Une métrique de sélection de variables appliquée à la centralité et à la détection des rôles communautaires

Nicolas Dugué, Jean-Charles Lamirel

Une métrique de sélection de variables appliquée à la centralité et à la détection des rôles communautaires

Nicolas Dugué*, Jean-Charles Lamirel**

* LIUM - University du Maine, 72000, Le Mans, France
** LORIA - SYNALP, 54000, Nancy, France

Résumé. La F-Mesure de trait est une métrique de sélection de variables statistique sans paramètres qui a montré de bonnes performances pour la classification, l'étiquetage de clusters ou encore la mesure de qualité des clusters. Dans cet article, nous proposons d'évaluer son utilisation dans le contexte des graphes de terrain et de leur structure communautaire pour bénéficier de son système sans paramètres et de ses performances bien évaluées. Nous étudions donc sur des graphes synthétiques réalistes les corrélations qui existent entre la F-Mesure de trait et certaines mesures de centralité, mais surtout avec des mesures destinées à caractériser le rôle communautaire des nœuds. Nous montrons ainsi que cette mesure est liée à la centralité des nœuds du réseau, et qu'elle est particulièrement adaptée à la mesure de leur connectivité au regard de la structure de communautés. Nous observons par ailleurs que les mesures usuelles de détection des rôles communautaires sont fortement dépendantes de la taille des communautés alors que celles que nous proposons sont par définition liées à la densité de la communauté, ce qui rend les résultats comparables d'un réseau à un autre. Ceci offre donc la possibilité d'applications comme le suivi temporel de la structure des communautés. Enfin, le processus de sélection appliqué aux nœuds permet de disposer d'un système universel, contrairement aux seuils fixés auparavant empiriquement pour l'établissement des rôles communautaires.

1 Introduction

De nombreux systèmes du réel sont modélisés et étudiés sous forme de réseau. L'université de Koblenz (Kunegis (2013)) recense et compile ainsi par exemple des réseaux sociaux, biologiques, routiers, lexicaux, de collaboration, etc. Ces réseaux fournissent une source importante de données pour étudier les systèmes qu'ils modélisent. Pour mieux exploiter ces données, les chercheurs d'un grand nombre de domaines (physiciens, sociologues, informaticiens, etc) ont développé des outils théoriques capables d'explorer et de fouiller ces données structurées en réseaux (Newman (2003)). Par exemple, la notion de *centralité* est étudiée depuis les années cinquante et l'impact de la centralité d'un individu dans une organisation a été discuté entre autres dans le cadre de la perception du leadership, de l'efficacité organisationnelle, de la diffusion de l'innovation technologique (Freeman (1979)). Cette notion reste très utilisée pour l'étude des nœuds d'un réseau, de leur position, de leur importance ou de leur influence, nous en rappelons donc les définitions essentielles.

Une métrique de sélection de variables appliquée à la centralité et aux rôles communautaires.

Centralité. Soit un graphe $G = (V, E)$ avec V l'ensemble de ses sommets, et E l'ensemble des arêtes. On note $N(u) = \{v \in V : \{u, v\} \in E\}$ le *voisinage* du nœud u, i.e. l'ensemble des nœuds connectés à u dans G. Le degré d'un nœud u est noté $d(u) = |N(u)|$.

La centralité d'*intermédiarité* $C_b(u)$ d'un nœud u est définie en fonction du nombre de plus court chemins entre chaque paire de nœuds du graphe qui passent par u (Freeman (1979)) :

Définition 1 (Centralité d'intermédiarité)**.**

$$C_b(u) = \sum_{v,w \in V, v<w} \frac{\sigma_{vw}(u)}{\sigma_{vw}}$$

où σ_{vw} est le nombre total de plus court chemins entre les nœuds v et w, et $\sigma_{vw}(u)$ est le nombre de plus court chemins entre v et w qui passent par u.

La centralité de *proximité* $C_c(u)$ est définie en fonction de la distance du nœud au reste du graphe :

Définition 2 (Centralité de proximité)**.**

$$C_c(u) = \frac{1}{\sum_{v \in V} dist(u, v)}$$

où $dist(u, v)$ est la distance géodesique entre les nœuds u et v, i.e. la taille du plus court chemin entre ces nœuds.

Le *pagerank* (Brin et Page (2012)) d'un nœud est une variante de la centralité de *vecteur propre*, également très utilisée. Son calcul est basé sur l'idée que plus les voisins d'un nœud sont importants et nombreux, plus celui-ci est considéré comme important. Il est ainsi défini (ici pour un réseau orienté) en fonction du pagerank des voisins du nœud comme suit :

Définition 3 (Pagerank d'un nœud)**.** *Soit $G = (V, A)$ un graphe orienté avec A l'ensemble des arêtes orientées.*

$$PR(u) = (1 - d) + d \sum_{v \in N_u^-} \frac{PR(v)}{|N^+(v)|}$$

où d est un paramètre entre 0 et 1, N_u^- resp. N_u^+ désignent les voisins entrants resp. sortants de u, les nœuds v tels que $(v, u) \in A$ resp. $(u, v) \in A$.

Ces mesures, certes pertinentes, ne tiennent pas compte d'une réalité plus récente. En effet, avec l'avènement de l'internet mais aussi grâce à la rapide amélioration des moyens informatiques, les réseaux à étudier sont de plus en plus grands. Bien heureusement, il est démontré que pour la plupart, ces réseaux possèdent une structure *communautaire* (Newman et Girvan (2004)) qu'il est possible de mettre en évidence en utilisant des algorithmes dont les performances sont bien évaluées. Ceci permet ainsi d'étudier les réseaux non plus au niveau global, mais au niveau *mésoscopique*, et ainsi de pouvoir recueillir des informations en lien avec la communauté d'un nœud. Des mesures ont ainsi récemment été proposées pour apprécier la position, l'importance, la centralité d'un nœud dans sa communauté, mais également ses liens avec les communautés externes, et ainsi lui attribuer un *rôle communautaire*.

Rôles communautaires. Guimerà et Amaral (2005) définissent les rôles communautaires en s'inspirant de la notion d'équivalence structurelle proposée par Lorrain et White (1971). Pour cela, ils définissent deux mesures qui permettent d'évaluer la connectivité d'un nœud avec sa communauté d'une part (*degré intra-module*), et d'autre part avec le reste du réseau (*coefficient de participation*).

On considère C une partition de l'ensemble V des nœuds du réseau en communautés. Le *degré intra-module* évalue la connectivité d'un nœud à sa communauté relativement à celle des autres nœuds de sa communauté sous forme de z-score du degré interne du nœud :

Définition 4 (Degré intra-module)**.**

$$Z_i(u) = \frac{d_i(u) - \mu_i(d_i)}{\sigma_i(d_i)}$$

avec $u \in c_i \subset C$, où $d_i(u)$ est le degré de u dans la communauté c_i, i.e. la somme des arêtes (u, v) telles que $v \in c_i$; $\mu_i(d_i)$ et $\sigma_i(d_i)$ dénotent respectivement la moyenne et l'écart-type du degré des nœuds appartenant à la communauté c_i dans cette communauté.

Le coefficient de participation quantifie la connectivité d'un nœud avec les nœuds des communautés 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).

Définition 5 (Coefficient de participation)**.**

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

où $d(u)$ est le degré du nœud u et $d_i(u)$ son degré dans la communauté c_i.

Guimerà et Amaral calculent donc ces deux mesures pour chaque nœud du réseau. Ensuite, ils appliquent des seuils définis *empiriquement* pour proposer une classification du nœud en 7 rôles comme le montre le Tableau 1.

Degré intra-module		**Coefficient de participation**	
Hub	≥ 2.5	Provincial	≤ 0.30
		Connecteur	$]0.30; 0.75]$
		Ultra-connecteur	> 0.75
Non-Hub	< 2.5	Ultra-périphérique	≤ 0.05
		Périphérique	$]0.05; 0.62]$
		Connecteur	$]0.62; 0.80]$
		Ultra-connecteur	> 0.80

TAB. 1 – *Rôles communautaires et leurs seuils empiriques (Guimerà et Amaral (2005)).*

Leur approche leur permet, notamment sur des données biologiques, de mettre en évidence qu'il existe une correspondance entre les connectivités interne (dans sa communauté) et externe (vers les autres communautés) d'un nœud et son *rôle biologique*.

Néanmoins, ces mesures et les seuils *empiriques* que Guimerà et Amaral supposent universels ont été critiqués par Dugué et al. (2015), qui réimplémentent la notion de rôle communautaire pour l'étude du réseau issu des abonnements entre utilisateurs Twitter. Ils mettent ainsi en évidence que la distribution statistique du degré intra-module sur le réseau qu'ils étudient est sensiblement différente de celle constatée par Guimerà et Amaral, ce qui rend leur seuil inutilisable. Par ailleurs, ils montrent que les nœuds de degré très élevé sont très connectés aux communautés externes (forte participation aux communautés externes), mais de façon très *hétérogène*, ce qui aboutit à des valeurs du coefficient de participation relativement faible. Dugué et al. proposent ainsi de réviser l'évaluation de la connectivité externe en introduisant trois nouvelles mesures et de détecter les seuils entre rôles de façon automatique, par clustering des nœuds en utilisant les mesures comme attributs. Malgré des résultats encourageants, les seuils obtenus par clustering avec si peu d'attributs peuvent se révéler *peu stables* et les frontières entre rôles se trouver ainsi mal délimitées. Par ailleurs, la multiplication des mesures et le fait qu'elles soient exprimées sous forme de z-score ne simplifient pas leur interprétation. En effet, Klimm et al. (2014) montrent que des mesures exprimées ainsi ne permettent pas d'obtenir des chiffres comparables ni d'un réseau à un autre, ni d'une communauté à une autre. Ils proposent à la place des mesures basées sur la *densité* du réseau, mais sans fournir de seuils permettant de traduire ces mesures en rôles communautaires. Ceci constitue pourtant une étape cruciale qui permet de rendre possible la lecture des résultats sur des données très volumineuses : Dugué et al. (2015) étudient par exemple plus de 50 millions de nœuds.

Nos contributions. Nous proposons ainsi dans cet article, d'évaluer la possibilité de transposer une méthode de sélection de variables basée sur la *F-Mesure de trait* au problème de la détection des rôles communautaires dans des réseaux non orientés. Cette méthode a deux avantages majeurs. Premièrement, elle a été testée, évaluée et a montré son efficacité pour de nombreuses applications en classification et en clustering. Deuxièmement, elle est sans paramètres et propose un système de seuillage automatique stable. Ces deux atouts font en effet défaut à l'état de l'art actuellement. Nous commençons donc par introduire la F-Mesure de trait Section 2. Puis, Section 3, nous mettons en évidence sur des réseaux synthétiques réalistes (Lancichinetti et al. (2008)) les corrélations qui existent entre la F-Mesure de trait et les mesures classiques présentées dans cette introduction, telles que la centralité d'intermédiarité, le pagerank ou encore le degré intra-module et le coefficient de participation. Ceci montre la pertinence de la *F-Mesure de trait* dans le contexte des réseaux. Nous montrons également que cette approche est indépendante de la taille des communautés, peu corrélée au degré du nœud, et qu'elle produit des scores plus interprétables, et dont les valeurs sont comparables d'un réseau à un autre. Cette possibilité de comparer les valeurs d'un réseau à l'autre est particulièrement intéressante pour des applications telles que le suivi temporel des communautés d'un réseau comme le font Dugué et al. (2015) par une méthode diachronique.

2 F-mesure de Trait

La F-mesure de trait est définie pour évaluer l'importance d'une variable dans une classe ou dans un cluster. Elle a montré son intérêt pour la classification (Lamirel et al. (2015)), l'étiquetage de classes (Dugué et al. (2016)), le clustering (Lamirel et al. (2015)) ou encore la mesure de qualité d'un clustering (Lamirel et al. (2016)). Son calcul passe par la combinaison

sous forme de moyenne harmonique de la Prédominance de trait, qui apprécie la capacité d'un trait (une variable numérique en général) à décrire une classe (Éq. 1), et du Rappel de trait, qui évalue la typicité et la saillance d'un trait dans une classe (Éq. 2). On considère un ensemble de données D décrites par des traits F et une partition de ces données C.

$$FP_c(f) = \frac{W_c^f}{W_c} \quad (1) \qquad FR_c(f) = \frac{W_c^f}{W_D^f} \quad (2)$$

avec $c \subset C$, W_c^f la somme des valeurs du trait $f \in F$ des données attachées à c, W_c la somme des valeurs des traits des données affectées à la classe c, et W_D^f la somme des valeurs du trait f pour toutes les données D.

Définition 6 (F-Mesure de trait)**.**

$$FF_c(f) = 2\left(\frac{FR_c(f) \times FP_c(f)}{FR_c(f) + FP_c(f)}\right)$$

Dans les applications précédemment citées, elle est associée à un processus de sélection de variables où une variable f est sélectionnée (dans S_c) pour décrire la classe c si :

$$S_c = \left\{ f \in F_c \mid FF_c(f) > \overline{FF}(f) \text{ and } FF_c(f) > \overline{FF}_D \right\} \text{ avec} \quad (3)$$

$$\overline{FF}(f) = \Sigma_{c' \in C} \frac{FF_{c'(f)}}{|C_{/f}|} \text{ and } \overline{FF}_D = \Sigma_{f \in F} \frac{\overline{FF}(f)}{|F|} \quad (4)$$

où $C_{/f}$ représente le sous-ensemble de C où le trait f est présent (non nul).

Un exemple. Dans la Table 2, l'ensemble de données D est constitué de 6 individus décrits par 3 traits et qui appartiennent soit à la classe M des hommes, soit à la classe F des femmes.

Taille Pieds	**Longueur Cheveux**	**Taille Nez**	**Classe**
9	**5**	**5**	M
9	**10**	**5**	M
9	**20**	**6**	M
5	15	5	F
6	25	6	F
5	25	5	F

TAB. 2 – *Un exemple pour la F-Mesure de trait sur des données jouet.*

Les valeurs de Feature Predominance et de Feature Recall sont définies ainsi pour le trait *Taille des pieds* dans la classe *M* :

$$W_M^{TaillePieds} = \mathbf{27} \qquad FR_M(TaillePieds) = \frac{27}{43}$$

$$W_D^{TaillePieds} = 43$$

$$W_M = \mathbf{78} \qquad FP_M(TaillePieds) = \frac{27}{78}$$

Les valeurs de F-Mesure de trait finalement obtenues pour chaque classe sont décrites Table 3.

Une métrique de sélection de variables appliquée à la centralité et aux rôles communautaires.

Taille Pieds	Longueur Cheveux	Taille Nez	
0.46	0.39	0.3	$FF_M(f)$
0.22	0.66	0.24	$FF_F(f)$
0.34	0.53	0.27	$\overline{FF(f)}$

TAB. 3 – *Les valeurs de F-mesure de trait sur les données de la Table 2*

Adaptation aux réseaux non orientés. La transcription de la F-mesure de trait aux réseaux est immédiate en remplaçant la fréquence d'un trait dans une classe par le nombre de liens qu'un nœud possède avec les nœuds d'une communauté.

Définition 7 (Rappel de nœud)**.**

$$NR_i(u) = \frac{d_i(u)}{d(u)}$$

où $d_i(u)$ est le degré du nœud u dans la communauté c_i, i.e. la somme des arêtes (u, v) telles que $v \in c_i$.

Définition 8 (Prédominance de nœud)**.**

$$NP_i(u) = \frac{d_i(u)}{d_{c_i}}$$

où d_{c_i} vaut pour la somme des liens dans la communauté c_i.

La *Prédominance de nœud* est ainsi utilisée pour caractériser la connectivité du nœud dans sa communauté. Plus elle est élevée, plus le nœud est connecté à sa communauté. On remarque d'ailleurs qu'elle est similaire à l'*enchâssement* proposé par Lancichinetti et al. (2010) pour caractériser la structure interne des communautés. Le *Rappel de nœud* sert à évaluer la connectivité du nœud avec les nœuds extérieurs à sa communauté. Plus ce rappel est faible, plus le nœud est connecté avec l'extérieur.

Dans le cadre de la procédure de détection des rôles communautaires, la méthode de sélection de variables basée sur la *F-mesure de trait* permet notamment de détecter les hubs provinciaux, très ancrés dans leur communauté mais peu connectés avec l'extérieur. Néanmoins, il est aisé d'envisager d'en étendre le principe pour le reste des rôles communautaires.

3 La F-Mesure de trait appliquée à des réseaux synthétiques

3.1 Les réseaux du LFR

Pour évaluer la pertinence de la méthode de la F-Mesure de trait pour la détection des rôles communautaires, nous utilisons l'outil de génération de réseaux artificiels introduit par Lancichinetti et al. (2008) que nous appelons *LFR*. L'outil permet de faire varier des paramètres importants et ainsi de générer des réseaux artificiels *réalistes* avec une structure de communautés. Dans nos expérimentations, nous nous intéressons particulièrement à la connectivité des nœuds relativement à la structure de communauté. Pour cela, les paramètres topologiques (distributions de degré, taille, degré moyen) des réseaux générés sont fixés, mais nous faisons en

revanche varier la façon dont est structurée la communauté. Ainsi, l'exposant de la loi de puissance suivie par les degrés des nœuds du réseau est laissée à sa valeur par défaut 2 et celui de la loi de puissance suivie par la taille des communautés à détecter est fixée à 1, valeur usuelle. Nous générons des réseaux de 1000 nœuds avec un degré moyen de 15 et dans lequels le degré maximal est fixée à un tiers du nombre de nœuds, ce qui correspond aux valeurs observées sur les graphes de terrain (Kunegis (2013)).

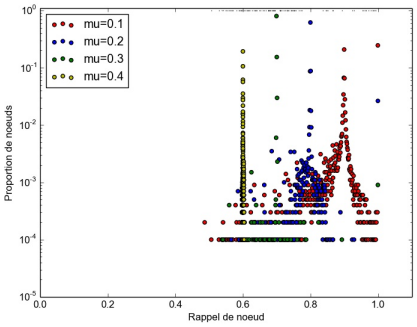

FIG. 1 – *Distribution du Rappel de nœuds en faisant varier mu.*

Nous faisons ensuite varier le paramètre de mélange μ qui sert à définir la netteté des communautés : quand μ est petit (resp. grand), les communautés sont faciles (resp. difficiles) à détecter. Nous générons 10 réseaux pour chaque valeur de μ : μ allant de 0.1 à 0.4, par pas de 0.1. Cela nous donne un total de 40 réseaux. La Figure 1, qui décrit la distribution du *Rappel de nœud* en fonction de μ sur ces réseaux nous permet de constater que pour $\mu = 0.4$, les réseaux générés ont une connectivité externe trop homogène, la valeur de Rappel de nœuds est quasiment constante. Les valeurs du *coefficient de participation* sont également très homogènes, ce biais n'est donc pas dû au Rappel de nœud, mais à la structure des réseaux générés. Ainsi, dans la suite, nous considérons uniquement les réseaux générés avec $0.1 \leq \mu \leq 0.3$, soit 30 réseaux au final.

3.2 Corrélation aux mesures de centralité

Sur les réseaux synthétiques décrits précédemment, nous nous intéressons dans un premier temps aux liens entre mesures de centralité définies globalement sur le graphe (sans considération pour la structure de communautés) et la *F-mesure de trait*.

Sur la Figure 2, on observe une corrélation entre la Prédominance de nœud, d'une part, et la centralité de degré et le pagerank, d'autre part. Pour obtenir cette corrélation, il est néanmoins nécessaire de multiplier la Prédominance de nœud par la taille de la communauté à laquelle le nœud appartient. Cela semble indiquer qu'un nœud fortement ancré dans une communauté de grande taille est particulièrement central et occupe une position importante dans le réseau.

Une métrique de sélection de variables appliquée à la centralité et aux rôles communautaires.

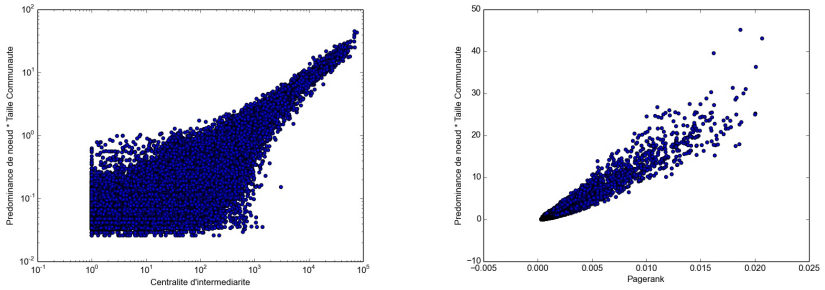

FIG. 2 – *À gauche, la Prédominance de nœud (Déf. 8) multipliée par la taille de communauté en fonction de la centralité d'intermédiarité (Déf. 1) à l'échelle logarithmique. À droite, idem en ordonnée mais le pagerank (Déf. 3) en abscisse.*

En revanche, on observe une quasi-indépendance entre la centralité de proximité (Déf. 2) et la Prédominance de nœud. De même, la valeur maximale du k du *k-core* [1] dans lequel le nœud se trouve semble indépendante de notre mesure. Cette valeur est importante pour l'étude de la propagation d'infection, ou de manière plus générale, pour celle de la diffusion d'information dans un réseau (Kitsak et al. (2010)). Le degré intra-module n'est pas non plus corrélé à ces mesures.

3.3 Corrélation avec les mesures de rôles communautaires

Après avoir montré les liens entre la *F-Mesure de trait* et les mesures de centralité, nous considérons maintenant le réseau à son niveau mésoscopique. Nous étudions les liens entre la *F-Mesure de trait* et les autres mesures destinées à évaluer la connectivité au regard de la structure communautaire, en particulier les mesures introduites par Guimerá et Amaral. Nous comparons donc, d'une part, la Prédominance de nœud (Déf. 8) au Degré intra-module (Déf. 4), qui évaluent la connectivité à l'intérieur de la communauté, et, d'autre part, le Rappel de nœud (Déf. 7) au Coefficient de participation (Déf. 5), qui évaluent la connectivité externe.

Degré intra-module et Prédominance de nœud. Tout d'abord, la Figure 3 met en évidence le fait que la *Prédominance de nœud* est plus indépendante du degré du nœud que le *Degré intra-module*. C'est notamment dû à la normalisation sous forme de z-score utilisée dans le *Degré intra-module*. La *Prédominance de nœud* semble donc plus pertinente pour mesurer la connectivité d'un nœud dans sa communauté.

Néanmoins, on remarque Figure 4 que la Prédominance de nœuds est corrélée au Degré intra-module lorsqu'elle est multipliée par la taille de la communauté. Ceci tend à montrer la dépendance qui existe entre le Degré intra-module et la taille de la communauté, dépendance confirmée par la Figure 5 et dûe à la normalisation par z-score.

1. Un sous-graphe $H = (C, E|C)$ induit par $C \subset V$ tel que le degré de tout $v \in C$ induit dans H est $\geq k$

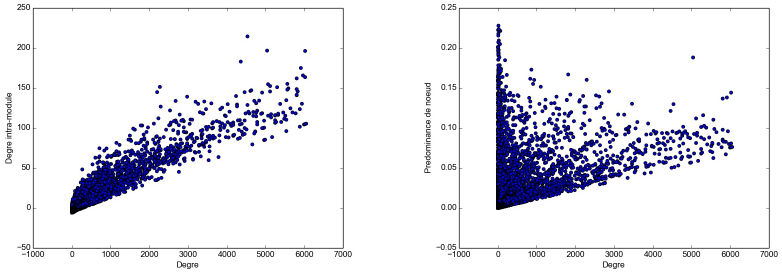

FIG. 3 – *À gauche, le Degré intra-module (Déf. 4) en fonction du degré. À droite, la Prédominance de nœud (Déf. 8) en fonction du degré.*

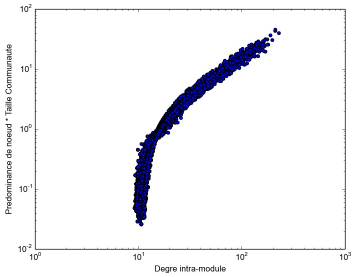

FIG. 4 – *La Prédominance de nœuds (Déf. 8) multipliée par la taille de la communauté en fonction du Degré intra-module (Déf. 4) à l'échelle logarithmique.*

On constate donc que la *Prédominance de nœud*, contrairement au degré intra-module est indépendante de la taille de la communauté dans lequel le nœud se trouve. En ce sens, elle rejoint le *Hub index* proposé par Klimm et al. (2014) pour caractériser la contribution d'un nœud à sa communauté. Celle-ci est conçue pour être absolue, et non relative à un graphe ou à une de ses communautés, ce qui rend possible de comparer des nœuds d'une communauté à une autre, ou d'un réseau à un autre. Pour cela, Klimm et al. (2014) proposent dans le *Hub index* (Équation 5) de comparer le degré du nœud à la distribution de degrés obtenue dans un réseau aléatoire équivalent au sous-graphe considéré (la communauté), i.e. avec le même nombre de nœuds et la même densité. La distribution de degrés d'un réseau aléatoire suit une loi de poisson : sa moyenne vaut $(N-1)\cdot\rho$ et son écart-type est la racine de sa moyenne (Barabási (2016)) avec ρ la densité du sous-graphe, N le nombre de nœuds du sous-graphe.

$$h(u) = \frac{d(u) - \mu(d_R)}{\sigma(d_R)} = \frac{d(u) - (N-1)\cdot\rho}{\sqrt{(N-1)\cdot\rho}} \tag{5}$$

avec d_R la distribution des degrés d'un graphe aléatoire équivalent au sous-graphe étudié (N nœuds, densité ρ). Or, comme pour le *Hub index*, notre mesure de *Prédominance de nœud*

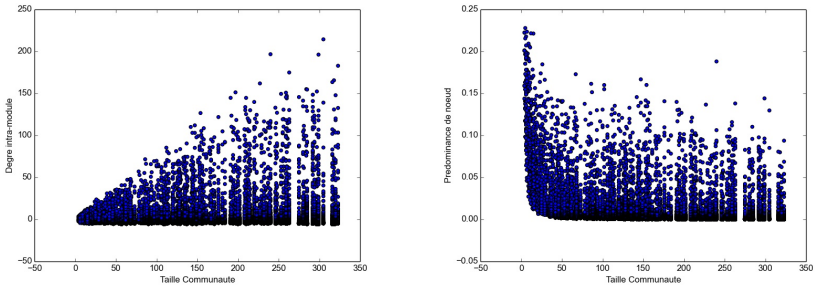

FIG. 5 – *Le Degré intra-module (à gauche) et la Prédominance de nœuds (à droite) en fonction de la taille des communautés.*

s'exprime en fonction de la densité puisque $d_{c_i} = \rho \cdot n \cdot (n-1)$. Il est donc possible de comparer les valeurs des nœuds d'un réseau à un autre, ce qui est important pour des applications telles que le suivi de l'évolution d'un réseau d'un point de vue communautaire : les correspondances entre communautés peuvent être détectées via les nœuds prédominants (Dugué et al. (2015)).

Coefficient de participation et Rappel de nœud. On observe Figure 6 une corrélation entre la métrique que nous proposons et celle introduite par Guimerá et Amaral. Celle-ci semble plus nette lorsque la structure de communautés est bien définie, i.e. lorsque μ est proche de 0. En effet, les deux mesures semblent diverger l'une de l'autre avec l'augmentation de μ.

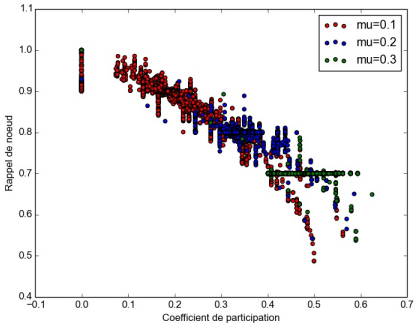

FIG. 6 – *Le Rappel de nœud (Déf. 7) en fonction du Coefficient de participation (Déf. 5).*

Ceci est probablement dû au fait que le *Coefficient de participation* encapsule plusieurs aspects de la connectivité externe : le nombre de liens vers l'extérieur mais également comment ceux-ci sont disséminés à travers les communautés. Ceci pose d'ailleurs problème dans le cadre de l'étude de grands réseaux où les nœuds de degré élevé ont tendance à être connectés à beaucoup de communautés externes, mais de façon hétérogène et donc à obtenir des scores

de participation faibles (Dugué et al. (2015)). Le *Rappel de nœud* fournit donc une alternative stable au *Coefficient de participation*, et se combine efficacement à la *Prédominance de nœud* et à la méthode de sélection de variables performante basée sur la *F-mesure de trait*.

4 Conclusion et perspectives

Dans cet article, nous montrons qu'il est pertinent d'envisager l'utilisation de la F-mesure de trait dans le cadre de l'étude des réseaux complexes, en particulier pour l'évaluation de la connectivité des nœuds au regard de la structure communautaire de ces réseaux. En effet, cette mesure montre des propriétés intéressantes : elle est corrélée à la centralité d'intermédiarité et au pagerank, mais également aux mesures introduites par Guimerà et Amaral pour la caractérisation des rôles communautaires. Par ailleurs, la mesure est par définition liée à la densité de la communauté dans laquelle elle est calculée. Cela la rend indépendante de la taille de la communauté, contrairement aux mesures de Guimerà et Amaral, ce qui permet des comparaisons de valeurs entre communautés, et même entre différents réseaux. Ceci ouvre des perspectives intéressantes comme l'utilisation de cette mesure pour la caractérisation de l'évolution des communautés d'un réseau dans le temps. De plus, la procédure de sélection de variables sans paramètres est extensible pour la détection des rôles communautaires, répondant ainsi partiellement au flou de l'état de l'art quant aux seuils à utiliser. Nous prévoyons donc à court terme d'adapter complètement cette procédure et de proposer ainsi tous les seuils nécessaires à une cartographie fonctionnelle des réseaux du réel. Par ailleurs, nous souhaitons valider l'efficacité de cette mesure pour le suivi diachronique de l'évolution des communautés de chercheurs dans le temps en étendant les travaux de Dugué et al. (2015). Nous pensons que les nœuds détectés par la méthode de sélection de variables associée à la F-mesure de trait sont particulièrement pertinents pour suivre l'évolution des communautés qui semblent se structurer et évoluer autour de ces nœuds. Enfin, il serait intéressant d'étudier l'utilisation de ces mesures dans le contexte des communautés recouvrantes qui constituent une réalité de terrain.

Références

Barabási, A.-L. (2016). *Network Science*. Cambridge University Press.

Brin, S. et L. Page (2012). Reprint of : The anatomy of a large-scale hypertextual web search engine. *Computer networks 56*(18), 3825–3833.

Dugué, N., V. Labatut, et A. Perez (2015). A community role approach to assess social capitalists visibility in the twitter network. *Social Network Analysis and Mining 5*(1), 1–13.

Dugué, N., J.-C. Lamirel, et P. Cuxac (2016). Keep track of your clusters ! In *Research Challenges in Information Science (RCIS)*.

Dugué, N., A. Tebbakh, P. Cuxac, et J.-C. Lamirel (2015). Feature selection and complex networks methods for an analysis of collaboration evolution in science : an application to the istex digital library. In *ISKO-MAGHREB 2015*.

Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social networks 1*(3), 215–239.

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

Kitsak, M., L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, et H. A. Makse (2010). Identification of influential spreaders in complex networks. *Nature Physics 6*(11), 888–893.

Klimm, F., J. Borge-Holthoefer, N. Wessel, J. Kurths, et G. Zamora-López (2014). Individual node's contribution to the mesoscale of complex networks. *New J. of Physics 16*(12).

Kunegis, J. (2013). Konect : the koblenz network collection. In *Proceedings of the 22nd International Conference on World Wide Web*, pp. 1343–1350. ACM.

Lamirel, J.-C., P. Cuxac, A. S. Chivukula, et K. Hajlaoui (2015). Optimizing text classification through efficient feature selection based on quality metric. *J. of I. IS 45*(3), 379–396.

Lamirel, J.-C., N. Dugué, et P. Cuxac (2016). New efficient clustering quality indexes. In *International Joint Conference on Neural Networks*.

Lamirel, J.-C., I. Falk, et C. Gardent (2015). Federating clustering and cluster labelling capabilities with a single approach based on feature maximization : French verb classes identification with igngf neural clustering. *Neurocomputing 147*, 136–146.

Lancichinetti, A., S. Fortunato, et F. Radicchi (2008). Benchmark graphs for testing community detection algorithms. *Physical review E 78*(4), 046110.

Lancichinetti, A., M. Kivelä, J. Saramäki, et S. Fortunato (2010). Characterizing the community structure of complex networks. *PloS one 5*(8), e11976.

Lorrain, F. et H. C. White (1971). Structural equivalence of individuals in social networks. *The Journal of mathematical sociology 1*(1), 49–80.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review 45*(2), 167–256.

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

Summary

The Feature F-measure is a statistical and parameter-free metric used in feature selection that performs well for classification, clustering, cluster labeling, also used to evaluate cluster quality. We evaluate its use in the complex networks framework. We are especially interested in evaluating its use to characterize the node connectivity regarding the community structure. This would allow to benefit from its parameter-free system of feature selection, and of its well-evaluated performance. We thus study on a benchmark of realistic synthetic graphs the correlations between Feature F-measure and classic centrality measures, but also between Feature F-measure and measures designed to characterize community roles of nodes. We show that Feature F-Measure is linked to node centrality, and that it is well-fitted to evaluate their connectivity w.r.t. the community structure. We also observe that community roles detection measures are dependent of the community size, whereas Feature F-Measure is tied to density, which makes results comparable from a network to another. This allows to consider using Feature F-Measure to study dynamic temporal network, using it for community matching.