Découverte de communautés dans les réseaux complexes

Ahlem Drif, Abdallah Boukerram, Yacine Slimani, Abdelouaheb Moussaoui

Découverte de communautés

dans les réseaux complexes

Ahlem DRIF* — Abdallah BOUKERRAM** — Yacine SLIMANI * — Abdelouaheb MOUSSAOUI*

* Département d'informatique
Université Ferhat Ababs Setif 1
19000 Setif - Algerie
E-mail: adrif.univsetif@gmail.com, slimani_y09@univ-setif.dz, moussaoui.abdel@gmail.com

** Département d'informatique
Université Abderrahmane Mira,
06000 Bejaia - Algerie
E-mail: aboukerram@hotmail.com

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.

ABSTRACT. The community detection in complex networks has attracted a growing interest and is the subject of several researches that have been proposed to understand the network structure and analyze the network properties. In this paper, we give a thorough overview of different community discovery strategies, we propose taxonomy of these methods, and we specify the differences between the suggested classes which helping designers to compare and choose the most suitable strategy for the various types of network encountered in the real world.

MOTS-CLÉS : Méthodes de découverte de communautés, réseaux complexes, réseaux sociaux, graphes, fonction de qualité

KEYWORDS : Community detection methods, complex networks, quality function, graph, social networks.

1. 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 [Sco12], les réseaux des réactions chimiques entre protéines dans le métabolisme d'un être vivant [HHJ03, JTAO00], les réseaux des pages web [AJAL99] qui contiennent plusieurs millions de noeuds, les réseaux trophiques [WM00], les réseaux de dictionnaires [BS02] ... 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 [WS98] et l'échelle-libre [BA99], des propriétés dynamiques tel que la diffusion [BP01, ESMS03] et des propriétés structurelle comme la structure de communauté [Moo01, FLGC02, GN02, LN04, KFMU03, GA05, PDFV05] qui paraît être commune à beaucoup de réseaux et permet de construire un pont pour 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 dans les réseaux complexes. Il s'agit d'une problématique proche des problématiques classiques de clustering de données et de partitionnement de graphes.

Une communauté peut être considérée comme un ensemble de noeuds qui sont très reliés entre eux au niveau d'un même groupe mais ont de faibles interactions avec des noeuds en dehors de ce groupe. 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 très variés ce qui a résulté d'une diversité des méthodes de découverte de communautés.

Notre travail discute les méthodes de découverte de communautés dans les réseaux complexes, son principal objectif, est de donner une étude synthétique, bibliographique, et comparative des différentes méthodes proposées dans la littérature et citer éventuellement leurs apports et leurs inconvénients. Nous proposons par la suite une taxonomie des méthodes de découverte de communautés tout en précisant la différence entre les différentes classes proposées.

2. 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, la disponibilité croissante des données précises 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 qui a été concernée par des problèmes des réseaux réels, mais son approche est orientée vers la conception et la technologie, 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 du réseau social pourrait avoir demandé : " quel est le noeud qui 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 pas du tout d'effet). Désormais, la question qui devrait être posée : " Quel pourcentage des sommets a besoin d'être 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 établi dans le processus de recherches le nouveau champ de la théorie des réseaux, ou la "science des réseaux" [AL02, Buc03, Wat04]. 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 [WS98] qui sont issus de domaines aussi variés que la biologie, l'économie ou la sociologie.

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

Cette science est distinguée 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 de système pour proposer des mesures appropriés à ces propriétés.

– Elle vise la création des modèles de réseau qui permettent la compréhension de toute signification des propriétés des réseaux du monde réel.

– Elle vise la prédiction de la dynamique de comportement des systèmes à la base des propriétés structurales mesurées et des règles locales régissant les différents sommets ou 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 [WS98, Str01, AB02, New03b, DM02]. 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 [WS98] comme la combinaison d'un fort coefficient de clustering et d'un petit diamètre. Cette propriété étudiée par le psychologue [Mil67] est vérifiée par le modèle de graphes aléatoires d'Erdös-Rényi [ER59]. Pour pallier aux limites de modèle d'Erdös-Rényi, plusieurs travaux ont été publiés tel que le travail [Bol98].

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 forte densité traduit le fait que les liens entre les sommets qui sont proches sont beaucoup plus probables que les liens entre les sommets éloignés. Cette densité est souvent capturée par le coefficient de clustering [WS98] qui compte la probabilité que deux voisins d'un même sommet soient eux-mêmes liés par une arête. Elle s'explique par la tendance des acteurs à se regrouper en modules ou communautés.

2.3.3. Distribution des degrés "Degree distributions "

Il existe dans les grands réseaux un nombre non négligeable de sommets possédant un très fort degré par rapport à une majorité de sommets possédant un très faible degré. Cette distribution est souvent bien approximée par une loi de puissance pour laquelle 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 $k = 10^6$). En 1999, Faloutsos et al [FFF99] ont observé que le réseau Internet présentait cette propriété. Par la suite, elle a également été observée dans des réseaux de pages web, et des réseaux de distribution d'électricité [New03b].

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 typique des chemins augmentera, et des paires de sommets devenus déconnectées et la communication entre elles deviendra impossible. Le niveau de résilience de réseau se varié selon la déconnexion des sommets. Un intérêt récent pour la résilience de réseau a été suscité par le travail d'Albert et al [AJAL00].

2.3.5. Mixing patterns

Dans la plupart des réseaux il existe quelques types différents de sommets et la probabilité qu'il existe un lien entre une paire de sommets différents dépend souvent du type en question. Cette propriété surprenante dans un réseau complexe explique la différence des types de ses sommets. Maslov et al [MSZ04] ont étudié l'existence de trois type de noeud dans le réseau Internet : les fournisseurs qui ont 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 noeuds précédents.

2.3.6. Corrélation entre degré "Degree correlations "

Comme le degré est lui-même une propriété qui représente la topologie du réseau, la corrélation de degré peut fournir des détails intéressants sur l'effet de la structure du réseau. En utilisant la propriété de corrélation de degré on veut savoir si les sommets de fort degré sont associé préférentiellement avec les sommets de haut degré ou bien avec

ceux de faible degré. Plusieurs études ont été proposées pour quantifier la corrélation de degré à l'exemple des travaux de Maslov et al [MSZ04, MS02].

2.3.7. Navigabilité "Network navigation"

Kleinberg en 2000 [Kle00] a introduit la notion de navigabilité. Il s'agit de caractériser non seulement la longueur des chemins, mais aussi la façon dont ils sont découverts. Dans l'expérience de Milgram, les individus n'utilisent que leurs contacts locaux pour acheminer la lettre ; il s'agit donc d'un routage décentralisé, en ce sens que l'on n'utilise qu'une vue locale du réseau pour transmettre le message. C'était aussi le cas de la navigation à travers le réseau des pages web il y a quelques années, qui se faisait d'une page à l'autre sans connaître la carte globale du réseau [AJAL99, Kai99]. Par ailleurs, la découverte des chemins de façon décentralisée est une nécessité pour les réseaux d'interactions réels qui comportent un très grand nombre de sommets et où une recherche classique des plus courts chemins n'est pas possible, car très coûteuse en temps.

2.3.8. Structure de communauté "Community structure"

Dans les réseaux complexes, la présence de groupes de sommets (communautés) fortement liés entre eux et faiblement liés avec l'extérieur fonde la propriété de structure de communauté. Les communautés sont des groupes de sommets qui partagent probablement des propriétés communes et/ou rôles semblables dans le réseau en question. Ainsi, les communautés peuvent correspondre aux groupes de pages Web traitant le même sujet [FLGC02], aux modules fonctionnels tels que les cycles et les voies dans les réseaux métaboliques [GA05, PDFV05], aux groupes d'individus relatifs dans les réseaux sociaux [GN02, LN04], et aux subdivisions dans les chaînes alimentaires [Pim79, KFMU03] ...ect. La figure 1 montre une visualisation du réseau d'amitié des enfants dans une école des USA pris d'une étude effectué par Moody [Moo01], le réseau semble avoir une forte structure de communauté qu'en fait les communautés apparaissent clairement sur la figure. D'ailleurs, quand Moody a coloré les sommets selon la race de chaque individu, comme le montre la figure 1, il apparaît immédiatement que la formation de communautés dans ce cas est due principalement à la propriété race d'individu.

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

Dans ce présent rapport, nous nous sommes intéressés à la découverte de la propriété de structure de communauté. Les objectifs principaux qui ont motivés les études portées sur la 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 permettent de donner un point de vue macroscopique sur la structure des graphes. Elles permettent par exemple de regrouper et d'identifier les sommets qui jouent potentiellement des rôles similaires. Par exemple, la détection de communautés dans le graphe du Web est une piste envisagée pour améliorer les moteurs de recherche [FLGC02].

– La détection de communautés peut aussi être utilisée pour la visualisation des graphes complexes [ACJM03].

– Les communautés sont utilisées pour diviser le graphe afin d'effectuer des calculs séparés moins coûteux sur chaque communauté. Ce procédé de parallélisme de calcul permet d'envisager des gains de complexité pour les algorithmes qui s'appliquent sur

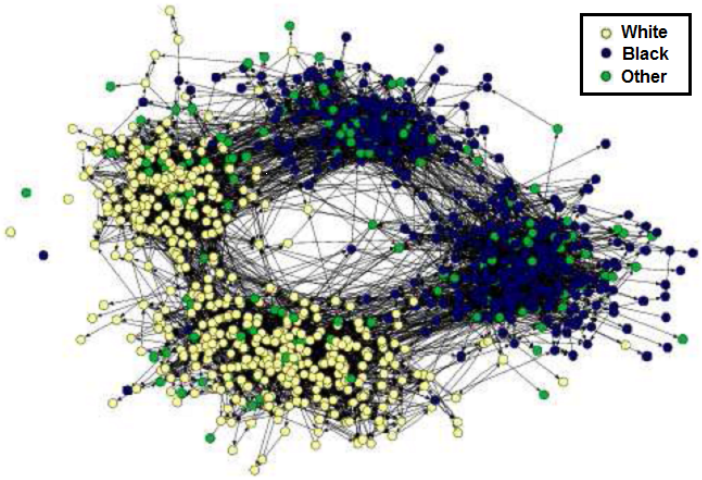

Figure 1. *Réseau d'amitié des enfants dans une école des USA. Les liens d'amitiés sont déterminées en demandant aux participants, et par conséquent ils sont orientés, puisque A peut dire que B est son ami mais pas forcement vice versa. Les sommets sont colorés selon la race. La division de haut en bas est entre l'école moyenne et l'école secondaire.*

certains grands graphes.

– La détection de communautés permet une classification des sommets, selon leur position topologique dans les groupes. Ainsi, les sommets possédant une position centrale dans leur clusters partagent un grand nombre de liens avec les autres groupes, ils peuvent avoir une importante fonction de contrôle et de stabilité au sein du groupe, quand aux sommets de frontières, 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 [Gra73, Bur76, Fre77].

– Les communautés peuvent être utilisées pour améliorer les méthodes de compression de graphes à l'exemple du travail [MKNG06].

3. Description des communautés

La notion de communauté dans un graphe est cependant difficile à définir formellement. C'est pourquoi toutes les approches récentes ont utilisé une notion intuitive des communautés. Une communauté est alors vue comme un ensemble de sommets dont la densité de connexions internes est plus forte que la densité de connexions vers l'extérieur. Le but est alors de trouver une partition des sommets en communautés vérifiant ce critère (sans savoir a priori le nombre de telles communautés).

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 : [WF94]

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 : [RCCLP04]

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

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

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

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

Tel que :

- $k_i^{in}(V) = \sum_{j \in V} A_{ij}$: est le nombre de liens reliant le noeuds i à d'autres noeuds appartenant à V ;
- $k_i^{out}(V) = \sum_{j \notin V} A_{ij}$: est le nombre de liens vers les noeuds qui n'appartiennent pas à V (le reste du réseau).

Définition d'une communauté au sens fort :

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

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

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 :

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

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 [BK73] pour trouver les cliques résulte d'un coût de calcul élevé (complexité exponentielle).

Définition 2 : [New06]

Une communauté est un sous graphe indivisible.

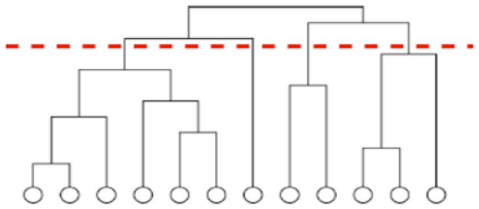

Figure 2. *Dendrogramme illustrant l'output d'un algorithme de détection de communautés*

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.

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 ? Quand est ce qu'on coupe le dendrogramme pour obtenir le niveau de division adéquat du réseau ou bien le nombre de communautés appropriées ?

Pour répondre à ces questions, Newman et al [NG04] ont introduit une mesure de la qualité d'une division particulière du réseau appelé "modularité". Cette mesure est basée sur la mesure d'assortative mixing proposée déjà par Newman dans [New03a].

Imaginer 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. Les auteurs considèrent tous les liens dans le réseau original.

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

Alors, on peut écrire : $e_{ij} = a_i.a_j$, c'est le nombre de liens intra-communauté qui sont prévus.

Ainsi, la mesure de modularité est définit comme suit :

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

Cette expression mesure la fraction des liens reliant des sommets de la même communauté sans la valeur prévue de la même quantité aux mêmes partitions de communauté mais en employant des connexions aléatoires entre les sommets.

Si une partition particulière ne possède plus de liens intra-communautés (une seule communauté) comme il a été prévu dans le modèle aléatoire, la modularité est $Q = 0$. Les valeurs autres que 0 indiquent des déviations de l'aspect aléatoire, et les valeurs supérieures à 0.3 semblent indiquer l'existence d'une structure de communauté significative. Pratiquement, les valeurs de modularité pour de tels réseaux sont souvent entre 0.3 à 0.7, alors que les valeurs les plus élevées sont rares. Cependant, il est possible que les partitions de meilleures modularités ne correspondent pas aux partitions en communautés les plus pertinentes [FB07].

4. 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 du réseau en question. 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 [WF94] dans lequel les acteurs sont des individus ou entités sociales (associations, entreprises, pays,...etc). 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 contact physique (deux individus sont reliés s'ils ont été physiquement en contact), les réseaux de collaboration (deux individus sont reliés s'ils ont travaillé ensemble, en particulier de nombreux travaux ont étudié les collaborations scientifiques [New01]), les réseaux d'appels téléphoniques [Res00] (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 [GLLB04] ou un courrier électronique [EMB02] par exemple), ...etc.

4.2. Réseaux biologiques

Les réseaux biologiques sont assez divers parmi lesquels il existe les réseaux métaboliques [JTAO00] (les sommets sont des gènes ou des protéines qui sont liés par leurs interactions chimiques), les réseaux de neurones (chaque neurone est connecté à plusieurs autres neurones) ou les réseaux trophiques [WM00] (les espèces d'un écosystème sont reliées pour représenter les chaînes alimentaire). Un exemple d'une chaîne alimentaire [Mar91] est illustré sur la figure 3.

4.3. Réseaux d'information

L'exemple classique du réseau d'information est le réseau de citation des papiers. La plupart des articles citent les travaux précédents des autres auteurs sur le même sujet. 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 du réseau de citation reflète

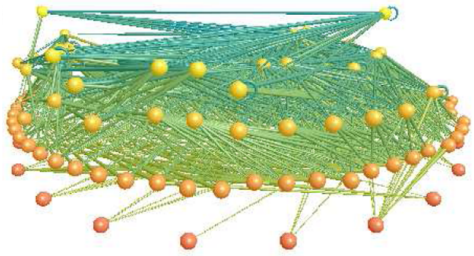

Figure 3. *Une chaîne alimentaire des interactions de prédateur-proie entre les espèce dans un lac d'eau [Mar91].*

la structure de l'information stockée dans ses sommets. Un autre exemple très important du réseau d'information est le réseau World Wide Web, dont les pages Web contenant l'information, liée ensemble par les lien hypertextes d'une page à l'autre [ER90].

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. Plusieurs études statistiques ont été menées sur la grille d'électricité par Watts et Strogatz [WS98, Wat99] et Amaral et al [ASBS00]. Nous pouvons aussi citer d'autres réseaux de distribution, qui ont été étudiés, tels que le réseau des itinéraires de ligne aérienne [ASBS00], les réseaux des routes [KSM03], réseau de chemins de fer [SDCS03, LM02]. Les réseaux de fleuve ont aussi pu être considérés comme forme d'occurrence des réseaux de distribution [DR00].

4.5. Réseaux linguistiques

Ces réseaux relient les mots d'un langage donné et regroupent entre autres les réseaux de synonymes (deux mots sont reliés s'ils sont synonymes), les réseaux de co-occurrences [iCS01] (deux mots sont reliés s'ils apparaissent dans une même phrase d'un ouvrage) ou encore les réseaux de dictionnaires [BS02] (deux mots sont liés si l'un est utilisé dans la définition de l'autre).

Figure 4. *Réseau des citations*

5. Méthodes de découverte de communautés dans les réseaux

La méthode proposée par Girvan et Newman [GN02] a marqué le début d'une nouvelle ère dans le domaine de la découverte de communautés dans les réseaux complexes. Depuis ce travail qui fait référence dans le domaine, le sujet a reçu une extraordinaire attention de la part de la communauté scientifique et de très nombreuses nouvelles approches sont 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.

La première, initialement introduite pour la parallélisation de processus, cherche à répartir des tâches, représentées par les sommets d'un graphe, tout en minimisant les échanges, représentés par les arêtes mais les algorithmes de partitionnement de graphe ne peut pas être utilisées pour la découverte de communauté car on ne connaît pas à l'avance le nombre de communautés que l'on cherche et leurs tailles.

La seconde thématique de clustering de données est une thématique générale plus vaste dans laquelle on cherche à regrouper des données possédant des caractères communs. 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 usuelles de clustering 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, principalement pour des questions de taille et de mesure de similarité adapté pour déterminer les clusters.

Nous allons lister ici les principales approches qui ont été proposées à ce jour. Bien que la liste soit importante, elle est non exhaustive. Nous avons retenu les approches qui ont reçu le plus d'attention de la part de la communauté scientifique. Notre but est de donner une vue d'ensemble des méthodes proposées, d'en illustrer la diversité, et de discuter leurs avantages et leurs inconvénients.

5.1. Betweenness-based algorithm

5.1.1. Idée fondamentale

Les algorithmes proposés par Newman et Girvan [GN02, NG04] diffèrent des algorithmes séparatifs existants dans le fait que ceux-ci ne se base 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 une nouvelle mesure appelé centralité "betweenness" qui se focalise sur les lien enter-communautés.

5.1.2. Mesure de communautés

5.1.2.1. Betweeness measure

La première mesure de betweenness, qui a été définit dans [NG04], consiste à trouver les plus courts chemins entre toutes paires de sommets et calculer l'implication de chaque lien le long de ces chemins.

La nouvelle approche de Newman et al est inspirée du travail de Freeman [Fre77]. 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 le long de leurs plus courts chemins de communication [Fre77].

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 $O(mn^2)$ [AMO94, CLRS01]. Newman a proposé [New01] un algorithme performant qui calcule le shortest path betweeness en $O(mn)$, l'algorithme se déroule comme suit :

1) La distance du sommet initial s est $d_s = 0$, son poids est $w_s = 1$,

2) Pour chaque sommet i adjacent au sommet s :
$d_i = d_s + 1$, le poids est $w_i = w_s = 1$,

3) Pour chaque sommet j adjacent à un de ces sommets i, faire :

a) Si aucune distance n'as pas été encore assignée au sommet j,
on assigne au j : $d_j = d_i + 1$, et $w_j = w_i$,

b) Si une distance a été assignée au sommet j et $d_j = d_i + 1$,
donc le poids wi est incrémenté à wi : $w_j = w_j + w_i$,

c) Si une distance a été assignée au sommet j et $d_j < d_i + 1$,
ne rien faire.

4) Répéter à partir de l'étape (3) jusqu'à ce qu'il ne reste aucun sommet dont il possède une distance et ses voisins n'ont pas une distance assignée.

Physiquement, le poids d'un sommet i représente le nombre des chemins distincts depuis le sommet source à ce sommet i. Ces poids sont utilisés pour calculer la mesure edge betweenness car si deux sommets i et j sont connectés, avec j est plus lointain que i du sommet source s, alors la fraction du chemin géodésique de j à s passant par i est donné par : w_i/w_j .

Ainsi, pour calculer la contribution de tous les plus courts chemins via la mesure de edge betweeness, partant de s, on applique les étapes suivantes :

1) Trouver chaque feuille t, c'est à dire un sommet qui n'a aucun chemin de s aux autres sommets et qui traverse t.

2) Pour chaque sommet i voisin de t,
assigner un score au lien de t à i de : $wi_/w_t$.

3) Ensuite, commençant par les liens qui sont les plus loin du sommet source s,
pour tous lien (i, j), tel que j est plus lointain que i de la source s,
assigner un score qui est égale à $1+$ la somme des scores des liens des voisins directs (les liens qui partage un sommet avec le lien (i, j)), cette valeur est multiplié par la fraction $wi_/w_t$.

4) Répéter l'étape (3) jusqu'à ce que s soit atteint.

Ce processus est répété pour chaque source de 1 à n en faisant la somme des résultats de scores qui représente le total de betweeness pour tous les liens. L'algorithme s'exécute en $O(n^3)$.

5.1.2.2. Random walk betweeness measure

Une autre mesure qui se base sur le signal qui circule à travers le réseau, 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 par le même taux constant, alors betweenness est une mesure du taux des signaux qui passent le long de 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 nouvelle mesure que Newman et al [NG04] l'appelé "random walk betweeness" (le nombre de périodes dans lesquelles le pas passe le long du 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 :

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

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 pas qui atteignent 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.

5.1.3. Méthode de détection des communautés

L'algorithme de détection de communautés proposé par Newman et al [NG04] se déroule comme suit :

1) Calcul des scores 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 à partir de l'étape (2) jusqu'à quand plus de liens restent.

La suppression d'un lien affecte seulement la mesure betweeness des liens qui appartiennent au même composant, donc on aura besoin de recalculer la mesure seulement dans ces composants.

5.1.4. Exemple d'expérimentation

L'application de l'algorithme de Newman et Girvan sur des réseaux du monde réel à prouver l'efficacité de l'algorithme en qualité de détection de communauté :

a) Sur le réseau de club de karaté de Zachary [Zac77] (Fig. 5), l'algorithme résulte de deux communautés qui correspondent presque parfaitement aux deux groupes du réseau réel. Le dendrogramme correspondant à cette division est donné sur Fig. 6. La modularité résultante de la division des deux version de l'algorithme (shortest path betweeness et random walk betweeness) est très élevés quand le réseau est divisé en deux communautés (environ 0.4), ce qui indique qu'il existe une forte division naturelle à ce niveau. Cependant, le sommet 3 a été mal classé par la version shortest path betweeness. Quand au random walk betweeness algorithm, tous les noeuds sont parfaitement classés.

b) Sur 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 $Q = 0.54$

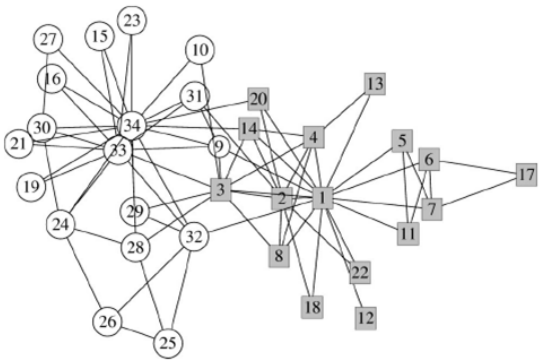

Figure 5. *Le réseau d'amitiés entre les individus dans l'étude de club de karaté de Zachary. L'administrateur et l'instructeur sont représentés par les noeuds 1 et 33, respectivement. Les carrés représentent les individus qui appartiennent au club administrateur après la fission du club et les cercles représentent ceux qui sont aligné avec l'instructeur.*

et correspond aux 11 communautés. Les communautés reflètent clairement la structure d'intrigue du livre (Jean Valjean et Javert forment le centre des communautés avec leurs adhérents respectifs).

5.1.5. Discussion

L'algorithme présente de meilleure qualité de division, 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. Aussi, ils effectuent la division du réseau même si aucune bonne division n'existe.

5.2. Fast algorithm

5.2.1. Idée fondamentale

Partant du principe qu'une grande valeur de modularité représente une bonne division en communautés, Newman [New04] a proposé d'optimiser cette valeur sur toutes les partitions possible afin de trouver la meilleur.

5.2.2. Mesure de communautés

Dans le Fast algorithm [New04], la mesure utilisée pour découvrir les communautés n'est que le changement de Q en joignant deux communautés.

5.2.3. Méthode de détection de communautés

Newman [New04] a proposé une méthode d'optimisation qui se base sur l'algorithme d'optimisation gloutonne. L'algorithme se déroule comme suit :

1) Initialement, on considère que 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)

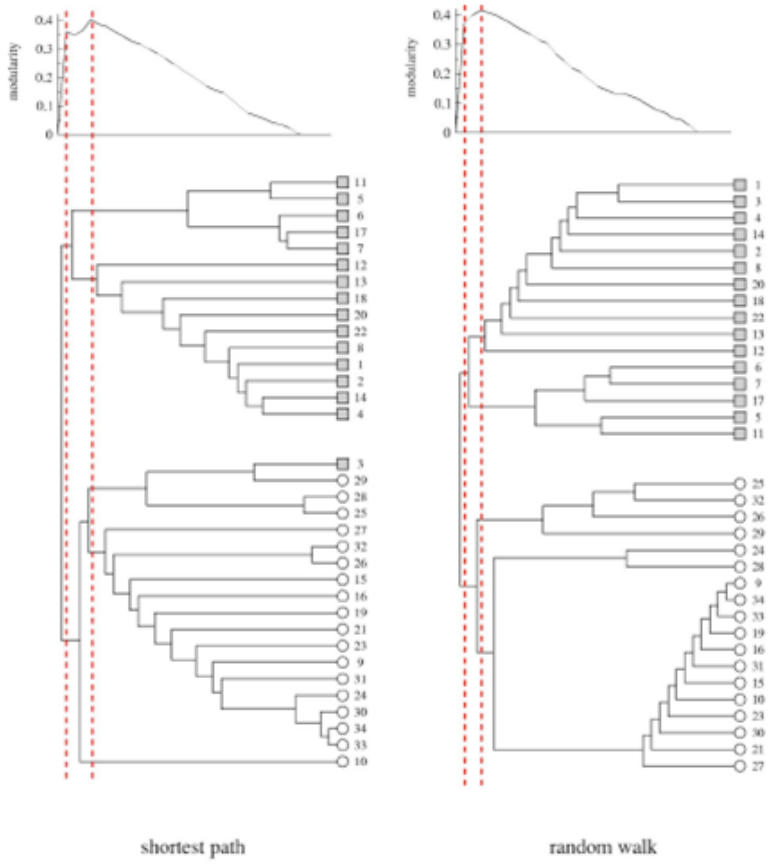

Figure 6. *Les dendrogrammes qui représentent les communautés généré par l'exécution des algorithme shortest path betweeness et random walk betweeness.*

3) La mis à jours des éléments de la matrice e_{ij} et de la matrice elle-même 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.

5.2.4. Exemple d'expérimentation

Afin d'étudier la performance des algorithmes de détection de communauté, plusieurs réseaux ont été utilisés, ces réseaux sont ceux du monde réel ou bien des graphes générés avec une structure de communautés connues.

Les auteurs ont généré un grand nombre de graphes avec $n = 128$ sommets, divisés en quatre communautés de 32 sommets chacune. Les liens ont été mis 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 distincts est P_{out}, 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 de la fraction des sommets correctement identifiés selon la variation de sont illustrés sur Fig. 8. Le Fast algorithm identifie correctement plus que 90% des sommets pour des valeurs $z_{out} \leq 6$. Cependant, quand les liens intra-communauté et les

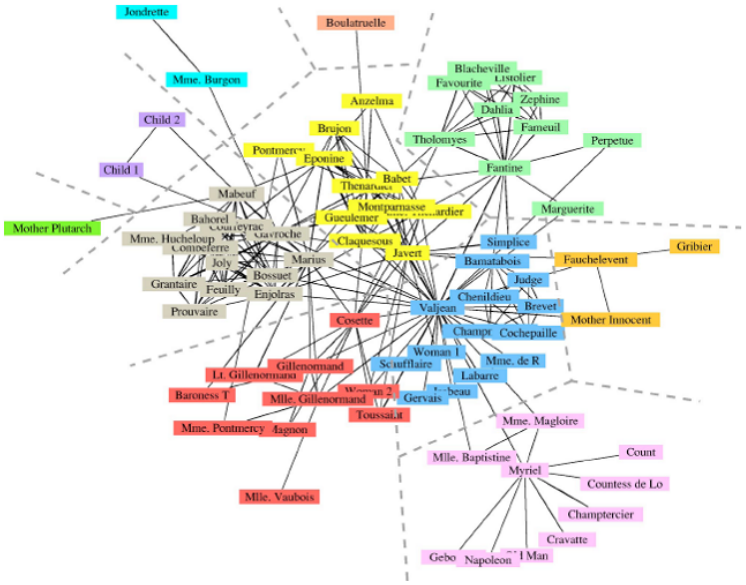

Figure 7. *Les communautés des personnages du roman Les Misérables de Victor Hugo.*

liens inter-communauté par sommet deviennent égaux (z_{out} est proche de la valeur 8) la performance de l'algorithme se dégrade. Quant à l'algorithme GN [New04] il identifie correctement les sommets mieux que fast algorithm [New04] pour des petites valeurs z_{out} (pour $z_{out} = 5$: GN détecte correctement 98.9% de sommets , fast algorithm détecte 97.4% de sommets). Pour des valeurs plus élevées de z_{out} le fast algorithm s'opère mieux que l'algorithme GN.

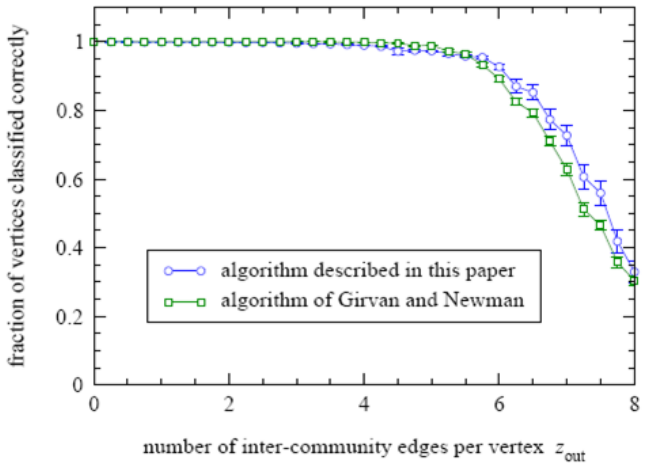

Figure 8. *Les résultats de la fraction des sommets correctement identifiés selon la variation de* z_{out}.

L'application de fast algoithm sur le réseau de Zachary donne des résultats presque similaires à ceux trouver par l'algorithme GN.

5.2.5. Discussion

Le fast algorithm base ses décisions sur les informations locales des différentes communautés, tandis que l'algorithme GN emploie les informations du réseau entier. Comme la structure de communauté est une quantité non local, les algorithmes qui se basent sur l'information non local tel l'algorithme GN trouve plus correctement cette structure.

En cas d'exécution des deux algorithmes sur un réseau de petite taille la différence du temps ne présentera pas un grand problème dans la plupart des situations pratiques, mais sur les systèmes de grande taille l'algorithme GN s'exécute en temps inadmissible. A titre d'exemple, l'exécution de fast algorithm sur le réseau des musiciens du jazz contenant 1275 noeuds se termine en une seconde du temps CPU, alors que GN s'exécute en 3 heures. Cependant, l'algorithme de Newman ne détecte pas les chevauchements entre les communautés.

5.3. Fast community detection algorithm based on a q-state Potts model

5.3.1. Idée fondamentale

Reichardt et al [RB04] ont combiné l'idée de Fu et Anderson [FA86] avec le modèle de clustering de Potts qui a été définit par Blatt et al [BWD96], ceci a permet de convertir les communautés du réseau vers le domaine magnétique. Dans [RB06], Reichardt et al ont proposé un framework pour détecter les communautés en déterminant l'état fondamental de "q-Potts model spin glass" [PMV87].

5.3.2. Mesure de communautés

Reichardt et al [RB04] 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 [KF69]. 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 [FA86] 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. Si on prend, par exemple, le problème de partitionnement de graphes en deux sous graphes, le nombre de liens qui existent entre les deux sous graphes égale à :

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

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 :

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

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

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

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 [RB04] ont modifié l'Hamiltonien de Potts à q états en ajoutant une contrainte globale :

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

Tel que :

E **:** est l'ensemble de liens,

σ_i **:** dénote les spins individuels $_{(i=1,\ldots,N)}$ qui peuvent prendre les valeurs $_{1,\ldots,q}$.

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.

5.3.3. Méthode de découverte de communautés

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. Les auteurs ont utilisé "Monte Carlo single spin flip heat-bath algorithm" pour déterminer l'état fondamental du système (structure de communauté) et l'optimisation de l'énergie du système (le deuxième terme de l'Hamiltonien) est faite par le recuit simulé. Cette minimisation d'énergie correspond à favoriser les liens intra-communauté et optimiser les liens inter-communauté.

5.3.4. Exemple d'expérimentation

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). Les auteurs ont définit deux mesures : Sensibilité et spécificité ; une paire de noeuds est positif (négatif) quand il est dans la même communauté (différente communauté), sensibilité (spécificité) désigne la fraction de tout les paires de noeuds positif (négatif) qui sont classifiés correctement par l'algorithme. Selon Fig. 9, les performances de l'algorithme sont aussi bien que la méthode de GN.

5.3.5. Discussion

L'algorithme proposé par Reichardt et al [RB04], par sa nature non déterministe et non hiérarchique, est capable de détecter l'affiliation des noeuds qui appartiennent aux

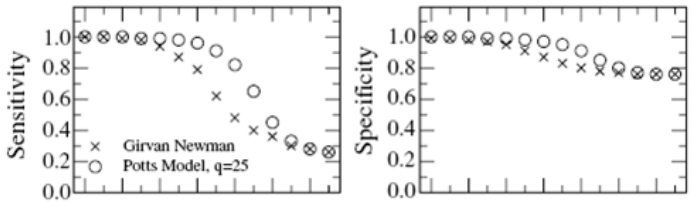

Figure 9. *Comparaison des résultats de la fraction des sommets correctement identifiés par les algorithmes GN et de Reichardt et al [RB04] .*

plusieurs communautés, il s'adapte bien aux structures de communautés confuses 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.

5.4. Extremal optimization algorithm

5.4.1. Idée fondamentale

Duch et al [DA05] 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.

5.4.2. Mesure de communautés

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 lié à la contribution de différents noeuds i dans la modularité, elle est normalisée dans l'intervalle $[-1, 1]$ et donnée comme suit :

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

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.

5.4.3. Méthode de découverte de communautés

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. Les mouvements change les partitions, donc la valeur de fitness doit être recalculée ;

3) Les liens entre les deux partitions sont supprimés et on procède de même pour chaque nouveau composant ;

4) Le processus est répété jusqu'à ce que la modularité Q ne puisse pas être amélioré (état optimal est atteint).

5.4.4. Exemple d'expérimentation

L'application 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. 10 et Fig. 11). La valeur de la modularité est 0.419 supérieur à la valeur 0.318 produite par l'algorithme de Newman [New04], à 0.406 reporté par l'algorithme de Reichardt et al [RB04] et à 0.412 reporté par Donetti et al [DM04].

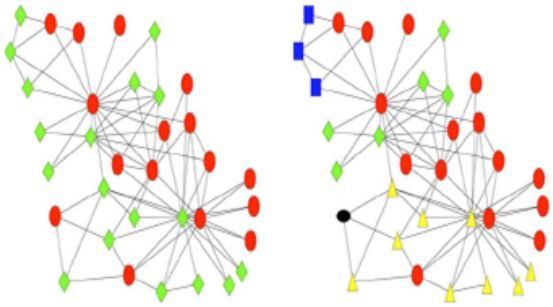

Figure 10. *Division initiale aléatoire du réseau de Zachary (le nombre de composants initial connecté dans les deux partitions est 5).*

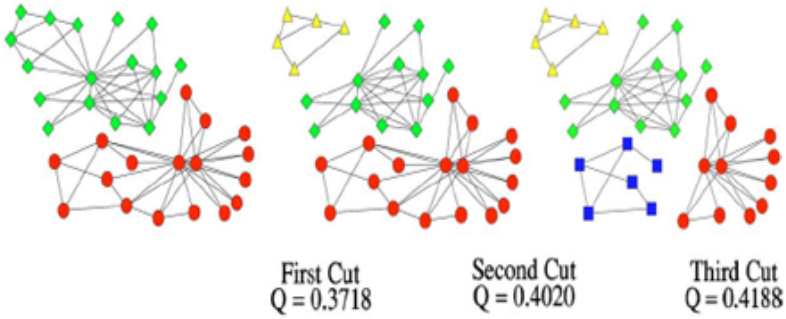

Figure 11. *Communautés obtenues après application de l'algorithme d'optimisation extrémale.*

5.4.5. Discussion

Bien que l'algorithme de Duch et al [DA05] produit une modularité optimale qui dépasse plusieurs algorithmes existants et tous ça en optimisant la complexité de calcul en $O(n^2 log(n))$, il n'offre pas une bonne partition en communautés et ne reflète pas le réseau réel car le résultat final dépend étroitement de l'étape d'initialisation du réseau en partition aléatoire. Ceci montre que l'optimisation en soi d'une fonction de qualité ne conduit pas forcément à la partition souhaitée.

5.5. Eigenvector-based algorithm

5.5.1. Idée fondamentale

Motivé par les similitudes et les différences entre les méthodes de découverte de communautés et de partitionnement des graphes, Newman [New06] 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.

5.5.2. Mesure de communautés

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 :

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

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 [NG04].

L'auteur a définit une nouvelle matrice symétrique de modularité B, qui s'écrit comme suit :

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

La somme des éléments de colonnes et de lignes est à zéro, de sorte qu'il ait toujours un vecteur propre (1,1,1,...) avec la valeur propre zéro. Cette observation est retenue de la matrice connue sous le nom de graphe de Laplacien 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 :

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

Le choix des éléments du vecteur s qui optimisent la modularité revient à résoudre un problème NP-hard similaire aux problèmes de partitionnement spectrale. Dans ce cas, une approximation très simple et efficace peut être trouvée en maximisant la valeur propre principale et ignorant complètement toutes les autres valeurs.

5.5.3. Méthode de détection de deux communautés

1) Calculer le vecteur propre principal de la matrice de modularité (u_1).

2) Diviser les noeuds en deux groupes selon le signe des éléments correspondants dans ce vecteur. Les noeuds qui ont une valeur positive sont mis dans un groupe et les autres noeuds dans le deuxième groupe).

5.5.4. Méthode de détection de plusieurs communautés

L'auteur [New06] 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 :

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

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 en question 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 division.

4) Arrêter le processus de division si la modularité devient nulle ou négative (le sous graphe est indivisible).

5) Si tous les sous graphes sont indivisibles arrêter l'algorithme.

5.5.5. Exemple d'expérimentation

Newman a appliqué son algorithme sur le réseau des 105 récent livres de la politique américaine qui sont vendus sur le site Amazon.com. Sur la figure ci-dessous (Fig. 12) , 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" écoule dans les deux restantes communautés. Donc ces livres forment des communautés d'achat qui sont alignées étroitement avec les vues politiques, ce qui démontre que l'algorithme proposé par Newman est capable d'extraire des résultats significatifs.

5.5.6. Discussion

Newman [New06] a proposé une approche différente des approches existantes basée sur une reformulation de la modularité en termes de propriétés spectrales du réseau complexe. Ces propriétés ont permet d'apporter plusieurs avantages à l'algorithme proposé. L'algorithme fonctionne même si les tailles des communautés ne sont pas précisées. 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 et elle sépare les noeuds selon une échelle continue qui répond à la question : combien un noeud fait partie d'un groupe ou d'un autre ? Cependant, le coût de calcul des vecteurs propres est élevé.

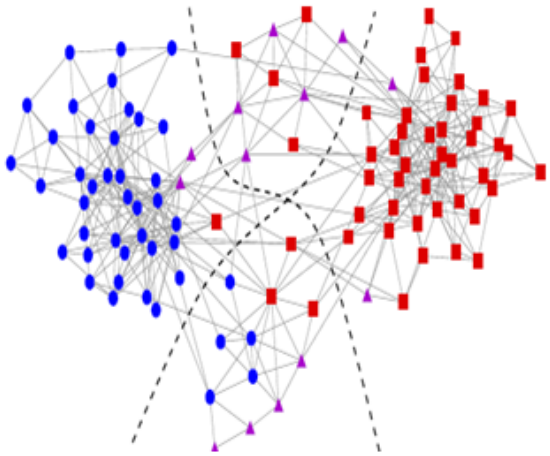

Figure 12. *Krebs' network of books on American politics.*

5.6. Recursive filtration algorithm

5.6.1. Idée fondamentale

Shen et al [SPWL08] ont proposé une méthode de détection de communautés en supprimant plusieurs liens simultanément dans chaque opération de filtration, et ont définit un coefficient récursif de communauté pour quantifier la qualité de division au lieu d'utiliser la modularité.

5.6.2. Mesure de communautés

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.

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

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, 0 pour les autres).

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

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

$$C_{ij} = A_{ij} - B_{ij}$$

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

Les auteurs ont proposé un coefficient récursif de communauté (CRC), dénoté M, afin de quantifier l'effet de division de réseau.

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

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

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

En cas où le réseau est indivisible par l'équation de filtration, les auteurs ont proposé une distribution du réseau en utilisant la matrice de Laplace.

5.6.3. Méthode de détection des communautés

Répéter

1) Construire le modèle aléatoire du réseau.

2) Diviser le réseau par l'opération de filtration 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 la filtration et aller à l'étape 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 le sous-réseau serait considéré en tant que communauté locale et arrêt de sa division.
Sinon considérer chaque sous-réseau en tant qu'une nouvelle communauté et aller à l'étape 1.

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

5.6.4. Discussion

La méthode récursive de filtration proposée par Shen et al [SPWL08] offre un gain en complexité de calcul $O(m^2+(c+1)m)$, pour un réseau de m liens et c communautés. La méthode peut détecter les communautés locales selon les densités de leurs liens externes dans l'ordre croissant en particulier dans les grands réseaux. En revanche, cette méthode devient lente et imprécise quand la densité des liens entre les communautés s'approche à la densité des liens au sein des communautés.

5.7. Local algorithm

5.7.1. Idée fondamentale

Bagrow et al [BB05a] 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.

5.7.2. Mesure de communauté

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 [BB05a] ont utilisé la notion de *shell* et ont définit deux mesures de liens : Shell : est définit en tant qu'un ensemble de sommets sur une distance géodésique depuis un sommet de départ (sommet origine). Le premier shell inclut les plus proches voisins du sommet d'origine et le deuxième inclut les prochains voisins des plus proches voisins du sommet d'origine 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 origine j. $K_i^e(j)$ 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 :

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

Avec S_j^l : est l'ensemble de tous les sommets exactement 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 :

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

5.7.3. Méthode de détection d'une seule communauté locale

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 est atteint ou bien le composant global connecté est ajouté à la liste de la communauté.

5.7.4. Méthode de détection de communautés

Pour récupérer une idée sur la structure globale du réseau, les auteurs 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.

5.7.5. Exemple d'expérimentation

Une illustration sur le fameux réseau de club de karaté (voir Fig. 10) montre que l'algorithme atteint le résultat souhaité quand $\alpha = 1.2$, ici trois noeuds ne sont pas correctement détectés (3,14,20) comme le montre la matrice d'adhésion (Fig. 13). Ces noeuds se situent à la frontière des deux groupes existants et sont presque identiquement relié aux deux communautés, comme le montre le dendrogramme (Fig. 14). Le dendrogramme définit par Bagrow et al [BB05a] explique bien la distance entre les sous communautés.

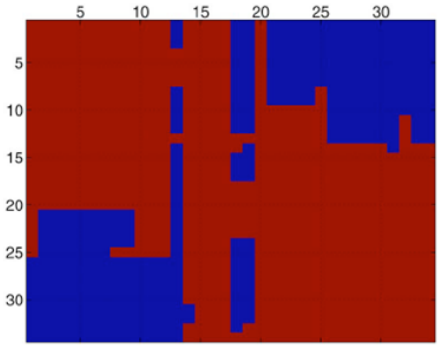

Figure 13. *Matrice d'adhésion du réseau de Zachary ($\alpha = 1.2$)*

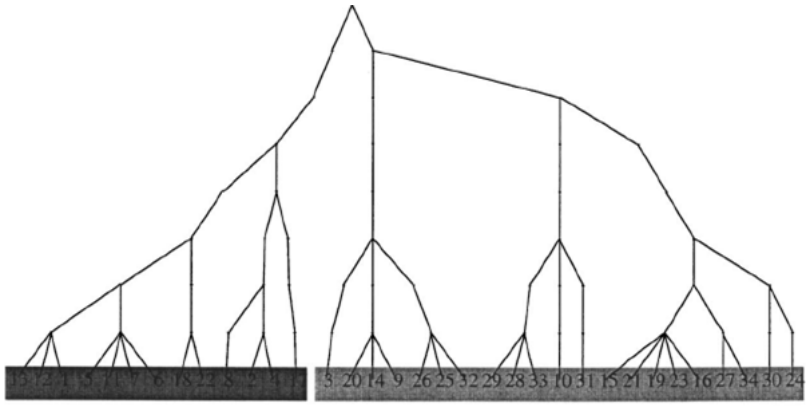

Figure 14. *Dendrogramme du réseau de Zachary utilisant la matrice d'adhésion ($\alpha = 1.2$).*

5.7.6. Discussion

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. De plus, l'efficacité de l'algorithme dépend du choix du seuil de changement ; une petite valeur de α s'étend les shells vers une seule partition principale alors qu'une grande valeur tronque les shells avant qu'ils aient eu une chance d'extraire les sous communautés des sommets de départs.

5.8. Community structure algorithm in directed networks

5.8.1. Idée fondamentale

La plupart des approches de découverte de communautés existantes dans la littérature ignorent l'orientation de lien et sont destinées à la découverte de communautés dans les réseaux non orientés tout en écartant l'information potentiellement utile contenue dans les orientations de liens. A cet effet, plusieurs auteurs ont étudié les réseaux orientés de données [NL07, GSPA07, ADFG07, RB08], mais le premier travail qui entame la problématique de découverte de communautés dans les réseaux orientés c'était celui proposé par Leicht et Newman [LN08] dans lequel la fonction de modularité a été généralisé afin d'incorporer l'information contenue dans l'orientation de lien.

5.8.2. Mesure de communauté

Vue que plusieurs réseaux complexes sont orientés, y compris le World Wide Web, food webs, beaucoup de réseaux biologiques et même quelques réseaux sociaux, Leicht et Newman [LN08] ont proposé une extension de la méthode d'optimisation spectrale de la modularité qui a été définit par Newman [New06] et ils l'ont adaptée aux réseaux complexes orientés.

Dans [New06], Newman a écrit la fonction de modularité sous sa forme matricielle comme suit :

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

(Voir section 5.5) Pour réécrire la fonction de modularité destinée aux réseaux orientés, Leicht et Newman [LN08] ont procédé comme suit :
Considérant deux sommet i et j. Le sommet i a un degré externe élevé et un faible degré interne, et inversement pour le noeud j. Donc, la probabilité qu'un lien partant d'un sommet i est orienté vers j est : $\frac{k_i^{in} k_j^{out}}{m}$. Cette proposition permet de définir la fonction de modularité comme suit :

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

La matrice de modularité est donnée par :

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

Pour la rendre symétrique, Q s'écrit :

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

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 :

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

Tel que $B^{(g)}$ est la matrice de modularité du sous graphe g.

5.8.3. Méthode de détection de communautés

L'algorithme se déroule comme suit :

1) Construire $(B + B^T)$ la matrice de modularité du graphe 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 mais 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 atteindre cet état l'algorithme s'arrête.

5.8.4. Exemple d'expérimentation

Leicht et Newman [LN08] 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

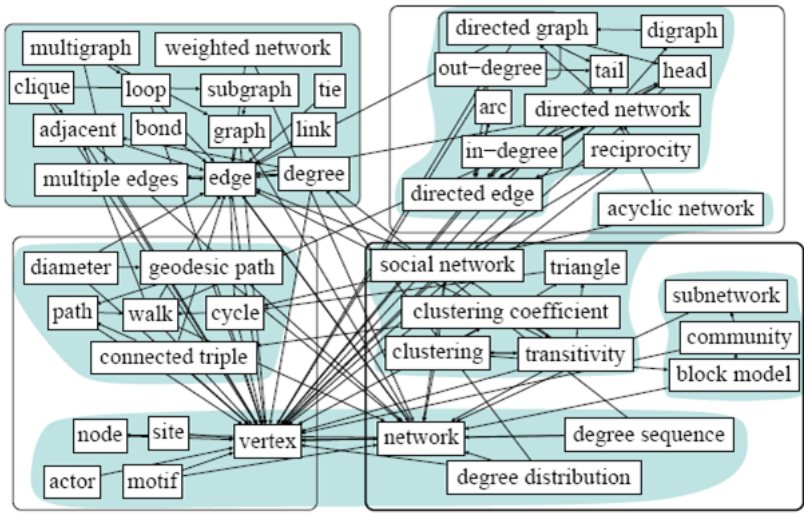

Figure 15. *Réseau des termes techniques illustre la découverte de communautés en appliquant l'algorithme de modularité orienté (groupes en bleu) et l'algorithme d'optimisation de modularité (groupes encadrés).*

dans un glossaire dérivé des papiers publiés récemment par Newman [New03b] et Boccaletti et al [BLMC06].

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. 15 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 système étudié dans les papiers [New03b] et [BLMC06].

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.

5.8.5. Discussion

Les méthodes de découverte de communautés destinée aux 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.

5.9. RCCLP method or Parisi method

5.9.1. Idée fondamentale

Radicchi et al [RCCLP04] 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 [RCCLP04] et d'Auber et al [ACJM03] basées sur le clustering d'arêtes. La détection des arêtes intercommunautaires est ici basée sur le fait que de telles arêtes sont dans des zones peu clustérisées. Radicchi et al [RCCLP04] proposent un coefficient de clustering (d'ordre g) d'arêtes. Il est défini comme étant le nombre de cycles de longueur g passant par l'arête divisé par le nombre total de tels cycles possibles (étant donné les degrés des extrémités de l'arête). Cet algorithme retire donc à chaque étape l'arête de plus faible clustering.

5.9.2. Discussion

Dans la méthode de Radicchi et al [RCCLP04], chaque suppression d'arête ne demande qu'une mise à jour locale des coefficients de clustering, ce qui lui permet d'être bien plus rapide que plusieurs algorithmes. En revanche, cet algorithme se fonde sur la présence des triangles dans le réseau ; quand un réseau a peu de triangles, le coefficient de clustering d'arête sera petit pour toutes les arêtes, et l'algorithme sera incapable à détecter les communautés.

5.10. Divisive algorithm of bipartite networks

5.10.1. Idée fondamentale

Les réseaux bipartis sont un type important de réseau complexe. 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, mais aussi les graphes d'occurrence des mots dans les phrases d'un livre, ou encore le graphe des auteurs de publications scientifiques.

Watts et Strogatz ont introduit en 1998 la notion formelle de coefficient de clustering [WS98]. Il s'agit de la moyenne, sur tous les noeuds 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). Zhang et al [ZWL08] ont modifie la définition de coefficient de clustering, ils l'ont adapté aux réseaux bipartis puis ils 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.

5.10.2. Mesure de communauté

Un réseau biparti est composé de deux ensemble de sommets distincts. Lind et al [LGH05] 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 [LP49] ne peut pas être utilisé. Au lieu de cela, ils ont définit un coefficient donné par la fraction de cycles à quatre dimensions.

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 [WS98]. Le coefficient de clustering $C_4(i)$, qui a été définit par Lind et al [LGH05], est la fraction entre le nombre de carrés existants

et le nombre total de tous les carrés possibles. Dans la langue des réseaux sociaux, C_4 représente la probabilité que vos amis ont des amis communs hormis vous.

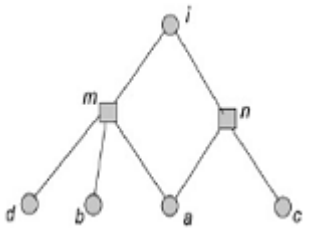

Figure 16. *Exemple pour démontrer le calcul des carrés.*

Pour le noeud i (Fig. 16), 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 :

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

Tel que :

m,n **:** Des voisins du noeud i ;

m,n **:** Le nombre de carrés qui incluent les trois noeuds ;

$\eta_{imn} = 1 + q_{imn}$.

Le dénominateur détermine le nombre de carrés possibles (les carrés existants et sous-jacents).

La définition de Lind considère d'éventuel coïncidence des noeuds, quand à la définition de Zhang et al [ZWL08] considère d'éventuels coïncidence des liens en se basant sur les normes du coefficient de clustering des réseaux binaires (aucun lien ne peut exister entre le noeud m et n, ni entre leurs voisins).

Radicchi et al [RCCLP04] ont proposé un algorithme séparatif de détection de communauté en introduisant un nouveau concept appelé coefficient de clustering d'arête. Zhang et al [ZWL08] 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 :

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

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 second 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 aux lesquels ce lien appartient, il s'écrit comme suit :

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

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.

5.10.3. Exemple d'expérimentation

Les figures ci-dessous (Fig. 17) montrent un réseau biparti qui contient 6 noeuds en haut et 6 noeuds en bas.

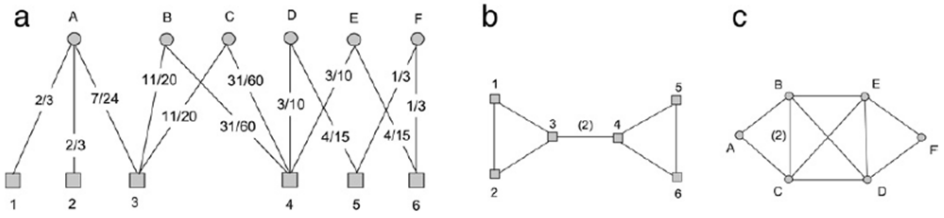

Figure 17. *Exemple d'un réseau biparti et sa projection.*

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 $\{F, 5, 6\}$.

5.10.4. Discussion

L'algorithme donne beaucoup plus de détails à propos de la structure de la communauté et les résultats sont plus conformes au réseau biparti original. L'application de l'algorithme sur un réseau biparti d'Econophysists [LWFD07] qui se compose de 818 auteurs et de 777 papiers a permis de détecter 20 communautés. La modularité de cette partition

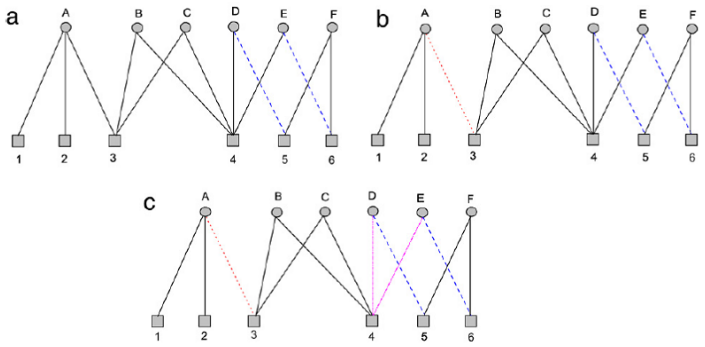

Figure 18. *L'identification des communautés par l'algorithme basé sur le coefficient.*

est : $M_B(p)_1 = 0.351$ (la fonction de modularité des réseaux bipartis qui a été définit dans [GSPA07]). Tandis que l'application d'extermal optimization algorithm [DA05] sur ce même réseau, après sa projection, résulte d'une modularité de $M_B(p)_1 = 0.351$. Donc, l'algorithme séparatif des réseaux bipartis produit une structure de communautés plus significatives que celle produite en appliquant les algorithmes existants destinés aux réseaux non bipartis.

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.

5.11. Clique Percolation Method

5.11.1. Idée fondamentale

Palla et al [PDFV05, DPV05] 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 se rapporte au 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. 19. 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.

Une extension de l'algorithme CPM a été proposée pour les réseaux pondérés [FbPV07] et les réseaux orientés [PFPD07].

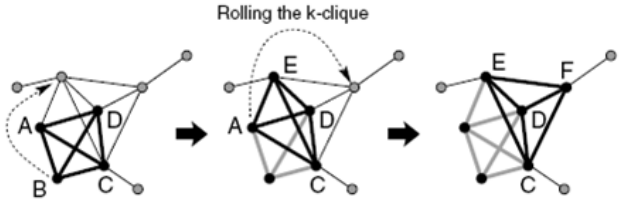

Figure 19. *Illustration de CPM [PDFV05, DPV05] un $k-clique$ template roulé sur un petit graphe non orienté ($k=4$).*

5.11.2. Discussion

L'algorithme CPM est efficace pour la détection des communautés qui se chevauchent existantes dans les réseaux (un noeud peut être un membre de plusieurs différentes communautés en même temps), et donc chaque communauté peut avoir un grand nombre de contacts avec d'autres communautés, juste comme se produit dans des situations réelles. Pour détecter les communautés à partir de $k-clique$, il faut tout d'abord calculer les cliques maximales dont la complexité de temps 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.

5.12. AgglomerativE hierarchicAl clusterinG based on maximaL clique "EAGLE" algorithm

5.12.1. Idée fondamentale

Dans [SCCH09], Shen et al ont présenté un algorithme pour détecter à la fois la hiérarchie des communautés ainsi que leurs chevauchements et ceci en se basant sur l'ensemble de cliques maximales tout en employant un processus agglomératif.

5.12.2. Mesure de communauté

La similarité entre une paire de communauté C_1 et C_2 est calculé comme suit :

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

Tel que :

A_{vw} **:** La matrice d'adjacence ;

k_v **:** Le degré du sommet v ;

m **=** Le nombre total de liens du réseau ($m = \frac{1}{2} \sum_{vw} A_{vw}$).

5.12.3. Méthode de détection de communautés

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. Dans les réseaux du monde réel, le seuil k prend typiquement une valeur entre 3 et 6. L'algorithme se déroule en deux étapes. Dans la première étape, l'algorithme résulte d'un dendrogramme qui représente toutes les division possible :

1) Trouver toutes les cliques maximales dans le réseau en utilisant l'algorithme de Bron-Kerbosch [BK73]. Ignorer les cliques maximales subordonnés et déclarer les autres cliques en tant que communautés initiales. Chaque sommet subordonné est également pris en tant que 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 n'y a plus qu'une seule communauté correspondant au graphe entier.

Shen et al ont définit une extension de modularité pour déterminer la qualité de division en communautés en tenant compte des sommets subordonnée.

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

Tel que :

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

Dans la deuxième étape, l'algorithme détermine la coupe appropriée du dendrogramme selon la valeur maximale de EQ.

5.12.4. Exemple d'expérimentation

Les figures ci-dessous comparent les divisions obtenues en appliquant fast algorithm de Nweman [New04], $k - clique$ de Palla et al [PDFV05] et EAGLE algorithm de Shen [SCCH09] sur un réseau de collaboration scientifique.

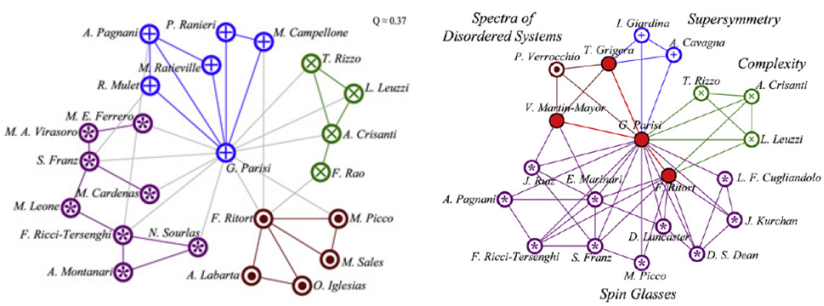

Figure 20. *Structure de communautés découvertes par fast algorithm (à droite) et l'algorithme k-clique. (Les noeuds et les liens qui appartiennent aux deux ou plusieurs communautés sont colorés en rouge)*

L'algorithme AIGLE et l'algorithme de Newman résultent d'un nombre de communautés presque identique à chaque niveau hiérarchique sauf que la taille de ces commu-

nautés 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.

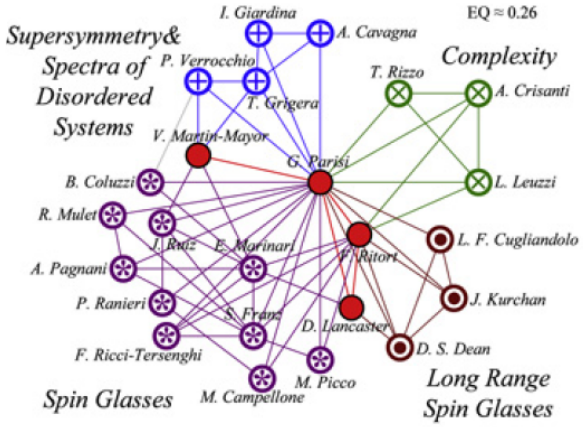

Figure 21. *Structure de communautés découvertes par EAGLE algorithm.*

5.12.5. Discussion

EAGLE algorithm n'as pas seulement la capacité de détecter les sous communautés jusqu'à aucune d'elles ne peut être divisé (graphe indivisible), mais aussi il détecte les communautés qui se chevauchent sans perte de leurs détails hiérarchiques. Cependant, le coût de calcul généré par l'algorithme lors de la recherche de cliques maximales est très élevé et les auteurs, en faite, visent son optimisation dans des futurs travaux.

5.13. Community detection algorithm using Random Walks

5.13.1. Idée fondamentale

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, ainsi plusieurs approches de détection de communautés se basent sur ces comportements.

Pons et al [PL06] 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).

5.13.2. Mesure de communauté

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 : $P_{ij} = \frac{A_{ij}}{d(i)}$, c'est la matrice de transition du processus de marche aléatoire.

P_{ij}^t : 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 al [PL06] ont basé alors sur le principe que deux marches qui partent d'une même communauté vont avoir des comportements similaires et se concentrer sur les mêmes zones du graphe. Pour comparer deux sommets i et j et définir une distance entre eux, les auteurs utilisent les remarques suivantes :

- Si deux sommets i et j sont dans la même communauté, la probabilité P^t_{ij} sera sûrement élevée, cependant une probabilité P^t_{ij} importante ne signifie pas forcément que i et j sont dans la même communauté.

- La probabilité P^t_{ij} est influencée par le degré $d(j)$ du sommet d'arrivée puisqu'il est plus facile d'atteindre les sommets de fort degré par une marche aléatoire.

- Deux sommets i et j d'une même communauté tendent à "voir" les autres sommets de la même manière. Quand i et j sont de la même communauté, ils vont avoir : $\forall k, P^t_{ik} \approx P^t_{jk}$.

Ces considérations permettent de définir une distance r_{ij} entre les sommets du graphe comme suit :

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

Tel que D est la matrice diagonale des degrés, r est une distance Euclidienne vue sur l'ensemble R^n. La généralisation de cette distance en une distance entre communautés est définit comme suit :

$$P^t_{C\bullet} = \frac{1}{|C|} \sum_{i\in C} P^t_{i\bullet}$$

Tel que : $C \subset V$ est une communauté de sommets, le vecteur de probabilité de position $P^t_{C\bullet}$ correspond à une marche de longueur t partant uniformément des sommets de la communauté C. Ainsi, la distance entre deux communautés C_1 et C_2 est donnée par :

$$r_{C_1C_2} = \|D^{-\frac{1}{2}} P^t_{C_1\bullet} - D^{-\frac{1}{2}} P^t_{C_2\bullet}\|$$

La distance r est directement reliée aux propriétés spectrales de la matrice de transition P. La distance r est alors relié aux approches spectrales qui sont basées sur le fait que deux sommets proches (d'une même communauté) ont des composantes similaires sur les vecteurs propres principaux.

5.13.3. Méthode de détection de communautés

Le problème de détection de communautés peut être réduit à détecter séparément des communautés dans chaque composante connexe. Pour cela, Pons et al supposent qu'un pré-calcul de composantes connexes est effectué, et appliqueront ensuite l'algorithme de détection de communautés séparément sur les sous graphes connexes. L'algorithme se déroule comme suit :

1) A partir d'une partition initiale en communauté $\wp^0 = \{\{v\} \in V\}$, formée des n communautés composées d'un seul sommet ;

2) Calcul initial des distances r_{ij} entre chaque paire de sommets adjacents, ceci correspond donc à m distances calculées.

3) Choisir deux communautés C_1 et C_2 de $\wp^k$. Ce choix des communautés à fusionner repose sur les distances r entre les communautés adjacentes de la partition courante en choisissant à chaque étape k la paire de communautés produisant la plus faible variation $\Delta\sigma = \sigma_{k+1} - \sigma_k$. Tel que σ_k est la moyenne des distances au carré de chaque sommet à sa communauté :

$$\sigma_k = \frac{1}{n} \sum_{C \in P^k} \sum_{i \in C} r_{iC}^2$$

4) Fusionner ces deux communautés en une nouvelle communauté $C_3 = C1 \cup C2$ et créer la nouvelle partition : $\wp^{k+1} = \{\wp^{k+1} | \{C_1 \cup C_2\} = \{C_3\}\}$

5) Mettre à jour les distances r qui ont été changé, c'est à dire les distances r_{C_3C} pour toutes les communautés C adjacentes de la nouvelle communauté C_3.

Pour déterminer la meilleure partition en communautés sortie par l'algorithme, il suffit de déterminer la partition qui maximise la fonction de qualité choisie à partir du dendrogramme.

5.13.4. Discussion

L'approche proposée par [PL06] possède le grand avantage de tirer partie de propriétés spectrales sans avoir recours à un calcul explicite de valeurs et vecteurs propres. Elle permet de tirer profit de ces propriétés spectrales tout en gardant une complexité raisonnable $O(nmlog(n))$ grâce aux calculs de marches aléatoires. Cette approche peut traiter des graphes allant jusqu'à un million de sommets. 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. De même les très courtes marches (de longueur $t < 3$) ne donnent pas les meilleurs résultats car elles n'ont pas le temps de recueillir suffisamment d'information autour de chaque sommet. Un bon choix de la longueur des marches fournis meilleure qualité des résultats.

5.14. Community identification algorithms using Brownian motion

5.14.1. Idée fondamentale

Les méthodes proposées par Zhou [Zho03b, Zho03a] et Zhou et Lipowsky [ZL04] 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.

5.14.2. Mesure de communauté

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 commu-

nauté. Soit un réseau connecté de N noeuds et M liens, A est sa matrice d'adjacence tel que :

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

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 sa position actuelle (i) vers la position du plus proche voisins (j). La matrice de transfert est donné par :

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

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 :

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

Tel que :

I **:** La matrice d'identité ;

$B(j)$ **:** La matrice de transfert P avec $B_{lj}(j) = 0$ pour tous $l \in V$.

Prenant un sommet i comme sommet origine du réseau, l'ensemble $\{d_{i1}, ..., d_{i,i-1}, d_{i,i+1}, ..., d_{iN}\}$ mesures à quelle distance tous les autres sommets sont situés de l'origine. Par conséquent, la perspective du réseau entier est distingué à partir du sommet i. Supposons que les sommets i et j sont des plus proches voisins, la différence dans leurs perspectives du réseau peut être quantitativement mesuré. Zhou [Zho03a] a définit un indexe de dissimilitude comme suit :

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

Si deux plus proche voisins i et j appartiennent à la même communauté, la distance moyenne d_{ik} de i à n'importe quel autre sommet $k(k \neq i, j)$, va être 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 différentes communautés.

5.14.3. Méthode de détection de communautés

Dans [Zho03b], 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 en tenant compte aux 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.

De même les auteurs ont définit les communautés basées sur un attracteur global, tel que chaque noeud a une forte probabilité d'être dans la même communauté G de

son attracteur global. Pour de grands réseaux, chaque communauté G peut contenir plusieurs L communautés en tant que ses sous-groupes. Zhou a identifié également le centre d'une communauté (s'il existe) en tant que le noeud attracteur global de lui-même. Dans [Zho03a], 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 et prend 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 est mis dans l'ensemble des amis avec qui il a une forte interaction. Après cette opération, les sommets de la communauté 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 [ZL04] 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.

5.14.4. Discussion

Les algorithmes proposés par Zhou [Zho03b, Zho03a] et Zhou et Lipowsky [ZL04] 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.

5.15. Community identification algorithms using Spectral analysis

5.15.1. Idée fondamentale

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.

Comme les méthodes classiques de partitionnement de graphe [KL70, PSL90, Fie73] requièrent une connaissance préalable du nombre de communautés recherchées ainsi que de leurs tailles, elles ne conviennent pas totalement à la détection de communautés, plusieurs approches récentes ont été alors proposée afin de partitionner le graphe et découvrir sa structure de communauté.

Donetti et al [DM04] 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 [DM05].

Dans [JDY09], 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.

5.15.2. Discussion

Les méthodes spectrales ont montré expérimentalement de meilleurs résultats mais elles sont pénalisées en terme de complexité car la détermination des valeurs et vecteurs propres d'une matrice creuse nécessite un temps de calcul en $O(n^3)$. Cette complexité peut rapidement devenir inabordable dès lors que la taille du graphe dépasse quelques milliers de sommets.

5.15.2.1. Remarque :

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, d'en voici une liste non exhaustive que nous avons consulté afin de proposer la classification adéquate mais nous n'avons pas citée ces méthodes dans ce rapport de synthèse ([FLM04, GSPA04, LSH08, CNM04, BILPR07, NL07, EM02, SPGMA07, ZZZ08]).

6. Etude comparative des algorithmes de découverte de communautés

Pour évaluer la qualité d'une partition trouvée par un algorithme et comparer les partitions trouvées aux partitions de références, la similarité entre la partition trouvée et la partition de référence doit être mesurée (taux des noeuds identifies correctement). Cette mesure permettre aussi de comparer les performances en qualité de partition des différents algorithmes.

Danon et al [**?**] ont utilisé une mesure empruntée de la théorie de l'information : "normalized mutual information" , la similarité de deux partitions A et B est donnée par l'expression suivante :

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

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 :

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

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.

Plusieurs mesures ont été définit dans la littérature, le papier [GHL06] présente les différents critères pour comparer les partitions détectées.

Tandis 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 ci-après récapitule le temps d'exécution des déférentes méthodes. Par exemple, l'algorithme de référence dans le domaine GN génère un coût de calcul très élevé ce qui empêche son application sur de grands réseaux. Le temps d'exécution de l'algorithme "eigenvector-based algorithm" est relative avec la taille du système, il est de $O(n^2 log(n))$ pour un cas typique d'un graphe dense, cela est considérablement mieux que $O(n^3)$ de GN "betweenness-based algorithm" et partiellement mieux que $O(n^2 log^2(n))$ de l'algorithme d'optimisation extrémale mais il n'est pas assez mieux que $O(nlog^2(n))$ de greedy algorithm. GN ne permet pas de traiter des graphes de plus de 1000 sommets alors que fast algorithm convient pour de grands réseaux.

En conclusion, comparer les différentes techniques est difficile car il est possible que les méthodes de meilleures temps d'exécution ne peuvent pas identifient les partitions en communautés les plus pertinentes. La recherche d'un compromis entre la qualité des résultats et le coût de calcul reste toujours le souci des auteurs. La difficulté réside dans le fait de minimiser les coûts (en temps et en espace) du calcul tout en maximisant la qualité des partitions en communautés trouvées. Suivant les contextes, différents compromis sont acceptables, ce qui justifie en partie la prolifération de ces méthodes.

7. Classification des méthodes de découverte de communautés dans les réseaux

Selon l'étude synthétique que nous avons faite, ils existent, dans la littérature, une panoplie des méthodes de détection de communautés dans les réseaux complexes mais aucune taxonomie de ces méthodes n'a été proposée auparavant. Dans [FC08] et [DDGDA05],

Tableau 1. *Récapitulatif de complexité en temps des différentes méthodes.*

Algorithme	**Référence**	**Complexité en temps**
Random-walk algorithm	Newman et al [NG04]	$O(n^3)$
Betweenness-based algorithm	Girvan et al [GN02] Newman et al [NG04]	$O(m^2n)$
Extremal optimization algorithm	Duch et al [DA05]	$O(n^2log^2(n))$
Fast algorithm	Newman [New04]	$O(nlog^2(n))$
Simulated annealing based algorithm	Guimerà et al [GSPA04]	Inconnue
Q-state Potts model based algorithm	Reichardt et al [RB04, RB06]	Dépend des paramètres
Local algorithm	Bagrow et al [BB05b]	$O(n^3)$
RCLP algorithm	Radicchi et al [RCCLP04]	$O(n^2)$
Eigenvector-based algorithm	Newman [New06]	$O(n^2log(n))$
Greedy algorithm	Clauset et al [CNM04]	$O(d.mlog(n))$
Directed modularity maximization algorithm	Leicht et Newman [LN08]	$O(n^2log())$
Divisive algorithm of bipartite networks	Zhang et al [ZWL08]	Inconnue
Recursive filtration algorithm	Shen et al [SPWL08]	$O(m^2 + (c+1)m)$
Biclique algorithm	Lehmann et al [LSH08]	$O(n^2)$
k-clique "CPM"	Palla et al [PDFV05]	Inconnue
Markov cluster algorithm	Van Dongen [VD00]	$O(n.k^2)$
EAGLE algorithm	Shen et al [SCCH09]	$O(n^2.s)$

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 [VD00] ;
d : La profondeur du dendrogramme ;
s : Le nombre de cliques maximale.

les auteurs ont donné une synthèse de quelques méthodes de découverte de communautés en les mentionnant selon le groupe auquel il appartiennent, toutefois ces classifications sont imprécises, incomplètes (absence des critères permettant de bien différentier ces méthodes) et contribuent très peu au choix de tel ou tel méthode de découverte de communautés.

Dans cette partie et d'après nos constats et nos analyses, nous venons de proposer une classification des méthodes de découverte de communautés dans l'objectif de positionner ces différentes approches les unes par rapport aux autres. Il semble que cette taxonomie 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 donné et pourquoi, ce qui apporte une flexibilité considérable en terme temps d'exécution et qualité de partition. Notre classification repose sur les trois points de vue suivants que nous avons adoptés :

1- Selon la manière de regroupement des noeuds en groupes, Jain et Dubes [JD88] 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 (même si la plupart des auteurs s'intéressent dans leurs approches à la technique utiliser pour détecter les communautés mais un simple constat peut clarifier l'approche employée).

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 temps d'exécution et qualité de partition. A cet effet, nous croyons que c'est fort important de distinguer les méthodes déterministes de celles stochastiques.

3- Dans la plupart des approches examinées jusqu'ici, les communautés ont été caractérisées et découverte, directement ou indirectement, par une certaine propriété globale du graphe, comme intermédiarité, centralité, etc., ou par un certain processus comme les promenades aléatoires, la synchronisation, etc... Mais les communautés peuvent être également interprété en tant qu'une forme d'organisation topologique du graphe. Toutes ces remarques nous ont permet de classifier les méthodes existantes selon la technique ou bien le processus utilisé au cours de la découverte de communautés.

7.1. Méthodes agglomératives

Dans les méthodes agglomératives [Zho03b, PL06, ZL04, GSPA04, RB04, RB06, ADGPV06, BILPR07, New04, CNM04, SCCH09, PDFV05, PFPD07, FbPV07, LSH08, DM04, DM05, JDY09, EM02, BB05b, ZZZ08], les métriques de similarité entre les pairs de sommets sont calculées au moyen de plusieurs méthodes, et par conséquent les liens sont ajoutés itérativement au réseau initial non connecté partant du lien qui relie les pairs de sommets de forte similarité. Ce processus d'ajout de liens peut être arrêté à n'importe quel point et les composants obtenus représentent les communautés.

Cependant ces méthodes ont tendance de trouver les noyaux de communautés et n'inclure pas les périphéries. Les noeuds de noyau dans une communauté ont souvent forte similarité, et par conséquent sont reliés tôt dans le processus agglomératif, mais les noeuds périphériques qui n'ont aucune forte similarité aux d'autres noeuds tendent à être négligés (ils ne sont pas mis dans la communauté appropriée)[NG04].

7.2. Méthodes séparatives

Les méthodes séparatives [Zho03b, VD00, DA05, NL07, New06, LN08, ZWL08, RCCLP04, SPWL08, GN02, NG04, FLM04, SPGMA07] scindent le graphe en plusieurs communautés en retirant progressivement les arêtes reliant deux communautés distinctes. Ces méthodes essayent de trouver les pairs de noeuds qui sont reliés par liens de faible similarité et enlèvent alors au fur et à mesure ces liens. Ainsi, le réseau est divisé en plusieurs composants représentants les communautés. Ce processus de suppression de liens peut être arrêté à n'importe quelle étape et les composants obtenus sont considérés comme des communautés.

Chaque méthode est destinée aux graphes complexes qui possèdent des propriétés bien précises : il s'agit des méthodes applicables sur des réseaux orientés, non orientés, pondérés, non pondérés, ou même les réseaux possédant des noeuds qui se chevauchent. Nous pouvons ainsi donner un tableau de synthèse qui clarifie le contexte d'exécution des méthodes de découverte de communautés comme suit :

Tableau 2. *Contexte d'application des différentes méthodes de découverte de communautés.*

	non orienté	orienté	pondéré	non pondéré	Overlapped communities
Bagrow et al [BB04]	V			V	
Newman [N06]	V			V	
Reichardt et al [RB04, RB06]	V	V		V	V
Newman et al [NL07]	V	V			
Shen [SCCH09]	V				V
Palla et al [PDFV05]	V				V
Palla et al [PFPD07]		V		V	V
Farkas et al [FAPV07]	V		V		V

8. Conclusion et perspectives de recherche

L'étude des réseaux complexes est une activité en plein essor. Notre objectif est de proposer une taxonomie complète des méthodes de découverte de communautés dans les réseaux complexes. Pour ce faire, nous avons présenté, à travers une étude synthétique, les principales méthodes de découverte de communautés, en discutant les apports et les inconvénients de chacune d'elles et proposant une classification des différentes approches.

Notons que de nouvelles directions se développent, par exemple vers l'étude des réseaux pondérés, des réseaux dynamiques, l'introduction de nouvelles fonctions de qualité et la recherche d'une nouvelle formalisation de la notion de communauté différente de la formalisation apportée par les fonctions de qualité.

9. Bibliographie

[AB02] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1) :47, 2002.

[ACJM03] David Auber, Yves Chiricota, Fabien Jourdan, Guy Melançon, and others. Multiscale Visualization of Small World Networks. In *InfoVis03*, volume 3, pages 75–81. IEEE Computer Society, 2003.

[ADFG07] Alex Arenas, Jordi Duch, Alberto Fernández, and Sergio Gómez. Size reduction of complex networks preserving modularity. *New Journal of Physics*, 9(6) :176, 2007.

[ADGPV06] Alex Arenas, Albert Díaz-Guilera, and Conrad J. Pérez-Vicente. Synchronization reveals topological scales in complex networks. *Physical review letters*, 96(11) :114102, 2006.

[AJAL99] Réka Albert, Hawoong Jeong, and Barabási Albert-László. Internet : Diameter of the world-wide web. *Nature*, 401(6749) :130–131, 1999.

[AJAL00] Réka Albert, Hawoong Jeong, and Barabási Albert-László. Attack and error tolerance of complex networks. *Nature*, 406(6794) :378–382, 2000.

[AL02] Barbási Albert-Laszlo. *Linked : the new science of networks*. Perseus, 2002.

[AMO94] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. Network flows : Theory, algorithms, and applications. *Journal of the Operational Research Society*, 45(11) :1340–1340, 1994.

[ASBS00] Luis A. Nunes Amaral, Antonio Scala, Marc Barthelemy, and H. Eugene Stanley. Classes of small-world networks. *Proceedings of the national academy of sciences*, 97(21) :11149–11152, 2000.

[BA99] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439) :509–512, 1999.

[BB05a] James P. Bagrow and Erik M. Bollt. Local method for detecting communities. *Physical Review E*, 72(4) :046108, 2005.

[BB05b] James P. Bagrow and Erik M. Bollt. Local method for detecting communities. *Physical Review E*, 72(4) :046108, 2005.

[BILPR07] S. Boccaletti, M. Ivanchenko, V. Latora, A. Pluchino, and A. Rapisarda. Detecting complex network modularity by dynamical clustering. *Physical Review E*, 75(4) :045102, 2007.

[BK73] Coenraad Bron and Joep Kerbosch. Finding all cliques of an undirected graph (algorithm 457). *Commun. ACM*, 16(9) :575–576, 1973.

[BLMC06] Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez, and D.-U. Hwang. Complex networks : Structure and dynamics. *Physics reports*, 424(4) :175–308, 2006.

[Bol98] Béla Bollobás. Random graphs. In *Modern Graph Theory*, pages 215–252. Springer, 1998.

[BP01] Sven Bilke and Carsten Peterson. Topological properties of citation and metabolic networks. *Physical Review E*, 64(3) :036106, 2001.

[BS02] Vincent D. Blondel and Pierre P. Senellart. Automatic extraction of synonyms in a dictionary. In *the SIAM Workshop on Text Mining*, volume 1. Vertex, 2002.

[Buc03] Mark Buchanan. *Nexus : small worlds and the groundbreaking theory of networks*. WW Norton & Company, 2003.

[Bur76] Ronald S. Burt. Positions in networks. *Social forces*, 55(1) :93–122, 1976.

[BWD96] Marcelo Blatt, Shai Wiseman, and Eytan Domany. Super paramagnetic clustering of data. *Physical review letters*, 76(18) :3251, 1996.

[CLRS01] Thomas H. Cormen, Charles Eric Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to algorithms*, volume 6. MIT press Cambridge, 2001.

[CNM04] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6) :066111, 2004.

[DA05] Jordi Duch and Alex Arenas. Community detection in complex networks using extremal optimization. *Physical review E*, 72(2) :027104, 2005.

[DDGDA05] Leon Danon, Albert Diaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *Journal of Statistical Mechanics : Theory and Experiment*, 2005(09) :09008, 2005.

[DM02] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of Networks : from Biological Nets to the Internet and WWW. 2003. *Oxford University Press*, 2002.

[DM04] Luca Donetti and Miguel A Muñoz. Detecting network communities : a new systematic and efficient algorithm. *Journal of Statistical Mechanics : Theory and Experiment*, 2004(10) :10012, 2004.

[DM05] Luca Donetti and Miguel A. Muñoz. Improved spectral algorithm for the detection of network communities. In *Modeling Cooperative Behavior in the Social Sciences*, volume 779, pages 104–107, 2005.

[DPV05] Imre Derényi, Gergely Palla, and Tamás Vicsek. Clique percolation in random networks. *Physical review letters*, 94(16) :160–202, 2005.

[DR00] Peter Sheridan Dodds and Daniel H. Rothman. Geometry of river networks. *Physical Review E*, 63(1) :016115–016117, 2000.

[EM02] Jean-Pierre Eckmann and Elisha Moses. Curvature of co-links uncovers hidden thematic layers in the world wide web. *Proceedings of the national academy of sciences*, 99(9) :5825–5829, 2002.

[EMB02] Holger Ebel, Lutz-Ingo Mielsch, and Stefan Bornholdt. Scale-free topology of e-mail networks. *Physical review E*, 66(3) :035103, 2002.

[ER59] Paul Erdös and Alfréd Rényi. On random graphs. *Publicationes Mathematicae (Debrecen)*, 6 :290–297, 1959.

[ER90] Leo Egghe and Ronald Rousseau. Introduction to informetrics : Quantitative methods in library, documentation and information science. *Elsevier*, 1990.

[ESMS03] Kasper Astrup Eriksen, Ingve Simonsen, Sergei Maslov, and Kim Sneppen. Modularity and extreme edges of the Internet. *Physical review letters*, 90(14) :148701, 2003.

[FA86] Yaotian Fu and Philip W. Anderson. Application of statistical mechanics to NP-complete problems in combinatorial optimisation. *Journal of Physics A : Mathematical and General*, 19(9) :1605, 1986.

[FB07] Santo Fortunato and Marc Barthelemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1) :36–41, 2007.

[FbPV07] Illés Farkas, Dániel Ábel, Gergely Palla, and Tamás Vicsek. Weighted network modules. *New Journal of Physics*, 9(6) :180, 2007.

[FC08] S. Fortunato and C. Castellano. *Community structure in graphs. Encyclopedia of Complexity and System Science*. Springer, 2008.

[FFF99] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the internet topology. In *ACM SIGCOMM computer communication review*, volume 29, pages 251–262. ACM, 1999.

[Fie73] Miroslav Fiedler. Algebraic connectivity of graphs. *Czechoslovak mathematical journal*, 23(2) :298–305, 1973.

[FLGC02] Gary William Flake, Steve Lawrence, C. Lee Giles, and Frans M. Coetzee. Self-organization and identification of web communities. *IEEE Computer*, 35(3) :66–70, 2002.

[FLM04] Santo Fortunato, Vito Latora, and Massimo Marchiori. Method to find community structures based on information centrality. *Physical review E*, 70(5) :056104, 2004.

[Fre77] Linton C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

[GA05] Roger Guimera and Luis A. Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028) :895–900, 2005.

[GLLB04] Jean-Loup Guillaume, Matthieu Latapy, and Stevens Le-Blond. Statistical analysis of a p2p query graph based on degrees and their time-evolution. In *International Workshop on Distributed Computing*, pages 126–137. Springer, 2004.

[GN02] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12) :7821–7826, 2002.

[Gra73] Mark S. Granovetter. The strength of weak ties. *American journal of sociology*, pages 1360–1380, 1973.

[GSPA04] Roger Guimera, Marta Sales-Pardo, and Luís A. Nunes Amaral. Modularity from fluctuations in random graphs and complex networks. *Physical Review E*, 70(2) :025101, 2004.

[GSPA07] Roger Guimerà, Marta Sales-Pardo, and Luís A. Nunes Amaral. Module identification in bipartite and directed networks. *Physical Review E*, 76(3) :036102, 2007.

[HHJ03] Petter Holme, Mikael Huss, and Hawoong Jeong. Subnetwork hierarchies of biochemical pathways. *Bioinformatics*, 19(4) :532–538, 2003.

[iCS01] Ramon Ferrer i Cancho and Richard V. Solé. The small world of human language. *Proceedings of the Royal Society of London B : Biological Sciences*, 268(1482) :2261–2265, 2001.

[JD88] Anil K. Jain and Richard C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[JDY09] Jeffrey Q. Jiang, Andreas WM Dress, and Genke Yang. A spectral clustering-based framework for detecting community structures in complex networks. *Applied Mathematics Letters*, 22(9) :1479–1482, 2009.

[JTAO00] Hawoong Jeong, Bálint Tombor, Réka Albert, Zoltan N. Oltvai, and Barabási Albert-László. The large-scale organization of metabolic networks. *Nature*, 407(6804) :651–654, 2000.

[Kai99] Jocelyn Kaiser. It's a small Web after all. *Science*, 285(1815), 1999.

[KF69] PW Kasteleyn and CM Fortuin. Physica (utrecht) 57, 536 (1972); pw kasteleyn and cm fortuin. *Physical Society of Japan Journal Supplement*, 26(11), 1969.

[KFMU03] Ann E. Krause, Kenneth A. Frank, Doran M. Mason, Robert E. Ulanowicz, and William W. Taylor. Compartments revealed in food-web structure. *Nature*, 426(6964) :282–285, 2003.

[KL70] Brian W. Kernighan and Shen Lin. An efficient heuristic procedure for partitioning graphs. *Bell system technical journal*, 49(2) :291–307, 1970.

[Kle00] Jon Kleinberg. The small-world phenomenon : An algorithmic perspective. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 163–170. ACM, 2000.

[KSM03] V. K. Kalapala, V. Sanwalani, and C. Moore. The structure of the United States road network. *Preprint, University of New Mexico*, 2003.

[LGH05] Pedro G. Lind, Marta C. González, and Hans J. Herrmann. Cycles and clustering in bipartite networks. *Physical review E*, 72(5) :056127, 2005.

[LM02] Vito Latora and Massimo Marchiori. Is the Boston subway a small-world network ? *Physica A : Statistical Mechanics and its Applications*, 314(1) :109–113, 2002.

[LN04] David Lusseau and Mark EJ Newman. Identifying the role that animals play in their social networks. *Proceedings of the Royal Society of London B : Biological Sciences*, 271(Suppl 6) :S477–S481, 2004.

[LN08] Elizabeth A. Leicht and Mark EJ Newman. Community structure in directed networks. *Physical review letters*, 100(11) :118703, 2008.

[LP49] R. Duncan Luce and Albert D. Perry. A method of matrix analysis of group structure. *Psychometrika*, 14(2) :95–116, 1949.

[LSH08] Sune Lehmann, Martin Schwartz, and Lars Kai Hansen. Biclique communities. *Physical Review E*, 78(1) :016108, 2008.

[LWFD07] Menghui Li, Jinshan Wu, Ying Fan, and Zengru Di. Econophysicists Collaboration Networks : Empirical Studies and Evolutionary Model. In *Econophysics of Markets and Business Networks*, pages 173–182. Springer, 2007.

[Mar91] Neo D. Martinez. Artifacts or attributes ? Effects of resolution on the Little Rock Lake food web. *Ecological Monographs*, 61(4) :367–392, 1991.

[Mil67] Stanley Milgram. The small world problem. *Psychology today*, 2(1) :60–67, 1967.

[MKNG06] Alireza Mahdian, Hamid Khalili, Ehsan Nourbakhsh, and Mohammad Ghodsi. Web graph compression by edge elimination. In *Data Compression Conference (DCC'06)*, pages 1–pp. IEEE, 2006.

[Moo01] James Moody. Race, school integration, and friendship segregation in america. *American journal of Sociology*, 107(3) :679–716, 2001.

[MS02] Sergei Maslov and Kim Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569) :910–913, 2002.

[MSZ04] Sergei Maslov, Kim Sneppen, and Alexei Zaliznyak. Pattern detection in complex networks : Correlation profile of the Internet. *Physica A*, 333 :529–540, 2004.

[New01] Mark EJ Newman. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Physical review E*, 64(1) :016132, 2001.

[New03a] Mark EJ Newman. Mixing patterns in networks. *Physical Review E*, 67(2) :026126, 2003.

[New03b] Mark EJ Newman. The structure and function of complex networks. *SIAM review*, 45(2) :167–256, 2003.

[New04] Mark EJ Newman. Fast algorithm for detecting community structure in networks. *Physical review E*, 69(6) :066133, 2004.

[New06] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23) :8577–8582, 2006.

[NG04] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2) :026113, 2004.

[NL07] Mark EJ Newman and Elizabeth A. Leicht. Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences*, 104(23) :9564–9569, 2007.

[PDFV05] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043) :814–818, 2005.

[PFPD07] Gergely Palla, Illes J. Farkas, Peter Pollner, Imre Derenyi, and Tamás Vicsek. Directed network modules. *New journal of physics*, 9(6) :186, 2007.

[Pim79] Stuart L. Pimm. The structure of food webs. *Theoretical population biology*, 16(2) :144–158, 1979.

[PL06] Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. *J. Graph Algorithms Appl.*, 10(2) :191–218, 2006.

[PMV87] Giorgi Parisi, M. Mézard, and M. A. Virasoro. Spin glass theory and beyond. *World Scientific, Singapore*, 187 :202, 1987.

[PSL90] Alex Pothen, Horst D. Simon, and Kang-Pu Liou. Partitioning sparse matrices with eigenvectors of graphs. *SIAM journal on matrix analysis and applications*, 11(3) :430–452, 1990.

[RB04] Jörg Reichardt and Stefan Bornholdt. Detecting fuzzy community structures in complex networks with a Potts model. *Physical Review Letters*, 93(21) :218701, 2004.

[RB06] Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74(1), 2006.

[RB08] Martin Rosvall and Carl T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4) :1118–1123, 2008.

[RCCLP04] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9) :2658–2663, 2004.

[Res00] Mauricio GC Resende. Detecting dense subgraphs in massive graphs. In *17th International Symposium on Mathematical Programming*, 2000.

[SCCH09] Huawei Shen, Xueqi Cheng, Kai Cai, and Mao-Bin Hu. Detect overlapping and hierarchical community structure in networks. *Physica A : Statistical Mechanics and its Applications*, 388(8) :1706–1712, 2009.

[Sco12] John Scott. *Social network analysis : A Handbook*. SAGE publications, 2012.

[SDCS03] Parongama Sen, Subinay Dasgupta, Arnab Chatterjee, P. A. Sreeram, G. Mukherjee, and S. S. Manna. Small-world properties of the Indian railway network. *Physical Review E*, 67(3) :036106, 2003.

[SPGMA07] Marta Sales-Pardo, Roger Guimera, André A. Moreira, and Luís A. Nunes Amaral. Extracting the hierarchical organization of complex systems. *Proceedings of the National Academy of Sciences*, 104(39) :15224–15229, 2007.

[SPWL08] Yi Shen, Wenjiang Pei, Kai Wang, Tao Li, and Shaoping Wang. Recursive filtration method for detecting community structure in networks. *Physica A : Statistical Mechanics and its Applications*, 387(26) :6663–6670, 2008.

[Str01] Steven H. Strogatz. Exploring complex networks. *Nature*, 410(6825) :268–276, 2001.

[VD00] S. Van Dongen. *Graph clustering by flow simulation.* phd thesis, University of Utrecht, The Netherlands, 2000.

[Wat99] D. J. Watts. Small Worlds. *Princeton University Press*, 1999.

[Wat04] Duncan J. Watts. *Six degrees : The science of a connected age*. WW Norton & Company, 2004.

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

[WM00] Richard J. Williams and Neo D. Martinez. Simple rules yield complex food webs. *Nature*, 404(6774) :180–183, 2000.

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

[Zac77] Wayne W. Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, pages 452–473, 1977.

[Zho03a] Haijun Zhou. Distance, dissimilarity index, and network community structure. *Physical review e*, 67(6) :061901, 2003.

[Zho03b] Haijun Zhou. Network landscape from a Brownian particle's perspective. *Physical Review E*, 67(4) :041908, 2003.

[ZL04] Haijun Zhou and Reinhard Lipowsky. Network brownian motion : A new method to measure vertex-vertex proximity and to identify communities and subcommunities. In *International conference on computational science*, pages 1062–1069. Springer, 2004.

[ZWL08] Peng Zhang, Jinliang Wang, Xiaojia Li, Menghui Li, Zengru Di, and Ying Fan. Clustering coefficient and community structure of bipartite networks. *Physica A : Statistical Mechanics and its Applications*, 387(27) :6869–6875, 2008.

[ZZZ08] Junhua Zhang, Shihua Zhang, and Xiang-Sun Zhang. Detecting community structure in complex networks based on a measure of information discrepancy. *Physica A : Statistical Mechanics and its Applications*, 387(7) :1675–1682, 2008.

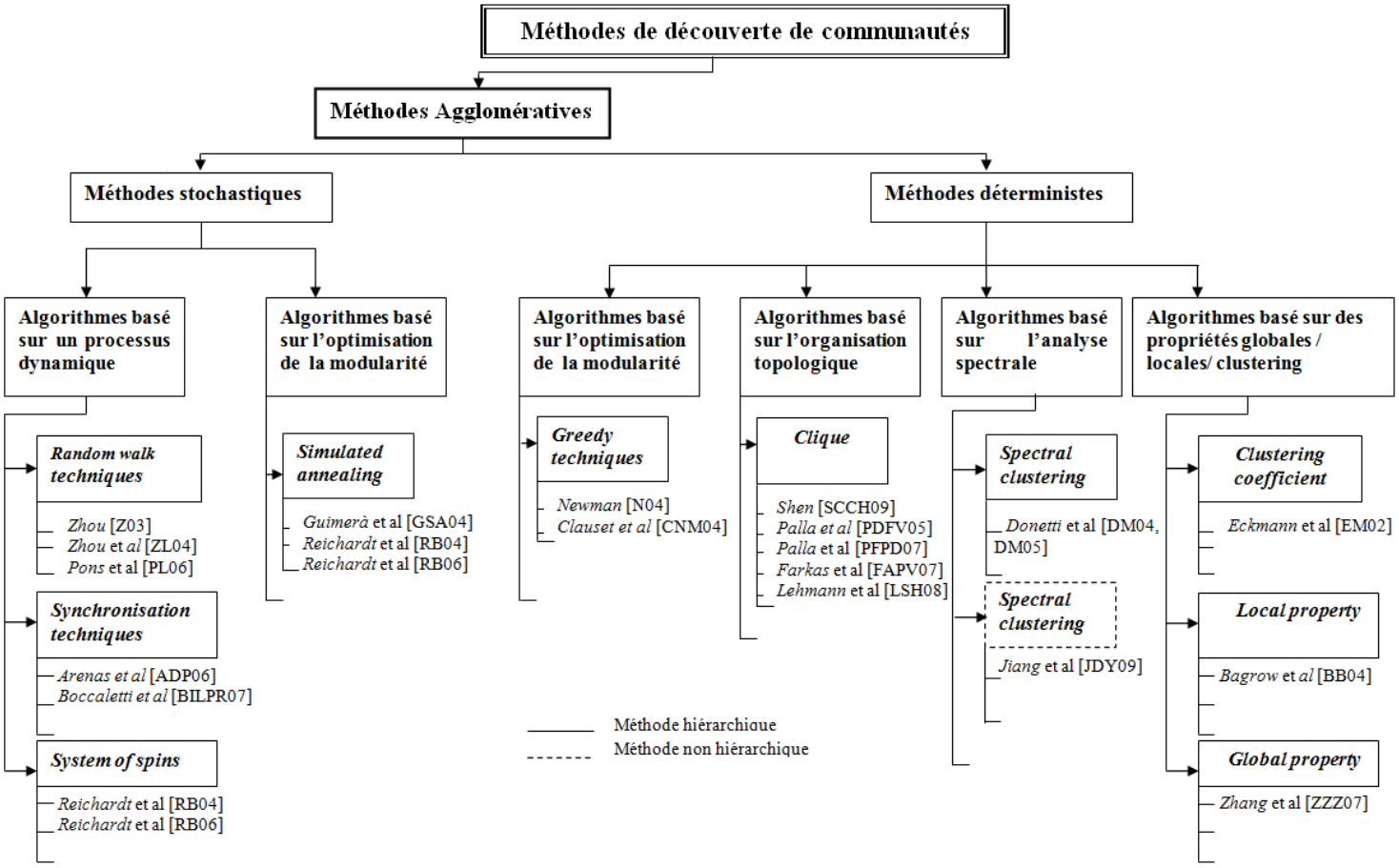

Figure 22. *Taxonomie des méthodes de découverte de communautés dans les réseaux complexes : 1. Méthodes agglomératives.*

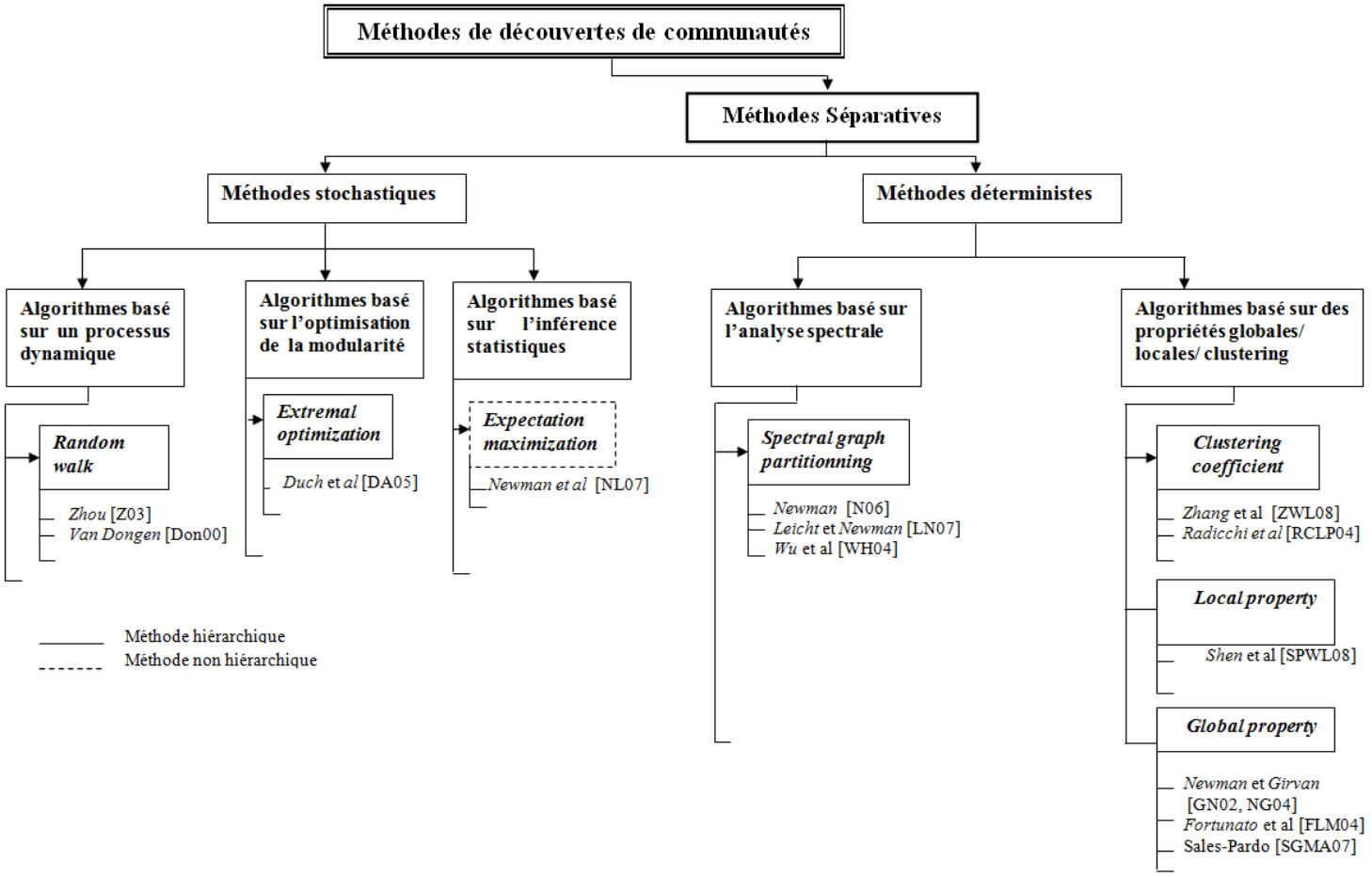

Figure 23. *Taxonomie des méthodes de découverte de communautés dans les réseaux complexes : 2. Méthode séparatives.*