HAL open science

Une méthode pour caractériser les communautés des réseaux dynamiques à attributs

Günce Keziban Orman, Vincent Labatut, Marc Plantevit, Jean-François Boulicaut

To cite this version:

Günce Keziban Orman, Vincent Labatut, Marc Plantevit, Jean-François Boulicaut. Une méthode pour caractériser les communautés des réseaux dynamiques à attributs. 14ème Conférence Extraction et Gestion des Connaissances (EGC), Jan 2014, Rennes, France. pp.101-112. hal-00918181

HAL Id: hal-00918181

https://hal.science/hal-00918181v1

Submitted on 13 Dec 2013

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.

Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

Une méthode pour caractériser les communautés des réseaux dynamiques à attributs

Günce Keziban Orman*,** Vincent Labatut** Marc Plantevit*** Jean-François Boulicaut*

*Université de Lyon, CNRS, INSA-Lyon, LIRIS UMR5205, F-69621, France
jean-francois.boulicaut@insa-lyon.fr,
**Université Galatasaray, Département d'informatique, Ortaköy 34349, Istanbul, Turquie
korman@gsu.edu.tr, vlabatut@gsu.edu.tr
*** Université de Lyon, CNRS, Université Lyon 1, LIRIS UMR5205, F-69622, France
marc.plantevit@liris.cnrs.fr

Résumé. De nombreux systèmes complexes sont étudiés via l'analyse de réseaux dits complexes ayant des propriétés topologiques typiques. Parmi celles-ci, les structures de communautés sont particulièrement étudiées. De nombreuses méthodes permettent de les détecter, y compris dans des réseaux contenant des attributs nodaux, des liens orientés ou évoluant dans le temps. La détection prend la forme d'une partition de l'ensemble des nœuds, qu'il faut ensuite caractériser relativement au système modélisé. Nous travaillons sur l'assistance à cette tâche de caractérisation. Nous proposons une représentation des réseaux sous la forme de séquences de descripteurs de nœuds, qui combinent les informations temporelles, les mesures topologiques, et les valeurs des attributs nodaux. Les communautés sont caractérisées au moyen des motifs séquentiels émergents les plus représentatifs issus de leurs nœuds. Ceci permet notamment la détection de comportements inhabituels au sein d'une communauté. Nous décrivons une étude empirique sur un réseau de collaboration scientifique.

1 Introduction

Un *réseau complexe* est la représentation d'un *système complexe* sous forme de graphe. Ils sont devenus très populaires en tant qu'outil de modélisation durant la dernière décennie car ils permettent de mieux comprendre le fonctionnement et la dynamique de certains systèmes . Un réseau complexe *ordinaire* ne contient que des nœuds et les liens existant entre eux ; cependant il est possible de l'enrichir avec différents types de données : orientation et/ou poids des liens, dimension temporelle, attributs associés aux nœuds ou aux liens, etc. Cette souplesse a permis d'utiliser les réseaux complexes pour étudier les systèmes du monde réel dans de nombreux domaines : sociologie, physique, génétique, informatique, etc. (Newman, 2003).

La nature complexe des systèmes modélisés entraine la présence de propriétés topologiques non-triviales au sein des réseaux correspondants. Parmi celle-ci, la *structure de communautés* est l'une des plus répandue et des plus étudiées. Informellement, on peut définir une communauté comme un groupe de nœuds densément interconnectés relativement au reste du réseau

(Newman, 2003). Cependant, dans la littérature, cette notion est formalisée de très nombreuses différentes façons (Fortunato, 2010). Il existe en fait des centaines d'algorithmes destinés à détecter les structures de communautés, caractérisés par l'utilisation d'une définition et/ou d'un traitement différents. Certains sont basés sur la mesure de modularité, une mesure similarité entre nœuds, le principe de compression des données, la notion de signification statistique, les mécanismes de diffusion de l'information, la percolation de cliques, etc. (cf. (Fortunato, 2010) pour une revue détaillée). La plupart des méthodes existantes traitent des réseaux ordinaires mais de nouvelles méthodes apparaissent pour analyser les réseaux les plus riches en exploitant les directions et poids des liens, puis le temps, et plus récemment les attributs des nœuds (Ruan et al., 2013; Tian et al., 2008; Zhou et al., 2009). Ces dernières se concentrent sur la recherche de groupes de nœuds denses en termes de liens, et dont les attributs sont homogènes. Même si cela n'est pas toujours indiqué explicitement, ces méthodes exploitant les attributs s'appuient sur l'hypothèse que les nœuds d'une même communauté doivent être similaires en termes d'attributs. Bien que les algorithmes diffèrent en termes de nature des communautés détectées, de complexité algorithmique, de qualité du résultat et d'autres aspects (Fortunato, 2010), leur production peut toujours être essentiellement décrite comme une liste de groupes de nœuds. Plus précisément, dans le cas de communautés mutuellement exclusives, il s'agit d'une partition de l'ensemble des nœuds. D'un point de vue applicatif, la question est alors de donner un sens à ces groupes relativement au système étudié. L'interprétation manuelle de petites communautés est possible, mais la méthode ne s'applique pas bien à de très grands réseaux.

Seuls quelques travaux ont essayé de s'attaquer explicitement à ce problème. Dans (Lancichinetti et al., 2010), les communautés sont caractérisées en comparant les distributions de plusieurs mesures purement topologiques. Dans (Tumminello et al., 2011) et (Labatut et Balasque, 2012), les auteurs se concentrent sur les attributs nodaux, et identifient les plus représentatifs pour chaque communautés au moyen de divers outils statistiques classiques. Les méthodes de détection de communautés exploitant les attributs sont généralement aussi en mesure de donner ce type d'information, car ces attributs sont identifiés pendant le processus de détection. Cependant, aucune des méthodes citées ne prend en compte toutes les données qu'un réseau riche peut contenir (attributs, topologie et dimension temporelle), qui plus est de manière systématique. Il existe donc un besoin pour un tel procédé, qui permettrait de caractériser les communautés des réseaux complexes riches. Dans ce travail, nous proposons une solution à ce problème, sous la forme d'une méthode d'analyse des réseaux attribués dynamiques. Pour cela, nous détectons les changements communs dans les mesures topologiques et les valeurs d'attribut sur une période de temps donnée. Plus précisément, nous cherchons à trouver les motifs séquentiels fréquents les plus représentatifs pour chaque communauté. Ces motifs peuvent ensuite être utilisées à la fois pour caractériser la communauté, et pour identifier ses anomalies, i.e. ses nœuds ayant un comportement non-standard. Les motifs fréquents représentent la tendance générale des nœuds dans la communauté considérée, alors que les anomalies peuvent correspondre à des nœuds ayant un rôle spécifique dans la communauté, ou situés à sa frontière. Nous illustrons notre méthode en l'appliquant à un réseau dynamique de co-auteurs extrait de la base de données bibliographiques DBLP [1].

Notre première contribution est de considérer la caractérisation de communauté comme un problème spécifique, distinct de celui de la détection de la communauté. La méthode à

1. http://www.informatik.uni-trier.de/~ley/db/

appliquer doit être indépendante de la technique utilisée pour détecter les communautés, se fonder sur une approche systématique facilement reproductible, et être la plus automatisée possible. Notre deuxième contribution est l'introduction d'une nouvelle représentation des réseaux attribués dynamiques. Elle prend la forme d'une base de données contenant des séquences de mesures topologiques, d'attributs nodaux et d'information communautaire, pour plusieurs tranches temporelles. Ce type de représentation avait précédemment été utilisé pour la représentation de données naturelles (Mabroukeh et Ezeife, 2010), mais pas celle de graphes. Notre troisième contribution est la définition d'une méthode basée sur l'extraction sous contraintes de motifs séquentiels qui tire parti de cette représentation pour caractériser les communautés. Enfin, notre dernière contribution concerne une application à un réseau du monde réel. Dans la section suivante, nous donnons une description détaillée de notre méthode. Dans la section 3, nous présentons nos résultats expérimentaux obtenus sur les données DBLP. La section 4 décrit les travaux connexes, et la section 5 présente les extensions possibles de notre travail.

2 Méthode

Un réseau dynamique attribué est constituée de différentes tranches temporelles, chacune représentée par un sous-réseau distinct, contenant les liens entre les nœuds pour un intervalle de temps donné. Ces tranches temporelles sont séquentielles. Habituellement, les nœuds et leurs attributs sont les mêmes pour chaque tranche, tandis que les liens entre eux et les valeurs des attributs peuvent changer. Nous proposons de caractériser les communautés d'un réseau dynamique attribué en fonction de l'évolution commune des mesures topologiques et des attributs de leurs nœuds. Le processus que nous proposons inclut 4 étapes. La première consiste à identifier une structure de communautés de référence. La seconde vise à créer la structure de données permettant une représentation séquentielle des mesures topologiques et des attributs des nœuds. Nous calculons d'abord les valeurs de toutes les mesures topologiques sélectionnées, puis nous discrétisons les valeurs des mesures et des attributs. Lors de la troisième étape, nous recherchons des motifs séquentiels fréquents et nous extrayons les nœuds qui supportent chaque motif. La quatrième étape consiste à choisir les motifs les plus représentatifs pour caractériser les communautés selon différents critères.

Détection des communautés. Pour détecter comment les nœuds évoluent en fonction de l'appartenance communautaire, nous avons d'abord besoin d'une structure de communautés de référence. Il serait possible d'appliquer une méthode dynamique, cependant cela entraîne des complications dues aux fusions, séparations, disparitions et apparitions de communautés au cours du temps. Pour cette raison, dans ce premier travail, nous avons décidé d'utiliser des communautés statiques, détectés sur une version intégrée du réseau.

Nous créons d'abord un nouveau réseau en agrégeant tous les liens dans le temps. Un poids est attribué à chaque lien en fonction de son nombre d'occurrences, afin de représenter sa stabilité dans le temps. Nous appliquons ensuite un algorithme de détection de communautés classique pour identifier nos communautés statiques. À cette fin, nous avons sélectionné Louvain (Blondel et al., 2008), qui est une méthode reconnue pour obtenir des structures de communautés de bonne qualité. La complexité temporelle de cette algorithme est $O(n \log n)$ où n est nombre de nœuds. La structure de communautés en résultant est utilisée dans le reste

de notre analyse. Bien que les communautés soient statiques, les changements dans la structure du réseau seront néanmoins considérés lors du traitement des mesures topologiques des nœuds.

Constitution de la base de données. La deuxième étape consiste à représenter le réseau d'une manière appropriée pour l'extraction des motifs séquentiels fréquents. Un réseau dynamique attribué $\mathcal{G} = \langle G_1, \ldots, G_\theta \rangle$ est constitué d'une séquence de tranches temporelles, chacune prenant la forme d'un réseau attribué séparé G_j $(1 \leq j \leq \theta)$, représentant une période de temps donnée. Chaque réseau statique G_j contient n nœuds et correspond aux connexions entre ces nœuds pour la période j. Les nœuds et leurs attributs disponibles sont supposés être statiques, c'est à dire rester les mêmes pour tout G_j. Au contraire, les valeurs des attributs et la structure du réseau peuvent changer au cours du temps. Un descripteur de nœud est soit un attribut nodal, soit une mesure topologique nodale.

Nous avons sélectionné les mesures topologiques nodales les plus répandues (degré, transitivité locale), ainsi que certaines mesures permettant de caractériser les nœuds en termes de position dans leur communauté (degré interne, coefficient de participation et enchâssement). Le *degré* d d'un nœud est son nombre total des voisins directs et sob *degré interne* d_{int} est le degré calculé dans sa seule communauté. La *transitivité locale*, $T = \triangle/(d(d-1))$, d'un nœud est la densité des triangles auxquels il appartient. Plus précisément, il s'agit du rapport entre le nombre de triangles auxquels le nœud appartient effectivement ($\triangle$), et le nombre maximal de triangles auxquels il pourrait appartenir, étant donné son degré (dénominateur). Le *degré interne normalisé* z et le *coefficient de participation* P sont deux mesures définies par Guimerá et Amaral (2005). La première exprime combien le nœud est connecté à sa propre communauté, relativement aux autres nœuds de sa communauté. Il s'agit du z-score de son degré interne, i.e. du nombre de voisins directs dans sa propre communauté. Un *hub* au sens de Guimerá et Amaral (2005) est un nœud dont le degré interne est élevé relativement au autres nœuds de sa communauté. La seconde mesure caractérise la distribution des voisins du nœud parmi toutes les communautés : $P = 1 - \sum_c (d_c/d)^2$, où d_c est le nombre de liens entre le noeud considéré et la communauté c. Il s'agit plus précisément de mesurer l'hétérogénéité de cette distribution. On obtient une valeur proche de 1 si tous les voisins sont répartis uniformément entre toutes les communautés, et 0 s'ils sont tous réunis dans la même communauté. L'*enchâssement*, $e = d_{int}/d$, évalue à quel point les voisins d'un nœud appartiennent à sa propre communauté (Lancichinetti et al., 2010). À la différence du degré interne normalisé, l'enchâssement est donc normalisé par rapport au nœud, et non pas à sa communauté. Toutes ces mesures topologiques peuvent être calculées en un temps linéaire par rapport à n.

Soit $D = \{D_1, D_2, \ldots, D_k\}$ l'ensemble des descripteurs. Chacun peut prendre plusieurs valeurs discrètes, définies sur son *domaine* $\mathcal{D}_i$ $(1 \leq i \leq k)$. Toutes nos mesures topologiques sont à valeurs réelles, nous avons donc dû les discrétiser pour les adapter à cette définition. Un *item* $I = \{(D_i, u) \in D \times \mathcal{D}_i\}$ est un couple constitué d'un descripteur D_i et d'une valeur u de son domaine $\mathcal{D}_i$. Une séquence nodale est notée $\nu = \langle (l_{11}, \ldots, l_{k1}), \ldots, (l_{1\theta}, \ldots, l_{k\theta}) \rangle$ où l_{ij} est l'item du i^{eme} descripteur pour la tranche temporelle j. Une base de données M est l'ensemble des tuples (η, ν_η, C_η) où η est l'identifiant du nœud, ν_η est sa séquence et C_η est l'étiquette de sa communauté (tels que définis à l'étape 1). La création de M revient à réaliser séquentiellement le calcul des mesures topologiques et la détection de communautés, donc sa complexité temporelle est aussi $O(n \log n)$.

Fouille de motifs séquentiels émergents. La troisième étape a pour but de trouver des motifs séquentiels émergents permettant de caractériser les communautés. Pour chaque communauté, nous recherchons les motifs séquentiels supportés par la majorité de ses membres. Avant de présenter la façon dont nous traitons la fouille, nous donnons quelques définitions nécessaires à la bonne compréhension de l'algorithme que nous avons utilisé.

Un *itemset* T est un sous-ensemble de I. Une *séquence* $s = \langle t_1, \ldots, t_m \rangle$ est une liste d'itemsets ordonnés dans le temps, où m est la longueur de la séquence. Une séquence $\alpha = \langle a_1, \ldots, a_m \rangle$ est une *sous-séquence* d'une autre séquence $\beta = \langle b_1, \ldots, b_n \rangle$ ssi $\exists i_1, i_2, \ldots, i_m$ tels que $1 \leq i_1 < i_2 < \ldots < i_m \leq n$ et $a_1 \sqsubseteq b_{i1}, a_2 \sqsubseteq b_{i2}, \ldots, a_m \sqsubseteq b_{im}$ On dit également que β est une *super-séquence* de α. La taille d'une communauté C est le nombre total de nœuds qu'elle contient. Le support d'une séquence s pour une communauté donnée C est le rapport du nombre de nœuds supportant s à la taille de C : $sup(s, C) = |\{\eta \in C | s \sqsubseteq \nu_\eta\}|/|C|$. Le taux de croissance d'une séquence s pour une communauté donnée C est le rapport du support de s dans C au support de s dans $\overline{C}$, où $\overline{C}$ est le complémentaire de C dans le réseau (i.e. tous les nœuds du réseau sauf ceux de C) : $Gr(s, C) = sup(s, C)/sup(s, \overline{C})$. Le taux de croissance mesure l'émergence de la séquence s dans la communauté C. Plus il est élevé, et plus la séquence s est caractéristique de la communauté C.

Compte tenu d'un seuil de support minimal min_{sup}, un motif séquentiel fréquent est une séquence dont le support est supérieur ou égal à min_{sup}. Un motif séquentiel fréquent *fermé* pour une communauté donnée est un motif séquentiel qui n'a pas de super-séquence pour le support minimal spécifié. Dans notre problème, nous caractérisons chaque communauté selon le motif séquentiel fréquent fermé. La méthode Clospan (Yan et al., 2003) a été définie pour identifier tous les motifs séquentiels fermés existants, pour un seuil donné, avec une complexité en $O(n^2)$. Cependant, nous voulons que les motifs séquentiels identifiés soient représentatifs des communautés dans lesquelles ils sont identifiés. Il est donc nécessaire de prendre en compte le taux de croissance des motifs afin de tenir compte de leur émergence. Pour ce faire, nous appliquons la méthode de post-traitement définie dans (Plantevit et Cremilleux, 2009) pour calculer les taux de croissance de séquences d'article classés. Au final, notre approche est donc la suivante : d'abord nous identifions les motifs séquentiels fermés d'une communauté pour un support minimum donné. Puis, nous calculons les supports de ces motifs pour le reste du réseau entier. Enfin, nous en déduisons le taux de croissance de chacun des motifs séquentiels fermés.

Sélection des motifs séquentiels. Une fois que les motifs séquentiels ont été extraits pour chaque communauté comme expliqué à l'étape 3, nous devons sélectionner les plus représentatifs afin de caractériser la communauté. Pour cela, nous extrayons d'abord les nœuds supportant chaque motif. Dans un premier temps, nous avons décidé d'effectuer ce calcul de façon naive, ce qui entraine une complexité en $O(rn)$, où r est le nombre de motifs. Mais il faut souligner que ce calcul pourrait être accéléré en l'intégrant à Clospan. On choisit ensuite les motifs selon deux approches distinctes : (1) motif dont le support est le plus élevé et (2) motif dont le taux de croissance est le plus élevé. Les résultats de l'étape 3 nous donnent directement le support et le taux de croissance. Cependant, dans certaines communautés, le motif de taux de croissance maximal ne couvre pas une partie significative des nœuds constituant la communauté. Il est alors nécessaire d'identifier d'autres motifs, dans un souci de représentativité. Nous sélectionnons alors le motif dont l'ensemble des nœuds le supportant est le plus différent possible de celui du premier motif, afin de couvrir au maximum la partie de la communauté

ignorée par celui-ci. La distance entre les ensembles des nœuds supportant est calculée au moyen du coefficient de Jaccard. En cas d'égalité, nous utilisons le taux de croissance comme critère secondaire. Si la couverture totale est toujours insuffisante, on réitère en sélectionnant d'autres motifs selon les mêmes modalités. Cette itération se poursuit jusqu'à ce qu'au plus cinq nœuds ne soient pas couverts. Nous qualifions de *déviant* ces nœuds ne supportant aucun motif représentatif de la communauté. Comme nous devons comparer chaque paire de motifs pour chaque communauté, le temps nécessaire à la sélection des motifs représentatifs est de l'ordre de $O(r^2/p)$ où p est le nombre de communautés. La complexité totale de notre méthode est donc $O(n^2 + r^2/p)$. Il faut noter que r dépend grandement des nombres de tranches temporelles θ et d'items $|I|$.

3 Résultats

Nous présentons les expérimentations réalisées sur des données réelles. Nous avons choisi de traiter le réseau dynamique attribué de co-auteurs décrit dans (Desmier et al., 2012) et extrait de la base de données DBLP. Chacun des 2145 nœuds représente un auteur. Deux nœuds sont reliés si les auteurs correspondants ont publié au moins un article ensemble. Chaque tranche temporelle correspond à une période de 5 ans. Il y a au total 10 tranches temporelles allant de 1990 à 2012. Les périodes consécutives ont un chevauchement de 3 ans pour des raisons de stabilité. Pour chaque auteur, à chaque tranche temporelle, la base de données fournit le nombre de publications pour 43 conférences et journaux. Nous utilisons ces informations pour définir les 43 attributs nodaux correspondants, et nous en rajoutons deux : le nombre total de publications dans les conférences et dans les journaux. On a donc un total de 45 attributs. Nos descripteurs sont ces attributs, ainsi que les mesures topologiques décrites à la section 2.

Les mesures topologiques sont discrétisées différemment, en fonction de leur nature. Pour le degré, nous utilisons les seuils 3, 10 et 30. Pour la transitivité, qui est définie sur $[0; 1]$, il s'agit de $0,35$, $0,5$ et $0,7$. Pour l'enchâssement, lui aussi défini sur $[0; 1]$, les intervalles sont $0,3$ et $0,7$. Ces intervalles ont été déterminés de manière à tenir compte des distributions de ces mesures sur l'ensemble des nœuds et des tranches temporelles : les différents seuils correspondent aux zones de faible densité. Pour les mesures de Guimerá et Amaral (2005), nous utilisons les seuils définis dans l'article original, c'est-à-dire : $2,5$ pour z et $0,05$, $0,6$, et $0,8$ pour P. Le seuil utilisé pour z permet de distinguer les hubs $(z > 2,5)$ des non-hubs $(z \leq 2,5)$ communautaires. Pour les deux attributs relatifs aux nombres de publications dans des conférences et revues, nous considérons les valeurs $1, 2, 3, 4$ et > 5. Pour le nombre total de publications, nous avons identifié les seuils 5, 10, 20 et 50 comme étant les plus pertinents.

L'algorithme Louvain identifie 127 communautés dans le réseau pondéré global, pour une modularité de $0,59$. Cette valeur signifie que ce réseau est hautement modulaire. 96 de ces communautés ne contiennent qu'un seul nœud. Parmi les communautés restantes, 17 contiennent plus de 10 nœuds, la plus grande en ayant 335. Nous recherchons ensuite les motifs séquentiels de ces communautés, pour un support minimum de $0,3$. Il ne nous a pas été possible de descendre en dessous de cette valeur en raison du coût spatial de l'algorithme Clospan utilisé. Pour chacune des communautés dont la taille est supérieure à 40, nous détections plus de 5000 motifs, dont la plupart ne comprennent que des mesures topologiques.

Les motifs les plus supportés sont toujours une séquence de $z < 2,5$ pour toutes les communautés, avec des longueurs variables. Cela signifie que la majorité des nœuds de chaque

Communauté	Taille	Longueur du motif	Support
38	335	2	0,99
40	43	8	0,97
42	109	5	1,00
45	227	3	1,00
55	39	7	1,00
61	204	3	1,00
75	140	4	0,99
77	41	7	1,00
86	111	3	1,00
98	113	5	1,00
106	134	5	1,00
115	125	1	1,00
125	79	3	1,00

TAB. 1 – *Longueur des motifs les plus supportés pour une sélection de communautés.*

communauté ont un rôle de non-hub communautaire. Pour mémoire, Amaral & Guimerà définissent un hub communautaire comme un nœud dont le degré interne est largement supérieur au degré interne moyen de sa communauté. Ainsi, le motif détecté signifie que la majorité des nœuds ont un degré interne relativement faible (ce qui est attendu), et ce durablement (ce qui ne l'est pas forcément). Bien que ce type de motif soit présent dans toutes les communautés, on peut faire une distinction en considérant la longueur de la séquence, qui mesure cette durée. Dans le Tableau 1, nous listons la longueur des motifs séquentiels les plus supportés, en indiquant leur communauté, la taille de celle-ci et la valeur de support du motif. Les communautés dont les tailles sont comprises entre 39 et 45 (communautés 40, 55 et 77) obtiennent de longues séquences (resp. $8, 7$ et 7). Surtout, les supports pour les communautés 55 et 77 atteignent même la valeur maximale de 1. Cela signifie que, dans ces communautés, aucun hub n'existe ; ou bien si un hub apparait, il disparait rapidement. Cette observation est particulièrement intéressante, et traduit l'absence d'un meneur communautaire qui structurerait la communauté par ses connexions multiples. Pour la communauté 115, la longueur de la séquence est 1 et sa valeur de support est également 1. Cela signifie que tous les nœuds de cette communauté ont tenu le rôle de non-hub simultanément au moins une fois au cours du temps, mais qu'il existe des hubs pour le reste des tranches temporelles. Pour les communautés 38, 40 et 75, le support est inférieur à 1, ce qui signifie que si une écrasante majorité de nœuds tient le rôle de non-hub pour de longue durées, en revanche un petit nombre de nœuds occupe la place de hub, éventuellement par intermittence. Nous extrayons les auteurs qui ne suivent pas les motifs les plus supportés pour ces trois dernières communautés. Pour la communauté 38, il s'agit de *Philip S. Yu*, *Jiawei Han* et *Beng C. Ooi*. Comme supposé, ces nœuds ont un nombre de connexions remarquablement élevé à l'intérieur de leurs communautés, et les auteurs qu'ils représentent ont effectivement des rôles de meneurs dans leur domaine. Une analyse plus approfondie des données montre également qu'ils publient un total de plus de 10 articles par tranche temporelle. En outre, ils n'occupent jamais de rôle non-hub. Les nœuds déviants pour les communautés 40 et 75 sont respectivement *Hans-Peter Kriegel* et *Divesh Srivastava*. Là

aussi, il s'agit d'auteurs importants dans leur communauté. Les séquences qui les caractérisent confirment qu'ils sont productifs et n'occupent jamais le rôle non-hub au fil du temps.

Pour les communautés dont les tailles sont comprises entre 39 et 45, nous ne constatons aucune motif émergent contenant une conférence ou revue. Les motifs les plus émergents ont un taux de croissance supérieur à $1,79$, ce qui signifie qu'il n'existe pas de motif séquentiel très distinctif pour ces communautés. Pour la plupart des grandes communautés, le motif le plus émergent inclut une conférence ou une revue spécifique, ce qui peut s'interpréter en termes de thématique de la communauté. Les autres descripteurs constituant le motif sont des mesures topologiques. Comme pour les motifs les plus supportés, l'item $z < 2,5$ apparait le plus souvent dans les motifs détectés. Cependant, ces motifs les plus émergents ne permettent pas de couvrir la majorité des nœuds des communautés concernées. C'est pourquoi, comme nous l'avons expliqué dans la section 2, nous recherchons des motifs supplémentaires en minimisant l'intersection de leurs supports. Ces motifs sont généralement constitués de mesures topologiques, et n'ont pas un taux de croissance très élevé. Dans la suite, nous nous concentrons sur les communautés donnant les résultats les plus intéressants. Pour chacune, nous décrivons le motif le plus émergent et nous présentons les nœuds déviants, qui ne supportent ni le motif le plus émergent ni les motifs supplémentaires. Chaque motif est représenté formellement entre crochets, comme une séquence d'itemsets, eux-mêmes représentés entre parenthèses.

Pour la communauté 61, le motif le plus émergent est *<(ICML PUB. NUM = 1) (DEGRE 3-10, Z < 2,5)>* avec un taux de croissance $3,52$ et support $0,30$. Ce motif fait référence aux auteurs qui sont publiés une fois dans ICML, après quoi leur degré se stabilise entre 3 et 10 et ils occupent tous le rôle de non-hubs. Nous extrayons 7 motifs supplémentaires afin de couvrir tous les nœuds de la communauté. Parmi eux, les plus intéressants sont *<(Z < 2,5) (Z < 2,5) (Z < 2,5, TOTALE CONF PUB. NUM 1-5) (AAAI PUB. NUM = 1)>* avec un taux de croissance de $1,69$ et un support de $0,30$ et *<(PART. COEFF 0.05-0.6, CIKM PUB. NUM = 1)>* avec un taux de croissance de $1,40$ et un support de $0,30$. Le premier motif se réfère à des nœuds qui restent des nœuds non-hubs pendant un certain temps, puis commencent à publier dans des conférences, avant de publier à AAAI tout en perdant leur statut de non-hub (sans pour autant devenir massivement des hubs). Le second n'a pas de dimension temporelle, mais nous montre l'existence ponctuelle de nœuds publiant à CIKM tout en occupant une position périphérique dans leur communauté, i.e., en étant significativement connecté à d'autres communautés. Les nœuds déviants pour cette communauté sont *Alex A. Freitas, Claire Cardie , Edwin P. D. Pednault*. Parmi ces auteurs, Alex A. Freitas ne publie pas pendant les 8 premières tranches temporelles, puis commence à publier de manière très active, à différentes conférences telles qu'ICML ou AAAI, et dans des journaux. Pour les deux autres auteurs, tandis que Claire Cardie publie régulièrement à ICML au cours des 6 premières tranches temporelles, Edwin P. D. Pednault ne publie ni à ICML, ni AAAI ou CIKM .

Le motif *<(PODS PUB. NUM = 1)>* est le plus émergent dans la communauté 75. Son taux de croissance est $3,59$ et son support $0,40$. Ce motif montre que $\%40$ des auteurs de cette communauté publie au moins une fois à PODS et ce groupe est émergent par rapport au reste du réseaux pour cette communauté. Quatre motifs supplémentaires sont nécessaires pour couvrir le reste de la communauté. Ces motifs se réfèrent à des nœuds occupant des positions de non-hub périphérique, et dont la transitivité est très élevée. Les nœuds déviants sont *Ninghui Li, Li Feifei* et *Abdullah Mueen*, qui ont la particularité de ne jamais publier dans PODS. Le motif le plus émergent de la communauté 106 est *<(Z < 2,5) (Z < 2,5) (Z < 2,5) (Z < 2,5) (Z < 2,5)*

(PART. COEFF 0.05-0.6, KDD PUB. NUM=1)> avec un taux de croissance $2,87$ et support $0,40$. Ce motif fait référence à des nœuds durablement non-hubs, qui deviennent périphériques tout en publiant à KDD. Nous identifions 4 motifs supplémentaires pour finir de couvrir la communauté. Ceux-ci concernent les nœuds à la fois ultrapériphériques et bien intégrés à leur propre communauté. Les nœuds déviants sont *Stan Matwin*, qui publie plus d'un article par tranche temporelle à KDD, et qui n'est pas durablement non-hub ; et *Hua-Jun Zeng* qui n'a jamais publié dans KDD. Ce dernier est aussi caractérisé par un fort accroissement du nombre d'articles produits, alors qu'il ne publie pas pendant les 5 premières tranches temporelles.

Le motif le plus émergent de la communauté 45 est <*(VLDB PUB. NUM=3) (DEGRE 3-10 Z < 2,5)*> avec un taux de croissance de $6,40$ et un support de $0,30$. Cette séquence nous dit qu'il y a un groupe remarquable d'auteurs qui ont publié 3 fois en conférence VLDB, puis dont le degré s'est stabilisé entre 3 et 10 et qui ont occupé des positions non-hubs dans leur communauté. Nous avons dû identifier 6 autres motifs pour couvrir le reste de la communauté. L'un d'eux est <*(Z < 2,5, TOTAL CONF. PUB. NUM 1-5) (Z < 2,5, EMBED 0.3-0.7, ICDE PUB. NUM= 1)*>, avec un taux de croissance de $2,30$ et un support de $0,30$. Ce motif couvre les nœuds non-hubs qui ont publié entre 1 et 5 fois dans des conférences, puis dont l'enchâssement s'est stabilisé à une valeur relativement élevée, tout en gardant leur position de non-hub et en publiant à ICDE . Les nœuds déviants sont *Ingmar Weber* et *Anastasia Ailamaki*, qui ne publient pas pour les sept premières tranches temporelles, puis deviennent de plus en plus productifs au cours des trois dernières tranches.

Pour résumer nos observations, le motif le plus émergent dans à peu près toutes les communautés comprend habituellement le fait d'être non-hub et d'avoir un petit nombre de publications dans divers conférence ou journaux. En fonction des quelques conférences ou journaux apparaissant dans ces motifs, il est possible de déduire le thème principal des communautés. Pour certaines communautés, cependant, les motifs séquentiels émergents sont purement topologiques (pas d'attributs). On peut alors supposer que les membres de ces communautés ne publient pas de façon suffisamment homogène pour que cela transparaisse dans les motifs. Une autre raison peut être simplement que les membres de la communauté sont reliés pour des raisons autres que thématiques, auquel cas cela n'apparait pas dans les attributs sélectionnés pour notre étude. En ce qui concerne les nœuds déviants, on peut distinguer différents types de profils. Certains peuvent correspondre à des auteurs dont la thématique principale est différente de celle de la communauté dans laquelle ils ont été placés. Dans certains cas, nous avons détecté des auteurs qui avaient visiblement changé de thématique, ou bien qui débutaient dans une thématique donnée. Il peut également s'agir d'auteurs actifs dans un autre domaine, dont les conférences et journaux ne font pas partie de ceux retenus dans les données que nous avons considérées ici. Un autre profil est celui du chercheur en train de monter en charge, et dont la position communautaire et le nombre de publications sont en train d'évoluer de conjointement.

4 Travaux connexes

Il existe de nombreuses méthodes de détection de communautés dans les réseaux complexes ordinaires (i.e., ne contenant que des nœuds et des liens), basées uniquement sur l'information topologique. Une revue détaillée de ces méthodes est donnée dans (Fortunato, 2010). Dans le cas des réseaux plus riches, contenant des attributs nodaux, une communauté est essentiellement définie comme un groupe de nœuds à la fois densément interconnectés, et similaires

en termes d'attributs. Certaines méthodes tentent d'estimer des partitions optimisant directement ces deux critères simultanément (Cruz et al., 2011). D'autres transforment encodent d'abord les attributs sous forme d'information structurelle, et détectent ensuite les communautés de façon plus classique (Ruan et al., 2013; Zhou et al., 2009). Il existe également des méthodes de recherche de motifs (ici de petits sous-graphes densément connectés tels que des cliques) possédant des attributs homogènes (Gunnemann et al., 2010; Moser et al., 2009; Mougel et al., 2010).

Il existe aussi des travaux qui se concentrent sur la détection de communautés dans les réseaux dynamiques (Robardet, 2009). Leur but principal est l'observation de l'évolution de la structure de communautés, à travers des évènements tels que la formation, la dissolution, la croissance, la diminution et la fusion. Pour des réseaux dynamique attribuées, il existe des méthodes pour trouver les groupes de nœuds structurellement similaires dont les attributs changement de la même façon au cours du temps (Desmier et al., 2012). Ces méthodes ne cherchent pas des communautés. L'objectif de ces travaux est la découverte de motifs intéressants en tenant compte de la structure des liens et des attributs pour les réseaux dynamiques attribuées. Pour regrouper les nœuds, ces méthodes utilisent des contraintes, comme par exemple une distance limite entre nœuds de la même motif, ou la nécessité de partager les mêmes voisins.

Bien qu'il existe de nombreuses techniques différentes pour identifier les structures de communautés, il y a peu d'auteurs qui travaillent à caractériser les communautés obtenues. Dans (Lancichinetti et al., 2010), les auteurs comparent les distributions de certaines mesures topologiques afin de comprendre la forme générale des communautés, et tentent de les caractériser en fonction du type de système modélisé (biologique, informatique, social, etc.). Ils n'utilisent pas d'attributs nodaux et ne traitent que des réseaux statiques. Dans (Tumminello et al., 2011), les auteurs proposent une méthode statistique inspirée d'approches utilisées en génétique, pour caractériser les communautés en termes d'attributs surexprimés. Cependant, cette étude n'a pas recours à des mesures topologiques et se limite également aux réseaux statiques. Dans (Labatut et Balasque, 2012), les auteurs interprètent les communautés d'un réseau attribué social. Ils utilisent la régression statistique et l'analyse discriminante pour identifier les valeurs des attributs les plus caractéristiques de chaque communauté. Cependant, ici encore, le réseau est statique et l'information structurelle n'est pas exploitée.

Enfin, le travail présenté dans (Prado et al., 2013) n'est pas directement concerné par la détection ou la caractérisation de communautés, mais est néanmoins lié à notre approche. Prado et al. (2013) ont introduit l'idée d'utiliser les mesures topologiques et les attributs nodaux ensemble dans une perspective d'exploration de données. Leur but était de trouver des motifs reflétant la covariation de mesures topologiques et d'attributs sur l'ensemble du réseau. Ils se sont concentrés sur des réseaux statiques et n'ont pas considéré de structures communautaires.

5 Conclusion

Nous traitons le problème de la caractérisation des communautés dans des réseaux complexes dynamiques et attribués. Nous proposons une nouvelle représentation de l'information encodée dans le réseau, permettant de stocker simultanément l'information topologique, les attributs nodaux et la dimension temporelle. Nous utilisons cette représentation pour effectuer une fouille de motifs séquentiels fréquents. Chaque communauté peut ensuite être caractéri-

sée par ses motifs les plus distinctifs. Nous tirons également parti des motifs pour détecter et caractériser les nœuds déviants dans chaque communauté. Nous appliquons notre méthode à un réseau de collaboration scientifique construit à partir des données publiques de la base DBLP. Les résultats montrent que notre méthode est capable de caractériser les communautés en fonction, notamment, de leur thématique. Les nœuds déviants identifiés correspondent à différents types de profils, tels que des chefs de file communautaires, des chercheurs émergents, ou d'autres en train de changer de thématique de recherche.

A notre connaissance, il s'agit de la première formulation de la caractérisation de communautés comme un problème de fouille de données. Notre but était de surmonter les limitations des rares travaux existants (Lancichinetti et al., 2010; Tumminello et al., 2011; Labatut et Balasque, 2012) en proposant une approche systématique, tenant compte à la fois de la structure, des attributs nodaux et du temps. La représentation des données que nous utilisons n'avait jamais été appliquée au traitement de graphes. Le processus proposé pour extraire les motifs les plus pertinents en se basant sur une recherche de motifs séquentiels sous contraintes est original et nous avons montré la richesse des interprétations apportées sur un cas réel de réseau. Certaines étapes de l'implémentation présentée ici étaient relativement naives, nous comptons les améliorer afin de réduire le temps nécessaire au calcul. En particulier, l'identification des ensembles de support pourrait être intégrée à Clospan.

Afin de restreindre la complexité conceptuelle de cette première approche, nous avons volontairement limité notre méthode d'analyse en ne considérant pas l'évolution des communautés au cours du temps. Dans des travaux ultérieurs, nous envisageons d'appliquer un algorithme de détection de communautés approprié en insérant cette information dans la base de données utilisée pour la fouille de motifs. Nous comptons également appliquer notre méthode d'analyse à d'autres types de réseaux pour explorer ses capacités de caractérisation. Comme autre perspective, nous pouvons aussi mieux exploiter nos représentations de réseaux attribués dynamiques. Ici, nous nous sommes seulement intéressés à l'extraction de séquences fréquentes. Cependant, notre représentation des données du réseau peut également être utilisée pour interroger les nœuds selon certaines mesures ou attributs topologiques spécifiques. Ainsi, dans nos expériences, nous avons vu qu'il y avait beaucoup de nœuds qui n'appartiennent pas à une communauté. Il serait intéressant de bien regarder ces nœuds et mieux comprendre en quoi ils sont différents des autres pour, par exemple, formuler des hypothèses sur leur isolement.

Références

Blondel, V., J.-L. Guillaume, R. Lambiotte, et E. Lefebvre (2008). Fast unfolding of communities in large networks. *J. Stat. Mech. 2008*(10), P10008.

Cruz, J., C. Bothorel, et F. Poulet (2011). Entropy based community detection in augmented social networks. In *CASoN 2011*, pp. 163–168.

Desmier, E., M. Plantevit, C. Robardet, et J.-F. Boulicaut (2012). Cohesive co-evolution patterns in dynamic attributed graphs. In *DS*, Volume 7569 of *LNCS*, pp. 110–124.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports 486*(3-5), 75–174.

Guimerá, R. et L. N. Amaral (2005). Cartography of complex networks : modules and universal roles. *J. Stat. Mech. 2005*(02), P02001.

Gunnemann, S., I. Farber, B. Boden, et T. Seidl (2010). Subspace clustering meets dense subgraph mining : A synthesis of two paradigms. In *ICDM*, pp. 845–850.

Labatut, V. et J.-M. Balasque (2012). Detection and interpretation of communities in complex networks : Practical methods and application. In *CSoNet*, pp. 81–113. Springer.

Lancichinetti, A., M. Kivela, J. Saramaki, et S. Fortunato (2010). Characterizing the community structure of complex networks. *PLoS ONE 5*(8), e11976.

Mabroukeh, N. R. et C. I. Ezeife (2010). A taxonomy of sequential pattern mining algorithms. *ACM Comput. Surv. 43*(1), 3 :1–3 :41.

Moser, F., R. Colak, A. Rafiey, et M. Ester (2009). Mining cohesive patterns from graphs with feature vectors. *SIAM SDM*, 593–604.

Mougel, P.-N., M. Plantevit, C. Rigotti, O. Gandrillon, et J.-F. Boulicaut (2010). Constraint-based mining of sets of cliques sharing vertex properties. In *ACNE - ECML/PKDD*.

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

Plantevit, M. et B. Cremilleux (2009). Condensed representation of sequential patterns according to frequency-based measures. In *IDA*, Volume 5772 of *LNCS*, pp. 155–166.

Prado, A., M. Plantevit, C. Robardet, et J.-F. Boulicaut (2013). Mining graph topological patterns : Finding covariations among vertex descriptors. *TKDE 25*(9), 2090–2104.

Robardet, C. (2009). Constraint-based pattern mining in dynamic graphs. In *ICDM'09*, pp. 950–955.

Ruan, Y., D. Fuhry, et S. Parthasarathy (2013). Efficient community detection in large networks using content and links. In *WWW*, pp. 1089–1098.

Tian, Y., R. A. Hankins, et J. M. Patel (2008). Efficient aggregation for graph summarization. In *SIGMOD*, pp. 567–580.

Tumminello, M., S. Micciche, F. Lillo, J. Varho, J. Piilo, et R. N. Mantegna (2011). Community characterization of heterogeneous complex systems. *J. Stat. Mech. 2011*(01), P01019.

Yan, X., J. Han, et R. Afshar (2003). Clospan : Mining closed sequential patterns in large datasets. In *SDM '03*, pp. 166–177.

Zhou, Y., H. Cheng, et J. Yu (2009). Graph clustering based on structural/attribute similarities. *Proc. VLDB Endow. 2*, 718–729.

Summary

Many complex systems are modeled through complex networks whose analysis reveals typical topological properties. Amongst those, the community structure is one of the most studied. Many methods are proposed to detect communities, not only in plain, but also in attributed, directed or even dynamic networks. A community structure takes the form of a partition of the node set, which must then be characterized relatively to the properties of the studied system. We propose a method to support such a characterization task. We define a sequence-based representation of networks, combining temporal information, topological measures, and nodal attributes. We then characterize communities using the most representative emerging sequen-

tial patterns of its nodes. This also allows detecting unusual behavior in a community. We describe an empirical study of a network of scientific collaborations.