Nouvelle approche de clustering par kernel-pattern via la densité en triades

Optimisation de la métrique Kernel Degree Clustering

Félicité Gamgne Domgue* — **Norbert Tsopze*** — **Arnaud S. R. M. Ahouandjinou****

* *Université de Yaoundé I - Cameroun*
** *Université du Littoral de la Cote d'Opale(ULCO),62228 Calais, France*

RÉSUMÉ. *La détection des communautés est devenue un domaine de recherche majeur ces dernières années. Plusieurs algorithmes appliqués aux graphes orientés ont été developpés. Ces derniers se focalisent sur la densité de liens à l'intérieur des communautés et considèrent la relation entre les nœuds comme symmétrique, car ils ignorent l'orientation des liens, ce qui biaise les résultats en produisant des communautés non-significatives. Ce document propose un algorithme basé sur l'extraction des kernels via la distribution des triades, utilisant l'optimisation de la nouvelle métrique Kernel Degree Clustering (KDC), et trouve des communautés plus sémantiques que la modularité, en accord à la notion de centralisation de l'information. Les expérimentations montrent que la nouvelle approche produit les résultats préconisés que ceux produits par certains algorithmes de détection de communautés de l'état de l'art.*

ABSTRACT. *Community detection has become a major active area of research in recent years. A plethora of relevant methods have been implemented for directed graphs. Most of them focus on the density of links, and consider the relationship between nodes as symmetric by ignoring links directionality during their clustering step, this leading to non-semantic results. This paper propose an efficient method based on the extraction of kernels through the distribution of triads in the graph, using Kernel Degree Clustering (KDC) a novel metric to judge the quality of a community partitioning, demonstrated to yield superior results over other commonly used metrics like modularity in conformity with centrality. To validate our approach, we conduct experiments on some networks which show that it has better performance over some of the other state-of-the-art methods and uncovers expected communities.*

MOTS-CLÉS : *Réseaux orientés, Détection des communautés kernel, Clusters basés sur la structure, Triade.*

KEYWORDS: *Directed graphs, Community kernel detection, Pattern-based clusters, Triad.*

1. Introduction

La détection des communautés dans les graphes orientés apparait comme l'un des objectifs majeurs des domaines de la recherche d'informations et de l'analyse des réseaux. Dans son sens premier, la notion de communauté correspond à un ensemble de nœuds densément connectés entre eux et faiblement connectés avec les autres nœuds du réseau (Fortunato, 2010). La détection des communautés peut par exemple aider à faire du marketting viral ; ou dans un réseau de produits fréquemment achetés ensemble, la détection des communautés peut être utilisée pour faire de la recommandation. Au vu de ces diverses perceptions sur la manière dont les objets sont semblables ou similaires, il existe différents algorithmes de clustering formalisant ces pensées respectives. A ce titre, la similarité entre les nœuds d'une même communauté ne tiendra plus seulement compte de la densité de liens, mais aussi des caractéristiques structurelles des nœuds dans les graphes orientés. Ainsi, alors que certains algorithmes de détection de communautés implémentés pour les graphes orientés ignorent l'orientation des liens, d'autres techniques transforment le graphe orienté en graphe non-orienté unipartite et valué (Fortunato, 2010) ou bipartite, et ensuite appliquent les algorithmes de détection de communautés sur les graphes non-orientés pour extraire leurs communautés.

Ces techniques ne sont pas satisfaisantes et ne produisent pas des résultats significatifs parce que la sémantique portée par les liens n'est pas prise en compte. Par exemple, dans un graphe de citation dans lequel les articles sont représentés par les nœuds et les relations telles que "*un article cite un autre article*" sont représentées par les liens orientés. Supposons qu'un article *i* cite un autre *j* (relation *père (j)-fils(i)*)mais pas l'inverse. D'après ces méthodes, la relation de réciprocité ou de symétrie est introduite entre les articles *i* et *j*, ce qui favorise la perte de l'information selon laquelle *j* soit cité par *i*. Dans le but de garder cette sémantique d'orientation des liens, une définition plus générique de la notion de communauté a été introduite par (Malliaros et Vazirgiannis, 2013) comme étant un ensemble de nœuds possédant des caractéristiques homogènes (plus précisément "ensemble de nœuds centrés autour d'autres nœuds, ces derniers possédant les intérêts communs"). Étant entendu que la majorité des graphes réels sont de grande taille et deviennent de plus en plus denses, au vu des multiples et divers outils de manipulation de l'information à l'ère du numérique qui révolutionne la vie quotidienne, il devient plus difficile voire infaisable de les traiter, suite à la taille limitée de la mémoire des machines. Pour ces deux raisons, la complexité et la sémantique portée par les liens, les approches basées sur les kernels semblent être indiquées pour résoudre le problème de détection des communautés dans les grands réseaux. Notre approche se base sur l'extension de l'idée selon laquelle à l'intérieur des "bonnes"communautés se trouvent des nœuds influents, *kernels*, qui centralisent l'information afin qu'elle soit aisément accessible. Les nœuds influents dits nœuds centraux sont traversés par un nombre maximal de triades dans la communauté. Une triade peut se définir comme étant un sous graphe de 3 nœuds impliquant deux liens. Ainsi, les triades constituent les bases de plusieurs structures de communautés (Klymko *et al.*, 2014). Ce travail s'attardant sur l'orientation des liens

dans les triades, les contributions spécifiques y afférentes sont entre autres :

– La définition d'un nouveau concept nommé kernel Degree Clustering(KDC) qui mesure la puissance de similarité qui existe entre les nœuds du kernel, et une nouvelle sémantique donnée à la notion de communauté basée sur le voisinage des nœuds du kernel via l'appartenance triadique.

– L'implémentation d'un nouvel algorithme basé sur l'optimisation du KDC pour découvrir les kernels et par la suite les communautés qui en découlent.

– L'amélioration de la qualité des structures obtenues par rapport aux méthodes existantes.

La suite du document est structuré de la manière suivante : La Section 2 est une introduction aux méthodes existantes relatives à cette approche. Dans la Section 3, nous définissons formellement les différents concepts utilisés dans l'approche de clustering proposée. La section 4 présente une forme détaillée de l'implémentation de la nouvelle méthode, suivie de la section 5 qui présente les expérimentations faites pour étudier et évaluer les résultats obtenus. Et enfin la section 6 conclut notre étude.

2. Etude de l'art

Plusieurs approches de détection de communautés se focalisent sur les modèles symétriques qui perdent la sémantique de l'orientation des liens entre les nœuds, un facteur clé distinguant les réseaux orientés de ceux non-orientés. Pour détecter les communautés dans les réseaux orientés, (Malliaros et Vazirgiannis, 2013) présentent des méthodes de transformation du graphe orienté en un graphe non orienté valué, permettant ainsi d'utiliser les concepts éprouvés ainsi que la complexité des modèles existants pour la détection des communautés dans les graphes non-orientés. Ainsi, pour mesurer la qualité de la partition obtenue, ils utilisent une fonction "objectif" parmi plusieurs, dont la plus répandue est la modularité. Cette mesure a pour but de caractériser la qualité d'une partition des sommets d'un graphe au regard de la densité des liens à l'intérieur des groupes et du nombre de liens entre groupes distincts, via la distribution des degrés des sommets. Plusieurs méthodes d'optimisation de la modularité ont été proposées, à l'instar de la méthode d'agglomération gloutonne de (Clauset *et al.*, 2004),et dont la plus répandue et la plus sûre étant celle de Louvain (Blondel *et al.*, 2008). Si elle a eu un succès dans la détection de communautés dans les graphes, il a néanmoins été montré que la modularité possède une limite de résolution (Fortunato et Barthelemy, 2007) qui restreint la possibilité de disposer de petites communautés qui soient bien définies, car plus la taille du graphe croit, plus la qualité de la partition décroit considérablement Il existe également des méthodes basées sur les marches aléatoires (Pons et Latapy, 2005) et (Rosvall et Bergstrom, 2008), consistant en la recherche d'une forme de description des nœuds et les liens, permettant de représenter les marches aléatoires. D'après ces auteurs, la description nécessitant le moins de mémoire via le taux de compression le plus élevé de la marche, est celui sélectionné. En 2010, divers modèles probabilistes de détection de communautés ont été

proposés (Malliaros et Vazirgiannis, 2013). Parmi eux, les modèles de bloc stochastiques semblent avoir eu le plus de succès en termes d'extraction de communautés sémantiques, avec des bonnes performances, et offrant des interprétations plausibles. Cependant, leur complexité en pratique parait énorme pour la raison selon laquelle au delà de 20 itérations, l'algorithme s'interrompt et les résultats deviennent invraisemblables. Pour pallier à cette limite de complexité, certains auteurs à l'instar de (Wang *et al.*, 2011) déterminent des kernels afin d'effectuer un traitement local de la détection de communautés avant de l'étendre au graphe tout entier.

Un kernel peut être assimilé à un ensemble de nœuds centraux ou influents à l'intérieur d'un groupe, appelés nœuds graines ou nœuds coeurs par (Kanawati, 2013). Un exemple typique d'algorithmes s'adaptant au type d'approche centrée-noeud sont ceux basés sur la propagation des labels (Raghavan *et al.*, 2007). Dans ce type d'approches, chaque vertex est initialisé par une étiquette ; elles définissent certaines règles simulant la propagation de ces étiquettes tel que l'établit le principe d'infection. La méthode de propagation de label possède l'avantage d'être asymptotiquement efficiente, mais aucune garantie n'est donnée sur la qualité de la partition, précisément dans les réseaux dans lesquels les communautés sont mal structurées. Certaines méthodes explorent le problème de détection de communautés dans les buts suivants : soit réduire le nombre d'itérations de réalisation des actions de l'algorithme, et par conséquent la complexité temporelle des algorithmes définis pour les grands réseaux, soit découvrir la communauté. (Wang *et al.*, 2011) identifie ces membres influents appelés kernel et ensuite propose un algorithme efficient pour déterminer la structure des communautés kernels. Lors de l'exécution de l'algorithme, le noeud initiateur du kernel est choisit aléatoirement parmi tous les nœuds du graphe, et la taille des communautés est fixée, ce qui mène à des résultats arbitraires de communautés. Pour pallier à cette limite, (Klymko *et al.*, 2014) a prouvé que les triangles jouent un rôle important dans la formation des réseaux complexes structurés et convertit un graphe orienté en un autre non-orienté et valué. Cette transformation, bien qu'efficace perd la sémantique portée par les liens au sein d'un réseau orienté. Nous proposons une méthode qui extrait les kernels via les triades et le voisinage des nœuds constituant les propriétés structurelles (orientées "pattern") dans les grands graphes réels.

3. Formalisation de la méthode

Nous proposons dans cette section le modèle à base de la communauté kernel et une définition des différents concepts y afférents, ainsi que les notations et formulations nécessaires, à base du modèle.

3.1. *Modèle de la communauté Kernel*

(Newman et Girvan, 2004) dans ses travaux initie l'étude des méthodes de détection de communautés basés sur la densité des liens entre les sommets d'un graphe ; ainsi une communauté dans son sens éthymologique correspond à un ensemble de

nœuds possédant le plus de relations entre eux qu'avec les autres nœuds du graphe. Cette définition typique de la notion de communautés est la plus répandue des méthodes de clustering dans les graphes non-orientés. Cependant celles-ci ne peuvent pas capturer des structures sémantiques, qui gardent le sens donné à l'orientation d'un lien entre les nœuds d'un graphe, contrairement aux méthodes de clustering pour lesquelles le critère cohésif de la mise en communauté des nœuds serait non pas la densité, mais la topologie accordée à une formulation bien définie de la notion de communauté. Plus précisément, les nœuds dans un graphe orienté pourraient être également groupés selon le critère de voisinage en commun (caractéristique sémantique ou structurelle) qu'ils possèdent et non pas seulement selon la densité de liens (caractéristique relationnelle) reliant les différents nœuds de cette communauté. Par exemple, le réseau de Co-citation signifiant qu'un ensemble de nœuds A, relié à un ensemble de nœuds B, implique une similarité entre les membres de chaque groupe, i.e. les nœuds de A possèdent un comportement similaire vis-à-vis des nœuds membres de B. Dans les graphes orientés, l'orientation des liens donne une impressionnante sémantique au graphe dans son ensemble, et au flux de circulation d'informations en particulier. La Figure 1 exhibe deux situations de structures représentant différents types de pattern orientés "densité en triades" d'une part et "4-cycles" d'autres part (tel que le présentent les zones d'ombre de la figure). Dans un réseau Twitter par exemple, la notion d'autorité est mise en exergue tel qu'illustré par la figure 2(a), à cause de la relation entre un ensemble de nœuds autoritaires appelées blogs Hub (nœuds u et v) et un ensemble de nœuds non populaires appelés "followers"(nœuds x) tel que présenté dans les Figures 2(b) et 2(c).

Ce concept d'autorité (ou de centralité) se traduit par l'optimisation de la notion kernel degree Clustering. La figure 2(a) est une visualisation d'un extrait du réseau de Twitter contenant deux kernels : d'une part les acteurs (Ashton Kutcher, Demi Moore, Opray Winfrey) et d'autre part les politiciens (Barack Obama, Al Gore). Ils constituent en d'autres termes les leaders alors que les nœuds situés à gauche de la figure correspondent à leurs fans ou followers, tel qu'étudié par (Gamgne et Tsopze, 2014). Les communautés kernel décrivent les nœuds possédant le même voisinage entrant (nœuds les plus connectés à un kernel et non pas à un autre). Nous considérons dans ce papier, les liens entrant vers les kernels pour exprimer la puissance de similarité qu'ils possèdent, conformément aux types de graphes qui y sont manipulés (réseau de citation) ; pour mieux illustrer cette formulation, le réseau de Twitter est structuré de pages hub "tweetées"ou aimées par un ensemble de visiteurs, et non l'inverse ; dans un réseau de Citation par exemple, les pionniers d'un domaine de recherche bien précis sont le plus cités par les chercheurs juniors. Initialement, un kernel est constitué d'un ensemble de nœuds centraux via leur degré entrant, obtenus par application de la notion de "triade ". Ce dernier constitue l'idée de base de cette approche, tout en s'inspirant de la notion d' "appartenance triadique" qui stipule que si deux amis ont un ami en commun, il est fort probable qu'ils soient du même groupe.

3.2. *Terminologie et concepts à base du modèle*

Étant donné un graphe $G(V, E)$ de $n = |V|$ sommets et $m = |E|$ liens. Soit Γ_u l'ensemble des voisins du noeud u. Nous définissons les notions et concepts à base de notre modèle :

Definition 1 *(Puissance de similarité). La puissance de similarité définit le critère ou le degré selon lequel deux ou plusieurs nœuds possèdent le plus grand nombre de voisins communs.*

Definition 2 *(Kernel). Un kernel correspond à un ensemble de nœuds possédant le plus grand nombre de voisins en commun. Ainsi, plus les nœuds d'un kernel possèdent des voisins en commun, plus la puissance de similarité qui les lie est important.*

Definition 3 *(Poids de la Triade). Le Poids de la triade pour chaque paire de nœuds (u, v) dans le graphe G peut être représenté par TW_{uv}. Nous utiliserons l'expression Δ_{uv} pour décrire le nombre de triades(cardinalité de triades) impliquant les nœuds u et v selon le schème présenté par les figures 2(b) et 2(c).*

$$TW_{uv} = \frac{|\Delta_{uv}|}{|\Delta_v|} \quad [1]$$

Où $|\Delta_v|$ correspond au nombre de triades impliquant le noeud v.

Definition 4 *(Chevauchement de voisinage). Étant donné deux nœuds u et v. Soit Γ_u l'ensemble des nœuds appartenant au voisinage du noeud u, soit Δ_v l'ensemble des nœuds appartenant au voisinage du nœud v. Notons NO_{uv} l'ensemble des nœuds voisins que u et v possèdent en commun.*

$$NO_{uv} = \frac{|\Gamma_v \cap \Gamma_u|}{|\Gamma_v \cup \Gamma_u| - \theta} \quad [2]$$

où θ peut prendre différentes valeurs fonction de la connectivité existant entre les nœuds u et v (0 lorsque les nœuds ne sont pas liés, 1 lorsqu'il existe un arc entre les nœuds, et 2 lorsqu'il existe une relation de réciprocité entre eux).

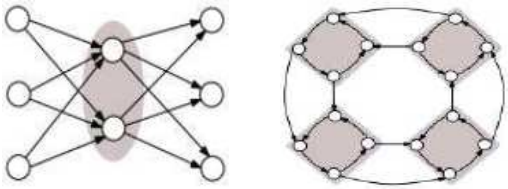
Definition 5 *(kernel Degree Clustering). Le Kernel Degree Clustering d'un couple de sommets u et v est défini par :*

$$KDC_{uv} = TW_{uv} * NO_{uv} \quad [3]$$

KDC_{uv} peut mesurer de manière particulière le degré de similarité du couple de nœuds (u, v) et de manière générale la puissance ou "force"(notion de similarité) d'un kernel à posséder des voisins en communs.

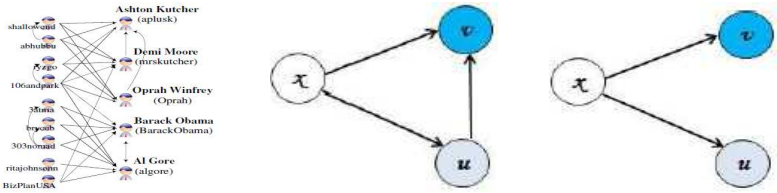
Definition 6 *(Communauté Kernel). La communauté kernel est un ensemble de nœuds possédant le plus grand voisinage commun, tel que ces voisins centrés autour du kernel par des liens entrant optimisent la mesure kernel Degree Clustering* KDC_{uv}.

Definition 7 *(Appartenance triadique). Cette notion stipule que si deux individus possèdent un ami en commun, alors il est très probable qu'ils fassent partie du même groupe (ou kernel) sans pour autant devenir absolument des amis.*



((a)) Cluster orienté Citations ((b)) Cluster orienté flot d'informations

Figure 1 – Exemples de clusters basés sur la topologie dans les graphes orientés. Le graphe de gauche (a) represente un cluster de pionniers d'un domaine de recherche donné dans un réseau de citation. Celui de droite (b) expose un graphe de 4 cycles dans un réseau de flots d'informations.



((a)) Illustration du réseau de Twitter ((b)) Triade fermée ((c)) Triade ouverte

Figure 2 – Structures à base du modèle de la communauté kernel.

4. Méthode d'extraction des communautés

La nouvelle approche de détection de communautés est structurée en deux étapes : l'identification des kernels qui sont les nœuds centraux de la communauté, et la migration des autres nœuds vers les kernels avec lesquels ils sont le plus liés. L'algorithme d'extraction des kernels, TRICA (Triads Cardinality Algorithm) que nous proposons ici fait usage du nouveau concept Kernel Degree Clustering (KDC), via l'optimisation de ce dernier, et pour lequel la valeur optimale détermine le degré de similarité des nœuds de la partition kernel. Cette mesure se base sur l'appartenance triadique

permettant d'exprimer la sémantique portée par les liens entre les différents membres d'une communauté, favorisant ainsi un accès certain à l'information à travers le réseau. Au lieu de mesurer la qualité de la partition entière de communautés comme le font (Clauset *et al.*, 2004) et (Blondel *et al.*, 2008) dans leur méthode, cette métrique s'applique aux kernels tout en effectuant une optimisation en local du graphe, tel que étudié par (Van Laarhoven et Marchiori, 2016). Nous nous attardons sur la cardinalité des triades communs aux vertex du kernel, correspondant au nombre de voisins en communs que ces derniers possèdent.

4.1. *Algorithme TRICA*

Ce paragraphe propose un algorithme évolutif, basé sur l'optimisation de la métrique KDC défini à la fois pour les graphes orientés et non-orientés. En s'inspirant des propriétés des réseaux réels, l'idée de base sous-jacente à cette métrique est la suivante : les nœuds d'un même kernel doivent favoriser une densité en triades plus importante dans les communautés dont ils sont le centre, en s'appuyant sur leur voisinage entrant. En effet la communauté engendrée par les nœuds centrés autour du kernel devrait contenir un nombre important de triades entre ses membres, et un nombre assez faible de triades entre les nœuds de ces communautés avec les autres nœuds extérieurs à la communauté. Ainsi, la qualité d'un kernel est défini comme la cohésion moyenne de chacun de ses membres avec les autres nœuds du kernel. La cohésion entre un vertex v et un ensemble de nœuds $u \in S$ dont la valeur du NO_{uv} est supérieure à un certain seuil ($\varepsilon = 0.5$ expérimentalement choisit et dont la valeur se verra discutée dans un prochain article) se définit par :

$$KDC_{uv}(u, v \in S) = TW_{uv} * NO_{uv}. \quad [4]$$

Le terme de gauche TW_{uv} calcule la proportion en triades à l'intérieur des kernels entre les nœuds pris deux à deux ; et celui de droite NO_{uv} détermine la proportion de voisinage en commun que deux nœuds d'un même kernel possèdent. Intuitivement, la métrique Kernel Degree Clustering mesure la force de similarité des membres d'un kernel. La qualité de la partition des kernels correspond à la qualité moyenne de chaque vertex dans son kernel. Ainsi, pour un ensemble S correspondant à un kernel, $KDC(S)$ se définit comme étant la moyenne $\forall x \in S$ de $KDC(x, S)$, et la valeur finale de KDC correspondant à la partition kernel $P = K_1, ..., K_n$ de K (l'ensemble des nœuds membres des kernels) formulé par :

$$KDC(P) = \frac{1}{|V|} \Sigma_{S \in P} \Sigma_{x \in S} KDC(x, S). \quad [5]$$

Supposons que le réseau à analyser est représenté par un graphe connexe, non-valué G de $n = |N|$ nœuds et $m = |E|$ liens . Cette étape d'identification des kernels se décompose en 4 sous-étapes comme suit :

1) Extraire une liste triée de nœuds centraux selon le critère de leur degré entrant, dans le graphe.

2) Déterminer le voisinage commun de chaque couple (u,v) par le biais d'une variante du coefficient de Jaccard(Fortunato, 2010) tel que décrit par NOuv dans la formule 2

3) Déterminer le poids des triades TW_{uv} (tel que décrit par la formule 1) pour chaque u, v dont le $NO_{uv} \geq \varepsilon$.

4) Calculer la moyenne des KDC_{uv} pour chaque couple (u, v) de la liste des nœuds éligibles à appartenir à un kernel.

5) Stocker v dans le kernel pour lequel KDC_{uv} est optimal.

Ces différentes étapes se répètent jusqu'à l'obtention d'une valeur de KDC_{uv} optimale.

La première étape consiste en la détermination d'une liste de nœuds triée par leur degré entrant. L'opération de tri permet de simplifier la réalisation de la deuxième étape consistant à supprimer aisément de cette liste les nœuds possédant moins de deux voisins, car ce type de nœuds serait probablement disqualifié dans l'idée de faire partie des nœuds centraux dans les kernels, à cause de leur degré entrant variant entre 1 et 0.

Après extraction de cette liste triée et épurée de nœuds classés par ordre décroissant de leur valeur de degré entrant, suit le calcul des NO_{uv} pour chaque couple de nœuds (u, v) pour déterminer ceux des nœuds éligibles à faire partie des kernels, correspondant ainsi aux nœuds dont le couple a pour valuation de NO_{uv} une valeur supérieure à un seuil ε.

L'étape de calcul des poids des triades quant à elle se décrit de la manière suivante : d'une part, il est question primo de compter pour chaque couple de nœuds (u, v) le nombre total de triades dans lesquels ils sont impliqués dans le graphe, secundo de supprimer tous les nœuds n'appartenant à aucune triade ; d'autre part l'on compte pour chaque vertex v le nombre total de triades dans lesquels il est impliqué (Δ_v). Cette phase de filtrage aide à l'amélioration des performances et permet de simplifier les hypothèses dans les choix futurs d'un noeud quelconque, dans sa décision à passer d'une communauté à une autre. Notons que ces deux valeurs sont des constantes pendant le processus de détermination des kernels et peuvent être calculées simultanément. Étant donné deux nœuds u et v, une manière classique de dénombrer les triades dans lesquels ils sont impliqués consiste en l'intersection de leur liste d'adjacence en vue de compter leur voisins en communs. Si les nœuds u et v ne possèdent aucun voisin en commun, ils sont mentionnés comme devant appartenir à des communautés distinctes dans la partition résultante du graphe, car ce type de nœuds n'affecte pas la détermination de KDC. Pour dénombrer tous les triades impliquant v, le précédent processus effectué pour les couples, est appliqué pour chaque voisin u de v, v étant le noeud central de degré entrant maximal.

La 5^e étape et la plus importante de ce processus de clustering des kernels consiste en l'optimisation de la mesure KDC_{uv}. L'idée de base sous-jacente au calcul de KDC_{uv} est celle permettant à chaque vertex de mettre à jour de manière répétée les kernels, via une heuristique d'amélioration, tout en évaluant l'ensemble des KDC_{uv}

entre chacune des mises à jour ; et après un certain nombre pré-spécifié d'étapes pour lesquelles KDC_{uv} ne croit plus jusqu'à un certain seuil, le processus s'interrompt. Cette heuristique basée sur le calcul de la moyenne des KDC_{uv} sur l'ensemble des nœuds u, v pour lesquels $NO_{uv} \geq \varepsilon$ permet de fixer une marge dans laquelle l'optimisation de cette métrique se verra varier. En effet, pour une valeur donnée de KDC_{uv} inférieure à la borne inférieure de cette marge (moyenne des KDC_{uv}), certes en deçà de la valeur optimale (borne supérieure), u se verra supprimé du kernel courant pour etre un noeud non-kernel. Sinon u restera dans son kernel courant. Et il migrera du kernel courant vers un autre kernel pour lequel la valeur optimale est atteinte. En fait, après avoir initialisé le kernel par un noeud central v, la combinaison d'autres nœuds u du graphe avec v via le calcul de KDC_{uv} peut conduire aux deux états ci-dessous :

– Migrer : Le vertex migre d'un kernel vers le kernel d'un des nœuds parmi ceux centraux, situé dans son voisinage le plus proche.

– Rester : Le vertex demeure dans son kernel.

Dans le but d'améliorer les performances de l'approche, le vertex doit choisir parmi les actions ci-dessus, celle qui conduirait à l'obtention d'une meilleure valeur optimale de KDC_{uv}. Le pseudo-code associé à TRICA est présenté dans l'algorithme 1.

Algorithm 1 Implementation de la méthode d'Extraction des kernels TRICA

Entrées: Graphe orienté $G = (V, E)$
Sorties: K Kernels
1: Initialisation : $K \leftarrow \emptyset$
2: $L = Sort(v/d^{in}(v) = max\{d^{in}(t), \forall t \in V\})$;
3: Calculer NO_{uv} et TW_{uv} pour chaque $(u, v) \in V$ tel que $NO_{uv} > \varepsilon$;
4: Calculer la moyenne(KDC_{uv})
5: **pour** Chaque $u \in L$ **faire**
6: $v = argmax\{d^{in}(t), \forall t \in L\}$;
7: $KDC^* \leftarrow Moyenne(KDC_{uv})$;
8: **répéter**
9: Calculer KDC_{uv}
10: **Si** $KDC_{uv} > KDC^*$ **alors**
11: $S \longleftarrow S \cup u$;
12: **Fin si**
13: $KDC^* \longleftarrow KDC_{uv}$
14: **jusqu'à** $KDC_{uv} < KDC^*$
15: $K \longleftarrow K \cup S$
16: **fin pour**
17: Retourner K ;

Algorithm 2 Pseudo code de l'algorithme de migration des nœuds non-kernels

Entrées: Communautés kernels $K = K_1, K_2, ..., K_t$
Sorties: Communautés globales $G_k = G_{k1}, G_{k2}, ..., G_{kt}$
1: $\forall i \in 1, ..., t, G_{ki} \leftarrow \emptyset$
2: **répéter**
3: $\forall i \in 1, ..., t, R_i \leftarrow K_i \cup G_{ki}$
4: **pour** $i \leftarrow 1$ to t **faire**
5: $S \leftarrow v \notin \cup R_i | \forall j \in 1, ..., t,$
6: $|Connexions(v, R_i)| \geq |Connexions(v, R_j)| > 0$
7: $G_{ki} \leftarrow G_{ki} \cup S$
8: **fin pour**
9: **jusqu'à** Plus de nœuds non-kernels
10: Retourner G_k ;

4.2. *Déduction des communautés globales*

Après l'extraction des communautés kernel, il est question de faire migrer les nœuds non-kernels, ceux n'appartenant à aucun kernel, vers les kernels avec lesquels ils sont le plus liés, via l'orientation "entrante"des liens de ce noeud. Ces nœuds non-kernel migreront vers les kernels et formeront ainsi des "Communautés globales". Le processus de génération des communautés globales (communautés contenant à la fois les nœuds kernels et non-kernels) consiste en l'exécution des étapes suivantes : initiallement, on étiquette chaque noeud non-kernel comme étant non-associé. Pour chaque noeud non-associé, le ranger dans le kernel avec lequel il possède le plus grand nombre de connexions ; le kernel change ainsi d'état pour devenir une communauté globale grandissante. Ce processus est repété jusqu'à ce qu'il n'y ait plus de nœuds non-kernel, tel que décrit par l'algorithme 2.

5. Expérimentation et évaluation de la nouvelle méthode

5.1. *Description des méthodes et des jeux de données*

Afin de montrer la performance de cette approche, l'évaluation empirique s'est focalisé sur une comparaison entre les résultats produits par plusieurs autres méthodes de l'état de l'art parmi lesquels : Walktrap (Pons et Latapy, 2005), Edge Betweenness (Newman et Girvan, 2004), Label Propagation (Raghavan *et al.*, 2007) et Louvain (Krings et Blondel, 2011). Notre méthode peut s'appliquer autant aux graphes orientés que non-orientés. L'expérimentation s'est appuyée sur trois niveaux d'évaluation : le premier concerne la densité en triades ou triad cardinality rate (TCR), dans les communautés résultantes, le second se base sur les valeurs de la modularité dans les partitions obtenues par chacune des méthodes, et la dernière sur le nombre final de communautés, en s'appuyant sur l'intuition selon laquelle plus une partition possède

Tableau 1 – Caractéristiques des jeux de données

Jeu de données	Nombre de nœuds	Nombre de liens
Extrait de Twitter	14	32
Celegansneural	297	2,345
Polblogs	1,490	19,090
Citeseer	3,327	4,732

de communautés, moins elle est dense (en triades). TCR correspond au pourcentage de triades dans la partition toute entière, tel que définit dans la formule ci-dessous.

$$TCR = \frac{\Sigma_i |\Delta_i|}{|\Delta|} \quad [6]$$

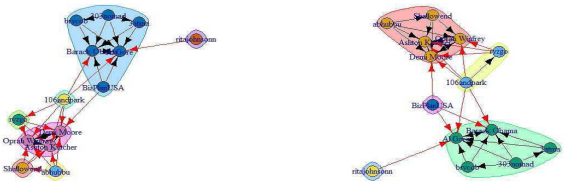
où i correspond à une communauté quelconque et $|\delta|$ le nombre de triades dans le graphe tout entier.

5.2. *Evaluation de la performance des méthodes de détection*

Les performances des différentes méthodes de détection de communautés sur les quatre jeux de données sont respectivement présentées dans les tableaux 2, 3, 4 et 5.

Walktrap détermine la distance (l'homogénéité) entre les communautés et fusionne celles qui sont moins distantes pour produire une nouvelle communauté résultante. L'idée de Walktrap et celle de la nouvelle approche sont semblables dans la mesure où elles ont en commun la notion de fusion des groupes de nœuds voisins (l'un sur la base de la distance, et l'autre sur la base du nombre de voisins en commun) ; ainsi les résultats des deux méthodes sont dans l'ensemble convergentes, tel que présenté ci-dessous : Dans le réseau extrait de Twitter présenté dans le Tableau 2, TRICA et Walktrap obtiennent la même valeur de TCR, soit 0.6428, mais la valeur de la modularité 0.401 obtenue par Walktrap est plus petite que celle obtenue par TRICA, soit 0.410. Les deux méthodes découvrent le même nombre 2 de communautés. La méthode Louvain détecte également 2 communautés, avec une veleur plus petite de la modularité égale à 0.395. Cependant, les méthodes Label Propagation et Edge Betweenness découvrent respectivement 5 et 7 communautés avec des faibles taux de triades ainsi que de valeurs de modularité, ce qui traduit l'insuffisance de ces approches sur l'idée de clusteriser les nœuds appartenant au meme voisinage d'un ensemble de nœuds kernels.

En ce qui concerne le jeu de données Celegansneural network : la méthode Edge Betweenness détermine le lien de centralité d'intermédiarité maximale, c'est à dire celui traversé par le plus grand nombre de géodésiques (plus courts chemins) et se charge de supprimer ce lien, et de façon récursive obtient des communautés. Elle détecte le plus grand nombre de communautés (194), avec un faible taux de triades dans



((a)) Partition Edge-betweeness ((b)) Partition Label Propagation

((c)) Partition Walktrap, Louvain ((d)) Partition Kernel-Pattern

Figure 3 – Visualisation des partitions obtenues par les différentes approches.

Figure 4 – Partition Walktrap sur le réseau Polblogs, avec des nœuds bruits appelés outliers.

la partition toute entière (0.0857). Contrairement aux autres méthodes, la nouvelle approche détermine le nombre de communautés attendues (5), car en tant que benchmark, (Klymko *et al.*, 2014) obtiennent cette meme valeur. Par ailleurs, avec un taux élevé de TCR égal à 0.3211), ceci démontre la performance de la nouvelle méthode sur la qualité de la partition résultante, tel que présenté par le Tableau 3.

La méthode Propagation de label consiste à faire déplacer un noeud d'une communauté vers une autre si ses voisins appartiennent à cette communauté de destination. De cette manière, pour le jeu de données polblogs, elle détecte le plus grand nombre de communautés (244) avec le plus faible taux de triades (0.0026). Cependant les mé-

Tableau 2 – Performance des méthodes de détection de communautés sur le réseau Extrait de Twitter, où les meilleures performances sont en gras.

Méthode	TCR	Modularité	Communautés
Edge-Betweenness	0.0857	0.187	7
Walktrap	**0.6428**	0.401	**2**
Label Propagation	0.34	0.306	5
Louvain	0.531	0.395	**2**
Kernel-pattern	**0.6428**	**0.410**	**2**

thodes Walktrap et Kernel-pattern produisent de meilleurs résultats sur tous les critères d'évaluation soit 12 et 34 communautés respectivement, avec les valeurs maximales de TCR (0.67 et 0.5732 respectivement) et les meilleures valeurs de modularité, soit 0.4302 et 0.429 respectivement, tel que présentés dans le Tableau 4. Ces résultats indiquent que les modèles basés sur l'idée de "communauté croissante centrée autour de kernels" est d'une manière ou d'une autre en accord avec la notion d'optimisation de la mesure Kernel Degree Clustering ; en effet, le nombre minimum 12 de communautés avec une valeur de TCR maximale de 0.67 est preuve que la topologie en triades via la méthode Walktrap est la mieux structurée. Cependant, cette dernière ne saurait être meilleure, à cause des données bruitées représentées par les deux nœuds **singletons**, tel que présenté dans la Figure 4, indiquant que la méthode capture des nœuds que (Ester et al., 1996) appelle "outliers", qui constituent des nœuds anormaux ou **bruits** de la partition.

Bien que la méthode Louvain produise la plus grande valeur de modularité (0.886) pour le corpus Citeseer tel que présenté dans le Tableau 5, son nombre de communautés est plus important que celui produit par la nouvelle approche. Ainsi, la méthode Louvain détermine une valeur de TCR (0.213) moins importante que celle obtenue par l'algorithme Kernel-pattern (0.407), ce qui montre la validité de la nouvelle approche sur la sémantique des liens. Citeseer, contrairement aux autres corpus, suit une distribution de la loi de puissance exponentielle, dû au fait qu'il s'agisse d'un réseau de citation dans lequel l'on pourrait être en possession de nœuds de centralité de degré plus importante que les autres nœuds (l'on parle de nœuds "hub"). C'est ce qui expliquerait la valeur nulle des TCR pour les trois méthodes du tableau 5.

Figure 3 permet de visualiser la plausibilité de la méthode d'extraction des communautés sur la base des kernels, TRICA sur le jeu de données extrait de Twitter.

6. Conclusion

Dans ce document, nous nous sommes focalisés sur le problème d'extraction de communautés basé sur les kernel-pattern, une communauté se ramenant à un ensemble de nœuds centrés autour d'un sous groupe de nœuds graines, initiateurs de la commu-

Tableau 3 – Performance des méthodes de détection de communautés sur le réseau Celegansneural, où les meilleures performances sont en gras.

Méthode	TCR	Modularité	Communautés
Edge-Betweenness	0.0004	0.081	194
Walktrap	0.0458	0.363	21
Label Propagation	0.0135	0.0027	29
Louvain	0.2951	**0.398**	6
Kernel-pattern	**0.3211**	0.359	**5**

Tableau 4 – Performance des méthodes de détection de communautés sur le réseau Polblogs, où les meilleures performances sont en gras.

Méthode	TCR	Modularité	Communautés
Edge-Betweenness	0.0064	0.1872	55
Walktrap	**0.67**	**0.4302**	**12**
Label Propagation	0.0026	0.386	244
Louvain	0.1289	0.427	276
Kernel-pattern	**0.5732**	**0.429**	**34**

nauté, possédant quasiment le même voisinage commun. Un kernel correspond ainsi à un outil favorisant la compréhension du rôle et de la structure d'un réseau. Nous avons principalement orienté ce travail dans l'extraction des kernels qui sont considérés comme étant des nœuds influents du réseau. La nouvelle approche proposée se base sur l'optimisation de la mesure Kernel Degree Clustering (KDC) qui définit la puissance de similarité existant entre nœuds d'un même kernel, via la notion de triade représentant les caractéristiques structurelles des grands réseaux réels. Les expérimentations sur la nouvelle approche prouvent que la méthode Kernel-pattern permet de détecter les communautés efficaces attendues, et réalise de meilleurs valeurs de

Tableau 5 – Performance des méthodes de détection de communautés sur le réseau Citeseer, où les meilleures performances sont en gras.

Méthode	TCR	Modularité	Communautés
Edge-Betweenness	0.0	0.5344	738
Walktrap	0.0	0.811	593
Label Propagation	0.0	0.491	842
Louvain	0.213	**0.886**	466
Kernel-pattern	**0.407**	0.707	**121**

qualité des communautés, par rapport à certaines méthodes de l'état de l'art. Cependant, elle ne s'applique pas aux graphes valués. Nos travaux futurs consisteront ainsi à prendre en compte cette propriété de valuation des graphes orientés et s'attardera sur la programmation en parallèle, afin d'améliorer la complexité du modèle.

7. Bibliographie

Blondel V. D., Guillaume J.-L., Lambiotte R., Lefebvre E., « Fast unfolding of communities in large networks », *Journal of statistical mechanics : theory and experiment*, vol. 2008, n° 10, p. P10008, 2008.

Clauset A., Newman M. E., Moore C., « Finding community structure in very large networks », *Physical review E*, vol. 70, n° 6, p. 066111, 2004.

Fortunato S., « Community detection in graphs », *Physics reports*, vol. 486, n° 3, p. 75-174, 2010.

Fortunato S., Barthelemy M., « Resolution limit in community detection », *Proceedings of the National Academy of Sciences*, vol. 104, n° 1, p. 36-41, 2007.

Gamgne D. F., Tsopze N., « Communautes et roles dans les reseaux sociaux », *Proceedings of the 12th Conference Africaine sur la Recherche en Informatique et Mathematiques appliquees*, vol. 12, n° 1, p. 122-188, 2014.

Kanawati R., « Détection de communautés dans les grands graphes d'interactions (multiplexes) : état de l'art », 2013.

Klymko C., Gleich D., Kolda T. G., « Using triangles to improve community detection in directed networks », *arXiv preprint arXiv :1404.5874*, 2014.

Krings G., Blondel V. D., « An upper bound on community size in scalable community detection », *arXiv preprint arXiv :1103.5569*, 2011.

Malliaros F. D., Vazirgiannis M., « Clustering and community detection in directed networks : A survey », *Physics Reports*, vol. 533, n° 4, p. 95-142, 2013.

Newman M. E., Girvan M., « Finding and evaluating community structure in networks », *Physical review E*, vol. 69, n° 2, p. 026113, 2004.

Pons P., Latapy M., « Computing communities in large networks using random walks », *International Symposium on Computer and Information Sciences*, Springer, p. 284-293, 2005.

Raghavan U. N., Albert R., Kumara S., « Near linear time algorithm to detect community structures in large-scale networks », *Physical review E*, vol. 76, n° 3, p. 036106, 2007.

Rosvall M., Bergstrom C. T., « Maps of random walks on complex networks reveal community structure », *Proceedings of the National Academy of Sciences*, vol. 105, n° 4, p. 1118-1123, 2008.

Van Laarhoven T., Marchiori E., « Local network community detection with continuous optimization of conductance and weighted kernel k-means », *Journal of Machine Learning Research*, vol. 17, n° 147, p. 1-28, 2016.

Wang L., Lou T., Tang J., Hopcroft J. E., « Detecting community kernels in large social networks », *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, IEEE, p. 784-793, 2011.