



**HAL**  
open science

## Détection des communautés dans les réseaux orientés à l'aide des concepts formels

Norbert Tsope, Gamgne Domgue Félicité

► **To cite this version:**

Norbert Tsope, Gamgne Domgue Félicité. Détection des communautés dans les réseaux orientés à l'aide des concepts formels. 2016. hal-01308690

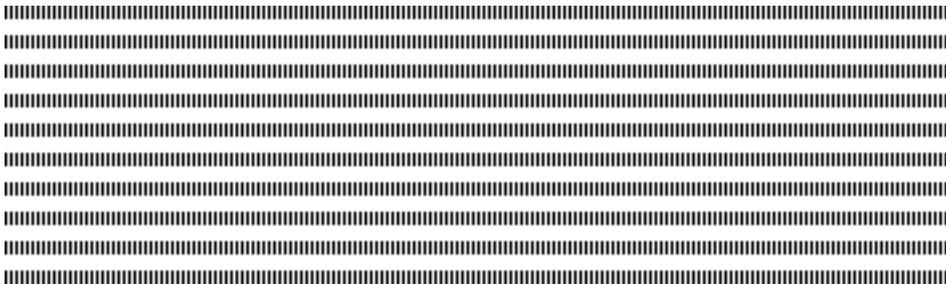
**HAL Id: hal-01308690**

**<https://hal.science/hal-01308690>**

Preprint submitted on 28 Apr 2016

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

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



Rubrique

## Analyse des réseaux orientés

### Détection des communautés dans les réseaux orientés à l'aide des concepts formels

Tsopze Norbert\* — Gamgne Domgue Félicité\*

\* Département d'Informatique - Université de Yaoundé I  
BP 812 Yaoundé - Cameroun  
tsopze@uy1.uninet.cm, fgamgne@uy1.uninet.cm



**RÉSUMÉ.** La détection des communautés dans les réseaux orientés ou non est une problématique actuelle. Elle consiste à regrouper les nœuds ou acteurs en groupes homogènes. Plusieurs chercheurs ont proposé des algorithmes pour résoudre ce problème; ces algorithmes fournissent des résultats acceptables en supposant que la relation entre les acteurs peut être symétrique, alors que dans beaucoup de situations réelles, la relation entre acteurs ne peut pas être symétrique. Dans ce travail, nous proposons une méthode basée sur les concepts formels permettant de prendre en compte la nature symétrique ou non de la relation entre les acteurs. Cette approche consiste à extraire des concepts noyaux et d'en déduire les communautés. Cette approche illustrée sur un cas de réseau sur le cas de réseau ayant relation symétrique, trouve le résultat est semblable à ceux des méthodes de l'état de l'art et sur le cas non symétrique, trouve les communautés attendues contrairement aux méthodes existantes. Les résultats expérimentaux sur des jeux de données pris sur Internet sont meilleurs que ceux de certains algorithmes existants.

**ABSTRACT.** Community detection in directed networks appears as one of dominant research works in network analysis. Most existing models for community detection are symmetric, in which incoming and outgoing links are treated equally. In this paper, we propose a method based on Formal Concepts that takes into account both symmetric and non-symmetric properties of links. To validate our approach, experiments on known networks with symmetric property of links in directed graphs show that outputs detect the same communities as some of state-of-art methods; while, with non-symmetric property of links, only this approach uncovers expected communities. Its Experimental results on benchmark are also better than those of many existing algorithms.

**MOTS-CLÉS :** Graphes orientés, Détection des communautés, Concepts formels, Noyau de la communauté

**KEYWORDS :** Directed graph, Community detection, Formal concepts, Community kernel



---

## 1. Introduction

La détection des communautés a pris une place très importante dans la recherche avec le développement des réseaux sociaux. Un réseau étant souvent modélisé par un graphe où les nœuds sont des acteurs et les liens entre eux la relation entre ces acteurs. Extraire les communautés est un processus consistant en la séparation des nœuds en des groupes homogènes (fonction du critère de similarité prédéfini). Ce processus peut s'appliquer à des graphes orientés et non-orientés. Certains algorithmes d'extraction de communautés dans les graphes orientés ne tiennent pas compte de l'orientation des liens, ou transforment ces graphes en des graphes bipartites. D'autres algorithmes supposent que la relation représentée par les liens est de nature symétrique (ou réciproque); alors que dans plusieurs situations réelles, cette relation ne l'est pas. Un exemple de graphe dans lequel l'orientation des liens serait capitale est la chaîne alimentaire où les acteurs ou nœuds représentent les espèces vivantes tandis que les liens symbolisent la relation selon laquelle une espèce (proie) a pour prédateur une autre.

Plusieurs auteurs définissent la notion de communauté comme étant un ensemble de nœuds densément connectés entre eux et faiblement connectés avec les autres nœuds du graphe [24, 29, 7, 6]. Cette définition n'est valable que lorsque l'on suppose que la relation entre les acteurs est réciproque ou symétrique comme c'est le cas dans les graphes d'amitié, de correspondances téléphoniques (appels, envois d'emails ou de SMS...). Dans les situations où cette relation n'est pas symétrique, comme c'est le cas dans la relation *A est prédateur de B*, les acteurs appartenant à une communauté n'ont généralement plus de liens entre eux. De ce fait, il n'est pas assez correct de définir une communauté comme un groupe d'acteurs densément connectés. Une communauté pourrait donc être définie comme un ensemble d'acteurs possédant des propriétés communes ou ayant un même comportement, un groupe homogène d'acteurs, à l'instar des herbivores, des carnivores... La notion d'homogénéité dépend de la connaissance qu'on a du réseau ou que l'on souhaiterait extraire du réseau.

Pour former les communautés en tenant compte la propriété symétrique (ou non) de la relation entre acteurs dans un réseau, nous proposons un algorithme basée sur les concepts formels [19, 46] plus précisément les facteurs booléens [5]. Cette méthode extrait d'abord les concepts noyaux, puis en fonction de la nature de la relation représentée sur les liens, définit les noyaux des communautés et enfin associe à chaque noyau les autres acteurs de la communauté. L'algorithme approximatif de génération des facteurs booléens proposé dans [5] a été modifié pour l'extraction des concepts noyaux. La modification consiste à éliminer à chaque étape du contexte les objets déjà affectés à un concept noyau (dans l'intention ou l'extension). Cette approche est différente des méthodes existantes basées sur les treillis de Galois [4, 1, 17] par le fait qu'elle est non seulement basée sur les facteurs booléens, mais aussi qu'elle ne génère pas les cliques et le contexte intermédiaire (à partir duquel ces méthodes déduisent les communautés). Le test sur des cas simples de graphes montre que pour des réseaux orientés ayant les liens de nature symétrique, cette méthode détecte les mêmes communautés que celles obtenues par les méthodes connues de l'état de l'art, et que dans le cas où les liens sont de nature non symétrique, seule cette nouvelle méthode détecte les communautés acceptables. Les mesures d'évaluation basées sur la densité des liens comme la modularité ne sont pas adaptées dans l'évaluation des communautés dans les réseaux ayant une relation non symétrique entre les acteurs. Les résultats expérimentaux sur les exemples *polblogs* (relation symétrique) et *elegans neural networks* (non symétrique) sont meilleurs que ceux de certaines méthodes existantes.

Dans la prochaine section, nous présenterons succinctement les méthodes de détection de communautés et l'application des treillis de Galois dans la recherche des communautés. La troisième section aura pour objet l'extraction des concepts noyaux et la formation des communautés. Le test de cette méthode fera l'objet de la quatrième section, suivie d'une conclusion.

---

## 2. Détection des communautés

La définition de la notion de communauté reste un challenge pour la communauté de recherche de ce domaine. Plusieurs auteurs donnent une définition à cette notion fonction de l'approche formelle mathématique qu'ils implémentent. Ces définitions ont en commun l'idée de *densité* entre entités de la même communauté [16, 35, 22]. Ainsi, la définition la plus utilisée correspond à celle de Yang [47] : "une communauté est un groupe de nœuds d'un réseau à l'intérieur duquel les liens connectant les nœuds sont plus denses que ceux les connectant avec les nœuds extra-communautaires". Fortunato [16] identifie trois variantes de définitions : la définition orientée locale, la définition orientée globale et celle basée sur la similarité des nœuds. Dans la première variante, la qualité intrinsèque de chaque communauté prise séparément et indépendamment de l'ensemble du graphe est évaluée. La deuxième variante assigne un critère de qualité à l'ensemble du graphe pour vérifier que l'ensemble des communautés est cohérent en tant que tout. La troisième variante se base sur la notion de similarité des nœuds. Encore appelée *cluster* ou *module*, la notion de communauté se verra attribuée une définition fonction de l'implémentation fixée. Ainsi, l'on distingue entre autres les cliques, n-cliques, k-plex, etc. communautés déterminées via les algorithmes basés sur les mesures de qualité [33].

Les mesures de qualité sont utilisées pour comparer les performances de différents algorithmes de détection de communauté qui n'optimisent pas la même fonction "objectif". Cette approche d'implémentation basée sur les mesures de qualité se déroule généralement en deux étapes :

- La première étape vise à spécifier la mesure qui capturera au mieux la structure du réseau

- La deuxième consiste à optimiser la mesure de qualité préalablement définie, afin d'obtenir la partition maximisant cette dernière.

Plusieurs mesures de qualité ont été définies dans la littérature : Typiquement, certaines d'entre elles se basent sur la densité de liens à la fois intra qu'extra-communautaire, à l'instar de la coupe ratio, la coupe normalisée [41], la conductance, la couverture et la performance [8]. D'autres mesures se basent uniquement sur l'un des deux types de densité. La modularité, dont le précurseur est Newman [31] en est l'une des plus usuelles. Elle mesure "le nombre de liens à l'intérieur d'une communauté versus le nombre attendu de liens dans cette communauté si les liens apparaissent aléatoirement dans le graphe tout en respectant la distribution des degrés des nœuds" [31]. Cependant elle présente une limite de résolution [15] qui stipule que sa maximisation ne conduit pas toujours à l'obtention de tous les modules du graphe.

### 2.1. Méthodes de détection des communautés

Quelques études [33] menées sur les méthodes de détection de communautés permettent de les classer en trois grandes catégories d'approches :

- Les approches de partitionnement de graphe,

- Les approches de clustering hiérarchique,
- Les approches de recouvrement.

Les approches de partitionnement de graphe [26] qui ont pour but de grouper les sommets d'un graphe en un nombre fixe de parties, de tailles connues, tout en minimisant le nombre d'arêtes entre les différents groupes. Cette approche a comme limite de requérir une connaissance préalable du nombre de communautés recherchées ainsi que de leurs tailles. La méthode de bissection spectrale [36] par exemple consiste à calculer le vecteur propre correspondant à la plus petite valeur propre non nulle de la matrice Laplacienne du graphe  $L = D - A$  où  $D$  est la matrice diagonale ( $n \times n$ ) avec  $d_{ii}$  le degré du nœud  $i$ . Le graphe est alors séparé en deux parties en fonction du signe de leur composante selon ce vecteur propre. Cette méthode fonctionne en temps  $\theta(n^3)$  et obtient de bons résultats lorsque le graphe possède effectivement deux grandes communautés de tailles similaires, ce qui n'est qu'un cas particulier dans la détection de communautés.

Pour pallier à cette limite des approches de partitionnement, d'autres études ont été faites afin de dénicher des approches de remédiation, à savoir celles hiérarchiques non supervisées, car ne nécessitent pas a priori une connaissance de la taille et du nombre de communautés, et produisent en résultat un ensemble de groupes distincts de nœuds dans lequel chaque entité appartient à une et une seule communauté. Elles sont divisées en algorithmes "séparatifs" à l'instar de l'algorithme de Girvan et Newman [31], et algorithmes "agglomératifs" à l'instar de l'algorithme de Louvain [6] et ces méthodes basées sur l'optimisation [11] d'une fonction "objectif", pour évaluer la qualité d'une partition donnée. Les classes à fusionner sont celles qui promettent une modularité maximum. Ces approches présentent la limite qu'elles ne tiennent pas compte du caractère d'ubiquité des individus (appartenance à plusieurs communautés).

Les approches de recouvrement permettent de remédier et d'aboutir à des communautés chevauchantes [21]. Cette catégorie d'approche de détection des communautés est très intéressante pour la majorité des graphes car permet de refléter véritablement la réalité (un individu peut appartenir à la fois à la communauté des collègues, des amis et de la famille), et de tels graphes peuvent également être orientés. En effet, ceux-ci traduisent au mieux la sémantique portée par les liens qu'il serait incorrect de négliger, et possèdent pour caractéristique principale de décrire effectivement les réseaux réels tels que les réseaux d'information, les réseaux de publication scientifique [20], les réseaux biologiques [43], les chaînes alimentaires [42], car leurs structures sont loin d'être aléatoires (comme ceux des graphes de Erdos-Renyi [14]).

Les approches de recouvrement peuvent être implémentés par plusieurs techniques, à l'instar des hypergraphes et des treillis de Galois, tels que présentés dans la section ci-dessous.

## 2.2. Méthodes de détection dans les graphes orientés

Plusieurs méthodes de détection des communautés dans les réseaux ont été étudiées [30]. Etant donné le caractère d'ubiquité associé aux réseaux, l'on observe plusieurs types de réseaux du monde réel, à savoir les réseaux technologiques incluant Internet, les grilles électriques, les réseaux téléphoniques, les réseaux routiers qui constituent une partie importante des activités quotidiennes des individus. La caractéristique commune de ces réseaux est la circulation d'un point  $A$  vers un point  $B$  des éléments. Par ailleurs, dans les réseaux d'interaction entre les entités (Hommes, êtres vivants, sites web), l'orientation des liens montre l'orientation de la relation entre ces entités. Ne pas tenir compte de cette orientation fait perdre cette information.

Déterminer les communautés dans les graphes orientés constitue un challenge captivant, avec plusieurs domaines d'applications. L'analyse de l'étude faite par [30] permet de scinder ces méthodes en deux familles :

- La famille des méthodes de transformation de la méthodologie initiale du graphe
- La famille des méthodes alternatives.

### 2.2.1. Méthodes de transformation de la méthodologie initiale du graphe

Dans cette famille, les graphes orientés sont transformés en un autre type de graphe : soit unipartite non-orienté, valué ou alors bipartite [12], ensuite les algorithmes de détection de communautés dans les graphes non-orientés leur sont appliqués. Cette démarche a priori semble intéressante, cependant la sémantique portée par les liens n'est pas mise en évidence, et l'utilité de l'information n'est pas prise en considération, à cause de notion de densité entre les nœuds. Autrement dit, les approches de clustering doivent être basées non seulement sur le critère de densité, mais également sur les propriétés des liens entrants et sortants. En effet, ces propriétés permettent de décrire au mieux la circulation de l'information dans les graphes orientés. C'est dans cette veine que [30] distingue les approches basées sur la **densité** des approches basées sur les  **patrons de structure**. Pour illustrer cette idée, dans l'exemple de la chaîne alimentaire [42], les prédateurs et leur proie ne saurait appartenir à la même communauté (le contraire tiendrait avec l'idée de densité dans la mesure où il y a plus d'interaction entre ces deux types d'espèces), par contre, une communauté "sensée" serait constituée d'espèces se nourrissant du même type de proie.

Par ailleurs, la transformation consiste en l'extension ou en la reformulation des fonctions de qualité pour les graphes orientés, et des outils algorithmiques d'analyse des graphes non orientés en des "versions orientées". Ainsi, l'on distingue entre autres :

- la modularité pour les graphes orientés (**équation 2.**) proposée par Arenas [3], qui avait été introduite pour les graphes non-orientés par Newman et Grivan [31] tel qu'exprimé dans (**équation 1.**).

$$Q = \frac{1}{2m} \sum [A_{ij} - \frac{k_i k_j}{2m}] \delta(c_i, c_j) \quad [1]$$

Où  $A_{ij}$  est la matrice d'adjacence du graphe,  $k_i$  représente le degré du sommet  $i$ ,  $m$  la taille du graphe (somme des arêtes), et  $\delta$  la fonction de Kronecker qui vaut 1 si les nœuds  $i$  et  $j$  appartiennent à la même communauté, et 0 sinon.

$$Q_u = \frac{1}{m} \sum [A_{ij} - \frac{k_i^{out} k_j^{in}}{m}] \delta(c_i, c_j) \quad [2]$$

- la matrice Laplacienne pour les graphes non-orientés [10] qui est un outil de clustering spectral basé sur la coupe normalisée [41], se verra modifiée pour les graphes orientés par Li et Zhang [28]. Elle s'exprime par l'équation 3.

$$L_d = I - \frac{\Pi^{1/2} P \Pi^{-1/2} + \Pi^{-1/2} P^T \Pi^{1/2}}{2} \quad [3]$$

où  $P$  est la matrice de transition, c'est-à-dire que  $P_{ij} = \frac{A_{ij}}{k_i^{out}}$  et  $\Pi$  la matrice diagonale avec probabilité de trouver une marche aléatoire sur chaque nœud.

- la conductance orientée de R. Anderson [2]. Cette mesure est fondée sur l'observation que si un groupe est bien connecté, alors un nombre important d'arêtes doivent être enlevées pour le diviser en deux parties égales. Les auteurs prouvent que le calcul d'un vecteur PageRank personnalisé dans les graphes orientés va d'un nœud "semence" contenu dans

un ensemble de nœuds  $S$  de conductance  $\alpha$ , pour obtenir un ensemble  $S'$  dont la conductance est définie par l'équation 4.

$$\Phi_M(S') = O(\sqrt{\alpha \log |S'|}) \quad [4]$$

### 2.2.2. Méthodes alternatives

Contrairement à la famille précédente dans laquelle les méthodes effectuent une transformation du graphe orienté en graphe non-orienté ou des outils d'analyse des graphes non-orientés en des versions orientées, dans cette catégorie de méthodes se retrouvent les techniques avec des méthodologies variées. Il s'agit des méthodes pouvant être classées soit dans la typologie de la théorie de l'information, soit dans la typologie des méthodes probabilistes, dans la mesure où après étude du phénomène sur un échantillon du graphe menant à une formule de corrélation, les mesures probabilistes sont appliquées pour étendre le comportement sur l'ensemble du graphe. A ce titre elles peuvent être groupées en trois classes :

- L'approche basée sur la théorie de l'information est fondée sur l'idée de compression des bits via des marches aléatoires et via des matrices binaires.

- L'approche par probabilité

- L'approche basée sur l'utilisation des concepts de treillis

Dans l'approche basée sur la théorie de l'information, pour ce qui est de la compression via les marches aléatoires, Rosvall et al. [39] implémentent un outil de détection, Isomap, à travers lequel ils décrivent la reconnaissance des types de flux d'information (coulée de l'information dans le réseau) entre les nœuds du graphe, par application du concept de la compression des marches aléatoires qui consiste en la minimisation de la taille de ces derniers. Selon eux, la reconnaissance d'un *module* est définie par la rapidité de circulation de l'information entre les nœuds du réseau. Leur objectif est celui de la détermination de la partition  $M$  minimisant la taille des marches aléatoires. En ce qui concerne la compression via une matrice binaire, l'objectif de Chakrabarti [9] est de grouper les bits de la matrice d'adjacence du graphe en des rectangles (blocs homogènes de bits indiquant la densité de connexion entre les nœuds d'un même groupe). Son algorithme nommé AutoPart est considéré comme un outil de *co-clustering* (les lignes et les colonnes de la matrice sont manipulées simultanément dans le processus de clustering) des matrices binaires. La qualité de la partition est évaluée en fonction du coût de compression. La meilleure compression est définie par le nombre de blocs produits et leur degré d'homogénéité. Au meilleur des cas, l'on obtient un seul bloc sensiblement homogène, et au pire des cas  $n^2$  blocs homogènes de taille 1 (chaque cellule de la matrice). Ainsi le meilleur clustering est celui qui minimise le cout de compression. La complexité de AutoPart est  $\theta(Imk^2)$  où  $I$  est le nombre d'itérations pour la minimisation du cout de compression,  $k$  le nombre optimal de clusters, et  $m$  la taille (nombre d'arêtes) du graphe.

Dans cette catégorie, Wang et al. [44] se basent sur le même principe de bloc pour détecter les communautés. Ils utilisent une approche par noyau qui est une approche structurée en deux étapes. La première consiste à regrouper les nœuds noyaux appelés *kernels* qui sont les nœuds les plus influents des communautés. Ces *kernels* peuvent être des triangles [27], ou des nœuds possédant une certaine centralité dans le réseau. Ces auteurs proposent la notion de *communauté kernel* pour explorer la structure des grands réseaux complexes (Par exemple, dans le réseau de "fans-followers" de Twitter, il existe certaines célébrités qui constituent les nœuds influents). Et la deuxième étape consiste à faire mi-

grer les autres nœuds "non-kernels" vers les communautés kernels avec lesquelles ils ont le plus grand nombre de connexions.

L'approche probabiliste présente deux formulations de solution au problème de détection de communautés dans les réseaux : la première faisant référence à l'application de l'inférence statistique et l'autre décrivant une modélisation stochastique par blocs. Dans la première formulation, l'inférence statistique est un procédé de représentation des données par un ensemble de variables. L'idée est la suivante : soit  $\pi_r$  la variable aléatoire représentant la fraction des nœuds dans une communauté  $r$  et  $\theta_{r,i}$  la probabilité qu'il existe un lien orienté d'un nœud quelconque de la communauté  $r$  vers un autre nœud  $i$ . Le problème de détection revient à maximiser la probabilité que les données soient générées par un modèle donnée en respectant des paramètres du modèle. Dans la modélisation par bloc il s'agit en fait d'un procédé de clustering dans lequel les nœuds du réseau sont groupés en fonction de la définition accordée à l'équivalence qu'ils possèdent. Holland et al. [23] définissent l'équivalence stochastique ainsi : « deux nœuds  $a$  et  $b$  sont stochastiquement équivalents si et seulement si la probabilité pour tout événement du réseau reste inchangée après permutation des nœuds  $a$  et  $b$  ». C'est dans le même sillage que Yang et al. [47] proposent un modèle par bloc, *Popularity and Productivity Link model (PPL)*, qui tient compte simultanément des liens entrants et sortants. Pour cela ils définissent deux variables, la *Productivité* et la *popularité* pour extraire respectivement les liens sortants et entrants. L'approche *PPL* modélise la probabilité qu'il existe un lien d'un nœud  $i$  vers un nœud  $j$ , en se servant de la variable productivité  $\alpha_i$  d'un nœud  $i$  qui décrit comment un lien provient de  $i$ , et la variable popularité  $\beta_j$  d'un nœud  $j$  qui décrit comment un lien arrive à  $j$ . [45, 38] proposent d'autres méthodes de modélisation par bloc.

La majorité des méthodes ne s'intéressent pas à la raison d'être des relations entre deux nœuds. Pour y remédier, certains auteurs se sont intéressés à l'approche basée sur les concepts de treillis ; *Comment extraire la connaissance dans un réseau social pour détecter des communautés plus sémantiques ?* C'est dans cette veine que Plantié et Crampes [33] étudient une approche de modélisation de la structure du graphe par les concepts des treillis sur la base des propriétés d'un hypergraphe [32] ; et pour ce faire, ils utilisent des algorithmes basés sur les treillis de Galois proposé par Freeman [18]. Les treillis de Galois sont des structures plus complexes que les graphes simples, mais fournissent plus de sémantique à la structure du réseau. L'idée ici est de partir d'un graphe bipartite et de le transformer en une hiérarchie du treillis, de telle sorte que les communautés correspondent aux extensions tandis que la structure résultante décrit un hypergraphe dont les hyperliens sont représentés par les intentions. Dans de telles structures, la connaissance au sein d'une communauté est préservée à cause du partage des propriétés communes par plusieurs nœuds.

La complexité de ces méthodes est présentée dans le tableau 1 selon la notation suivante :  $m$  est la taille du graphe (somme des arêtes) et  $n$  le degré du graphe (nombre de nœuds).

## 2.3. Application des treillis à la détection des communautés

### 2.3.1. Analyse des concepts formels

**Définition 1** Un contexte formel est un triplet  $C = (O, A, I)$  où  $O$  est un ensemble non vide d'objets et  $A$  un ensemble non vide d'attributs et  $I$  est une relation binaire entre les éléments de  $O$  et ceux de  $A$  (formellement  $I \subseteq \{(o, a) / o \in O \text{ et } a \in A\}$ ).

Méthodes	Fonctions de qualité	Caractéristiques	Type de graphe	Complexité
Partitionnement local[2]	Conductance	Densité de liens	Orienté	Optimisation spectrale
Bissection spectrale [36]		Matrice Laplacienne	Non orienté	$\theta(n^3)$
Algorithme de Girvan et Newman [20]	Centralité d'intermédiarité	Matrice d'adjacence	Non orienté	$\theta(m^2n)$
L'algorithme de Newman d'optimisation de la modularité [31]	Modularité	Fusion des nœuds si gain de la modularité	Non orienté	$\theta(mn)$
Algorithme de Louvain [6]	Modularité	Fusion des nœuds si gain de la modularité	Non orienté	$\theta(m)$
Modularité Orientée [31]	Modularité orientée	Fusion des nœuds si gain de la modularité	Orienté	Optimisation spectrale
Laplacien Orienté [28]	Coupe normalisée	Matrice Laplacien	Orienté	Optimisation spectrale
AutoPart [9]	Taille de la description	Compression	Non orienté/ Orienté	$\theta(Imk^2)$
Structures en treillis de Galois [33]		Concepts formels	Orienté	Complexité exponentielle
Communautés kernel [44]		Degré du kernel	Non orienté/Orienté	Linéaire
PPL [47]		Productivité et Popularité	Orienté	Linéaire

Tableau 1 – Récapitulatif des caractéristiques des approches de détection de communautés. La colonne 2 présente les métriques de qualité appliquées par chaque méthode (fonctions objectif et procédés de fusion ou division des nœuds du graphe), la colonne suivante présente les caractéristiques décrivant les graphes, la colonne 4 quant à elle spécifie les types de graphes auxquels s'appliquent les algorithmes et enfin la dernière colonne présente la complexité de chacune des approches de détection des communautés. Précisément, la complexité de l'optimisation spectrale repose sur celle de la décomposition des valeurs propres

Un contexte  $C$  peut être représenté comme une matrice binaire  $M$  (où  $M_{i,j} = 1 \iff (o, a) \in I, 0$  sinon) ou comme une base de transactions.

**Définition 2** Soient  $f$  et  $g$  deux applications définies par :

$$\begin{aligned} f &: 2^O \longrightarrow 2^A \\ f(X) &= X \uparrow = \{a \in A / \forall o \in X, (o, a) \in I\}, \\ g &: 2^A \longrightarrow 2^O \\ g(Y) &= Y \downarrow = \{o \in O / \forall a \in Y, (o, a) \in I\}. \end{aligned}$$

La fonction  $f$  retourne la liste des attributs que l'ensemble des objets appartenant à  $X$  ont en commun alors que  $g$  retourne l'ensemble des objets qui ont en commun l'ensemble des attributs de  $Y$ . Le couple  $(f, g)$  forme une correspondance de Galois. La fonction  $f$  (resp.  $g$ ) définit sur un ensemble de nœuds  $A$  d'un graphe (représenté par sa matrice d'adjacence) est l'ensemble des voisins sortants (resp. entrants) communs à tous les éléments de  $A$ .

**Définition 3** Une paire  $c = (X, Y)$  est appelée **concept formel** ssi  $X = Y \downarrow$  et  $Y = X \uparrow$ .  $X$  (resp.  $Y$ ) est appelé **extension** ou *extent*( $c$ ) (resp. **intention** ou *intent*( $c$ )) du concept.

**Définition 4** Soit  $X \subseteq O$  (resp.  $Y \subseteq A$ ), on dit que  $X$  (resp.  $Y$ ) est fermé si et seulement  $g \circ f(X) = X$  (resp.  $f \circ g(Y) = Y$ ).

Les concepts formels sont alors constitué des couples de fermés. Il s'agit des rectangles maximaux de '1' extraits dans le contexte binaire.

Soient  $C_1 = (X_1, Y_1)$  et  $C_2 = (X_2, Y_2)$  deux concepts formels et  $\leq$  une relation définie par  $C_1 \leq C_2$  si  $X_1 \subset X_2$  (ou encore  $Y_2 \subset Y_1$ ).

**Définition 5** La relation  $\leq$  définie une relation d'ordre sur l'ensemble des concepts, et l'ensemble des concepts extraits d'un contexte formel muni de cette relation est un treillis de Galois.

L'extraction des concepts d'un contexte donné a une complexité exponentielle dû au fait que l'extension (rep. intention) sous les sous ensembles de  $O$  (resp. de  $A$ ). Face à cette situation, les mesures telles que la fréquence, la stabilité sont appliquées pour sélectionner les concepts. Le lecteur peut consulter les références [19] et [46] pour plus de détails sur les treillis de Galois.

Les domaines d'application des treillis de Galois sont nombreuses : extraction des connaissances à partir des données, classification, détection des communautés,...

### 2.3.2. Treillis de Galois et détection des communautés

L'utilisation des treillis de Galois pour la détection des communautés dans un réseau  $G = (V, E)$  suit généralement les étapes suivantes [1, 18] :

- 1) Extraire les cliques  $K$  du réseau  $G$
- 2) Construire un treillis de Galois à partir du nouveau contexte  $F = (V, K, I)$  où pour  $v \in V$  et  $k \in K$   $(v, k) \in I$  si  $v$  est un nœud dans la clique  $k$
- 3) Détecter les cliques intermédiaires et supprimer les acteurs appartenant à ces cliques puis en déduire les communautés.

Plusieurs autres approches ont été proposées permettant d'appliquer les treillis de Galois à la recherche des communautés :

– Adapter l’optimisation de la modularité à travers une nouvelle formule appelée *Group Node Modularity* [1] pour améliorer les communautés partielles obtenues à partir des treillis. Les communautés finales sont obtenues des communautés partielles en incluant tous les acteurs dans des communautés par maximisation du *Group Node Modularity*.

– considérer l’extension d’un concept formel comme une communauté où les membres ont en commun les attributs appartenant à l’intention [13]. Cette idée a été utilisée dans [40] pour extraire les communautés épistémiques (ensemble d’acteurs partageant le même domaine de connaissances et travaillant pour son évolution). Cette manière de procéder est plus adaptée aux graphes bipartis où l’ensemble  $O$  représente les acteurs, l’ensemble  $A$  les attributs et il n’existe aucune entre les éléments d’un même ensemble.

– combiner l’approche basée sur l’optimisation des fonctions et les treillis de Galois ; la méthode basée sur l’optimisation des fonctions permet de trouver les ensembles de nœuds fortement connectés alors les treillis de Galois permettent d’étudier l’évolution des groupes. Les auteurs mettent en évidence les avantages des treillis de Galois qui contrairement aux méthodes basées sur l’optimisation de la modularité comme la méthode de Louvain qui obtiennent les communautés non déterministes, les treillis de Galois fournissent les communautés déterministes, et aussi recouvrantes étiquetées avec leurs extensions et intentions [4].

---

### 3. Extraction des concepts et formation des communautés

Comme un réseau est représenté par un graphe, la matrice d’adjacence de ce graphe peut être déduite et considéré comme un contexte binaire. Nous représentons alors le graphe  $G = (V, E)$  par sa matrice d’adjacence  $M$ .  $V$  et  $E$  sont respectivement l’ensemble des nœuds et l’ensemble des liens. Pour un couple de nœuds  $(u, v)$ ,  $M_{u,v} = 1$  s’il existe un lien de  $u$  à  $v$  dans  $G$  et 0 sinon. Le graphe  $G$  correspond donc à un contexte binaire  $(V, V, M)$  à partir duquel les concepts peuvent être extraits. Un concept est une paire d’ensembles  $(A, B)$  de nœuds où tous les éléments  $A$  ont comme voisins (sortant) tous les éléments de  $B$ .

**Définition 6** Soient  $D$  un ensemble de nœuds et  $j \notin D$  un nœud, la fonction d’ajout de  $j$  à  $D$  notée  $\oplus$  est le concept défini l’équation 5.

$$D \oplus j = |(D \cup j) \downarrow, (D \cup j) \downarrow \uparrow| \quad [5]$$

La fonction d’ajout permet d’ajouter un nouvel élément à l’extension un concept et d’en déduire un nouveau concept.

**Définition 7** La fonction d’utilité (équation 6) définie sur un concept  $C$  est calculée par l’équation

$$U(C) = \frac{|extent(C)| \times |intent(C)|}{|E|} \quad [6]$$

La fonction  $U(C)$  permet de calculer la proportion des liens à l’intérieur du concept  $C$ .

**Définition 8** Un concept noyau est un concept  $C$  qui maximise la fonction d’utilité (équation 6) c’est-à-dire il n’existe pas un concept  $C'$  tel que  $extent(C) \subset extent(C')$  et  $U(C') \geq U(C)$ .

Les nœuds appartenant à l'union de l'intention et de l'extension du concept noyau sont des nœuds noyaux.

**Définition 9** La relation représentée par les liens dans un réseau est dite **non symétrique** si et seulement si  $\forall x, y \in V$ , si  $(x, y) \in E$  alors  $(y, x) \notin E$

**Proposition 1** Si la relation entre les nœuds est non symétrique alors pour tout concept  $C$  extrait du contexte formel  $(V, V, M)$  correspondant à la matrice d'adjacence du réseau,  $extent(C) \cup intent(C) = \emptyset$ .

La validité de cette proposition est due au fait pour toute relation non symétrique et pour deux nœuds  $x$  et  $y$  de  $V$ , si  $(x, y) \in E$  alors  $(y, x) \notin E$ . Donc tout ensemble de nœuds  $A$ , on aura  $A \cap f(A) = \emptyset$  et  $A \cap g(A) = \emptyset$ .

L'approche proposée se déroule en deux étapes :

1) Extraction des concepts noyaux. Un concept noyau est une biclique de taille maximale que l'on peut extraire du contexte. Les nœuds n'appartenant pas au noyau sont appelés nœuds non noyaux.

2) Déduction des noyaux et formation des communautés. En fonction de la nature symétrique ou non de la relation entre les acteurs, nous formons les noyaux à partir des concepts noyaux, puis les nœuds non noyaux sont affectés à ces noyaux pour constituer la communauté.

### 3.1. Extraction des concepts noyaux

Une approche simple serait d'extraire l'ensemble des concepts formels du contexte et de sélectionner les concepts noyaux. Mais la recherche des concepts a une complexité exponentielle. Pour cela, nous proposons une approche glouton qui consiste à ajouter un à un les nœuds à un ensemble  $D$  (en utilisant la fonction de l'équation 5) et qui maximise la fonction d'utilité (équation 6). Cette fonction représente la surface du rectangle constitué des "1" dans le contexte binaire représentant la matrice d'adjacence. Cette fonction croît de 0 à partir de zéro (car  $D$  est vide initialement) et atteint son maximum lorsque les opérandes de la multiplication tendent à s'égaliser, puis décroît. Le concept obtenu maximise la fonction et est sauvegardé, et  $D$  est encore réinitialisé au vide et le processus se répète. Le paramètre  $\varepsilon$  fixé permet d'éviter d'extraire les concepts de taille très faible.

La complexité de l'algorithme 1 est polynomiale car le calcul de  $D \downarrow$  ou de  $D \uparrow$  est linéaire. La sortie de l'algorithme 1 représente un ensemble de concepts formant ainsi les noyaux des communautés. La tâche suivante est d'en déduire les communautés en associant les nœuds n'appartenant pas aux noyaux (nœuds non noyaux) à l'un des noyaux.

### 3.2. Déduction des communautés

La propriété *symétrie* de relation entre les acteurs est prise en compte. Deux cas sont envisageables :

#### 3.2.1. Relation symétrique en acteurs

Si la relation est symétrique alors les noyaux des communautés sont formés respectivement par l'union de l'extension et de l'intention de chaque concept noyau. C'est le cas lorsque les nœuds sont les pages web et les liens hypertextes correspondent à la relation entre ces nœuds. Chaque nœud non noyau est alors affecté à la communauté avec laquelle il a le plus nombre de voisins. L'algorithme 2 implémente cette proposition.

---

**Algorithm 1** Extraction des concepts noyaux

---

**Entrées:** contexte binaire (matrice d'adjacence)  $K = (V, V, M)$  et un seuil d'utilité  $\varepsilon$

**Sorties:** concepts formels noyaux

- 1:  $L \leftarrow \emptyset$
  - 2: **répéter**
  - 3:  $C \leftarrow \emptyset$
  - 4:  $k_0 \leftarrow 0$
  - 5: **répéter**
  - 6:  $k \leftarrow k_0$
  - 7:  $C \leftarrow C \cup \{ArgMax_j | C \oplus j\}$
  - 8:  $k_0 \leftarrow |C \uparrow \downarrow| \times |C \uparrow|$
  - 9:  $q \leftarrow k_0 / |M|$
  - 10: **jusqu'à** ( $q < \varepsilon$  ou  $k_0 < k$ )
  - 11:  $V \leftarrow V - (C \uparrow \downarrow \cup C \uparrow)$
  - 12:  $L \leftarrow L \cup (C \uparrow \downarrow, C \uparrow)$
  - 13: **jusqu'à** ( $q < \varepsilon$ )
  - 14: return  $L$
- 

### 3.2.2. Relation non symétrique.

Suivant la proposition 1, il est évident que les nœuds appartenant respectivement à l'extension et à l'intention d'un concept appartiennent à deux communautés disjointes. Dans le cas d'une chaîne alimentaire (exemple de la figure 2), il est prudent de tenir compte de la nature non symétrique afin d'éviter que proie et prédateur ne retrouvent dans la même communauté. Pour cela nous proposons de définir les noyaux comme étant respectivement les extensions et les intentions des concepts noyaux obtenus à la fin de l'algorithme 1. Un acteur non noyau est simplement affecté à la communauté avec qui il a le plus grand nombre de voisins sortants (ou entrants). Cette proposition est formalisée par l'algorithme 3.

---

## 4. Illustration et expérimentations

Dans cette section, nous ferons d'abord une application de cette méthode à deux cas (un cas avec une relation symétrique et un autre avec une relation non symétrique); ensuite nous ferons l'expérimentation sur les benchmarks pris sur Internet.

---

**Algorithm 2** Communautés dans le cas d'une relation symétrique

---

**Entrées:**  $L = \{c_i\}$ , liste des concepts noyaux; contexte binaire  $K = (V, V, M)$

**Sorties:** Communautés

- 1:  $COM \leftarrow \{extent(c_i) \cup intent(c_i)\} \{ \text{Pour tout } c_i \in L \}$
  - 2: **pour**  $x \notin com$  **faire**
  - 3:  $C_{max} \leftarrow ArgMax [(com \cap \{x\} \uparrow), (com \cap \{x\} \downarrow)]$
  - 4: affecter  $x$  à  $com_{max}$
  - 5: **fin pour**
  - 6: retourner  $COM$
-

---

**Algorithm 3** Communautés dans le cas d'une relation non symétrique

---

**Entrées:**  $L = \{c_i\}$ , liste des concepts noyaux ; contexte binaire  $K = (V, V, M)$

**Sorties:** Communautés

- 1:  $COM \leftarrow \{extent(c_i)\} \cup \{intent(c_i)\}$  {pour tout  $c_i \in L$ }
  - 2: **pour**  $x$  non noyau **faire**
  - 3:   **Si**  $\{x\} \uparrow \cap \{x\} \downarrow \neq \emptyset$  **alors**
  - 4:     Eliminer les noeuds appartenant aux noyaux et calculer dans le contexte restant le concept  $c$  contenant  $x$  qui maximise l'équation 6 dans le contexte ;
  - 5:     créer une nouvelle communauté avec  $c$
  - 6:   **sinon**
  - 7:      $C_{max} \leftarrow ArgMax [(com \cap \{x\} \uparrow), (com \cap \{x\} \downarrow)]$
  - 8:     **Si**  $C_{max} = intent(c)$  **alors**
  - 9:       affecter  $x$  à la communauté de noyau  $extent(c)$
  - 10:    **sinon**
  - 11:     affecter  $x$  à la communauté de noyau  $intent(c)$
  - 12:    **Fin si**
  - 13: **Fin si**
  - 14: **fin pour**
  - 15: retourner  $COM$
- 

#### 4.1. Illustration

Pour illustrer cette méthode, nous allons l'exécuter sur les deux graphes de la figure 3. La figure 2<sup>1</sup> représentant une chaîne alimentaire est un réseau où la relation entre les acteurs n'est pas symétrique. Pour cet exemple, les acteurs sont renard, chouette, hérisson, grenouille, hanneton, végétal, papillon, araignée, rouge-gorge, lapin, musaraigne. La figure 1 est un extrait du réseau social twitter pris dans [44] ; la relation entre les acteurs peut être symétrique (exemple entre les acteurs *Barack Obama* et *Al Gore*).

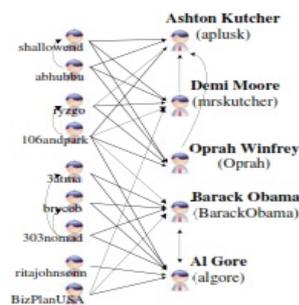


Figure 1 – Extrait du réseau social tweeter [44]

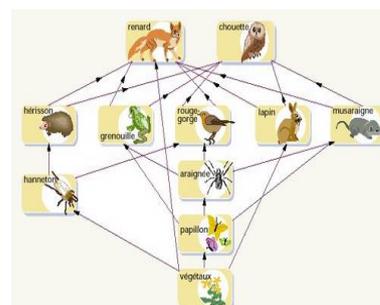


Figure 2 – Extrait d'une chaîne alimentaire

Figure 3 – Exemples de réseaux orientés

Pour le premier cas, les concepts noyaux obtenus sont :  $((Al\ Gore, Obama), (BPUSA, 3atma, 303nomad, brycob))$  et  $((DMoore, Oprah, Ahton), (Shalom, abhubbu, ryzgo, 106Park))$  ; le nœud non noyau est *ritaJohn* (coloré en violet dans la sous figure 4). Le premier noyau

---

1. cette image a été prise sur le site web <http://www.conservation-nature.fr/article1.php?id=310>, consulté le 1 mars 2016.

est constitué des politiciens avec comme extension (coloré en jaune dans la sous figure 4) sont *Al Gore* et *Obama* et l'intention constituée des suiveurs (coloré en vert dans la sous figure 4). Le deuxième concept noyau est composé en extension des *Stars* (nœuds bleus dans la sous figure 4) et leurs fans (nœuds colorés en rouge dans la sous figure 4). Après l'exécution de l'algorithme 2 de recherche des concepts noyaux, l'obtention des communautés finales se fait en associant le nœud non noyau *ritaJohn* à une communauté. Ce nœud est alors associé à la communauté des politiciens et le résultat final est présenté graphiquement par la sous figure 5.

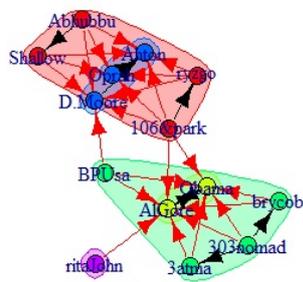


Figure 4 – Concepts noyaux

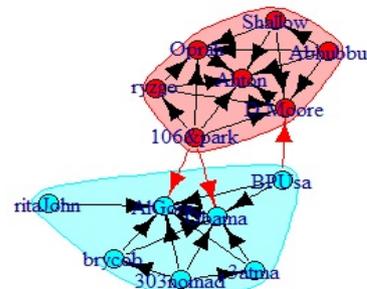


Figure 5 – Communautés finales

Figure 6 – Illustration sur le réseau 1.

La seconde illustration est le cas de la sous figure 2. Dans cet exemple la relation entre les acteurs est non symétrique et doit être prise en compte dans le traitement. Les concepts noyaux extraits par l'algorithme 1 sont ((renard, chouette), (hérisson, grenouille, rouge-gorge, lapin, musaraigne)) et ((hanneton, papillon), (végétal)). Seule l'araignée est le nœud non noyau. De cet exemple l'algorithme 3 met en sortie cinq communautés en créant une pour l'araignée (coloré en violet). La représentation graphique est présentée à la figure 7.

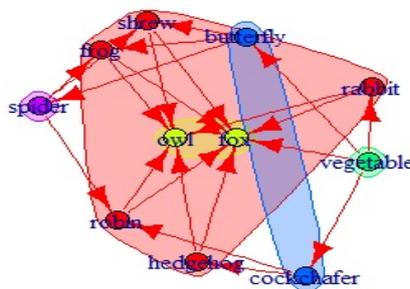


Figure 7 – Illustration sur la sous figure 2.

## 4.2. Expérimentations

Nous avons expérimenté cette nouvelle approche sur deux jeux de données<sup>2</sup> :

2. disponible sur <http://www-personal.umich.edu/~mjejn/netdata/>

1) *Elegan neural "celegansneural.gml"* qui est un réseau de neurones artificiels. Dans un réseau de neurones, les neurones sont connectés les uns aux autres et peuvent être organisés en couches de telle sorte qu'il n'existe aucune connexion entre les neurones d'une même couche. Dans cette expérimentation, nous avons supposé que la relation entre les neurones est non symétrique. Ce réseau contient 297 nœuds et 2359 connexions.

2) *Politic blogs "Polblogs"*, dans ce réseau, les nœuds sont les blogs des politiciens américains et les arcs les liens hypertextes entre ces blogs. Nous avons supposé que la relation entre les blogs est symétrique. Ce réseau compte 1490 nœuds et 19090 connexions. Sur le réseau "celegansneural", notre approche a trouvé quatre grandes communautés et neuf communautés à un nœuds. Ces quatre communautés peuvent être vues comme des couches du réseau de neurones. Sur les données "Polblogs", cette approche extrait deux grandes communautés et quelques blogs isolés. Ces deux grandes communautés peuvent être interprétées comme les deux principaux partis politiques américains.

### 4.3. Comparaison

Nous avons expérimenté quelques méthodes de l'état de l'art et cette section présente un récapitulatif de ces tests.

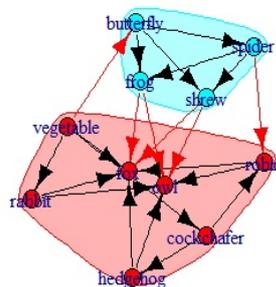


Figure 8 – Sortie de Walk-trap Community

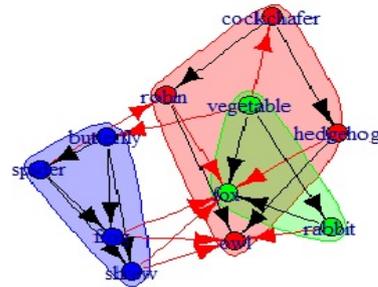


Figure 9 – Sortie de Springglass Community

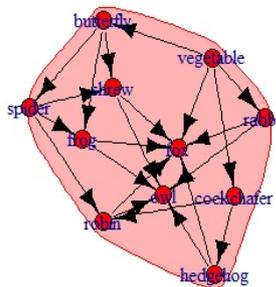


Figure 10 – Sortie de Edge-betweenness

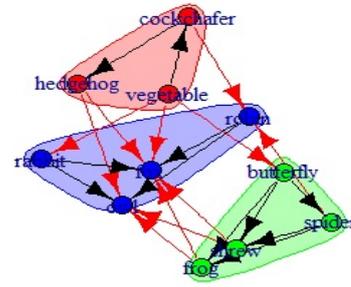


Figure 11 – Sortie de Optimal Communities

Figure 12 – Communautés obtenues des méthodes existantes sur la figure 2.

La modularité définie comme une fonction du nombre de liens internes à des communautés obtenues est la mesure la plus utilisée pour évaluer les algorithmes de détection des communautés. Le tableau 2 présente les modularités de ces algorithmes sur ces deux

réseaux. Ces trois algorithmes (notre approche, Walktrap [34], l’algorithme de Louvain [6] et Optimal communities [7]) ont une modularité maximale (0.402) et leurs sorties est conforme à la sortie désirée dans [44]. L’algorithme Edge-betweeness [20] a une modularité minimale. Sur la sous figure 2, Springlass [37] a une modularité maximale alors que pour la figure 2, elle est classée troisième avec les communautés mal structurées. Sa sortie est constituée de trois communautés avec dans la même communauté le lapin et les végétaux ; ce qui est impossible. Ceci montre l’une des limites de l’optimisation de la modularité comme énoncée dans [25].

Algorithmes	Modularité sur la figure 1	Modularité sur la figure 2
<b>FCA</b>	0.402	0.190
<b>Walktrap</b>	0.402	0.190
<b>Edge-betweeness</b>	0.154	0.00
<b>Louvain</b>	0.402	0.19
<b>Spring glass</b>	0.276	0.210
<b>Optimal communities</b>	0.402	0.148

Tableau 2 – Modularité des méthodes existantes sur les exemples 3

Le tableau 3 présentent une comparaison avec quelques méthodes de l’état de l’art. Dans ce tableau, la méthode *Edge-betweeness* obtient la modularité la plus élevée avec 194 communautés.

## 5. Conclusion

La détection des communautés est l’une des tâches importantes de l’étude des réseaux. Mais les méthodes existantes ne tiennent pas compte des connaissances sur la nature de la relation représentée par des liens entre les acteurs. Nous proposons une approche basée sur les concepts formels permettant de prendre en compte la nature symétrique ou non de la relation entre les acteurs. L’application de cette méthode sur un exemple de réseau où la relation entre acteur peut être symétrique montre les mêmes résultats que ceux des

Algorithme	Celegansneural		Polblogs	
	Modularité	nombre de communautés	Modularité	Nombre de communautés
Approche FCA	0.20	13	0.32	268
Walktrap	0.36	21	0.4	278
Edge-betweeness	0.80	194	-	-
Spring glass	0.38	5	-	-
Optimal communities	-	-	-	-
Louvain	0.45	6	0.42	278

Tableau 3 – Résultats expérimentaux.

méthodes existantes ; et dans un cas non symétrique, seule cette méthode a détecté les communautés acceptables.

---

## 6. Bibliographie

- Selmane Sid Ali, Fadila Bentayeb, Rokia Missaoui, and Omar Boussaid. An efficient method for community detection based on formal concept analysis. In Troels Andreasen, Henning Christiansen, Juan-Carlos Cubero, and Zbigniew W. Ras, editors, *Foundations of Intelligent Systems*, volume 8502 of *Lecture Notes in Computer Science*, pages 61–72. Springer International Publishing, 2014.
- Reid Andersen, Fan Chung, and Kevin Lang. Local partitioning for directed graphs using pagerank. In *Algorithms and Models for the Web-Graph*, pages 166–178. Springer, 2007.
- Alex Arenas, Jordi Duch, Alberto Fernández, and Sergio Gómez. Size reduction of complex networks preserving modularity. *New Journal of Physics*, 9(6) :176, 2007.
- Marie-Aude Aufaure and Bénédicte Le Grand. Advances in fca-based applications for social networks analysis. *Int. J. Concept. Struct. Smart Appl.*, 1(1) :73–89, 2013.
- R. Belohlávek and V. Vychodil. Discovery of optimal factors in binary data via a novel method of matrix decomposition. *Journal of Computer System*, 76(2) :3–20, 2010.
- Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics : theory and experiment*, 2008(10) :P10008, 2008.
- Ulrik Brandes, Daniel Delling, Marco Gaertler, Robert Gorke, Martin Hofer, Zoran Nikoloski, and Dorothea Wagner. On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20(2) :172–188, 2008.
- Ulrik Brandes and Thomas Erlebach. *Network analysis : methodological foundations*, volume 3418. Springer Science & Business Media, 2005.
- Deepayan Chakrabarti. Autopart : Parameter-free graph partitioning and outlier detection. In *Knowledge Discovery in Databases : PKDD 2004*, pages 112–124. Springer, 2004.
- Fan RK Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.
- Aaron Clauset, Mark EJ Newman, and Christopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6) :066111, 2004.
- Michel Crampes and Michel Plantié. Partition et recouvrement de communautés dans les graphes bipartis, unipartis et orientés. In *IC-24èmes Journées francophones d’Ingénierie des Connaissances*, 2013.
- Michel Crampes and Michel Plantié. Partition et recouvrement de communautés dans les graphes bipartis, unipartis et orientés. In *IC - 24èmes Journées francophones d’Ingénierie des Connaissances*, 2013.
- Paul Erdős and A Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hungar. Acad. Sci*, 5 :17–61, 1960.
- Santo Fortunato. Resolution limit in community detection. *Physics Reports*, 104 :36–41, 2006.
- Santo Fortunato. Community detection in graphs. *Physics Reports*, 486 :75–174, 2010.
- Linton C. Freeman. Cliques, galois lattices, and the structure of human social groups. *Social Networks*, 18(3) :173 – 187, 1996.
- Linton C. Freeman and Douglas R. White. Using galois lattices to represent network data. In *Sociological methodology*, pages 127–146, 1993.

- B. Ganter and R. Wille. *Formal Concepts Analysis : Mathematical foundations*. Springer - Verlag, 1999.
- 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.
- Steve Gregory. Finding overlapping communities in networks by label propagation. *New Journal of Physics*, 12(10) :103018, 2010.
- Natali Gulbahce and Sune Lehmann. The art of community detection. *BioEssays*, 30(10) :934–938, 2008.
- Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels : First steps. *Social networks*, 5(2) :109–137, 1983.
- Hong Jin, Shuliang Wang, and Chenyang Li. Community detection in complex networks by density-based clustering. *Physica A : Statistical Mechanics and its Applications*, 392(19) :4606 – 4618, 2013.
- Rushed Kanawati. Détection de communautés dans les grands graphes d’interactions (multiplexes) : état de l’art. Technical report, LIPN - Université de Paris 13, 2013. 52.
- Brian W Kernighan and Shen Lin. An efficient heuristic procedure for partitioning graphs. *Bell system technical journal*, 49(2) :291–307, 1970.
- Christine Klymko, David Gleich, and Tamara G Kolda. Using triangles to improve community detection in directed networks. *arXiv preprint arXiv :1404.5874*, 2014.
- Yanhua Li and Zhi-Li Zhang. Random walks on digraphs, the generalized digraph laplacian and the degree of asymmetry. In *Algorithms and models for the web-graph*, pages 74–85. Springer, 2010.
- Fragkiskos D Malliaros and Michalis Vazirgiannis. Clustering and community detection in directed networks : A survey. *Physics Reports*, 533(4) :95 – 142, 2013.
- Fragkiskos D Malliaros and Michalis Vazirgiannis. Clustering and community detection in directed networks : A survey. *Physics Reports*, 533(4) :95–142, 2013.
- Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2) :026113, 2004.
- Symeon Papadopoulos, Yiannis Kompatsiaris, Athena Vakali, and Ploutarchos Spyridonos. Community detection in social media. *Data Mining and Knowledge Discovery*, 24(3) :515–554, 2012.
- Michel Plantié and Michel Crampes. Survey on social community detection. In *Social Media Retrieval*, pages 65–85. Springer, 2013.
- Pascal Pons and Matthieu Latapy. Computing communities in large networks using random walks. *J. of Graph Alg. and App. bf*, 10 :284–293, 2004.
- Mason A Porter, Jukka-Pekka Onnela, and Peter J Mucha. Communities in networks. *Notices of the AMS*, 56(9) :1082–1097, 2009.
- 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.
- Joerg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74 :016–110, 2006.
- Jörg Reichardt and Douglas R White. Role models for complex networks. *The European Physical Journal B*, 60(2) :217–224, 2007.
- Martin Rosvall and Carl T Bergstrom. An information-theoretic framework for resolving community structure in complex networks. *Proceedings of the National Academy of Sciences*, 104(18) :7327–7331, 2007.

- Camille Roth and Paul Bourguine. Epistemic communities : Description and hierarchic categorization. *Mathematical Population Studies*, 12(2) :107–130, 2005.
- Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8) :888–905, 2000.
- DB Stouffer, J Camacho, R Guimera, CA Ng, and LA Nunes Amaral. Quantitative patterns in the structure of model and empirical food webs. *Ecology*, 86(5) :1301–1311, 2005.
- Peter Uetz, Loic Giot, Gerard Cagney, Traci A Mansfield, Richard S Judson, James R Knight, Daniel Lockshon, Vaibhav Narayan, Maithreyan Srinivasan, Pascale Pochart, et al. A comprehensive analysis of protein–protein interactions in *saccharomyces cerevisiae*. *Nature*, 403(6770) :623–627, 2000.
- Liaoruo Wang, Tiancheng Lou, , Jie Tang, and John E. Hopcro. Detecting community kernels in large social networks. In *ICDM '11 Proceedings of the 2011 IEEE 11th International Conference on Data Mining*, pages 784–793, 2011.
- Yuchung J Wang and George Y Wong. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82(397) :8–19, 1987.
- R. Wille. Restructuring lattice theory : An approach based on hierarchies of concepts. *I. Rival(Ed) Ordered Sets*, 95 :445–470, 1982.
- Bo Yang, Dayou Liu, and Jiming Liu. Discovering communities from social networks : methodologies and applications. In *Handbook of Social Network Technologies and Applications*, pages 331–346. Springer, 2010.