Evolution temporelle de communautés représentatives : mesures et visualisation

Haolin Ren*, Marie-Luce Viaud*, Guy Mélançon**

*INA, 4 avenue de l'Europe Bry/Marne
{hren,mlviaud}@ina.fr
**CNRS UMR 5800 LaBRI, Université de Bordeaux
Guy.Melancon@u-bordeaux.fr

Résumé. La problématique de ce papier est d'identifier dans un graphe dynamique les communautés les plus représentatives sur une période donnée, de mesurer leur stabilité, et d'en visualiser les évolutions majeures. Notre cas d'usage concerne l'étude de la visibilité médiatique des communautés et des individus grâce aux données relatives aux émissions télévisuelles et radiophoniques entre 2011 et 2015. A partir d'une détection de communautés sur l'intégralité de la période, nous proposons des mesures de stabilité et d'activité des communautés et proposons une visualisation de leur évolution temporelle.

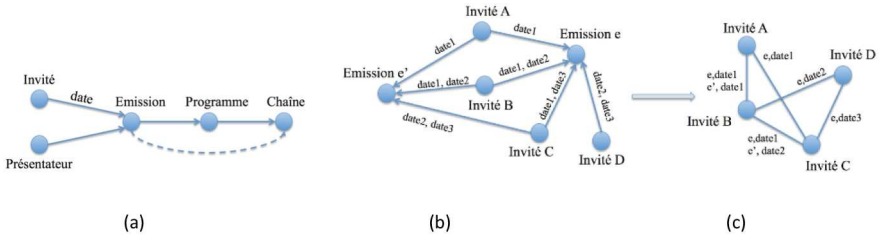
1 Introduction

La présence médiatique d'une personnalité ou d'un groupe de personnalités dans les média "main stream" participe de la vie sociale et politique d'un pays. Dans le cadre de l'analyse des médias entreprise par l'INA avec le projet OTMedia [1], nous nous intéressons à la présence médiatique des personnalités à la radio et à la télévision. Nous voulons détecter et visualiser les communautés disposant d'une grande visibilité ainsi que les principaux *événements* qui les affectent. Ces *événements* incluent : l'apparition ou la disparition d'une communauté ou les éléments qui la quittent ou qui la rejoignent. La visibilité médiatique est par essence réactive car elle dépend des événements médiatiques (élections, sorties de livre, faits d'actualités, ...). Mais l'hypothèse des experts des médias est qu'il existe néanmoins une stabilité produite par "la machine média et ses contraintes". C'est à l'observation de ce phénomène qu'il s'agit d'apporter un appui, afin de mieux l'étudier.

Les données disponibles pour cette étude sont des données structurées sur 4 ans (2011-2015), et rassemble 490000 émissions. Nous considérerons dans cet article le graphe formé des liens entre *invités* et *émissions* (dérivé à partir du graphe d'origine, voir figure). Deux invités sont liés l'un à l'autre s'ils ont participé à au moins une même émission, les liens entre invités emportant avec eux les attributs de *date* des liens d'origine afin de pouvoir tenir compte du temps et de la fréquence de co-apparition dans les émissions.

Les travaux relatifs aux graphes temporels se scindent en 3 méthodologies principales. Les méthodes qui cherchent à coupler des communautés obtenues sur des graphes statiques représentant des temps précis ou des tranches de temps (D. Greene, 2010) (M. Oliveira, 2010),

1. Voir `http://www.otmedia.fr/`



(a) (b) (c)

les méthodes qui utilisent l'information temporelle pour effectuer une meilleure détection des communautés (disparition, apparition de nœuds et d'arêtes) (Lin et al., 2008) (Tantipathananandh et Berger-Wolf, 2011) et les algorithmes itératifs, adaptés aux données qui évoluent en temps réel (Falkowski, 2009).

2 Détection de communautés : quelques définitions

2.1 Graphe des participants : définitions

Soit $G = (V, E)$, où V est l'ensemble des participants aux émissions radiophoniques et télévisuels ; une émission se déroule à une date donnée et connue. Désignons par S l'ensemble des émissions. On définit une propriété $\sigma : V \to \mathcal{P}(S)$ qui associe à $v \in V$ l'ensemble $\sigma(v) \subset S$ émissions accueillant l'invité $v \in V$. L'ensemble des arêtes du graphe G est défini en posant :

$$e = (v, v') \in E \iff \sigma(v) \cap \sigma(v') \neq \emptyset$$

On étend alors naturellement la propriété σ sur les arêtes en posant $\sigma(e) = \sigma(v) \cap \sigma(v')$. On associe aussi à une arête $e \in E$ sa *force* $|\sigma(e)|$ (voir prochaine section).

Donné deux dates $t_i < t_j$, on définit le graphe $G_{t_i,t_j}(V', E')$ induit de G à partir des invités $v \in V$ ayant participé à une émission se déroulant pendant la période $[t_i, t_j]$.

2.2 Squelette et communautés du graphe G des participants

La densité du graphe G est telle qu'il est difficile d'en produire un dessin lisible Pour le manipuler, en isoler des communautés, et le visualiser, il est d'abord crucial d'en simplifier la structure. Cette simplification, afin d'être interprétable, doit toutefois préserver un "squelette" du graphe G ; c'est précisément ce que fait l'approche de (Nick et al., 2013). En deux mots, cette approche propose de filtrer les arêtes en associant deux critères, un critère de force global et un critère de voisinage local. La force d'une arête $e \in E$ est $|\sigma(e)|$: deux invités sont liés plus fortement s'ils ont co-participé à un plus grands nombres d'émissions. Deux indivdus, liés par une arête "forte" restent liés dans le graphe filtré dès qu'ils partagent une partie suffisamment grande de leur voisinage "fort". Le filtrage fait intervenir 2 paramètres m et n, contraignant plus ou moins fortement le voisinage commun. Fixés empiriquement, m et n jouent sur l'amplitude du filtrage sans bouleverser les communautés comme le montre la figure 1.

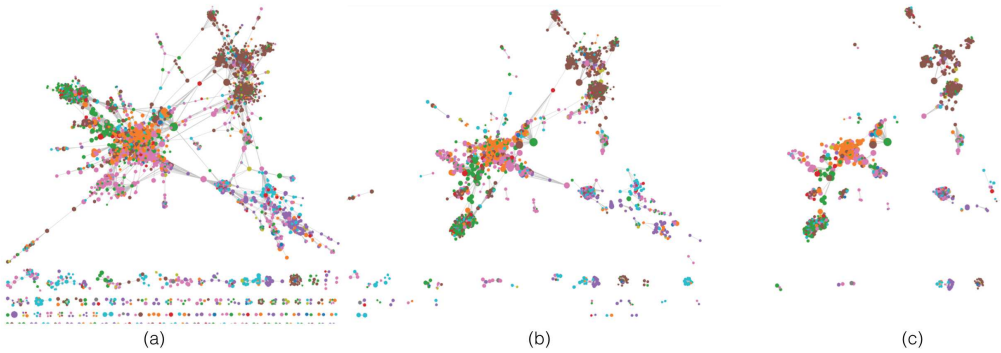


FIG. 1 – *Filtrage effectué par l'algorithme de (Nick et al., 2013) avec des valeurs différentes de m et n (a) $m = 6, n = 24$, 3815 noeuds, 61753 arêtes (b) $m = 8, n = 18$, 2580 noeuds, 42030 arêtes (c) $m = 9, n = 16$, 1775 noeuds, 22385 arêtes. Les communautés sont plus ou moins importantes mais stables*

Le squelette ainsi calculé G^s capture les zones du graphe les plus denses en terme d'émissions mutuellement partagées, et arrive ainsi à exhiber la structure petit monde du graphe. Ensuite, les communautés sont détectées sur G^s à l'aide de l'algorithme de Louvain (Blondel et al., 2008). Par définition, les communautés détectées par l'algorithme de Louvain forment une partition $\mathbf{C} = \{C_1, C_2, \ldots, C_k\}$

3 Mesures de stabilité et d'activité d'une communauté

Les émissions représentent un signal discret contrairement à une relation entre deux personnes (cas des graphes sociaux). Aussi, la continuité d'une relation ne peut-elle s'exprimer que sur un intervalle de temps. C'est pourquoi nous élaborons deux mesures qui nous permettent de définir les notions d'activité et de stabilité d'une communauté sur un intervalle de temps.

3.1 Mesure d'activité d'une communauté

Nous allons définir l'*indice d'activité* d'une communauté C_p, noté $I_{ac}(C_p)$ ou plus simplement I_{ac}, variant dans $[-1, -1]$. L'ensemble C_p étant fixé, nous n'allons considérer sur une intervalle de temps $[t_i, t_j]$ que les arêtes faisant apparaître une émission se déroulant pendant la période $[t_i, t_j]$. Cet indice mesure combien les membres d'une communauté sont prioritairement liés les uns aux autres sur l'intervalle de temps $[t_i, t_j]$ ou, au contraire, combien l'activité de la communauté est tournée vers l'extérieur (impliquant des invités qui ne sont pas dans C_p).

Sur l'intervalle $[t_i, t_j]$, nous considérons le sous graphe de G_{t_i,t_j} *engendré* par C_p et noté $G_{p,[t_i,t_j]} = (V_p, E_p)$. Pour rappel, le graphe $G_{p,[t_i,t_j]}$ est le plus petit graphe contenant toutes les arêtes incidentes à au moins un sommet de C_p ; par conséquent on a $C_p \subset V_p$.

Soit E_{int} l'ensemble des arêtes *internes* (à C_p), $e = (v, v') \in E_p$, qui sont telles que v, $v' \in C_p$ et E_{ext} ; de manière similaire, on définit l'ensemble des arêtes *externes*, qui sont telles que $e \notin E_{int}$ (c'est-à-dire les arêtes dont l'une des extrémités est dans $V_p \setminus C_p$).

Soit $F = \bigcup_{e \in E_p} \sigma(e)$ l'ensemble des émissions f apparaissant sur les arêtes de $G_{p,[t_i,t_j]}$. Désignons par F_{int} l'ensemble des émissions *internes*, en posant $F_{int} = \bigcup_{e \in E_{int}} \sigma(e)$; on définit aussi l'ensemble des émissions *externes* $F_{ext} = F - F_{int}$. L'indice I_{ac} est définie par :

$$I_{ac} = \frac{\sum_{e \in E_{int}} |\sigma(e)| - \sum_{e \in E_{ext}} \sum_{f \in F_{ext}} 1}{\sum_{e \in E} |\sigma(e)|}$$

Nous noterons que les émissions portées à la fois par des arêtes internes et externes, correspondant à des activités regroupant des invités de la communauté et de l'extérieur n'interviennent qu'au dénominateur pour faire baisser la valeur de l'indice. La formulation adoptée ici admet diverses variantes ; après expérimentation, la formulation arrêtée nous est apparue être la plus discriminante.

3.2 Mesure de stabilité d'une communauté

Le degré pondéré d'un individu $u \in C_p$ est défini par $d_u = \sum_{v \in N_{[t_i,t_j]}(u)} |\sigma(u, v)|$, où $N_{[t_i,t_j]}(u)$ désigne le voisinage de u dans $G_{[t_i,t_j]}$. On définit aussi $d_{u,C_k} = \sum_{v \in N_{[t_i,t_j],C_k}(u)} |\sigma(u, v)|$ où $N_{[t_i,t_j],C_k}(u) = N_{[t_i,t_j]}(u) \cap C_k$ désigne l'ensemble des voisins de u qui sont dans C_k. On peut définir par extension le degré pondéré d'une communauté $C_{p,[t_i,t_j]}$ en posant $d_{C_p} = \sum_{u \in C_p} d_u$, et $d_{C_p,C_k} = \sum_{u \in C_p} d_{u,C_k}$.

Le coefficient de participation du sommet u, par rapport aux communautés $(C_1, C_2, \ldots)$ est défini par (Guimerà et al., 2005) :

$$p(u) = 1 - \sum_k \left(\frac{d_{u,C_k}}{d_u}\right)^2 \quad (1)$$

Dans l'équation (1), la somme inclut la communauté C_p auquel u appartient. Lorsque u n'est connecté qu'à des sommets de sa communauté C_p, on a $p(u) = 0$. Le coefficient $p(u)$ croît et approche 1 avec la diversité des connections de u. On étend le coefficient de participation aux communautés C_p en posant :

$$p(C_p) = 1 - \sum_k \left(\frac{d_{C_p,C_k}}{d_{C_p}}\right)^2 \quad (2)$$

Le coefficient de participation d'une communauté témoigne de l'*homogénéité* des interactions entre les invités qui la forment avec les invités des autres communautés, sur l'intervalle de temps $[t_i, t_j]$ considéré. Une valeur de 0 témoigne d'une activité de la communauté totalement refermée sur une communauté ; une valeur approchant 1 témoigne, d'une certaine manière, d'un groupe d'invité s'étant retrouvé en interaction de manière quasiment fortuite sur la période.

4 Animation, Exploration et Analyse des données

Les mesures d'activité et les coefficients de participation sont calculées à la volée pour des graphes temporels $G_{[t_i,t_j]}$, selon une fenêtre de temps $[t_i, t_j]$ glissante.

L'interactivité permet à l'utilisateur de sélectionner une communauté en cliquant sur l'un de ses membres ou au lasso ; la sélection est matérialisée par une enveloppe convexe permettant de visualiser le contour de la communauté et donnant un retour sur ses indices de stabilité et d'activité communautaire.

Plus la valeur d'activité est proche de 1 plus la teinte associée dont est intense. elle devient grise lorsque la valeur est négative.

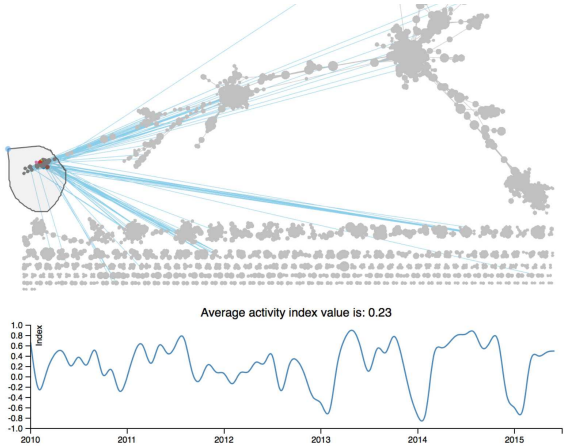


FIG. 2 – *Si l'utilisateur sélectionne une communauté,la courbe d'activité communautaire apparaît. Cette communauté partage bien des activités sur les années 2011, 2013 et 2014 mais moins sur 2012. Les arêtes en bleu sont les arêtes de G filtrées par l'algorithme* **??** *et réintégrées pour observer avec quelles communautés ce groupe interagit*

Le scénario d'exploration de l'expert vise à déterminer, pour un individu u, quelle communauté C_i lui donne la visibilité la plus forte et sur quel intervalle de temps $[t_i, t_j]$. Une animation est réalisée avec le déplacement continu de la fenêtre de temps sur l'axe temporel. Si la communauté C_p dans laquelle se trouve u dans l'intervalle considéré est différente celle où il se trouvait dans l'intervalle précédent alors alors u migre vers cette nouvelle communauté.

Sans surprise, les communautés les plus actives sont les "communautés centrales", celles qui émergent de la composante connexe la plus importante de G et qui rassemblent surtout des personnalités politiques. Ce sont aussi celles qui présentent le plus de diversité en terme de programmes. Lorsqu'on la considère sur l'intégralité de la période 2011-2015, cette composante connexe montre une grande stabilité et possède un coefficient de participation approchant 1.

L'animation montre que les trajectoires de migration des sommets sont le plus souvent des trajectoires qui relient toutes les communautés aux communautés centrales, quelle que soit la granularité de l'intervalle choisi. Nous observons fréquemment des phénomènes d'oscillation sur ces trajectoires.

5 Conclusion et Perspectives

Ce travail constitue les prémisses d'un outil d'analyse de la visibilité médiatique des individus et des communautés. En l'état, la méthodologie (indicateurs et visualisation) intéresse déjà les chercheurs en sciences politiques et infocom avec qui nous collaborons, et auprès de qui nous validons nos choix de conception (indicateurs calculés, représentations des réseaux et communautés, variables visuelles, interactions).

Il apparait utile de revenir sur la méthode de filtrage du graphe de départ qui repose pour l'heure sur l'approche de (Nick et al., 2013). A l'inverse de ce qui est fait ici, il serait intéressant de détecter les communautés de faible stabilité, sans être trop instable, pour lesquelles nos premières expérimentations montrent des phénomènes d'oscillation. Si ces mouvements s'avèrent pérennes dans le temps, alors une évolution plus classique d'apparition, disparition, séparation ou combinaison des communautés serait plus riche en terme de perception.

Références

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

D. Greene, D. D. (2010). Tracking the evolution of communities in dynamic social networks. pp. 1–13.

Falkowski, T. (2009). *Community analysis in dynamic social networks*. Sierke.

Guimerà, R., S. Mossa, A. Turtschi, et L. A. N. Amaral (2005). The worldwide air transportation network : anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences of the United States of America 102*(22), 7794–7799.

Lin, Y.-R., Y. Chi, S. Zhu, et H. Sundaram (2008). Facetnet : a framework for analyzing communities and their evolutions in dynamic networks. pp. 685–694.

M. Oliveira, J. G. (2010). Bipartite graphs for monitoring clusters transitions. pp. 114–124.

Nick, B., C. Lee, P. Cunningham, et U. Brandes (2013). Simmelian backbones : Amplifying hidden homophily in facebook networks. In *Advances in Social Network Analysis and Mining (ASONAM)*, pp. 525–532.

Tantipathananandh, C. et T. Y. Berger-Wolf (2011). Finding communities in dynamic social networks. In *2011 IEEE 11th International Conference on Data Mining*, pp. 1236–1241. IEEE.

Summary

The identification of communities in graphs has been largely addressed in the literature, although the case of dynamic graphs still remains open. This article proposes an approach to compute representative communities in dynamic graphs, measure their stability and level of activity, and visualize their evolution over time.