Clusters dans les réseaux sociaux : intersections entre liens conceptuels fréquents et communautés

Erick Stattner, Martine Collard

Laboratoire LAMIA
Université des Antilles et de la Guyane, France
{estattne, mcollard}@univ-ag.fr

Résumé. La recherche de liens conceptuels fréquents (FCL) est une nouvelle approche de clustering de réseaux, qui exploite à la fois la structure et les attributs des noeuds. Bien que les travaux récents se soient déjà intéressés à l'optimisation des algorithmes de recherche des FCL, peu de travaux sont aujourd'hui menés sur la complémentarité qui existe entre les liens conceptuels et l'approche classique de clustering qui consiste en l'extraction de communautés. Ainsi dans ce papier, nous nous intéressons à ces deux approches. Notre objectif est d'évaluer les relations potentiellement existantes entre les communautés et les FCL pour comprendre la façon dont les motifs obtenus par chacune des méthodes peuvent correspondre ou s'intersecter ainsi que la connaissance utile résultant de la prise en compte de ces deux types de connaissance. Nous proposons pour cela un ensemble de mesures originales, basées sur la notion d'homogénéité, visant à évaluer le niveau d'intersection des FCL et des communautés lorsqu'ils sont extraits d'un même jeu de données. Notre approche est appliquée à deux réseaux et démontre l'importance de considérer simultanément plusieurs types de connaissance et leur intersection.

1 Introduction

L'identification de groupes est une des tâches les plus courantes dans le domaine de la fouille de données. En effet, dans de nombreux systèmes naturels ou sociaux, les agents ont souvent tendance à s'organiser en différents groupes. L'identification de ces groupes est devenu un domaine de recherche actif pour comprendre les potentielles relations impliquant les interactions sous-jacentes et le rôle au sein du système.

Pourtant, si de nombreux travaux ont été menés pour concevoir de nouveaux algorithmes d'identification de groupes, ou des adaptations d'algorithmes existants, nous observons que ces travaux sont souvent menés séparément, sans prendre en compte la complémentarité avec les groupes extraits par d'autres méthodes. C'est par exemple le cas des méthodes qui s'intéressent à la recherche de groupes au sein de réseaux sociaux. Alors que les méthodes traditionnelles de clustering dédiées aux réseaux exploitent uniquement la structure du réseau pour extraire des **communautés** (Fortunato, 2009; Blondel et al., 2008), les approches récentes se sont, elles, intéressées à la fois à la structure du réseau et aux attributs des noeuds pour identifier des **liens**

conceptuels fréquents (Stattner et Collard, 2012b, 2013). Or, à notre connaissance, il n'existe aucune étude menée sur la correspondance ou les intersections de ces deux types de motifs.

Ainsi dans ce papier, nous nous intéressons à ces deux approches de clustering de réseaux sociaux : (i) la recherche de **communautés** qui vise à partitionner le réseau en différents groupes de noeuds fortement connectés et (ii) l'extraction de **liens conceptuels fréquents**, une nouvelle approche dont l'objectif est d'identifier les liens fréquents entre des groupes des noeuds partageant des caractéristiques communes.
Notre objectif est d'évaluer les relations potentiellement existantes entre les communautés et les FCL pour comprendre la façon dont les motifs obtenus par chacune des méthodes peuvent correspondre ou s'intersecter, ainsi que la connaissance utile résultant de la prise en compte de ces deux types de connaissance. Nous proposons pour cela un ensemble de mesures originales, basées sur la notion d'homogénéité, visant à évaluer le niveau d'intersection des FCL et des communautés lorsqu'ils sont extraits à partir d'un même jeu de données.
En appliquant nos mesures aux motifs extraits de deux jeux de données (un réseau de contacts de proximité et un réseau d'achats extrait du site de e-commerce Amazon.com), nous montrons, à travers diverses expériences, la connaissance utile qui peut-être obtenue en considérant simultanément ces deux types de connaissance et leurs intersections.

Le papier est organisé comme suit : La Section 2 présente les deux familles de méthodes que nous abordons dans ce papier. La Section 3 définit formellement les concepts de communautés et de liens conceptuels fréquents, et discute les questions soulevées par ces deux approches. La Section 4 décrit les mesures que nous proposons. La Section 5 est consacrée aux expériences ainsi qu'aux résultats obtenus. Enfin, la Section 6 conclut l'article et présente nos travaux futurs.

2 Travaux antérieurs

Cette section présente les deux méthodes de clustering de réseaux sociaux que nous abordons dans ce papier : (i) l'extraction de communautés et (ii) la recherche de liens conceptuels fréquents.

2.1 Extraction de communautés

La recherche de communautés dans les réseaux sociaux, également appelée *clustering basé sur les liens*, fait référence à une famille de méthodes qui effectuent un regroupement des noeuds en tenant uniquement compte de la structure topologique du réseau (Fortunato, 2009). Ces méthodes ont pour objectif de partitionner le réseau en plusieurs composantes (appelées groupes, clusters ou communautés) de sorte que les noeuds au sein de chaque composante possède une forte densité de connexions.

Ainsi, le principe de base consiste à identifier les groupes qui maximisent les liens intra-communautaires tout en minimisant les liens inter-communautaires. Dans cet objectif, la mesure de *modularité* introduite par M. E. J. Newman (Newman, 2006), est couramment utilisée pour évaluer la qualité du partitionnement.

Plusieurs classifications des algorithmes de recherche de communautés ont été proposés. La plus courante consiste à identifier deux catégories de méthodes : (i) les méthodes agrégatives, qui fusionnent les groupes de noeuds de manière itérative si leur score de similarité est

suffisamment élevé (Pons et Latapy, 2005; Blondel et al., 2008; Shen et al., 2009) et (ii) les méthodes séparatives, dans lesquelles les groupes de noeuds sont itérativement scindés en supprimant les liens entre les noeuds possédant un score de similarité faible (Girvan et Newman, 2002; Newman et Girvan, 2004; Rattigan et al., 2007).

2.2 Recherche de liens conceptuels fréquents

L'extraction de liens conceptuels fréquents est une nouvelle approche qui combine les informations disponibles sur la structure du réseau et les attributs des noeuds dans le but d'extraire les groupes de noeuds les plus connectés du réseau et au sein desquels les noeuds partagent des caractéristiques communes (Stattner et Collard, 2012b). Conformément à l'analyse de concepts formels (Ganter et al., 2005), "*conceptuel*" signifie ici que les liens ainsi identifiés sont des connexions entre des groupes de noeuds qui peuvent être assimilés à la notion de *concepts*, puisqu'ils partagent des propriétés communes.

L'extraction des liens conceptuels fréquents implique deux étapes clés : (i) la phase de clustering des noeuds, qui regroupe les noeuds sur la base d'attributs communs et (ii) la phase d'identification des liens conceptuels, qui évalue la fréquence des liens entre les clusters identifiés en phase (i).

Bien que cette approche soit relativement nouvelle, plusieurs algorithmes ont été proposés pour optimiser le processus d'extraction des liens conceptuels fréquents. Par exemple, dans (Stattner et Collard, 2012a) l'algorithme FCL-Min (Frequent Conceptual Link Mining) optimise l'extraction en réduisant progressivement l'espace de recherche. Une amélioration de cet algorithme est l'algorithme MFCL-Min (Maximal Frequent Conceptual Link Mining) proposé dans (Stattner et Collard, 2012b) qui s'intéresse, lui, uniquement aux liens conceptuels fréquents maximaux, c'est-à-dire ceux qui ne possèdent pas de sur-motifs qui soient également fréquents.

3 Concepts de communautés et de liens conceptuels

Dans cette section, nous présentons formellement les concepts de communautés et de liens conceptuels fréquents et nous discutons des questions qu'ils soulèvent.

Posons tout d'abord $G = (V, E)$ un réseau, dans lequel V est l'ensemble des noeuds et E l'ensemble des liens avec $E \subseteq V \times V$.

3.1 Communautés

Soit C l'ensemble des communautés extraites du réseau G. Nous supposons qu'il n'existe pas de chevauchement des communautés, c'est-à-dire qu'un noeud appartient à une seule communauté. Notons $F : V \rightarrow C$, la fonction qui renvoie, pour un noeud v donné, la communauté à laquelle il appartient.

Les communautés sont extraites de façon à maximiser la mesure de modularité Q définie comme suit :

$$Q = \frac{1}{2|E|} \sum_{ij} [W_{ij} - \frac{k_{v_i} k_{v_j}}{2|E|}] \, \delta(F(v_i), F(v_j))$$

avec W_{ij} qui représente le poids du lien entre les noeuds v_i et v_j, k_{v_i} qui correspond au degré du noeud v_i et la fonction δ qui est égale à 1 si $F(v_i) = F(v_j)$ et 0 autrement. Dans nos expériences, nous utilisons l'algorithme d'extraction proposé par Blondel et al. (Blondel et al., 2008), basé sur l'optimisation de la modularité.

3.2 Liens conceptuels fréquents

Nous définissons l'ensemble V comme une relation $R(A_1, ..., A_p)$ où chaque A_i est un attribut. Chaque noeud $v \in V$ est ainsi défini par un tuple $(a_1, ..., a_p)$ où $\forall k \in [1..p],\ v[A_k] = a_k$. Ainsi, a_k est la valeur de l'attribut A_k dans v et $p = |R|$ correspond au nombre d'attributs. Un item est une expression logique $A = x$, où A est un attribut et x une valeur. L'item vide est noté $\emptyset$. Un itemset est une conjonction d'items par exemple ($A_1 = x \quad$ et $\quad A_2 = y \quad$ et $A_3 = z$). Quand un itemset est une conjonction de k items non-vides, on parle de k-itemset.

Soient m et sm deux itemsets. Si $sm \subseteq m$, on dit que sm est un sous-itemset de m et que m est un sur-itemset de sm. Par exemple, $sm = xy$ est un sous-itemset de $m = xyz$.
Nous notons I_V l'ensemble des itemsets construits à partir de V.

Soit $G = (V, E)$ un *réseau dirigé uniparti*. Pour tout itemset $m \in I_V$, nous notons V_m l'ensemble des noeuds de V qui satisfont m et nous définissons :

- Le *m-ensemble de liens de G du coté gauche*, LE_m, comme l'ensemble des liens de E qui partent de noeuds vérifiant m, i.e. $LE_m = \{e \in E\ ;\ e = (a, b) \quad a \in V_m\}$
- Le *m-ensemble de liens de G du coté droit*, RE_m, comme l'ensemble des liens de E qui arrivent à des noeuds vérifiant m, i.e. $RE_m = \{e \in E\ ;\ e = (a, b) \quad b \in V_m\}$

Définition 1. Lien conceptuel. Soient deux itemsets m_1 et m_2 appartenant à I_V. Le *lien conceptuel* (m_1, m_2) de G représente l'ensemble des liens connectant les noeuds de V_{m_1} aux noeuds de V_{m_2}, i.e. $(m_1, m_2) = LE_m \cap RE_m = \{e \in E\ ;\ e = (a, b) \quad a \in V_{m_1}\ et\ b \in V_{m_2}\}$.

Définition 2. Support d'un lien conceptuel. Nous appelons *support* du lien conceptuel $l = (m_1, m_2)$, le pourcentage de liens dans E qui appartiennent à l, i.e. $supp(l) = \frac{|(m_1,m_2)|}{|E|}$.
Remarquons que pour tout itemset m et tout lien conceptuel l, si $l = (\emptyset, m)$ ou $l = (m, \emptyset)$, alors $supp(l) = 0$.

Définition 3. Lien conceptuel fréquent (FCL). Un lien conceptuel l est dit *fréquent* si son support est supérieur à un seuil minimum de support β, i.e. $supp(l) \geq \beta$.

Nous notons FL_V l'ensemble des liens conceptuels fréquents de G selon un seuil de support donné β.

$$FL_V = \bigcup_{m_1 \in I_V, m_2 \in I_V} \{\ l = (m_1, m_2) \in I_V^2\ ;\ supp(l) > \beta\ \} \tag{1}$$

Définition 4. Sous et Sur-lien conceptuel. Soient les itemsets sm_1, sm_2, sous-itemsets respectifs de m_1 et m_2 dans I_V. Le lien conceptuel (sm_1, sm_2) est appelé un *sous-lien conceptuel* de (m_1, m_2). De même, (m_1, m_2) est appelé un *sur-lien conceptuel* de (sm_1, sm_2). Nous notons $(sm_1, sm_2) \subseteq (m_1, m_2)$.

Définition 5. Lien conceptuel fréquent maximal (MFCL). Soit β un seuil minimum de support donné, nous appelons *lien conceptuel fréquent maximal*, tout lien conceptuel fréquent l, tel qu'il n'existe aucun sur-lien conceptuel l' de l qui soit également fréquent.
Plus formellement, l est maximal si et seulement si $\nexists l' \in FL_V$ tel que $l \subset l'$.

L'ensemble des MFCL fournit une *vue conceptuelle* du réseau, dans le sens où ces motifs apportent une connaissance sur les groupes de noeuds partageant des propriétés *internes* communes (*les concepts*) et étant les plus connectés du réseau. Plus précisément, la vue conceptuelle est une structure de graphe dans laquelle chaque noeud est associé à un itemset (on parle de *meta-noeud*) et chaque lien correspond à un MFCL.

Définition 7. Vue conceptuelle du réseau social. Soient $G = (V, E)$ un réseau social et β un seuil de support minimum. Nous définissons le graphe $G^*_\beta = (M, L)$, comme la vue conceptuelle de G obtenue avec le seuil β, dans lequel M est l'ensemble des itemsets union des itemsets des liens conceptuels, appelés *meta-noeuds* et L est l'ensemble des liens conceptuels fréquents maximaux.

La Figure 1 montre les deux types de motifs résultants de l'extraction des communautés et la recherche des liens conceptuels fréquents à partir d'un réseau de référence.

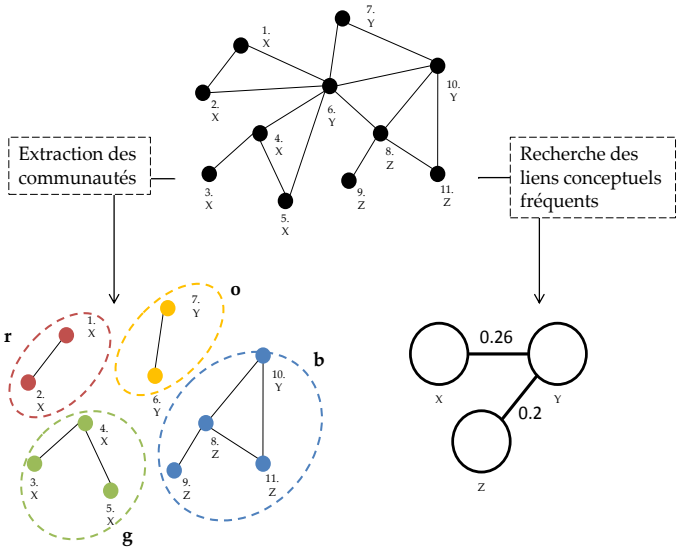


FIG. 1 – *Exemple de motifs extraits lors la recherche de communautés et l'extraction de liens conceptuels fréquents à partir d'un réseau de référence*

Les groupes extraits par les deux méthodes soulèvent plusieurs questions intéressantes sur les relations existantes entre les deux types de motifs :

1. Les noeuds au sein d'une même communautés partagent-ils des attributs communs ?
2. Les meta-noeuds sont-ils composés de noeuds appartenant à la même communauté ?
3. Les liens conceptuels fréquents connectent-ils des noeuds appartenant à la même communauté ou des noeuds de communautés différentes ?

Pour répondre à ces questions, nous présentons dans la section suivante un ensemble de mesure visant à évaluer les intersections entre les communautés et les liens conceptuels fréquents maximaux.

4 Mesures d'homogénéité

Cette section est consacrée aux mesures que nous proposons pour évaluer l'intersection des deux types de motifs : communautés et liens conceptuels fréquents.

Soit $G = (V, E)$ un réseau social dans lequel V est l'ensemble des noeuds et $E \subseteq V \times V$ l'ensemble des liens. Le cardinal des ensembles V et E, noté respectivement $|V|$ et $|E|$, donne le nombre de noeuds et de liens dans G.

Soit C l'ensemble des communautés identifiées dans le réseau en utilisant un algorithme de recherche classique (Fortunato, 2009). $|C|$ donne le nombre total de communautés extraites de G. La fonction F (cf. Section 3) est la fonction qui renvoie, pour un noeud $v \in V$ donné, la communauté $c = F(v)$ assignée à v. Nous notons V_c l'ensemble des noeuds de V qui appartiennent à la communauté c, i.e. $V_c = \{v \in V \ ; \ F(v) = c\}$.

Enfin, soit $G^*_\beta = (M, L)$ la vue conceptuelle obtenue par extraction des liens conceptuels fréquents maximaux à partir de G. L'ensemble M est l'ensemble des meta-noeuds et $L \subseteq M \times M$ est l'ensemble des liens conceptuels fréquents maximaux. L'extraction des MFCL est effectuée en utilisant l'algorithme MFCL-Min (Stattner et Collard, 2012b). Soit $m \in M$ un itemset, nous rappelons que V_m est l'ensemble des noeuds de V qui satisfont la propriété m.

Comme notre objectif est d'évaluer les éventuelles correspondances et intersections qui existent entre les communautés et les liens conceptuels fréquents, trois objets doivent être considérés : (i) les communautés, (ii) les meta-noeuds et (iii) les liens conceptuels. Nous présentons ainsi trois mesures liées à la notion d'homogénéité au sein de chacun de ces objets.

4.1 Homogénéité dans une communauté

Le taux H_c d'*homogénéité dans une communauté*, est une mesure qui indique, pour une communauté donnée $c \in C$, sa capacité à agréger des noeuds qui appartiennent au même meta-noeud, c'est à dire des noeuds partageant des propriétés communes. Cette mesure correspond au pourcentage de meta-noeuds non-représentés dans la communauté c.

$$H_c = 1 - \frac{|\{m \in M \ ; \ \exists v \in V \ avec \ F(v) = c \ et \ v \in V_m\}|}{|M|} \tag{2}$$

Si $H_c = 0$, tous les meta-noeuds sont présents dans la communauté c. Plus précisément, les noeuds de la communauté c satisfont toutes les propriétés impliquées dans les liens conceptuels. Inversement, une valeur élevée H_c indique que les noeuds dans la communauté c tendent à être très similaires du point de vue de leurs propriétés.
Par exemple, le taux d'homogénéité dans la communauté r est $H_r = 0.66$, alors qu'il est de $H_b = 0.33$ pour la communauté b (cf. Figure 2).

Pour prendre en compte le poids d'un meta-noeud au sein d'une communauté, nous introduisons $H_{c/m}$, le taux d'*homogénéité d'un meta-noeud m dans une communauté c*. Il correspond au pourcentage de noeuds vérifiant la propriété m dans la communauté c.

$$H_{c/m} = \frac{|\{v \in V \ ; \ F(v) = c \ \ et \ \ v \in V_m\}|}{|\{v \in V \ ; \ F(v) = c\}|} \tag{3}$$

Ainsi, si $H_{c/m} = 0$, les noeuds dans le meta-noeud m ne sont pas présents dans c. En d'autres termes, les noeuds vérifiant la propriété m n'appartiennent pas à la communauté c. Inversement, si $H_{c/m}$ tend vers 1, un fort pourcentage de noeuds vérifiant la propriété m appartient à

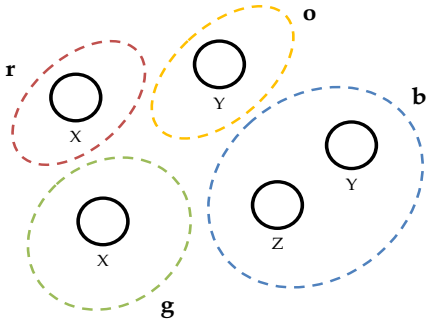


FIG. 2 – *Représentation des méta-noeuds dans les communautés à partir de l'exemple Figure 1*

c.
Dans l'exemple de la Figure 2, le taux d'homogénéité du meta-noeud X dans la communauté r est $H_{r/X} = 1$, alors que le taux d'homogénéité du meta-noeud Z dans la communauté b $H_{b/Z} = 0.75$.

4.2 Homogénéité dans un meta-noeud

Le taux H_m d'*homogénéité dans un meta-noeud* est une mesure qui indique, pour un meta-noeud donné $m \in M$, sa capacité à agréger des noeuds de la même communauté. Il correspond plus précisément au pourcentage de communautés non-présentes dans m.

$$H_m = 1 - \frac{|\{c \in C \; ; \; \exists v \in V_m \; avec \; F(v) = c\}|}{|C|} \tag{4}$$

Ainsi, si $H_m = 0$, toutes les communautés sont représentées dans le meta-noeud m. Autrement dit, toutes les communautés contiennent des noeuds qui vérifient la propriété m. Inversement, quand H_m tend vers 1, seul un faible pourcentage de communautés est présent dans m, i.e. le meta-noeud contient des noeuds de la même communauté.
La Figure 3 montre par exemple la représentation des communautés au sein des méta-noeuds, obtenue à partir du réseau de référence de la Figure 1. Le taux d'homogénéité dans le meta-noeud X est $H_X = 0.5$, alors que le taux d'homogénéité dans le meta-noeud Z est $H_Z = 0.75$.

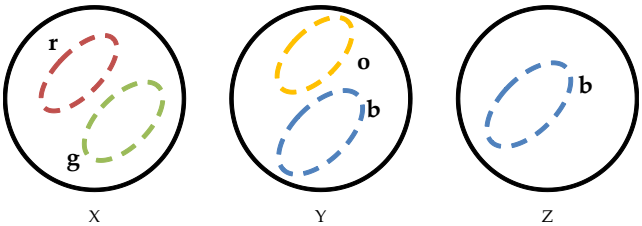


FIG. 3 – *Représentation des communautés dans les méta-noeuds à partir de l'exemple Figure 1*

Pour prendre en compte la taille des ensembles, nous introduisons $H_{m/c}$, qui est le taux d'*homogénéité de la communauté c dans le meta-noeud m*. Cette mesure indique le pourcen-

tage de noeuds dans c vérifiant la propriété m.

$$H_{m/c} = \frac{|\{v \in V_m \ ; \ F(v) = c\}|}{|V_m|} \tag{5}$$

Ainsi, si $H_{m/c} = 0$, les noeuds de la communauté c ne sont pas présents dans m. En d'autres termes, aucun noeud dans la communauté c ne vérifie la propriété m. Inversement, quand $H_{m/c}$ tend vers 1, un fort pourcentage de noeuds dans la communauté c vérifie la propriété m.
Par exemple, partant de l'exemple de la Figure 1, le taux d'homogénéité de la communauté r dans le meta-noeud X est $H_{X/r} = 0.4$ (cf. Figure 3). De la même façon, le taux d'homogénéité de la communauté b dans le meta-noeud Z est $H_{Z/b} = 1$.

4.3 Homogénéité dans un lien conceptuel fréquent

Le taux H_l d'*homogénéité dans un lien conceptuel* mesure, pour un lien conceptuel fréquent donné $l = (m_1, m_2)$, sa capacité à connecter des noeuds appartenant aux mêmes communautés. Il correspond au pourcentage de communautés similaires représentées des deux cotés du lien conceptuel fréquent.
$T_1 = \{c \in C \ ; \ \exists v \in V \text{ avec } F(v) = c \text{ et } v \in V_{m_1}\}$
$T_2 = \{c \in C \ ; \ \exists v \in V \text{ avec } F(v) = c \text{ et } v \in V_{m_2}\}$

$$HL_l = \frac{|(T_1 \cap T_2)|}{|(T_1 \cup T_2)|} \tag{6}$$

Ainsi, pour un lien conceptuel fréquent $l = (m_1, m_2)$, une valeur HL_l faible indique que les noeuds impliqués des deux cotés du lien appartiennent à des communautés différentes. Inversement, une valeur HL_l élevée indique que les mêmes communautés sont représentées dans m_1 et m_2.
Dans l'exemple de la Figure 4, le taux d'homogénéité dans le lien conceptuel (Z, Y) est $H_{(Z,Y)} = 0.5$. De la même façon, $H_{(X,Y)} = 0$.

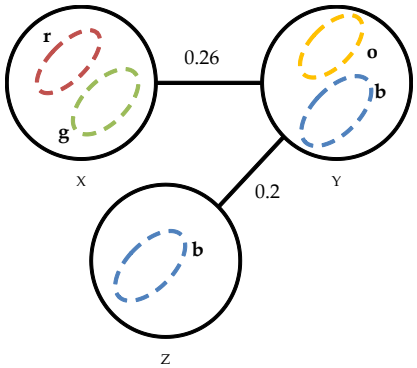


FIG. 4 – *Représentation des communautés dans les FCL à partir de l'exemple Figure 1*

Comme précédemment, nous introduisons $H_{l/c}$, le taux d'*homogénéité d'une communauté c dans un lien conceptuel $l = (m_1, m_2)$* ; c'est-à-dire la différence de représentation de c dans

les meta-noeuds m_1 et m_2.

$$H_{l/c} = 1 - \frac{|H_{m_1/c} - H_{m_2/c}|}{max(H_{m_1/c}, H_{m_2/c})} \quad (7)$$

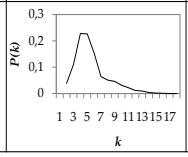
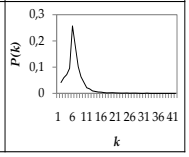
Ainsi, soit $l = (m_1, m_2)$ un lien conceptuel, $H_{l/c} = 1$ indique que le pourcentage de noeuds appartenant à c est identique dans m_1 et m_2.
Par exemple, le taux d'homogénéité de la communauté b dans le lien conceptuel (Z, Y) est $H_{(Z,Y)/b} = 1 - \frac{0.7}{1} = 0.33$ (cf. Figure 4).

5 Résultats expérimentaux

Cette section est consacrée aux résultats obtenus. La Section 5.1 décrit les jeux de données utilisés. La Section 5.2 s'intéresse aux motifs pertinents, identifiés à l'aide des mesures présentées dans la section précédente.

5.1 Jeux de données

Deux jeux de données ont été utilisés pour nos expériences. Le premier est un réseau de contacts de proximité obtenu avec l'outil *EpiSims* (Barrett et al., 2008), qui reproduit les déplacements d'individus dans la ville de Portland. Le second est un réseau d'achats extrait d'Amazon (Leskovec et al., 2007), dans lequel deux produits sont connectés quand ils sont achetés conjointement. Les principales propriétés de ces réseaux et des groupes extraits sont décrits sur la Figure 5.

	EpiSims	Amazon
#Noeuds	1043	5001
#Attributs	6	7
#Liens	2382	14981
Densité	0,0044	0,0012
Coeff. Clust	0,7091	0,4874
#Composante	1	1
Degré max	15	64
Degré Moyen	4,56759348	5,99120176
Distribution des degrés		

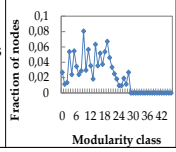
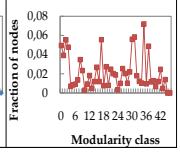
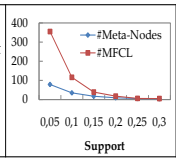
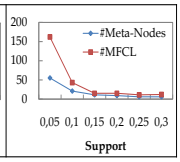
		EpiSims	Amazon
Communautés	Modularité	0,864	0,883
	#Communautés	29	45
	Distribution des tailles de communautés		
Liens conceptuels	#meta-nœuds et #mfcl selon différents seuils de support		

FIG. 5 – *Propriétés et groupes extraits des jeux de données utilisés*

Le réseau EpiSims contient 1043 noeuds, 2382 liens et 29 communautés. Chaque noeud est caractérisé par 6 attributs : (1) classe d'age, (2) sexe, (3) statut professionnel, (4) type de relation avec le chef de famille (5) classe de contact, (6) degré de sociabilité.

Le réseau Amazon contient 5001 noeuds, 14981 liens et 45 communautés. Chaque noeud est identifié par 7 attributs : (1) groupe (Livre, DVD, etc.), (2) nombre de produits similaires

achetés conjointement, (3) famille (entier), (4) catégorie principale (Literature & Fiction, Arts & Photography, Sport, etc.), (5) sous-catégorie, (6) nombres d'avis, (7) note ($\in [1..5]$).

Les communautés ont été extraites en utilisant l'algorithme de Blondel et al. (Blondel et al., 2008) et les liens conceptuels fréquents maximaux ont été identifiés en utilisant l'algorithme MFCL-Min avec $\beta = 0.1$. La Figure 6 montre (a) les communautés et (b) les liens conceptuels fréquents extraits du réseau Amazon. Par simplicité, nous notons les meta-noeuds (<att 1 >, <att 2 >, ..., <att n >) ou <att i > correspond à la valeur de l'attribut i. Le caractère '$*$' signifie que l'attribut peut avoir une valeur quelconque.

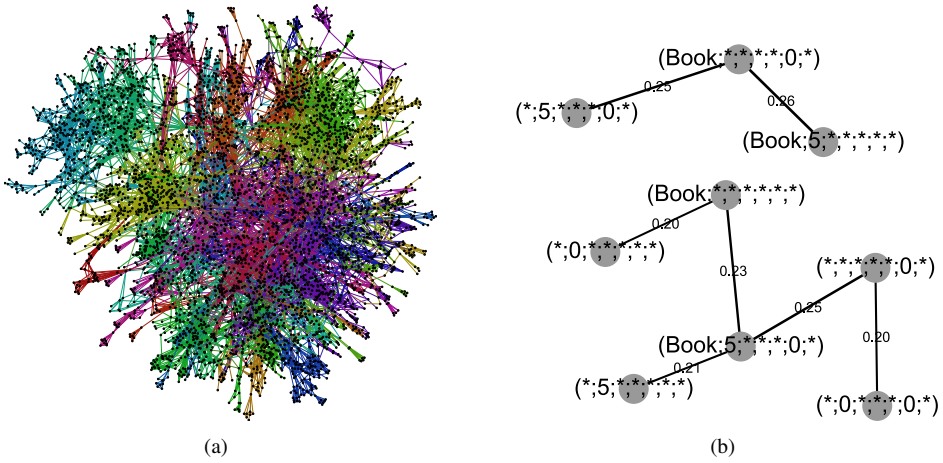


FIG. 6 – *Communautés (a) et liens conceptuels fréquents (b) extraits du réseau Amazon*

5.2 Résultats

Nous avons utilisé nos mesures pour identifier les situations qui maximisent les taux d'homogénéité. La Figure 7 montre des exemples de telles situations pour les deux réseaux.

Par exemple, la 1e ligne du tableau indique que, dans le réseau Amazon, l'ensemble des livres achetés seuls, i.e. sans produit similaire, $(Book; 0; *; *; *; *; *)$ intervient dans un lien conceptuel fréquent et est composé à 8% de noeuds appartenant à la communauté 35.

D'un autre point de vue, la 5e ligne indique que dans la communauté 5 du réseau Amazon, 68% des noeuds sont des livres qui n'ont pas reçu d'avis $(Book; *; *; *; *; 0; *)$.

La 7e ligne du tableau indique par exemple que dans le réseau Amazon, les livres sont fréquemment achetés conjointement avec des produits généraux, puisqu'il existe un lien conceptuel fréquent entre ces deux groupes $((Book; *; *; *; *; *; *), (*; *; *; *; General; *; *))$. De plus, nous observons que les noeuds dans la communauté 2 sont équitablement répartis dans ces deux groupes comme l'indique la valeur de $H_{c/l}$ très élevée.

Il est important de rappeler que ces résultats ont été obtenus avec notre environnement de test. De toute évidence, la quantité et la qualité des motifs (communautés et FCL) peuvent varier en fonction de la nature réseau et du seuil β de support des liens conceptuels utilisé.

	Meta-nœud ou MFCL	Communautés	Mesures	Valeur
Amazon	(Book;0;*;*;*;*;*)	35		0,0820
	(Music;5;*;*;*;*;*)	12	HMc/m	0,0803
	(Book;0;*;*;*;0;*)	8		0,0801
	(Book;*;*;*;*;*;*)	25		0,8400
	(Book;*;*;*;*;0;*)	5	HCm/c	0,6800
	(Book;5;*;*;*;*;*)	36		0,5833
	((Book;*;*;*;*;*;*),(*;*;*;General;*;*))	2		0,9954
	((Book;5;*;*;*;*;*),(Book;5;*;*;*;0;*))	21	HLc/l	0,9773
	((*;5;*;*;*;*;*),(*;0;*;*;*;0;*))	44		0,9211
EpiSims	(*;2;2;2;*;*)	9		0,1027
	(1;1;2;2;*;*)	19	HMc/m	0,1008
	(1;1;2;2;*;*)	20		0,1176
	(*;1;2;2;*;*)	24		0,5000
	(*;2;2;2;*;*)	1	HCm/c	0,4167
	(*;1;2;2;*;*)	20		0,3958
	((*;*;2;2;*;*),(*;1;2;2;*;*))	23		0,9343
	((*;1;2;2;*;*),(*;*;2;2;*;*))	14	HLc/l	0,9788
	((1;*;2;2;*;*),(*;*;2;*;*;*))	9		0,9350

FIG. 7 – *Exemples d'intersections intéressantes*

6 Conclusion

Dans ce papier, nous nous sommes intéressés à l'intersection des groupes obtenus par deux méthodes de clustering de réseaux sociaux : l'extraction de communautés et la recherche de liens conceptuels fréquents maximaux.

Nous avons pour cela proposé un ensemble de mesures, basées sur la notion d'homogénéité, qui permettent de quantifier les intersections entre les deux types de groupes. Nos résultats expérimentaux, menés sur deux jeux de données, ont d'une part démontré que de telles intersections pouvaient être trouvées, mais montrent plus généralement, comment la connaissance extraite d'un réseau est enrichie en considérant simultanément plusieurs types de motifs.

Références

Barrett, C. L., K. R. Bisset, S. G. Eubank, X. Feng, et M. V. Marathe (2008). Episimdemics : an efficient algorithm for simulating the spread of infectious disease over large realistic social networks. In *Proceedings of the 2008 ACM/IEEE conference on Supercomputing*.

Blondel, V., J. L. Guillaume, R. Lambiotte, et E. Lefebvre (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics : Theory and Experiment 2008*, P10008.

Fortunato, S. (2009). Community detection in graphs. *Physics Reports 486*, 75–174.

Ganter, B., G. Stumme, et R. Wille (2005). Formal concept analysis, foundations and applications. *Lecture Notes in Computer Science 3626*.

Girvan, M. et M. E. Newman (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences 99*(12), 7821–7826.

Leskovec, J., L. A. Adamic, et B. A. Huberman (2007). The dynamics of viral marketing. *ACM Trans. Web 1*.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences 103*(23), 8577–8582.

Newman, M. E. et M. Girvan (2004). Finding and evaluating community structure in networks. *Physical review E 69*(2), 026113.

Pons, P. et M. Latapy (2005). Computing communities in large networks using random walks. In *Computer and Information Sciences-ISCIS 2005*, pp. 284–293. Springer.

Rattigan, M. J., M. Maier, et D. Jensen (2007). Graph clustering with network structure indices. In *Proceedings of the 24th international conference on Machine learning*, pp. 783–790. ACM.

Shen, H., X. Cheng, K. Cai, et M.-B. Hu (2009). Detect overlapping and hierarchical community structure in networks. *Physica A : Statistical Mechanics and its Applications 388*(8), 1706–1712.

Stattner, E. et M. Collard (2012a). Flmin : An approach for mining frequent links in social networks. *International Conference on Networked Digital Technologies*.

Stattner, E. et M. Collard (2012b). Social-based conceptual links : Conceptual analysis applied to social networks. *International Conference on Advances in Social Networks Analysis and Mining*.

Stattner, E. et M. Collard (2013). Towards a hybrid algorithm for extracting maximal frequent conceptual links in social networks. *IEEE International Conference on Research Challenges in Information Science*, 1–8.

Summary

The search for frequent conceptual links (FCL) is a new clustering approach of networks, that exploit both structure and content of nodes. While recent studies have focused on the optimization of the FCL extraction process, very few works are now conducted on the complementarity between the conceptual links and classical clustering approach that consists on extracting communities. In this work, our objective is to evaluate the possible relations existing between these two kinds of patterns, in order to understand how the patterns extracted by each family of methods may match or intersect. For this purpose we propose a set of original measures based on the notion of homogeneity into a pattern. We apply our measures on two datasets and demonstrate the interest to consider simultaneously several kinds of knowledge and their potential intersections.