HAL
open science

Découverte de communautés dans les réseaux complexes

Yacine Slimani, Ahlem Drif

To cite this version:

Yacine Slimani, Ahlem Drif. Découverte de communautés dans les réseaux complexes. 2016. hal-01389844v2

HAL Id: hal-01389844

https://hal.science/hal-01389844v2

Preprint submitted on 1 Jan 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

JOURNAL OF INTERDISCIPLINARY METHODOLOGIES AND ISSUES IN SCIENCE

Méthodes de découverte de communautés dans les réseaux complexes

Yacine SLIMANI[*1] **and Ahlem DRIF**[2]

[1]Laboratoire des Systèmes Intelligents (L.S.I), Université Ferhat Abbas Sétif 1, Algérie
[2]Laboratoire de Réseau et Systèmes Distribuées (L.R.S.D), Université Ferhat Abbas Sétif 1, Algérie

*Correspondance : slimani_y09@univ-setif.dz

DOI : 10.18713/JIMIS-ddmmyy-v-a
Soumis le 01 Janvier 2019 - Publié le Jour Mois-en-lettres Année
Volume : **5** - Année : **2019**
Titre du numéro : **Analyse de graphes et réseaux**
Éditeurs : *Vincent Labatut, Didier Josselin*

Résumé

La détection de communautés dans les réseaux complexes fait l'objet de plusieurs recherches qui ont été proposées pour découvrir la structure du réseau et d'analyser les propriétés du réseau. Dans cet article, nous donnons un aperçu complet des différentes stratégies de découverte de communauté, nous proposons une taxonomie de ces méthodes, et les différences entre ces dernières qui aident les concepteurs à comparer et choisir la stratégie la plus appropriée pour les différents types de réseaux rencontrés dans le monde réel.

Mots-Clés

découverte de communautés ; réseaux complexes

I INTRODUCTION

Beaucoup de systèmes complexes du monde réel peuvent être représentés et étudiés en tant que réseaux. Les réseaux complexes recouvrent ainsi des réseaux aussi divers que le réseau Internet, les réseaux des contacts sociaux entre individus Scott (2012), les réseaux des réactions chimiques entre protéines dans le métabolisme d'un être vivant Holme *et al.* (2003); Jeong *et al.* (2000), les réseaux des pages web Albert *et al.* (1999) qui contiennent plusieurs millions de noeuds, les réseaux trophiques Williams et Martinez (2000), les réseaux de dictionnaires Blondel et Senellart (2002) ... et bien d'autres.

Les études menées sur la signification physique et les propriétés mathématiques des réseaux complexes ont constaté que ces réseaux partagent des propriétés macroscopiques. Parmi ces propriétés, on cite des propriétés prototypes telle que l'effet petit-monde Watts et Strogatz (1998) et l'échelle-libre Barabási et Albert (1999), des propriétés dynamiques tel que la diffusion Bilke et Peterson (2001); Eriksen *et al.* (2003) et des propriétés structurelle comme la

structure de communauté Moody (2001); Flake *et al.* (2002); Girvan et Newman (2002); Lusseau et Newman (2004); Krause *et al.* (2003); Guimera et Amaral (2005); Palla *et al.* (2005). La propriété de structure de communautés paraît être commune à beaucoup de réseaux complexes et permet de comprendre la relation entre un simple noeud dans la microscopie et des groupes dans la macroscopie. Par conséquent, la découverte de structure de communautés a fait l'objet de plusieurs récents efforts. Il s'agit d'une problématique proche des problématiques classiques de clustering de données et de partitionnement de graphe.

Les méthodes de découverte de communautés supposent que le réseau se divise naturellement en un ensemble de sous-groupes et visent la détection de ces groupes (communautés). Les critères utilisés pour détecter correctement la structure de communautés sont très cruciaux et divers ce qui justifie le nombre important des méthodes de découverte de communautés proposées. Cet article est un état de l'art des méthodes existant pour la détection de communautés .

II CONTEXTE ET MOTIVATIONS

2.1 La théorie des graphes classique est-elle appropriée aux réseaux du monde réel ?

L'étude des réseaux sous la forme de théorie des graphes est l'un des piliers fondamentaux des mathématiques discrètes. La résolution d'Euler, en 1735, du problème de ponts de Königsberg est considérée comme le premier théorème de la théorie des graphes. Au 20 ème siècle la théorie des graphes s'est développée en tant que domaine substantiel de la connaissance et les graphes sont également devenus extrêmement utiles comme représentation d'une grande variété de systèmes dans différents secteurs tels que les réseaux biologiques, sociaux, technologiques, et de l'information. Ainsi, l'analyse des graphes est devenue cruciale pour comprendre ces réseaux du monde réel. Les réseaux ont été également étudiés intensivement dans les sciences sociales en se basant sur l'usage des graphes dont les sommets représentent les individus ou les organisations sociales et les liens désignent les interactions sociales entre eux. Des études sont, par exemple, menées sur les propriétés de centralité et de connectivité.

Ces dernières années, vue la disponibilité croissante des données à grande échelle, l'étude des réseaux a été changé de l'analyse des simples graphes et des propriétés des sommets individuels à l'analyse des propriétés statistiques des graphes complexes. La théorie des graphes classique a été concernée par des problèmes des réseaux réels mais son approche qui est orientée vers la conception n'est pas approprié aux réseaux surgissant dans le monde réel. Plusieurs questions qui ont été précédemment posées dans les études de petits réseaux ne peuvent pas être utile dans des grands réseaux, par exemple, l'analyste d'un réseau social pourrait demandé : "quel noeud affecte-il la connectivité du réseau s'il est retiré ?", mais une telle question a peu de signification dans des réseaux qui contiennent des millions de sommets (car dans des tels réseaux la suppression d'un seul sommet n'aura aucun effet). Désormais, la question qui devrait être posée : " Quel est le pourcentage des sommets a enlevé pour affecter considérablement la connectivité du réseau ?" et ce type de questions statistiques a une concrète signification dans les réseaux du monde réel.

C'est ainsi qu'un groupe divers de scientifiques, y compris des mathématiciens, physiciens, informaticiens, sociologues, et biologistes, avaient activement poursuivi ces questions et avaient fondé le nouveau champ de la théorie des réseaux, ou la "science des réseaux" Albert-Laszlo (2002); Buchanan (2003); Watts (2004). Une littérature significative s'est déjà accumulée dans ce nouveau domaine interdisciplinaire qui se penche sur l'étude et la découverte des propriétés que partagent un grand nombre de grands réseaux complexes Watts et Strogatz (1998).

2.2 Quelle opportunité y a-t-il pour une nouvelle science interdisciplinaire des réseaux ?

Cette science se distingue des travaux précédents sur les réseaux de trois manières importantes :

— Elle se focalise sur les propriétés des réseaux du monde réel telles que la longueur des chemins, le degré de distribution, et le comportement du système pour proposer des mesures appropriés à ces propriétés.
— Elle vise l'extraction des modèles qui permettent la compréhension approfondie des propriétés des réseaux du monde réel.
— Elle étudie la prédiction de la dynamique de comportement des systèmes en considérant les propriétés du réseaux complexes influant les différents acteurs des réseaux.

2.3 Quelles sont les propriétés que partagent un grand nombre de réseaux complexes ?

Les réseaux complexes que l'on peut rencontrer dans les différentes disciplines n'ont, à première vue, pas de raison de se ressembler. Cependant, plusieurs études ont révélé l'existence de caractéristiques communes et significatives Watts et Strogatz (1998); Strogatz (2001); Albert et Barabási (2002); Newman (2003b); Dorogovtsev et Mendes (2002). Nous citons brièvement les propriétés communes les plus étudiées dans les réseaux complexes :

2.3.1 L'effet petit monde "The small-world effect"

L'effet petit monde tient son nom de l'expression populaire "le monde est petit" désignant la surprise de constater que deux connaissances d'un même individu, a priori sans rapport, se connaissent entre elles. La notion petit monde est définie, dans certains articles Watts et Strogatz (1998) comme la combinaison d'un fort coefficient de clustering et d'un petit diamètre. Cette propriété étudiée par le psychologue Milgram Milgram (1967) est vérifiée par le modèle de graphes aléatoires d'Erdös-Rényi Erdös et Rényi (1959). Pour pallier aux limites de modèle d'Erdös-Rényi, plusieurs travaux ont été publiés Bollobás (1998).

2.3.2 Clustering "Transitivity or clustering"

Une des propriétés essentielles des réseaux complexe est l'existence d'une forte densité locale qui s'oppose à la faible densité globale du graphe. Cette densité est souvent mesuré par le coefficient de clustering Watts et Strogatz (1998). Il s'agit de la moyenne, sur tous les nœuds u, du ratio du nombre de voisins de u qui sont reliés entre eux sur le nombre total de liens qui pourraient potentiellement exister entre ces voisins (probabilité que deux voisins de u soient reliés). Cette propriété illustre la tendance des acteurs à se regrouper en modules ou communautés.

2.3.3 Distribution des degrés "Degree distributions "

Le modèle usuel utilisé au départ pour ce domaine d'étude était un réseau aléatoire uniforme, sur lequel on observe un effet de seuil pour la transmission d'un virus, c'est-à- dire qu'en dessous d'une fraction d'individus infectés, le virus cesse de se répandre. Mais une étude similaire menée sur un modèle présentant une distribution de degrés en loi de puissance a donné des résultats différents, en particulier l'effet de seuil disparaît. En 1999, Faloutsos et al Faloutsos *et al.* (1999) ont observé que le réseau Internet présentait cette propriété. Une telle observation a donc remis en cause les mécanismes mis en place pour freiner la propagation des virus et les modèles utilisés jusqu'alors. Des modèles suivant une loi de puissance sont donc utilisés, car c'est cette distribution qui est retrouvée la plupart du temps dans tous les réseaux réels Newman (2003b). Deux distributions de degrés sont connues : une distribution homogène des degrés des noeuds (selon une loi de Poisson), et une distribution hétérogène des degrés des noeuds (selon une loi de Puissance). La distribution selon une loi de puissance est donnée comme suit : le

nombre P_k de sommet de degré k est proportionnelle à $k^{-\alpha}$, pour une constante $\alpha > 0$ sur un intervalle de plusieurs ordres de grandeur (par exemple entre $k = 10$ et $k = 10^6$).

2.3.4 *Résilience des réseaux "Network resilience"*

La propriété de résilience des réseaux est liée à la distribution de degrés. Quand il y a une suppression de sommets, la longueur des chemins augmentera, en conséquence, des paires de sommets devenus déconnectées et la communication entre elles deviendra impossible. Le niveau de résilience de réseau se varié selon la connectivité des sommets. Un intérêt particulier pour l'étude de la résilience de réseau a été soulevé par le travail d'Albert et al Albert *et al.* (2000).

2.3.5 *Mixing patterns*

Dans la plupart des réseaux il existe des types différents de sommets et la probabilité qu'il existe un lien entre une paire de sommets différentes dépend souvent du type de la relation. Autrement dit, mixing patterns se réfère aux tendances systématiques d'un type de nœuds dans un réseau pour se connecter à un autre type par exemple Maslov et al Maslov *et al.* (2004) ont étudié l'existence de trois type de nœud dans le réseau Internet : les fournisseurs qui ont une forte connectivité, les consommateurs qui sont les utilisateurs finaux, et les providers de services Internet qui jouent le rôle de relais entre les deux types de nœuds précédents.

2.3.6 *Degré de corrélation "correlation degree"*

La corrélation de degré peut fournir des détails intéressants sur la structure du réseau. En fait, cette propriété nous permet de savoir si les sommets ayant un degré élevé sont de préférence connectés à d'autres sommets avec un degré élevé, ou sont plutôt connectées à des sommets ayant un faible degré. Plusieurs études ont été proposées pour quantifier le degré de corrélation à l'exemple des travaux de Maslov et al Maslov *et al.* (2004); Maslov et Sneppen (2002).

2.3.7 *Navigabilité "Network navigation"*

Kleinberg Kleinberg (2000) a proposé le premier modèle de petit monde présentant la propriété de navigabilité, c'est-à-dire le premier modèle de graphe dont le diamètre est polylogarithmique en nombre de noeuds et dont des chemins polylogarithmiques peuvent être découverts par un algorithme décentralisé entre tout couple de sommets. A l'exemple de la navigation à travers le réseau des pages Web qui se faisait d'une page à l'autre sans connaître la carte globale du réseau Albert *et al.* (1999); Kaiser (1999). Par ailleurs, la découverte des chemins de façon décentralisée est très rentable pour les réseaux d'interactions qui contiennent un très grand nombre de noeuds car une recherche classique des plus courts chemins est très coûteuse en temps.

2.3.8 *Structure de communautés "Community structure"*

Dans les réseaux complexes, la présence de groupes de sommets fortement liés entre eux et faiblement liés avec l'extérieur fonde la propriété de structure de communautés. Les communautés sont des groupes de sommets qui partagent probablement des propriétés communes ou des rôles semblables dans le réseau complexe. Ainsi, les communautés peuvent correspondre aux groupes de pages Web traitant le même sujet Flake *et al.* (2002), aux modules fonctionnels dans les réseaux métaboliques Guimera et Amaral (2005); Palla *et al.* (2005), aux groupes d'individus dans les réseaux sociaux Girvan et Newman (2002); Lusseau et Newman (2004), et aux subdivisions dans les chaînes alimentaires Pimm (1979); Krause *et al.* (2003), ...ect. La figure 1 montre une visualisation du réseau d'amitié des enfants dans une école au USA selon l'étude de Moody. Moody Moody (2001) a coloré les sommets selon la race de chaque individu.

Ce réseau semble avoir une forte structure de communauté qui en résulte principalement à cause de la propriété de race des individus.

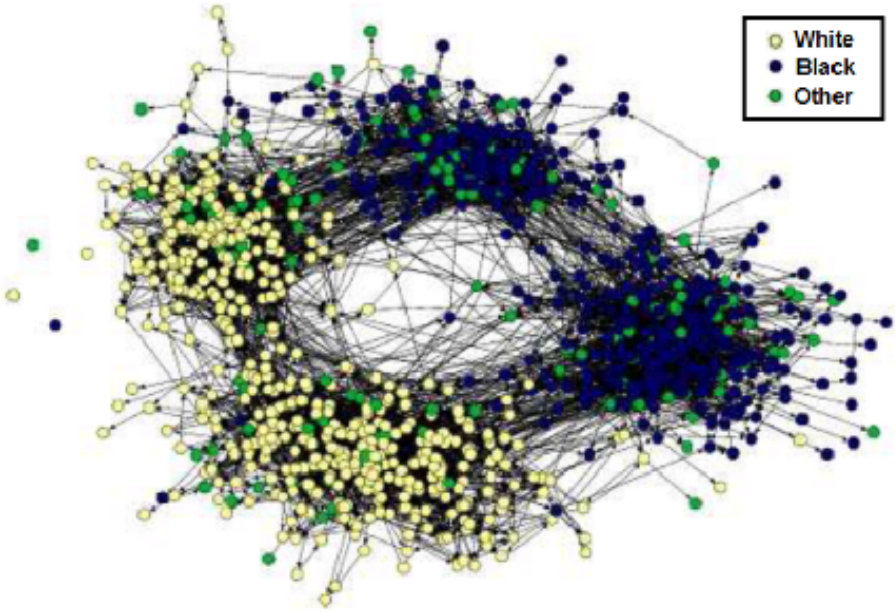

FIGURE 1 – Réseau d'amitié des enfants dans une école aux USA. Les liens d'amitiés sont déterminées par les participants, et par conséquent ils sont orientés. Les sommets sont colorés selon la race.

2.4 Quel sont les objectifs du processus de découverte de communautés dans les réseaux complexes ?

Les principaux objectifs qui ont motivés les études sur les méthodes de découverte de communautés dans les réseaux du monde réel sont les suivants :

— La détection de communautés est un outil important pour la compréhension des structures et des fonctionnements des systèmes complexes.

— Les communautés donnent un point de vue macroscopique sur la structure des graphes. Elles permettent par exemple de regrouper et d'identifier les sommets qui jouent des rôles similaires. Par exemple, la détection de communautés dans le graphe du Web est une piste envisagée pour améliorer les moteurs de recherche Flake *et al.* (2002).

— La détection de communautés peut aussi être utilisée pour la visualisation des graphes complexes Auber *et al.* (2003).

— Les méthodes de détection de communautés sont utilisées pour le partitionnement du graphe afin d'effectuer des calculs séparés moins coûteux sur chaque communauté. Ce procédé de parallélisme permet d'envisager des gains en complexité du temps pour les grands graphes.

— La détection de communautés permet une classification des sommets, selon leur position topologique dans le graphe. Ainsi, les sommets de position centrale dans leur clusters partagent un grand nombre de liens avec les autres groupes, en conséquence, ils peuvent avoir une importante fonction de contrôle et de stabilité au niveau du groupe. Quand aux sommets de frontière, ils jouent un rôle important de relais entre les différentes communautés. Une telle classification est très utiles dans les réseaux sociaux et les réseaux métaboliques Granovetter (1973); Burt (1976); Freeman (1977).

— Les communautés peuvent être utilisées pour améliorer les méthodes de compression de graphes à l'exemple du travail Mahdian *et al.* (2006).

III DESCRIPTION DES COMMUNAUTÉS

Une définition naturelle des communautés stipule qu'une communauté est dense, c'est-à-dire que ses membres sont fortement connectés entre eux et que, dans le même temps, ils sont peu liés à des membres en dehors de la communauté. Le problème de la détection de communautés est donc naturellement formalisé en la recherche d'une partition d'un graphe en sous-groupes denses peu connectés entre eux. Définir les communautés comme les parties d'une partition est fréquemment admis, mais cela implique qu'un nœud n'appartient qu'à une et une seule communauté. Il existe des définitions de communautés où un nœud peut appartenir à diverses communautés. On appelle de telles communautés des communautés recouvrantes. Néanmoins, aucune définition ne fait aujourd'hui consensus, il existe peu d'algorithmes pour les détecter

3.1 Définition des communautés

En dépit de la grande quantité d'étude dans ce domaine, un consensus sur ce qui est la définition d'une communauté n'a pas été atteint. Conceptuellement, les définitions de communauté se basent sur la notion de sous graphe et peuvent être séparées en deux catégories : les définition comparatives et les définition de référence individuel. Dans ce qui suit, nous en citons quelques exemples :

3.1.1 Définitions comparatives

La comparaison est effectué le plus souvent en terme de liens internes et externes dans chaque communauté et parfois des auteurs comparent des critères de similarités pour pouvoir détecter la structure de communautés.

Définition 1 : Wasserman et Faust (1994)

Une communauté peut être décrite comme collection de sommets dans un graphe qui sont fortement reliés entre eux-mêmes mais faiblement relié du reste du graphe.

Définition 2 : Radicchi *et al.* (2004)

Soit A_{ij} la matrice d'adjacence du graphe G ; Le degré k_i d'un noeud $i \in G$ est :

$$k_i = \sum_{j \in G} A_{ij} \tag{1}$$

Soit un sous graphe $V \subset G$ et $i \in V$, le degré total est donné par :

$$k_i(V) = k_i^{in}(V) + k_i^{out}(V) \tag{2}$$

Tel que :

Le nombre de liens reliant le noeud i à d'autres noeuds appartenant à V :

$$k_i^{in}(V) = \sum_{j \in V} A_{ij} \tag{3}$$

Le nombre de liens vers les noeuds qui n'appartiennent pas à V (le reste du réseau) :

$$k_i^{out}(V) = \sum_{j \notin V} A_{ij} \tag{4}$$

Définition d'une communauté au sens fort :

Le sous-graphe V est une communauté au sens fort si :

$$k_i^{in}(V) > k_i^{out}(V), \forall i \in V \tag{5}$$

Une communauté est définie en tant qu'un ensemble de noeuds dans lequel chaque noeud a plus de connexions au sein de cette communauté qu'avec le reste du réseau.

Définition d'une communauté au sens faible :

Le sous graphe V est une communauté au sens faible si :

$$sum_{i \in V} k_i^{in}(V) > \sum_{i \in V} k_i^{out}(V) \tag{6}$$

Une communauté est définie comme un ensemble de noeuds dont le nombre total de liens internes est supérieur au nombre total des liens vers l'extérieur.

Définition 3 :

Les communautés sont des groupes de sommets qui sont similaires les uns aux autres. Un critère est choisi pour l'évaluation de la similarité.

3.1.2 Définitions de référence individuelle

Définition 1 : La communauté est une clique, définie en tant que sous-groupe d'un graphe contenant plus de deux noeuds où tous les noeuds sont reliés entre eux au moyen de liens dans les deux directions (c'est un sous graphe entièrement connecté). Les triangles sont les cliques les plus simples, et sont fréquentes dans les réseaux du monde réel mais les plus grandes cliques sont rares, ainsi elles ne sont pas de bons modèles de communautés. En outre, l'utilisation de l'algorithme de Bron-Kerbosch Bron et Kerbosch (1973) pour trouver les cliques résulte d'un coût de calcul élevé (complexité exponentielle).

Définition 2 : Newman (2006) Une communauté est un sous graphe indivisible.

3.2 Représentation graphique des communautés

La théorie des graphes est employée pour représenter le réseau et même les communautés dans le réseau en question. Les dendrogrammes sont aussi souvent utilisés pour illustrer la progression entière de l'algorithme de découverte de communautés et le regroupement des sommets depuis le graphe initial au graphe résultant partitionné en communauté comme le montre la figure 2. Les coupes horizontales à travers l'arbre hiérarchique représentent clairement toutes les divisions possibles en communautés à chaque niveau.

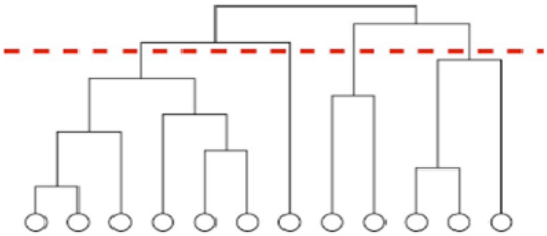

FIGURE 2 – Dendrogramme d'un algorithme de détection de communautés

3.3 Mesures de la qualité de partition d'un réseau en communauté

Comment savoir si les communautés détectées sont bonnes ou non et comment évaluer une telle partition ? Quelle est la meilleure partition pour le réseau en question ? A quel niveau on coupe le dendrogramme pour obtenir la partition adéquate du réseau ou bien le nombre de communautés appropriées ? Pour répondre à ces questions, Newman et al Newman et Girvan (2004) Newman (2003a). ont introduit une mesure de la qualité de partition du réseau appelé "modularité".

Supposons une partition particulière d'un réseau en k communautés. Soit e une matrice symétrique $k \times k$, ses éléments e_{ij} représente la fraction de tous les liens dans le réseau qui relient les sommets de la communauté i aux sommets de la communauté j.

La trace de la matrice e : $Tr_e = \sum_i e_{ii}$ représente la fraction de tous les liens qui relient les sommets dans les mêmes communautés. Une valeur élevée de la trace indique une bonne partition en communautés.

La somme de n'importe quelle ligne (ou colonne) de e : $a = \sum_j e_{ij}$ correspond à la fraction de tous les liens reliés aux sommets de la communauté i.

Si le réseau ne possède pas la propriété de structure de communauté, la valeur prévue des fractions des liens dans une partition peut être estimée. C'est la probabilité qu'un sommet d'extrémité d'un lien soit dans la communauté i, donc a_i, multiplier par la fraction des liens qui se termine par un sommet dans la communauté i, donc a_i . On peut alors écrire : $e_{ij} = a_i.a_i$, ce qui représente le nombre des liens intra-communautés prévus.

Ainsi, la mesure de modularité est définit comme suit :

$$Q = \sum_i (e_{ii} - a_i^2) = Tr_e - \|e^2\| \tag{7}$$

La modularité permet de comparer deux partitions d'un même graphe mais pas vraiment des partitions de graphes différents. Elle n'est pas une mesure absolue de qualité, dans le sens où la meilleure partition pour un graphe n'aura pas la même modularité que la meilleure partition pour un autre graphe. Cependant, il est possible que les partitions de meilleures modularités ne correspondent pas aux partitions en communautés les plus pertinentes Fortunato et Barthelemy (2007).

IV DOMAINES D'APPLICATION DES MÉTHODES DE DÉCOUVERTE DE COMMUNAUTÉS

La nature interdisciplinaire de la nouvelle théorie de réseaux vient de la diversité des réseaux du monde réel. Ces réseaux complexes possèdent des propriétés communes et soulèvent des problématiques similaires. Une de ces problématiques est la découverte d'une structure significative de communautés qui constitue un backbone fondamental pour bien comprendre les interactions des réseaux complexes. Dans cette section, nous citons quelques exemples des réseaux complexes qui sont caractérisés par une structure de communautés.

4.1 Réseaux sociaux

Les réseaux sociaux constituent un champ d'application ancien et important Wasserman et Faust (1994) dans lequel les acteurs sont des individus ou entités sociales (associations, entreprises, pays,...etc)et les liens entre eux peuvent être de différentes natures. Il existe plusieurs types de réseaux sociaux : les réseaux de connaissance (deux individus sont reliés s'ils se connaissent), les réseaux de collaboration (deux individus sont reliés s'ils ont travaillé ensemble) , en particulier, de nombreux travaux ont étudié les collaborations scientifiques Newman (2001)), les réseaux d'appels téléphoniques Resende (2000) (deux individus ou numéros de téléphones sont reliés s'il y a eu un appel entre eux), les réseaux d'échanges (deux entités sont reliées si elles ont échangé un fichier Guillaume *et al.* (2004) ou un courrier électronique Ebel *et al.* (2002) par exemple), ...etc.

4.2 Réseaux biologiques

Les réseaux biologiques sont assez divers parmi lesquels il existe les réseaux métaboliques Jeong *et al.* (2000) (les sommets sont des gènes ou des protéines qui sont liés selon leurs interactions chimiques), les réseaux de neurones (chaque neurone est connecté à plusieurs autres neurones) ou les réseaux trophiques Williams et Martinez (2000) (les espèces d'un écosystème sont reliées pour représenter les chaînes alimentaire). Un exemple d'une chaîne alimentaire Martinez (1991) est illustré sur la figure 3.

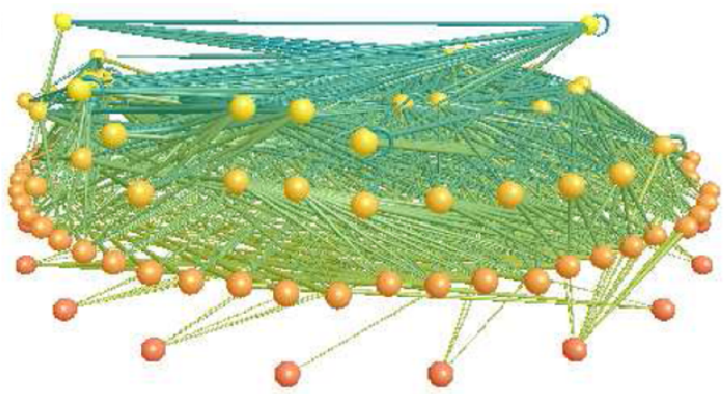

FIGURE 3 – Une chaîne alimentaire des interactions de prédateur-proie entre les espèce dans un lac.

4.3 Réseaux d'information

L'exemple classique du réseau d'information est le réseau de citation des travaux scientifiques. La plupart des articles citent les travaux précédents des auteurs sur le même sujet (voir 4). Ces citations forment un réseau dont les sommets sont des articles ; un lien orienté de l'article A vers l'article B indique que A cite B. La structure d'un réseau de citation reflète la structure de l'information stockée dans ses sommets.

FIGURE 4 – Réseau de citations

Un autre exemple très important du réseau d'information est le réseau World Wide Web, dont les pages Web contenant l'information, reliée ensemble par les lien hypertextes d'une page vers l'autre Egghe et Rousseau (1990).

4.4 Réseaux technologiques

Les réseaux technologiques sont des réseaux synthétiques conçus typiquement pour la distribution d'un certain produit ou ressource, telle que l'électricité. La grille d'énergie électrique est un bon exemple des réseaux technologiques. Plusieurs études statistiques ont été menées sur la grille d'électricité par Watts et Strogatz Watts et Strogatz (1998); Watts (1999) et Amaral et al Amaral *et al.* (2000). Nous pouvons aussi citer d'autres réseaux de distribution tels que le réseau des itinéraires de ligne aérienne Amaral *et al.* (2000), les réseaux des routes Kalapala *et al.* (2003), réseau de chemins de fer Sen *et al.* (2003); Latora et Marchiori (2002), et les réseaux sans fil mobiles Drif *et al.* (2014). Les réseaux de fleuve sont aussi considérés comme des réseaux de distribution Dodds et Rothman (2000).

4.5 Réseaux linguistiques

Ces réseaux relient les mots d'un langage donné, à l'exemple des réseaux de synonymes (deux mots sont reliés s'ils sont synonymes), des réseaux de co-occurrences Ramon et Solé (2001) (deux mots sont reliés s'ils apparaissent dans une même phrase d'un ouvrage) ou encore des réseaux de dictionnaires Blondel et Senellart (2002) (deux mots sont liés si l'un est utilisé dans la définition de l'autre).

V CLASSIFICATION DES MÉTHODES DE DÉCOUVERTE DE COMMUNAUTÉS

La méthode proposée par Girvan et Newman Girvan et Newman (2002) a marqué le début d'une nouvelle ère dans le domaine de la découverte de communautés dans les réseaux complexes. Depuis ce travail de référence, le sujet a reçu une extraordinaire attention de la part de la communauté scientifique et de très nombreuses nouvelles approches ont été sans cesse proposées. La détection de communautés s'approche des deux thématiques classiques en informatique qui sont le partitionnement de graphe et le clustering de données. Le problème de détection de communautés peut être vu comme un problème de clustering de données pour lequel il faut choisir une distance adéquate. Cependant, les graphes considérés par les applications de clustering usuelles ne possèdent pas les caractéristiques spécifiques des graphes complexes. Par conséquent de nombreuses approches classiques de clustering de données sont inadaptées pour la détection de communautés. Dans cette section, nous décrivons les approches de détection de communautés existantes. Notre but est de donner une classification des méthodes proposées, d'en illustrer la diversité, et de discuter leurs avantages et leurs inconvénients.

L'étude de découverte de communauté dans les réseaux complexes connaît une continuelle évolution et les auteurs proposent sans cesse des nouvelles approches pour la détection des communautés Fortunato *et al.* (2004); Guimera *et al.* (2004a); Lehmann *et al.* (2008); Clauset *et al.* (2004); Boccaletti *et al.* (2007a); Newman et Leicht (2007); Eckmann et Moses (2002a); Sales-Pardo *et al.* (2007); Zhang *et al.* (2008); Fortunato et Castellano (2008); Danon *et al.* (2005). Bien que la liste des méthodes présentées dans ce chapitre est importante elle n'est pas exhaustive. de ce fait, nous avons essayé de généraliser notre classification selon les différentes démarches utilisées. Notre classification permet d'expliciter les différents choix d'une approche possible pour la découverte de communautés et décider la quelle est la plus convenable pour un réseau étudié. Un choix convenable de la méthode de découverte à utilisée apporte un gain considérable en terme de temps d'exécution et de qualité de partition. Notre classification repose sur les trois points de vue suivants :

1- Selon la manière de regroupement des noeuds en groupes, Jain et Dubes Jain et Dubes (1988) ont distingué deux approches pour ce faire : agglomérative et séparative. De même, nous avons remarqué que toutes les méthodes de détection de communauté utilisent soit une approche séparative soit une approche agglomérative pour regrouper les noeuds en communautés.
2- Dans le contexte d'évaluation des performances des méthodes de détection de communautés, nous avons constaté que les méthodes déterministes et stochastiques se diffèrent dans leurs apports en terme de complexité en temps et de qualité de partition. A cet effet, il est fort important de distinguer les méthodes déterministes de celles stochastiques.
3- Les approches de découverte de communautés caractérisent les communautés directement ou indirectement, par des propriétés globale du graphe, comme l'intermédiarité, la centralité,.. etc., ou par l'emploie de certains processus comme les promenades aléatoires, la synchronisation,..ect . Les communautés peuvent être également interprété en tant qu'une forme d'organisation topologique du graphe. Ainsi, les différentes démarches pour caractériser les communautés nous ont permis de classifier les méthodes existantes selon la technique utilisée au cours de la découverte de communautés.

VI MÉTHODES AGGLOMÉRATIVES

Dans les méthodes agglomératives Zhou (2003b); Pons et Latapy (2006); Zhou et Lipowsky (2004); Guimera *et al.* (2004a); Reichardt et Bornholdt (2004, 2006); Arenas *et al.* (2006); Boccaletti *et al.* (2007a); Newman (2004); Clauset *et al.* (2004); Shen *et al.* (2009); Palla *et al.* (2005, 2007); Farkas *et al.* (2007); Lehmann *et al.* (2008); Donetti et Munoz (2004); Donetti et Muñoz (2005); Jiang *et al.* (2009); Eckmann et Moses (2002a); Bagrow et Bollt (2005); Zhang *et al.* (2008), les métriques de similarité entre les pairs de sommets sont calculées au moyen de plusieurs méthodes, et par conséquent les liens , qui relie les pairs de sommets de forte similarité, sont ajoutés progressivement au réseau initial. Ce processus d'ajout de liens peut être arrêté à n'importe quel niveau et les composants connectés obtenus représentent les communautés. La figure 5 illustre la classification des méthodes agglomératives. Cependant ces méthodes identifient les noyaux de communautés et n'incluent pas les noeuds périphériques. Les noeuds de noyau dans une communauté ont souvent une forte similarité, et par conséquent, ils sont reliés tôt dans le processus agglomératif, mais les noeuds périphériques sont négligés (ils ne sont pas mis dans la communauté appropriée)Newman et Girvan (2004).

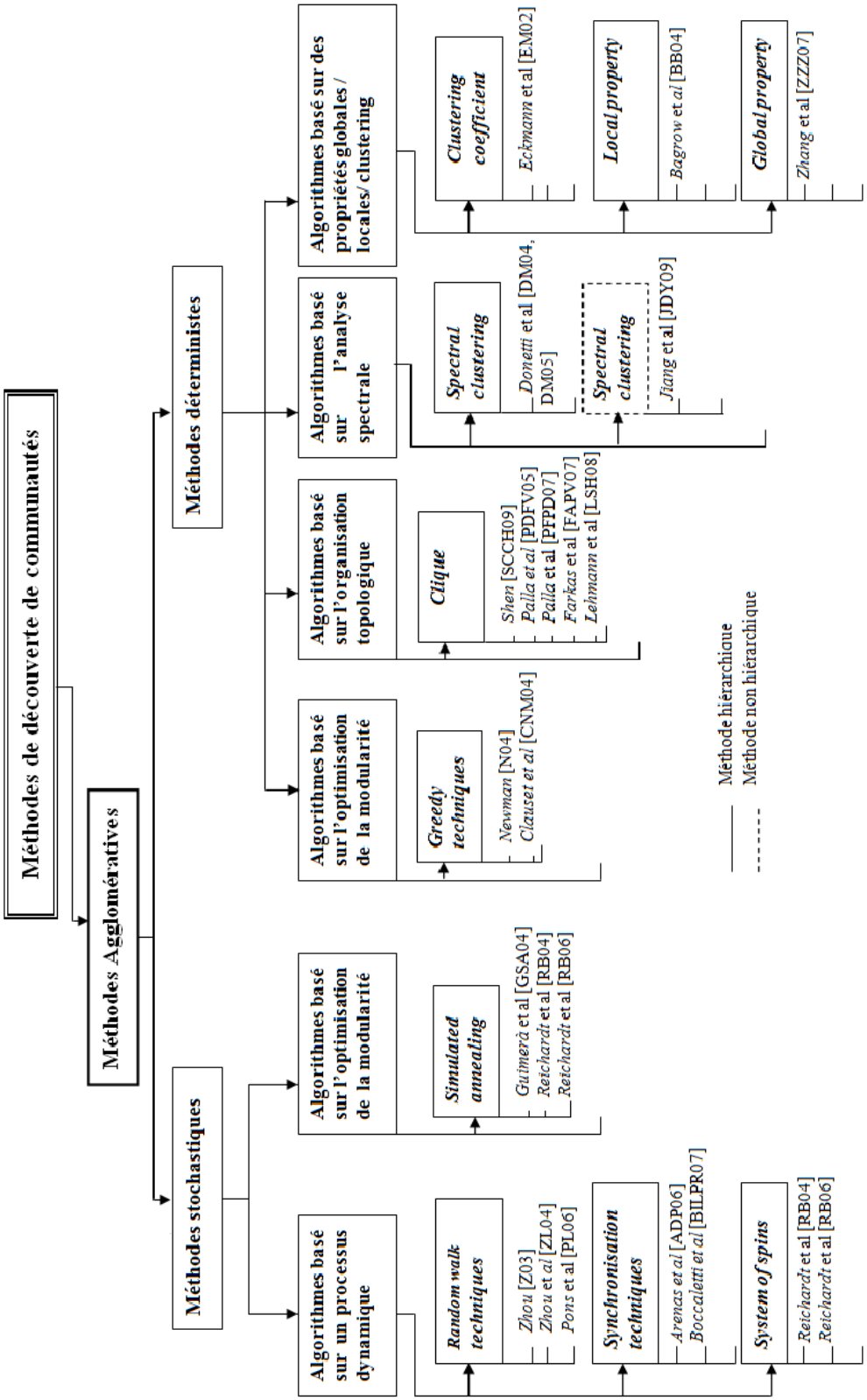

FIGURE 5 – Taxonomie des méthodes de découverte de communautés dans les réseaux complexes : 1. Méthodes agglomératives.

6.1 Méthodes basées sur l'optimisation de la modularité

Dans le travail Newman (2003b), les auteurs ont montré que des valeurs élevées de modularité indiquent de bonnes partitions en communautés, ce qui a motivé la proposition de nombreuses méthodes pour la maximisation de la modularité.

6.1.1 Les méthodes gloutonnes

Newman Newman (2004) a proposé d'optimiser la valeur de modularité sur toutes les partitions possible afin de trouver la meilleur partition en communautés. Dans le Fast algorithm Newman (2004), la mesure utilisée pour découvrir les communautés n'est que le changement de Q quand on joint deux communautés. Cette méthode d'optimisation emploie un algorithme d'optimisation glouton. L'algorithme se déroule comme suit :

1. Initialement, on considère chaque communauté se compose d'un seul sommet.
2. Joindre la paire des communautés qui résulte d'une division qui maximise la valeur de modularité, mais ne pas joindre la paire des communautés entre lesquelles il n'existe pas de liens. (Utilisation d'un algorithme glouton)
3. La mis à jours des éléments de la matrice e_{ij} en ajoutant des lignes et des colonnes qui correspondent aux communautés jointes.
4. Répéter l'étape 2 jusqu'à ce que la modularité Q ne puisse pas être améliorée.

Afin d'étudier la performance des algorithmes de détection de communauté, plusieurs réseaux du monde réel ont été utilisés. Le travail Newman (2004) utilise un graphe de $n = 128$ sommets divisés en quatre communautés de 32 sommets chacune. Les liens ont été établit aléatoirement entre les paires de sommets : la probabilité qu'un lien relie les sommets dans la même communauté est P_{in}, la probabilité qu'un lien relie les sommets dans des communautés distinctes est P_{out}, et les valeurs de probabilité ont été choisit d'une façon que le degré prévu de chaque sommet est égale à 16. Le nombre moyen des liens (inter-communauté) qui relient un sommet aux sommets de n'importe quelle autre communauté est égale à z_{out}. Les résultats des sommets correctement identifiés selon la variation de z_{out} sont illustrés sur Fig. 6.
L'algorithme Fast identifie correctement plus que 90% des sommets pour des valeurs $z_{out} \leq 6$. Cependant, quand les liens intra-communauté et les liens inter-communauté par sommet deviennent égaux (z_{out} est proche de la valeur 8), nous constatons une dégradation des la performances de l'algorithme. L'algorithme GN Newman (2004) identifie correctement les sommets mieux que fast algorithm Newman (2004) pour des petites valeurs z_{out} (pour $z_{out} = 5$: GN détecte correctement 98.9% de sommets et l'algorithme fast détecte 97.4% de sommets). l' algorithm fast s'opère mieux que l'algorithme GN Pour des valeurs plus élevées de z_{out}.

Les décisions de l'algorithme fast se base sur les informations locales des différentes communautés, tandis que l'algorithme GN emploie les informations du réseau entier.

6.1.2 Méthode de recuit simulé

Guimer et al. Guimera *et al.* (2004b) ont utilisé un algorithme de recuit simulé pour l'optimisation de la modularité. L'idée est d'effectuer un mouvement selon une distribution de probabilité qui dépend de la qualité des différents voisins ; les meilleurs voisins ont une probabilité plus élevée alors que les moins bons ont une probabilité plus faible. L'algorithme de recuit simulé converge rapidement vers la solution optimale mais il est plus efficace pour les petits réseaux.

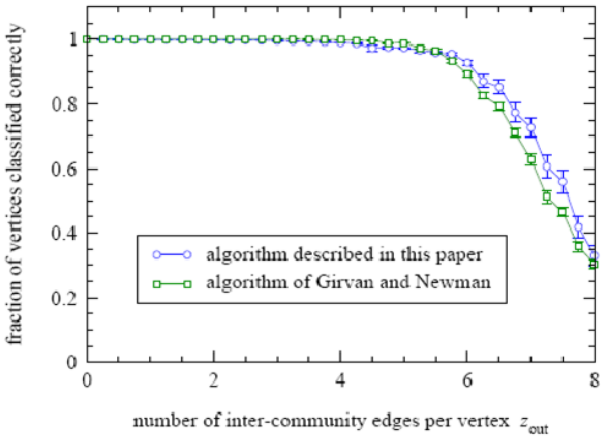

FIGURE 6 – Les résultats des noeuds correctement identifiés selon la variation de z_{out}.

6.2 Méthodes basées sur un processus dynamique

6.2.1 Les techniques de marche aléatoire

Les marches aléatoires dans les graphes sont des processus aléatoires dans lesquels un marcheur est positionné sur un sommet du graphe et peut à chaque étape se déplacer vers un des sommets voisins. Le comportement des marches aléatoires est étroitement lié à la structure du graphe. A cet effet, plusieurs approches de détection de communautés se basent sur ces comportements.

Pons et al Pons et Latapy (2006) ont proposé l'algorithme "Walktrap" en utilisant le constat intuitif que les marches aléatoires vont se faire piéger dans des zones denses, en d'autres termes lorsqu'un marcheur sera dans une communauté il possédera une forte probabilité de rester dans la même communauté à l'étape suivante (grâce à la forte densité de liens internes et la faible densité de liens externes). Une marche aléatoire dans un graphe G est un processus en temps discret sur l'ensemble des sommets V. Sa matrice de transition P est donnée par : $P_{ij} = \frac{A_{ij}}{d(i)}$.

P^t_{ij} : est la probabilité d'aller d'un sommet i à un sommet j par une marche aléatoire de longueur t. Le temps est discrétisé ($t = 0, 1, 2, ...$) et un marcheur est localisé à chaque instant t sur un sommet du graphe G. Le marcheur se déplace à chaque instant aléatoirement et uniformément vers l'un de ses sommets voisins. La suite des sommets visités constitue une marche aléatoire. Ainsi un marcheur possède de grandes chances de rester lors d'une marche de courte distance dans sa communauté d'origine.
L'idée pour comparer la proximité de deux sommets est alors de comparer les distributions de probabilité des marches aléatoires partant de ces deux sommets. La propriété de réversibilité des marches aléatoires indique que les probabilités P^t_{ij} et P^t_{ji} sont directement reliées ; elles sont donc porteuses de la même information. Toute l'information des marches aléatoires concernant un sommet donné $i \in V$ est contenue dans les probabilités $\left(P^t_{kj}\right)_{k \in V}$. Ces probabilités correspondent à la i^{eme} ligne de la matrice P^t, et sont notées par un vecteur colonne $P^t_{i\bullet}$. Pons et Latapy (2006) ont définit une distance r_{ij} pour comparer les sommets du graphe comme suit :

$$r_{ij} = \sqrt{\sum_{k=1}^{n} \frac{(P^t_{ik} - P^t_{jk})^2}{w(k)}} = \|D^{-\frac{1}{2}} P^t_{i\bullet} - D^{-\frac{1}{2}} P^t_{j\bullet}\| \tag{8}$$

Tel que :
D **:** la matrice diagonale des degrés ;
r **:** une distance Euclidienne $\in R^n$.
La distance r est directement liée aux propriétés spectrales de la matrice de transition P (deux sommets appartiennent à la même communauté ont des composantes similaires sur les vecteurs propres principaux).
Le problème de détection de communautés peut être réduit à détecter séparément des communautés dans chaque composante connexe. Cette approche permet de tirer profit de ces propriétés spectrales tout en gardant une complexité raisonnable $O(nmlog(n))$ grâce aux calculs de marches aléatoires. Lorsque la longueur des marches devient importante, la qualité des résultats diminue. Ceci est expliqué par le fait que les marches aléatoires atteignent rapidement leur état stationnaire limite.

Les méthodes proposées par Zhou Zhou (2003b,a) et Zhou et Lipowsky Zhou et Lipowsky (2004) sont basées sur le nombre moyen d'étapes pour qu'une particule brownienne (mouvement aléatoire d'une particule) atteigne un sommet donné en partant d'un autre sommet.
La distance entre les sommets mesurée par une particule brownienne est utilisée pour identifier la structure de communauté et identifier le noeud central de chaque communauté. Soit un réseau connecté de N noeuds et M liens, A est sa matrice d'adjacence tel que :

$$A_{ij} = \begin{cases} 0 & S'il\ n'existe\ pas\ de\ lien\ entre\ i\ et\ j \\ A_{ij} = A_{ji} > 0 & Sinion,\ cette\ valeur\ designe\ la\ force\ d'interaction \end{cases} \tag{9}$$

L'ensemble des plus proches voisins du noeud i est dénotés par E_i, une particule brownienne continue à se déplacer sur le réseau, et à chaque étape elle fait un pas à partir de sa position actuelle (i) vers la position du plus proche voisins (j). La matrice de transfert est donné par :

$$P_{ij} = \frac{A_{ij}}{\sum_{l=1}^{N} A_{il}} \tag{10}$$

La distance d_{ij} est le nombre moyen d'étapes pour qu'une particule brownienne se déplace du noeud i au noeud j est calculé comme suit :

$$d_{ij} = \sum_{l=1}^{N} \left(\frac{1}{I - B(j)} \right)_{il} \tag{11}$$

Tel que :
I **:** La matrice d'identité ;
$B(j)$ **:** est la matrice formée en remplaçant la $j^e me$ colonne de la matrice P par une colonne de zéros.

Considérant un sommet i comme le sommet origine du réseau, l'ensemble $\{d_{i1}, d_{i,2}, ..., d_{iN}\}$ mesure à quelle distance tous les autres sommets sont situés de l'origine. Par conséquent, la perspective du réseau entier est identifiée à partir du sommet i. Supposons que les sommets i et j sont des voisins, la différence dans leurs perspectives du réseau peut être quantitativement mesuré. Zhou Zhou (2003a) a définit l'indexe de dissimilitude suivant :

$$\Lambda(i,j) = \frac{\sqrt{\sum_{k \neq i,j}^{N} [d_{ik} - d_{jk}]^2}}{(N-2)} \tag{12}$$

Quand les noeuds i et j appartiennent à la même communauté, la distance moyenne d_{ik} de i à n'importe quel autre sommet $k(k \neq i, j)$ soit presque similaire à la distance moyenne d_{jk} (de j à k). La valeur de dissimilitude $\Lambda(i,j)$ est petite si i et j appartiennent à la même communauté et grande s'ils appartiennent aux communautés différentes.

Dans Zhou (2003b), l'attracteur global d'un sommet i est le sommet le plus proche à i, tandis que l'attracteur local de i est son voisin le plus proche. Deux types de communautés ont été définit, selon les attracteurs locaux ou globaux. Une communauté L basée sur un attracteur local est identifiée selon les considérations suivantes :

1. Si le noeud $i \in L$ et j est un attracteur local du noeud i, alors $j \in L$.
2. Si $i \in L$ et i est un attracteur local d'un noeud k, alors mettre k dans L.
3. Un sous ensemble de L ne produit pas une communauté.

Zho et al ont définit les communautés basées sur un attracteur global, tel que chaque noeud a une forte probabilité d'être dans la même communauté C de son attracteur global. Dans Zhou (2003a), les communautés sont identifiées en utilisant une procédure séparative dont les étapes sont les suivantes :

1. Initialement, le graphe entier est considéré comme une seule communauté. Un seuil maximal de dissimilitude θ_{upp} est attribué à cette communauté ;
2. Pour chaque communauté, un paramètre θ de seuil de résolution est introduit avec la valeur initiale θ_{upp} de cette communauté. Si $\Lambda(i,j) \leq \theta$, les sommets i et j sont marqués comme "amis".
3. Décrémenter la valeur de θ : tous les liens dans la communauté sont examinés pour voir si deux plus proches voisins sont des amis. Différent ensemble d'amis sont alors formées, chacun contient tous les amis des sommets dans l'ensemble. Un sommet qui n'a aucun ami rejoint l'ensemble des amis avec qui il a une forte interaction. Après cette opération, les sommets des communautés sont distribués en un certain nombre de communautés disjointes.
4. Un processus d'ajustement local est exécuté pour déplacer les noeuds qui n'ont pas été correctement classifié.
5. Si les sommets de la communauté n'ont pas été divisés, alors retourner à l'étape (3). Si les sommets sont divisé en deux ou plusieurs communautés, on assigne à la communauté père un seuil inférieur de dissimilitude θ_{low} équivalente à θ. A chaque nouvelle communauté est assignée une valeur θ_{upp} équivalente à la valeur courante de θ. Répéter l'algorithme à partir de l'étape (2) pour traiter les communautés identifiées.

6. Après que toutes les communautés soient traitées, le dendrogramme est dessiné pour démontrer le rapport entre les différentes communautés aussi bien que les seuils de dissimilitude supérieure et inférieure de chaque communauté.

Zhou et Lipowsky Zhou et Lipowsky (2004) ont aussi utilisé le mouvement brownien pour définir l'algorithme Netwalk algorithm (NW) qui emploi la mesure de proximité structurelle (indexe de proximité) de deux sommets en appliquant une méthode de détection de communauté de clustering hiérarchique. Les algorithmes proposés par Zhou Zhou (2003b,a) et Zhou et Lipowsky Zhou et Lipowsky (2004) peuvent identifier une structure de communauté significative. Cependant, ces algorithmes sont lents car le calcul des distances entre tous les pairs de sommets se fait en $O(n^3)$. Ce qui rend l'application de ces approches sur des grands graphes inadmissible.

6.2.2 Les systèmes à état de spins

Reichardt et al Reichardt et Bornholdt (2004) ont combiné l'idée de Fu et Anderson Fu et Anderson (1986) avec le modèle de clustering de Potts qui a été définit par Blatt et al Blatt *et al.* (1996), ceci a permet de convertir les communautés du réseau vers le domaine magnétique. Dans Reichardt et Bornholdt (2006), Reichardt et al ont proposé un framework pour détecter les communautés en déterminant l'état fondamental de "q-Potts model spin glass" Parisi *et al.* (1987).

Reichardt et al Reichardt et Bornholdt (2004) ont proposé un algorithme de découverte de communauté qui se base sur le modèle de Potts à Q états. Le modèle de Potts est l'un des modèles les plus utilisés en physique statistique afin de décrire le comportement des corps magnétiques Kasteleyn et Fortuin (1969). Il correspond à modéliser ces corps comme des spins à Q états situés aux noeuds d'un réseau et qui sont en interaction entre voisins de façon à s'aligner pour un corps ferromagnétique, ou bien à être en opposition pour un corps antiferromagnétique, selon le signe de la constante de couplage.

Fu et Anderson Fu et Anderson (1986) ont démontré par analogie qu'il existe une relation entre l'énergie des systèmes physiques (représenté par l'Hamiltonien) et la fonction de coût dans un problème d'optimisation combinatoire. Soit le problème de partitionnement de graphes en deux sous graphes, le nombre de liens qui existent entre les deux sous graphes égale à :

$$\sum_{i>j} \frac{a_{ij}}{4}(\mu_i - \mu_j)^2 \tag{13}$$

tel que a_{ij} est le nombre de liens entre les deux sommets i et j, $\mu_i = \pm 1$ est une variable qui indique la partition à laquelle le sommet i appartient. La différence entre le nombre de sommets des deux sous graphes est égale à : $\sum_i \mu_i$.

Ainsi, la fonction de coût s'écrit comme suit :

$$C = \sum_{i>j} (\lambda - \frac{a_{ij}}{2})\mu_i \mu_j \tag{14}$$

La fonction de coût a la même forme que l'Hamiltonien d'un spin qui est donnée par :

$$H = \sum_{i>j} (J_0 - J_{ij}) s_i s_j \tag{15}$$

Tel que un spin s_i à deux orientations "haut, bas" qui correspondent aux $\mu_i = 1$ et $\mu_i = -1$ respectivement. Pour les système de magnétisme aléatoire l'Hamiltonien est composé de deux termes : un composant ferromagnétique avec la constante de couplage J_{ij} et un composant antiferromagnétique avec la constante de couplage J_0. Les auteurs Reichardt et Bornholdt (2004) ont modifié l'Hamiltonien de Potts à q états en ajoutant une contrainte globale :

$$H = -J \sum_{(i,j)\in E} \delta_{\sigma_i,\sigma_j} + \gamma \sum_{s=1}^{q} \frac{n_s(n_s-1)}{2} \tag{16}$$

Tel que :
E **:** est l'ensemble de liens,
σ_i **:** dénote les spins individuels $(i = 1, ..., N)$ qui peuvent prendre les valeurs $\sigma_{1,...,q}$.
n_s **:** dénote le nombre de spins correspondant au spin s, avec $\sum_{s=1}^{q} n_s = N$.
J **:** est la force d'interaction ferromagnétique,
γ **:** est un paramètre positif
δ **:** est le symbole de Kronecker.
Chaque sommet est caractérisé par un spin prenant q valeurs possibles. La première somme est le terme ferromagnétique de Potts qui représente une distribution homogène des spins dans le réseau, et est minimisé par : $H_{ferr} = -JM$. Le deuxième terme additionne tous les pairs de spins qui sont égaux, ce qui représente la diversité de la configuration de spins ou bien les classes de spins existantes.
Pour définir la structure de communauté ça revient à trouver l'état fondamental du système. Les communautés correspondent aux classes de sommets ayant des valeurs de spin égales. 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. Ce travail emploie l'algorithme de "Monte Carlo single spin flip heat-bath algorithm" pour déterminer l'état fondamental du système (structure de communauté). L'optimisation de l'énergie du système (le deuxième terme de l'Hamiltonien) correspond à favoriser les liens intra-communauté et optimiser les liens inter-communauté.

Une comparaison a été effectué avec l'algorithme de Girvan et Newman (le graphe de $n = 128$ sommets, divisés en quatre communautés de 32 sommets chacune). Deux mesures ont été définit : Sensibilité et spécificité. Sensibilité : une paire de noeuds est positif (négatif) quand il est dans la même communauté (différente communauté), la spécificité désigne la fraction de tout les paires de noeuds positif (négatif) qui sont classifiés correctement par l'algorithme. Selon Fig. 7, les performances de l'algorithme sont aussi bien que la méthode de GN.

L'algorithme proposé par Reichardt et al Reichardt et Bornholdt (2004) est capable de détecter l'affiliation des noeuds qui appartiennent aux plusieurs communautés, il s'adapte bien aux structures de communautés qui se chevauchent et permet la quantification de la stabilité des communautés. Cependant, le recuit simulé n'est pas une méthode d'optimisation globale efficace et l'algorithme ne pas être appliqué sur des grands réseaux.

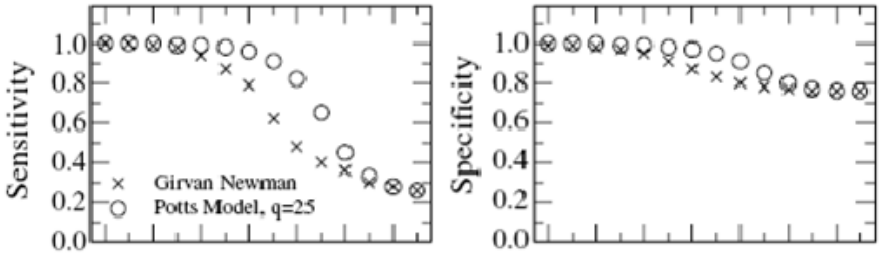

FIGURE 7 – Comparaison des résultats des sommets correctement identifiés par les algorithmes GN et l'algorithme de Reichardt et al Reichardt et Bornholdt (2004) .

6.2.3 *Les techniques de synchronisation*

Plusieurs études ont proposé un processus de synchronisation pour découvrir les communautés en prenant en considération la relation entre la structure topologique et les changements temporels Arenas *et al.* (2006), Boccaletti *et al.* (2007b). Il a été montré que les ensembles d'oscillateurs qui, sont fortement interconnectés et sont placés sur des nœuds de graphe, se synchronisent plus facilement pour former des clusters locaux Arenas *et al.* (2006). La complexité en temps de l'algorithme est $o(mn)$, ou n^2 sur des grands graphes Boccaletti *et al.* (2007b).

6.3 Méthodes basées sur l'analyse spectrale

Newman Newman (2006) a réécrit la mesure de modularité Q sous forme matricielle. La méthode proposée vise l'optimisation de la modularité tout en choisissant une division appropriée du réseau.
Soit un réseau de n noeuds. Prenons $s_i = 1$ si le noeud i appartient au groupe 1, $s_i = -1$ si le noeud appartient au groupe 2.

La modularité s'écrit comme suit :

$$Q = \frac{1}{4m}\sum_{ij}(A_{ij} - \frac{k_i k_j}{2m})s_i s_j = \frac{1}{4m}s^T B s \tag{17}$$

Tel que :
A_{ij} **:** Matrice d'adjacence ;
k_i **:** Degré du noeud i ;
m **:** le nombre total des liens dans le réseau ($m = 1/2\sum_i k_i$) ;
$\frac{k_i k_j}{2m}$ **:** le nombre prévu des liens entre les noeuds i et j s'il sont placés aléatoirement ;
$\frac{1}{4m}$: un facteur conventionnel : il est inclut pour la compatibilité avec la définition antérieure de modularité Newman et Girvan (2004).
Une nouvelle matrice symétrique de modularité B a été définit comme suit :

$$B_{ij} = A_{ij} - \frac{k_i k_j}{2m} \tag{18}$$

Cette matrice est une matrice Laplacienne qui est la base de toutes les méthodes les plus connues de partitionnement de graphes.

La modularité peut être écrite en fonction des vecteurs propres u_i de B :

$$Q = \sum_i a_i u_i^T B \sum_j a_j u_j = \sum_{i=1}^{n} (u_i^T s)^2 \beta_i \tag{19}$$

Le choix des éléments du vecteur s qui optimisent la modularité revient à résoudre un problème NP-hard similaire aux problèmes de partitionnement spectrale.

1. Calculer le vecteur propre principal de la matrice de modularité (u_1).
2. Diviser les nœuds en deux groupes selon le signe des éléments correspondants dans ce vecteur. Les nœuds qui ont une valeur positive sont mis dans un groupe et les autres nœuds dans le deuxième groupe.

L'auteur Newman (2006) a proposé une extension de sa première méthode pour diviser le réseau en plusieurs communautés. Ainsi, la matrice de modularité est décrite par l'équation suivante :

$$B_{ij}^{(g)} = A_{ij} - \frac{k_i k_j}{2m} - \delta_{ij}[k_i^{(g)} - k_i \frac{d_g}{2m}] \tag{20}$$

Tel que :
$B^{(g)}$ **:** Matrice de modularité d'un sous graphe g ;
d_g **:** La somme de degrés du sous graphe g.
L'algorithme se déroule comme suit :

1. Construire la matrice de modularité du graphe et trouver sa principale valeur propre et son principal vecteur propre.
2. Diviser le réseau en deux groupes selon les signes des éléments de vecteur propre.
3. Pour chaque groupe obtenu lors de l'étape 2 répéter le même algorithme de partitionnement.
4. Arrêter le processus de division si la modularité est nulle ou négative (le sous graphe est indivisible).
5. Quand tous les sous graphes sont indivisibles, le critère d' arrêter de l'algorithme est atteint.

Cet algorithme est testé sur le réseau des 105 livres de la politique américaine qui sont vendus sur le site Amazon.com. Sur la figure 8, les liens connectent des paires de livres qui sont fréquemment achetés par le même acheteur, les formes représentent l'alignement politique des livres : les cercles sont "liberal books", les carrés sont "conservative books" et les triangles sont "centrist books". Le résultat de l'algorithme est la détection de quatre communautés (marqué par des lignes pointillées).

On observe qu'une de ces communautés est composé entièrement des livres libérales et une autre se compose entièrement des livres conservateurs. Quand à la majorité des "centrist books" ont été identifiés dans les deux communautés restantes. En conséquence, ces livres forment des communautés d'achat qui sont alignées étroitement avec les points de vues politiques, ce qui démontre que l'algorithme proposé par Newman est capable d'extraire des résultats très significatifs. Les propriétés spectrales ont apporté plusieurs avantages à l'algorithme proposé.

Cette méthode n'a pas seulement la capacité de diviser le réseau efficacement, mais également de refuser de le diviser quand aucune bonne division n'existe. Cependant, le coût de calcul des vecteurs propres est élevé.

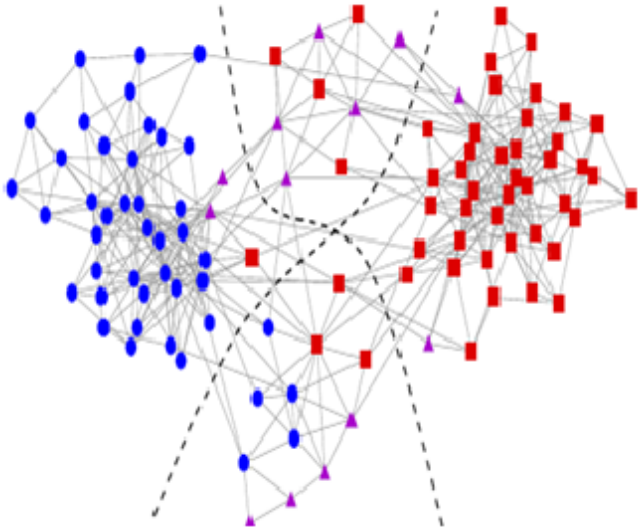

FIGURE 8 – "Krebs' network" réseau des livres d ela politique américaine.

6.4 Méthodes basées sur la structure topologique

Plusieurs approches de détection de communautés sont basées sur l'observation que la communauté peut être interprétée comme un sous-graphe complet (entièrement connecté). Dans cette section, nous décrivons les méthodes les plus référencées qui font partie des méthodes d'organisation topologique.

6.4.1 Cliques

Palla et al Palla *et al.* (2005); Derényi *et al.* (2005) ont définit une nouvelle méthode de percolation de clique (CPM) pour détecter les communautés jointes des réseaux. CPM utilise l'information locale qui est la densité de liens. Les auteurs se sont basés sur l'observation qu'une communauté peut être interprétée comme union de plus petits sous graphes complets qui partagent des noeuds entre eux. De tels sous graphes complets dans un réseau s'appellent les $k - cliques$, où k est le nombre de noeuds dans le sous graphe. Deux $k - cliques$ seraient adjacentes si elles partagent $(k - 1)$ noeuds, et une communauté est définie en tant que l'union de toutes les $k - cliques$ qui peuvent être atteintes par une série de $k - cliques$ adjacentes. Ces communautés peuvent être mieux visualisées à l'aide d'un modèle "template" $k - clique$ (un objet isomorphe pour un graphe complet de $k - sommets$). Cet objet peut être placé sur un $k - clique$ dans le graphe et roulé vers un $k - clique$ adjacent en changeant un de ses sommets et en gardant ses autres $(k - 1)$ sommets. Ainsi, les communautés ($k - clique$ percolation cluster) sont tous ces sous-graphes qui peuvent être entièrement exploré en roulant l'objet $k - clique$ sur eux, comme il est illustré sur Fig. 9. Initialement le template est placé sur $A - B - C - D$, puis il est roulé sur le sous graphe $A - C - D - E$. A chaque étape, seulement un des noeuds est déplacée et les deux $4 - cliques$ (avant et après le roulement) partage $k - 1 = 3$ noeuds . À l'étape finale le template atteint le sous-graphe $C - D - E - F$, et l'ensemble de noeuds visités pendant le processus $A - B - C - D - E - F$ sont considérés comme la communauté identifiée par le CPM.

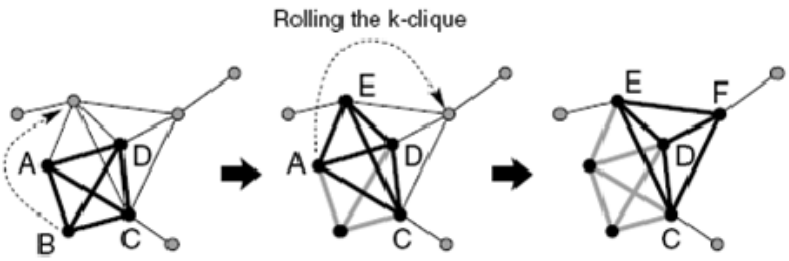

FIGURE 9 – Illustration de CPM Palla *et al.* (2005); Derényi *et al.* (2005) un $k-clique$ template roulé sur un petit graphe non orienté ($k = 4$).

Une extension de l'algorithme CPM a été proposée pour les réseaux pondérés Farkas *et al.* (2007) et les réseaux orientés Palla *et al.* (2007). L'algorithme CPM est efficace pour la détection des communautés qui se chevauchent (un noeud peut être un membre de plusieurs différentes communautés en même temps). Pour détecter les communautés à partir de $k-clique$, il faut tout d'abord calculer les cliques maximales. La complexité de temps pour trouver les cliques est exponentielle et directement proportionnelle à la taille du graphe, néanmoins Palla et al ont prouvé que l'algorithme peut s'exécuter en un temps admissible et cela sur des réseaux du monde réels qui peuvent aller jusqu'au 10^5 sommets. Cependant, la méthode CPM suppose que le graphe a un grand nombre de cliques, ainsi elle peut échouer à détecter des partitions significatives pour des graphes contenant juste quelques cliques (faible densité), comme dans les réseaux technologiques.

Dans Shen *et al.* (2009), Shen et al ont présenté un algorithme pour détecter à la fois la hiérarchie des communautés ainsi que leurs chevauchements en se basant sur l'ensemble de cliques maximales tout en employant un processus agglomératif. La similarité entre une paire de communauté C_1 et C_2 est calculé comme suit :

$$M = \frac{1}{2m} \sum_{v \in C_1, w \in C_2, v \neq w} \left[A_{vw} - \frac{k_v k_w}{2m} \right] \tag{21}$$

Tel que :
A_{vw} **:** La matrice d'adjacence ;
k_v **:** Le degré du sommet v ;
m **=** Le nombre total de liens du réseau ($m = \frac{1}{2} \sum_{vw} A_{vw}$).
Une clique maximale est une clique qui n'est pas un sous-ensemble d'aucune autre clique. Les cliques maximales, dont les sommets font partie à d'autres plus larges cliques maximales, s'appellent cliques maximales subordonnés. Les cliques maximales subordonnés peuvent dégradé le fonctionnement de l'algorithme et devraient être éliminées. La plupart des cliques maximales subordonnés ont de petites tailles. Ainsi, l'élimination de ces cliques se fait en respectant un seuil k et en négligeant toutes les cliques maximales dont la taille est inférieure à k. L'algorithme se déroule en deux étapes :

1. Trouver toutes les cliques maximales dans le réseau en utilisant l'algorithme de Bron-Kerbosch Bron et Kerbosch (1973). Ignorer les cliques maximales subordonnés et marquer les autres cliques en tant que communautés initiales. Chaque sommet subordonné est également considéré communauté initiale comportant un seul sommet. Calculer la similarité entre chaque paire de communautés.
2. Choisir la paire de communautés qui a une similarité maximale, les fusionner en une nouvelle et calculer la similarité entre la nouvelle communauté et les autres communautés.
3. Répéter l'étape 2 jusqu'à ce qu'il 'y a qu'une seule communauté qui correspond au graphe entier.

Shen et al ont définit une extension de modularité pour déterminer la qualité de partitionnement de l'algoritjme CPM :

$$EQ = \frac{1}{2m} \sum_{i} \sum_{v \in C_i, w \in Ci} \frac{1}{O_v O_w} \left[A_{vw} - \frac{k_v k_w}{2m} \right] \tag{22}$$

Tel que :

O_v **:** Le nombre de communautés auxquelles le sommet v appartient.

Dans la deuxième étape, l'algorithme détermine où couper le dendrogramme selon la valeur maximale de EQ. Les figures 10 11 comparent les résultats obtenues en appliquant fast algorithm de Nweman Newman (2004), $k - clique$ de Palla et al Palla *et al.* (2005) et EAGLE algorithm de Shen Shen *et al.* (2009) sur un réseau de collaboration scientifique.

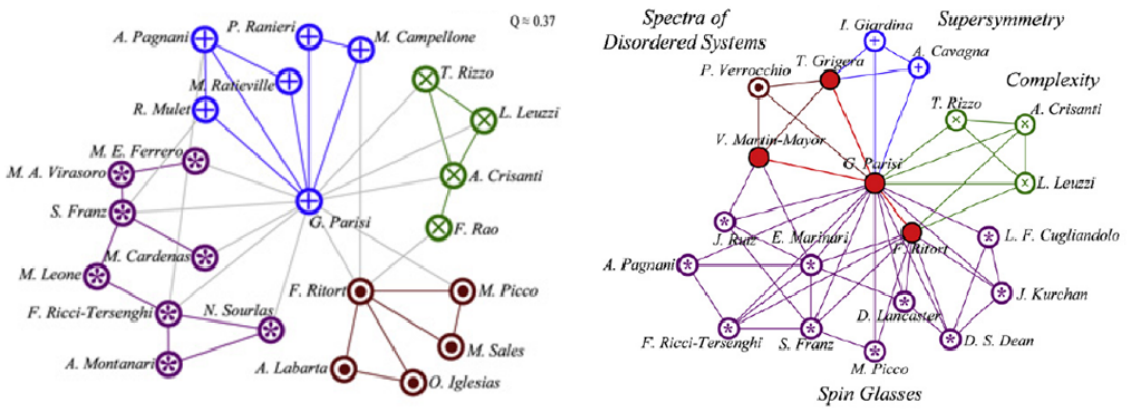

FIGURE 10 – Structure de communautés découvertes par fast algorithm (à droite) et l'algorithme k-clique. (Les noeuds et les liens qui appartiennent aux deux ou plusieurs communautés sont colorés en rouge)

L'algorithme AIGLE et l'algorithme de Fast résultent d'un nombre de communautés presque identique à chaque niveau hiérarchique sauf que la taille de ces communautés est un peu différente. L'algorithme AIGLE a détecté une communauté supplémentaire qui représente les liens et les noeuds appartiennent aux communautés jointes. De même, l'algorithme CPM détecte le chevauchement des communautés mais il ne découvert pas la hiérarchie complète des communautés. Cependant, le coût de calcul généré par l'algorithme AIGLE lors de la recherche de cliques maximales est très élevé.

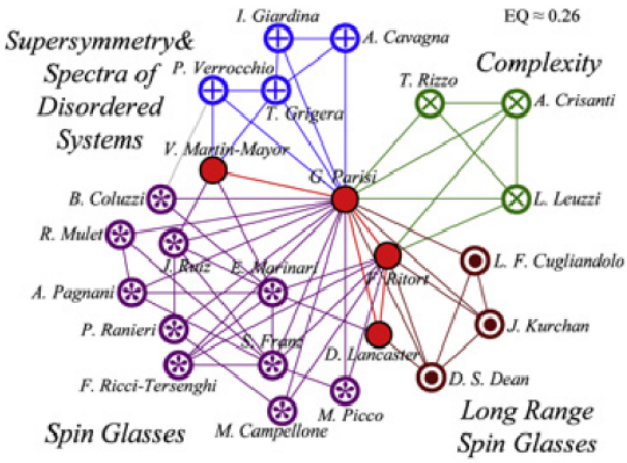

FIGURE 11 – Structure de communautés découvertes par EAGLE algorithm.

6.4.2 *Motifs*

La forte densité des arêtes au sein d'une communauté détermine la forte corrélation entre les nœuds ce qui est désigné par la présence de motifs. Arenas et al Arenas *et al.* (2008) ont montré comment les motifs peuvent être utilisés pour définir des classes générales de nœuds, y compris les communautés, en réécrivant l'expression mathématique de la modularité de Newman-Girvan. Ils ont défini la modularité du motif comme la fraction de motifs à l'intérieur des communautés moins la fraction dans un réseau aléatoire.

6.5 Méthode basée sur des propriétés locales

Bagrow et al Bagrow et Bollt (2005) ont proposé un algorithme de détection de communautés qui utilise l'information locale sans avoir une connaissance globale sur le réseau entier. Cette information locale représente le nombre de lien interne et externe d'un groupe de sommets se trouvant sur une distance géodésique depuis un sommet de départ. Afin de détecter les communautés localement et en se basant sur un simple critère qui emploie le nombre de liens interne et externe d'un groupe de sommets, Bagrow et al Bagrow et Bollt (2005) ont utilisé la notion de *shell* et ont définit deux mesures de liens : Shell qui est définit en tant qu'un ensemble de sommets sur une distance géodésique depuis un sommet de départ. Le premier shell inclut les plus proches voisins du sommet de départ et le deuxième inclut les prochains voisins des plus proches voisins de ce sommet et ainsi de suite (jusqu'au l shell). $K_i^e(j)$ est le degré émergent d'un sommet i qui représente le nombre de liens qui relient ce sommet i et les sommets d'un shell partant d'un sommet de départ j. K_j^l le degré émergent total d'un shell de profondeur l partant d'un sommet j. Il est clair que le nombre total des liens reliant les sommets d'un shell l est égale à la somme des degré émergent de tous les sommets ayant un lien vers le l shell :

$$K_j^l = \sum_{i \in S_j^l} k_i^e(j) \tag{23}$$

Avec S_j^l : est l'ensemble de tous les sommets à l pas du sommet j.

En outre, le changement de degré émergent totale d'un shell de profondeur l partant d'un sommet j, s'écrit comme suit :

$$\Delta K_j^l = \frac{K_j^l}{K_j^{l-1}} \tag{24}$$

L'algorithme étend le nombre de shell, à chaque itération, en ajoutant les sommets qui se trouvent à l pas du sommet j, tout en respectant un seuil de changement α tel que : $\Delta K_j^l < \alpha$. Pour un sommet de départ j faire :

1. Initialiser l shell, $l = 0$, depuis le sommet j, calculer K_j^0 (degré de sommet j) et ajouter j à la liste des membres de communauté.
2. Incrémenter le nombre de shell, $l = 1$, ajouter les sommets se trouvant sur le 1 shell à la liste des membres de communauté et calculer K_j^1.
3. Calculer ΔK_j^1 : Si $\Delta K_j^1 < \alpha$ une communauté a été détectée et le processus s'arrête, sinon, répéter l'algorithme à partir de l'étape 2 pour le prochain shell jusqu'à ce que la contrainte de seuil soit satisfaite ou bien le composant global connecté est ajouté à la liste des communautés.

Bagrow et al Bagrow et Bollt (2005) ont définit une matrice d'adhésion M qui regroupe les vecteurs v_i représentant les communautés de chaque sommet de départ, puis selon la distance entre les vecteurs, un processus est exécuté afin de permuter les lignes qui ont une plus courte distance entre eux, ce qui permet de regrouper les sous communautés appartenant au même communauté. Ce regroupement permet de produire le dendrogramme correspondant à la structure de communautés.

Une illustration sur le réseau de club de karaté (voir Fig 18) montre que l'algorithme atteint le résultat souhaité quand $\alpha = 1.2$, nous constatons que trois noeuds ne sont pas correctement détectés (3,14,20) comme le montre la matrice d'adhésion (Fig. 12.a). Ces noeuds se situent à la frontière des deux groupes existants et sont identiquement relié aux deux communautés, comme le montre le dendrogramme (Fig. 12.b).

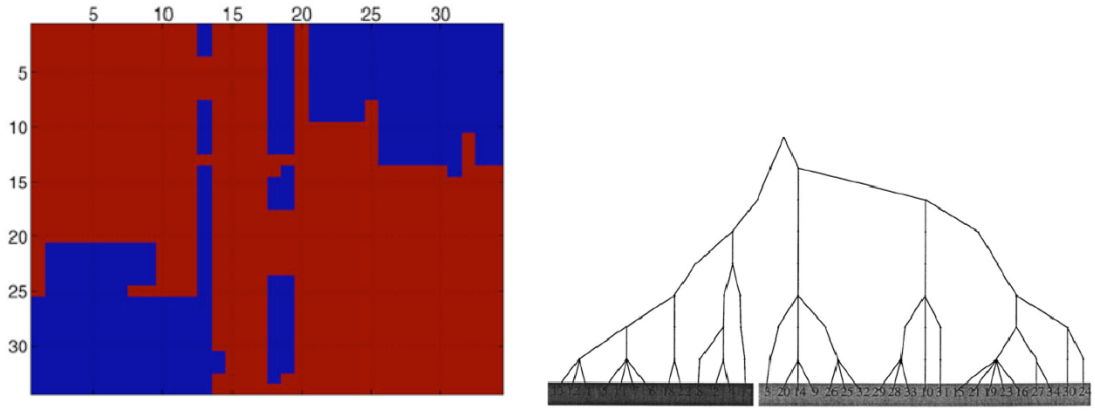

FIGURE 12 – (a) : Matrice d'adhésion du réseau de Zachary ($\alpha = 1.2$) ; (b) : Dendrogramme

En raison de sa nature locale, l'algorithme est très rapide et efficace dans certaines situations où il s'agit de l'identification d'une seule communauté mais sa version globale génère un coût de calcul élevé $O(N^3)$ (N $\leq$nombre de sommets du réseau). Cependant, le shell peut être s'étaler sur un ou plusieurs autres communautés et cela dépend étroitement de la localisation de sommet de départ surtout si ce dernier est proche à un ou des sommets qui n'appartiennent pas à la communauté du sommet de départ.

6.6 Méthodes basées sur la propriété de clustering

Eckmann et al Eckmann et Moses (2002b) ont proposé une méthode basée sur les propriétés de clustering. L'idée est d'utiliser le coefficient de regroupement des nœuds comme une quantité pour distinguer les sous graphe de neouds connectés. Plus il y a de triangles dans le sous-graphe, plus la distance moyenne est courte.

VII MÉTHODES SÉPARATIVES

Les méthodes séparatives Zhou (2003b); Van Dongen (2000); Duch et Arenas (2005); Newman et Leicht (2007); Newman (2006); Leicht et Newman (2008); Zhang *et al.* (2008); Radicchi *et al.* (2004); Shen *et al.* (2008); Girvan et Newman (2002); Newman et Girvan (2004); Fortunato *et al.* (2004); Sales-Pardo *et al.* (2007) scindent le graphe en plusieurs communautés en retirant progressivement les arêtes reliant deux communautés distinctes. La figure 13 illustre la classification des méthodes séparatives. Ces méthodes trouvent les pairs de nœuds qui sont reliés par des liens de faible similarité et les enlèvent au fur et à mesure. Ainsi, ce processus de suppression de liens peut être arrêté à n'importe quelle étape et le réseau est divisé en plusieurs composants représentants les communautés.

7.1 Méthodes de coefficient de clustering

Radicchi et al Radicchi *et al.* (2004) ont proposé un algorithme séparatif de détection de communauté en introduisant un nouveau concept appelé coefficient de clustering d'arête. Ils ont définit le coefficient de clustering d'arête par analogie avec le coefficient de clustering de noeud.

Les approches de Radicchi et al Radicchi *et al.* (2004) et d'Auber et al Auber *et al.* (2003) basées sur le clustering d'arêtes. Radicchi et al Radicchi *et al.* (2004) proposent un coefficient de clustering (d'ordre g) d'arêtes. Il est défini comme étant le nombre de cycles de longueur g passant par l'arête divisé par le nombre total de tels cycles possibles. Cet algorithme retire donc à chaque étape l'arête de plus faible clustering.
Dans la méthode de Radicchi et al Radicchi *et al.* (2004), chaque suppression d'arête ne demande qu'une mise à jour locale des coefficients de clustering, ce qui améliore la performance de l' algorithme. Cependant, cet algorithme se base sur la présence des triangles dans le réseau ; quand un réseau a peu de triangles, le coefficient de clustering d'arête est de petite valeur pour toutes les arêtes et l'algorithme est incapable de détecter les communautés.

Les réseaux bipartis sont un type important de réseau complexe. Un réseau biparti est composé de deux ensemble de sommets distincts. En fait, beaucoup de réseaux du monde réels sont naturellement bipartis, on cite souvent l'exemple du graphe reliant les acteurs aux films dans lesquels ils jouent, les graphes d'occurrence des mots dans les phrases, ou encore le graphe des auteurs de publications scientifiques. Watts et Strogatz ont introduit en 1998 la notion formelle de coefficient de clustering Watts et Strogatz (1998). Il s'agit de la moyenne du ratio du nombre de voisins de u qui sont reliés entre eux sur le nombre total de liens qui pourraient potentiellement exister entre ces voisins. Zhang et al Zhang *et al.* (2008) ont modifie la définition de coefficient de clustering. Ils l'ont adapté aux réseaux bipartis puis ont proposé un algorithme de détection de communauté pour les réseaux bipartis dont le principe est de retirer les liens qui ont la plus petite valeur de coefficient de lien.
Lind et al Lind *et al.* (2005) ont étudié le coefficient de clustering dans des réseaux bipartis où il n'y a pas de cycles de dimension trois , et par conséquent, la définition standard de coefficient de clustering donné dans Luce et Perry (1949) ne peut pas être utilisé. De ce fait, ils ont définit un coefficient donné par la fraction de cycles à quatre dimensions.

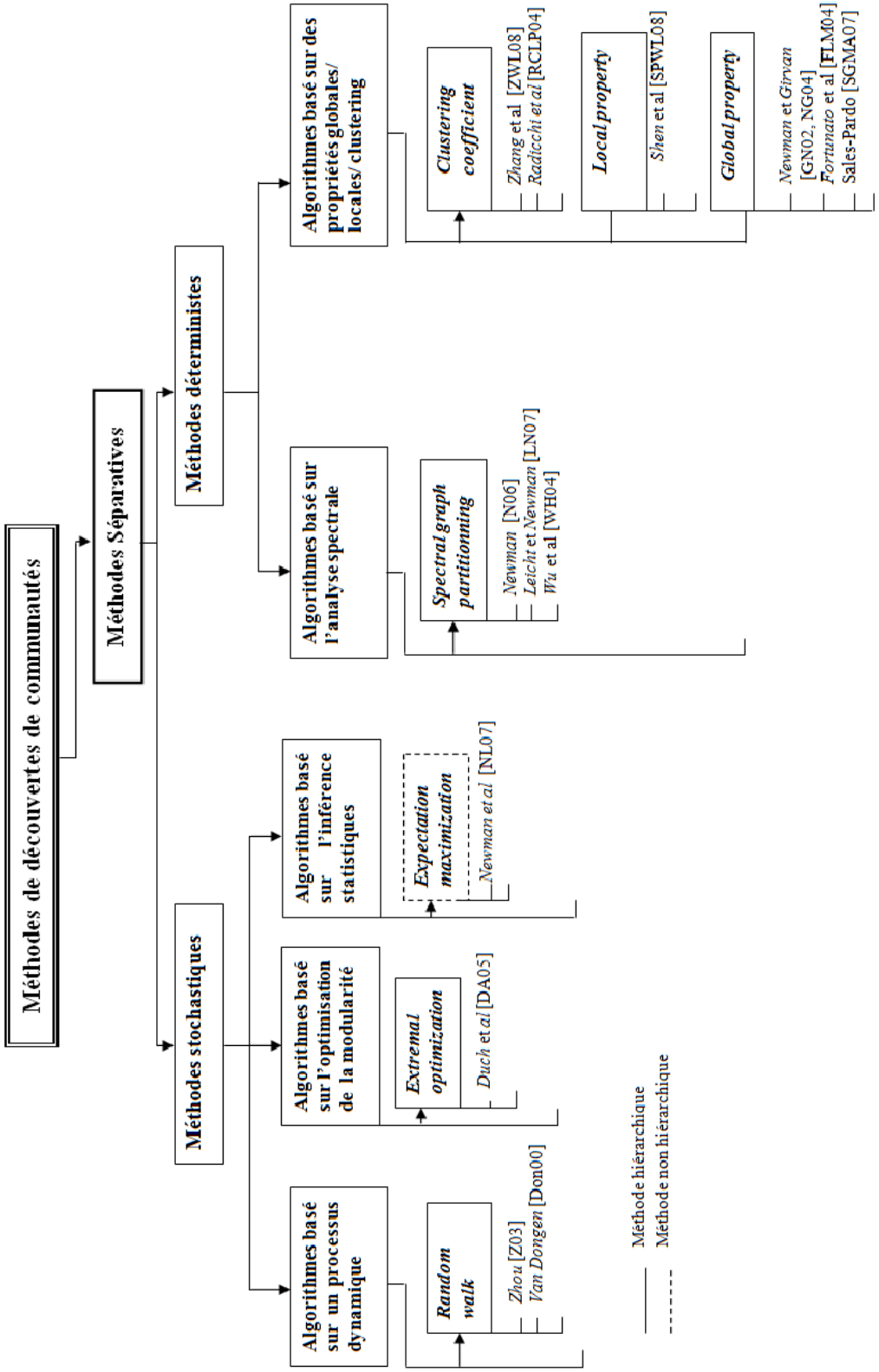

FIGURE 13 – Taxonomie des méthodes de découverte de communautés dans les réseaux complexes : 2. Méthode séparatives.

Le calcul des triangles possibles dans un réseau binaire prend en considération tous les liens éventuels entre les voisins les plus proches ; Donc $C_3(i)$ décrit la probabilité que les amis du noeud i soient des amis Watts et Strogatz (1998). Le coefficient de clustering $C_4(i)$, qui a été définit par Lind et al Lind *et al.* (2005), est la fraction entre le nombre de carrés existants et le nombre total de tous les carrés possibles (C_4 représente la probabilité que vos amis ont des amis communs hormis vous).

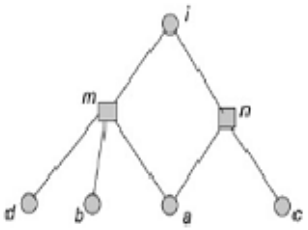

FIGURE 14 – Exemple pour illustrer le calcul des carrés.

Pour le noeud i (Fig. 14), le nombre de carrés (quadruplet de noeuds) possibles est donné par le nombre de voisins communs entre ses voisins m et n. Tandis que les carrés sous-jacents peuvent être calculées en ajoutant des liens éventuels, par exemple b est l'ami de m mais pas de n ; on l'appelle : un ami commun sous jacent du noeud m et n. Si nous considérons le lien potentiel entre b et n, nous pouvons obtenir un carré sous-jacent $imbn$.

Selon cette considération, l'équation de C_4 est définit comme suit :

$$C_{4,mn} = \frac{q_{imn}}{(k_m - \eta_{imn}) + (k_n - \eta_{imn}) + q_{imn}} \tag{25}$$

Tel que :
m,n **:** Des voisins du noeud i ;
q_m, n **:** Le nombre de carrés qui incluent les trois noeuds (les carrés existants et sous-jacents).
$\eta_{imn} = 1 + q_{imn}$.
La définition de Lind considère d'éventuel coïncidence des noeuds, quand à la définition de Zhang et al Zhang *et al.* (2008) considère d'éventuels coïncidence des liens en se basant sur les normes du coefficient de clustering des réseaux binaires. Radicchi et al Radicchi *et al.* (2004) ont proposé un algorithme séparatif de détection de communauté en introduisant un nouveau concept appelé coefficient de clustering d'arête. Zhang et al Zhang *et al.* (2008) ont aussi définit le coefficient de clustering de lien LC_4 et LC_3 pour les réseaux bipartis.

Ainsi, LC_4 s'écrit comme suit :

$$LC_{4,iX} = \frac{q_{iX}}{(k_i - 1)(k_X - 1) + k_i^{(2)} + k_X^{(2)}} \tag{26}$$

Tel que :
q_{iX} **:** Le nombre de carré auxquels le lien l_{iX} appartient ;
k_i **:** Le degré du noeud i ;
$k_i^{(2)}$ **:** Le nombre des voisins des voisins du noeud i sans les noeuds qui sont des premiers voisins du noeud X.

Dans les réseaux bipartis, les triples sont l'unité de base qui exprime la relation entre deux noeuds du même ensemble. Ainsi, les auteurs ont définit le coefficient de clustering de lien LC_3 basé sur les triples. LC_3 d'un lien l_{iX} représente la moyenne de la similitude de liens obtenues de tout les triples à lesquels ce lien appartient. Il s'écrit comme suit :

$$LC_{3,iX} = \frac{1}{k_i + k_X - 2}\left(\sum_{m=2}^{k_X} \frac{t_{mi}}{k_m + k_i - t_{mi}} + \sum_{N=2}^{k_i} \frac{t_{NX}}{k_N + k_X - t_{NX}}\right) \tag{27}$$

Tel que :
m **et** i **:** sont du même ensemble ;
i **et** X **:** n'appartiennent pas au même ensemble ;
t_{mi} **:** le nombre de triples qui contient les noeuds i et m (de même pour les noeuds N et X).
Lors de la détection des communautés dans les réseaux bipartis, le lien avec la petite valeur LC_4 (ou LC_3) est retiré et cela à chaque étape. Les figures ci-dessous (Fig. 15) montrent un réseau biparti qui contient 6 noeuds en haut et 6 noeuds en bas.

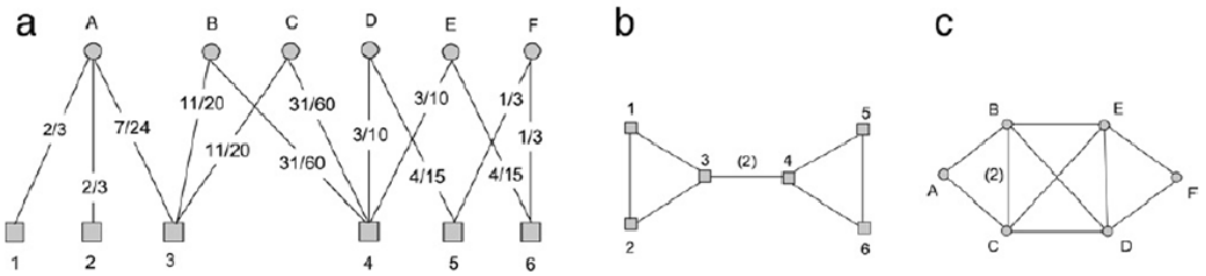

FIGURE 15 – Exemple d'un réseau biparti et sa projection.

L'exécution de l'algorithme basé sur le coefficient $LC3$ a résulté de :

a) Dans une première étape, les liens $D5$ et $E6$ sont supprimés et les communautés obtenues sont : $\{A, B, C, D, E, 1, 2, 3, 4\}$ et $\{F, 5, 6\}$,
b) Puis le lien $A3$ est retiré et les communautés obtenues sont : $\{A, 1, 2\}$, $\{B, C, D, E, 3, 4\}$, et $\{F, 5, 6\}$;
c) Dans la troisième étape, les liens $D4$ et $E4$ sont retirés. Il en résulte 4 communautés différentes : $\{A, 1, 2\}$, $\{B, C, 3, 4\}$, $\{D, E\}$, et $\{F, 5, 6\}$.

L'exécution de l'algorithme sur un réseau biparti d'Econophysists Li *et al.* (2007) qui se compose de 818 auteurs et de 777 papiers a permis de détecter 20 communautés. La modularité de cette partition est : $M_B(p)_1 = 0.351$ (en utilisant la fonction de modularité des réseaux bipartis qui a été définit dans Guimerà *et al.* (2007)). L'algorithme résulte de communautés pertinentes. Cependant, le nombre de communauté à détecter doit être déterminé d'avance et il n'y a pas un critère d'arrêt bien précis pour arrêter la division du réseau en communautés.

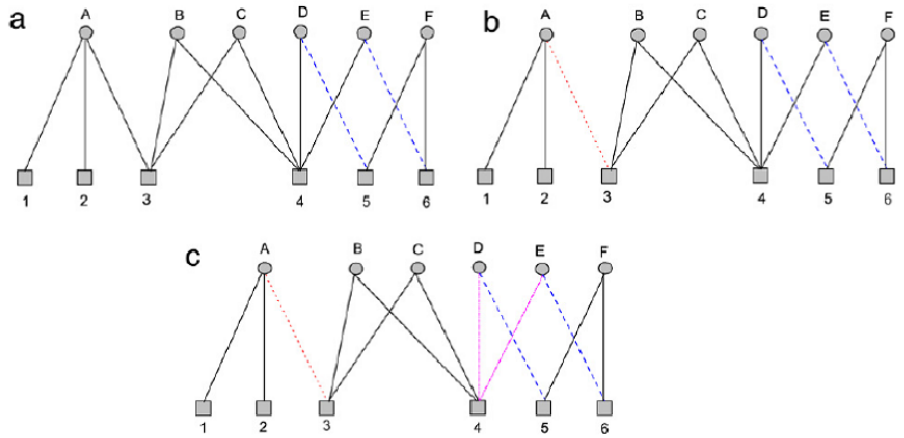

FIGURE 16 – L'identification des communautés par l'algorithme basé sur le coefficient de lien.

7.2 Méthodes basées sur des propriétés locales

Shen et al Shen *et al.* (2008) ont proposé une méthode de détection de communautés en supprimant plusieurs liens simultanément dans chaque opération de filtrage, et ont définit un coefficient récursif de communauté pour quantifier la qualité de division au lieu d'utiliser la modularité.

Soit un réseau complexe de m liens et de n noeuds, P_{ij} représente la probabilité pour qu'il y ait des liens entre chaque paire de noeuds i et j.

$$P_{ij} = \frac{k_i k_j}{2m} \tag{28}$$

Tel que k_i et k_j sont respectivement les degrés des noeuds i et j.

Selon l'ordre décroissant de la valeur P_{ij}, les éléments de la matrice d'adjacence B_{ij} sont attribués (1 pour des valeurs supérieur, sinon 0).

Si $B_{ij} \neq B_{ji}$ alors : $B_{ij} = B_{ji} = 0$.

Certains liens entre les communautés peuvent être supprimées par l'opération de filtrage décrite par l'équation suivante :

$$C_{ij} = A_{ij} - B_{ij} \tag{29}$$

Si $C_{ij} = -1$ Alors $C_{ij} = 0$

Shen et al Shen *et al.* (2008) ont proposé un coefficient récursif de communauté (CRC), dénoté M, afin de quantifier l'effet de division de réseau.

Soit un réseau, avec n noeuds et m liens, filtré par l'équation de C_{ij} et divisé en c sous réseaux. Soit n_k $(k = 1, ..., c)$ le nombre de noeuds dans le k^{ieme} sous réseau.

M s'écrit comme suit :

$$M = \frac{\frac{1}{2}\sum_{k=1}^{c}\sum_{ij}^{n_k} C_{ij}\delta(n_{w_i}, n_{w_j})}{\frac{1}{2}\sum_{ij}^{n} A_{ij}} \tag{30}$$

Tel que : $\delta(n_{w_i}, n_{w_j}) = 1$ si i et j sont dans le même sous graphe, sinon 0.

Répéter

1. Construire le modèle aléatoire du réseau.
2. Diviser le réseau par l'opération de filtrage donnée par l'équation de C_{ij}.
   Si le réseau est divisible aller à l'étape 3,
   sinon appliquer une nouvelle distribution du réseau avant le filtrage et aller à 1.
3. Calculer le coefficient CRC de chaque sous réseaux obtenu à l'étape 2,
   si le CRC est plus petit que celui de son réseau père alors : considérer le sous-réseau en tant que communauté locale et arrêter sa division.
   Sinon considérer chaque sous-réseau en tant qu'une nouvelle communauté et aller à 1.

Jusqu'à ce que toutes les communautés locales soient construites.

La méthode récursive de filtrage proposée par Shen et al Shen *et al.* (2008) offre un gain en complexité de calcul $O(m^2 + (c+1)m)$, pour un réseau de m liens et c communautés. Cependant, cette méthode devient imprécise quand la densité des liens itra-communautés se rapproche à la densité des liens inter-communautés.

7.3 Méthodes basées sur des propriétés globales

Les algorithmes proposés par Newman et Girvan Girvan et Newman (2002); Newman et Girvan (2004) se différencient des algorithmes séparatifs existants dans le fait qu'ils ne se basent pas sur l'enlèvement des liens de faible similarité entre les pairs de sommets, mais sur l'enlèvement des liens de forte similarité en introduisant la mesure appelé centralité "betweenness" qui se focalise sur les lien enter-communautés.

La mesure de betweenness, qui a été définit dans Newman et Girvan (2004), consiste à trouver les plus courts chemins entre toutes paires de sommets et calculer l'implication de chaque lien le long de ces chemins. L'approche de Newman et al est inspirée du travail de Freeman Freeman (1977). La conception intuitive d'un point central dans la communication, qui se base sur la propriété structurelle "betweenness", a été proposé par Freeman qui a définit ce point comme étant le point qui relie entre d'autres points tout au long de leurs plus courts chemins de communication Freeman (1977). Le calcul de plus court chemin entre n'importe quelle pair de sommets peut être effectué en utilisant l'algorithme recherche en largeur d'abord "breadth-first search" (BFS) en temps $O(mn^2)$ Ahuja *et al.* (1994); Cormen *et al.* (2001). Newman a proposé Newman (2001) un algorithme performant qui calcule le shortest path betweeness en $O(mn)$.

Une autre mesure qui se base sur le signal qui circule à travers le réseau a été définit. Si le signal passe de la source à la destination tout au long des géodésiques chemins et tous les sommets envoient des signaux à tous les autres noeuds par le même taux constant, alors betweenness est la mesure du taux des signaux qui traversent chaque lien. Supposons que le signale ne circule pas le long des plus courts chemins, mais il fait une promenade aléatoire dans le réseau jusqu'à ce qu'ils atteint sa destination. Cela permet de définir une mesure que Newman et al Newman et Girvan (2004) l'ont appelé "random walk betweeness" (le nombre de périodes durant lesquelles

le signal traverse un lien dans une seule direction). Un random walk betweeness d'un lien (v, w) est définit comme suit : $|V_v - V_w|$

Tel que V est donné par la formule suivante :

$$V = D_t^{-1}(I - M_t)^{-1}s = (D_t - A_t)^{-1}s \tag{31}$$

Tel que :

t : Sommet destination ;

s : Le vecteur de la source s ;

A_t : La matrice d'adjacence sans la t^{ieme} ligne et colonne
Enlever le sommet t du graphe car on s'intéresse aux nombre de pas nécessaire pour atteindre t ;

D_t : La matrice diagonale sans la t^{ieme} ligne et colonne ;

k_i : : Le degré du noeud i,
La probabilité de transition de j à i est : A_{ij}/k_j ,$M = A.D^{-1}$

$k_v^{-1}[(I - M_t)^{-1}]$: est le nombre moyen des périodes d'un pas de n'importe quelle longueur qui traverse le lien de v à w.

L'algorithme de détection de communautés proposé par Newman et al Newman et Girvan (2004) se déroule comme suit :

1. Calculer les scores de centralité d'intermédiarité "betweeness" pour tous les liens du réseau.
2. Trouver le lien de plus fort score et le retirer du réseau.
3. Recalculer le score betweeness entre tous les liens restants.
4. Répéter l'algorithme à partir de l'étape (2) jusqu'à ce qu'il n'y ait plus de liens à retirer.

Nous citons quelques exemples de l'exécution de l'algorithme de Newman et Girvan sur des réseaux du monde réel :

a) Prenons le réseau de club de karaté de Zachary Zachary (1977) (Fig. **??**). L'algorithme résulte de deux communautés qui correspondent parfaitement aux deux groupes du réseau réel. Le dendrogramme correspondant à cette division est donné sur la figure **??**. La modularité résultante de la division des deux version de l'algorithme (shortest path betweeness et random walk betweeness) est élevés et le réseau est divisé en deux communautés (environ 0.4). Cependant, le sommet 3 a été mal classé par la version shortest path betweeness. Quand à l'algorithme random walk betweeness produit une classification correcte de tous les noeuds.

b) Prenons le réseau des personnages du roman Les Misérables de Victor Hugo. L'apparence des personnages ensemble dans une ou plusieurs scènes a été étudiée. La modularité la plus élevée généré par la version de shortest path betweeness algorithm est $Q = 0.54$ et correspond aux 11 communautés. Les communautés reflètent clairement la structure du livre (Jean Valjean et Javert forment le centre des communautés avec leurs adhérents respectifs).

L'algorithme présente de meilleure qualité de partition, notamment dans les réseaux de taille moyenne. Cependant, les deux version de l'algorithme shortest path betweeness et random walk betweeness sont très coûteux en calculs et s'exécute en $O(n^3)$ à cause de nombre de calculs répétés à chaque suppression d'un lien ce qui est inadmissible dans des applications critiques.

FIGURE 17 – Les communautés des personnages du roman Les Misérables de Victor Hugo.

7.4 Méthode basées sur l'optimisation de la modularité

Duch et al Duch et Arenas (2005) ont proposé une procédure de recherche heuristique pour optimiser la valeur optimale de la modularité. Ils considèrent que la modularité globale Q est la somme de la modularité local sur chaque sommet. La variable global à optimiser est la modularité $Q = \sum_r (e_{rr} - a_r^2)$. La définition de la variable locale dans le problème d'optimisation extrémal doit être liée à la contribution de différents noeuds i dans la modularité, elle est normalisée dans l'intervalle $[-1, 1]$ et donnée par :

$$\lambda_i = \frac{q_i}{k_i} = \frac{k_r(i)}{k_i} - a_r(i) \tag{32}$$

Tel que :

λ_i **:** La division de la modularité locale de chaque noeud sur son degré ;

k_i **:** Le degré du noeud i.

$k_r(i)$ **:** Le nombre de liens entre le sommet i et des sommets qui appartiennent à la même communauté r.

La procédure de recherche heuristique proposée pour trouver la valeur optimale de la modularité se déroule comme suit :

1. Initialement, les noeuds des graphes sont divisés en deux partitions aléatoires contenant le même nombre de noeuds ;
2. À chaque itération, le système s'auto-organise en déplaçant le noeud de plus faible valeur de fitness à une autre partition. Ces déplacement modifient les partitions, donc la valeur de fitness doit être recalculée ;

3. Les liens entre les deux partitions sont supprimés et on répète l'étape 2 et 3 pour chaque nouveau composant ;
4. Le processus est répété jusqu'à ce que la modularité Q ne puisse pas être amélioré (l'état optimal est atteint).

L'exécution de l'algorithme d'optimisation extrémale sur le réseau de Zachary produit une valeur de modularité optimale, après trois itérations, et divise le réseau en 4 communautés (Fig. 18 et Fig. 19). La valeur de la modularité est 0.419 supérieur à la valeur 0.318 produite par l'algorithme de Newman Newman (2004). Aussi, cette valeur est supérieur à 0.406 donné par l'algorithme de Reichardt et al Reichardt et Bornholdt (2004) et supérieur aux résultats de l'algorithme de Donetti et al Donetti et Munoz (2004) (on a 0.412 sur le réseau de Zachary).

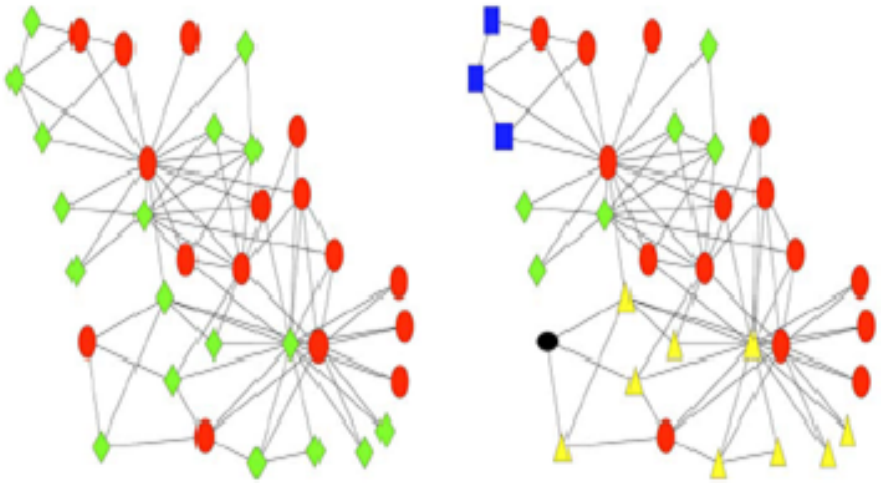

FIGURE 18 – Division initiale aléatoire du réseau de Zachary (le nombre de composants initial connecté dans les deux partitions est 5).

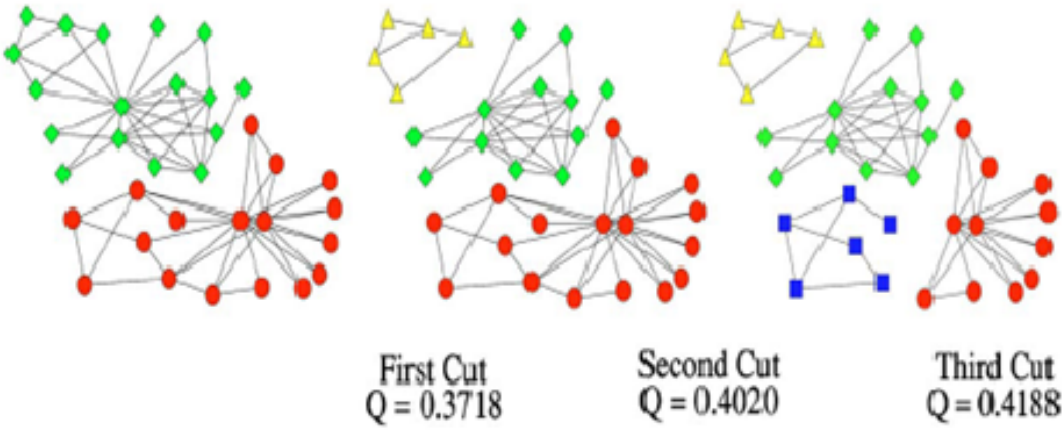

FIGURE 19 – Communautés obtenues après l'exécution de l'algorithme d'optimisation extrémale.

Bien que l'algorithme de Duch et al Duch et Arenas (2005) produit une modularité optimale qui dépasse plusieurs algorithmes existants tout en optimisant la complexité de calcul en $O(n^2 log(n))$, il dépend étroitement de l'étape d'initialisation du réseau en partition aléatoire. De ce fait, l'optimisation de la modularité ne conduit pas forcément à la partition souhaitée.

©JIMIS, Creative Commons
Volume : 5 - Année : 2019, DOI : 10.18713/JIMIS-ddmmyy-v-a

7.5 Méthodes basées sur l'analyse spectrale

Plusieurs auteurs ont étudié les réseaux orientés Newman et Leicht (2007); Guimerà *et al.* (2007); Arenas *et al.* (2007); Rosvall et Bergstrom (2008). Dans le travail proposé par Leicht et Newman Leicht et Newman (2008) la fonction de modularité a été généralisé afin d'incorporer l'information utile contenue dans l'orientation de lien. Vue que plusieurs réseaux complexes sont orientés, tels que le World Wide Web, food webs, beaucoup de réseaux biologiques et les réseaux sociaux, Leicht et Newman Leicht et Newman (2008) ont proposé une extension de la méthode d'optimisation spectrale de la modularité pour les réseaux complexes orientés.

Dans Newman (2006), Newman a écrit la fonction de modularité sous sa forme matricielle comme suit :

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta_{c_i, c_j} \tag{33}$$

Leicht et Newman Leicht et Newman (2008) ont réécrit la fonction de modularité des réseaux orientés. Considérons deux sommets i et j. Supposons que le sommet i a un degré sortant élevé et un faible degré entrant, et inversement pour le noeud j. Donc, la probabilité qu'un lien partant d'un sommet i est orienté vers j est : $\frac{k_i^{in} k_j^{out}}{m}$. Ainsi, la modularité est définit comme suit :

$$Q = \frac{1}{m} \sum_{ij} \left[A_{ij} - \frac{k_i^{in} k_j^{out}}{m} \right] \delta_{c_i, c_j} \tag{34}$$

La matrice de modularité est donnée par :

$$B_{ij} = A_{ij} - \frac{k_i^{in} k_j^{out}}{m} \tag{35}$$

Pour obtenir la matrice symétrique, Q est donné par :

$$Q = \frac{1}{4m} s^T (B + B^T) s = \beta_i (v_i^T s)^2 \tag{36}$$

Tel que β_i est la valeur propre de $(B + B^T)$ qui correspond au vecteur propre v_i.

Pour diviser le réseau en deux communautés, on calcule le vecteur propre correspondant à la plus grande valeur propre positive de la matrice symétrique $(B + B^T)$ et on assigne alors les communautés en se basant sur les signes des éléments du vecteur propre.

Pour diviser le réseau en plusieurs communautés, une généralisation de la matrice de modularité a été donnée par la formulation suivante :

$$B_{ij}^{(g)} = B_{ij} - \delta_{ij} \sum_{k \in g} B_{ik} \tag{37}$$

Tel que $B^{(g)}$ est la matrice de modularité du sous graphe g. L'algorithme se déroule comme suit :

1. Construire la matrice de modularité du graphe $(B + B^T)$ et trouver la plus grande valeur propre positive et son principal vecteur propre.
2. Diviser le réseau en deux groupes selon les signes des éléments de ce vecteur.
3. Un processus d'ajustement local est exécuté pour déplacer les noeuds qui n'ont pas été correctement classifié.
4. Pour chaque communauté obtenue lors de l'étape 2 répéter le même algorithme de division en utilisant la matrice de modularité généralisée.
5. Si l'algorithme ne trouve aucune division qui peut maximiser la modularité d'une communauté donnée, donc la communauté ne peut pas être divisée en des sous communautés. Quand toutes les communautés atteindraient cet état l'algorithme s'arrête.

Leicht et Newman Leicht et Newman (2008) ont utilisé plusieurs exemples des réseaux pour démontrer l'efficacité de leur algorithme. Un exemple du réseau qui représente la relation entre un ensemble de termes techniques, telles que "vertex", et "edge" et "community", contenu dans un glossaire dérivé des papiers publiés récemment par Newman Newman (2003b) et Boccaletti et al Boccaletti *et al.* (2006).

Les sommets dans ce réseau représentent les termes techniques et il y a un lien orienté d'un sommet vers l'autre si le premier terme a été employé dans la définition du deuxième terme. Fig. 20 illustre le résultat de l'exécution de l'algorithme de modularité orienté, il en résulte 6 communautés. Chaque communauté regroupe les termes commun à un concept donné (on trouve par exemple une communauté qui représente les termes de réseau orienté). L'algorithme a pu trouver une structure de communauté significative permettre de bien comprendre le contexte étudié dans les papiers Newman (2003b) et Boccaletti *et al.* (2006).

L'algorithme de modularité dans sa version non orientée a été également appliqué sur ce même réseau, ce qui résulte de quatre groupes. Deux de ces derniers sont étroitement semblables à ceux trouvés par l'algorithme orienté. Cependant, les autres groupes contiennent un mélange de termes qui ne correspondent pas strictement aux mêmes concepts de réseau, avec des mots tel que "vertex," "diameter," "cycle," et "motif" ont été regroupé ensemble. Les méthodes de découverte de communautés des réseaux non orientés sont le plus souvent incapables de détecter une partie très significative de la structure de communautés puisqu'elles ignorent l'information contenue dans l'orientation de lien. La méthode d'optimisation spectrale de la modularité dans sa version destinée aux réseaux complexes orientés extrait l'information d'orientation de liens pour identifier la structure de communautés, ce qui donne une structure de communautés significative. Aussi, son coût de calcul qui est $O(n^2 log(n))$ rend son utilisation très bénéfique.

Les méthodes spectrales consistent à plonger le graphe dans un espace euclidien de sorte que les sommets fortement reliés soient représentés dans une même partie de l'espace et les sommets sans ou avec peu de connexions soient représentés à distance. Donetti et al Donetti et Munoz (2004) ont proposé une approche basée sur les propriétés spectrales de la matrice Laplacienne

du graphe. Les coordonnées i et j des vecteurs propres correspondant aux plus petites valeurs propres non nulles sont corrélées lorsque les sommets i et j sont dans la même communauté. Une distance (distance euclidienne ou distance angulaire) entre sommets est alors calculée à partir de ces vecteurs propres, cette distance étant ensuite utilisée dans un algorithme de clustering hiérarchique. Le nombre de vecteurs propres à considérer est à priori inconnu. Plusieurs calculs sont successivement effectués en prenant en compte différents nombres de vecteurs propres, et le meilleur résultat est retenu. Les performances de l'algorithme sont limitées par les calculs des vecteurs propres qui se fait en $O(n^3)$ pour une matrice creuse. Une amélioration de cette approche a été proposée en utilisant une version normalisée de la matrice Laplacienne Donetti et Muñoz (2005).

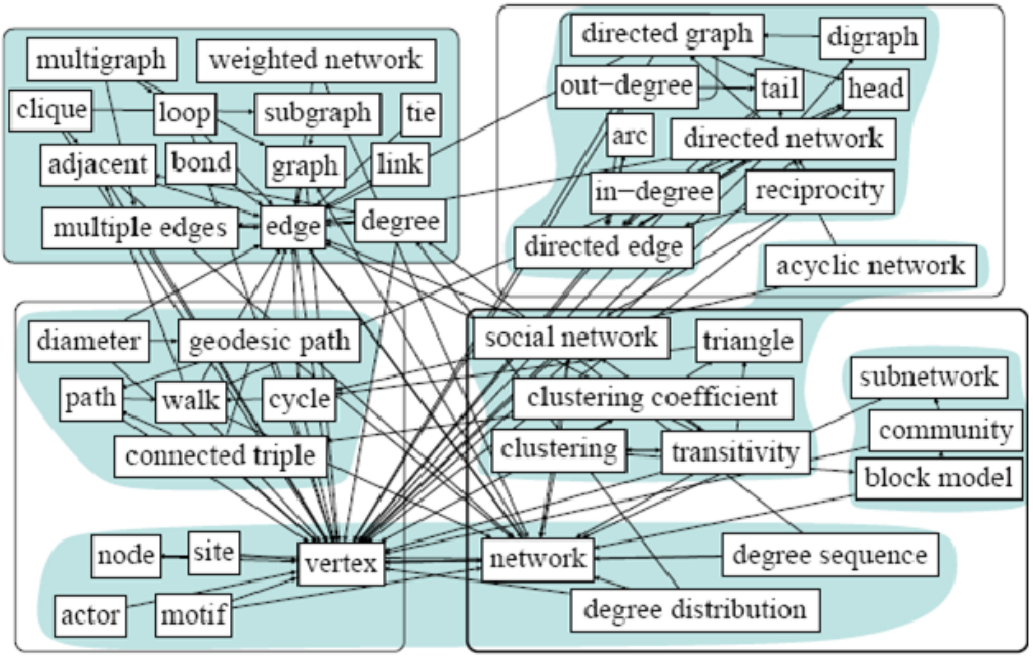

FIGURE 20 – Réseau des termes techniques illustre la découverte de communautés en appliquant l'algorithme de modularité orienté (groupes en bleu) et l'algorithme d'optimisation de modularité (groupes encadrés).

Dans Jiang *et al.* (2009), les auteurs ont reformulé la mesure de modularité "Q" en utilisant le clustering spectral afin de maximiser la modularité et en conséquence détecter correctement la structure de communauté du réseau. Les méthodes spectrales ont montré expérimentalement de meilleurs résultats mais elles sont coûteuses en terme de complexité car la détermination des valeurs et vecteurs propres d'une matrice creuse nécessite un temps de calcul en $O(n^3)$.

VIII ÉTUDE COMPARATIVE DES ALGORITHMES DE DÉCOUVERTE DE COMMUNAUTÉS

La disponibilité d'une partition de référence a permis de proposer différentes mesures pour évaluer la qualité d'une partition identifiée par un algorithme de découverte de communautés. Le travail Gustafsson *et al.* (2006) présente une revue des différents critères pour comparer les partitions détectées.

Nous citons un exemple des mesures qui sont basées sur l'information mutuelle. Une version

normalisée de l'information mutuelle (NMI) est introduite dans Danon *et al.* (2005) est donnée par :

$$N\ M\ I(A,B) = \frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B} N_{ij} log(\frac{N_{ij}N}{N_{i\bullet}N_{\bullet j}})}{\sum_{i=1}^{C_A} N_{i\bullet} log(\frac{N_{i\bullet}}{N}) + \sum_{j=1}^{C_B} N_{\bullet j} log(\frac{N_{\bullet j}}{N})} \tag{38}$$

Tel que :

N **:** La matrice de confusion. Les éléments N_{ij} représentent le nombre de sommets dans la communauté de référence i qui sont également dans la communauté détectée j.

C_A **:** Le nombre de communautés dans la partition de référence A

C_B **:** Le nombre de communautés dans la partition de référence B

Si les communautés identifiées sont identiques aux communautés de référence, alors $NMI(A,B)$ prend sa valeur maximale 1.

Une autre mesure utile de similarité entre les partitions est l'indice de Jaccard qui s'écrit comme suit :

$$I_j(A,B) = \frac{n_{11}}{n_{11} + n_{01} + n_{10}} \tag{39}$$

Tel que :

n_{11} **:** Le nombre de paires de sommets qui sont mis dans la même communauté dans les deux partitions A et B ;

n_{10} **:** Le nombre de paires de sommets qui sont mis dans la même communauté dans la partition A et dans différentes communautés dans B.

n_{01} **:** Le nombre de paires de sommets qui sont mis dans la même communauté dans la partition B et dans différentes communautés dans A.

Bien que la qualité de partition est une considération essentielle pour choisir une méthode de découverte de communautés, la rapidité d'exécution est aussi un facteur important particulièrement pour les grands réseaux. Le tableau 1 récapitule la complexité en temps des déférentes méthodes. Par exemple, la complexité de l'algorithme de Newman Newman et Girvan (2004) est $O(n^3)$ ce qui génère un coût de calcul très élevé sur de grands réseaux (des graphes de plus de 1000 sommets). Le temps d'exécution de l'algorithme "eigenvector-based algorithm" est relative à la taille du système, et il est de $O(n^2 log(n))$ pour un graphe dense, cela est considérablement meilleur que le temps d'exécution de "betweenness-based algorithm". La complexité de l'algorithme d'optimisation extrémale est de $O(n^2 log^2(n))$ ce qui offre un gain de performance. L'algorithme "fast" apporte plus de performance ($O(nlog^2(n))$) et convient aux grands réseaux complexes.

TABLE 1 – Récapitulatif de complexité en temps des différentes méthodes.

Algorithme	**Référence**	**Complexité en temps**
Random-walk algorithm	Newman et al Newman et Girvan (2004)	$O(n^3)$
Betweenness-based algorithm	Girvan et al Girvan et Newman (2002) Newman et al Newman et Girvan (2004)	$O(m^2n)$
Extremal optimization algorithm	Duch et al Duch et Arenas (2005)	$O(n^2log^2(n))$
Fast algorithm	Newman Newman (2004)	$O(nlog^2(n))$
Simulated annealing based algorithm	Guimerà et al Guimera *et al.* (2004a)	Inconnue
Q-state Potts model based algorithm	Reichardt et al Reichardt et Bornholdt (2004, 2006)	Dépend des paramètres
Local algorithm	Bagrow et al Bagrow et Bollt (2005)	$O(n^3)$
RCLP algorithm	Radicchi et al Radicchi *et al.* (2004)	$O(n^2)$
Eigenvector-based algorithm	Newman Newman (2006)	$O(n^2log(n))$
Greedy algorithm	Clauset et al Clauset *et al.* (2004)	$O(d.mlog(n))$
Directed modularity maximization algorithm	Leicht et Newman Leicht et Newman (2008)	$O(n^2log())$
Divisive algorithm of bipartite networks	Zhang et al Zhang *et al.* (2008)	Inconnue
Recursive filtration algorithm	Shen et al Shen *et al.* (2008)	$O(m^2 + (c+1)m)$
Biclique algorithm	Lehmann et al Lehmann *et al.* (2008)	$O(n^2)$
k-clique "CPM"	Palla et al Palla *et al.* (2005)	Inconnue
Markov cluster algorithm	Van Dongen Van Dongen (2000)	$O(n.k^2)$
EAGLE algorithm	Shen et al Shen *et al.* (2009)	$O(n^2.s)$

n : Le nombre des noeuds du réseau ;
m : Le nombre des liens du réseau ;
c : Le nombre de communauté dans le réseau ;
$k \leq n$: Le nombre maximal des éléments non nuls par colonne Van Dongen (2000) ;
d : La profondeur du dendrogramme ;
s : Le nombre de cliques maximale.

La comparaison des différentes techniques est difficile car il est possible que les méthodes les plus performant en termes de temps d'exécution ne peuvent pas identifient des partitions en communautés les plus pertinentes. La recherche d'un compromis entre la qualité des résultats et le coût de calcul reste toujours un défi à relever.

IX CONCLUSION

L'étude des réseaux complexes est une activité en plein essor. Un problème central dans l'étude des réseaux complexes est celui de la détection des communautés : des sous-graphes denses faiblement connectés entre eux. La découverte de la structure communautaire d'un graphe permet d'enrichir nos connaissance sur la structure interne des schéma des interactions mais aussi nous renseigner sur les possibilités d'évolution du graphe. Dans ce chapitre, nous avons passer en revue les principales méthodes de découverte de communautés dans les réseaux complexes. Nous avons discuté leurs apports et leurs inconvénients et proposé une classification des différentes approches.

Références

Ahuja R. K., Magnanti T. L., Orlin J. B. (1994). Network flows : Theory, algorithms, and applications. *Journal of the Operational Research Society 45*(11), 1340–1340.

Albert R., Barabási A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics 74*(1), 47.

Albert R., Jeong H., Albert-László B. (1999). Internet : Diameter of the world-wide web. *Nature 401*(6749), 130–131.

Albert R., Jeong H., Albert-László B. (2000). Attack and error tolerance of complex networks. *Nature 406*(6794), 378–382.

Albert-Laszlo B. (2002). *Linked : the new science of networks*. Perseus.

Amaral L. A. N., Scala A., Barthelemy M., Stanley H. E. (2000). Classes of small-world networks. *Proceedings of the national academy of sciences 97*(21), 11149–11152.

Arenas A., Diaz-Guilera A., Pérez-Vicente C. J. (2006). Synchronization reveals topological scales in complex networks. *Physical review letters 96*(11), 114102.

Arenas A., Duch J., Fernández A., Gómez S. (2007). Size reduction of complex networks preserving modularity. *New Journal of Physics 9*(6), 176.

Arenas A., Díaz-Guilera A., Pérez-Vicente C. J. (2006). Synchronization reveals topological scales in complex networks. *Physical review letters 96*(11), 114102.

Arenas A., Fernandez A., Fortunato S., Gomez S. (2008, June). Motif-based communities in complex networks. *Journal of Physics A : Mathematical and Theoretical 41*(22), 224001.

Auber D., Chiricota Y., Jourdan F., Melançon G., others (2003). Multiscale Visualization of Small World Networks. In *InfoVis03*, Volume 3, pp. 75–81. IEEE Computer Society.

Bagrow J. P., Bollt E. M. (2005). Local method for detecting communities. *Physical Review E 72*(4), 046108.

Barabási A.-L., Albert R. (1999). Emergence of scaling in random networks. *Science 286*(5439), 509–512.

Bilke S., Peterson C. (2001). Topological properties of citation and metabolic networks. *Physical Review E 64*(3), 036106.

Blatt M., Wiseman S., Domany E. (1996). Super paramagnetic clustering of data. *Physical review letters 76*(18), 3251.

Blondel V. D., Senellart P. P. (2002). Automatic extraction of synonyms in a dictionary. In *the SIAM Workshop on Text Mining*, Volume 1. Vertex.

Boccaletti S., Ivanchenko M., Latora V., Pluchino A., Rapisarda A. (2007a). Detecting complex network modularity by dynamical clustering. *Physical Review E 75*(4), 045102.

Boccaletti S., Ivanchenko M., Latora V., Pluchino A., Rapisarda A. (2007b). Detecting complex network modularity by dynamical clustering. *Physical Review E 75*(4), 045102.

Boccaletti S., Latora V., Moreno Y., Chavez M., Hwang D.-U. (2006). Complex networks : Structure and dynamics. *Physics reports 424*(4), 175–308.

Bollobás B. (1998). Random graphs. In *Modern Graph Theory*, pp. 215–252. Springer.

Bron C., Kerbosch J. (1973). Finding all cliques of an undirected graph (algorithm 457). *Commun. ACM 16*(9), 575–576.

Buchanan M. (2003). *Nexus : small worlds and the groundbreaking theory of networks*. WW Norton & Company.

Burt R. S. (1976). Positions in networks. *Social forces 55*(1), 93–122.

Clauset A., Newman M. E., Moore C. (2004). Finding community structure in very large networks. *Physical review E 70*(6), 066111.

Cormen T. H., Leiserson C. E., Rivest R. L., Stein C. (2001). *Introduction to algorithms*, Volume 6. MIT press Cambridge.

Danon L., Diaz-Guilera A., Duch J., Arenas A. (2005). Comparing community structure identification. *Journal of Statistical Mechanics : Theory and Experiment 2005*(09), 09008.

Derényi I., Palla G., Vicsek T. (2005). Clique percolation in random networks. *Physical review letters 94*(16), 160–202.

Dodds P. S., Rothman D. H. (2000). Geometry of river networks. *Physical Review E 63*(1), 016115–016117.

Donetti L., Munoz M. A. (2004). Detecting network communities : a new systematic and efficient algorithm. *Journal of Statistical Mechanics : Theory and Experiment 2004*(10), 10012.

Donetti L., Muñoz M. A. (2005). Improved spectral algorithm for the detection of network communities. In *Modeling Cooperative Behavior in the Social Sciences*, Volume 779, pp. 104–107.

Dorogovtsev S. N., Mendes J. F. F. (2002). Evolution of Networks : from Biological Nets to the Internet and WWW. 2003. *Oxford University Press*.

Drif A., Boukerram A., Slimani Y. (2014). Community Discovery Topology Construction for Ad Hoc Networks. In *International Wireless Internet Conference*, Volume 146, Lisbon (Portugal), pp. 197–208. Springer. doi:10.1007/978-3-319-18802-7_28.

Duch J., Arenas A. (2005). Community detection in complex networks using extremal optimization. *Physical review E 72*(2), 027104.

Ebel H., Mielsch L.-I., Bornholdt S. (2002). Scale-free topology of e-mail networks. *Physical review E 66*(3), 035103.

Eckmann J.-P., Moses E. (2002a). Curvature of co-links uncovers hidden thematic layers in the world wide web. *Proceedings of the national academy of sciences 99*(9), 5825–5829.

Eckmann J.-P., Moses E. (2002b). Curvature of co-links uncovers hidden thematic layers in the world wide web. *Proceedings of the national academy of sciences 99*(9), 5825–5829.

Egghe L., Rousseau R. (1990). Introduction to informetrics : Quantitative methods in library, documentation and information science. *Elsevier*.

Erdös P., Rényi A. (1959). On random graphs. *Publicationes Mathematicae (Debrecen) 6*, 290–297.

Eriksen K. A., Simonsen I., Maslov S., Sneppen K. (2003). Modularity and extreme edges of the internet. *Physical review letters 90*(14), 148701.

Faloutsos M., Faloutsos P., Faloutsos C. (1999). On power-law relationships of the internet topology. In *ACM SIGCOMM computer communication review*, Volume 29, pp. 251–262. ACM.

Farkas I., Ábel D., Palla G., Vicsek T. (2007). Weighted network modules. *New Journal of Physics 9*(6), 180.

Flake G. W., Lawrence S., Giles C. L., Coetzee F. M. (2002). Self-organization and identification of web communities. *IEEE Computer 35*(3), 66–70.

Fortunato S., Barthelemy M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences 104*(1), 36–41.

Fortunato S., Castellano C. (2008). *Community structure in graphs. Encyclopedia of Complexity and System Science*. Springer.

Fortunato S., Latora V., Marchiori M. (2004). Method to find community structures based on information centrality. *Physical review E 70*(5), 056104.

Freeman L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 35–41.

Fu Y., Anderson P. W. (1986). Application of statistical mechanics to NP-complete problems in combinatorial optimisation. *Journal of Physics A : Mathematical and General 19*(9), 1605.

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

Granovetter M. S. (1973). The strength of weak ties. *American journal of sociology*, 1360–1380.

Guillaume J.-L., Latapy M., Le-Blond S. (2004). Statistical analysis of a p2p query graph based on degrees and their time-evolution. In *International Workshop on Distributed Computing*, pp. 126–137. Springer.

Guimera R., Amaral L. A. N. (2005). Functional cartography of complex metabolic networks. *nature 433*(7028), 895.

Guimera R., Sales-Pardo M., Amaral L. A. N. (2004a). Modularity from fluctuations in random graphs and com-

plex networks. *Physical Review E 70*(2), 025101.

Guimera R., Sales-Pardo M., Amaral L. A. N. (2004b). Modularity from fluctuations in random graphs and complex networks. *Physical Review E 70*(2), 025101.

Guimerà R., Sales-Pardo M., Amaral L. A. N. (2007). Module identification in bipartite and directed networks. *Physical Review E 76*(3), 036102.

Gustafsson M., Hörnquist M., Lombardi A. (2006). Comparison and validation of community structures in complex networks. *Physica A : Statistical Mechanics and its Applications 367*, 559–576.

Holme P., Huss M., Jeong H. (2003). Subnetwork hierarchies of biochemical pathways. *Bioinformatics 19*(4), 532–538.

Jain A. K., Dubes R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc.

Jeong H., Tombor B., Albert R., Oltvai Z. N., Albert-László B. (2000). The large-scale organization of metabolic networks. *Nature 407*(6804), 651–654.

Jiang J. Q., Dress A. W., Yang G. (2009). A spectral clustering-based framework for detecting community structures in complex networks. *Applied Mathematics Letters 22*(9), 1479–1482.

Kaiser J. (1999). It'sa small Web after all. *Science 285*(1815).

Kalapala V. K., Sanwalani V., Moore C. (2003). The structure of the United States road network. *Preprint, University of New Mexico*.

Kasteleyn P., Fortuin C. (1969). Physica (utrecht) 57, 536 (1972) ; pw kasteleyn and cm fortuin. *Physical Society of Japan Journal Supplement 26*(11).

Kleinberg J. (2000). The small-world phenomenon : An algorithmic perspective. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pp. 163–170. ACM.

Krause A. E., Frank K. A., Mason D. M., Ulanowicz R. E., Taylor W. W. (2003). Compartments revealed in food-web structure. *Nature 426*(6964), 282–285.

Latora V., Marchiori M. (2002). Is the Boston subway a small-world network ? *Physica A : Statistical Mechanics and its Applications 314*(1), 109–113.

Lehmann S., Schwartz M., Hansen L. K. (2008). Biclique communities. *Physical Review E 78*(1), 016108.

Leicht E. A., Newman M. E. (2008). Community structure in directed networks. *Physical review letters 100*(11), 118703.

Li M., Wu J., Fan Y., Di Z. (2007). Econophysicists Collaboration Networks : Empirical Studies and Evolutionary Model. In *Econophysics of Markets and Business Networks*, pp. 173–182. Springer.

Lind P. G., González M. C., Herrmann H. J. (2005). Cycles and clustering in bipartite networks. *Physical review E 72*(5), 056127.

Luce R. D., Perry A. D. (1949). A method of matrix analysis of group structure. *Psychometrika 14*(2), 95–116.

Lusseau D., Newman M. E. (2004). Identifying the role that animals play in their social networks. *Proceedings of the Royal Society of London. Series B : Biological Sciences 271*(Suppl 6), S477–S481.

Mahdian A., Khalili H., Nourbakhsh E., Ghodsi M. (2006). Web graph compression by edge elimination. In *Data Compression Conference (DCC'06)*, pp. 1–pp. IEEE.

Martinez N. D. (1991). Artifacts or attributes ? Effects of resolution on the Little Rock Lake food web. *Ecological Monographs 61*(4), 367–392.

Maslov S., Sneppen K. (2002). Specificity and stability in topology of protein networks. *Science 296*(5569), 910–913.

Maslov S., Sneppen K., Zaliznyak A. (2004). Pattern detection in complex networks : Correlation profile of the Internet. *Physica A 333*, 529–540.

Milgram S. (1967). The small world problem. *Psychology today 2*(1), 60–67.

Moody J. (2001). Race, school integration, and friendship segregation in america1. *American Journal of Sociology 107*(3), 679—-716.

Newman M. E. (2001). Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality.

Physical review E 64(1), 016132.

Newman M. E. (2003a). Mixing patterns in networks. *Physical Review E 67*(2), 026126.

Newman M. E. (2003b). The structure and function of complex networks. *SIAM review 45*(2), 167–256.

Newman M. E. (2004). Fast algorithm for detecting community structure in networks. *Physical review E 69*(6), 066133.

Newman M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences 103*(23), 8577–8582.

Newman M. E., Girvan M. (2004). Finding and evaluating community structure in networks. *Physical review E 69*(2), 026113.

Newman M. E., Leicht E. A. (2007). Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences 104*(23), 9564–9569.

Palla G., Derényi I., Farkas I., Vicsek T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature 435*(7043), 814–818.

Palla G., Farkas I. J., Pollner P., Derenyi I., Vicsek T. (2007). Directed network modules. *New journal of physics 9*(6), 186.

Parisi G., Mézard M., Virasoro M. A. (1987). Spin glass theory and beyond. *World Scientific, Singapore 187*, 202.

Pimm S. L. (1979). The structure of food webs. *Theoretical population biology 16*(2), 144–158.

Pons P., Latapy M. (2006). Computing communities in large networks using random walks. *J. Graph Algorithms Appl. 10*(2), 191–218.

Radicchi F., Castellano C., Cecconi F., Loreto V., Parisi D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America 101*(9), 2658–2663.

Ramon F.-i.-C., Solé R. V. (2001). The small world of human language. *Proceedings of the Royal Society of London B : Biological Sciences 268*(1482), 2261–2265.

Reichardt J., Bornholdt S. (2004). Detecting fuzzy community structures in complex networks with a Potts model. *Physical Review Letters 93*(21), 218701.

Reichardt J., Bornholdt S. (2006). Statistical mechanics of community detection. *Physical Review E 74*(1), 016110.

Resende M. G. (2000). Detecting dense subgraphs in massive graphs. In *17th International Symposium on Mathematical Programming*.

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

Sales-Pardo M., Guimera R., Moreira A. A., Amaral L. A. N. (2007). Extracting the hierarchical organization of complex systems. *Proceedings of the National Academy of Sciences 104*(39), 15224–15229.

Scott J. (2012). *Social network analysis : A Handbook*. SAGE publications.

Sen P., Dasgupta S., Chatterjee A., Sreeram P. A., Mukherjee G., Manna S. S. (2003). Small-world properties of the Indian railway network. *Physical Review E 67*(3), 036106.

Shen H., Cheng X., Cai K., Hu M.-B. (2009). Detect overlapping and hierarchical community structure in networks. *Physica A : Statistical Mechanics and its Applications 388*(8), 1706–1712.

Shen Y., Pei W., Wang K., Li T., Wang S. (2008). Recursive filtration method for detecting community structure in networks. *Physica A : Statistical Mechanics and its Applications 387*(26), 6663–6670.

Strogatz S. H. (2001). Exploring complex networks. *Nature 410*(6825), 268–276.

Van Dongen S. (2000). *Graph clustering by flow simulation.* phd thesis, University of Utrecht, The Netherlands.

Wasserman S., Faust K. (1994). *Social network analysis : Methods and applications*, Volume 8. Cambridge university press.

Watts D. J. (1999). Small Worlds. *Princeton University Press*.

Watts D. J. (2004). *Six degrees : The science of a connected age*. WW Norton & Company.

Watts D. J., Strogatz S. H. (1998). Collective dynamics of 'small-world'networks. *Nature 393*(6684), 440–442.

Williams R. J., Martinez N. D. (2000). Simple rules yield complex food webs. *Nature 404*(6774), 180–183.

Zachary W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of anthropological research*, 452–473.

Zhang J., Zhang S., Zhang X.-S. (2008). Detecting community structure in complex networks based on a measure of information discrepancy. *Physica A : Statistical Mechanics and its Applications 387*(7), 1675–1682.

Zhang P., Wang J., Li X., Li M., Di Z., Fan Y. (2008). Clustering coefficient and community structure of bipartite networks. *Physica A : Statistical Mechanics and its Applications 387*(27), 6869–6875.

Zhou H. (2003a). Distance, dissimilarity index, and network community structure. *Physical review e 67*(6), 061901.

Zhou H. (2003b). Network landscape from a Brownian particle's perspective. *Physical Review E 67*(4), 041908.

Zhou H., Lipowsky R. (2004). Network brownian motion : A new method to measure vertex-vertex proximity and to identify communities and subcommunities. In *International conference on computational science*, pp. 1062–1069. Springer.

A BIOGRAPHIE

Yacine SLIMANI est maître de conférences en informatique à la faculté de technologie, université Ferhat Abbes, Sétif1, Algérie. Il a obtenu le diplôme d'ingénieur en Informatique à l'université Ferhat Abbes de Sétif en 1997, ensuite le diplôme de Magistère en Informatique à l'université Ferhat Abbes de Sétif en 2006. Il est membre au laboratoire de systèmes intelligents (LIS). Il a obtenu le diplôme de doctorat sciences en Décembre 2018. Cette thèse a été réalisée sous la direction de Prof Moussaoui Abdelouhab à l'université Farhat Abbes, en collaboration avec Professeur Yves Lechevalier, équipe axis d'INRIA Paris, France. Ses domaines d'intérêt en recherche sont : le web usage mining, les réseaux complexes, la découverte de communautés, les algorithmes méta-heuristique et l'intelligence artificielle.

Ahlem DRIF diplôme de doctorat sciences en informatique et elle est actuellement maître de conférences en informatique à la faculté des sciences, université Ferhat Abbes, Sétif1, Algérie. Elle a obtenu le diplôme d'ingénieur en Informatique à l'université Ferhat Abbes de Sétif en 2002, ensuite le diplôme de Magistère en Informatique à l'université Ferhat Abbes de Sétif en 2006. Il est membre au laboratoire réseaux et système distribués (LRSD). Ses domaines d'intérêt en recherche sont : les réseaux complexes, la découverte de communautés, la diffusion d'information dans les réseaux sociaux et l'apprentissage profond et l'intelligence artificielle.