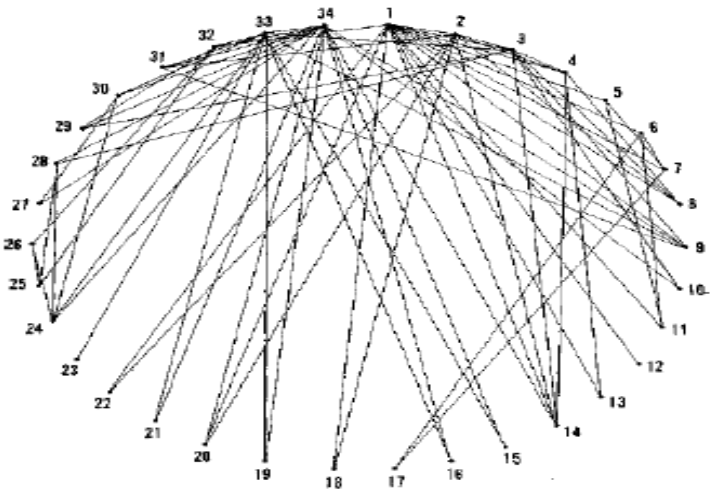


Figure 1.2 **:** Un exemple de réseau social : Réseau d'amitié du Zachary Karaté Club [Zachary,1977]

Au cours de cette étude s'est déroulé un phénomène inattendu qui a été la division de ce club en deux à cause de divergences sur l'organisation du club. Zachary a alors remarqué que cette division s'est caractérisée par une coupure du graphe représentatif des membres du club par sa coupe minimale, séparant ainsi les personnes ayant des opinions différentes sur le problème. Cette particularité illustre l'existence d'une clusterisation de ces graphes : les personnes se groupent en ***communautés***.

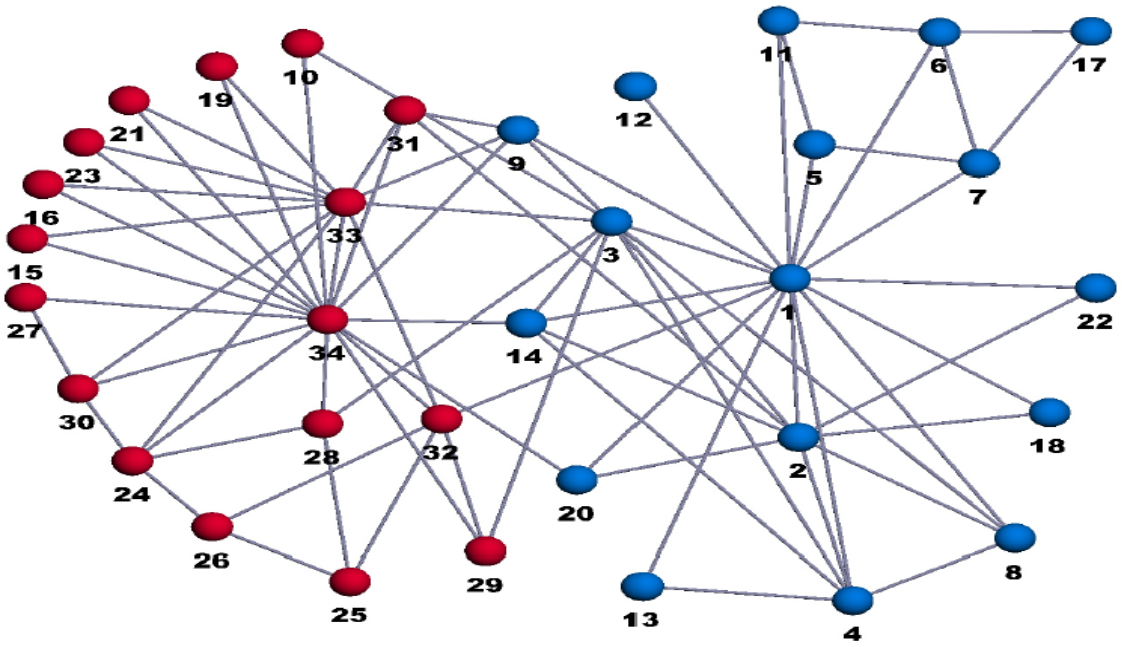


Figure 1.3 **:** Apparition de deux groupes dans le réseau du Zachary Karaté Club **[Nicolas,2012]**

1.4 Propriétés des réseaux sociaux

1.4.1 Définition

- ***Réseau social :***

Structure par laquelle des individus sont liés entre eux par un lien. Un tel réseau est généralement représenté par un graphe dont les nœuds sont les acteurs du réseau et dont les liens illustrent les relations entre ces acteurs.

- ***Communautés, Clusters et Groupes :***

La notion de communautés dans les graphes n'a pas de définition formelle. Cependant, l'existence de zones plus densément connectées que d'autres est le résultat d'une présence de structures de graphes dont les nœuds se sont regroupés en communautés du fait de leur ressemblance ou de leurs intérêts communs. Cette ressemblance ou ce partage d'intérêt peut avoir des interprétations différentes selon la nature et le type du réseau d'interaction considéré (réseaux sociaux, réseaux biologiques, *etc*).

Nous allons donner ici deux définitions des communautés, l'une sémantique et l'autre structurelle.

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

Définition structurelle :*Une communauté est un ensemble de nœuds fortement liés entre eux et faiblement liés avec les autres nœuds du graphe.*

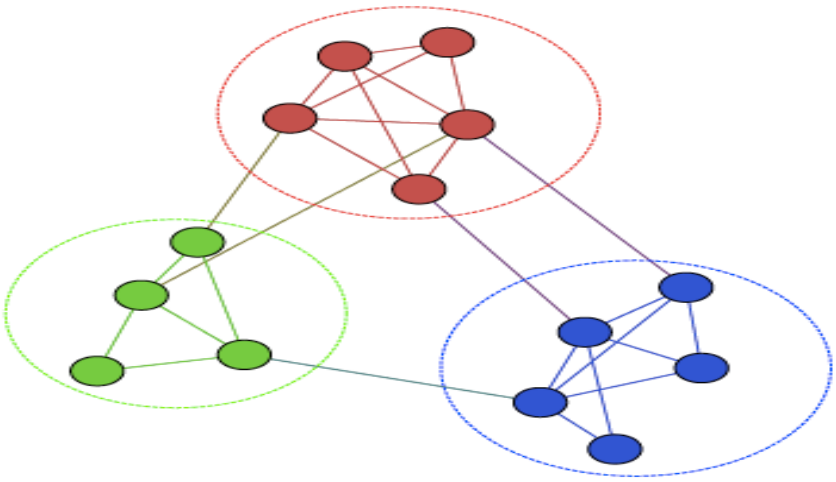


Figure1.4 : Communautés dans un graphe

1.4.2 Caractéristique d'un réseau social

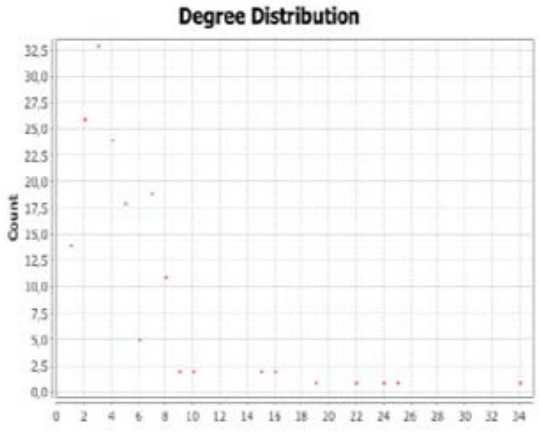
Tout d'abord, il faut savoir que les études sur les graphes de réseaux sociaux ont montré que ces graphes possèdent des caractéristiques particulières.

1.4.2.1 L'Effet petit-monde

La principale caractéristique d'un réseau social est ***«l'effet petit monde»***, mise en évidence historiquement par la fameuse expérience de Milgram [Milgram,1969],et qui exprime le fait que les graphes ont souvent des diamètres très faibles. Le plus court chemin entre deux nœuds dans un réseau social de taille n est de l'ordre de $log\ (n)$. Newman a montré que les individus d'un même réseau social possèdent la faculté de trouver facilement ces plus courts chemins.

1.4.2.2 Distribution hétérogène de degrés :

Une autre caractéristique d'un réseau social est la loi de distribution des degrés de ces graphes. En effet, il a été observé par Price [RS8] puis vérifié plus tard avec expérimentation que dans un graphe représentant les réseaux sociaux, les degrés des nœuds suivent une distribution de type loi de puissance, $P(k) = \alpha\, k^{-\propto}$ avec k les degrés d'un nœud. C'est à dire que, plus on considère un degré élevé, plus le nombre de sommets ayant ce degré dans un même réseau est faible.



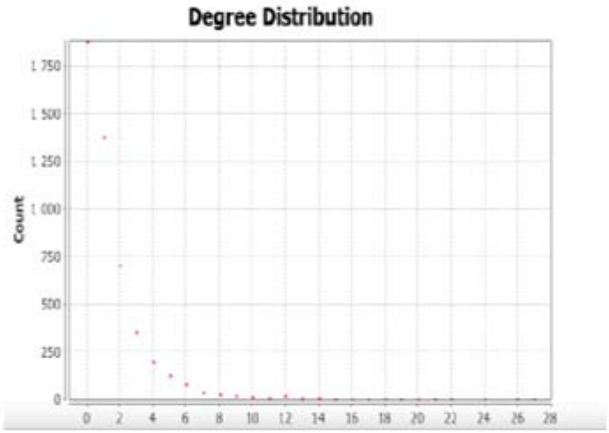


Figure 1.5: Distribution de degrés du Réseau de la Figure 1.1 [Kanawati,2014]	***Figure 1.6:*** Distribution de degrés du réseau de la Figure 1.2 [Kanawati,2014]

1.4.2.3 Un cœfficient de clustering local élevé

En effet, les individus dans un tel réseau ont tendance à se socialiser en se regroupant en ***Communautés***. Si un nœud S_1 est connecté à un autre nœud S_2 qui est lui même connecté à un nœud S_3, alors il y a une forte probabilité pour que S_1 soit aussi connecté à S_3. En d'autres termes, la façon avec laquelle les nœuds sont dans un réseau social favorise l'émergence de structures de graphe de type ***Triangle***. Ces communautés possèdent une forte densité locale et une faible densité globale. Le coefficient de clustering est donné par la formule suivant :

$$\sum \frac{3 \times \Delta}{\nabla}$$

Où Δ est le nombre de triangles dans le graphe et ∇ est le nombre de triades. Noter que dans un graphe aléatoire le coefficient de clustering sera de l'ordre de la probabilité de l'existence d'un lien.

1.4.2.4 Clusterisation :

Un réseau social se démarque également par son haut taux de clusterisation, qui peut se mesurer par la proportion de **« mes »**, amis qui sont amis entre eux. Ce taux, pour un nœud ***i*** est calculé de la manière suivante :

$$C_i = \frac{2e_i}{k_i(k_i-1)},\ C_i \in [0,1]$$

Avec e_i est le nombre de liens entre les voisins de ***i*** et avec k_i le degré de ***i***. Le taux de clusterisation peut également être calculé pour un graphe en effectuant la moyenne de tous les coefficients de chaque nœud :

$$C = \frac{1}{N}\sum_i^N C_i$$

Cette propriété de clustering peut être expliquée par la propriété de similarité (les amis de x lui sont similaires) et par une transitivité de la similarité où si **x** et **y** sont amis avec **z** alors **x** et **z** sont similaires, de même entre **y** et **z** d' où par transitivité **x** et **y** sont similaires ou encore **x** et **y** sont amis. Cette propriété est appelée **homophily**.

1.5 Modélisation des réseaux sociaux

L'intérêt d'une telle modélisation est de pouvoir effectuer des simulations de pannes, d'attaques, de propagation et d'autres évènements qui peuvent survenir sur les réseaux réels. Dans ce qui suit, nous allons essayer de donner quelques présentations de différents modèles de modélisation de réseaux sociaux.

1.5.1 Modélisation d'un réseau social par génération aléatoire

Des modèles de réseaux sociaux ont été élaborés dans le but d'améliorer la compréhension globale de ces réseaux. Le premier modèle proposé est celui de P. Erdös et A. Rényi [Erdös et Rényi ,1960] qui est généré aléatoirement. Ce modèle est aussi simple, et permet effectivement d'obtenir l'existence de courts chemins mais ne permet pas d'obtenir des degrés suivant la bonne distribution de degrés ni de retrouver une clusterisation caractéristique des petits mondes.

En effet, nous pouvons rappeler que les graphes aléatoires ont des degrés suivant une distribution de loi de Poisson qui indique que la plupart des nœuds ont approximativement le même nombre de liens qui ne correspondant pas à ce qui est caractéristique aux réseaux sociaux, même si leur diamètre est bien de l'ordre de ***O(log n).***

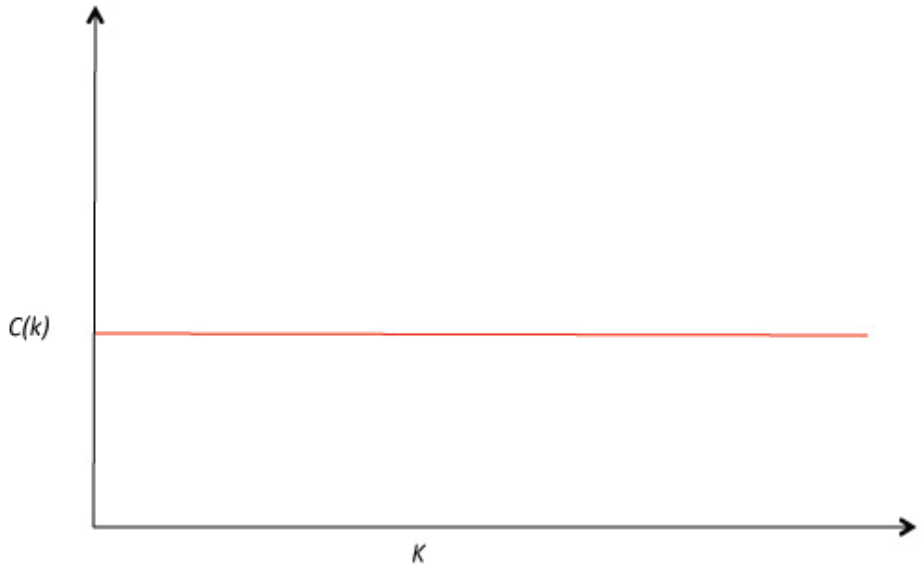


Figure 1.7: Evolution du coefficient de clustering d'un graphe aléatoire

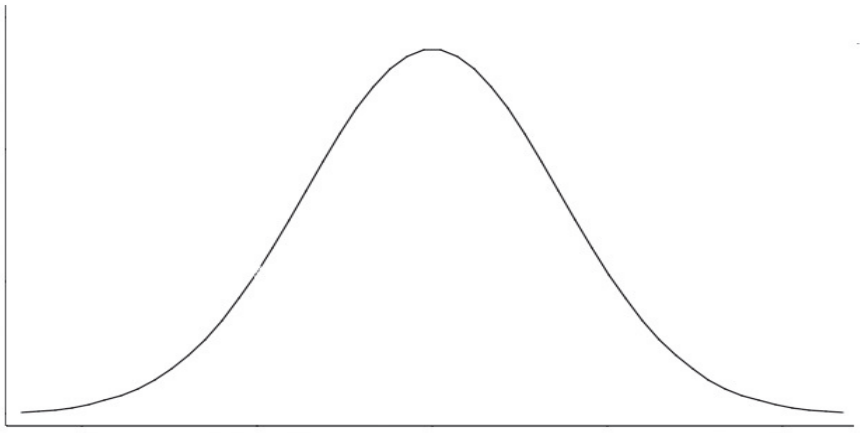


Figure 1.8 : Distribution de degrés d'un graphe aléatoire

1.5.2 Modélisation d'un réseau social par les petits mondes

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, ce sont les caractéristiques d'un graphe de petits mondes.

1.5.2.1 Modèle de Watts & Strogatz [Watts et Strogatz,1965]

Le modèle proposé par Watts & Strogatz consiste à générer des graphes petits mondes (*Figure 8*). Partant d'un anneau régulier à $\boldsymbol{n}$ nœuds où chaque nœud est relié à ses $\boldsymbol{2k}$ plus proches voisins ($\boldsymbol{k}$ voisins de chaque côté). Le coefficient de clustering d'un nœud $\boldsymbol{u}$ de l'anneau régulier est assez important : $\boldsymbol{C(u) = \frac{3(k-2)}{4(k-1)}}$ et la distance moyenne, dans un anneau régulier, est elle aussi très élevée.

Le principe des auteurs est de parcourir chacun des nœuds de ce graphe et de choisir, pour chaque nœud , un lien que l'on va, ou non, reconnecter avec une autre extrémité selon une probabilité $\boldsymbol{p}$ à définir.

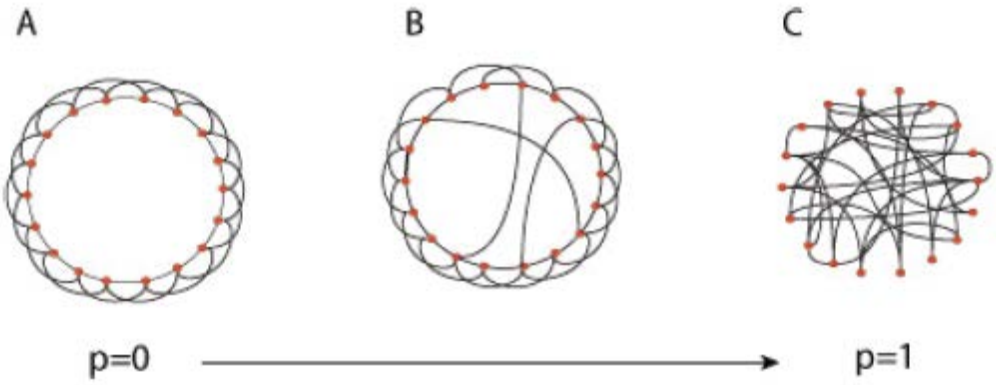


Figure 1.9 :Modèle de Watts et Strogatz. [Watts et Strogatz,1965]

A : *p=0*, uniquement les nœuds voisins proches sont connectés. Le coefficient de clustering et la distance moyenne entre n'importe quels deux nœuds sont grands.

B : *p < 1* et *p > 0*, peu de courts chemins sont introduits dans les réseau, ce qui réduit la distance moyenne.

C : *p* est grand et le réseau est équivalent à un graphe aléatoire. Dans ce cas, la distance moyenne et le coefficient de clustering sont trop faibles.

1.5.2.2 Modèle de Kleinberg [Duncan,1998]

L'idée de Kleinberg est de disposer ***n*** individus sur une grille de dimension ***k*** et de connecter chaque individu avec, d'une part, leurs voisins les plus proches mais aussi avec d'autres personnes éloignées avec la probabilité de $\mathbf{Pr}[\boldsymbol{u} \rightarrow \boldsymbol{v}]\boldsymbol{\alpha} \frac{\mathbf{1}}{\boldsymbol{d(u,v)^{\alpha}}}$

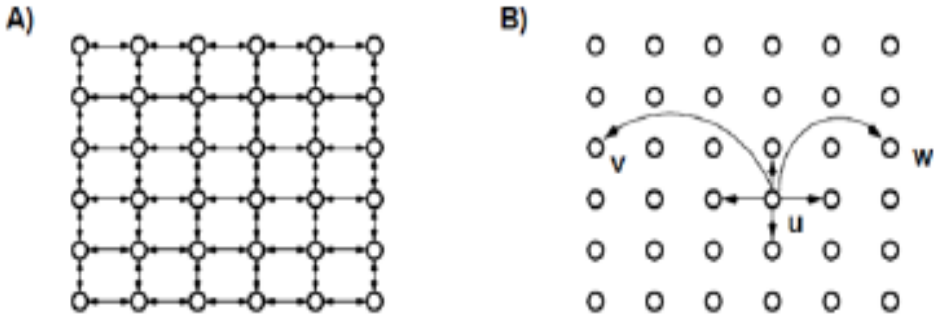


Figure 1.10: Construction d'un graphe selon le modèle de Kleinberg

[Kleinberg,2000]

1.6 Conclusion

Dans ce chapitre, Nous avons introduit les concepts de réseaux sociaux, et les 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. Nous avons terminé le chapitre par la définition de ces réseaux, de définir leurs propriétés et de citer les modélisations des réseaux sociaux d'interaction les plus utilisées.

Dans le chapitre suivant en va présenter une brève revue de littérature sur la détection de communautés. Parmi les nombreuses approches proposées, nous allons retenir celles ayant reçu le plus d'intérêt de la part de la communauté scientifique.

Chapitre 2

Etat de l'art

Sommaire

2.1 Introduction

Le problème de la détection de communauté dans les réseaux est un sujet relativement récent, mais qui a très rapidement conduit à une grande quantité de travaux.

Lorsque l'on étudie des réseaux de terrain, de grande taille et/ou représentent des données complexes tel que les réseaux sociaux, le nombre de groupes que l'on cherche à obtenir ne peut être connu à l'avance. Ce qui avons à nous intéresser à un autre problème plus complexe, celui de la détection de communautés.

On peut définir le problème de la manière suivante : pour un réseau donné, comment le décomposer en un nombre inconnu de groupes de nœuds de manière à ce que ces groupes de nœuds satisfasse efficacement le problème de la minimisation des liens inter-communautés, et la maximisation des liens intra-communautés.

Ce chapitre présente une brève revue de littérature sur la détection de communautés. Comme il existe de nombreuses approches proposées, nous allons retenir celles ayant le plus d'intérêts de la part de la communauté scientifique. Ces approches illustrent aussi la diversité de méthodologies et donnent une vue d'ensemble des techniques proposées selon leurs principes méthodologiques.

2.2 Classification des approches de détection de communautés

L'existence dans les grands graphes de terrain de zones plus densément connectés que d'autres constitue une des caractéristiques non triviales que l'on retrouve dans de nombreux cas. Ces zones sont appelées communautés (par analogie avec les réseaux sociaux) et correspondent intuitivement à des groupes de nœuds plus fortement connectés entre eux qu'avec les autres nœuds.

Les méthodes de détection de communautés ont fait l'objet de nombreux travaux. Lorsque l'on étudie des réseaux de grande taille et/ou représentent des données complexes (réseaux sociaux, réseaux biologiques, etc.), le nombre de communautés et leurs tailles sont inconnus et, le plus important, est de pouvoir reconnaître les réseaux qui ne possèdent pas une structure modulaire.

Nous allons lister ici quelques approches qui ont été proposées à ce jour.

2.2.1 Les approches statiques, sans recouvrement

Aujourd'hui, décomposer un graphe en communautés consiste souvent à partitionner l'ensemble des nœuds et la qualité de la décomposition est ensuite évaluée par une fonction de qualité. La fonction de qualité la plus utilisée est la modularité définie par Girvan et Newman de 2002 qui se calcule comme la différence entre la proportion de liens internes aux communautés et la proportion de liens qu'auraient des communautés aléatoires de même taille.

2.2.1.1 Les approches hiérarchiques

Les méthodes hiérarchiques partent d'une structure dans laquelle chaque nœud est identifié comme une petite communauté. On calcule des distances entre communautés et on fusionne les deux communautés les plus proches en une nouvelle communauté. Le nombre de communautés est réduit de 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 représentée sous une forme arborescente appelée dendrogramme.

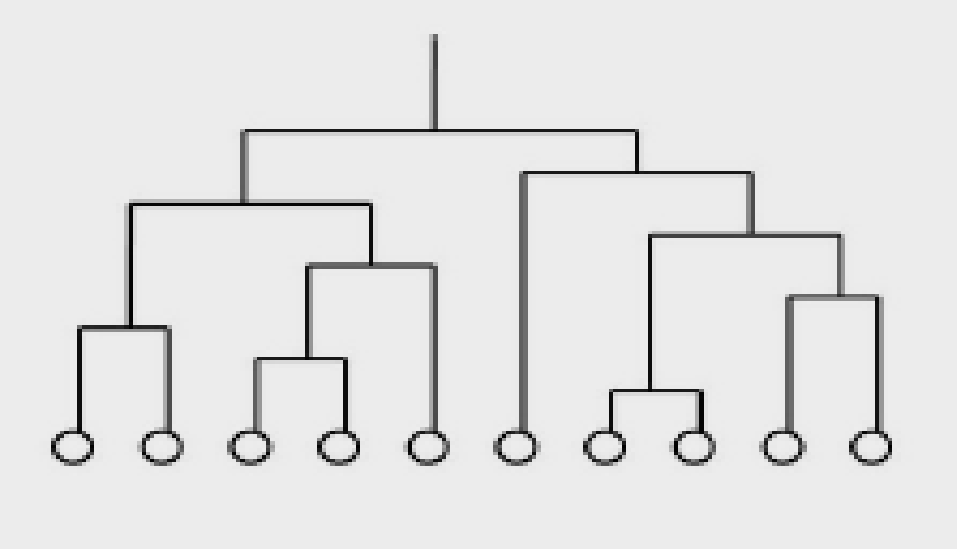


Figure 2.1 **:** Exemple d'un dendrogramme

Les méthodes hiérarchiques peuvent être divisées en deux : les méthodes hiérarchiques ascendantes (agglomératives) et les méthodes hiérarchiques descendantes (séparatives). Les deux méthodes sont basées sur la définition d'une mesure de similarité entre chaque paire de nœuds.

A. Approches hiérarchiques ascendantes (agglomeratives)

Le principe global de ces méthodes consiste à regrouper les nœuds itérativement en communautés. Au début, chaque nœud constitue une communauté à part (il y a autant de communautés que de nœuds). Les communautés sont fusionnés deux à deux jusqu'à avoir une grande communauté représentant l'ensemble des nœuds du graphe. A 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. Le résultat de ces approches est un dendrogramme reprenant l'historique des jointures de communautés à chaque étape.

- ***L'algorithme glouton de Newman*** [Newman,2004]

Il s'agit d'un algorithme d'optimisation de la modularité, cette métrique permet de chercher directement le découpage en communautés correspondant à la valeur maximale de la modularité pour un graphe donné.

Initialement, chaque nœud est une communauté. Pour toutes les paires de communautés voisines, la modification de la modularité en cas de fusion est calculée et les deux communautés qui apportent le gain le plus important sont réunies dans une seule. La métrique utilisée dans cet algorithme est la *modularité* qui est calculée à chaque jointure de deux communautés *i* et *j*:

$$\Delta Q = 2(e_{ij} - a_i^2)$$

Où, e_{ij} est la proportion d'arêtes à l'intérieur des communautés et $a_i$la proportion d'arêtes attendue dans le graphe aléatoire de G.

Le calcul est effectué de manière itérative jusqu'au moment où aucun gain n'est plus possible. L'inconvénient principal de cet algorithme est sa complexité qui est de l'ordre de $O(n^3)$.

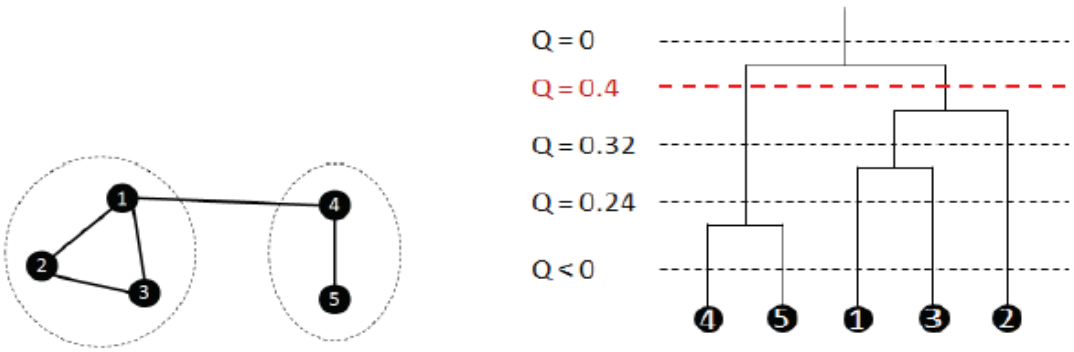


Figure 2.2 : Un graphe à deux communautés avec le dendrogramme résultant de l'application d'un algorithme d'optimisation de la modularité **[Lemmouchi,2010]**

- ***L'algorithme Lauvin*** [Blondel,2008]

Comme dans la version de Newman, au début chaque nœud est mis dans une communauté différente. Les nœuds sont parcourus dans un ordre aléatoire et, pour chacun d'entre eux, on regarde si le placement dans la communauté d'un de ses voisins apporte un gain en modularité. Si c'est le cas, le nœud est déplacé dans cette communauté, sinon il reste dans l'ancienne communauté. Lorsqu'aucun gain n'est plus possible, la deuxième étape de l'algorithme commence. Un nouveau graphe est créé, dont les nœuds correspondent aux communautés de l'étape précédente, et la première étape est répétée sur ce graphe. Les itérations s'arrêtent au moment où une nouvelle étape n'apporte plus une croissance de la modularité.

La méthode de Louvain possède des avantages significatifs vis-à-vis d'autres méthodes de détection de communautés, notamment sa rapidité évaluée à ***O(n log n)***qui lui permet de traiter des graphes ayant jusqu'à plusieurs milliards de liens, son aspect multi échelle, qui lui permet de découvrir des communautés à différentes échelles, et son excellente précision par rapport à d'autres méthodes gloutonnes.

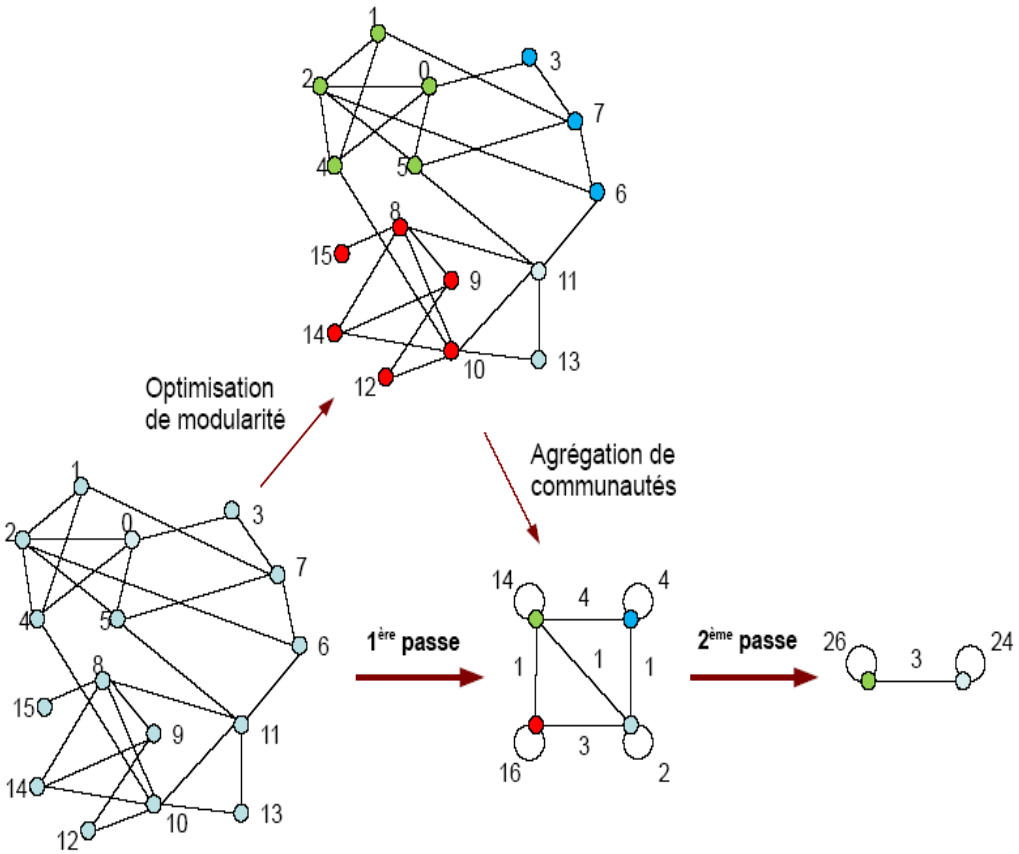


Figure 2.3 : Exemple d'exécution de la méthode de Louvain **[Blondel,2008]**

B. Approches hiérarchiques descendantes (séparatives)

Les approches séparatives divisent le graphe en plusieurs communautés en retirant progressivement les arêtes reliant les communautés distinctes. Les arêtes sont retirées une à une, et à chaque étape les composantes connexes du graphe obtenu sont identifiées à des communautés. Ce processus est répété jusqu'au retrait de toutes les arêtes. On obtient alors une structure hiérarchique de communautés (dendrogramme), comme pour les méthodes hiérarchiques ascendantes. Les méthodes existantes diffèrent par la façon de choisir les arêtes à retirer. Dans la suite, nous citons quelques méthodes les plus connues dans la littérature de cette classe.

- ***L'algorithme de* la centralité d'intermédiarité** [Givan et Newman,2002]

C'est la méthode séparative la plus classique qui introduit une mesure de centralité appelée centralité d'intermédiarité des liens **(Edge-Betweenness Centrality)**. Cette mesure est similaire à la centralité d'intermédiarité d'un nœud et, pour un lien donné (*u*; *v*) peut être calculé ainsi :

$$C_B\big((u,v)\big) = \sum_{i,j,i\neq j} \frac{\sigma_{ij}\big((u,v)\big)}{\sigma_{ij}}$$

Où σ_{ij} ((*u*; *v*)) est le nombre de plus courts chemins allant de i à j passant par le lien (*u*; *v*) et σ_{ij} le nombre total de plus courts chemins allant de *i* à *j*.

Cette mesure définie comme étant le nombre des plus courts chemins du graphe qui passent par cette arête. Il existe en effet peu d'arêtes reliant les différentes communautés et les plus courts chemins entre deux nœuds de deux communautés différentes ont de grandes chances de passer par ces arêtes. L'algorithme proposé calcule la centralité d'intermédiarité pour chaque lien du graphe, puis enlevé le lien possédant la plus forte centralité d'intermédiarité. Ensuite, la centralité d'intermédiarité pour tous les liens du graphe résultant est recalculée et le processus est itéré jusqu'à ce que tous les liens aient été enlevés et on aura autant de communautés que de nœuds dans le graphe.

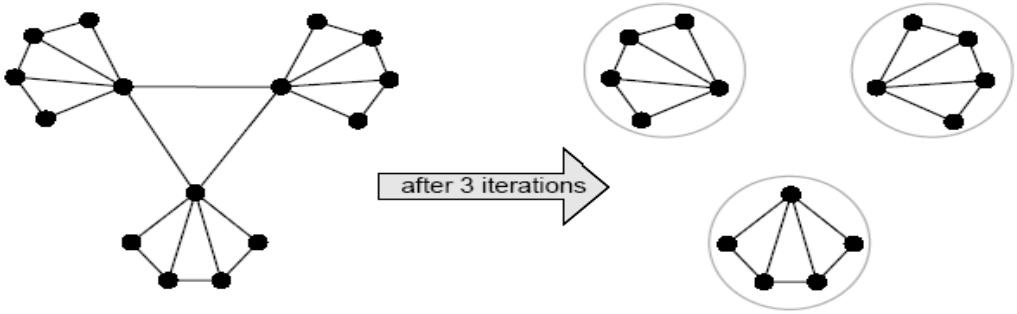


Figure 2.4 : Détection de communautés par la méthode de Edge-Betweenness **[Tang,2010]**

Cependant, cet algorithme nécessite un calcul des centralités d'intermédiaritée coûteux en temps et possède une complexité en $O(m^2n)$ soit $O(n^3)$ pour les graphes creux. Il n'est donc exploitable que sur de petits graphes.

- ***L'algorithme de Radicchi et al.*** [Radicchi,2004]

Les auteurs ont défini une autre métrique pour détecter les arêtes inter-communautés à supprimer. Les auteurs s'inspirent de la notion du coefficient de clustering pour donner des poids aux arêtes. La fonction de score d'une arête est calculée en divisant le nombre de triangles construits par cette arête sur le nombre maximum de triangles possibles. Cet algorithme retire donc à chaque étape l'arête de plus faible clustering.

L'avantage de cet algorithme est que le calcul de ce coefficient requiert des calculs locaux seulement (seul le score des arêtes voisines sera recalculé),contrairement à la centralité d'intermédiarité. L'algorithme est moins coûteux en temps, mais ne donne pas de résultats assez satisfaisants. La complexité de calcul de l'algorithme est de l'ordre de $O(n^2)$.

- ***L'algorithme de Fortunato et al.*** [Fortunato,2004]

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 l'efficacité du réseau due à la suppression de cette arête. L'efficacité d'une arête $e(i, j)$ est définie comme l'inverse de la distance entre le nœud i et le nœud j. Pour toute arête, on calcule l'efficacité puis la diminution du réseau due à la suppression de cette arête, et on affecte cette diminution au score de l'arête.

Cette approche donne de meilleurs résultats que l'approche de Girvan et Newman mais le calcul de l'efficacité du réseau est coûteux en temps. La complexité de cet algorithme est de $O(m^3n)$.

2.2.1.2 Approches utilisant des marches aléatoires :

Les marches aléatoires dans les graphes sont des processus aléatoires dans lesquelles un marcheur est positionné sur un nœud du graphe et peut à chaque étape se déplacer vers un des nœuds voisins. L'idée est que, si on commence une marche aléatoire à partir du nœud i, la probabilité P^t_{ij} de se trouver au nœud j après t pas est plus grande si i et j sont dans la même communauté (si t ne dépasse pas une certaine limite).

Notons que beaucoup des algorithmes présentés dans cette section sont aussi des algorithmes agglomératifs qui auraient pu être classés dans la section précédente.

- ***L'algorithme Walktrap*** [Pons et Latapy,2005]

Cet algorithme est basé sur l'idée qu'une marche aléatoire partant d'un nœud a plus de probabilité à rester piégée pendant un certain temps dans la communauté du nœud de départ. Supposons que nous effectuons une marche aléatoire courte sur le graphe partant d'un nœud ***i,*** la probabilité d'accéder à chacun de ces voisins en une étape est de $1/|P_i|$. On peut donc calculer de la même manière, la probabilité de se trouver au sommet ***j*** en partant de ***i*** après avoir effectué aléatoirement ***k*** pas. Cette probabilité permet de définir une distance entre les paires de nœuds du graphe dans laquelle deux nœuds ***i*** et ***j*** sont proches si leurs vecteurs de probabilité d'atteindre les autres nœuds sont similaires. Une fois ces probabilités calculées pour toutes les paires de nœuds, l'algorithme les utilise pour partitionner le graphe par l'intermédiaire d'une méthode de clustering hiérarchique. Commençant par ***n*** communautés ne contenant chacune qu'un seul nœud, l'algorithme cherche les deux communautés les plus proches, les fusionne, recalcule les distances, puis effectue une nouvelle fusion et ainsi de suite, jusqu'à n'obtenir qu'une seule communauté recouvrant tout le graphe.

L'inconvénient de cet algorithme est le paramètre ***k*** qui représente le nombre de pas ou de liens entre les nœuds . En effet, l'algorithme doit fixer à l'avance cette valeur. La complexité de *Walktrap* est de ***O(nmH)*** où *m* est le nombre de liens, *n* le nombre de nœuds et *H* est la hauteur du dendrogramme.

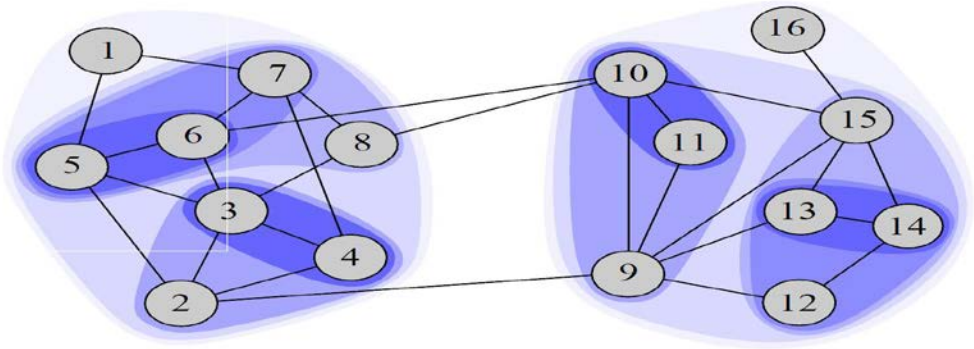


Figure 2.5: Structure hiérarchique de communautés trouvée par Walktrap **[Pons&Latapy,2005]**

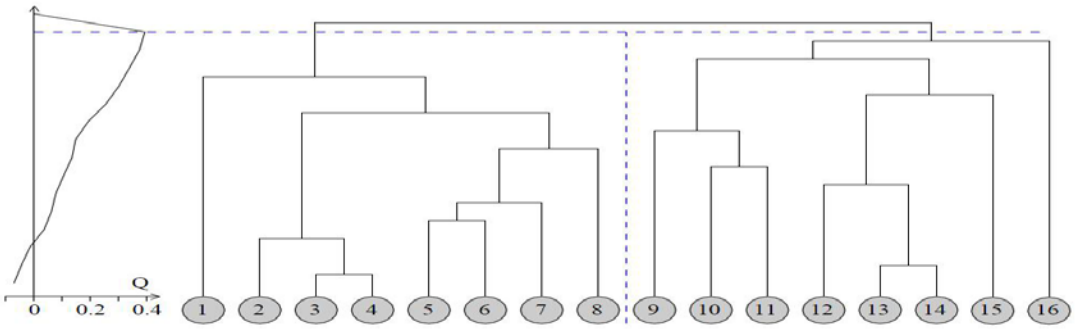


Figure 2.6: Dendrogramme associé aux communautés de la ***Figure2.5*** trouvées par Walktrap

- ***L'algorithme Infomap*** [*Rosvall et Bergstrom*,2008]

Son principe est que si l'on considère un marcheur aléatoire qui se déplace sur un réseau, et que ce réseau a une structure en communauté, alors le marcheur aléatoire va avoir tendance à rester à l'intérieur des communautés. Ce principe découle de la définition intuitive des communautés, selon laquelle elles sont des groupes de nœuds fortement connectés et plus faiblement connectés au reste du réseau. Une conséquence intéressante de cette définition est que si le graphe étudié est un graphe aléatoire, un marcheur aléatoire ne restera pas «coincé» dans une communauté, et InfoMap sera donc capable de dire que le graphe n'est pas divisible en communautés pertinentes, contrairement aux méthodes basées sur une optimisation de la modularité qui trouvent toujours des communautés quel que soit le graphe étudié.

Infomap prouve son efficacité démontrée sur des réseaux de terrain par sa rapidité lui permettant de traiter de très grands graphes, ainsi que par sa possibilité d'évaluer la qualité du découpage fourni.

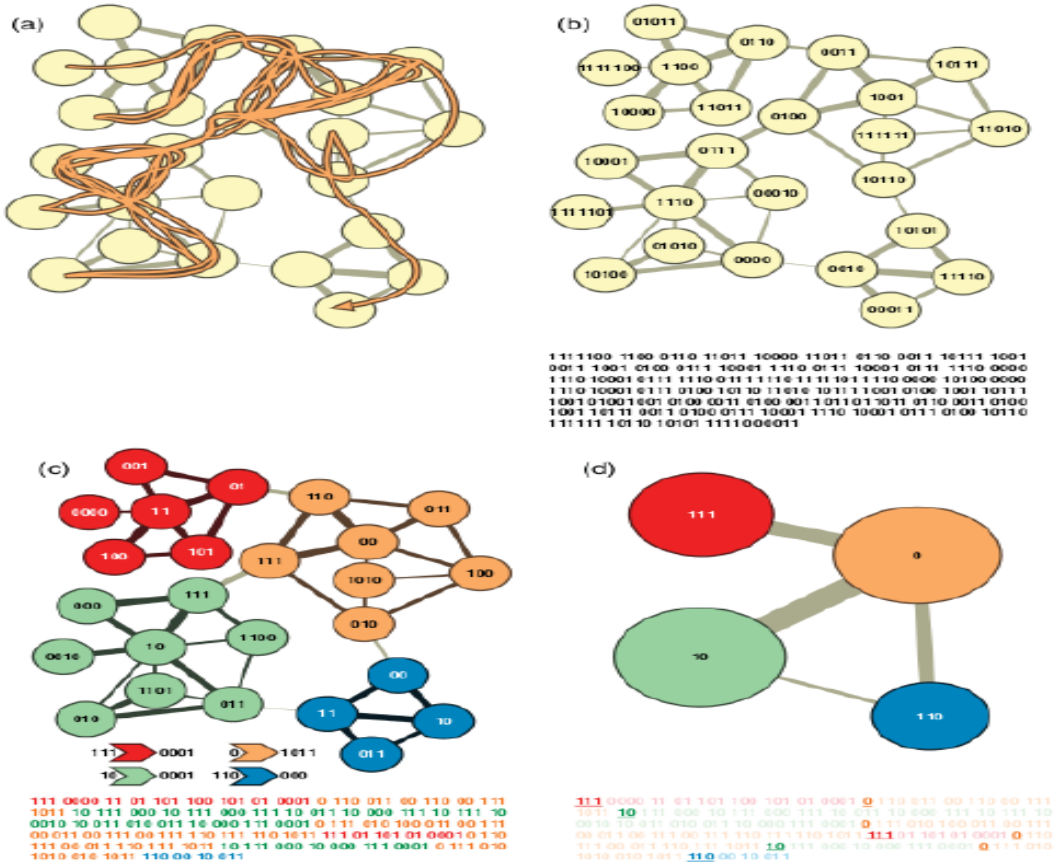


Figure 2.7 : Illustration de fonctionnement d'Infomap **[*Rosvall et Bergstrom*,2008]**
a) Représentation de parcours d'un marcheur aléatoire que l'on souhaite encoder
b) Un codage binaire de l'ensemble des parcours avec le moins de caractères possibles**.**
c) Un codage à deux niveaux (sur les groupes puis sur les nœuds) permettant à un même code d'être utilisé pour des nœuds différents.
d) Une visualisation des groupes sous forme de graphe quotient.

- ***L'algorithme Markov Cluster*** [Dongen,2000]

Cette approche calcule des probabilités de transition entre tous les sommets du graphe en partant de la matrice de transition des marches aléatoires. Deux opérations matricielles sont successivement itérées. La première calcule les probabilités de transition par des marches aléatoires de longueur fixée ***r*** et correspond à une élévation de la matrice à la puissance ***r***. La seconde consiste à amplifier les différences en augmentant les transitions les plus probables et en diminuant les transitions les moins probables. Les transitions entre sommets d'une même communauté sont alors favorisées et les itérations successives des deux opérations conduisent à une situation limite dans laquelle seules les transitions entre sommets d'une même communauté sont possibles. La complexité totale de l'algorithme est en $O(n^3)$

2.2.1.3 Approches spectrales :

Elles consistent à calculer le vecteur propre correspondant à la plus petite valeur propre non nulle de la matrice Laplacienne du graphe, $L = D - A$. L'inconvénient de ces algorithmes vient de sa complexité de $O(n^3)$ et obtient 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.

- ***L'algorithme de Donetti et Munoz*** [*Donetti et Munoz*,2004]

Donetti et Munoz a proposé un algorithme de détection des communautés basé basé sur l'analyse des vecteurs propres de **L** .

Au lieu d'utiliser un seul vecteur propre, ils utilisent les vecteurs propres associés aux **D** plus petites valeurs propres non nulles. Chaque nœud est représenté dans un espace de dimension **D** en utilisant les composantes qui lui correspondent dans les **D** vecteurs. Pour trouver les communautés, les auteurs utilisent un algorithme hiérarchique basé sur la distance angulaire entre les nœuds dans cet espace. Le meilleur partitionnement du dendrogramme est celui qui a la plus grande modularité.

Pour des structures très modulaires, l'algorithme de Donetti et Munoz donne de bons résultats même pour D = 1 ou D = 2 (voir figure 07) mais, dans le cas général, la qualité du partitionnement augmente avec D.

L'inconvénient de l'algorithme vient de sa complexité de $O(n^3)$ et du fait que le nombre D de vecteurs propres qu'on doit analyser n'est pas connu à priori.

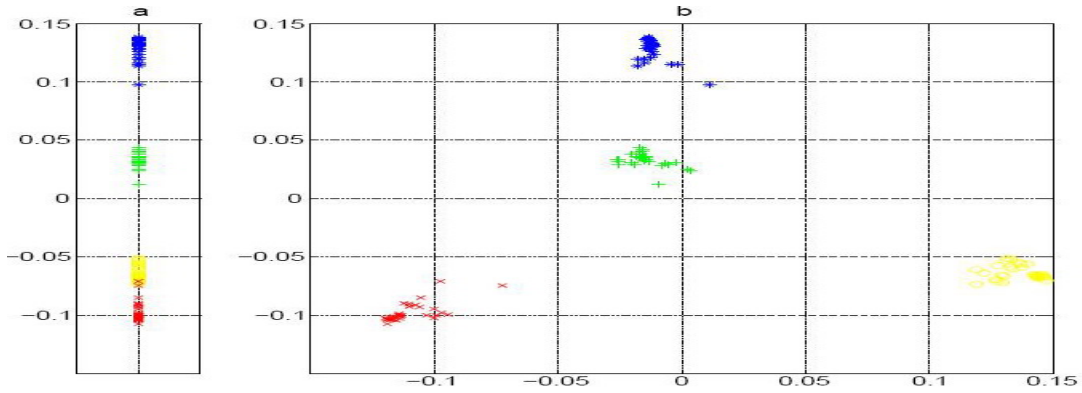


Figure 2.8 : Résultat de l'algorithme de Donetti et Munoz pour un graphe de 04 communautés ***a)*** pour **D=1*b)*** pour **D=2 [*Donetti et Munoz*,2004]**

2.2.1.4 Autres approches

- ***L'algorithme LICOD*** [Kanawati,2011]

Le principe de l'algorithme Licod est quelles communautés se forment autour de nœuds spécifiques appelés leaders. Licod se résume en trois étapes :

1. Identification de l'ensemble des Leaders L. Tout nœud ayant une centralité supérieur à celle de ces voisins est déclaré leader. Après, l'ensemble des leaders est réduit en un ensemble de communautés de Leaders. Deux leaders sont regroupés s'ils ont un nombre de voisins communs élevé.

2. Pour chaque nœud un vecteur de préférence est formé. Les communautés sont triées en ordre décroissant, et le degré d'appartenance d'un nœud à une communauté appartient à l'ensemble, il est simplement donné par la distance minimale entre ce nœud et tout l'ensemble de communautés.

3. Une phase d'intégration où le vecteur de préférence d'un nœud est fusionné avec ceux de ses voisins directs. Ceci permet de favoriser la classe dominante dans l'ensemble des nœuds voisins.

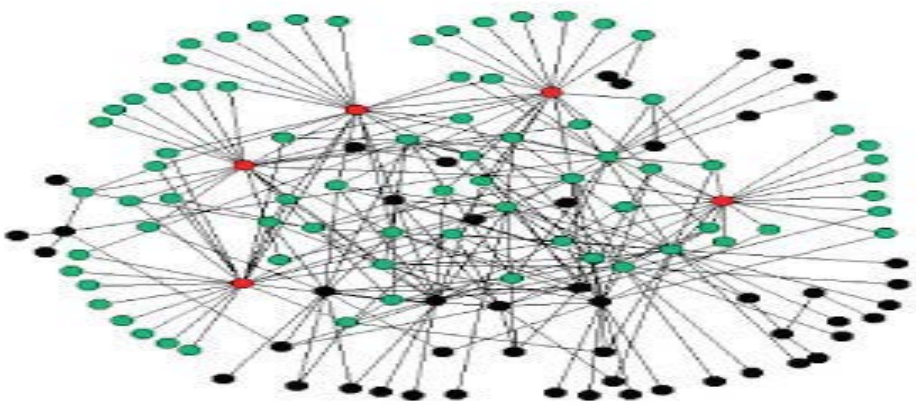


Figure 2.9: Exemple de réseau contenant des nœuds Leaders

- ***L'algorithme de Reichardt et Bornholdt*** [Reichardt,2004]

Cette approche est basée sur un modèle physique de Potts. Chaque nœud est caractérisé par un spin prenant ***q*** valeurs possibles, et les communautés correspondent aux classes de nœuds ayant des valeurs de spin égales. Une énergie du système est définie et doit être minimisée par recuit simulé. Cette minimisation favorise les paires de nœuds liés par une arête et possédant le même état de spin tout en pénalisant les trop grandes classes de spins. Le nombre ***q*** de spins possibles correspond au nombre maximal de communautés que l'on peut trouver et doit être choisi de manière à ce qu'il soit supérieur au nombre effectif de communautés.

2.2.2 Les approches statiques, avec recouvrement

Le recouvrement de communautés est une propriété, observée depuis longtemps dans les réseaux sociaux, et présente dans la plupart des réseaux de terrain. Dans un réseau d'individus dans lesquels les liens ne sont pas limités à un type précis d'interaction (comme on peut le trouver dans les réseaux sociaux Web 2.0 tels que Facebook), il est évident que chaque personne appartient à plusieurs groupes, ou communautés : les personnes de sa famille, ses collègues de travail, ses amis de l'université, etc. Il est probable qu'au moins l'une des personnes de sa famille fasse aussi partie de son cercle d'amis de l'université, ou autre appartenance multiple.

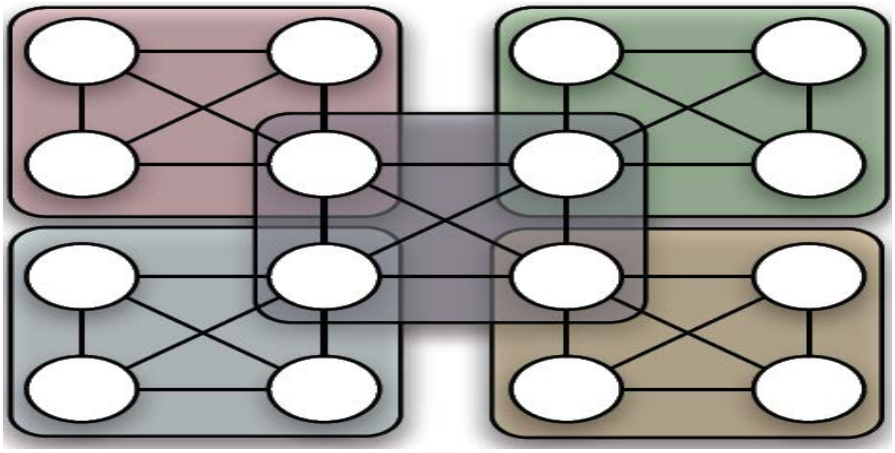


Figure 2.10 : Exemple d'un graphe avec communauté recouvrante

De nouvelles solutions sont proposées prenant en compte efficacement le recouvrement , et on va présenter quelques méthodes publiées à ce jour.

2.2.2.1 Approches basées sur des cliques:

Une communauté est définie comme une chaîne de ***k-cliques*** adjacentes. Une ***k-clique*** est un sous-ensemble de ***k*** nœuds tous adjacents les uns aux autres, et deux ***k-cliques*** sont adjacentes si elles partagent ***k-1*** nœuds. L'idée de ces algorithmes est que, à partir de ***k-cliques***, de construire petit à petit les communautés. L'avantage immédiat d'une telle approche est la détection de communautés avec recouvrement, un nœud pouvant appartenir à plusieurs ***k-cliques*** non forcément adjacentes**.**

- ***L'algorithme CFinder*** [Palla,2007]

L'algorithme est structuré en trois principales étapes :

1. Calculer l'ensemble de cliques de taille ***k***(paramètre de l'algorithme) dans le graphe cible ***G***.

2. Construire un graphe de cliques ou chaque clique est représentée par un nœud. Deux nœuds sont connectés par un lien si les deux cliques associées partagent k-1 nœuds dans le graphe G.

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

Une limite de cet algorithme est qu'il nécessite un paramétrage : la valeur de ***k*** (la taille des communautés à considérer).

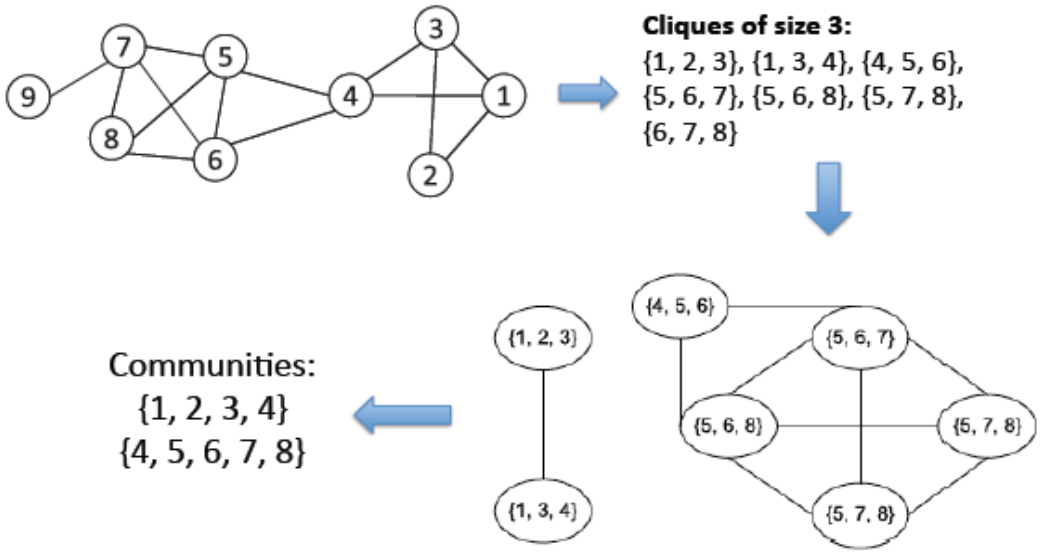


Figure 2.11 : Exemple de l'algorithme de percolation de cliques avec ***k***=3
[Tang et Liu, 2010]

De plus, cette méthode est sensible à certaines configurations. Par exemple, si l'on imagine une suite de cliques de taille k, ayant k - 1 nœuds en commun, et formant une

chaine, notre méthode les détectera comme une communauté, alors que cela n'est généralement pas pertinent.

- ***L'algorithme EAGLE*** [Shen,2009]

Il s'agit d'un algorithme qui utilise un dendrogramme de cliques. Il commence par identifier toutes les cliques maximales qui sont les communautés initiales. Ensuite, les communautés ayant le plus fort taux de similarité sont fusionnées, formant de nouvelles communautés, qui, à leur tour, pourront être fusionnées avec des communautés semblables.

La coupe optimale du dendrogramme est déterminée en utilisant une version modifiée de la modularité, définie comme :

$$Q_{ov}^{E} = \frac{1}{2m} \sum_{c} \sum_{i,j \in c} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \frac{1}{O_i O_j}$$

Avec O_i est le numéro de communauté de nœud i.

La complexité estimée de cet algorithme est de $O(n^2+(h+n)s)$, avec s est le nombre de cliques maximal , et h est le nombre de paires des cliques maximales en voisins.

2.2.2.2 Approches basées sur la propagation de labels

Ces algorithmes nécessitent de parcourir tous les nœuds du graphe, et les partitionnements résultats dépendent beaucoup de l'ordre dans lequel on parcourt ces nœuds. Cela est vu comme un avantage par les auteurs qui suggèrent de faire tourner l'algorithme plusieurs fois et de détecter ainsi les nœuds qui appartiennent à plusieurs communautés.

Les résultats de ces algorithmes doivent donc être pris avec précaution si le réseau analysé a une structure avec des communautés de tailles hétérogènes, ce qui semble être le cas dans beaucoup de réseaux sociaux.

- ***L'algorithme de Raghavan et al.*** [Raghavan,2007]

C'est le premier algorithme qui implante l'idée de propagation de labels. C'est un algorithme itératif où à chaque itération un nœud envoie son label à ses voisins directs, et reçoit ceux de ses voisins. Chaque nœud détermine le label majoritaire qu'il adopte pour l'itération suivante. Ce processus itératif mène à un accord sur un label précis pour chaque groupe de nœuds.

L'avantage de cet algorithme est qu'il est le plus performant en pratique, avec une complexité de $O(n)$, mais il peut y avoir un problème de convergence lié à un échange infini de label entre deux nœuds.

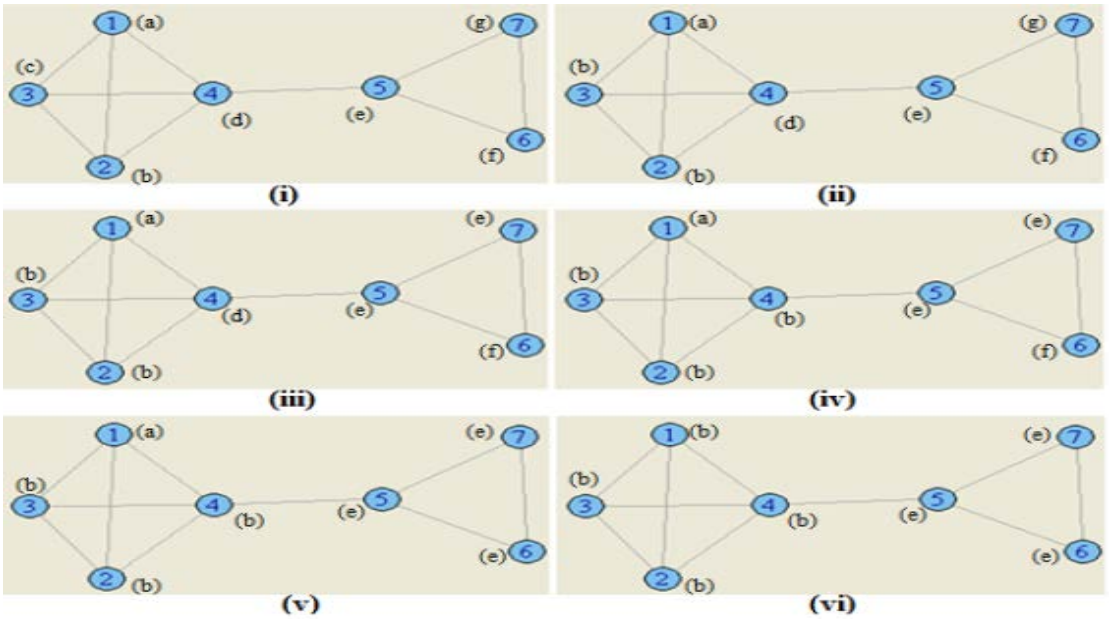


Figure 2.12: Exemple d'exécution de l'algorithme *Label Propagation*. **[Talbi,2013]**

- ***L'algorithme COPRA*** [Gregory,2010]

Il propose une adaptation de la méthode propagation de labels aux cas avec recouvrement .

Pour ce faire, il propose, de ne plus choisir seulement le label le plus courant chez ses voisins, mais de maintenir une liste des labels les plus courants dans son entourage. Un paramètre de l'algorithme fixe le nombre maximum de labels qu'un nœud peut retenir (sans quoi chaque label s'étendrait à l'infini).

La limite de cet algorithme est que le choix de nœud à traiter est aléatoire et avec une condition d'arrêt (non pas une mesure), plusieurs résultats finaux peuvent être obtenus.

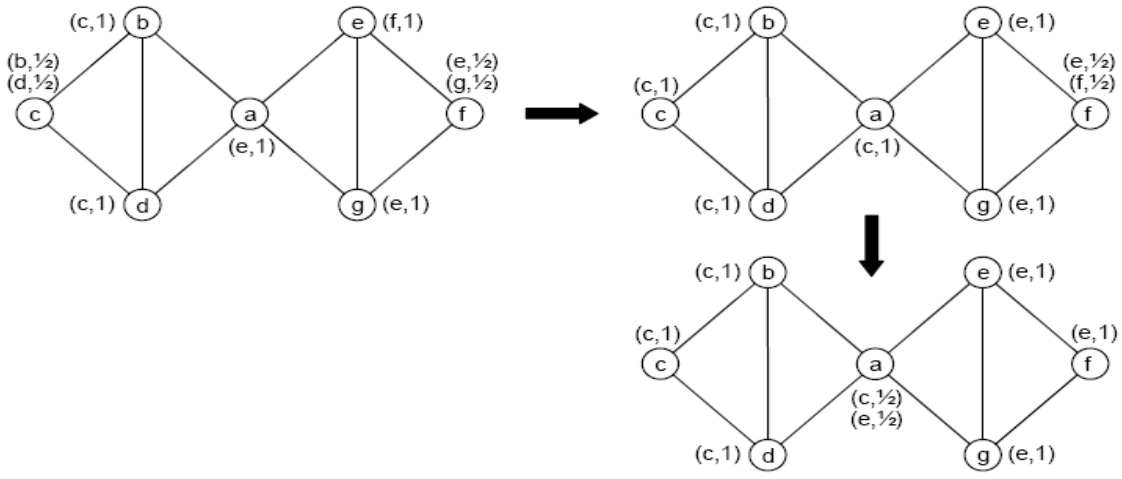


Figure 2.13 : Exemple de l'exécution de l'algorithme Copra **[Gregory,2010]**

2.2.2.3 Approches basées sur des graines

Le schéma général de ces méthodes est structuré en deux étapes :

- Déterminer un ensemble de nœuds ou groupes de nœuds dans le graphe qu'on désigne par des graines et qui constituent en quelque sorte les centres de communautés à retrouver.

- Appliquer une procédure d'expansion autour des graines afin d'identifier les communautés recouvrantes dans le réseau.

Différentes heuristiques de choix de graines ont été proposées. Une graine peut être composée d'un seul nœud sélectionné en utilisant les mesures classiques de centralité. Dans d'autres algorithmes la graine est composée d'un ensemble de nœuds qui ont une certaine connectivité.

Différentes stratégies d'expansion des graines sont aussi proposées. Dans beaucoup d'algorithmes on utilise les heuristiques développées pour l'identification de communautés locales .

Plusieurs solutions ont été proposées pour ce faire, nous détaillerons quelques-unes par la suite.

- ***L'algorithme de Wang et al.*** [Wang,2009]

Les auteurs définissent comme graines les communautés trouvées par un algorithme sans recouvrement. Ensuite, ils proposent d'ajouter ou de retirer des nœuds de ces communautés pour en augmenter la valeur de force (*community strength*), définie comme le ratio entre les liens internes sur la somme des degrés des nœuds de la communauté :

$$strength(C) = \frac{C^{int}}{(C^{int} + C^{ext})^{\propto}}$$

où α est un paramètre, de résolution, qui permet de faire varier la taille des communautés trouvées.

Cet algorithme a un inconvénient important, c'est qu'il ne permet pas de détecter toutes les communautés dans des cas de recouvrement très importants. Certaines communautés ne pourront pas être découvertes en partant de graines sans recouvrement. De telles communautés peuvent aussi, au final, être très semblables.

- ***L'algorithme OSLOM*** [Lancichinetti.2011]

Les auteurs de OSLOM proposent pour obtenir des graines plusieurs méthodes : pour une plus grande rapidité d'exécution, ils proposent d'utiliser InfoMap ou Louvain (des algorithmes rapides et efficaces mais ne tenant pas compte du recouvrement) pour obtenir un «premier jet» de communautés, qu'OSLOM va ensuite se charger d'améliorer en intégrant des nœuds périphériques, ou au contraire d'en réduire en rejetant les nœuds les moins pertinents.

Le principe est d'utiliser la probabilité pour un nœud d'être connecté à une communauté : connaissant le degré d'un nœud et le nombre de nœuds dans une communauté, on peut estimer combien ce nœud devrait avoir de connexions avec les nœuds de cette communauté dans un modèle non aléatoire. Si le nœud a plus de connexions que cette valeur moyenne, on aura tendance à l'ajouter à la communauté. S'il en a moins, on préférera le rejeter.

En plus de l'inconvénient de mentionner dans le premier algorithme ,au moins quatre paramètres modifiables, et une valeur seuil arbitraire, on peut s'interroger sur la robustesse des résultats.

Malgré que cet algorithme reproche d'être lent à l'exécution par la répétition de processus d'optimisation plusieurs fois pour chaque graine, il fournit des résultats plus pertinents.

2.2.3 Les approches dynamiques

Tous les algorithmes et définitions que nous avons présentées précédemment s'appliquent à des réseaux statiques. Or, la majorité des réseaux sont dynamiques.

Les réseaux dynamiques sont des réseaux évoluant dans le temps. Cette définition nous a conduit à une nouvelle définition de communautés dynamiques : est une succession de communautés statiques. L'évolution de communautés dynamique peut se faire de plusieurs manières :

- ***La croissance et la contraction*** : correspondant à l'ajout et au retrait de nœuds d'une communauté existante.

-La naissance et la mort de communautés : des nouvelles communautés peuvent apparaître, et d'anciennes communautés peuvent disparaître avec l'évolution de réseau.

- ***La fusion et la division de communautés*** :Deux communautés - ou plus - peuvent en effet, se fusionner en une seule au cours du temps. De manière semblable, une communauté peut se diviser en deux ou plus en communautés, plus petites que celle dont elles sont issues.

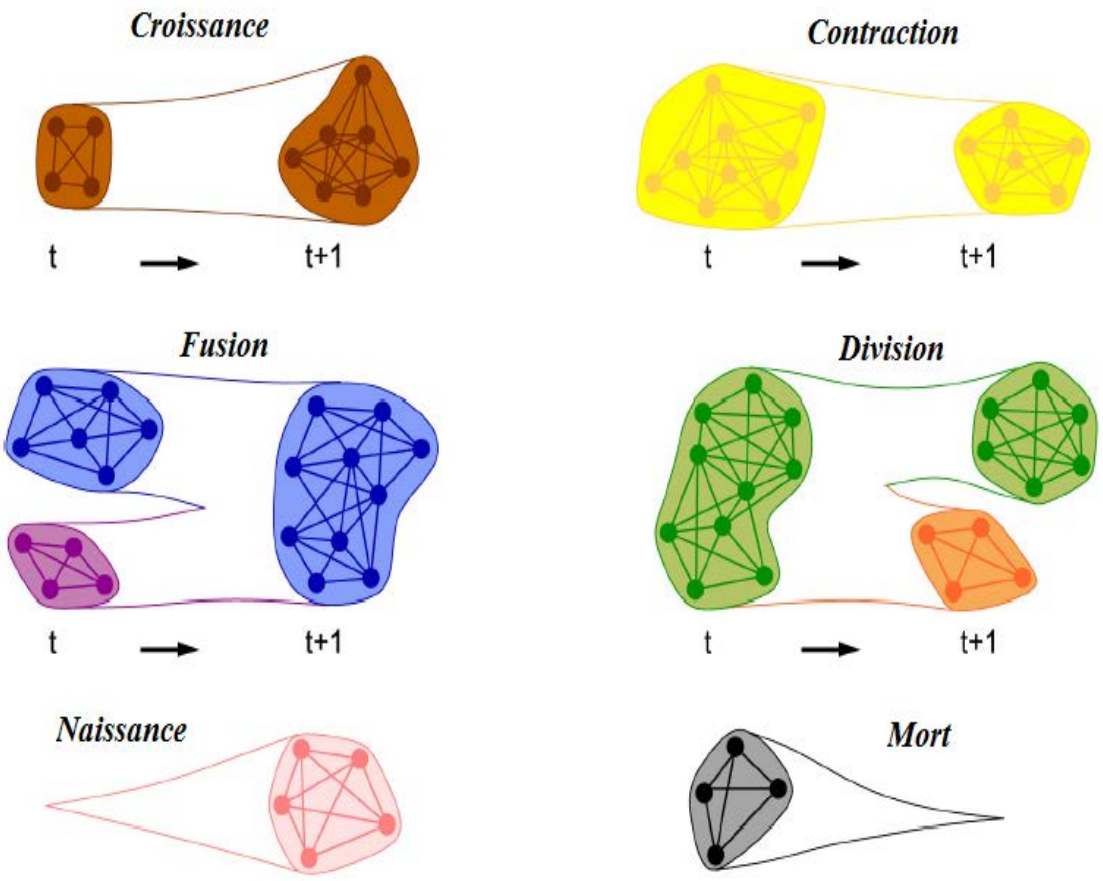


Figure 2.14 : Les différentes opérations possibles sur les communautés dynamiques **[Palla,2007]**

Actuellement, l'étude des graphes dynamiques a été menée dans deux grandes directions:

Premièrement :

découper les données en plusieurs graphes statiques (snapshots) et traiter chaque graphe de manière indépendante en utilisant un algorithme adapté pour les graphes statiques, puis de faire correspondre les communautés trouvées dans un graphe avec celles du graphe de l'instantané précédent.

Deuxièmement :

intégrer directement la dynamique dans la décomposition et non d'étudier une succession de graphes statiques.

Nous allons maintenant passer en revue les différentes solutions proposées pour la détection de communautés dans des graphes dynamiques.

2.2.3.1 Les approches par détections statiques successives

Plusieurs approches, avaient comme idée de considérer le graphe dynamique comme une succession d'instantanés indépendants. L'idée générale consiste à diviser le réseau dynamique en une série d'instantanés qui sont tous des graphes statiques puis la détection se fait en deux étapes :

Une première étape consiste donc à appliquer un algorithme statique sur chacun de ces instantanés, ce qui permet d'obtenir une série de partitions, une pour chaque instantané. Ensuite de trouver une correspondance (association) entre les communautés existant dans des instantanés consécutifs.

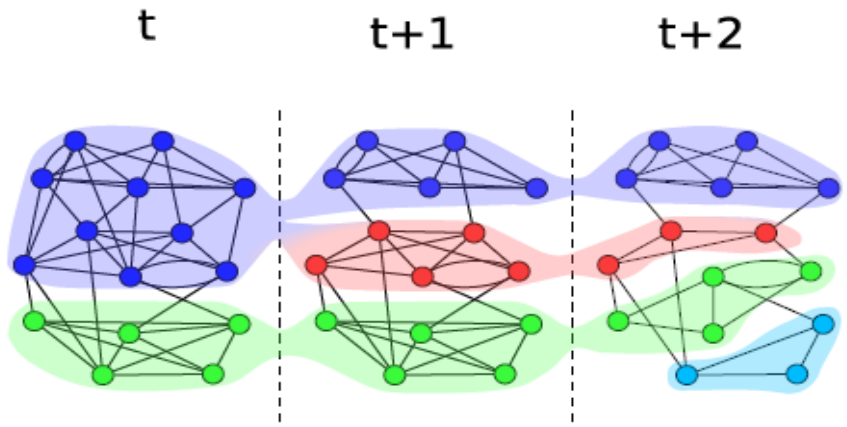


Figure 2.15 : Exemple de trois instantanés d'un réseau dynamique avec une association entre les communautés des différentes étapes. **[Aynaud,2011]**

A. Approches non recouvrantes

Ce sont les méthodes les plus simples, puisqu'elles utilisent les solutions statiques dans le cas dynamique. Cependant, toutes ces méthodes souffrent du problème de l'instabilité de la détection, liée aux algorithmes statiques utilisés.

- ***L'algorithme de Hopcroft et al*** [Hopcroft,2004]

Les auteurs utilisent la taille de l'intersection entre deux communautés successives : une communauté à l'instant ***t*** + **1** succède à une communauté à l'instant ***t*** si leur intersection est grande. Plus précisément, la valeur d'appariement ***match*** entre deux communautés est définie comme suit :

$$match\ (C_1, C_2\ ) = \min(\frac{|C_1 \cap C_2\ |}{|C_1|}\ , \frac{|C_1 \cap C_2\ |}{|C_2\ |}\ )$$

Cette valeur est comprise entre 0 et 1 et plus les communautés sont proches, plus l'intersection est grande. Il vaut 1 si elles sont égales et 0 si elles sont disjointes.

L'inconvénient de cet algorithme est dû principalement à l'instabilité des algorithmes de détection de communautés statiques. Par conséquent, les communautés sont initialement impossibles à suivre, car les changements sont causés par l'algorithme et non par une modification de la structure du réseau. Pour résoudre ce problème Hopcroft et al, décident de ne considérer que les communautés stables, qui sont les communautés qui continuent à exister même après une modification mineure.

Malgré que cette restriction élimine de nombreuses communautés intéressantes, elle a permis une première analyse de communautés dynamiques.

- ***L'algorithme de Wang et al.*** [Wang,2010]

Les auteurs utilisent le même type de solution basée sur des comparaisons d'intersections mais, au lieu de calculer le ***match*** entre des communautés, ils proposent de les suivre à l'aide de quelques nœuds importants appelés nœuds cœurs. Pour observer comment les communautés ont évolué, il suffit alors d'observer comment les nœuds cœurs, identifient le comportement des communautés.

Quand des nœuds cœurs dans la même communauté se séparent, c'est qu'un changement important a lieu. Les auteurs sélectionnent donc pour chaque communauté un certain nombre de représentants et suivent ces représentants au cours du temps.

Une limite ajoutée au problème de la définition des nœuds cœurs ,se résume par des difficultés liées au suivi des communautés .

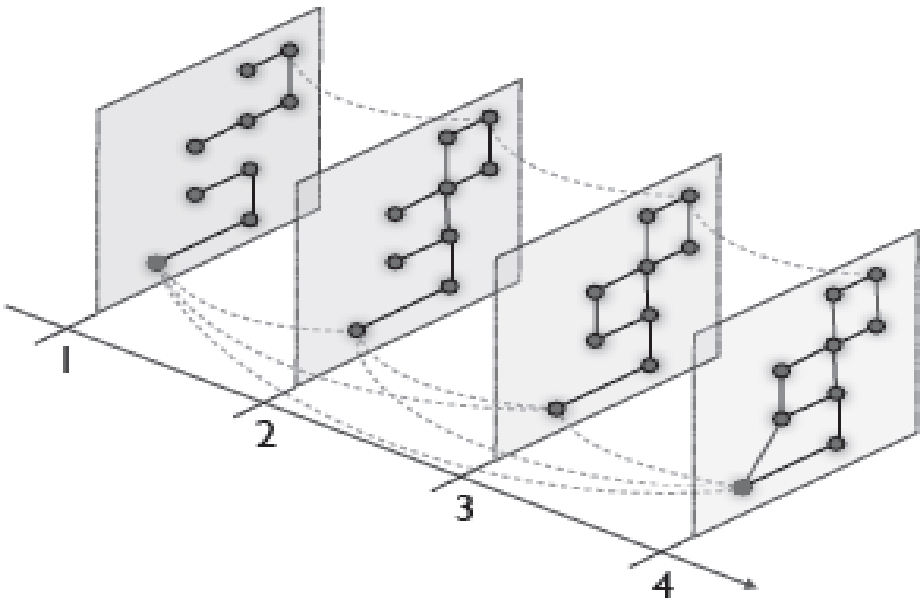


Figure 2.16: Exemple d'utilisation des nœuds cœurs pour identifier le comportement des communautés **[Cazabet,2013]**

B. Approches recouvrantes

- ***L'algorithme de Palla et al.*** [Palla,2007]

Les auteurs construisent un graphe contenant l'union des liens aux instants ***t*** et ***t*** +1. Les communautés sont détectées sur chaque instantané à l'aide de CPM adapté. Une des propriétés de l'algorithme utilisé pour la détection est que les communautés trouvées dans le graphe joint vont être l'union de communautés du graphe de l'instant ***t*** et de communautés du graphe de l'instant ***t*** +1. Cela permet d'obtenir l'évolution des communautés.

Les communautés du graphe à l'instant ***t*** deviennent celles de l'instant ***t*** +1 avec lesquelles elles sont regroupées dans le graphe joint (figure 16).

La propriété permettant cela est difficile à garantir, et la notion de communauté utilisée dans cet article est très restreinte et complexe à calculer.

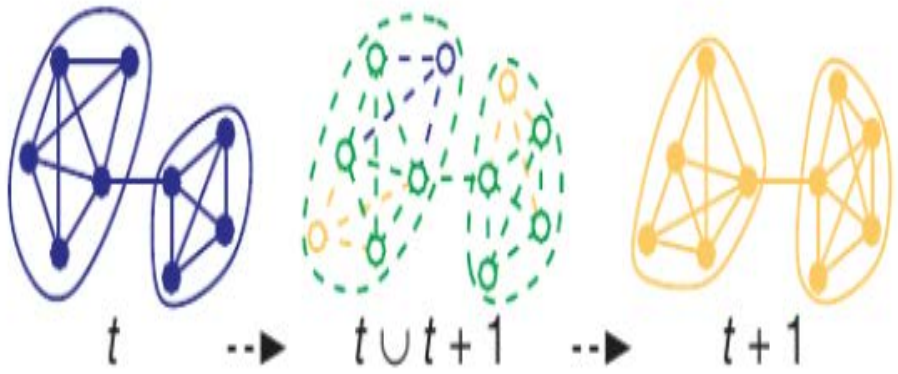


Figure 2.17 : Regroupement des graphes à deux instants pour trouver l'évolution des communautés entre ***t*** et ***t*** + 1 **[Palla,2007]**

- ***L'algorithme de Chen et al***. **[Chen,2010]**

Les auteurs utilisent des nœuds cœurs définis comme les nœuds existant aux instants ***t*** **- 1**, ***t*** et ***t*** **+ 1** pour réduire le nombre de communautés à considérer. Ils définissent les communautés initialement comme les cliques maximales et peuvent donc avoir des communautés recouvrantes. Néanmoins, le nombre de communautés peut être élevé et ils utilisent la notion de nœuds cœurs pour ne considérer que les communautés contenant des nœuds cœurs et donc restreindre leur nombre. Les nœuds cœurs permettent aussi dans un deuxième temps de faire du suivi en appliquant des règles classiques.

Evidemment, la faiblesse principale de cette méthode est qu'elle définit les communautés comme étant des cliques maximales, ce qui, d'une part, donne souvent des communautés peu pertinentes et, d'autre part, conduit à un nombre bien trop important de communautés dans des grands graphes ayant une densité élevée.

2.2.3.2 Les approches par détections statiques informées successives

Ces méthodes utilisent toujours les instantanés. Elles proposent de prendre en compte les résultats obtenus à l'instant ***t*** lors de la détection des communautés à l'instant ***t*** + 1. Ceci permet de réduire l'instabilité des algorithmes, en imposant le choix entre deux découpages différents, et pourraient prendre le plus semblable au découpage précédent.

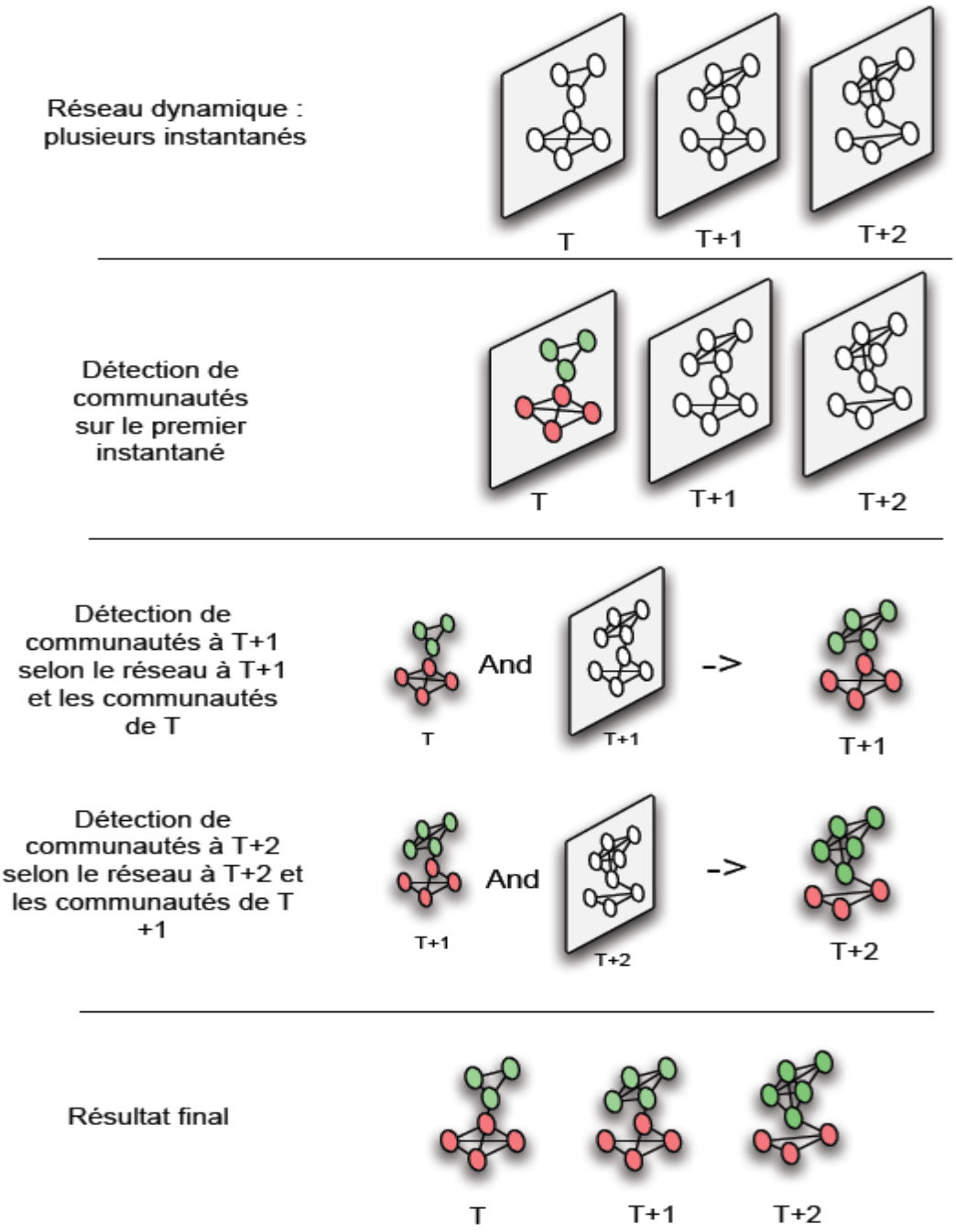


Figure 2.18: Représentation d'une méthode par détections statiques informées successives **[Cazabet,2013]**

A. Approches non recouvrantes

- ***L'algorithme de* Chakrabarti et al.** [Chakrabarti,2006]

Chakrabarti et al. proposent une fonction de qualité séparée en deux termes : un pour la qualité statique sur l'instantané considéré, et un second pour assurer la stabilité :

$$Q = Q_{instant} + \alpha Q_{stabilite}$$

où $Q_{instant}$ est une fonction de qualité statique (comme la modularité, mais d'autres fonctions sont utilisées), $Q_{stabilite}$est un terme qui évalue la distance entre la nouvelle partition et la précédente et α un paramètre caractérisant l'importance de la stabilité**.**

Cette nouvelle fonction de qualité permet d'obtenir une série de partitions plus intéressantes en réduisant le nombre d'artefacts causés par l'algorithme d'optimisation.

- ***L'algorithme de* Xu et al.** [Xu,2011]

Une approche similaire est décrite par Xu et al., où les auteurs tiennent compte de l'histoire non pas dans la fonction de qualité mais dans la matrice d'adjacence.

En considérant que la matrice d'adjacence à l'instant t est W_t, en tenant compte de la matrice d'adjacence W_{t+1} de l'instant t + 1.Les auteurs cherchent donc à maximiser la qualité sur un graphe ayant pour matrice d'adjacence :

$$W = \alpha W_t + (1- \alpha)W_{t+1}$$

B. Approches recouvrantes

- **L'algorithme de Lin et al.** [Lin.2009]

Les auteurs proposent une modification telle que pour étudier des réseaux multi-mode, c'est-à-dire des réseaux où les nœuds et les liens peuvent correspondre à des objets de types différents, une fonction est basée sur un modèle génératif probabiliste, consistant à formuler une fonction qui optimise conjointement la qualité et la stabilité des communautés.

L'avantage de cette méthode est qu'elle permet la détection de communautés recouvrantes, mais elle impose cependant de fortes contraintes : le nombre de communautés doit être connu à l'avance, et elle ne permet pas d'opérations telles que la fusion ou la division de communautés.

- **L'algorithme de Kim et al.** [Kim,2009]

Les auteurs proposent d'au lieu de modifier la fonction de qualité, la stabilité forcée en modifiant les distances entre les objets à classer. La distance considérée entre deux objets u et v, sert à classer les objets en communautés d'objets proches et modifiés à partir de la distance réelle à l'instant t, notée $dist_t(u; v)$en :

$$dist(u; v) = dist_{t+1}(u ; v) + \alpha dist_t (u ; v)$$

2.2.3.3 Les approches travaillant sur des réseaux temporels

Ces méthodes travaillent directement sur le réseau temporel qui représente les relations entre les communautés à chaque pas de temps. On peut alors décomposer ce réseau entre communautés pour obtenir des groupes de communautés appartenant à plusieurs pas de temps. L'algorithme est donc en deux phases :

Premièrement, tous les instantanés du réseau analysé sont décomposés en communautés statiques.

Deuxièmement, ces communautés deviennent les nœuds d'un réseau représentant les rapports entre les communautés à différents pas de temps.

Ce réseau est ensuite analysé, souvent à l'aide encore d'un algorithme de détection de communautés.

Nous appellerons les communautés obtenues initialement sur chaque instantané des ***instances*** et les communautés obtenues sur le réseau entre instances des ***chronologies***.

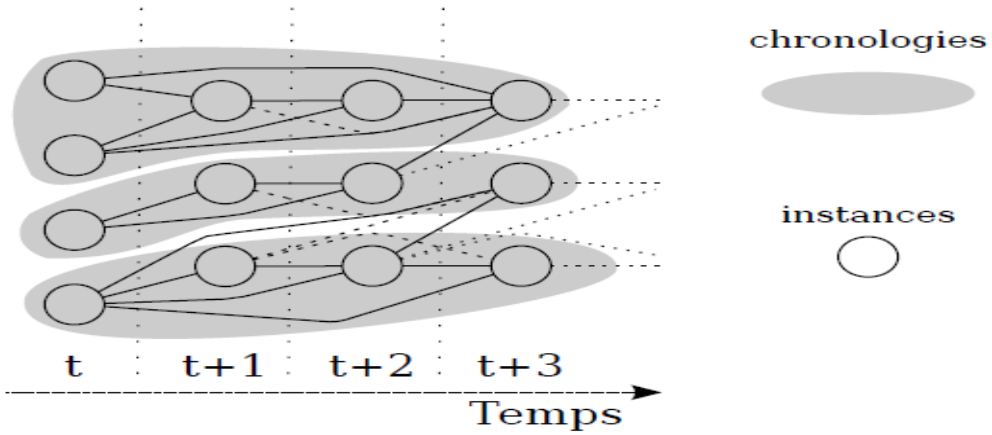


Figure 2.19 :Représentation d'un graphe d'***instances*** et de ***chronologies*** **[Aynaud,2011]**

A. Approches non recouvrantes

- ***L'algorithme de Li et al.*** [Li , 2012]

Itérativement, les nœuds affectés par les modifications faites au réseau (ajout ou suppression d'un lien) vont pouvoir changer de communautés avec laquelle ils ont le plus de liens, par rapport à l'étape précédente. L'algorithme fixe un paramètre sur lequel les communautés sont jugés différentes ou pas. Cependant, si la différence entre deux communautés est inférieure à ce paramètre, alors le nœud préférera rester dans la communauté dans laquelle il appartenait à l'étape précédente.

L'inconvénient de cet algorithme est qu'il ne permet ni recouvrement de communautés ni opération sur les communautés.

B. Approches recouvrantes

- **L'algorithme de Falkowski et al.** [Falkowski,2008]

Les auteurs définissent tout d'abord une distance entre les nœuds d'un réseau puis définissent le voisinage d'un nœud comme la boule topologique de rayon ***R*** **(** variant entre 0 et 1). Ils ne considèrent que les nœuds dont le voisinage est plus grand qu'une limite donnée ***S*** et les considèrent comme des nœuds cœurs. Les nœuds présents dans le voisinage d'un nœud cœur sont des nœuds frontières. Ils définissent ensuite les communautés comme les unions des voisinages partageant des nœuds. Cela définit un algorithme de détection de communautés statiques très similaire à l'algorithme de clustering standard DBSCAN. Ils proposent ensuite des techniques pour mettre à jour les voisinages et les communautés : à chaque nouvelle étape, les valeurs de distance entre les nœuds sont mises à jour. Si l'une de ces modifications fait apparaître un nouveau nœud cœur, celui-ci est intégré à une nouvelle communauté, ou en crée une nouvelle s'il n'a pas de nœud cœur dans son voisinage. De même, si des nœuds qui étaient dans le rayon **r** d'un nœud cœur passent au-delà de ce rayon, ils quittent la communauté, et ainsi de suite.

Cependant, le problème de cette solution est que la définition de communauté utilisée est très particulière, et est assez éloignée de ce que l'on considère généralement comme une bonne communauté. De plus, elle dépend toujours des valeurs choisies de paramètres ***R*** et ***S***.

2.3 Faiblesses des méthodes existantes pour la détection de communautés

2.3.1 Optimisation de la modularité

Les approches statiques (ou dynamiques basées sur des algorithmes statiques) fondées sur l'optimisation de la modularité souffrent d'un problème de limite de résolution dans le sens qu'ils ne peuvent pas distinguer des communautés plus petites d'une certaine taille limite .Pour des graphes non pondérés la maximisation de la modularité ne permet pas de distinguer des communautés ayant un nombre de liens inférieur à $\sqrt{\frac{m}{2}}$

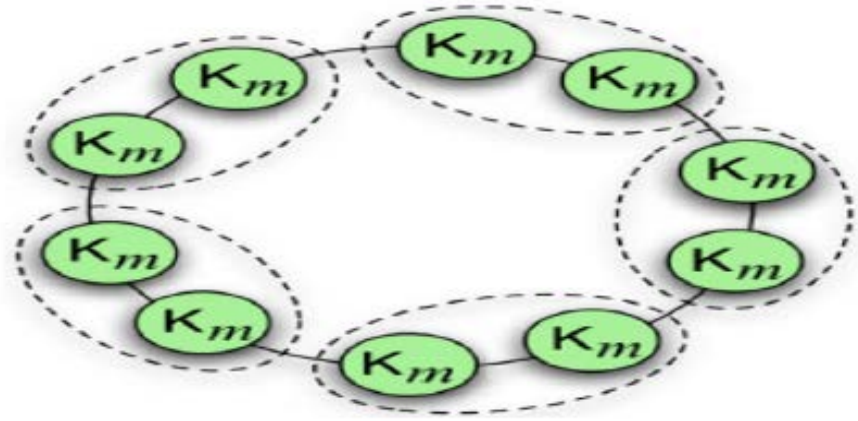


Figure 2.20: Exemple du problème de limite de la résolution de la modularité : La maximisation de la modularité conduit à grouper les cliques deux à deux

[Seifi,2012]

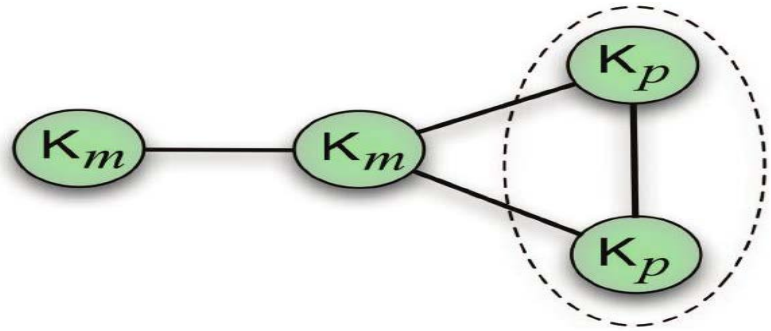


Figure 2.21: Un graphe formé de deux cliques de taille m et deux cliques de taille p. Si p << m les deux petites cliques Kp de taille p sont réunies en une seule communautés, bien qu'il n'y ait qu'un lien entre elles. **[Seifi,2012]**

En plus, dans les grands graphes il existe un grand nombre de partitions qui ont des valeurs de modularité qui sont très proches de la modularité maximale et correspondent pourtant à des partitions très différentes.

Cependant, malgré la popularité de la modularité, la significativité des résultats obtenus avec des algorithmes basés sur la modularité n'est pas évidente. En effet, une petite perturbation du graphe peut influer grandement sur la sortie de tels algorithmes.

2.3.2 L'instabilité

En effet, les algorithmes statiques (ou dynamiques basées sur des algorithmes statiques) de détection de communautés peuvent donner des résultats assez éloignés pour des graphes très proches. Ceci est lié à leur caractère heuristique, à leur non déterminisme ou au fait qu'il y a souvent plusieurs bonnes décompositions. Si les résultats varient trop, les événements suivis sont plus liés à l'algorithme qu'à un changement structurel et n'apportent donc pratiquement pas d'informations sur l'évolution du réseau.

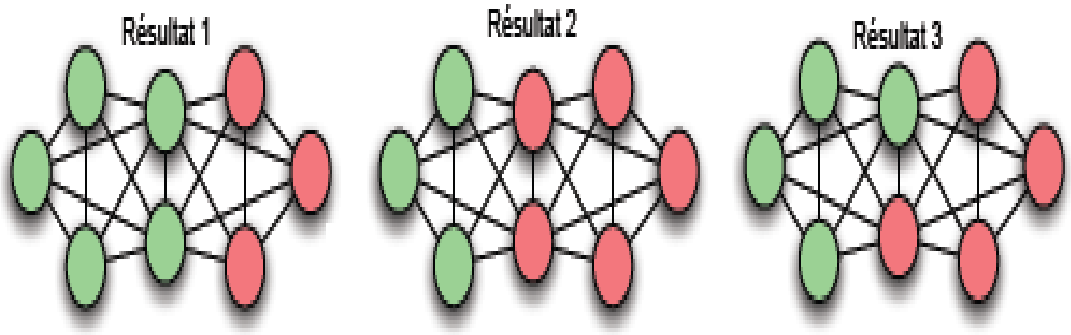


Figure 2.22: Exemple d'exécution d'un même algorithme plusieurs fois conduisant à des résultats différents .

2.3.4 Le recouvrement

L'existence d'un mécanisme à classer les nœuds en permettant à certains d'appartenir à plusieurs communautés désignes le recouvrement : c'est un comportement important connu depuis longtemps dans les réseaux sociaux.

Dans le cas d'un graphe dynamique, une grande partie des transformations se fait continuellement : un ensemble de nœuds d'une communauté la quitte pour en rejoindre une autre par exemple, il est impossible de représenter ceci sans recouvrement et cela impose des évolutions faites d'importantes actions comme la division ou la fusion d'une communauté. Ainsi, une compréhension profonde de la dynamique nous semble difficile à représenter un tel cas de figure sans recouvrement.

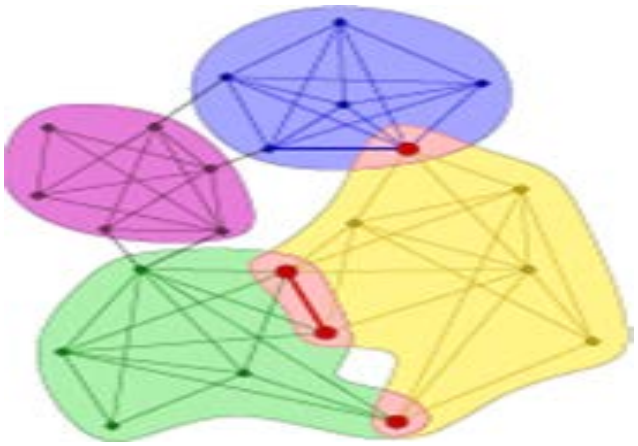


Figure 2.23 : Superposition de communautés **[Palla,2005]**

2.4 Conclusion

Dans ce chapitre nous avons détaillé quelques algorithmes appartenant à des catégories de méthodes de détection de communautés différentes. Aucune des techniques mentionnées n'est complètement satisfaisante à cause des résultats obtenus avec ces algorithmes qui ne sont pas toujours significatifs.

Dans le chapitre suivant, nous allons présenter une nouvelle méthode de détection de communautés qui se comporte mieux avec des réseaux sociaux où le nombre de liens inter-communautés est élevé et le nombre de liens intra-communautés est faible.

Chapitre 3

Méthode proposée

Sommaire

3.1 Introduction

Ce chapitre présente une nouvelle méthode de détection de communautés dans les réseaux sociaux. La méthode proposée est applicable sur des réseaux non orientés et non pondérés. La détection de communautés se fait en deux phases. La première phase vise à détecter tous les circuits afin de décomposer le réseau initial en petits groupes élémentaires. La deuxième phase vise à identifier une structure de communautés par la fusion de ces groupes élémentaires via un processus itératif. Les détails de chaque phase sont présentés dans les sous-sections 3.1.2 et 3.2.2

3.2 La première phase

La méthode proposée est inspirée de principes issus des méthodes basées sur des cliques. La détection des circuits - des cliques – se fait par l'exploration en profondeur du graphe cible en parcourant les nœuds un par un et à chaque fois qu'on tombe dans un nœud déjà visité , tous les nœuds situés dans ce parcours appartant au même circuit.

3.2.1 Procédure d'implémentation

L'implémentation de l'algorithme de cette phase nécessite l'utilisation d'une pile qui est capable de déterminer tous les nœuds formant un circuit : chaque fois que l'on parcourt un lien allant du nœud de pile ***A*** à un autre nœud ***B*** appartenant à la pile, tous les nœuds situés dans la pile depuis ***A*** jusqu'à ***B*** sont sur le même circuit. L'algorithme suivant décrit les étapes de la première phase.

3.2.2 Algorithme de la première phase :

Entrées : $G = (V , E)$: le graphe initial

Sorties : $G_1 , G_2 , G_3 , ...,G_N$: des sous graphes représentant les différents circuits détectés.

Variables :

Empiler() : une fonction qui sert à empiler les nœuds

Dépiler() : une fonction qui sert à dépiler les nœuds

Circuit ( ) : une fonction qui sert à recopier les nœuds formant un circuit

list_circuit : une liste qui contient les circuits détectés

Début :

Empiler(racine); // empiler le premier nœud

Répéter

Pour chaque *lien* N_{ij} **faire**

Empiler (***N***);

Si ***N*** $\in$ *Pile* **alors** :

Ajouter *Circuit* (***N***) dans *list_circuit* ;

Dépiler (***N***) ;

Fin si

Fin pour

Dépiler (***N***) ; // dépiler le nœud après visité tous ces liens

Jusqu'à *Pile* (***vide***);

Fin début

3.2.3 Complexité

Puisque chaque nœud est empilé exactement une fois. C'est donc un algorithme avec une complexité de ***0(n)***.

3.2.4 Un exemple illustratif

Le graphe qu'on va utiliser est constitué initialement de onze nœuds et dix-huit liens et il est connexe. Après le déroulement de la première phase la méthode a retrouvé six sous-graphes.

Le graphe évalué est représenté par la Figure 1.

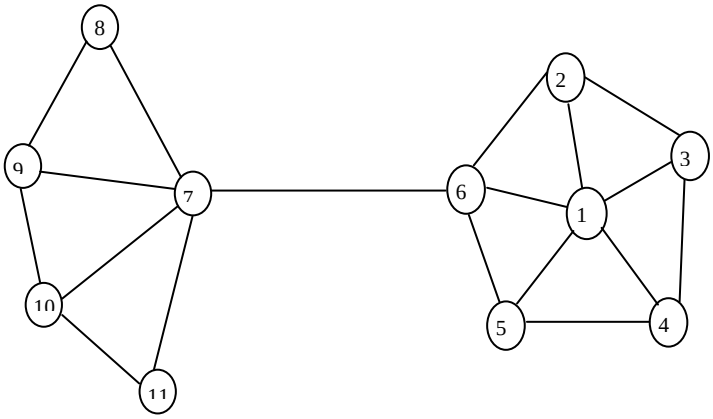


Figure 3.1 **-** Exemple de graphe connexe

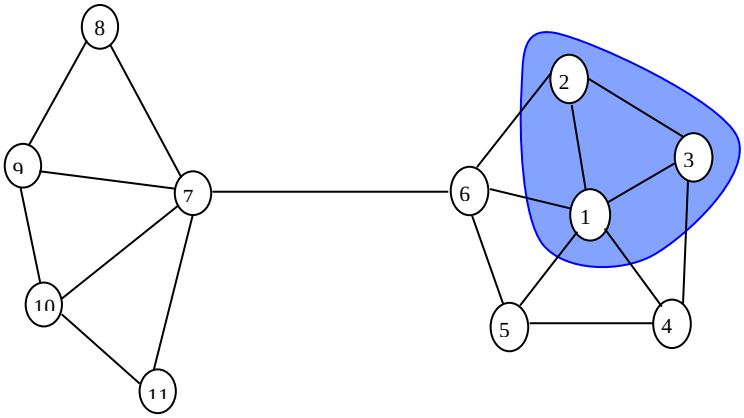


Figure 3.2 **-** Détection du premier circuit dans le graphe

Circuit 1: 1 – 2 – 3

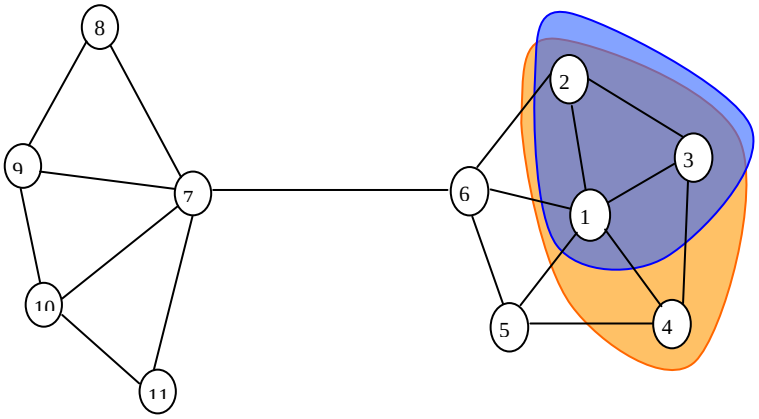


Figure 3.3 **-** Détection du deuxième circuit dans le graphe

Circuit 2 : 1 – 2 – 3 – 4

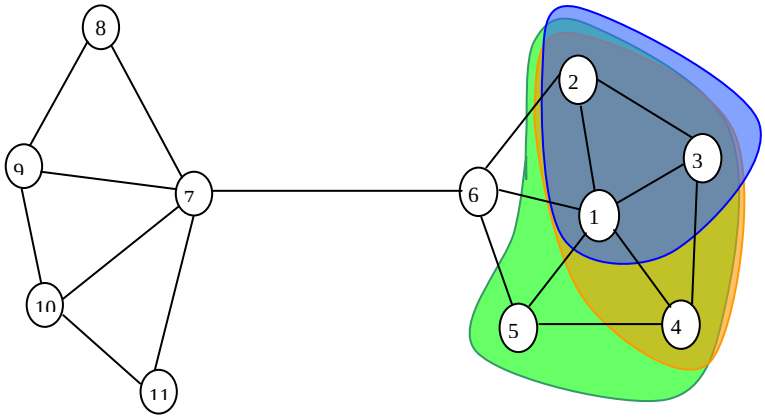


Figure 3.4 : Détection du troisième circuit dans le graphe

Circuit 3: 1 – 2 – 3 – 4 – 5

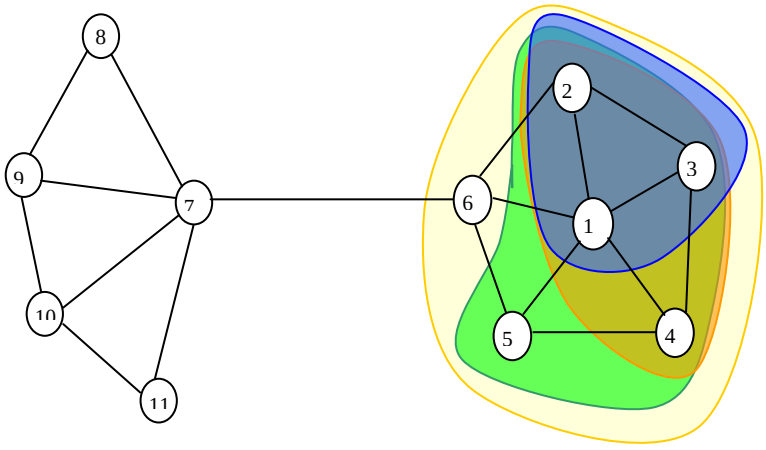


Figure 3.5 : Détection du quatrième circuit dans le graphe

Circuit 4: 1 – 2 – 3 – 4 – 5 – 6

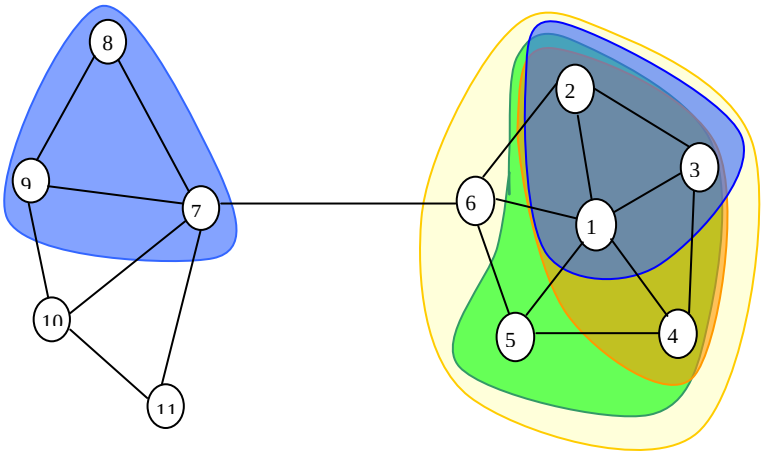


Figure 3.6 : Détection du cinquième circuit dans le graphe

Circuit 5: 7 – 8 – 9

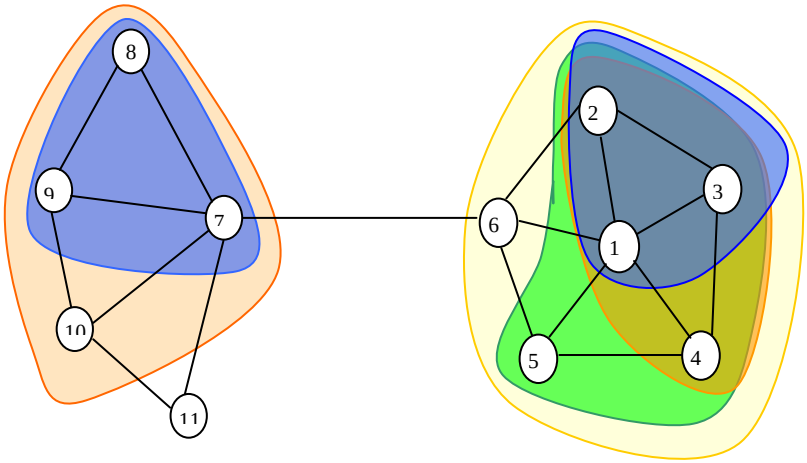


Figure 3.7: Détection du sixième circuit dans le graphe

Circuit 6: 7 – 8 – 9 – 10

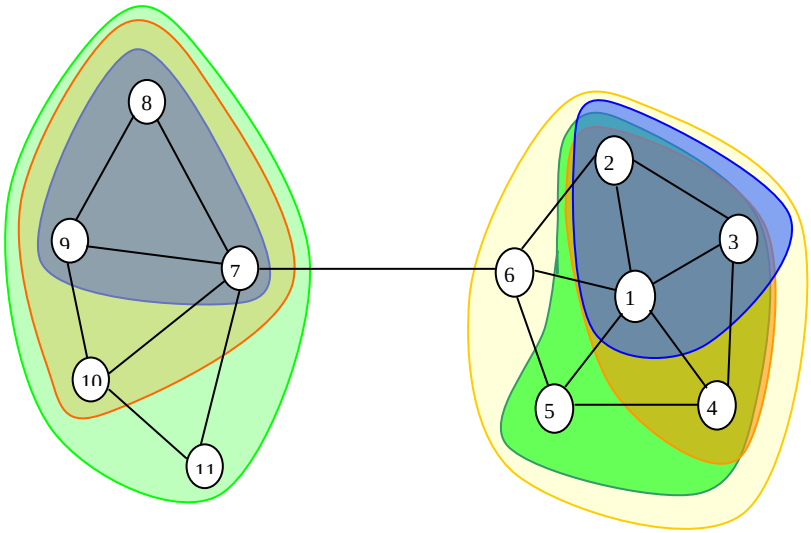


Figure 3.8 : Détection du septième circuit dans le graphe

Circuit 7: 7 – 8 – 9 – 10 – 11

3.3 La deuxième phase

La deuxième phase vise à fusionner les sous graphes – circuits- obtenus dans la première phase afin de trouver la structure de communautés optimale. Pour chaque sous graphe si il a ***n - 1*** nœuds en commun avec un autre, fusionner les deux sous graphes, même si les tailles des sous graphes sont différentes.

3.3.1 Procédure d'implémentation

La mise en œuvre de cette phase nécessite l'utilisation d'une liste afin de mémoriser tous les circuits détectés. Une autre procédure lancée s'occupe du raffinement des circuits détectés et de la suppression de redondances . L'algorithme suivant décrit les étapes de la deuxième phase.

3.3.2 Algorithme de la deuxième phase :

Entrées : G_1 , G_2 , G_3 , ...,G_N : des sous graphes obtenu à la première phase

Sorties : C_1 , C_2 , C_3 , ...,C_N : des sous graphes représentant les différentes communautés détectées.

Variables :

Test_graphe () : une fonction qui sert à tester si les deux sous graphes peuvent être fusionnés (avoir n-1 nœuds en communs)

Fusionner () : une fonction qui sert à fusionner deux sous graphes

*List_* comm: une liste qui contient les communautés détectées

Début :

Pour *i de 1 à n* **faire** // pour chaque sous graphe

Pour *j de i à n* **faire**

Si *Test_graphe*(G_i , G_j) **alors** :

C = Fusionné(G_i , G_j) ;

Fin si

Fin pour

Liste_comm (C) // ajouter communauté dans la liste

Fin pour

Fin début

3.3.3 Complexité

Dans ce cas, on ne peut pas faire le calcul de complexité exact puisque les sous graphes détectés dépendent du nombre de liens (densité) dans le réseau initial. Par expérimentation, on peut dire que la complexité est de **O (m – n)** où ***m*** est le nombre de liens et ***n*** celui des nœuds.

3.3.4 Un exemple illustratif

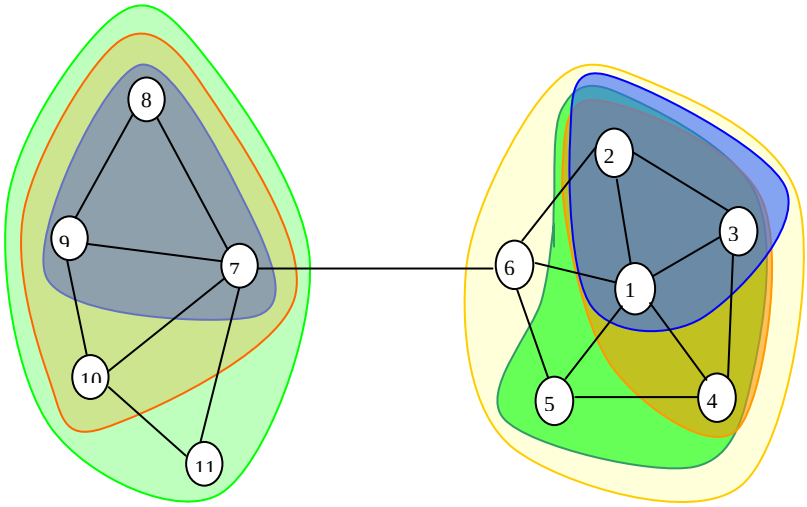


Figure 3.9 :
Les sous graphes obtenus dans la première phase

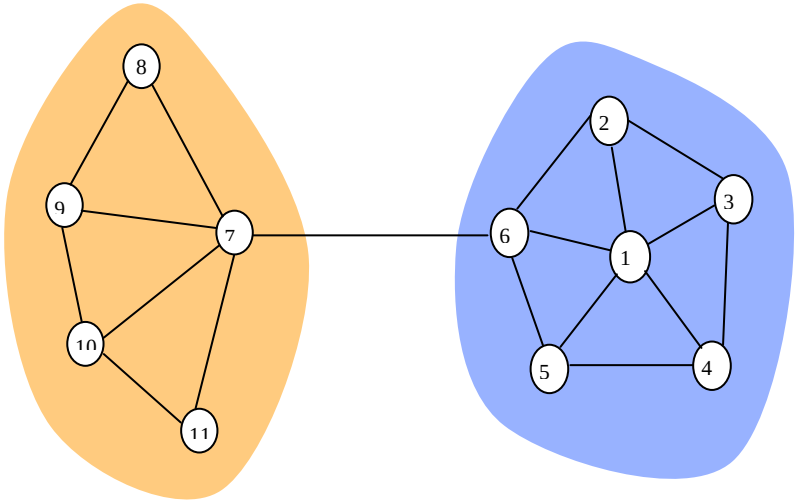


Figure 3.10 :
Communautés détectées après la deuxième phase

3.4 Complexité globale

La complexité globale estimée de notre méthode est égale à la somme de complexité des deux phases . La complexité de l'algorithme est donc :

Complexité = O(n) + O(m-n) = O(n+m-n) = O(m).

3.5 Discussion sur la méthode proposée

Dans cette partie nous allons présenter en détail les avantages et les limites de la méthode proposée. Puisque cette méthode est une inspiration de l'algorithme de percolation de cliques, qui est l'un des algorithmes les plus classiques de détection de communautés, alors elle hérite de la plupart de ses avantages et ses faiblesses.

3.5.1 Les avantages

- **La rapidité**

  La rapidité de l'algorithme est due au fait que sa complexité vis-à-vis d'autres méthodes de détection de communautés, lui permet de traiter des graphes ayant jusqu'à plusieurs millions de nœuds .

- **Le recouvrement**

  Notre méthode est capable de détecter des recouvrements très importants, parfois avec de nombreux nœuds en commun, ce qui est très important dans les réseaux sociaux.

- **Les paramètres**

  Il n'y a aucune valeur du paramètre requis à fixer , ce qui nous a évité de tomber dans le problème de la recherche du meilleur choix de ces paramètres.

- **La stabilité**

  Les communautés trouvées par notre méthode restant stables dans plusieurs exécutions sur le même réseau.

- **L'implémentation**

  La mise en œuvre de la méthode proposée est très facile. En plus de la pile utilisée, on attache pour chaque circuit détecté une liste chainée . Les structures utilisées sont faciles à implémenter et sont de manipulation simple.

3.5.2 Les limites

Comme toutes les méthodes basées sur les cliques, cette méthode souffre de certaines faiblesses : dans des graphes très denses, elle peut avoir une complexité et une consommation mémoire prohibitives.

De plus, sur des graphes peu dense ou des graphes ayant une structure arborescente, sans fermetures transitives, elle donne des résultats peu pertinentes.

3.6 Conclusion

Dans ce chapitre, nous avons présenté une nouvelle méthode de détection de communautés dans les réseaux sociaux. La méthode proposée contient deux phases.

Dans la première phase, la détection des circuits utilisée pour partitionner le graphe initial en des sous graphes . La deuxième phase consiste à trouver la structure de communautés optimale en fusionnant les sous graphes – circuits – selon la méthode de percolation de cliques sauf que les cliques ne sont pas forcément égaux.

Chapitre 4

Evaluation de la méthode proposée

Sommaire

4.1 Introduction

Ce chapitre présente une évaluation empirique de la méthode proposée de détection de communautés dans des réseaux sociaux.

Pour qu'un algorithme de détection de communautés soit considéré intéressant, il faut que les communautés qu'il trouve soient pertinentes, ce qui est souvent difficile à démontrer étant donné l'absence de définition précise de ce qu'est une bonne communauté.

Pour cela, nous avons considéré plusieurs réseaux sociaux. Tout d'abord nous l'avons testée sur un exemple cité dans le chapitre précédant, ensuite sur des réseaux sociaux issus du monde réel. La performance de notre méthode est comparée avec celles des algorithmes connus dans la littérature.

4.2 Expérimentations sur le réseau précédant

Dans cette section nous avons exécuté notre algorithme en basant sur le même exemple de graphe utilisé dans le chapitre précédant. Le graphe est constitué initialement de onze nœuds et dix-huit liens et il est connexe. La première phase a retrouvé six sous-graphes. Durant la deuxième phase, les sous-graphes qui ont été retrouvés sont fusionnés entre eux. Le graphe évalué est représenté par la figure suivante.

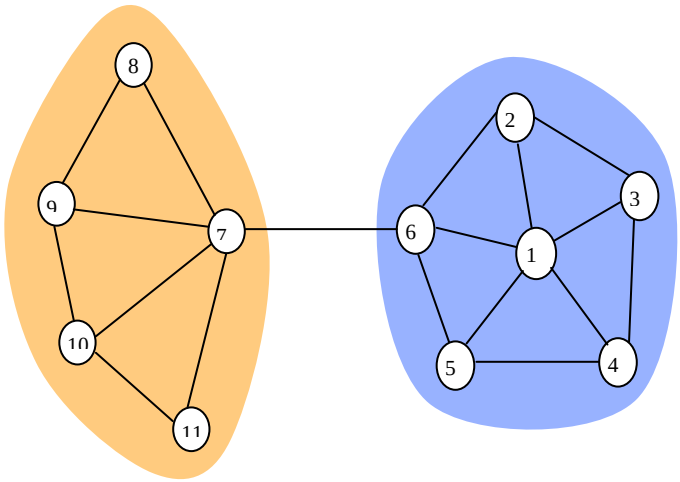


Figure 4.1 : Exemple de graphe utilisé précédemment

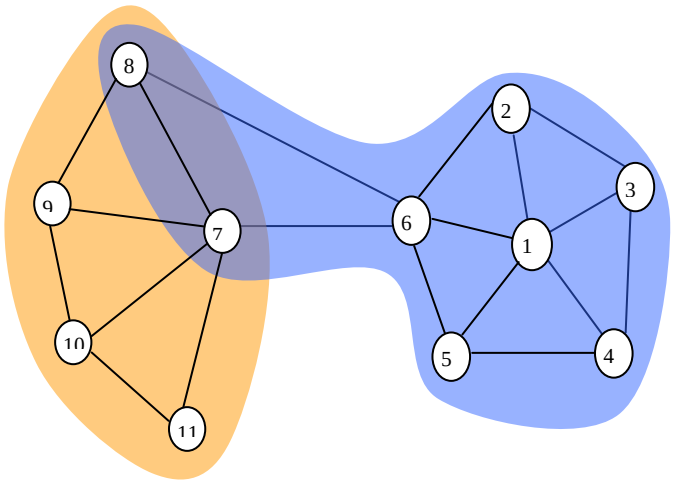


Figure 4.2 :
Réaction des communautés après l'ajout de l'arc (6-8)

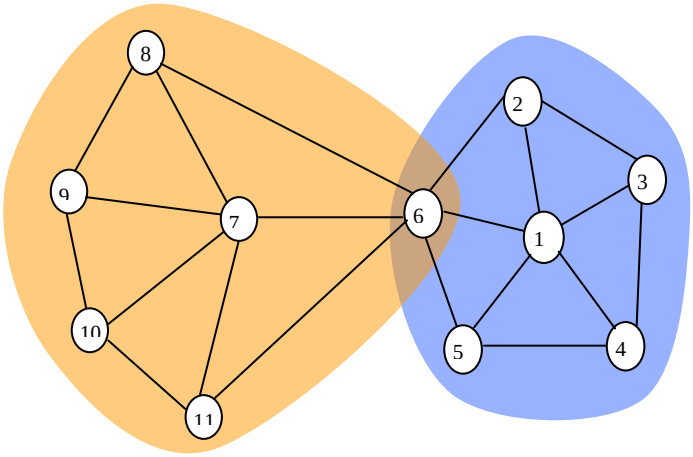


Figure 4.3 :
Réaction des communautés après l'ajout de l'arc (6-11)

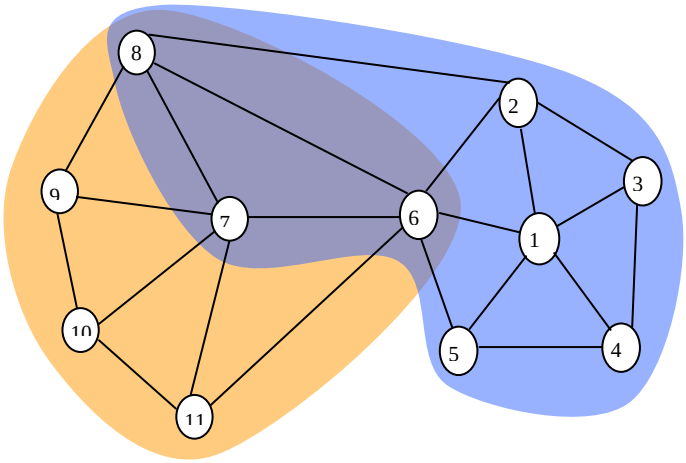


Figure 4.4 :
Réaction des communautés après l'ajout de l'arc (2-8)

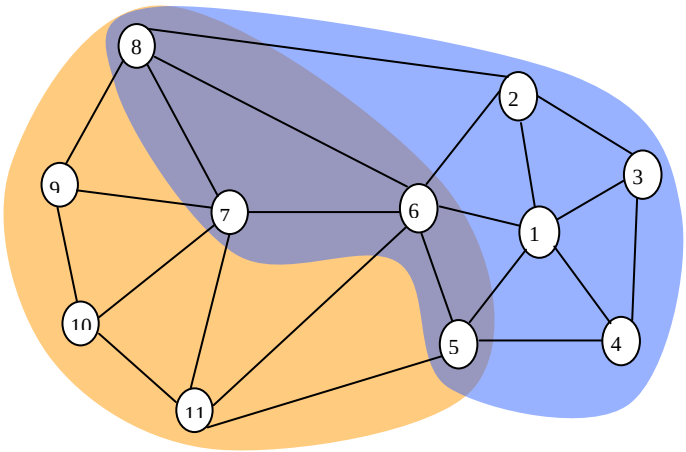


Figure 4.5 :
Réaction des communautés après l'ajout de l'arc (5-11)

4.3 Expérimentations sur des réseaux réels

Notre solution est aussi testée sur des réseaux réels. Pour les réseaux dont la structure de communautés est connue d'avance, nous avons testé quelques réseaux de références en comparant les résultats avec ceux des autres algorithmes connus dans la littérature, soit *Newman* [Newman,2004] , *Edge Betweenness* [Givan et Newman,2002] , *Label Propagation* [Raghavan,2007] *et Walktrap* [Pons et Latapy,2005].

4.3.1 Club de karaté de Zachary

Le premier exemple est le club de karaté de Zachary [Zachary,1977]. C'est un réseau construit à partir des relations entre 34 membres d'un club de karaté dans une université aux États Unis. Il s'agit d'un réseau très populaire et très utilisé par plusieurs algorithmes afin de tester leurs performances puisque sa structure de communautés est connue à l'avance. Ce réseau comporte deux communautés.

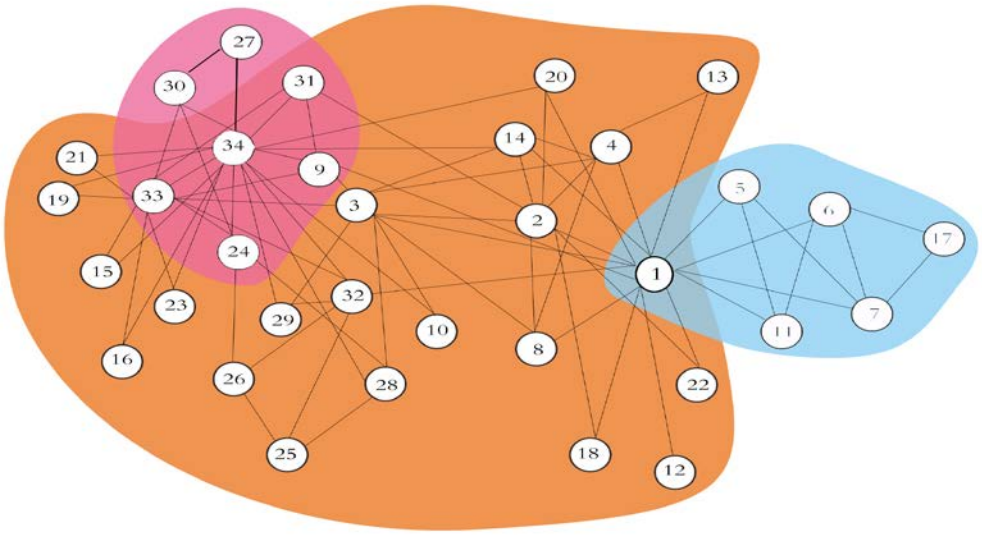


Figure 4.6 :
Structure de communautés trouvée par notre méthode pour le réseau de Zachary

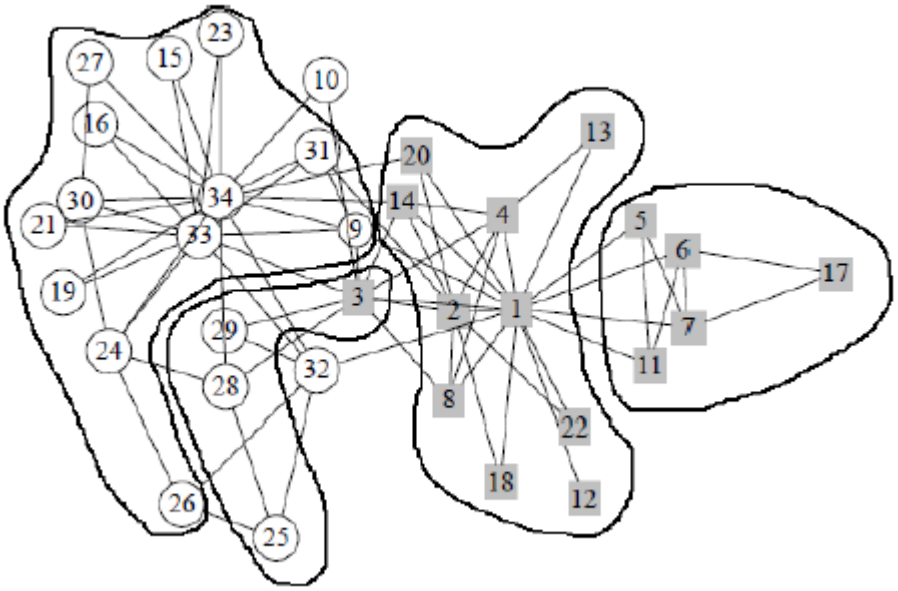


Figure 4.6 :
Structure de communautés trouvée par Givan et Newman pour le réseau de Zachary

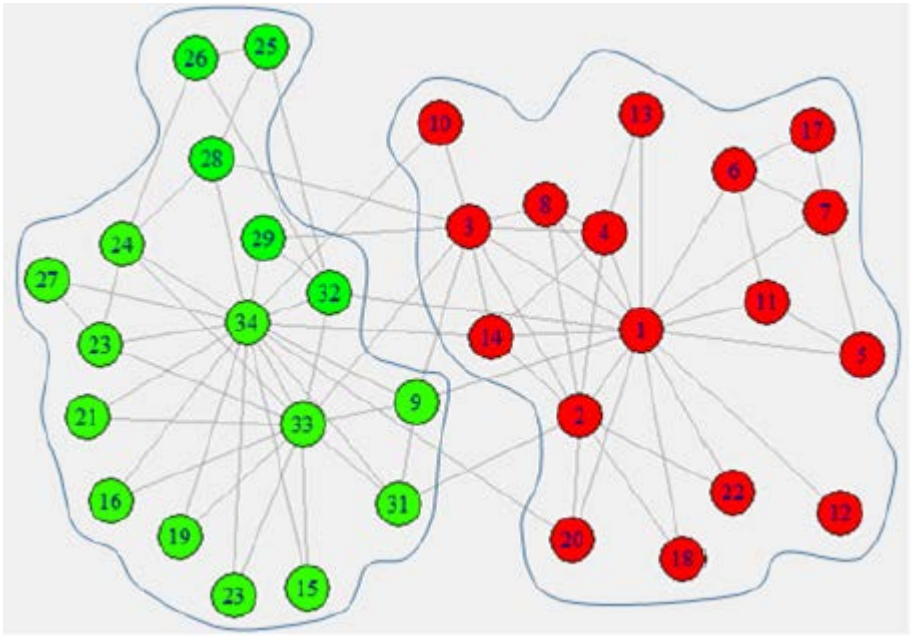


Figure 4.8 :
Structure de communautés trouvée par Fast Greedy pour le réseau de Zachary

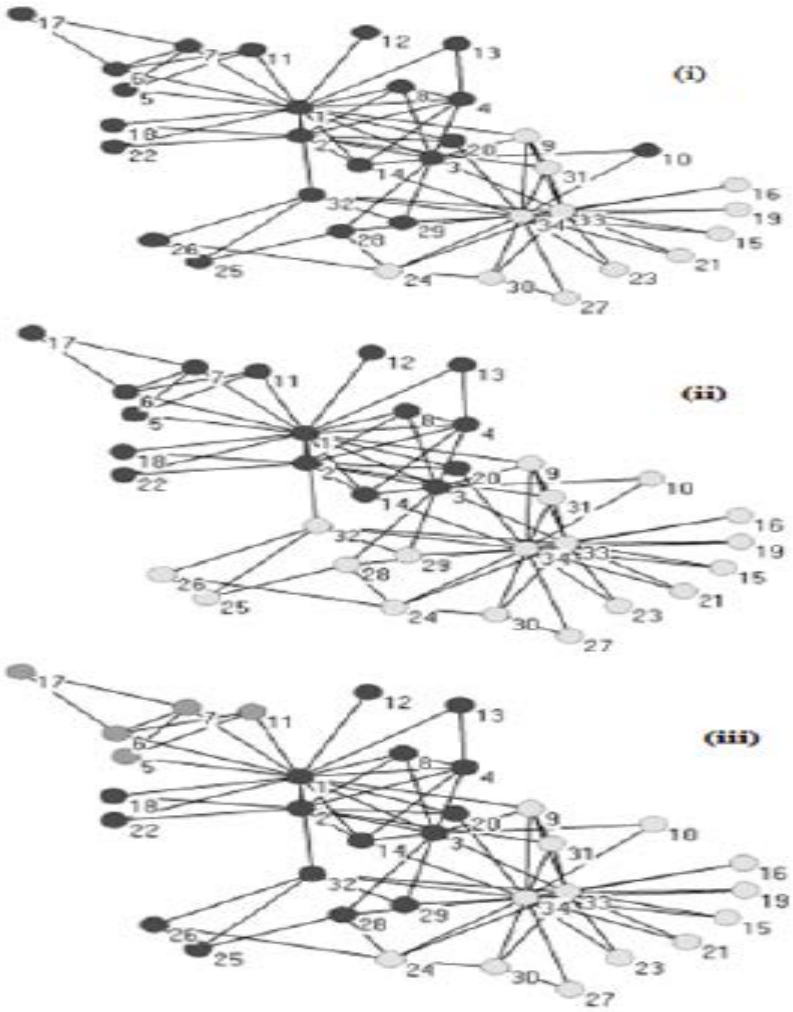


Figure 4.9 :
Différentes structures de communautés trouvées par Label Propagation pour le réseau de Zachary

Méthodes	Nombre de communautés	Nœuds non classés
Méthode proposée	3	
Newman	4	2
Edge Betweenness	5	
Label Propagation	3	
Walktrap	4	

Table 4.1 : Résultats de l'exécution des algorithmes sur le réseau de Zachary

4.3.2 Les dauphins de Lusseau

Le deuxième exemple à traiter est le réseau de dauphins de Lusseau [Lusseau,2003]. Ce réseau contient 62 nœuds et 159 liens. La méthode proposée a détecté quatre communautés avec des nœuds orphelines . Ce réseau est constitué essentiellement de deux communautés.

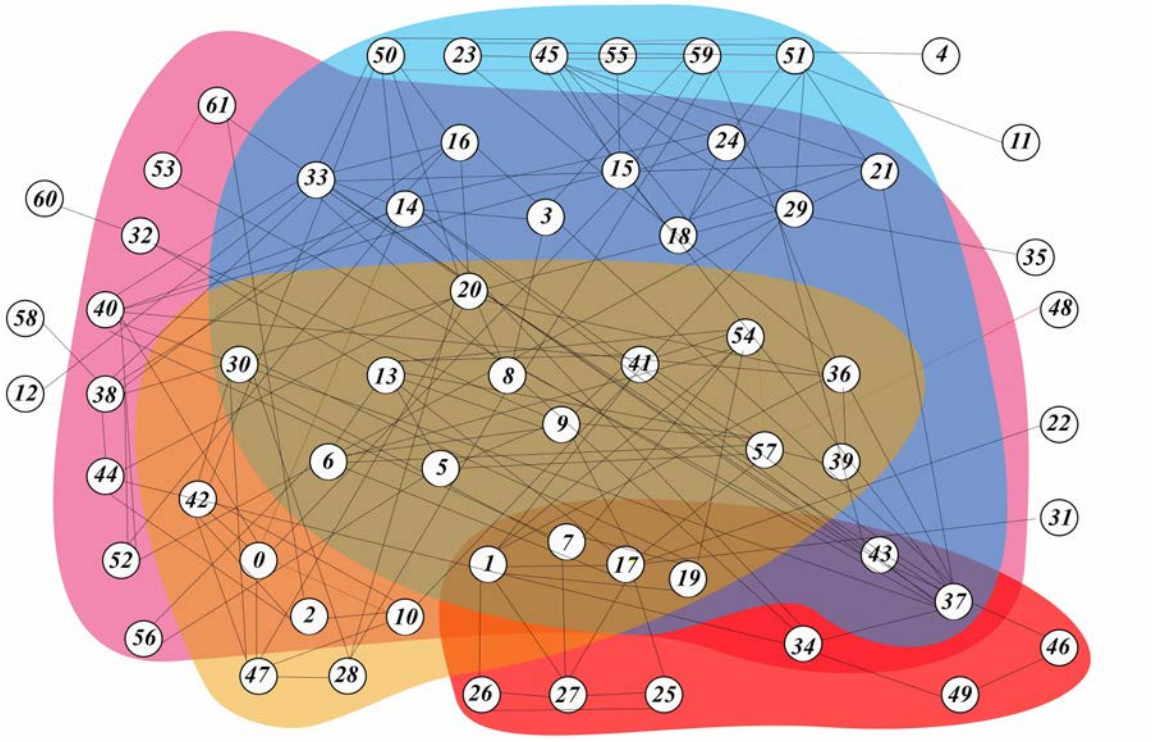


Figure 4.10 :

Les communautés détectées par la méthode proposée pour le réseau de dauphins de Lusseau

Méthodes	Nombre de communautés	Nœuds non classés
Méthode proposée	**4**	**9**
Newman	**5**	**2**
Edge Betweenness	**6**	
Label Propagation	**3**	
Walktrap	**4**	

Table 4.2 : Résultats de l'exécution des algorithmes sur le réseau de dauphins de Lusseau

4.3.3 Les livres politiques

Un troisième exemple que nous avons traité est le réseau de livres politiques [KREBS,2008]. Il ne s'agit pas de relations entre les êtres humains, mais plutôt de relations entre l'achat de plusieurs titres sur Amazon. Le réseau est constitué de 105 livres avec 441 liens se trouvant entre les livres achetés ensemble. Ce réseau a initialement trois groupes.

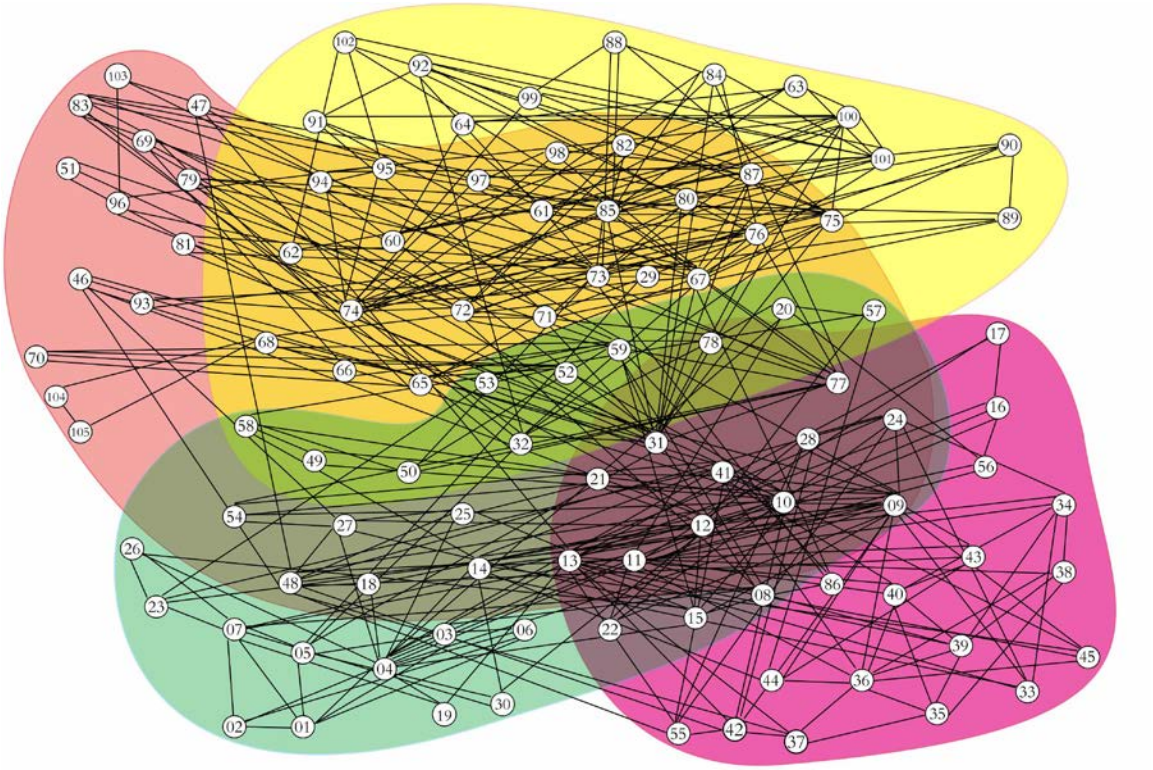


Figure 4.11 :

Les communautés détectées par la méthode proposée pour le réseau des Livres politiques

Méthodes	Nombre de communautés	Nœuds non classés
Méthode proposée	**4**	
Newman	**5**	
Edge Betweenness	**5**	
Label Propagation	**3**	
Walktrap	**4**	

Table 4.3 : Résultats de l'exécution des algorithmes sur le réseau de Livres politiques

4.3.4 Football américain

Un autre exemple de réseau réel étudié dans le cadre de ce mémoire est le réseau de jeux du football américain (*American football games*) [Park et Newman,2005]. Il représente le calendrier des matchs entre des équipes américaines de football durant l'année 2 000. Ce réseau est constitué de douze communautés, 115 nœuds et 613 liens.

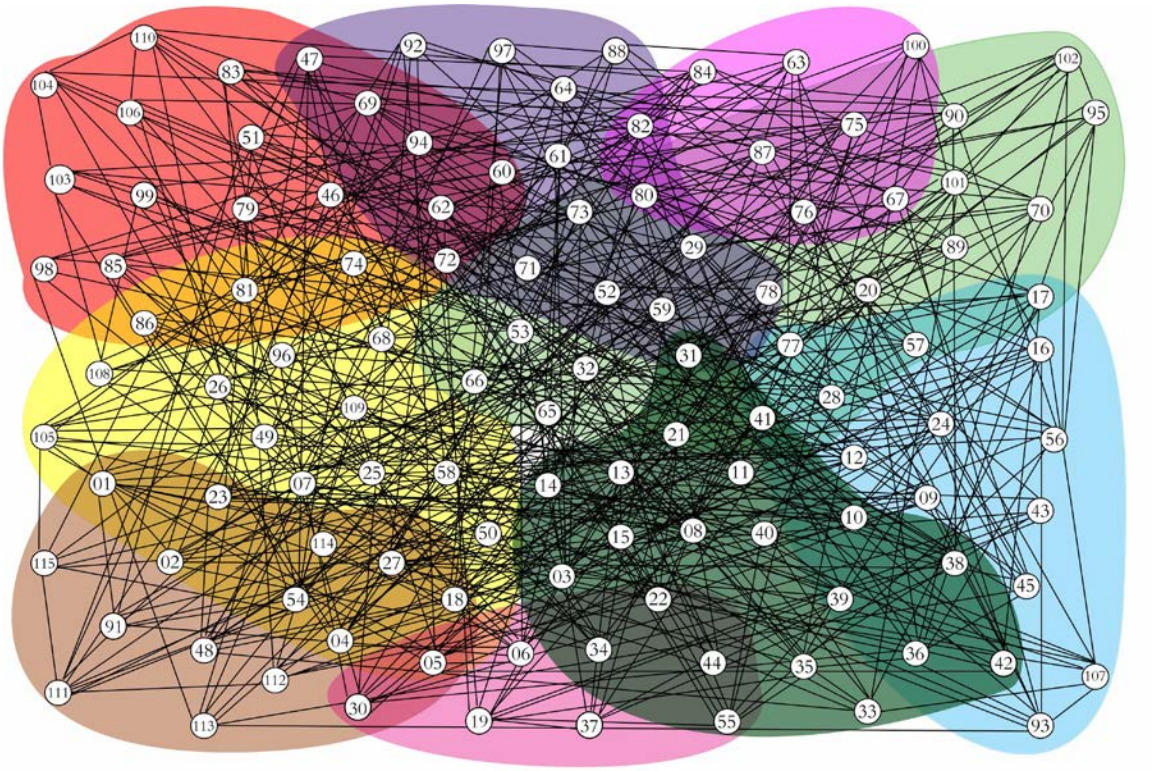


Figure 4.12 :

Les communautés détectées par la méthode proposée pour le réseau de Football américain

Méthodes	Nombre de communautés	Nœuds non classés
Méthode proposée	**9**	
Newman	**9**	
Edge Betweenness	**12**	
Label Propagation	**11**	
Walktrap	**10**	

Table 4.4 : Résultats de l'exécution des algorithmes sur le réseau du Football américain

4.4 Conclusion

Dans ce chapitre, nous avons exposé les résultats des expérimentations que nous avons réalisées. Nous avons montré, que les communautés trouvées par notre méthode étaient comparables à celles que pouvaient trouver d'autres algorithmes sur des réseaux sociaux.

Les tests réalisés montrent clairement que la détection de communautés est encore un problème complexe pour lequel aucun algorithme n'est le meilleur dans tous les cas.

En ce qui concerne le temps d’exécution, la méthode proposée a fait preuve de sa performance. En effet, elle est toujours parmi les plus rapide des algorithmes évalués, et la qualité de partitionnement était aussi bien pour des réseaux d'évaluations que pour des réseaux sociaux réels.

Conclusion générale

La détection de communautés est un domaine qui est encore dans une phase d'exploration, et pour lequel il faudra encore attendre quelques années avant d'arriver à un stade de maturation. Cette relative jeunesse du domaine a, d'une part, représenté un challenge et, d'autre part, a été un facteur de motivation important.

Dans le cadre de ce mémoire, nous avons présenté une méthode de détection de communautés dans les réseaux sociaux en procédant à une fonction de fusion entre les sous groupes et en exploitant la fonction de méthodes basées sur des cliques. Afin de valider notre travail, nous avons mené des comparaisons avec des algorithmes qui sont également très connus et considérés parmi les plus performants.

Le temps d'exécution de notre méthode est parmi les meilleurs des autres algorithmes . Nous étions toujours intéressés par la validation de nos résultats sur des réseaux sociaux, nous avons travaillé sur des Benchmark de références . Ceci nous a permis, d'une part, de prouver que les résultats trouvés par notre méthode étaient pertinents et exploitables et, d'autre part, que la détection de communautés avait des applications concrètes, et pouvait être très intéressante dans de nombreux domaines.

Dans le but d'améliorer ce travail, deux suggestions peuvent être émises. La première consiste à développer d'une nouvelle mesure de fusion. La deuxième, consiste en la prise en compte de la dynamique dans l'analyse de réseaux sociaux afin de permettre une meilleur détection des communautés grâce à cette dernière.

Bibliographique

[Newman,2004] M. E. J. Newman. Fast algorithm for detecting community structure in networks.*Physical Review E,* 69 :066133, 2004.

[Blondel ,2008] V.D. Blondel, J.L. Guillaume, R. Lambiotte et E. Lefebvre. Fast unfolding of communities in large networks. Journal of Statistical Mechanics : Theory and Experiment, vol. 2008, page P10008, 2008.

[Givan et Newman,2002] M. Girvan and M.E.J. Newman. Community structure in social and biological networks. Proceedings of the National Academy of Sciences, 99(12) :7821{7826, 2002.

[Radicchi,2004] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. PNAS, 101(9):2658–2663, 2004.

[Fortunato,2004] Santo Fortunato, Vito Latora, and Massimo Marchiori. Method to find community structures based on information centrality. Physical Review E, 70(5) :056104, 2004.

[Pons et Latapy,2005] P. Pons and M. Latapy. Computing communities in large networks using random walks. Computer and Information Sciences-ISCIS 2005, pages 284{293, 2005.

[*Rosvall et Bergstrom*,2008] M. Rosvall and C.T. Bergstrom. Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences,105(4) :1118{1123, 2008.

[Dongen,2000] Stijn van Dongen. Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht, May 2000.

[Donetti et Munoz,2004] L. Donetti and M. A. Mu˜noz. Detecting network communities : a new systematic and efficient algorithm. Journal of Statistical Mechanics, 2004(10) :10012, 2004.

[Kanawati,2011] Kanawati, R. (2011). Licod : Leaders identification for community detection in complex networks. In Social Com/PASSAT.

[Reichardt,2004] J. Reichardt and S. Bornholdt. Detecting fuzzy community structures in complex networks with a potts model. Physical Review Letters, 93(21) :218701, 2004.

[Palla,2007] G. Palla, A.L. Barabasi, and T. Vicsek. Quantifying social group evolution. Nature, 446(7136) :664{667, 2007.

[Tang et Liu, 2010] Tang, L. and Liu, H. (2010). Community Detection and Mining in Social Media. Synthesis Lectures on Data Mining and Knowledge Discovery. Morgan & Claypool Publishers.

[Shen,2009] H. Shen, X. Cheng, K. Cai, and M.B. Hu. Vgfvgfvg etect overlapping and hierarchical community structure in networks. Physica A : Statistical Mechanics and its Applications, 388(8) :1706{1712, 2009.

[Raghavan,2007] U.N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. Physical Review E, 76(3) :036106, 2007.

[Gregory,2010] S. Gregory. Finding overlapping communities in networks by label propagation. New Journal of Physics, 12(10) :103018, 2010.

[Wang,2009] X. Wang, L. Jiao, and J. Wu. Adjusting from disjoint to overlapping community detection of complex networks. Physica A : Statistical Mechanics and its Applications, 388(24) :5045{5056, 2009.

[Lancichinetti.2011] A. Lancichinetti, F. Radicchi, J.J. Ramasco, and S. Fortunato. Finding statistically significant communities in networks. PloS one, 6(4) :e18961, 2011.

[Hopcroft,2004] J. Hopcroft, O. Khan, B. Kulis, and B. Selman. Tracking evolving communities in large linked networks. Proceedings of the national academy of sciences of the United States of America, 101(Suppl 1) :5249{5253, 2004.

[Wang,2010] Q. Wang and E. Fleury. Mining time-dependent communities. In LAWDN - Latin-American Workshop on Dynamic Networks, 2010.

[Chen,2010] Z. Chen, K. a. Wilson, Y. Jin, W. Hendrix, and N. F. Samatova. Detecting and Tracking Community Dynamics in Evolutionary Networks. 2010 IEEE International Conference on Data Mining Workshops, pages 318–327, Dec. 2010.

[Chakrabarti,2006] D. Chakrabarti, R. Kumar, and A. Tomkins. Evolutionary clustering. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 554{560. ACM, 2006.

[Xu,2011] K. Xu, M. Kliger, and A. Hero. Tracking communities in dynamic social networks. Social Computing, Behavioral-Cultural Modeling and Prediction, 2011.

[Lin.2009] Y.R. Lin, Y. Chi, S. Zhu, H. Sundaram, and B.L. Tseng. Analyzing communities and their evolutions in dynamic social networks. ACM Transactions on Knowledge Discovery from Data (TKDD), 3(2) :8, 2009.

[Kim,2009] M. Kim and J. Han. A particle-and-density based evolutionary clustering method for dynamic networks. Proceedings of the VLDB Endowment, 2(1) :622–633, 2009.

[Li , 2012] J. Li, L. Huang, T. Bai, Z. Wang, and H. Chen. Cdbia : A dynamic community detection method based on incremental analysis. In Systems and Informatics (ICSAI), 2012 International Conference on, pages 2224{2228. IEEE, 2012.

[Falkowski,2008] T. Falkowski, A. Barth, and M. Spiliopoulou. Studying community dynamics with an incremental graph mining algorithm. AMCIS 2008 Proceedings, 2008.

[Lemmouchi,2010] S. Lemmouchi M. Haddad and H. Kheddouci. Robustesse de communautes emerges des graphes sociaux issus des reseaux de communication. *EGC*, 2010.

[Tang,2010] Tang, L. and Liu, H. (2010). Community Detection and Mining in Social Media. Synthesis Lectures on Data Mining and Knowledge Discovery. Morgan & Claypool Publishers.

[Talbi,2013] M. TALBI, Une nouvelle approche de detection de communautés dans les réseaux Sociaux, Phd these , UNIVERSITÉ DU QUÉBEC EN OUTAOUAIS , 2013.

[Aynaud,2011] T. Aynaud ,Détection de communautés dans les réseaux dynamiques, THÈSE DE DOCTORAT DE L'UNIVERSITÉ PIERRE ET MARIE CURIE ,2011.

[Cazabet,2013] R. CAZABET , Détection des communautés dynamiques dans des réseaux temporels, THÈSE DE DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE,2013.

[Seifi,2012] Seifi, M. (2012). Coeurs stables de communautès dans les graphes de terrain. PhD thesis, Université Pierre et marie Curie (paris 6).

[Palla,2005] Palla, Derényi, Farkas, and Vicsek. Uncovering the overlapping community structure of complex networks in nature and society, 2005.

[Moeno,1933] J.L. Moreno, Emotions mapped by new geography, New York Times (1933)

[Kanawati,2014] R. Kanawati Detection of communauty in interactions graphes 2014

[Zachary,1977] W.W. Zachary. An information flow model for conflict and fission in small groups. Journal of anthropological research, 1977.

[Lusseau,2003] David Lusseau, Karsten Schneider, Oliver J Boisseau, Patti Haase, Elisabeth Slooten, and Steve M Dawson. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, 54(4) :396–405, 2003.

[KREBS,2008] KREBS V., A network of books about recent US politics sold by the online bookseller amazon.com, http ://www.orgnet.com, 2008.

[Givan et Newman,2004] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69 :026113, 2004.

[Newman,2013] Newman real networks. http://www-personal.umich.edu/~mejn/ netdata/, 2013. consulté le : 10/07/2013.

[Park et Newman,2005] J. Park and M E J Newman. A network-based ranking system for us college football. *Journal of Statistical Mechanics : Theory and Experiment*, 2005(10) :P10014, 2005.

[John,1988] John Scott Social Network Analysis Sociology February 1988 vol. 22 no. 1 109-127

[Milgram,1969] Travers and Milgram. An experimental study of the small world problem, 1969.

[Erdös et Rényi ,1960] Erdös and Rényi. On the evolution of random graphs, 1960.

[Watts et Strogatz,1965] Price. Networks of scientific papers, 1965.

[Duncan,1998] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of small-world networks. 393 :440–442, 1998.

[Kleinberg,2000] Kleinberg. The small-world phenomenon : An algorithmic perspective, 2000.