Identification de rôles communautaires dans des réseaux orientés appliquée à Twitter

Nicolas Dugué, Vincent Labatut, Anthony Perez

Identification de rôles communautaires dans des réseaux orientés appliquée à Twitter

Nicolas Dugué, Vincent Labatut, Anthony Perez

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 a été dérivé sur cette base, afin de décrire le positionnement d'un nœud en fonction de sa connectivité communautaire. Cependant, les approches existantes sont restreintes aux réseaux non-orientés, elles 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 et étendant les mesures existantes, et en utilisant une méthode non-supervisée pour déterminer les rôles. Nous illustrons l'intérêt de notre méthode en l'appliquant à l'analyse du réseau de Twitter. Nous montrons que nos modifications permettent de mettre 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. Leurs propriétés topologiques ont récemment fait l'objet de nombreux travaux, dont un certain nombre s'est concentré sur leur structure de communautés (Fortunato, 2010). Dans sa forme la plus simple, cette structure est une partition de l'ensemble des nœuds, dont les parties (communautés) sont des groupes de nœuds densément interconnectés. La notion de communauté est particulièrement intéressante, car elle 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 concept de rôle communautaire illustre bien cette caractéristique : il décrit la position d'un nœud dans sa communauté. Il a été initialement introduit par Guimerà et Amaral (2005), puis indépendamment par Scripps et al. (2007). En se basant sur une estimation de la structure de communautés, ces auteurs caractérisent le positionnement communautaire de chaque nœud au moyen de plusieurs mesures topologiques *ad hoc*. Les nœuds sont ensuite catégorisés au moyen de seuils prédéfinis pour chaque mesure.

Ces approches peuvent être critiquées sur trois points. Premièrement, elles sont définies seulement pour des réseaux non-orientés. Pourtant, de nombreux systèmes contiennent des

relations asymétriques, et ne pas en tenir compte constitue une perte significative d'information. 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. Dans ce travail, 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, utilisant les mesures proposées.

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*. Le principe du capitalisme social est d'essayer d'obtenir un maximum de visibilité en utilisant diverses techniques. Cette notion a été mise en évidence dans un travail sur les comptes spammers de Twitter par Ghosh et al. (2012). Sur Twitter, deux principes relativement simples sont principalement utilisés pour accroître le nombre de followers et donc la 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. De tels utilisateurs peuvent être néfastes pour l'équilibre du réseau social, dans la mesure où leurs comptes gagnent en visibilité et leurs tweets sont bien classés par les moteurs de recherche du réseau, mais souvent sans réelle raison de contenu. L'amélioration de la qualité du service passe donc par une bonne compréhension de leur positionnement dans le réseau, et donc dans les communautés. On peut en effet se demander si ces utilisateurs sont ancrés dans leur communauté, étroitement liés aux autres utilisateurs, ou s'ils sont au contraire isolés. Une autre question est de savoir s'ils sont liés aux autres communautés, et, si oui, avec quelle intensité.

Dans la section suivante, nous décrivons l'approche originale de Guimerà & Amaral. Nous mettons ensuite en évidence ses limitations, et proposons notre propre approche en section 3. Dans la section 4, 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 Approche originale

Nous avons décidé de construire notre méthode à partir de celle de Guimerà et Amaral (2005), non seulement parce qu'elle est plus répandue que celle de Scripps et al. (2007), mais aussi parce qu'elle s'appuie plus fortement sur la structure de communauté. Pour caractériser les rôles des nœuds, Guimerà & Amaral définissent d'abord deux mesures complémentaires, 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 particulier. Dans cette section, nous décrivons d'abord les mesures, puis la méthode qu'ils utilisent pour identifier les rôles.

La première mesure, nommée *degré intra-module* (*within-module degree* en anglais) 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. Pour une fonction 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 :

$$Z_f(u) = \frac{f(u) - \mu_i(f)}{\sigma_i(f)}, \text{ avec } u \in C_i \tag{1}$$

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 appartenant à la communauté C_i. Le degré intra-module de Guimerà et Amaral, noté $z(u)$, 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). Le degré intra-module évalue la connectivité d'un nœud à sa communauté relativement à celle des autres nœuds de sa communauté. La seconde mesure, appelée *coefficient de participation*, traite de la connectivité externe du nœud, i.e. relative à toutes les communautés auquel il est lié. Elle est définie de la manière suivante :

$$P(u) = 1 - \sum_i \left(\frac{d_i(u)}{d(u)} \right)^2 \tag{2}$$

où $d_i(u)$ représente le nombre de liens que u possède vers des nœuds de la communauté C_i. Notons que dans le cas où C_i est la communauté de u, alors on a $d_i(u) = d_{int}(u)$. Le coefficient de participation représente combien les connexions d'un nœud sont diversifiées, en termes de communauté externes. Une valeur proche de 1 signifie que le nœud est connecté de façon uniforme à un grand nombre de communautés différentes. Au contraire, une valeur de 0 ne peut être atteinte que si le nœud n'est connecté qu'à une seule communauté (vraisemblablement la sienne).

Guimerà et Amaral (2005) proposent de caractériser 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 défini sur le degré intra-module z permet de distinguer ce que les auteurs appellent les *pivots communautaires* ($z \geq 2.5$) des autres nœuds ($z < 2.5$). Ces pivots (*hubs* en anglais) sont considérés comme fortement intégrés à leur communauté, par rapport au reste des nœuds de cette même communauté. Ces deux catégories (pivot et non-pivot) sont subdivisées au moyen d'une série de seuils définis sur le coefficient de participation P. En considérant les nœuds par participation croissante, Guimerà et Amaral les qualifient de *provinciaux* ou *(ultra-)périphériques*, *connecteurs* et *orphelins*. Les deux premiers rôles sont essentiellement connectés à leur communauté, les troisièmes, bien qu'eux aussi potentiellement bien connectés à leur propre communauté, sont également largement liés à d'autres communautés, et les derniers sont connectés à un grand nombre de communautés.

3 Méthode proposée

Dans cette section, nous décrivons les trois modifications que nous proposons pour résoudre les limitations précédemment mentionnées de l'approche de Guimerà & Amaral. Nous expliquons tout d'abord comment étendre les mesures pour le cas d'un réseau orienté, puis proposons des mesures supplémentaires permettant de mieux évaluer la connectivité externe des nœuds, et enfin une méthode non-supervisée pour déterminer les rôles nodaux.

3.1 Orientation des liens

Il est souvent assez simple de généraliser des mesures définies sur des graphes non-orientés vers des graphes orientés. En effet, la schéma classique consiste à distinguer les liens *entrants* des liens *sortants*. Dans notre cas, cela consister à utiliser 4 mesures au lieu de 2 : degrés intra-module entrant et sortant, ainsi que coefficients de participation entrant et sortant.

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^{in}_{int} 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^{in}_i 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^{in}_i 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^{out}_{int} et d^{out}_i.

Remarquons ici qu'il ne serait pas pertinent de calculer de telles mesures en se basant sur un ensemble de communautés détecté par un algorithme ne tenant pas compte de l'orientation des liens. Il est donc nécessaire d'utiliser un algorithme prenant en compte cette information. Un autre point important est la définition des rôles. Il n'y a a priori aucune raison pour que les seuils définis par Guimerà et Amaral (2005) soient toujours valables pour les versions orientées des mesures. Nous verrons cependant que la méthode d'identification des rôles non supervisée que nous proposons dans la Section 3.3 permet de résoudre ce problème.

3.2 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. Premièrement, on peut considérer sa *diversité*, c'est à dire le nombre de communautés concernées. Deuxièmement, il est possible de s'intéresser à son *intensité*, i.e. au nombre de liens concernés. Comme le montre la Figure 1, ces deux aspects ne sont pas pris en compte dans P. En effet, bien que la connectivité externe du nœud central soit différente sur chacune des trois figures, le coefficient de participation reste le même. Pour pallier cette limitation, nous proposons deux nouvelles mesures permettant de quantifier la diversité et l'intensité. De plus, afin d'obtenir un ensemble cohérent de mesures, nous révisons également P.

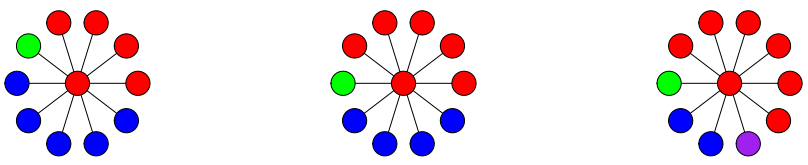

FIG. 1 – *Chaque couleur représente une communauté. Dans chaque cas, le coefficient de participation du nœud central est* $0,58$.

Diversité. Notre mesure de *diversité*, notée $D(u)$, évalue le nombre de communautés différentes auxquelles le nœud u est connecté , indépendamment de la densité de ces connexions. Soit $\epsilon(u)$ le nombre de communautés, autres que la sienne, auxquelles le nœud u est connecté. Alors la diversité est définie comme le z-score d'ϵ relativement à la communauté de u. C'est à dire qu'on l'obtient en substituant ϵ à f dans (1).

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, et relativement aux autres nœuds de sa communauté. Soit $d_{ext}(u)$ le degré externe de u, correspondant au nombre de liens que u possède avec des nœuds n'appartenant pas à sa communauté. Remarquons qu'on a alors $d = d_{int} + d_{ext}$. Nous définissons l'intensité externe comme le z-score du degré externe, c'est à dire qu'on l'obtient en substituant d_{ext} à f dans (1).

Hétérogénéité. L'*hétérogénéité* $H(u)$ quantifie combien le nombre de connexions externes du nœud u varie d'une communauté à l'autre. Nous utilisons pour cela l'écart-type du nombre de liens externes que le nœud possède par communauté, que nous notons $\lambda(u)$. L'hétérogénéité est alors le z-score de λ, relativement à la communauté de *u*, et on l'obtient donc en substituant λ à f dans (1). Cette mesure a une signification très proche de celle du coefficient de participation P de Guimerà et Amaral. Elle diffère en ce qu'elle est exprimée relativement à la communauté de u, et que les liens internes à cette même communauté sont exclus du calcul.

Intensité interne. Pour représenter la connectivité interne du nœud, nous conservons la mesure z de Guimerà et Amaral. En effet, celle-ci est construite sur la base du z-score, et est donc cohérente avec les autres mesures définies pour décrire la connectivité externe. De plus, il n'est pas nécessaire de lui adjoindre d'autres mesures, car les notions d'hétérogénéité et de diversité n'ont pas de sens ici (puisqu'on considère seulement une seule communauté). Cependant, en raison de sa symétrie avec notre intensité externe, nous désignons z sous le nom d'*intensité interne*, et la notons $I_{int}(u)$.

Pour chacune des 4 mesures présentées, nous utilisons deux variantes, l'une considère les liens entrants, l'autre les liens sortants.

3.3 Identification non-supervisée des rôles

Notre dernière modification de l'approche de Guimerà et Amaral (2005) concerne la manière dont les rôles sont définis. Guimerà et Amaral (2005) supposent l'existence de rôles universels, présents dans tous les systèmes. Ils supposent notamment que les seuils établis de façons empiriques pour définir les rôles sont indépendants des jeux de données utilisés. Cette dernière partie est sujette à discussion, dans la mesure où, parmi leurs deux mesures, seules P est normalisée sur un intervalle fixé. En effet, il n'y a aucune limitation pour z, et il n'y a donc aucune garantie que le seuil défini originellement pour cette mesure reste cohérent pour d'autres réseaux. Si nous considérons les mesures présentées dans la section 3.2, cet argument est d'autant plus fort que *toutes* nos mesures sont définies comme des z-scores. De plus, le fait que nous proposions 8 mesures fait croître le nombre de seuils nécessaires de manière significative, et rend la sélection de seuil originelle impossible à utiliser.

Afin de contourner ces problèmes, nous proposons d'appliquer une méthode automatique 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. Cette méthode présente l'avantage de ne pas être affectée par le nombre de mesures utilisé, et revient à ajuster les seuils pour le système considéré.

4 Résultats

Le réseau sur lequel nous avons travaillé a été collecté en 2009 par Cha et al. (2010). Il 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, à savoir les liens de "follow". La très grande taille de ces données a influencé le choix de nos outils d'analyse. La détection de communautés a été réalisée au moyen de l'algorithme de Louvain (Blondel et al., 2008), très efficace pour le traitement de grands réseaux. Nous avons repris le code mis à disposition par ses auteurs et l'avons adapté à la modularité orientée décrite par Leicht et Newman (2008). L'analyse de regroupement a alors 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). L'ensemble de notre code source est disponible à l'adresse `https://github.com/CompNet/Orleans`.

Pour valider les résultats obtenus, nous étudions la position des capitalistes sociaux dans les rôles détectés. Ceux-ci sont identifiés en utilisant la méthode proposée par Dugué et Perez (2013), basée sur l'utilisation de deux mesures topologiques spécifiques. Pour faciliter l'interprétation des résultats, à l'instar de Dugué et Perez, 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).

4.1 Approche originale

Nous avons tout 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 sur la connectivité des nœuds du réseau relativement à la structure de communautés. La perte d'information due au fait que la méthode originale ne tienne pas compte de l'orientation des liens peut expliquer ces résultats.

Nous avons ensuite appliqué l'approche originale adaptée aux graphes orientés, telle que décrite dans la Section 3.1. L'analyse de regroupement a identifié 6 rôles : un groupe de nœuds pivots (nœuds mieux connectés que les autres à leur communauté), et 5 groupes de nœuds non-pivots. Les nœuds non-pivots sont séparés selon la distribution de leurs liens externes en utilisant les coefficients de participation. On retrouve ainsi des non-pivots périphériques et

ultra-périphériques, considérés comme peu connectés aux communautés externes (de faibles P^{in} ou P^{out}), et des non-pivot orphelins connectés de façon homogène avec les communautés externes (P^{in} ou P^{out} élevés). La diversité des rôles obtenus montre l'intérêt des mesures orientées par rapport à l'approche non-orientée, qui avait considéré comme équivalents plusieurs de ces groupes. 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 étant non-pivots périphériques ou ultra-périphériques. Ces nœuds ayant un degré entrant supérieur à 10000, et pour certains un degré sortant supérieur, cela semble surprenant. En effet, si ces nœuds ne sont pas pivots, donc peu connectés, ils devraient néanmoins être connectés avec l'extérieur. Cela vient du fait que la participation évalue l'*hétérogénéité* des connexions aux communautés externes, sans tenir compte de l'intensité ou de la diversité de ces connexions, comme précisé dans la section 3.2. La classification ainsi obtenue ne sépare pas les capitalistes sociaux selon leurs degrés ou ratios. Ceux-ci se trouvent majoritairement dans les groupes de nœuds considérés comme périphériques et ultra-périphériques.

Afin de dépasser les limites inhérentes à la participation, nous appliquons donc nos mesures proposées sur les données et présentons les résultats obtenus dans la sous-section qui suit.

4.2 Groupes

Considérons tout d'abord les mesures obtenues sur l'ensemble des données traitées. On observe des corrélations positives pour l'ensemble des paires de mesures, allant de valeurs proches de 0 à $0,9$. Les deux variantes d'une même mesure (liens entrants contre liens sortants) sont peu corrélées, ce qui confirme une nouvelle fois l'intérêt de tenir compte de l'orientation dans notre étude. Trois mesures sont fortement corrélées : les intensités internes et externes et l'hétérogénéité (ρ allant de $0,78$ à $0,92$). Le lien entre les intensités interne et externe semble indiquer que les variations dans le degré total d'un nœud ont globalement le même effet sur ses degrés internes et externes. Autrement dit, la proportion entre ces deux types de liens ne dépend pas du degré du nœud. Le très fort lien observé entre hétérogénéité et intensité indique que seuls les nœuds de faible intensité sont connectés de façon homogène à des communautés externes, tandis que les nœuds possédant de nombreux liens sont connectés de façon hétérogène.

En ce qui concerne l'analyse de regroupement, nous obtenons la meilleure séparation pour $k = 6$ groupes, dont le Tableau 1 donne les tailles. Nous avons caractérisé les groupes relativement à nos huit mesures, afin d'en identifier les rôles et de les comparer à ceux définis par Guimerà et Amaral. Le Tableau 2 contient les valeurs moyennes obtenues pour chaque mesure dans chaque groupe. Les ANOVA que nous avons réalisées ont révélé des différences significatives pour toutes les mesures ($p < 0.01$). Un test post-hoc (t-test avec correction de Bonferroni) a montré que ces différences existaient entre tous les groupes, pour toutes les mesures.

Dans le groupe 1, toutes les mesures sont négatives mais proches de 0, à l'exception des deux variantes de la diversité, en particulier l'entrante, qui est proche de -1. Il ne peut pas s'agir de pivot au sens de Guimerà et Amaral (nœud largement connecté à sa communauté), puisque l'intensité interne est négative. De même, les mesures externes sont très faibles ce qui montre qu'il ne s'agit pas non plus de nœud qualifiés de connecteurs par Guimerà et Amaral (ayant une connexion privilégiée avec d'autres communautés que la leur). On peut donc considérer que ce groupe correspond au role des non-pivots ultra-périphériques. Ce groupe est le

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 de groupes détectés, et rôles correspondants dans la typologie de Guimerà et Amaral.*

plus grand (il contient à lui seul 47% des nœuds), ce qui confirme la correspondance avec ce rôle, dont les nœuds constituent généralement la masse du réseau. Relativement au système modélisé, ces nœuds sont caractérisés par le fait qu'ils sont particulièrement peu suivis par les autres communautés.

Le groupe 4 est extrêmement similaire au groupe 1, à la différence que sa diversité entrante est de $0,69$. Ces nœuds restent donc périphériques, car l'intensité externe est toujours négative, mais ils reçoivent néanmoins des liens provenant d'un nombre relativement élevé de communautés. Autrement dit, ils sont suivis par peu d'utilisateurs externes, mais ceux-ci sont situés dans un grand nombres de communautés distinctes. Le groupe 5 est lui aussi très proche du groupe 1, mais la différence est cette fois que les deux variantes de la diversité sont positives, avec une diversité sortante de $0,60$. À l'inverse du groupe 4, on peut donc dire ici que les utilisateurs concernés suivent (avec une faible intensité) des utilisateurs situés dans un grand nombre de communautés différentes. Les groupes 4 et 5 sont respectivement le deuxième (23%) et troisième (21%) plus grands groupes en termes de taille, ce qui porte le total des nœuds périphériques à 91%.

G	$\mathbf{I_{int}}$		D		$\mathbf{I_{ext}}$		H	
1	$-0,12$	$-0,03$	$-0,55$	$-0,80$	$-0,09$	$-0,04$	$-0,12$	$-0,06$
2	$94,22$	$311,27$	$7,18$	$88,40$	$113,87$	$283,79$	$112,79$	$285,57$
3	$5,52$	$1,40$	$5,60$	$3,10$	$5,28$	$1,43$	$6,76$	$2,34$
4	$-0,04$	$0,00$	$-0,37$	$0,69$	$-0,07$	$0,00$	$-0,10$	$-0,01$
5	$-0,03$	$-0,01$	$0,60$	$0,19$	$-0,03$	$-0,02$	$-0,04$	$-0,02$
6	$0,48$	$0,12$	$1,96$	$1,70$	$0,35$	$0,12$	$0,53$	$0,19$

TAB. 2 – *Mesures moyennes obtenues pour les* 6 *groupes. Pour chaque mesure, deux valeurs sont indiquées, correspondant respectivement aux deux variantes : liens sortants et entrants.*

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, même si ces nœuds sont mieux connectés à leur communautés que ceux des groupes précédents. L'intensité externe est elle aussi faible, mais le fait qu'elle soit positive, à l'instar des autres mesures externes, semble suffisante pour considérer ces nœuds comme des connecteurs au sens de Guimerà et Amaral (relativement bien reliés à d'autres communautés). La diversité est relativement élevée, aussi bien pour les liens entrants que sortants ($D > 1,7$). Ces nœuds sont donc plus fortement connectés à leur

communauté mais aussi à l'extérieur, et avec une plus grande diversité. Il s'agit du quatrième plus gros groupe, représentant $9,5\%$ des nœuds.

Toutes les mesures du groupe 3 sont largement positives : supérieures à $1,4$ pour celles basées sur les liens entrants, et supérieures à $5,2$ pour les liens sortants. L'intensité interne élevée permet d'associer ce groupe au rôle de pivot. Les valeurs externes montrent en plus 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. Ce groupe ne représente que $0,6\%$ des nœuds, il s'agit donc d'un rôle bien plus rare que ceux associés aux groupes précédents. Cette observation est encore plus caractéristique du groupe 2, qui représente bien moins de 1% des nœuds. Toutes les mesures y sont particulièrement élevées, la plupart dépassant 100. 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, et donc influents. Nous associons ce groupe au rôle de pivot orphelin defini par Guimerà et Amaral.

En conclusion de cette analyse des groupes, on peut constater que tous les rôles identifiés par Guimerà et Amaral ne sont pas présents dans le réseau étudié : on n'y trouve ni non-pivots orphelins, ni pivots provinciaux. Cette observation semble confirmer la nécessité d'une approche objective pour déterminer comment regrouper les nœuds en fonction des mesures. Elle est également consistante avec la forte corrélation observée entre les intensités interne et externe : les rôles manquants correspondraient à des nœuds possédant une forte intensité interne mais une faible intensité externe, ou vice-versa. Or, ceux-ci sont très peu fréquents dans notre réseau. De plus, le fait de distinguer les liens entrants et sortants permet d'obtenir une typologie plus fine. Ainsi, certains groupes distincts ont émergé (groupes 4 et 5) là où l'approche de Guimerà et Amaral aurait considéré ces nœuds comme équivalents.

4.3 Positionnement des capitalistes sociaux

Avec la méthode définie dans Dugué et Perez (2013), nous détectons près de 160.000 capitalistes sociaux. Nous étudions ici leur positionnement dans les 6 groupes identifiés par la méthode des k-moyennes. De plus, nous affinons notre analyse en structurant les capitalistes sociaux en différents groupes. Tout d'abord via le ratio, qui nous permet de mettre en évidence les comportements FMIFY et IFYFM. Ensuite, en utilisant le degré de ces utilisateurs. En effet, les capitalistes sociaux ayant accru le plus efficacement leur nombre de followers sont susceptibles d'avoir un placement ou un rôle différent au sein des communautés.

Chaque tableau présente ainsi sur la première ligne la proportion de capitalistes sociaux du réseau qui sont contenus dans chaque groupe, et sur la deuxième la proportion de nœuds du groupe qui sont des capitalistes sociaux.

Capitalistes sociaux de faible degré entrant.

Ces capitalistes sociaux se retroupent dans trois groupes : 3, 5 et 6. Les nœuds du groupe 3 sont des pivots connecteurs qui ont en particulier tendance à suivre plus d'utilisateurs du réseau que la normale. Même si le degré entrant des capitalistes sociaux est considéré comme faible ici, il reste élevé relativement au degré moyen du reste du réseau. Cela semble donc cohérent de voir qu'un grand nombre de capitalistes sociaux est plus connecté à la fois à leur communauté mais également aux autres communautés. Il semble également cohérent d'observer que les capitalistes sociaux de type **IFYFM** dont le degré sortant est supérieur au degré

Ratio	G1	G2	G3	G4	G5	G6
< 1	0.01%	0.00%	$\mathbf{23.10\%}$	3.42%	$\mathbf{18.28\%}$	$\mathbf{55.19\%}$
	$< 0.01\%$	0.00%	3.71%	0.14%	0.08%	0.54%
> 1	0.03%	0.00%	$\mathbf{18.78\%}$	0.48%	$\mathbf{14.31\%}$	$\mathbf{66.40\%}$
	$< 0.01\%$	0.00%	$\mathbf{6.61\%}$	$< 0.01\%$	0.14%	1.43%

TAB. 3 – *Répartition des capitalistes sociaux de faible degré dans les différents groupes.*

entrant sont près de deux fois plus présents dans ce groupe que les autres. La diversité sortante élevée du groupe 3 nous apprend également que ces capitalistes sociaux ont tendance à ne pas cibler uniquement leur communauté même s'ils y sont bien connectés, mais à appliquer leurs méthodes à travers de nombreuses communautés du réseau.

On observe que la large majorité des capitalistes sociaux de faible degré se place au sein du groupe 6, non-pivot connecteur. Ces nœuds, qui sont 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.

Enfin, on retrouve une faible proportion de capitalistes sociaux de faible degré dans le groupe 5, groupe de nœuds non-pivots périphériques. Un certain nombre de capitalistes sociaux sont ainsi isolés au sein de leur communauté et avec l'extérieur.

Capitalistes sociaux de degré entrant élevé.

Ratio	G1	G2	G3	G4	G5	G6
< 0.7	0.00%	$\mathbf{12.14\%}$	$\mathbf{87.29\%}$	0.00%	0.00%	0.57%
	0.00%	$\mathbf{21.05\%}$	0.15%	0.00%	0.00%	$< 0.01\%$
> 0.7 et < 1	0.00%	1.55%	$\mathbf{95.64\%}$	0.00%	0.00%	2.81%
	0.00%	$\mathbf{7.24\%}$	0.45%	0.00%	0.00%	$< 0.01\%$
> 1	0.00%	0.03%	$\mathbf{97.99\%}$	0.00%	0.00%	1.98
	0.00%	0.33%	1.22%	0.00%	0.00%	$< 0.01\%$

TAB. 4 – *Répartition des capitalistes sociaux de degré élevé dans les différents groupes.*

Les capitalistes sociaux de degré élevé se placent presque exclusivement dans les groupes 2 et 3. Ces groupes contiennent des nœuds pivots connecteurs et orphelins. Cela semble cohérent avec les degrés élevés de ces nœuds. Ceux-ci sont naturellement plus connectés avec leurs communautés et avec l'extérieur que les autres nœuds. On constate que les nœuds classés dans le groupe 2 sont ceux de ratio inférieur à 1 et particulièrement ceux de ratio inférieur à $0,7$ ayant beaucoup plus de followers que de followees. Cela correspond bien à la définition du rôle donné par nos mesures qui montre que ce groupe de nœuds est suivi par un grand nombre de nœuds provenant d'une large variété de communautés.

En conclusion, on observe ainsi que notre approche permet d'établir une nette séparation entre capitalistes sociaux de faible degré, majoritairement connecteurs et non-pivots et ceux de degré élevé, classé comme pivots. Par ailleurs, les rôles obtenus permettent également de discriminer les utilisateurs de ratios différents. Les capitalistes sociaux de degré élevé et de

ratio inférieurs à 1 sont par exemple les seuls à appartenir au groupe des pivots orphelins. Ce n'était pas le cas avec l'approche originale adaptée aux graphes orientés. Enfin, notre approche permet de mieux décrire les différents rôles obtenus grâce aux trois mesures utilisées pour caractériser la connectivité du nœud aux communautés auxquelles il n'appartient pas.

5 Conclusion

Dans cet article, notre but est de proposer une extension à la méthode définie par Guimerà et Amaral (2005) pour caractériser le role communautaire de nœuds dans des réseaux complexes. Nous définissons d'abord une version orientée des mesures originales, puis nous les étendons pour qu'elle tiennent compte des différents aspects de la connectivité des nœuds (diversité, intensité et hétérogénéité). Nous proposons ensuite une méthode non-supervisée pour déterminer les rôles à partir de ces mesures. Elle a l'avantage d'être indépendante du système étudié. Enfin, nous donnons un exemple d'application en utilisant nos outils pour analyser le rôle des capitalistes sociaux dans Twitter. Notre méthode met en lumière les rôles caractéristiques joués par les capitalistes sociaux. Ceux de degré entrant élevé sont considérés comme des pivots orphelins ou connecteurs, en fonction de leur ratio. Ceux de faible degré entrant sont pour la plupart des non-pivots connecteurs. La prise en compte de l'orientation des liens, notamment, permet d'obtenir des rôles plus pertinents, ce qui confirme l'intérêt d'exploiter cette information lors de l'étude des réseaux sociaux.

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 (Arora et al., 2012) ; de plus, de nombreux algorithmes existent pour les détecter (Xie et al., 2013). L'adaptation de nos mesures à ce contexte se ferait naturellement, en définissant des versions internes de l'hétérogénéité et de la diversité.

Références

Arora, S., R. Ge, S. Sachdeva, et G. Schoenebeck (2012). Finding overlapping communities in social networks : Toward a rigorous approach. In *ACM Conference on Electronic Commerce*.

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

Cha, M., H. Haddadi, F. Benevenuto, et K. Gummadi (2010). Measuring user influence in twitter : The million follower fallacy. In *ICWSM*.

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.

Ghosh, S., B. Viswanath, F. Kooti, N. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, et K. Gummadi (2012). Understanding and combating link farming in the twitter social network. In *WWW*, pp. 61–70.

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

Leicht, E. A. et M. E. J. Newman (2008). Community structure in directed networks. *Phys. Rev. Lett. 100*(11), 118703.

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

Scripps, J., P.-N. Tan, et A.-H. Esfahanian (2007). Node roles and community structure in networks. In *WebKDD/SNAKDD*, San Jose, US-CA, pp. 26–35. ACM.

Xie, J., S. Kelley, et B. Szymanski (2013). Overlapping community detection in networks : the state of the art and comparative study. *ACM Computing Surveys 45*(4).

Summary

The notion of community structure is particularly useful when analyzing complex networks, because it provides an intermediate level, compared to the more classic global (whole network) and local (node neighborhood) approaches. The concept of community role of a node was derived from this base, in order to describe the position of a node in a network depending on its connectivity at the community 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 by generalizing and extending the measures, and using an unsupervised approach to determine the roles. We then illustrate the applicability of our method by analyzing a Twitter network. We show how our modifications allow discovering the fact some particular users called *social capitalists* occupy very specific roles in this system.