Détection de structures de communauté dans les hyper-réseaux d'interactions

Cécile Bothorel, Mohamed Bouklit

Détection de structures de communauté dans les hyper-réseaux d'interactions

Cecile Bothorel[†] et Mohamed Bouklit[‡]

France Telecom R&D, 2 avenue Pierre Marzin 22307 Lannion

Dans cet article, nous nous intéressons au problème de la détection de communautés dans les hyper-réseaux d'interactions (*complex hyper-networks*). Plus précisément, nous proposons une généralisation du fameux algorithme de détection de communautés de Girvan et Newman aux hyper-réseaux d'interactions. Les résultats expérimentaux montrent que notre algorithme donne des résultats encourageants sur des hypergraphes de cooccurence de tags obtenus à partir du collectif en ligne de partage de photos Flickr.

Keywords: algorithmique des (hyper-)réseaux d'interactions, hypergraphes, détection de communautés, centralité d'intermédiarité, Flickr

1 Introduction

L'utilisation des réseaux d'interactions ne fournit pas toujours une description suffisamment fine de la structure des systèmes complexes étudiés. Ainsi, la représentation sous forme de graphe d'un réseau de collaboration scientifique nous permet par exemple uniquement de savoir si deux chercheurs ont collaboré. En revanche, elle ne permet pas de savoir si plus de trois chercheurs (reliés dans le réseau) ont co-écrit ensemble un article. Estrada et Rodríguez-Velásquez introduisent les hyper-réseaux d'interactions (*complex hyper-networks*) comme une généralisation *naturelle* des réseaux d'interactions [ERV06]. Les hyper-réseaux d'interactions sont des hypergraphes rencontrés en pratique qui permettent de modéliser la structure de certains systèmes complexes de manière beaucoup plus précise que les réseaux d'interactions. Dans un graphe, une arête relie seulement un couple de sommets tandis que les arêtes d'un hypergraphe connues sous le nom d'*hyperarêtes* peuvent relier des groupes de plusieurs sommets. Ainsi, on peut représenter notre réseau de collaboration scientifique par un hypergraphe dont les noeuds sont les auteurs et dont les hyperarêtes représentent les groupes d'auteurs ayant co-écrit des articles. Estrada et Rodríguez-Velásquez généralisent dans le même article le coefficient de clustering introduit par Watts et Strogatz [WS98] aux hyper-réseaux d'interactions [ERV06]. Nous nous intéressons dans cet article au problème de la détection de structures de communauté dans les hyper-réseaux d'interactions. A notre connaissance, il n'existe pas encore d'algorithme de detection de communautés dans le contexte des hyper-réseaux d'interactions.

Définitions. Un hypergraphe est représenté par un couple $H = (V,E)$ où $V = \{v_1, \cdots, v_n\}$ est l'ensemble des noeuds et $E = \{e_1, \cdots, e_m\}$ est l'ensemble des hyperarêtes telles que $e_i \neq \emptyset$ et $\bigcup_{i=1}^{i=m} e_i = V$ [Ber85]. Une hyperarête e_i est un sous-ensemble de V. La *taille* d'une hyperarête est le nombre de sommets qu'elle contient. Dans la suite de l'article, H désignera un hypergraphe non-orienté, non-pondéré et connexe composé de $n = |V|$ sommets et $m = |E|$ hyperarêtes. En raison de leur faible densité en pratique, nous avons choisi de représenter les hypergraphes sous la forme d'un graphe biparti reliant les sommets aux hyperarêtes (auxquelles ils appartiennent). La complexité de cette représentation coûte $O(m+n+k)$ espace (k désignant le nombre d'arêtes de ce graphe biparti). Comme nous considérons uniquement des hypergraphes connexes ($k \geq m+n-1$), on en déduit une complexité spatiale en $O(k)$. Après avoir détaillé notre algorithme de

[†] cecile.bothorel@orange-ftgroup.com
[‡] mohamed.bouklit@orange-ftgroup.com

detection de communautés dans les hyper-réseaux d'interactions, nous présentons les résultats sur des hypergraphes de cooccurence de tags obtenus à partir du collectif en ligne de partage de photos Flickr[§].

2 Description de l'algorithme

Notre algorithme est une généralisation du fameux algorithme de Girvan et Newman [NG04] aux hypergraphes. Leur algorithme retire itérativement les arêtes de plus forte centralité d'intermédiarité. Cette centralité est définie pour une arête comme le nombre de plus courts chemins passant par cette arête. Il existe en effet peu d'arêtes reliant les différentes communautés et les plus courts chemins entre deux sommets de deux communautés différentes ont de grandes chances de passer par ces arêtes. En supprimant ces arêtes, les composantes connexes du graphe résultant sont assimilées à des communautés. A chaque étape, la qualité de la partition du graphe est calculée. L'algorithme retourne la partition du graphe possédant la meilleure qualité.

2.1 Principe de l'algorithme

Nous allons partir de la partition de l'hypergraphe contenant une seule communauté (correspondant à l'hypergraphe entier) et scinder successivement les communautés jusqu'à obtenir n communautés contenant chacune un seul sommet de la façon suivante :

- Calculer la centralité d'intermédiarité de chaque sommet et de chaque hyperarête décrite en section 2.2 (complexité : $O(nk)$)
- Retirer l'hyperarête de centralité maximale (complexité : $O(k)$)
- Calculer une partition de l'hypergraphe en communautés[¶] (complexité : $O(k)$)
- Calculer et mémoriser un paramètre de qualité Q présenté en section 2.3 (complexité : $O(k\log k)$[‖])

Nous obtenons ainsi, après m itérations, une suite de m partitions des sommets en communautés $P_0, ...P_{m-1}$ parmi lesquelles il va falloir choisir la meilleure (maximisant Q). La complexité d'une itération étant en $O(nk+k\log k)$ temps, il en découle donc que la complexité totale de l'algorithme est en $O(m(nk+k\log k))$ temps dans le pire des cas.

2.2 Calcul de la centralité d'intermédiarité

La centralité d'intermédiarité d'un sommet ou d'une hyperarête u (que l'on notera $B(u)$) est le nombre de plus courts hyperchemins passant par u. Adoptant une approche similaire à Girvan et Newman, nous allons calculer pour chaque sommet et chaque hyperarête de l'hypergraphe sa centralité d'intermédiarité dite locale à v. La centralité d'intermédiarité locale à v d'un sommet ou d'une hyperarête u (que l'on notera $B_v(u)$) est le nombre de plus courts hyperchemins partant de v passant par u. On en déduit que : $B(u) = \sum_{v \in V} B_v(u)$. Cela revient en fait à faire circuler un flot égal à 1 le long des hyperchemins de u à v pour tout sommet u. Les valeurs de flots obtenues pour chaque sommet et chaque hyperarête correspondent aux centralités locales recherchées.

L'algorithme effectue donc n itérations correspondant au calcul de la centralité d'intermédiarité locale à v de chaque sommet et de chaque hyperarête de l'hypergraphe de la façon suivante :

1. on calcule dans un premier temps, en $O(k)$ temps, l'ensemble des plus courts hyperchemins partant de v à l'aide d'un parcours en largeur modifié de l'hypergraphe H. La figure 1.b montre l'ensemble des hyperchemins partant de a dans l'hypergraphe représenté à la figure 1.a. Plus précisément, on associe à chaque sommet et à chaque hyperarête de l'hypergraphe l'ensemble de ses prédécesseurs sur ces hyperchemins. On peut constater par exemple que l'hyperarête D a pour prédécesseurs c et d.
2. on calcule dans un second temps, en $O(k)$ temps, les centralités locales de chaque hyperarête et de chaque sommet qui sont respectivement initialisées à 0 et à 1. Plus précisément, on traite l'ensemble

[§] `http://www.flickr.com`

[¶] Les composantes connexes de l'hypergraphe étant identifiées à des communautés, il suffit d'effectuer un calcul de composantes connexes à l'aide d'un parcours en largeur.

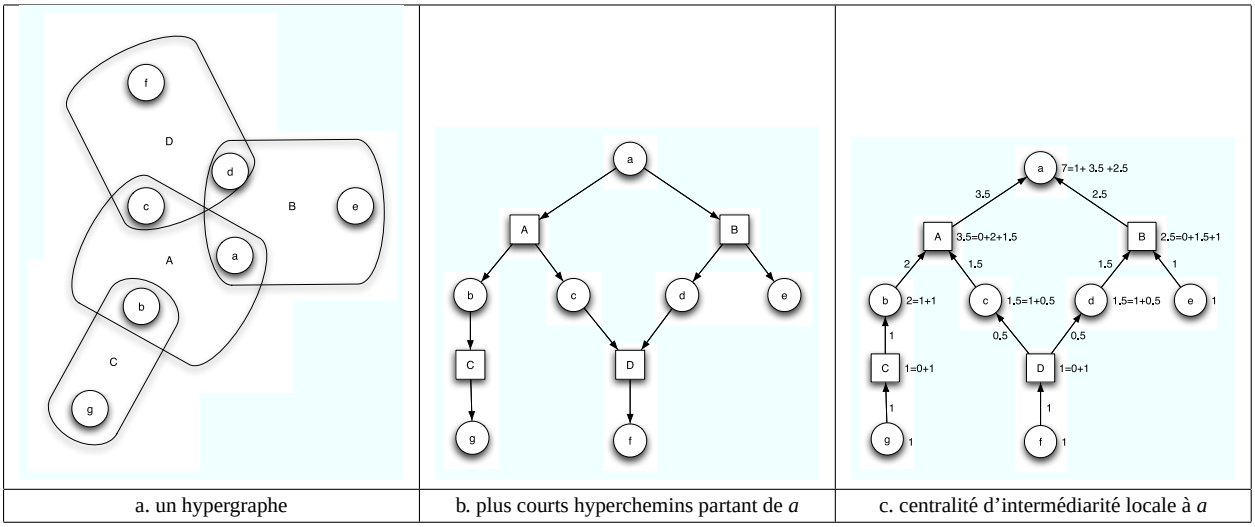
[‖] Le calcul de cette quantité nécessite un tri préalable des arêtes du graphe biparti (codant l'hypergraphe) que nous ne pouvons pas malheureusement détailler faute de place.

des sommets et des hyperarêtes u dans l'ordre inverse du parcours en largeur ($f\ g\ D\ C\ e\ d\ c\ b\ B\ A\ a$ dans notre cas représenté à la figure 1.c) :

(a) la centralité locale $B_v(u)$ est tout d'abord ajoutée à la centralité d'intermédiarité globale $B(u)$: $B(u) \leftarrow B(u) + B_v(u)$. Lorsqu'on traite par exemple l'hyperarête D, on ajoute sa centralité locale $B_a(D)$ (qui n'augmentera plus dans la suite du parcours) à sa centralité globale $B(D)$.

(b) $B_v(u)$ est ensuite distribuée de manière équitable entre ses predecesseurs w : $B_v(w) \leftarrow B_v(w) + \frac{B_v(u)}{n_u}$ où n_u désigne le nombre de prédécesseurs de u. L'hyperarête D distribue par exemple sa centralité locale $B_a(D) = 1$ équitablement entre ses prédécesseurs c et d qui recevront donc chacun 0.5.

Les figures 1.b et 1.c illustrent donc une itération de l'algorithme. Après n itérations, nous obtenons comme résultat de l'algorithme les centralités d'intermédiarité globales pour tous les sommets et les hyperarêtes de l'hypergraphe H. La complexité de cet algorithme est donc en $O(nk)$.



a. un hypergraphe | b. plus courts hyperchemins partant de a | c. centralité d'intermédiarité locale à a

FIG. 1: Calcul de la centralité d'intermédiarité

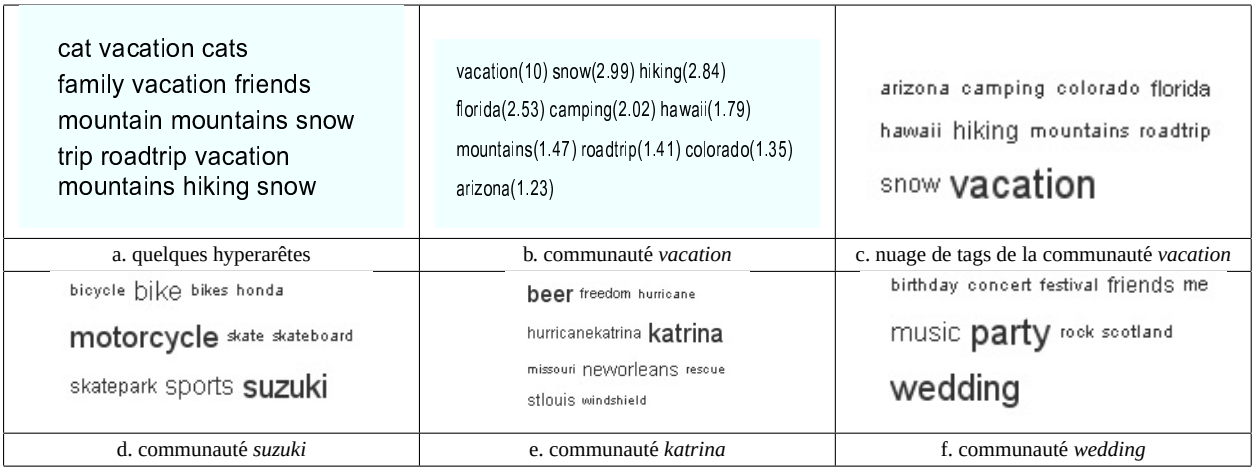
2.3 Evaluation de la qualité d'une partition d'un hypergraphe en communautés

Afin d'évaluer la qualité d'une partition P d'un hypergraphe H en communautés, nous proposons l'*hypermodularité* $Q(P) = \sum_{C \in P} \left[e(C) - \left(\sum_{t=2}^{t=n} a_t(C)^t \right) \right]$ où $e(C)$ est la fraction des hyperarêtes internes à la communauté C et $a_t(C)$ est la fraction des hyperarêtes de taille t ayant au moins une extrémité dans la communauté C. Cette mesure de qualité est une généralisation de la modularité introduite par Girvan et Newman dans leur algorithme de détection de communautés. Une hyperarête est dite *interne* à une communauté C si toutes ses extrémités sont dans la communauté C. Une hyperarête de taille t est dite *liée* à une communauté C si au moins l'une de ses t extrémités appartient à la communauté C. Ainsi, les hyperarêtes de taille 4 ayant 2 extrémités dans C comptent pour moitié ($\frac{2}{4}$) par rapport aux hyperarêtes de taille 4 ayant toutes leurs extrémités dans C.

L'objectif est d'avoir des communautés de forte densité interne mesurée par $e(C)$. Les grosses communautés ont cependant mécaniquement une proportion d'arêtes internes plus élevée : si C est un ensemble aléatoire de sommets et si les hyperarêtes sont aussi aléatoires alors la proportion d'hyperarêtes de taille t internes attendue est $a_t(C)^t$. Comme la modularité, l'hypermodularité compare la proportion effective d'hyperarêtes internes aux communautés à la proportion attendue selon ce schéma. Une communauté est d'autant plus pertinente que sa proportion d'hyperarêtes internes sera supérieure à sa proportion attendue d'hyperarêtes. Nous retenons donc comme résultat de notre algorithme la partition de l'hypergraphe H possédant la meilleure hypermodularité.

3 Résultats

Ce travail a été mené en concertation avec le projet pluridisciplinaire AUTOGRAPH** qui s'intéresse à la visualisation et à l'auto-organisation des collectifs en ligne à base coopérative sur Internet comme Flickr [PCB+08]. Une des questions fondamentales posées par les sociologues dans ce projet était de savoir si un schéma de catégorisation des tags émerge au niveau global dans Flickr même si les utilisateurs ne partagent pas un vocabulaire centralisé pour annoter leurs photos. C'est la raison pour laquelle nous avons souhaité fragmenter des hypergraphes de cooccurrence de tags extrait de Flickr à l'aide de notre algorithme. Les hypernoeuds représentent les tags. Les hyperarêtes correspondent aux ensembles de tags qui co-apparaissent fréquemment dans la description des photos (figure 2.a). Il s'agit plus précisément de vérifier s'il apparaît un consensus ou des conflits dans l'utilisation des tags parmi les communautés.



a. quelques hyperarêtes | b. communauté *vacation* | c. nuage de tags de la communauté *vacation*

d. communauté *suzuki* | e. communauté *katrina* | f. communauté *wedding*

FIG. 2: Exemple de nuages de tags

Les tags affichés sont les plus représentatifs des communautés selon le critère de centralité (figure 2.b). Pour une représentation sous forme de nuage de tags (figures 2.c à 2.e), la police de chaque tag est proportionnelle à sa centralité d'intermédiarité dans l'hypergraphe initial. Ces premiers résultats prometteurs constituent un premier pas qui semblent confirmer l'hypothèse des sociologues d'un consensus dans l'utilisation des tags.

Remerciements. Nous souhaitons remercier l'ensemble des participants du projet RNRT Autograph et tout particulièrement Dominique Cardon, Pascal Pons et Christophe Prieur pour nous avoir gracieusement fourni dans ce cadre leurs données volumineuses de qualité.

Références

[Ber85] Claude Berge. *Graphs and Hypergraphs*. Elsevier Science Ltd, 1985.

[ERV06] E. Estrada and J. A. Rodríguez-Velázquez. Subgraph centrality and clustering in complex hyper-networks. *Physica A : Statistical Mechanics and its Applications*, 364 :581–594, May 2006.

[NG04] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review*, E 69(026113), 2004.

[PCB+08] Christophe Prieur, Dominique Cardon, Jean-Samuel Beuscart, Nicolas Pissard, and Pascal Pons. The stength of weak cooperation : A case study on flickr. 2008. Communication personnelle avec C. Prieur.

[WS98] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393 :440–442, June 1998.

** Ce projet rassemble entre autres des chercheurs en informatique de l'Université Paris VII et de France Telecom, des chercheurs en sciences sociales de l'EHESS mais aussi des acteurs impliqués dans des collectifs en ligne comme Wikipédia.