Extraction de co-variations entre des propriétés de sommets et leur position topologique dans un graphe attribué

Adriana Prado*, Marc Plantevit**, Céline Robardet*, Jean-François Boulicaut*

*Université de Lyon, CNRS, INSA-Lyon, LIRIS UMR5205, F-69621, France
**Université de Lyon, CNRS, Université Lyon 1, LIRIS UMR5205, F-69622, France

Résumé. L'analyse de grands réseaux est très étudiée en fouille de données. Toutefois, les approches existantes proposent une analyse soit à un niveau macroscopique (étude des propriétés globales comme la distribution des degrés), soit à un niveau microscopique (extraction de sous-graphes fréquents ou denses). Nous proposons une nouvelle méthode qui effectue une analyse intermédiaire permettant de découvrir des motifs regroupant des propriétés microscopiques et macroscopiques du réseau. Ces motifs capturent des co-variations entre des propriétés numériques relatives aux sommets. Par exemple, un motif mésoscopique dans un réseau de co-auteurs peut être *plus le nombre de publications à EGC est important, plus la centralité des sommets correspondants dans le réseau l'est également*. Notre contribution est multiple. D'abord, ce travail est le premier à exploiter conjointement des propriétés locales et des propriétés topologiques. De plus, nous produisons de nouvelles avancées dans le domaine de l'extraction de co-variations en revisitant les motifs émergents dans ce contexte. Enfin, nous rapportons une analyse d'un réseau bibliographique réel issu de DBLP.

1 Introduction

De nombreux phénomènes réels peuvent être modélisés par des réseaux où les sommets représentent les entités, les arêtes représentant des relations entre elles. Des attributs sont souvent associés aux sommets, fournissant des informations supplémentaires. Ce type de données est devenu ubiquitaire. Par conséquent, permettre la découverte de connaissances dans de telles données, comme par exemple dans de grands réseaux sociaux ou biologiques, est devenu un défi pour la communauté fouille de données. Dans cet article, nous illustrons notre proposition sur un réseau de co-auteurs où les sommets décrivent les auteurs, les arêtes encodent la relation de co-publication. Les attributs rattachés aux sommets décrivent des caractéristiques des auteurs comme leur affiliation, leurs domaines d'expertise. De nombreux travaux se sont attaqués à la fouille de graphes. On peut classer les approches actuelles en deux types d'analyse : (i) celles qui considèrent le graphe à un niveau *macroscopique* se focalisant sur des caractéristiques statistiques des propriétés topologiques pour décrire de grands réseaux (voir par exemple Albert et Barabási (2000)) ; (ii) celles qui effectuent des analyses à un niveau *microscopique*, se focalisant sur l'extraction de motifs locaux comme des cliques ou des quasi-cliques (Liu et Wong (2008)), de sous-graphes fréquents dans une collection de graphes (Jiang et Pei (2009);

Yan et Han (2002)) ou dans un grand graphe (voir Bringmann et Nijssen (2008); Kuramochi et Karypis (2005)).

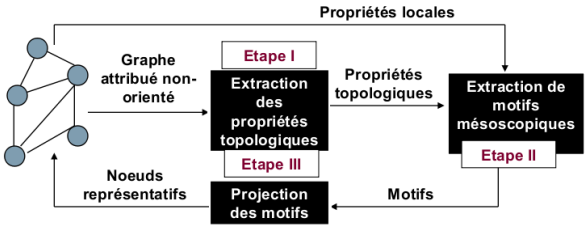


FIG. 1 – *Notre modèle MesoMining.*

Lorsque l'on considère de grands graphes « attribués » (i.e., les sommets possèdent plusieurs étiquettes ou attributs), un défi est de découvrir des régularités entre les attributs relatifs aux sommets et le rôle topologique des sommets dans le graphe. Pour résoudre ce problème, une solution est de se focaliser sur de petites parties du réseaux, ce qui conduit à une vue microscopique sur le réseau, alors qu'une autre solution est de considérer le réseau dans sa globalité, fournissant ainsi une vision macroscopique. Dans cet article, nous proposons de combiner ces deux niveaux d'analyse pour effectuer ce que nous définissons comme *analyse mésoscopique* de réseaux. Les travaux existants (Khan et al. (2010); Silva et al. (2010)) proposent seulement une vue microscopique en identifiant des ensembles de sommets qui partagent des attributs et qui se situent dans un même voisinage. Cependant, notre conviction est que le fait que des sommets possèdent des attributs locaux similaires peut être expliqué par leurs caractéristiques topologiques non nécessairement liées aux voisinages directs des sommets. En d'autres termes, même si des sommets n'appartiennent pas au même voisinage, ils peuvent partager des attributs locaux, ce qui peut s'expliquer par des propriétés topologiques, comme *avoir une forte « betweenness centrality »*.

Pour illustrer notre proposition et souligner son caractère novateur, considérons l'exemple jouet décrit par la Figure 2. Le tableau contient, pour chaque sommet (de A à P), les valeurs des attributs numériques X et Y, et aussi la valeur d'une propriété topologique, en l'occurrence une mesure de centralité (betweenness centrality notée BETW). Dans un réseau de co-auteurs, X et Y peuvent indiquer le nombre de fois que l'auteur a publié dans des conférences ou des journaux. Dans un tel graphe, on peut découvrir, par exemple, que plus la valeur de l'attribut X est importante et que plus celle de l'attribut Y est faible, alors la valeur de centralité d'un sommet est également plus forte. Dans notre proposition, un tel motif est noté $P = \{X^+, Y^-, \text{BETW}^+\}$. P combine des propriétés microscopiques (X et Y) et une propriété macroscopique $P = \{X^+, Y^-, \text{BETW}^+\}$.

Pour découvrir de tels motifs mésoscopiques, nous introduisons l'algorithme MesoMining. La Figure 1 illustre les différentes étapes de MesoMining qui prend en entrée un graphe attribué orienté ou non, comme par exemple celui de la Figure 2. Étant donné un tel graphe, il calcule d'abord un ensemble de propriétés topologiques pour chaque sommet. Dans la Figure 2, BETW est un exemple de propriété topologique. Dans une seconde étape, les motifs mésoscopiques sont extraits à partir des attributs relatifs aux sommets ainsi que leurs propriétés topologiques

Sommets	X	Y	BETW	Sommets	X	Y	BETW
A	3	3.1	0	**B**	7.3	1	1
C	3	3.2	0	**D**	10	-0.5	73
E	14	-2	166	**F**	8.8	0	52
G	7.1	1.4	0	**H**	6.5	2	0
I	12	-1.4	91	**J**	5	3	0
K	8	0.3	14	**L**	6	2.4	1
M	5.5	3	12	**N**	9	-0.1	54
O	2	4	0	**P**	1	4.5	0

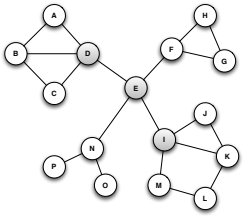


FIG. 2 – *Exemple jouet de graphe attribué.*

précédemment calculées. Dans la dernière étape, l'algorithme offre à l'utilisateur la capacité de visualiser chaque motif dans le graphe de départ en identifiant les sommets les plus représentatifs, c'est-à-dire ceux qui contribuent le plus au support du motif. Par exemple, $\{E, I, D\}$ sont les sommets représentatifs du motifs $\{X^+, Y^-, \text{BETW}^+\}$ dans l'exemple jouet.

Notre contribution est multiple. Nous proposons un nouveau type d'analyse de graphes qui exploite à la fois les attributs locaux et les propriétés topologiques des sommets dans les graphes. Pour effectuer une telle analyse, notre contribution se situe également au niveau de l'extraction de motifs capturant des co-variations. En considérant à la fois les variations positives et négatives, nous revisitons le cadre des motifs émergents dans ce nouveau contexte. Nous offrons aussi à l'utilisateur la possibilité de pouvoir visualiser les motifs dans le graphe original. Ensuite, nous rapportons une étude de cas sur un grand réseau réel de co-auteurs.

2 Étape I - Extraction des propriétés topologiques d'un réseau

Notre analyse mésoscopique requiert un graphe attribué non orienté [1] $G = (V, E, L)$ où V est l'ensemble des sommets, E l'ensemble des arêtes, et $L = \{l_1, \cdots, l_m\}$ est un ensemble de m attributs associés à chaque sommet de V et qui sont des propriétés des entités représentées par les sommets. Chaque sommet $v \in V$ est associé à un vecteur $(l_1(v), \cdots, l_m(v))$ où $l_j(v)$ est la valeur du sommet v sur l'attribut l_j. Ces attributs peuvent être de type catégoriel, ordinal, ou numérique et sont appelés attributs locaux dans le reste de l'article. D'importantes propriétés des sommets sont aussi encodées par les arêtes du graphe qui décrivent des relations entre les sommets. À partir des arêtes, nous pouvons calculer des propriétés topologiques qui synthétisent le rôle de chaque sommet dans le graphe. Le but de l'étape I est de calculer de telles propriétés topologiques qui sont représentées par l'ensemble T. Les différentes propriétés qui nous intéressent se situent à différents niveaux allant du niveau microscopique au niveau macroscopique. Les propriétés microscopiques relatives à un sommet sont celles qui s'appuient sur le sommet lui-même et son voisinage immédiat alors que les propriétés macroscopiques nécessitent quant à elles la prise en compte du graphe dans son ensemble comme par exemple le calcul du plus court chemin entre chaque paire de sommets. Ci-dessous, nous présentons succinctement ces propriétés.

1. Dans cet article, nous nous concentrons sur des graphes non orientés mais l'extension de l'approche à des graphes orientés est immédiate.

Les propriétés microscopiques s'appuient sur le voisinage direct d'un sommet. Par exemple le degré d'un sommet u. Quand celui-ci est normalisé par le nombre maximum d'arêtes ($|V|-1$), cette mesure est appelée la centralité de degré noté DEGREE. Le coefficient de clustering, noté CLUSCOEF, qui évalue la connectivité des voisins d'un sommet donné est une autre mesure de ce type. D'autres propriétés topologiques microscopiques s'appuient sur un voisinage un peu plus étendu. Pour mieux comprendre la structure du voisinage d'un sommet donné, nous considérons la taille de la plus grande quasi-clique (SZQCLIQ) dans laquelle le sommet est inclus et qui a une densité supérieure à un seuil donné. De plus, nous utilisons aussi le nombre de telles quasi-cliques dans lesquelles le sommet apparaît (NBQCLIQ).
Les propriétés macroscopiques caractérisent un sommet en considérant le graphe entier. Des communautés peuvent être calculées en s'appuyant sur la mesure de modularité de Newman (2004). Dans cet article, nous considérons la taille de la communauté dans laquelle le sommet appartient (SZCOM) comme une propriété macroscopique. Nous considérons aussi diverses mesures de centralité (voir Freeman (1977)) comme la « Closeness centrality »(CLOSE), la « Betweenness centrality »(BETW) et la centralité basée sur les vecteurs propres de la matrice d'adjacence du graphe (EGVECT). La dernière mesure que nous considérons est l'index PAGERANK qui s'appuie sur des marches aléatoires sur les sommets du graphe. Cet index reflète la probabilité qu'une marche aléatoire passe par le sommet lui-même.

À la fin de l'étape I, la relation encodée par E est résumée par un ensemble de propriétés T qui associe à chaque sommet $v \in V$ un vecteur $(t_1(v), \cdots, t_n(v))$ où $t_j(v)$ est la valeur de propriété topologique t_j pour le sommet v.

3 Étape II - Extraction de motifs mésoscopiques d'un réseau

En considérant simultanément les attributs locaux et les propriétés topologiques des sommets du graphe, le but de cette étape est de découvrir des motifs mésoscopiques, c'est-à-dire des co-variations possibles sur les propriétés $\mathcal{A} = L \cup T = \{A_1, \ldots, A_k\}$, où $k = m + n$. Un tel motif est noté $P = {A_1}^{s_1}, \cdots, {A_j}^{s_j}, \cdots, {A_\ell}^{s_\ell}$, où $A_j \in \mathcal{A}$ est un attribut et $s_j \in \{+, -\}$ le signe de variation.

3.1 Support d'un motif mésoscopique

Des attributs signés co-varient s'ils sont supportés par un nombre significatif de couples de sommets, ce qui est défini formellement de la façon suivante :

Définition 1 ($Supp_\tau$) *Le support d'un motif mésoscopique est calculé à partir d'une généralisation de la mesure du τ de Kendall comme suit :*

$$Supp_\tau(P) = \frac{|\{(u,v) \in V^2 \mid \forall A^s \in P : A(u) \rhd_s A(v)\}|}{\begin{pmatrix} |V| \\ 2 \end{pmatrix}}$$

Si $s = +$, $\rhd_s$ signifie $<$, autrement $>$. Comme le dénominateur est une constante, dans la suite nous utilisons

$Supp(P) = |\{(u,v) \in V^2 \mid \forall A^s \in P : A(u) \rhd_s A(v)\}|$

Cette mesure retourne le nombre de couples de sommets (u, v) tels que u est strictement inférieur à v sur les attributs positifs (ceux qui ont l'exposant $+$) et strictement supérieur à v sur les attributs négatifs (ceux avec l'exposant $-$). Comme indiqué dans Calders et al. (2006), $Supp$ est une mesure anti-monotone pour les attributs positifs. C'est encore le cas lorsqu'on traite des attributs négatifs : l'ajout d'un attribut négatif A^- à un motif P conduit à un support inférieur ou égal à celui de P puisque les couples (u, v) qui supportent P doivent aussi satisfaire la condition $A(u) > A(v)$.

En outre, le support de motifs contenant des attributs négatifs peut être déduit à partir de celui d'autres motifs porteur de la même sémantique mais sous une forme symétrique :

Propriété 1 (Support de motifs symétriques) *Soient P un motif mésoscopique et $\overline{P}$ son symétrique, c'est-à-dire $\forall A_j^{s_j} \in P$, $A_j^{\overline{s_j}} \in \overline{P}$, avec $\overline{s_j} = \{+, -\} \setminus \{s_j\}$. Si un couple (u, v) de V^2 participe au support de P, alors le couple (v, u) participe au support de $\overline{P}$. Par conséquent, on a $Supp(P) = Supp(\overline{P})$.*

Afin d'éviter le calcul inutile de motifs mésoscopiques dupliqués, nous exploitons la propriété 1. L'équation (1) indique le nombre de motifs possibles qui peuvent être construits sur $\mathcal{A}$ en évitant les formes dupliquées (les symétriques) :

$$2^{|\mathcal{A}|} - 1 + \sum_{k=2}^{|\mathcal{A}|} \left(\begin{array}{c} |\mathcal{A}| \\ k \end{array} \right) \times (2^{k-1} - 1) \tag{1}$$

$2^{|\mathcal{A}|} - 1$ est le nombre de motifs contenant uniquement des attributs positifs. Le reste représente ceux qui combinent des attributs positifs et négatifs. La tâche accomplie dans l'étape II est ainsi formellement définie ci-dessous :

Définition 2 (Extraction de Motifs Mésoscopiques) *Étant donnés un ensemble d'attributs $\mathcal{A}$ définis sur l'ensemble des sommets V de G, et un seuil minimum de support $minsup$, le problème de l'extraction de motifs mésoscopiques est de découvrir l'ensemble complet, mais sans duplication, des ensembles d'attributs signés qui co-varient sur au moins $minsup$ couples de sommets.*

3.2 Motifs mésoscopiques émergents

Il est bien connu que la fouille de motifs peut engendrer la production d'un trop grand nombre de motifs. Il est donc primordial de permettre à l'utilisateur de naviguer dans le processus de fouille de données en ciblant, par exemple, des motifs par rapport à un attribut spécifique. Dans cette optique, nous offrons à l'utilisateur la possibilité d'obtenir les motifs les plus discriminants par rapport à un attribut cible. Nous revisitons ainsi le cadre des motifs *émergents* de Dong et Li (1999) dans le contexte des motifs mésoscopiques. L'extraction de motifs émergents permet de mettre en évidence les motifs dont la fréquence est significativement plus forte (par rapport à un seuil noté $mingr$) dans une partie du graphe que dans le reste du graphe. Cette partie du graphe est identifiée par un ensemble de couples de sommets caractérisés par un attribut numérique ou catégoriel appelé attribut de classe.

Dans cet article, nous nous intéressons uniquement au cas où l'attribut de classe est de type numérique. Si la partie du graphe que l'on considère est spécifiée par un attribut de classe C

de type numérique, alors chaque couple de sommets (u, v) appartient à une seule classe parmi celles définies ci-dessous :

- $C^+ = \{(u, v) \mid (C(u) < C(v)) \wedge (u \neq v)\}$
- $C^- = \{(u, v) \mid (C(u) > C(v)) \wedge (u \neq v)\}$
- $C^= = \{(u, v) \mid (C(u) = C(v)) \wedge (u \neq v)\}$

Étant données ces classes, le taux de croissance (gr) d'un motif mésoscopique P par rapport à une classe $C^\star$ ($\star \in \{+, -, =\}$) est $gr(P, C^\star) = \frac{Supp(P \cup C^\star)}{Supp(P \cup C^{\overline{\star}})} \cdot \frac{Supp(C^{\overline{\star}})}{Supp(C^\star)}$ où $C^{\overline{\star}}$ est l'ensemble des couples de sommets (u, v), $u \neq v$ qui n'appartiennent pas à $C^\star$. L'intuition est d'identifier les motifs mésoscopiques qui sont principalement supportés par des couples qui sont positifs, négatifs ou stable sur l'attribut de classe.

4 Étape III - Projection du motif et résumé de l'approche

Si nous regardons l'ensemble des couples de sommets (u, v) qui contribuent au support d'un motif mésoscopique P comme un ensemble d'arcs (de u vers v), alors P est un motif supporté par un ensemble d'arcs E' qui forme avec V un nouveau graphe orienté $G' = (V, E')$. Ce graphe n'est pas nécessairement connexe ni un sous-graphe de G. C'est la caractéristique principale d'un motif mésoscopique : les couples de sommets participants au support ne sont pas contraints. Pour cette raison, l'utilisateur peut être intéressé par l'identification de sommets qui contribuent le plus au support d'un motif, lui donnant ainsi la possibilité de projeter le motif dans le graphe de départ. Dans ce but, on peut rechercher les sommets les plus représentatifs d'un motif, c'est-à-dire les sommets avec un haut degré entrant et un faible degré sortant dans G'. Nous pouvons considérer les sommets qui ont un degré entrant supérieur à leur degré sortant ou simplement se focaliser sur les k sommets qui ont le plus fort degré entrant pour avoir les k sommets plus dominants. Cet ensemble de sommets, que nous appelons les top-k sommets représentatifs, peut être facilement extrait lors du calcul du support d'un motif.

Nous avons présenté les principales étapes de notre analyse mésoscopique. D'un point de vue algorithmique, l'entrée de MesoMining est un graphe non orienté $G = (V, E, L)$ et deux paramètres : le seuil de support $minsup$ et k le nombre de sommets représentatifs souhaités pour chaque motif. Pour l'étape I de notre modèle, nous utilisons la librairie SNAP développée par Jure Leskovec [2] et QUICK, l'extracteur de quasi-cliques de l'état de l'art de Liu et Wong (2008). La partie restante de l'algorithme consiste à exécuter les étapes II et III en une seule tâche. Le calcul des motifs mésoscopiques, comme définis dans la Section 3, est fait de façon similaire à la stratégie utilisée dans l'algorithme ECLAT de Zaki. Plus précisément, tous les sous-ensembles d'un motif P sont toujours évalués avant le motif P lui-même. De cette façon, en stockant tous les motifs fréquents dans l'arbre de hachage $\mathcal{M}$, nous pouvons ainsi vérifier les contraintes anti-monotones à la volée. Nous commençons par énumérer les singletons contenant uniquement des attributs positifs pour éviter la génération de motifs dupliqués. Des motifs plus grands sont récursivement générés. La mesure de support étant définie comme le nombre de couples qui supportent un motif donné P, le calcul de cette mesure est quadratique par rapport au nombre de sommets. Toutefois, comme proposé dans Calders et al. (2006), une recherche dirigée de tous les sommets qui ont des valeurs supérieures ou inférieures sur tous les attributs de P est mise en œuvre en utilisant des « range trees ». Tout en calculant le support

2. `http://snap.stanford.edu/index.html`

d'un motif P, nous maintenons la liste des top k sommets représentatifs de P. Ceci est mis en œuvre dans un tas afin de maintenir cette liste en utilisant des opérations en $O(\log k)$. Enfin, la définition d'une borne supérieure sur le support, calculable en temps linéaire sur le nombre de sommets, permet d'éviter de calculer le support exact de certains motifs non fréquents.

5 Analyse d'un réseau bibliographique

Dans cette section, nous rapportons les résultats de l'analyse mésoscopique d'un réseau de co-auteurs construit à partir de la base DBLP [3]. Le but de cette étude expérimentale est d'évaluer l'impact de certaines publications sur les propriétés topologiques du graphe de co-auteurs correspondant. L'algorithme MesoMining a été implémenté en C en étendant celui de Calders et al. (2006). A partir de la base DBLP, nous avons construit un graphe de co-auteurs où les sommets représentent les auteurs qui ont publié au moins un article dans l'ensemble de conférences [4] ou journaux [5] entre janvier 1990 et février 2011. Une arête du graphe indique que les deux auteurs associés ont co-écrit un article. Nous avons retenu les principales conférences et revues des communautés de base de données (BD) et de fouille de données (FD). Les propriétés locales (L) sont le nombre de publications des auteurs dans chacune des conférences ou revues sélectionnées. Les propriétés topologiques (T) sont calculées dans la première étape de l'analyse. Les principales caractéristiques du graphe construit sont données dans le Tableau 1. Beaucoup de ces propriétés ont un écart-type plus grand que leur moyenne, ce qui suggère qu'elles suivent une loi de puissance. Nous avons également calculé la matrice de corrélation entre les propriétés. VLDB, ICDE et SIGMOD ont une corrélation positive supérieure à $0,7$. BETW, DEGREE et PAGERANK d'un côté et, SZQCLIQ et NBQCLIQ de l'autre, sont aussi fortement corrélées. Les propriétés SAC, Comm. of ACM, IEEE Int. Sys, CLOSE et CLUSCOEF sont quant à elles non corrélées aux autres (corrélation inférieure à $0,2$). Étant donné ce graphe, nous rapportons cette analyse par rapport aux questions suivantes : Comment notre algorithme se comporte sur un tel graphe ? Pouvons-nous trouver des motifs intéressants entre les publications uniquement ? Quels sont les motifs mésoscopiques intéressants ?

5.1 Motifs sur les publications

Considérons d'abord uniquement les propriétés locales des sommets (seulement les publications). Dans ce cas, l'extraction de tous les motifs mésoscopiques pour un seuil de support à 1% a pris 70 secondes, 263 motifs sont découverts dont 58 (22%) contenant des attributs négatifs. Nous avons réalisé un clustering de cette sortie en appliquant l'algorithme K-means avec la distance cosinus et en utilisant la mesure silhouette pour valider le nombre de groupes. Dix groupes, dont les principales caractéristiques sont données dans le Tableau 2, ont été trouvés. Nous pouvons observer que la majorité des groupes sont homogènes décrivant des publications soit en BD soit en FD. Par exemple, les groupes 1, 2, 6, et 9 sont liés à des publi-

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

4. Conférences sélectionnées : KDD, ICDM, ECML/PKDD, PAKDD, SIAM DM, AAAI, ICML, IJCAI, IDA, DASFAA, VLDB, CIKM, SIGMOD, PODS, ICDE, EDBT, ICDT, SAC.

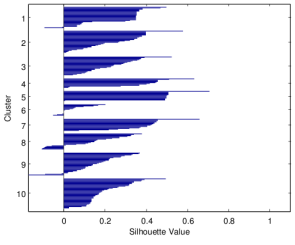
5. Journaux sélectionnés : *Data Min. Knowl. Discov.* (DMKD), IEEE TKDE, IEEE Int. Sys., SIGKDD Exp., Comm. ACM, IDA J., KAIS, SADM, PVLDB, VLDB J., ACM TKDD.

Propriétés globales du graphe	Valeurs			
#Sommets	42 252			
#Arêtes	210 320			
Densité	2×10^{-4}			
#Composant Connexes	577			
#Communautés	1016			
Propriétés associées aux sommets	**Min**	**Max**	**Mean**	**Std. Dev.**
Degré	0	304	9.73	14.22
DEGREE	0	7.3×10^{-3}	2.4×10^{-4}	3.4×10^{-4}
BETW	0	2.6×10^{6}	1.4×10^{5}	5.7×10^{5}
CLOSE	0	1	0.024	0.137
EGVECT	0	0.003	2.36×10^{-5}	9.91×10^{-5}
PAGERANK	0	21.53	0.98	0.98
CLUSCOEF	0	1	0.31	0.29
NBQCLIQ	0	4.6×10^{5}	2.2×10^{2}	7.8×10^{3}
SZQCLIQ	0	35	2.75	4.83
Propriétés sur les structures du graphes	**Min**	**Max**	**Mean**	**Std. Dev.**
Taille des composantes connexes	2	39767	71.50	1.65×10^{3}
Taille des quasi-cliques	0	35	16.26	2.50
SZCOM	0	9342	40.67	5×10^{2}

TAB. 1 – *Caractéristiques principales du graphe de co-auteurs.*

cations en FD alors que les groupes 3, 8, et 10 sont clairement associés au domaine des BD. Les autres groupes sont liés à des communications particulières.

Groupe	# patterns	Publications représentatives	Groupe	# patterns	Publications représentatives
1	32	SAC$^-$, IJCAI$^+$	6	17	KAIS$^+$, SDM$^+$, PAKDD$^+$, KDD$^+$
2	34	CIKM$^+$, PAKDD$^+$	7	18	IEEE TKDE$^+$
3	28	SIGMOD$^+$	8	24	VLDB$^+$, VLDBJ$^+$, PVLDB$^+$
4	15	AAAI$^+$	9	34	ICDM$^+$, PKDD$^+$, KDD$^+$
5	15	CommACM$^+$	10	46	ICDE$^+$, SIGMOD$^+$, TKDE$^+$, VLDB$^+$



TAB. 2 – *Les différents groupes de motifs mésoscopiques et la silhouette associée.*

Il est intéressant de remarquer que 20 de ces motifs contiennent l'attribut SAC^- comme {SAC$^-$, KDD$^+$}, {SAC$^-$,VLDB$^+$}, et {SAC$^-$, SIGMOD$^+$}. Ceci peut s'expliquer par le fait que les scientifiques qui publient à SAC ne font pas nécessairement partie des communautés BD ou FD. Puisque le spectre couvert par SAC est bien plus général que ces deux domaines, il n'est pas surprenant que beaucoup d'auteurs aient plusieurs publications à SAC et peu ou pas de publications en BD ou FD.

5.2 Motifs sur les propriétés topologiques.

Nous recherchons ici les motifs émergents par rapport au PAGERANK des auteurs. Avec un seuil de support de 1% et un taux de croissance minimum de 3, notre algorithme a découvert 310 motifs en 15 minutes. Le motif le plus émergent est {DEGREE$^+$, BETW$^+$, CLUSCOEF$^-$, NBQCLIQ$^+$, SZQCLIQ$^+$, PAGERANK$^+$}. Toutes les propriétés topologiques sont positives excepté CLUSCOEF. Comme nous l'avons énoncé, PAGERANK est fortement corrélé à DEGREE et

BETW. Il n'est donc pas surprenant de les retrouver ensemble dans ce motif. La présence de l'attribut CLUSCOEF$^-$ peut s'interpréter de la façon suivante : *plus fort est le* PAGERANK *des auteurs, plus faible est la connectivité de leurs co-auteurs.* En d'autres termes, les auteurs avec un fort PAGERANK ont beaucoup de co-auteurs qui n'ont jamais publié ensemble. Les auteurs qui publient avec de nombreux doctorants en sont des exemples typiques.

La Figure 3 montre les 30 sommets les plus représentatifs du motif ci-dessus. Nous pouvons voir que certains ont un voisinage commun (S. Abiteboul, H. Garcia Molina, et M. J. Carey) alors que d'autres sont plus « isolés ». La partie dense du graphe identifie la communauté BD tandis que les autres parties identifient d'autres domaines comme FD, l'apprentissage, le TAL, ou le web sémantique. Cette projection du motif dans le graphe illustre bien que le motif n'apparaît pas uniquement dans un voisinage mais qu'il est diffus dans tout le graphe.

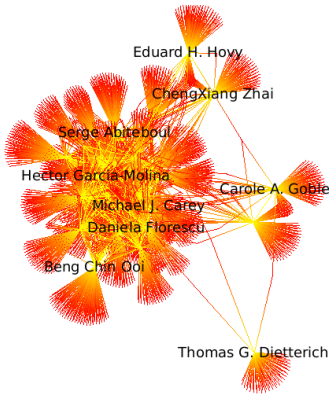


FIG. 3 – *Top* 30 *sommets représentatifs (en jaune) du motif {*DEGREE$^+$, BETW$^+$, CLUSCOEF$^-$, NBQCLIQ$^+$, SZQCLIQ$^+$, PAGERANK$^+$*} et leurs* 2544 *voisins (en rouge).*

5.3 Motifs mésoscopiques sur les propriétés locales et topologiques

Nous considérons encore les motifs émergents par rapport au PAGERANK des auteurs avec pour paramètres $minsup = 1\%$ et $mingr = 3$. Notre algorithme a pris 6 heures pour extraire 4313 motifs mésoscopiques. Les 5 motifs les plus émergents par rapport à PAGERANK$^+$ sont présentés dans le Tableau 3. Notons que dans tous, les propriétés topologiques sont similaires à celles présentées dans la sous-section précédente. Ici, nous pouvons conclure que l'effet de CLUSCOEF$^-$ est plus fort chez les auteurs d'articles ICDE et IEEE TKDE. Puisque BETW, DEGREE et PAGERANK sont corrélés, il n'est pas surprenant que {PAGERANK$^+$,BETW$^+$} et {PAGERANK$^+$,DEGREE$^+$} soient des motifs émergents avec des taux de croissance respectifs de 124.69 et 584.46. De plus, il est aussi intéressant de voir quelles sont les conférences qui ont le plus fort impact positif sur le taux de croissance. Dans ce but, pour chaque conférence C et pour chacun de ces deux motifs émergents notés X, nous calculons le ratio du taux de croissance de CX sur celui de X. Le Tableau 4(A) donne les 5 publications les plus « impactantes »(celles qui ont le plus fort ratio). Avec surprise, nous observons que les conférences de FD ont un plus fort impact sur {PAGERANK$^+$,DEGREE$^+$} alors que les conférences de BD

ont plus d'influence sur le taux de croissance de {PAGERANK$^+$,BETW$^+$}. Puisque la fouille de données est à l'intersection de nombreux domaines, ces résultats peuvent s'expliquer par le fait que les chercheurs en fouille de données publient dans d'autres domaines tels que les BD et l'apprentissage. D'un autre côté, la BD est un domaine plus ancien et mieux établi, cela peut être la raison pour laquelle les auteurs de BD ont une plus forte centralité dans le graphe. Pour la publication la plus impactante, nous avons identifié les 5 auteurs les plus représentatifs dans le Tableau 4 (B).

ICDE$^+$, DEGREE$^+$, BETW$^+$, CLUSCOEF$^-$, NBQCLIQ$^+$, SZQCLIQ$^+$
IEEE TKDE$^+$, DEGREE$^+$,BETW$^+$, CLUSCOEF$^-$, NBQCLIQ$^+$, SZQCLIQ$^+$
ICDE$^+$, DEGREE$^+$,CLOSE$^+$, BETW$^+$, CLUSCOEF$^-$, NBQCLIQ$^+$, SZQCLIQ$^+$
ICDE$^+$, DEGREE$^+$,BETW$^+$, EGVECT$^+$, CLUSCOEF$^-$, NBQCLIQ$^+$, SZQCLIQ$^+$
IEEE TKDE$^+$, DEGREE$^+$,CLOSE$^+$, BETW$^+$, CLUSCOEF$^-$, NBQCLIQ$^+$, SZQCLIQ$^+$

TAB. 3 – *Top* 5 *motifs mésoscopiques par rapport à* PAGERANK$^+$.

(A)

Rank	DEGREE		BETW	
	Publication	Factor	Publication	Factor
1	ECML/PKDD$^+$	2.5	PVLDB$^+$	5.67
2	IEEE TKDE$^+$	2.28	EDBT$^+$	5.11
3	PAKDD$^+$	2.21	VLDB J.$^+$	4.35
4	DASFAA$^+$	2.09	SIGMOD$^+$	4.25
5	ICDM$^+$	1.95	ICDE$^+$	3.42

(B)

PAGERANK$^+$ DEGREE$^+$ ECML/PKDD$^+$	PAGERANK$^+$ BETW$^+$ PVLDB$^+$
Christos Faloutsos	Gerhard Weikum
Jiawei Han	Jiawei Han
Philip S. Yu	David Maier
Bing Liu	Philip S. Yu
C. Lee Giles	Hector Garcia-Molina

TAB. 4 – *Top* 5 *publications « impactantes » dans l'émergence de {*DEGREE$^+$*} et {*BETW$^+$*} pour* PAGERANK$^+$ *(A) et les top* 5 *auteurs représentatifs (B).*

6 Travaux connexes

Il existe principalement deux grands types d'approches pour analyser des graphes. Soit les graphes sont étudiés à un niveau macroscopique en considérant des propriétés statistiques du graphe (diamètre, distribution des degrés) (voir Albert et Barabási (2000); Chakrabarti et al. (2004)), soit des propriétés plus subtiles sont découvertes en s'appuyant sur la fouille de motifs locaux. Des travaux récents traitent des graphes attribués qui sont porteurs de plus d'information. Dans de tels graphes, des informations supplémentaires sont disponibles sur des attributs propres aux sommets. L'approche pionnière de Moser et al. (2009) souligne bien le fait que les attributs des sommets contiennent des informations complémentaires qui ne peuvent pas être dérivées par la structure du graphe même. Les auteurs proposent ainsi une méthode pour trouver des sous-graphes denses homogènes (des sous-graphes qui partagent un ensemble suffisant de propriétés booléennes). Mougel et al. (2010) propose une nouvelle tâche de fouille qui vise à extraire des ensembles de cliques homogènes. Silva et al. (2010) propose d'extraire des paires de sous-graphes denses et d'ensembles de propriétés booléennes telles que les propriétés booléennes sont fortement associées à la densité du sous-graphe. Une autre approche présentée dans Khan et al. (2010) où un plus large voisinage est considéré grâce à une relaxation des contraintes portant sur la structure des sous-graphes. Une approche probabiliste est utilisée pour construire à la fois le voisinage d'un sommet mais également propager l'information

sur son voisinage. Notons que ces approches n'utilisent qu'un seul type d'information topologique basé sur le voisinage des sommets et ne permettent pas de traiter des attributs de type numérique comme dans notre proposition. Les approches basées sur une analyse statistique globale (voir Freeman (1977)) d'un graphe considèrent de nombreuses mesures pour décrire la topologie du graphe mais ne tirent pas avantage des informations portées par les attributs des sommets. En outre, les approches basées sur les motifs locaux ne considèrent pas les attributs numériques ni les propriétés topologiques macroscopiques. À notre connaissance, notre proposition est le premier essai visant à coupler analyse macroscopique et analyse microscopique grâce aux de motifs mésoscopiques (émergents).

Les motifs véhiculant des co-variations sont aussi connus sous le nom de motifs graduels (voir Do et al. (2010)) ou de « rank-correlated itemsets »(Calders et al. (2006)). Dans Do et al. (2010), les auteurs utilisent une mesure de support basée sur la longueur du plus long chemin entre les objets ordonnés. Cette mesure a quelques inconvénients à la fois sur des aspects calculatoires et sémantiques. Calders et al. (2006) introduisent une nouvelle mesure de support basée sur la mesure statistique bien établie du τ de Kendall. Toutefois, leur approche ne permet pas de combiner des co-variations mêlant attributs positifs et négatifs ainsi que la notion de motifs émergents.

7 Conclusion

Nous avons proposé MesoMining, un algorithme pour une analyse mésoscopique des réseaux. Les motifs découverts contiennent à la fois des propriétés locales et des propriétés topologiques des sommets. De plus, en revisitant les motifs émergent dans ce nouveau contexte et en identifiant les sommets représentatifs des motifs mésoscopiques, nous offrons une meilleure interaction avec l'utilisateur final. L'analyse d'un réseau de co-auteurs issu de DBLP illustre les plus-values de notre proposition. Des motifs qui font sens sont découverts et ne pourraient pas l'être par les approches actuelles. Notre travail débouche sur de nombreuses perspectives comme la fouille de multi-graphes. Une autre perspective intéressante est d'appliquer notre analyse mésoscopique au cas des graphes dynamiques, en identifiant par exemple les motifs mésoscopiques inattendus au cours du temps.

Remerciements

Ce travail est partiellement financé par l'Agence Nationale de la Recherche (ANR) via le projet FOSTER (ANR-2010-COSI-012-02). Les expériences ont été réalisées sur les ordinateurs du Centre de Calcul de l'IN2P3 (CC-IN2P3) (USR 6402).

Références

Albert, R. et A.-L. Barabási (2000). Topology of complex networks : Local events and universality. *Phys. Rev. 85*, 5234–5237.

Bringmann, B. et S. Nijssen (2008). What is frequent in a single graph ? In *PAKDD*, pp. 858–863.

Calders, T., B. Goethals, et S. Jaroszewicz (2006). Mining rank-correlated sets of numerical attributes. In *KDD*, pp. 96.

Chakrabarti, D., Y. Zhan, et C. Faloutsos (2004). R-MAT : A recursive model for graph mining. In *SIAM SDM*.

Do, T. D. T., A. Laurent, et A. Termier (2010). PGLCM : Efficient parallel mining of closed frequent gradual itemsets. In *IEEE ICDM*, pp. 138–147.

Dong, G. et J. Li (1999). Efficient mining of emerging patterns : Discovering trends and differences. In *KDD*, pp. 43–52.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry 40*(1), 35–41.

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

Khan, A., X. Yan, et K.-L. Wu (2010). Towards proximity pattern mining in large graphs. In *SIGMOD*, pp. 867–878.

Kuramochi, M. et G. Karypis (2005). Finding frequent patterns in a large sparse graph. *DMKD 11*, 243–271.

Liu, G. et L. Wong (2008). Effective pruning techniques for mining quasi-cliques. In *ECML/PKDD*, pp. 33–49.

Moser, F., R. Colak, A. Rafiey, et M. Ester (2009). Mining cohesive patterns from graphs with feature vectors. In *SIAM SDM*, pp. 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 *Workshop on Analysis of Complex NEtworks (ACNE'10) @ECML-PKDD 2010*.

Newman, M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Phys. Rev. E 69*, 066133.

Silva, A., J. Wagner Meira, et M. J. Zaki (2010). Structural correlation pattern mining for large graphs. In *8th Workshop on Mining and Learning with Graphs*.

Yan, X. et J. Han (2002). gSpan : Graph-Based Substructure Pattern Mining. In *IEEE ICDM*, pp. 721–724.

Summary

The analysis of large networks is a popular data mining task. However, in most of the graph mining approaches, the analysis is focused either on a macroscopic level (e.g., studying global properties as degree distributions) or on a microscopic level (e.g., looking for frequent subgraphs or dense subgraphs). We present an algorithm that performs an intermediate analysis of graphs by discovering patterns among microscopic and macroscopic properties over large networks. Such patterns capture co-variations between numerical vertex properties. For instance, such a pattern in a co-authorship graph could be *the higher the number of publications in EGC, the higher the centrality of the vertex within the graph*. Our contributions is threefold. First, this work is the first attempt to exploit local and topological properties over graphs. Next, we provide new insights into co-variation pattern mining. Finally, we report a case study on a large real-life co-authorship network.