Extraction sous Contraintes d'Ensembles de Cliques Homogènes

Pierre-Nicolas Mougel * **, Marc Plantevit* ***
Christophe Rigotti* **, Olivier Gandrillon****, Jean-François Boulicaut* **

*Université de Lyon, CNRS, INRIA
**INSA-Lyon, LIRIS Combining, UMR5205, F-69621, France
***Université Lyon 1, LIRIS Combining, UMR5205, F-69622, France
prénom.nom@liris.cnrs.fr,
****Université Lyon 1, Centre de Génétique Moléculaire et Cellulaire, UMR5534,
F-69622, Villeurbanne, France
gandrillon@cgmc.univ-lyon1.fr

Résumé. Nous proposons une méthode de fouille de données sur des graphes ayant un ensemble d'étiquettes associé à chaque sommet. Une application est, par exemple, d'analyser un réseau social de chercheurs co-auteurs lorsque des étiquettes précisent les conférences dans lesquelles ils publient. Nous définissons l'extraction sous contraintes d'ensembles de cliques tel que chaque sommet des cliques impliquées partage suffisamment d'étiquettes. Nous proposons une méthode pour calculer tous les Ensembles Maximaux de Cliques dits Homogènes qui satisfont une conjonction de contraintes fixée par l'analyste et concernant le nombre de cliques séparées, la taille des cliques ainsi que le nombre d'étiquettes partagées. Les expérimentations montrent que l'approche fonctionne sur de grands graphes construits à partir de données réelles et permet la mise en évidence de structures intéressantes.

1 Introduction

De nombreux problèmes peuvent être étudiés sous l'angle de l'analyse de grands graphes dont les sommets représentent des entités et les arêtes des relations ou interactions entre ces entités. Les contextes d'application sont innombrables et nous traitons ici deux graphes réels issus de l'étude d'un réseau social (coopération de chercheurs) et de réseaux d'interaction protéine/protéine. L'accumulation toujours plus facile de données sur de tels graphes mobilise largement la communauté des chercheurs en fouille de données. Deux approches complémentaires sont principalement étudiées. De nombreux travaux s'intéressent aux propriétés macroscopiques de ces graphes (e.g., valeurs moyennes des diamètres, nombre et taille des composantes connexes, degrés). D'autres se réclament davantage de la découverte de motifs locaux pour identifier des sous-graphes particuliers, par exemple, les cliques maximales.

Nous nous intéressons à des données qui correspondent à un graphe et tel qu'un ensemble d'étiquettes soit associé à chaque sommet. De telles données sont souvent disponibles mais la

plupart des méthodes actuelles ne traitent que l'information sur les relations (arêtes) entre les sommets. Ainsi, lorsque des vecteurs de propriétés sont associés aux sommets, Moser et al. (2009) proposent d'extraire des sous-graphes denses tel que tous les sommets de chaque sous-graphe partagent suffisamment de propriétés. Dans notre contexte, les étiquettes vont être utilisées pour représenter un ensemble de propriétés booléennes. Ce type de données se retrouve, par exemple, dans la modélisation d'un réseau social où les arêtes représentent des rencontres et où chaque individu serait associé à une liste de ses activités préférées. Sur ce type de graphe, la tâche de fouille de données que nous étudions consiste à extraire des *collections* de cliques tel que tous les sommets partagent suffisamment d'étiquettes. Plus précisément nous recherchons des collections de cliques qui satisfont des contraintes sur le nombre de cliques séparées, la taille des cliques et le nombre d'étiquettes partagées par tous les sommets. Nous proposons alors un algorithme efficace pour extraire *tous* les motifs ainsi caractérisés. Notez que pour nous, un motif n'est pas une clique mais une collection de cliques. Cette nouvelle tâche de fouille de données va permettre de révéler des interactions non triviales entre diverses cliques.

La recherche de communautés dans un réseau social est très étudiée. D'après Wasserman et Faust (1994) une communauté est souvent définie comme un ensemble d'individus fortement en interaction entre eux et ayant peu d'interaction avec l'extérieur. Il est donc possible de modéliser une communauté comme un sous-graphe fortement connecté, typiquement une clique ou une quasi-clique, dont les sommets sont moins connectés avec le reste du graphe. L'extraction d'Ensembles Maximaux de Cliques Homogènes (EMCHs) qui est formalisée dans cet article va permettre de découvrir des ensembles de communautés partageant certaines caractéristiques (au sens des étiquettes communes). Par exemple, il devient possible de trouver des communautés qui n'ont aucune interaction entre elles (collection de cliques non connectées) et l'étude de la distribution des tailles des communautés sera souvent informative (par exemple, tailles de communautés équivalentes ou très différentes bien qu'elles partagent les mêmes intérêts). Ces analyses peuvent permettre d'interpréter les possibilités et réalisations de fusion ou d'absorption. Sur ces mêmes motifs, nous allons pouvoir trouver des individus qui font partie de plusieurs communautés du motif et qui deviennent donc des acteurs privilégiés pour la propagation d'informations. La recherche de ce type d'individu est une tâche courante lorsque l'on prend en compte l'ensemble du graphe pour trouver des sommets ayant un grand degré (concept de "hub"). L'intérêt de notre approche est de trouver des individus en relation avec beaucoup d'individus dans leurs communautés, mais qui ne sont pas forcément en relation avec un grand nombre d'individus sur l'ensemble du jeu de données. Il ne sera justement pas possible de répondre à cet objectif avec les approches que nous avons qualifié de macroscopiques.

Nous avons validé notre méthode d'extraction des EMCHs via des expérimentations pour deux domaines différents. Nous traitons un réseau de collaborations scientifiques calculé à partir de la base de données DBLP [1] mais aussi un réseau d'interactions protéine/protéine. Ces expérimentations montrent que le processus d'extraction passe à l'échelle sur des graphes avec plusieurs centaines de milliers de sommets et que l'on obtient des informations non triviales en interprétant les motifs extraits. On sait que l'énumération de toutes les cliques est coûteuse en temps de calcul [2] à cause du grand nombre de cliques, et que, par conséquent, l'énumération de tous les ensembles de cliques est encore plus difficile. Cependant, nous avons étudié avec

1. Voir `http://dblp.uni-trier.de/` et Ley (2009).
2. Tomita et al. (2006) montrent que l'énumération des cliques maximales est en $\mathcal{O}(3^{n/3})$.

soin les propriétés des contraintes à exploiter au cours d'une recherche complète des EMCHs pour proposer une méthode qui passe à l'échelle sur des données réelles.

La Section 2 formalise la tâche de calcul des Ensembles Maximaux de Cliques Homogènes et la méthode d'extraction implémentée. Les expérimentations sont présentées dans la Section 3. Les travaux connexes sont discutés dans la Section 4. La Section 5 est une brève conclusion.

2 Ensemble Maximal de Cliques Homogènes

Dans cette section, nous définissons le langage de description des motifs que nous souhaitons extraire et nous proposons un algorithme pour les calculer.

2.1 Spécification du problème

Définition 2.1.1. (Jeu de données) *Soit $\mathcal{E}$ un ensemble d'étiquettes, un* jeu de données *est une paire $\langle G, f\rangle$, avec $G = \langle \mathcal{S}, \mathcal{A}\rangle$ un graphe simple non orienté (sommets $\mathcal{S}$ et arêtes $\mathcal{A}$), et f une fonction $f : \mathcal{S} \to \mathcal{P}(\mathcal{E})$ associant un ensemble d'étiquettes à chaque sommet.*

Un ensemble C de sommets est appelé *clique* de G si le sous-graphe induit par C est complet. La collection de toutes les cliques de G est notée $\mathcal{C}_G$.

Définition 2.1.2. (Ensemble de Cliques Homogènes) *Soit α, β, et κ trois entiers strictement positifs, un* Ensemble de Cliques Homogènes *(ECH) dans un jeu de données $\langle G, f\rangle$ est une collection M de cliques $\{C_1, \ldots, C_n\} \subseteq \mathcal{C}_G$ tel que les trois contraintes $\mathbb{C}^{lab}_{\alpha}$, $\mathbb{C}^{clique}_{\kappa,\beta}$ et $\mathbb{C}^{sep}$ soient satisfaites :*

- *$\mathbb{C}^{lab}_{\alpha}$: $|\bigcap_{C\in M}(\bigcap_{v\in C} f(v))| \geq \alpha$, i.e., les sommets partagent au moins α étiquettes ;*
- *$\mathbb{C}^{clique}_{\kappa,\beta}$: M est composé d'au moins κ cliques de taille au moins β ;*
- *$\mathbb{C}^{sep}$: $\forall C, C' \in M, C \neq C'$, on a $C \cup C' \notin \mathcal{C}_G$, i.e., les cliques de M sont* séparées.

On peut noter que les cliques d'un ECH ne sont pas forcément maximales dans G car cela serait trop contraignant. En effet, puisque la contrainte $\mathbb{C}^{lab}_{\alpha}$ porte sur l'ensemble des étiquettes en commun, une clique maximale de G peut ne pas vérifier cette contrainte alors que l'un de ses sous-ensembles la vérifie. De plus, la contrainte $\mathbb{C}^{sep}$ est nécessaire afin d'éviter que toutes les collections de κ sous-ensembles (de taille au moins β) d'une clique faisant partie d'un ECH forment également un ECH.

Tant que les contraintes $\mathbb{C}^{lab}_{\alpha}$ et $\mathbb{C}^{clique}_{\kappa,\beta}$ sont satisfaites, il est possible de construire à partir d'un ECH d'autres ECHs en supprimant des sommets d'une des cliques. Cette propriété fait que le nombre d'ECH a tendance à grandir rapidement. Nous nous focalisons donc sur les motifs maximaux qui peuvent être considérés comme les plus spécifiques.

Définition 2.1.3. (Ensemble Maximal de Cliques Homogènes et ordre partiel $\preceq$) *Un Ensemble Maximal de Cliques Homogènes (EMCH) est un ECH maximal par rapport à l'ordre partiel $\preceq$ défini comme suit : Soit M_1 et M_2 deux ECHs, $M_1 \preceq M_2$ si et seulement si pour tout $C_1 \in M_1$ il existe $C_2 \in M_2$ tel que $C_1 \subseteq C_2$.*

De façon générale il n'y a pas d'anti-symétrie pour $\preceq$ avec une collection quelconque d'ensembles. Cependant, dans le cas particulier d'une collection d'ECHs, $\preceq$ est un ordre partiel comme énoncé par le théorème suivant.

Théorème 2.1.1. *La relation $\preceq$ est un ordre partiel sur une collection d'ECHs.*

Démonstration. On voit immédiatement que la relation est réflexive et transitive. Pour montrer l'anti-symétrie, prenons M_1 et M_2 deux ECHs tel que $M_1 \preceq M_2$ et $M_2 \preceq M_1$. Supposons que $M_1 \neq M_2$, alors il existe une clique $C \in M_1$ qui est différente de toutes les autres cliques de M_2. De plus comme $M_1 \preceq M_2$, il existe $C' \in M_2$ tel que $C \subset C'$. De même, comme $M_2 \preceq M_1$, il existe C'' dans M_1 tel que $C' \subseteq C''$. Donc $C \subset C''$, et comme les cliques d'un ECH doivent être séparées, cela mène à une contradiction. □

Nous pouvons maintenant définir la tâche de fouille de données que nous voulons réaliser.

Tâche de fouille de données Étant donné un jeu de données formé par un graphe avec un ensemble d'étiquettes associé à chaque sommet, le problème consiste à extraire tous les motifs de type EMCH vérifiant la conjonction de contraintes $\mathbb{C}^{lab}_{\alpha} \wedge \mathbb{C}^{clique}_{\kappa,\beta} \wedge \mathbb{C}^{sep}$.

2.2 Extraction des EMCHs

Notons $vertices(M) = \bigcup_{C \in M} C$, i.e. l'ensemble des sommets d'une collection M d'ensembles de sommets. Le théorème suivant énonce que si M est un EMCH, alors connaître $vertices(M)$ est suffisant pour déterminer M.

Théorème 2.2.1. *Soit M un EMCH, alors M est la collection des cliques maximales dans le sous-graphe G_M de G induit par $vertices(M)$.*

Démonstration. (Esquisse). Soit S la collection des cliques maximales dans G_M, et supposons que $M \neq S$. Comme S contient toutes les cliques maximales et M contient des cliques qui sont séparées, alors $S \not\subseteq M$. Donc il existe $D \in S$ tel que $D \notin M$. Deux cas sont possibles. Si D est un sur-ensemble d'une clique C de M, alors puisque les cliques de M sont séparées, remplacer C par D dans M donne un motif qui satisfait $\mathbb{C}^{sep}$ (ainsi que $\mathbb{C}^{lab}_{\alpha}, \mathbb{C}^{clique}_{\kappa,\beta}$), et donc M n'est pas un ECH maximal. Si D n'est pas un sur-ensemble d'une clique de M, alors puisque D est une clique maximale de G_M toutes les cliques de $M \cup \{D\}$ sont séparées. Donc $M \cup \{D\}$ satisfait $\mathbb{C}^{sep}$ (ainsi que $\mathbb{C}^{lab}_{\alpha}, \mathbb{C}^{clique}_{\kappa,\beta}$), et de la même manière, M n'est pas un ECH maximal. □

Nous pouvons en déduire le corollaire suivant.

Corollaire 2.2.1. *Étant donné M un EMCH, alors il n'y a pas d'autre EMCH M' ($M \neq M'$) tel que $vertices(M) \subseteq vertices(M')$.*

Pour un jeu de données $\langle G, f \rangle$, f peut être représentée par une relation binaire $\mathcal{R} \subseteq \mathcal{S} \times \mathcal{E}$, définie comme $x\mathcal{R}y \Leftrightarrow y \in f(x)$, et mettant en relation chaque sommet avec ses étiquettes. Discutons des relations entre les EMCHs et les *ensembles clos* sur $\mathcal{R}$.

Considérons les fonctions g et h, définies comme suit, $g : \mathcal{P}(\mathcal{S}) \rightarrow \mathcal{P}(\mathcal{E}), g(X) = \{y \in \mathcal{E} | \forall x \in X, \mathcal{R}(x, y)\}$ et $h : \mathcal{P}(\mathcal{E}) \rightarrow \mathcal{P}(\mathcal{S}), h(Y) = \{x \in \mathcal{S} | \forall y \in Y, \mathcal{R}(x, y)\}$. Ces fonctions définissent une *correspondance de Galois* entre $\mathcal{P}(\mathcal{S})$ et $\mathcal{P}(\mathcal{E})$ (voir Zaki et Ogihara (1998); Pasquier et al. (1999)). Un ensemble de sommets $V \subseteq \mathcal{S}$ (resp. d'étiquettes $L \subseteq \mathcal{E}$) est *clos* dans $\mathcal{R}$ si $V = h(g(V))$ (resp. $L = g(h(L))$). En ne considérant que les ensembles clos, les fonctions g et h sont des anti-isomorphismes (i.e., bijections qui associent à A, B, telles que $A \subseteq B$, les images A', B' telles que $B' \subseteq A'$).

Théorème 2.2.2. *Soit M un EMCH, alors $vertices(M)$ est clos dans $\mathcal{R}$.*

Démonstration. (Esquisse). Soit $V = vertices(M), L = g(V), V' = h(L)$, et supposons que V n'est pas clos dans $\mathcal{R}$, alors $V \subset V'$. Soit M' la collection de toutes les cliques maximales dans le sous-graphe de G induit par V'. Cette collection satisfait $\mathbb{C}^{sep}$. Comme M est un EMCH, alors, d'après le Théorème 2.2.1, M est la collection des cliques maximales dans le sous-graphe de G induit par V, et donc $M \preceq M'$, et M' satisfait $\mathbb{C}^{clique}_{\kappa,\beta}$ puisque M la satisfait. Comme $V' = h(L)$, les sommets de V' partagent au moins autant d'étiquettes que les sommets de V. Comme M satisfait $\mathbb{C}^{lab}_{\alpha}$, M' la satisfait également. Donc M n'est pas un ECH maximal. □

Désignons par $maxCliques(G, V)$ la collection de cliques maximales dans le sous-graphe de G induit par V. Les Théorèmes 2.2.1 et 2.2.2 (et le Corollaire 2.2.1) permettent de proposer deux méthodes correctes pour trouver tous les EMCHs :

- Trouver les ensembles clos de sommets V dans $\mathcal{R}$, tel que $maxCliques(G, V)$ satisfasse $\mathbb{C}^{lab}_{\alpha}, \mathbb{C}^{clique}_{\kappa,\beta}$ et $\mathbb{C}^{sep}$. Parmi eux, sélectionner les maximaux et, pour chacun de ces ensembles maximaux V, retourner $maxCliques(G, V)$.
- Ou, en utilisant la correspondance de Galois, trouver les ensembles clos minimaux d'étiquettes L dans $\mathcal{R}$, tel que $maxCliques(G, h(L))$ satisfasse $\mathbb{C}^{lab}_{\alpha}, \mathbb{C}^{clique}_{\kappa,\beta}$ et $\mathbb{C}^{sep}$, et pour tous ces ensembles minimaux L, retourner $maxCliques(G, h(L))$.

Pour concevoir un algorithme complet efficace, l'essentiel de l'effort consiste à étudier les propriétés des contraintes pour savoir si les techniques d'élagage aujourd'hui bien maitrisées s'appliquent ou non. Il faut alors, le cas échéant, savoir comment le test de certaines contraintes peut s'appuyer sur des formes relaxées et lesquelles.

Propriétés des contraintes

Nous avons montré qu'étant donné un ensemble de sommets V (resp. ensemble d'étiquettes L), V (resp. L) satisfait les contraintes $\mathbb{C}^{lab}_{\alpha}, \mathbb{C}^{clique}_{\kappa,\beta}$ ou $\mathbb{C}^{sep}$ si et seulement si la collection $maxCliques(G, V)$ (resp. $maxCliques(G, h(L))$) satisfait les mêmes contraintes.

Considérons ces contraintes et leurs propriétés de monotonie (i.e., si A satisfait une contrainte, alors tous les sur-ensembles de A la satisfont également) et d'anti-monotonie (i.e., si A satisfait une contrainte, alors tous les sous-ensembles de A la satisfont également). Les propriétés suivantes sont immédiates.

- $\mathbb{C}^{lab}_{\alpha}$ est monotone (resp. anti-monotone) par rapport à l'ensemble des étiquettes (resp. l'ensemble des sommets) ;
- $\mathbb{C}^{clique}_{\kappa,\beta} \wedge \mathbb{C}^{sep}$ est anti-monotone (resp. monotone) par rapport à l'ensemble des étiquettes (resp. l'ensemble des sommets) ;

De plus, la conjonction des contraintes $\mathbb{C}^{clique}_{\kappa,\beta} \wedge \mathbb{C}^{sep}$ peut être exprimée de façon relaxée par la contrainte $\mathbb{C}^{vert}$ définie de la manière suivante :

M satisfait $\mathbb{C}^{vert}$ si $|vertices(M)| \geq \beta + \kappa - 1$ (i.e., pour contenir au moins κ cliques séparées de taille au moins β, M doit contenir au moins $\beta + \kappa - 1$ sommets).

On voit immédiatement que $\mathbb{C}^{vert}$ est anti-monotone (resp. monotone) par rapport à l'ensemble des étiquettes (resp. l'ensemble des sommets). Puisque la vérification de $\mathbb{C}^{vert}$ est simple et ne nécessite pas d'extraire les cliques maximales, elle peut être utilisée en premier : la vérification de $\mathbb{C}^{clique}_{\kappa,\beta} \wedge \mathbb{C}^{sep}$ n'est nécessaire que si $\mathbb{C}^{vert}$ est satisfaite.

Algorithme et implémentation

Afin d'extraire les EMCHs, il est possible d'énumérer soit l'ensemble des sommets soit l'ensemble des étiquettes en poussant les contraintes lorsque c'est possible. Pour les domaines d'application auxquels nous nous intéressons dans la partie 3, $\mathcal{E}$ est généralement plus petit que $\mathcal{S}$. Nous préférons donc énumérer sur l'ensemble des étiquettes dans $\mathcal{R}$. Une étude permettant de valider cette approche pour l'extraction des EMCHs sort du contexte de cet article.

Pour l'extraction des EMCHs en énumérant les ensembles clos d'étiquettes, nous pouvons facilement réutiliser la plupart des algorithmes d'extraction d'ensembles clos qui effectuent un parcours en largeur ou en profondeur, en utilisant $\mathbb{C}^{vert}$ comme une contrainte de *support* standard (anti-monotone), $\mathbb{C}^{lab}_{\alpha}$ comme une contrainte monotone, et $\mathbb{C}^{clique}_{\kappa,\beta} \wedge \mathbb{C}^{sep}$ comme une contrainte anti-monotone. Dans l'implémentation actuelle, nous avons utilisé un algorithme similaire à Closet de Pei et al. (2000). La conjonction de contraintes $\mathbb{C}^{clique}_{\kappa,\beta} \wedge \mathbb{C}^{sep}$ est *partiellement poussée*, en élaguant la branche d'énumération qui ne la satisfait pas. La contrainte $\mathbb{C}^{lab}_{\alpha}$ n'est pas poussée, elle est uniquement vérifiée, mais il serait possible de la pousser efficacement lors d'un développement futur, par exemple en utilisant la technique de réduction de données ExAnte proposée par Bonchi et al. (2003).

Puisque nous ne nous intéressons qu'aux ensembles clos d'étiquettes qui satisfont les contraintes (celles qui correspondent aux ECHs maximaux), lorsqu'un ensemble clos L vérifiant toutes les contraintes est trouvé, alors la branche d'énumération actuelle est élaguée, et L ainsi que $maxCliques(G, h(L))$ sont stockés. Finalement, une fois que l'exploration est terminée, un test est réalisé en post-traitement pour s'assurer de la minimalité des ensembles clos et $maxCliques(G, h(L))$ est retourné pour chaque ensemble minimal L.

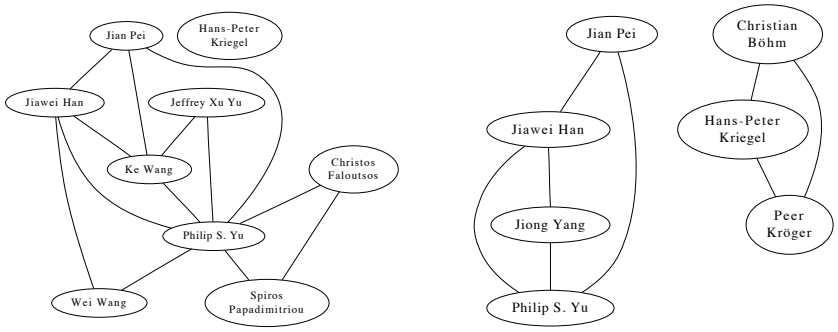
Pour extraire les cliques maximales (fonction $maxCliques$), nous avons implémenté l'algorithme de Tomita et al. (2006) qui a une complexité temporelle *optimale* en $\mathcal{O}(3^{n/3})$. Cet algorithme permet également de rechercher des cliques maximales dans un sous-graphe sans coût supplémentaire.

3 Expérimentations

L'algorithme a été implémenté avec Python 2.6. Toutes les expérimentations ont été effectuées sur un PC avec GNU/Linux, équipé d'un processeur 3 GHz Core 2 Duo et de 8 Go de mémoire (pas plus de 700 Mo ont été utilisés pendant les expérimentations). Les expérimentations ont été effectuées sur plusieurs jeux de données issues de sources bibliographiques et biologiques. Les performances de l'algorithme ont été étudiées sur un graphe provenant de DBLP contenant plus de 400 000 sommets. L'objectif était de répondre aux questions suivantes : Est-ce que le nouveau type de motif est pertinent ? Est-il possible pour les experts du domaine d'interpréter des motifs de la collection ? Est-ce que notre approche passe à l'échelle sur de grands jeux de données ?

3.1 Jeux de données bibliographiques

Les jeux de données bibliographiques ont été construits à partir de DBLP. Cette base de données contient des informations sur la plupart des conférences et journaux en informatique. Elle a déjà été largement utilisée comme jeu de test dans de nombreux travaux. Notons également que, d'après Newman et Girvan (2004), les réseaux de collaborations scientifiques ont



(a) 5 cliques avec les conférences ICDE, ICDM, KDD, PAKDD, SDM, et SIGMOD.

(b) 3 cliques avec les conférences EDBT, ICDE, ICDM, SDM, SIGMOD, et SSDBM

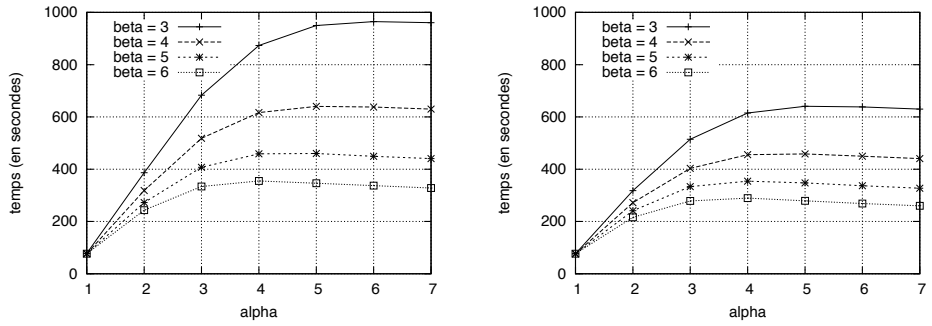
FIG. 1 – *Deux motifs extraits avec* $\alpha = 6$, $\beta = 3$, *et* $\kappa = 3$

des propriétés similaires à celle des autres réseaux sociaux. Nous avons construit notre jeu de données à partir de toutes les conférences depuis l'année 2000 incluse (les journaux n'ont pas été pris en compte). Un sommet représente un auteur et deux auteurs sont reliés par une arête s'ils ont co-écrit au moins deux articles ensemble. Un sommet est étiqueté avec les conférences dans lesquelles l'auteur correspondant a publié (e.g., ECML/PKDD, KDD).

Dans un premier jeu, nous avons voulu vérifier la pertinence des motifs extraits. Afin d'éviter de conserver les auteurs n'ayant publié qu'une seule fois, nous avons étiqueté les auteurs avec les conférences dans lesquels ils ont publié au moins deux fois. Les auteurs qui ne sont étiquetés par aucune conférence sont enlevés du graphe. Le graphe produit contient 117 526 sommets (auteurs), 233 863 arêtes (relation de co-auteurs sur au moins 2 articles) et 3 257 étiquettes différentes (conférences).

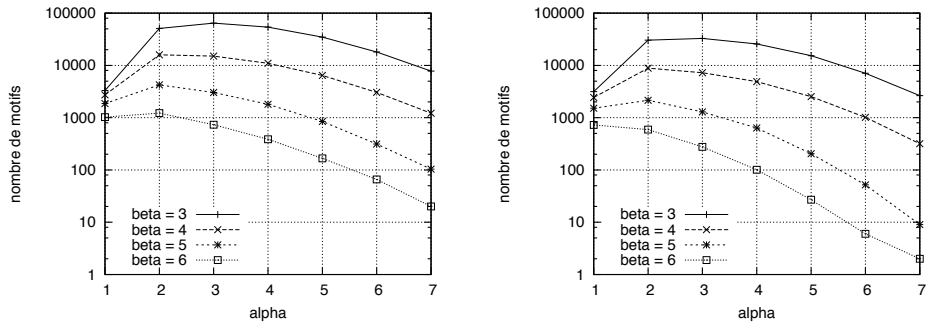
Sur ce jeu de données, nous avons recherché des EMCHs avec au moins 3 cliques contenant au moins 3 sommets et partageant 6 étiquettes ($\alpha = 6$, $\beta = 3$ et $\kappa = 3$). 80 motifs satisfont ces contraintes, et dans cette collection, 32 contiennent au moins une conférence en fouille de données parmi ICDM, KDD ou SDM. Nous discutons les deux motifs présentés sur la Figure 3.1. Le motif de la Figure 1(a) contient 5 cliques dont une formée par quatre auteurs : Jiawei Han, Jian Pei, Philip S. Yu et Ke Wang. Ces auteurs sont connus dans la communauté pour collaborer régulièrement et publier fréquemment. On peut noter une seconde information : le sommet correspondant à Philip S. Yu joue un rôle particulier dans le motif. Il fait partie de quatre des cinq cliques maximales du motif et est connecté à 7 des 8 auteurs du motif. Nous pensons que la découverte de sommets ayant un degré localement élevé est intéressante. L'autre motif présenté sur la Figure 1(b) est formé de 3 cliques. On remarque que l'une des cliques n'est pas connectée avec les deux autres. Christian Böhm, Hans-Peter Kriegel et Peer Kröger travaillent tous dans la même université en Allemagne, alors que les deux autres cliques sont formées par des chercheurs basés en Amérique du Nord et ayant déjà travaillé dans la même université. Ce type de structure est particulièrement intéressante car elle met en évidence des groupes qui ne sont pas en interaction mais qui partagent les mêmes intérêts.

Pour obtenir des résultats quantitatifs (nombre de motifs, performances en mémoire et en temps de calcul), nous avons construit un jeu de données avec plus de sommets et d'étiquettes



(a) $\kappa = 2$ (b) $\kappa = 3$

FIG. 2 – *Évolution du temps d'exécution en fonction de α, β et κ*



(a) $\kappa = 2$ (b) $\kappa = 3$

FIG. 3 – *Évolution du nombre de motifs (échelle logarithmique) en fonction de α, β et κ*

en étiquetant un auteur avec toutes les conférences dans lesquelles il a publié. Le jeu de données produit contient 479 067 sommets, 386 838 arêtes et 3 607 étiquettes différentes. Les résultats présentés sont obtenus en faisant varier les valeurs des paramètres α, β et κ.

La consommation maximale de mémoire de chaque extraction n'a jamais dépassé 700 Mo, et était de 657 Mo en moyenne sur toutes les extractions avec un écart type de 7 Mo. Concernant les performances en temps d'exécution, les Figures 2(a) et 2(b) montrent que les extractions peuvent être effectuées dans des temps raisonnables, même lorsque les contraintes sont peu sélectives. Le pire cas qui nécessite moins de 17 minutes est obtenu pour les valeurs $\alpha = 6$, $\beta = 3$, et $\kappa = 2$. Comme l'algorithme commence par énumérer tous les ensembles clos ayant au moins α étiquettes, les performances en temps d'exécution dépendent principalement du paramètre α lorsque α est petit. Lorsque α grandit, la contrainte $\mathbb{C}^{vert}$ dépendant de β et κ est utilisée pour éviter de générer tous les ensembles de α étiquettes et impacte donc de plus en plus la durée d'exécution. Enfin, concernant le nombre de motifs retournés, les Figures 3(a) et 3(b) montrent qu'il décroit rapidement lorsque les valeurs des paramètres augmentent. Pour

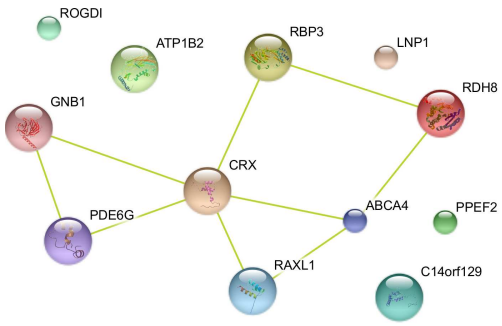


FIG. 4 – *Graphe d'interaction de gènes formant un EMCH avec 2 cliques de 3 gènes sur-exprimés dans 3 situations biologiques*

des valeurs du paramètre α supérieur à 1, lorsque β augmente de 2, le nombre de motifs retournés diminue d'au moins un ordre de magnitude.

3.2 Jeu de données biologiques

Le jeu de données réalisé à partir des résultats d'expérimentations biologiques utilise deux bases de données librement accessibles pour des travaux de recherche : STRING décrite par Jensen et al. (2009) et SQUAT décrite par Leyritz et al. (2008). STRING agrège des données d'interactions protéine/protéine à partir de différente sources (i.e., données génomiques, co-expression, bibliographie). Parmi ces interactions, nous n'avons conservé que celles ayant un seuil de confiance[3] supérieur ou égal à 400 (seuil de confiance par défaut dans STRING). SQUAT est une base de données de niveau d'expression des gènes dans différentes situations biologiques construite à partir d'expérimentations SAGE. Dans nos expérimentations nous n'avons utilisé que des gènes de l'espèce humaine. Comme SQUAT ne contient des informations que sur une partie des gènes de STRING, les protéines qui sont codées par des gènes non décrits dans SQUAT sont supprimées du graphe. Le jeu de données obtenu contient 4 923 sommets (gènes), 35 063 arêtes (interactions entre les protéines codées par les gènes correspondant) et 486 étiquettes différentes (situations biologiques).

L'intérêt principal de notre approche sur ce type de données est de pouvoir obtenir des ensembles de cliques non connectées mettant en évidence des groupes de gènes non liés fonctionnellement hormis par certaines protéines jouant le rôle de pont. Ce serait le cas par exemple pour un facteur de transcription qui active des gènes appartenant à différentes catégories fonctionnelles. Nous avons d'ailleurs extrait un motif de ce type (voir Figure 4) avec les paramètres $\alpha = 3, \beta = 3$, et $\kappa = 2$. Sa pertinence biologique qui ne peut être détaillée ici faute de place a été attestée par le chercheur biologiste co-auteur de l'article.

3. Cette confiance est une mesure spécifique à STRING variant entre 0 et 1000. Une confiance élevée indique que l'interaction est corroborée par plusieurs sources.

4 Travaux connexes

Deux types d'approches de fouille de données sont généralement utilisées sur les graphes dont les sommets sont décrits par des ensembles de propriétés (booléennes dans notre cas). Certaines comme Ge et al. (2008); Hanisch et al. (2002); Ulitsky et Shamir (2007) cherchent à réaliser un partitionnement des données pour en donner une vision globale. D'autres travaux se focalisent sur la recherche de motifs locaux, en utilisant généralement des approches de fouille sous contraintes. Moser et al. (2009) introduisent la recherche de motifs cohésifs dans des graphes dont les sommets sont associés à des vecteurs de propriétés. Les motifs cohésifs sont des sous-graphes qui satisfont des contraintes sur la cohésion du sous-graphe (i.e., ils partagent un ensemble suffisamment grand de propriétés), sa densité et sa connectivité. Miyoshi et al. (2009) étendent la tâche de recherche des motifs cohésifs à des propriétés quantitatives. Dans le travail de Berlingerio et al. (2009), des propriétés sont également associées aux sommets d'un graphe évoluant dans le temps. Les auteurs recherchent des règles caractérisant l'évolution du graphe à partir de motifs locaux. Par rapport à ces travaux, notre apport est de trouver des motifs impliquant plusieurs sous-graphes fortement connectés permettant de découvrir des relations non triviales entre eux.

Le problème de recherche d'EMCHs peut être vu comme une tâche d'extraction de motifs sous contraintes sur deux jeux de données simultanément (un graphe et une relation). De ce point de vue, notre approche est similaire à celle proposée par Crémilleux et al. (2009) définissant un cadre de travail générique pour travailler simultanément sur différents jeux de données. Cette approche n'a cependant pas été définie dans le cadre de l'extraction de motifs dans des graphes. Jiang et Pei (2009); Gallo et al. (2008) s'intéressent à d'autres problèmes sur le même type de données. Jiang et Pei (2009) recherchent des quasi-cliques qui persistent dans des collections de graphes. Gallo et al. (2008) traitent le problème de re-description des données en cherchant des sous-groupes ayant plusieurs descriptions. Nous pensons que ces approches sont complémentaires avec celle que nous proposons. Un Ensemble Maximal de Cliques Homogènes peut également être vu comme une collection où chaque composante (une clique) satisfait des contraintes à un niveau local alors que, au niveau global, d'autres contraintes caractérisent le motif (contraintes prenant en compte les cliques composantes). Suzuki (2004) propose les règles d'exception qui combinent trois motifs locaux (trois règles différentes). Récemment, plusieurs approches génériques - les "pattern teams" de Knobbe et Ho (2006), l'extraction d'ensembles de motifs basés sur des contraintes par De Raedt et Zimmermann (2007) et la programmation par contraintes pour les motifs n-aires de Khiari et al. (2010) - cherchent à trouver une petite collection de motifs intéressants à partir de motifs locaux initialement extraits.

5 Conclusion

Nous nous sommes intéressés à la tâche d'extraction dans un graphe dont les sommets sont associés à un ensemble d'étiquettes. Nous proposons de rechercher des Ensembles Maximaux de Cliques Homogènes qui sont des ensembles de cliques tel que tous les sommets d'un motif partagent suffisamment d'étiquettes. Nous avons proposé un algorithme complet exploitant les propriétés des contraintes utilisées pour formaliser ce nouveau domaine de motif. Pour finir, nous exposons les résultats d'expérimentations qui montrent que ces extractions peuvent

être réalisées sur des jeux de données réels et qu'elles produisent des motifs mettant en évidence des interactions non triviales entre les cliques. Une perspective immédiate de ce travail va consister à relaxer la contrainte de clique pour étudier l'extraction d'ensembles de quasi-cliques homogènes.

Remerciements : Ce travail est partiellement financé par le projet ANR-07-MDCO-014 Bingo2 (Knowledge Discovery For and By Inductive Queries).

Références

Berlingerio, M., F. Bonchi, B. Bringmann, et A. Gionis (2009). Mining graph evolution rules. In *European Conf. on Machine Learning and Princ. and Pract. of Knowl. Disc. in Databases (ECML/PKDD)*, pp. 115–130.

Bonchi, F., F. Giannotti, A. Mazzanti, et D. Pedreschi (2003). Exante : Anticipated data reduction in constrained patterns mining. In *European Conf. on Machine Learning and Princ. and Pract. of Knowl. Disc. in Databases (ECML/PKDD)*, pp. 59–70.

Crémilleux, B., A. Soulet, J. Klema, C. Hébert, et O. Gandrillon (2009). Discovering Knowledge from Local Patterns in SAGE data. In *Data Mining and Medical Knowledge Management : Cases and Applications*, pp. 251–267.

De Raedt, L. et A. Zimmermann (2007). Constraint-based pattern set mining. In *SIAM Data Mining Conf. (SDM)*.

Gallo, A., P. Miettinen, et H. Mannila (2008). Finding subgroups having several descriptions : Algorithms for redescription mining. In *SIAM Data Mining Conf. (SDM)*, pp. 334–345.

Ge, R., M. Ester, B. J. Gao, Z. Hu, B. Bhattacharya, et B. Ben-Moshe (2008). Joint cluster analysis of attribute data and relationship data : The connected k-center problem, algorithms and applications. *ACM Trans. Knowl. Discov. Data (TKDD) 2*(2), 1–35.

Hanisch, D., A. Zien, R. Zimmer, et T. Lengauer (2002). Co-clustering of biological networks and gene expression data. In *Int. Conf. on Intelligent Systems for Molecular Biology (ISMB)*, pp. 145–154.

Jensen, L. J., M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork, et C. von Mering (2009). STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic acids research 37*, 412–416.

Jiang, D. et J. Pei (2009). Mining frequent cross-graph quasi-cliques. *ACM Trans. Knowl. Discov. Data (TKDD) 2*(4), 1–42.

Khiari, M., P. Boizumault, et B. Crémilleux (2010). Combining CSP and Constraint-Based Mining for Pattern Discovery. In *Int. Conf. on Computational Science and Its Applications (ICCSA) (2)*, pp. 432–447.

Knobbe, A. J. et E. K. Y. Ho (2006). Pattern teams. In *European Conf. on Machine Learning and Princ. and Pract. of Knowl. Disc. in Databases (ECML/PKDD)*, pp. 577–584.

Ley, M. (2009). DBLP - Some Lessons Learned. *PVLDB 2*(2), 1493–1500.

Leyritz, J., S. Schicklin, S. Blachon, C. Keime, C. Robardet, J.-F. Boulicaut, J. Besson, R. Pensa G., et O. Gandrillon (2008). SQUAT : A web tool to mine human, murine and

avian SAGE data. *BMC Bioinformatics 9*(1), 378.

Miyoshi, Y., T. Ozaki, et T. Ohkawa (2009). Frequent pattern discovery from a single graph with quantitative itemsets. In *IEEE Int. Conf. on Data Mining (ICDM) Workshops*, pp. 527–532.

Moser, F., R. Colak, A. Rafiey, et M. Ester (2009). Mining Cohesive Patterns from Graphs with Feature Vectors. In *SIAM Data Mining Conf. (SDM)*, pp. 593–604.

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

Pasquier, N., Y. Bastide, R. Taouil, et L. Lakhal (1999). Efficient mining of association rules using closed itemset lattices. *Information Systems 24*(1), 25–46.

Pei, J., J. Han, et R. Mao (2000). Closet : An efficient algorithm for mining frequent closed itemsets. In *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pp. 21–30.

Suzuki, E. (2004). Undirected exception rule discovery as local pattern detection. In *Local Pattern Detection*, Volume 3539 of *LNCS*, pp. 207–216.

Tomita, E., A. Tanaka, et H. Takahashi (2006). The worst-case time complexity for generating all maximal cliques and computational experiments. *Theor. Comput. Sci. (TCS) 363*, 28–42.

Ulitsky, I. et R. Shamir (2007). Identification of functional modules using network topology and high-throughput data. *BMC Systems Biology 1*(1).

Wasserman, S. et K. Faust (1994). *Social network analysis : methods and applications*. Cambridge University Press.

Zaki, M. J. et M. Ogihara (1998). Theoretical foundations of association rules. In *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pp. 1–8.

Summary

We consider data mining methods on large graphs where a set of labels is associated to each vertex. A typical example concerns the social network of collaborating researchers where additional information concern their main publication targets (preferred conferences or journals). We investigate the extraction of sets of cliques such that the vertices in all subgraphs of a set share a large enough set of labels. We proposed a method to compute all *Maximal Homogeneous Clique Sets* that satisfy user-defined constraints on the number of separated cliques, on the size of the cliques, and on the number of labels shared by all the vertices. Empirical validation on real-life graphs illustrates the scalability of our approach and shows up interesting structures.