Identification de rôles communautaires dans des réseaux orientés appliquée à Twitter

Nicolas Dugué*, Vincent Labatut**, Anthony Perez*

*Université d'Orléans, LIFO EA 4022, F-45067 Orléans, France
**Galatasaray University, Computer Science Department,
Çırağan cad. n°36, Ortaköy 34357, İstanbul, Turquie

Résumé. La notion de structure de communautés est particulièrement utile pour étudier les réseaux complexes, car elle amène un niveau d'analyse intermédiaire, par opposition aux plus classiques niveaux local (voisinage des nœuds) et global (réseau entier). Le concept de rôle communautaire permet de décrire le positionnement d'un nœud en fonction de sa connectivité communautaire. Cependant, les approches existantes sont restreintes aux réseaux non-orientés, utilisent des mesures topologiques ne considérant pas tous les aspects de la connectivité communautaire, et des méthodes d'identification des rôles non-généralisables à tous les réseaux. Nous proposons de résoudre ces problèmes en généralisant les mesures existantes, et en utilisant une méthode non-supervisée pour déterminer les rôles. Nous illustrons l'intérêt de notre méthode en l'appliquant au réseau de Twitter. Nous montrons que nos modifications mettent en évidence les rôles spécifiques d'utilisateurs particuliers du réseau, nommés capitalistes sociaux.

1 Introduction

Les réseaux complexes sont des graphes modélisant des systèmes réels. La structure de communautés (Fortunato, 2010) d'un réseau complexe est une partition de l'ensemble des nœuds, dont les parties (communautés) sont des groupes de nœuds densément interconnectés. Cette structure permet l'étude du réseau à un niveau intermédiaire, par comparaison avec les plus classiques niveaux local (voisinage du nœud) et global (réseau entier). Le rôle communautaire décrit ainsi la position d'un nœud dans le réseau à ce niveau. Il a été initialement introduit par Guimerà et Amaral (2005). Ces auteurs caractérisent le positionnement communautaire de chaque nœud au moyen de deux mesures topologiques *ad hoc*. Les nœuds sont ensuite catégorisés au moyen de seuils prédéfinis pour ces mesures. Cette approche peut être critiquée sur trois points. Premièrement, elle est définie seulement pour des réseaux non-orientés. Deuxièmement, les mesures utilisées ne prennent pas en compte tous les aspects de la connectivité communautaire d'un nœud. Troisièmement, rien ne garantit que les seuils fixés empiriquement pour définir les rôles soient pertinents pour d'autres données. Nous proposons des solutions à ces trois problèmes. Pour le premier, nous adaptons les mesures de Guimerà & Amaral aux réseaux orientés. Pour le deuxième, nous définissons des mesures supplémentaires distinguant trois aspects de la connectivité communautaire : diversité des communautés, hétérogénéité de

la distribution des liens, et intensité de la connexion. Pour le troisième, nous proposons une méthode non-supervisée de définition des rôles. Afin d'illustrer l'intérêt de notre méthode, nous l'appliquons à l'étude du rôle communautaire d'un type particulier d'utilisateur de Twitter, appelé *capitaliste social*. Ces utilisateurs mettent en oeuvre deux principes simples pour accroître leur nombre de followers et donc leur visibilité. *Follow Me, I Follow You* (FMIFY) : le capitaliste promet aux utilisateurs qui le suivent de les suivre en retour. *I Follow You, Follow Me* (IFYFM) : le capitaliste suit un maximum d'utilisateurs, en espérant être suivi en retour.

Dans la section suivante, nous décrivons l'approche originale de Guimerà & Amaral. Nous mettons ensuite en évidence ses limitations et décrivons les trois modifications que nous proposons pour les résoudre. Dans la section 3, nous présentons les rôles obtenus sur le réseau Twitter et discutons du positionnement des capitalistes sociaux. Enfin, nous concluons en indiquant les perspectives ouvertes par ce travail.

2 Méthode proposée

2.1 Approche originale

Pour caractériser les rôles des nœuds, Guimerà & Amaral définissent deux mesures, qui leur permettent de placer chaque nœud dans un espace bidimensionnel. Puis, ils proposent plusieurs seuils pour discrétiser cet espace, chaque zone ainsi définie correspondant à un rôle. La première mesure, le *degré intra-module* traite de la connectivité interne du nœud, i.e. des liens avec sa propre communauté. Elle est basée sur la notion de z-score. Comme celle-ci sera réutilisée plus loin, nous la définissons ici de façon générique.

Équation 1 (Z-score). Pour une mesure nodale quelconque $f(u)$, permettant d'associer une valeur numérique à un nœud u, le z-score $Z_f(u)$ par rapport à la communauté de u est : $Z_f(u) = \frac{f(u)-\mu_i(f)}{\sigma_i(f)}$, avec $u \in C_i$ où C_i représente une communauté, et $\mu_i(f)$ et $\sigma_i(f)$ dénotent respectivement la moyenne et l'écart-type de f sur les nœuds de la communauté C_i.

Le degré intra-module $z(u)$ correspond au z-score du degré interne, calculé pour la communauté du nœud considéré. On l'obtient donc en substituant le degré interne d_{int} à f dans l'équation (1). La seconde mesure, appelée *coefficient de participation*, traite de la connectivité externe du nœud, i.e. relative à toutes les communautés auxquelles il est lié.

Équation 2 (Coefficient de participation). $P(u) = 1 - \sum_i \left(\frac{d_i(u)}{d(u)}\right)^2$ où $d_i(u)$ représente le nombre de liens que u possède vers des nœuds de la communauté C_i, et $d(u)$ le degré de u.

Guimerà et Amaral (2005) caractérisent le rôle d'un nœud dans un réseau en se basant sur ces deux mesures. Pour ce faire, ils définissent sept rôles différents en discrétisant l'espace à deux dimensions formé par z et P. Un premier seuil sur le degré intra-module z permet de distinguer les *pivots* des *non-pivots*. Ces pivots (*hubs* en anglais) sont fortement intégrés à leur communauté, par rapport au reste des nœuds de cette même communauté. Ces deux catégories sont subdivisées au moyen d'une série de seuils définis sur le coefficient de participation P. En considérant les nœuds par participation croissante, Guimerà & Amaral les qualifient de *provinciaux* ou *(ultra-)périphériques*, *connecteurs* et *orphelins*.

2.2 Orientation des liens

On se propose d'abord d'adapter les mesures originales au cas orienté. Nous notons d^{in} le degré entrant d'un nœud, i.e. le nombre de liens entrants connectés à ce nœud. Nous pouvons ainsi définir le *degré entrant interne* d'un nœud, noté d_{int}^{in} et représentant le nombre de liens entrants que le nœud possède à l'intérieur de sa communauté. En calculant le z-score de cette valeur, nous obtenons ainsi le *degré intra-module entrant*, noté z^{in}. De manière similaire, nous définissons d_i^{in} comme le *degré communautaire entrant*, à savoir le nombre de liens entrants qu'un nœud a avec les nœuds de la communauté C_i. Cela nous permet de définir le *coefficient de participation entrant*, noté P^{in}, en remplaçant d par d^{in} et d_i par d_i^{in} dans l'équation (2). Le *degré intra-module sortant* z^{out} et le *coefficient de participation sortant* P^{out} sont obtenus de façon symétrique, en utilisant les contreparties sortantes des degrés entrants : d^{out}, d_{int}^{out} et d_i^{out}.

2.3 Aspects de la connectivité externe

Le coefficient de participation se concentre sur un aspect de la connectivité externe d'un nœud : l'*hétérogénéité* de la distribution de ses liens, relativement aux communautés auxquelles il est connecté. Mais il est possible de caractériser cette connectivité de deux autres manières. On peut considérer sa *diversité*, c'est à dire le nombre de communautés concernées, ainsi que son *intensité*, i.e. le nombre de liens concernés. Par souci de simplicité, nous présentons les mesures dans un contexte non-orienté. L'adaptation aux réseaux orientés peut se faire en distinguant les liens entrants et sortants, comme en section 2.2.

Diversité. La *diversité* $D(u)$ évalue le nombre de communautés différentes auxquelles le nœud u est connecté. Soit $\epsilon(u)$ le nombre de communautés, autres que la sienne, auxquelles u est connecté. $D(u)$ est définie comme le z-score d'ϵ relativement à la communauté de u.

Intensité externe. L'*intensité externe* $I_{ext}(u)$ mesure la force de la connexion de u à des communautés externes, en termes de nombre de liens. Soit $d_{ext}(u)$ le degré externe de u, i.e. le nombre de liens que u possède avec des nœuds n'appartenant pas à sa communauté. $I_{ext}(u)$ est alors définie comme le z-score du degré externe.

Hétérogénéité. L'*hétérogénéité* $H(u)$ quantifie la variation du nombre de connexions externes de u d'une communauté à l'autre. Nous utilisons l'écart-type du nombre de liens externes que u possède par communauté, noté $\lambda(u)$, et définission $H(u)$ comme le z-score de λ, relativement à la communauté de *u*.

Intensité interne. Pour représenter la connectivité interne du nœud, nous conservons la mesure z de Guimerà & Amaral. Celle-ci est construite sur la base du z-score, et est donc cohérente avec les mesures définies pour décrire la connectivité externe. Cependant, par symétrie avec notre intensité externe, nous désignons z sous le nom d'*intensité interne*, et la notons $I_{int}(u)$.

2.4 Identification non-supervisée des rôles

Guimerà et Amaral (2005) supposent que les seuils sur les mesures établis de façons empiriques pour définir les rôles sont indépendants des jeux de données utilisés. Pourtant, seule P est normalisée sur un intervalle fixé. En effet, z n'est pas limitée, et donc rien ne garantit que

le seuil défini pour cette mesure reste cohérent pour d'autres réseaux. Cet argument est d'autant plus fort que toutes nos mesures présentées en section 2.3 sont des z-scores. De plus, leur nombre élevé (8) rend l'utilisation de la typologie originale impossible. Afin de contourner ces problèmes, nous proposons d'appliquer une méthode de classification non supervisée. Dans un premier temps, nous calculons l'ensemble des mesures sur les données considérées. Ensuite, nous appliquons une analyse de regroupement. Chaque groupe ainsi identifié correspond a un rôle communautaire.

3 Résultats

Le réseau sur lequel nous avons travaillé comporte un peu moins de 55 millions de nœuds représentant les utilisateurs de Twitter, et près de 2 milliards d'arcs orientés qui matérialisent les abonnements entre utilisateurs. La détection de communautés a été réalisée au moyen de l'algorithme de Louvain (Blondel et al., 2008). L'analyse de regroupement a ensuite été menée au moyen d'une implémentation libre et distribuée de l'algorithme des k-moyennes (Liao, 2009). Nous avons appliqué cet algorithme pour des valeurs de k allant de 2 à 15, et avons sélectionné la meilleure partition d'après l'indice de Davies et Bouldin (1979). Pour valider les résultats obtenus, nous étudions les rôles détectés pour les capitalistes sociaux, identifiés via la méthode proposée par Dugué et Perez (2013). Nous distinguons différentes catégories de capitalistes sociaux en fonction de deux de leurs caractéristiques topologiques. La première est le *ratio*. Il s'agit du nombre de followees divisé par le nombre de followers. Ce critère permet de distinguer ceux qui appliquent la méthode FMIFY (ratio inférieur à 1) de ceux utilisant IFYFM (ratio supérieur à 1). La seconde est le degré entrant : nous séparons ceux de faible degré (entre 500 et 10000) et ceux de degré élevé (supérieur à 10000).

3.1 Calculs effectués

Nous avons d'abord appliqué l'approche originale (non-orientée) de Guimerà & Amaral sur nos données. Les valeurs de z obtenues sont bien supérieures à celles observées dans Guimerà et Amaral (2005). Le seuil défini pour z n'est ainsi plus utilisable pour l'identification des rôles. Nous avons donc procédé à une analyse de regroupement qui identifie 2 rôles, contenant chacun trop de nœuds pour obtenir une information pertinente. Nous avons ensuite appliqué l'approche originale orientée (section 2.2). L'analyse de regroupement a identifié 6 rôles, ce qui montre l'intérêt des mesures orientées. Néanmoins, lorsque l'on regarde le positionnement des capitalistes sociaux au sein de ces groupes, certaines incohérences apparaissent. Une large majorité des capitalistes sociaux de degré élevé est ainsi classée comme non-pivots périphériques ou ultra-périphériques. Ces nœuds ont pourtant un degré entrant supérieur à 10000. Si ces nœuds ne sont pas pivots, donc peu connectés en interne, ils devraient néanmoins être connectés avec l'extérieur. Cela vient des limitations de la participation, comme indiqué en section 2.3. Nous avons enfin appliqué le dernier groupe de 8 mesures, aboutissant aux résultats discutés dans le reste de cette section.

3.2 Étude des groupes

L'analyse de regroupement nous donne $k = 6$ groupes (Tableau 1). Nous caractérisons ces groupes relativement à nos huit mesures, afin d'en identifier les rôles. Dans les groupes 1, 4 et 5, presque toutes les mesures sont négatives et proches de 0. Il ne s'agit donc pas de pivots (nœud largement connecté à sa communauté) ni de de nœud qualifiés de connecteurs (ayant une connexion privilégiée avec d'autres communautés que la leur). La diversité entrante (resp. sortante) du groupe 4 (resp. 5) est l'unique mesure positive, ceci indique que ces noeuds reçoivent (resp. envoient) des liens d'un nombre relativement élevé de communautés et sont moins isolés. Toutes les mesures sont positives dans le groupe 6. L'intensité interne reste proche de 0, donc on ne peut toujours pas parler de pivot. L'intensité externe faible, mais positive, et la diversité élevée permettent de considérer ces nœuds comme des connecteurs. Toutes les mesures du groupe 3 sont largement positives L'intensité interne élevée associe ce groupe au rôle de pivot. Les valeurs externes montrent que ces nœuds sont connectés à de nombreux nœuds présents dans de nombreuses autres communautés. Toutefois, les liens sortants sont plus nombreux, ces nœuds correspondent donc à des utilisateurs plus suiveurs que suivis. Toutes les mesures du groupe 2 sont particulièrement élevées. Pour une mesure donnée, la variante concernant les liens entrants est toujours largement supérieure, ce qui signifie que les utilisateurs représentés par ces nœuds sont particulièrement suivis. Nous associons ce groupe au rôle de pivot orphelin.

Groupe	**Taille**	**Proportion**	**Rôle**
1	24543667	$46,68\%$	Non-pivot ultra-périphérique
2	304	$< 0,01\%$	Pivot orphelin
3	303674	$0,58\%$	Pivot connecteur
4	11929722	$22,69\%$	Non-pivot périphérique (entrant)
5	10828599	$20,59\%$	Non-pivot périphérique (sortant)
6	4973717	$9,46\%$	Non-pivot connecteur

TAB. 1: Tailles des groupes détectés et rôles dans la typologie de Guimerà & Amaral.

3.3 Positionnement des capitalistes sociaux

Avec la méthode définie par Dugué et Perez (2013), nous détectons près de 160000 capitalistes sociaux. Nous étudions leur positionnement dans les 6 groupes identifiés. Les capitalistes sociaux de faible degré se retrouvent dans trois groupes : 3, 5 et 6. Les nœuds du groupe 3 sont des pivots connecteurs qui suivent plus d'utilisateurs du réseau que la normale, ce qui est cohérent avec le comportement des capitalistes sociaux. Il semble également cohérent d'observer que les capitalistes sociaux dont le degré sortant est supérieur au degré entrant sont près de deux fois plus présents dans ce groupe que les autres. La majorité des capitalistes sociaux de faible degré se place au sein du groupe 6, non-pivot connecteur. Ces nœuds légèrement plus connectés au sein de leur communauté et avec l'extérieur que la moyenne, ont en revanche une diversité bien plus élevée. Les capitalistes sociaux qui s'y situent semblent ainsi avoir débuté l'application de leurs méthodes, en créant des liens avec de nombreuses autres communautés. Les capitalistes sociaux de degré élevé se placent presque exclusivement dans les groupes 2 et 3, pivots connecteurs et orphelins. Cela semble cohérent avec les degrés élevés de ces nœuds.

Les nœuds classés dans le groupe 2 sont ceux de ratio inférieur à 0,7 avec beaucoup plus de followers que de followees, ce qui correspond à la définition du rôle donné par nos mesures. Notre approche établit une séparation entre capitalistes sociaux de faible degré, majoritairement connecteurs et non-pivots et ceux de degré élevé, classés pivots.

Ratio	G3	G5	G6
< 1	$\mathbf{23.10\%}$	$\mathbf{18.28\%}$	$\mathbf{55.19\%}$
> 1	$\mathbf{18.78\%}$	$\mathbf{14.31\%}$	$\mathbf{66.40\%}$

Ratio	G2	G3
< 0.7	$\mathbf{12.14\%}$	$\mathbf{87.29\%}$
> 1	0.03%	$\mathbf{97.99\%}$

TAB. 2: Répartition des capitalistes sociaux de degré faible (gauche) et élevé (droite).

4 Perspectives

Le travail présenté peut s'étendre de différentes façons. Tout d'abord, certains des rôles définis dans Guimerà et Amaral (2005) n'apparaissent pas dans notre analyse. Il serait intéressant d'étudier d'autres réseaux afin de déterminer si cette observation reste valable. Une autre piste consiste à baser nos calculs sur des communautés recouvrantes (i.e. non-mutuellement exclusives). En effet, les réseaux sociaux que nous étudions sont réputés posséder ce type de structures, dans lesquelles un nœud peut appartenir à plusieurs communautés en même temps. L'adaptation de nos mesures à ce contexte se ferait naturellement, en définissant des versions internes de l'hétérogénéité et de la diversité.

Références

Blondel, V., J.-L. Guillaume, R. Lambiotte, et E. Lefebvre (2008). Fast unfolding of communities in large networks. *J. Stat. Mech. 10*, P10008.

Davies, D. et D. Bouldin (1979). A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell. 1*(2), 224–227.

Dugué, N. et A. Perez (2013). Detecting social capitalists on twitter using similarity measures. In *Complex Networks IV*, Volume 476 of *Studies in Computational Intelligence*, pp. 1–12. Springer.

Fortunato, S. (2010). Community detection in graphs. *Phys. Rep. 486*(3-5), 75–174.

Guimerà, R. et L. Amaral (2005). Functional cartography of complex metabolic networks. *Nature 433*, 895–900.

Liao, W.-K. (2009). Parallel k-means data clustering.

Summary

Community structure is useful when analyzing complex networks. It provides an intermediate level, compared to the more classic global (whole network) and local (node neighborhood) approaches. Community role describes the position of a node in a network depending on its connectivity at this level. However, the existing approaches are restricted to undirected networks, use topological measures which do not consider all aspects of community-related connectivity, and their role identification methods are not generalizable to all networks. We tackle these limitations and illustrate the applicability of our method by analyzing a Twitter network.