Vers une modularité pour données vectorielles

David Combe, Christine Largeron, Elöd Egyed-Zsigmond, Mathias Géry

Vers une modularité pour données vectorielles

David Combe*, Christine Largeron*
Előd Egyed-Zsigmond**, Mathias Géry*

*Université de Lyon, F-42023, Saint-Étienne, France,
CNRS, UMR 5516, Laboratoire Hubert Curien, F-42000, Saint-Étienne, France
Université de Saint-Étienne, Jean-Monnet, F-42000, Saint-Étienne, France
{david.combe, christine.largeron, mathias.gery}@univ-st-etienne.fr
**Université de Lyon
UMR 5205 CNRS, LIRIS
7 av J. Capelle, F 69100 Villeurbanne, France
elod.egyed-zsigmond@insa-lyon.fr

Résumé. La modularité, introduite par Newman pour mesurer la qualité d'une partition des sommets d'un graphe, ne prend pas en compte d'éventuelles valeurs associées à ces sommets. Dans cet article, nous introduisons une mesure de modularité complémentaire, basée sur l'inertie, et adaptée pour évaluer la qualité d'une partition d'éléments représentés dans un espace vectoriel réel. Cette mesure se veut un pendant pour la classification non supervisée de la modularité de Newman. Nous présentons également 2Mod-Louvain, une méthode utilisant ce critère de modularité basée sur l'inertie conjointement à la modularité de Newman pour détecter des communautés dans des réseaux d'information. Les expérimentations que nous avons menées ont montré qu'en exploitant à la fois les données relationnelles et vectorielles, 2Mod-Louvain détectait plus efficacement les communautés que des méthodes utilisant un seul type de données et qu'elle était robuste face à des dégradations des données.

1 Introduction

Dans le cas de la classification automatique comme pour la détection de communautés, après avoir construit une partition à l'aide d'une méthode, il convient d'évaluer sa qualité. Pour ce faire, on peut faire appel à des critères externes ou internes. Les premiers permettent de comparer le résultat obtenu avec un résultat attendu, par exemple une partition faisant office de "vérité terrain" alors que les seconds quantifient la qualité de la partition proposée. Parmi les critères externes, utilisables aussi bien en classification automatique que pour la détection de communautés, on peut citer le taux de bien classés, la pureté, l'indice de Rand ou sa version ajustée, l'entropie ou encore l'information mutuelle, éventuellement normalisée ou ajustée, mais aussi la V-mesure, l'homogénéité ou la complétude (Hubert et Arabie (1985); Vinh et al. (2010); Rosenberg et Hirschberg (2007)).

En classification automatique, les critères internes peuvent eux-mêmes être subdivisés en critères dont l'usage est spécifique à une distance ou à une méthode, comme par exemple l'iner-

tie intra ou interclasses, et en critères non spécifiques, comme les indices de Dunn, de Davies et Bouldin ou de Silhouette (Rousseeuw (1987); Dunn (1973); Davies et Bouldin (1979)) Dans le domaine de la détection de communautés, on pourra citer la couverture, la conductance, la performance ou encore le coefficient de clustering mais le critère le plus couramment employé est la modularité (Yang et Leskovec (2012)).

Introduite par Newman pour juger de la qualité d'un partitionnement des sommets d'un graphe, la modularité a ensuite été utilisée directement pour identifier des classes telles que les sommets à l'intérieur d'une classe soient fortement reliés et qu'ils aient peu de relations avec ceux des autres classes (Newman (2006)). Bien que des travaux récents aient souligné des problèmes liés à l'optimisation de ce critère pour déterminer un partitionnement optimal notamment, la limite de résolution rendant difficile la détection de classes de faible taille ou dans des graphes creux ou encore l'existence de partitions à forte modularité dans des graphes sans structure communautaire, c'est un critère qui a néanmoins produit de bons résultats dans la pratique, notamment pour la détection de communautés dans un réseau social (Guimera et al. (2004); Good et al. (2010); Lancichinetti et Fortunato (2011)). De plus, elle présente des propriétés intéressantes. Premièrement, elle est calculable sur des graphes valués ou non valués, et ne nécessite pas de normalisation préalable. Ensuite, elle repose sur des concepts intelligibles, où on cherche à former des classes entre sommets mieux reliés entre eux que dans une formation aléatoire. Enfin, contrairement à d'autres critères, la modularité permet de comparer des partitions n'ayant pas nécessairement le même nombre de classes. Cependant, la modularité ne peut pas être utilisée pour évaluer la qualité d'un partitionnement dans un contexte de classification automatique et, à notre connaissance, il n'existe pas de critère ayant les propriétés de la modularité de Newman qui soit adapté à des éléments décrits par des attributs vectoriels. C'est la raison pour laquelle, dans cet article, nous définissons un critère de mesure de la qualité d'une partition d'éléments représentés par des vecteurs, inspirée de la modularité et qui pourra être utilisée pour comparer deux partitions. Ce critère sera décrit dans la section suivante. La section 3 est consacrée à l'adaptation de ce nouveau critère à l'heuristique de la méthode de Louvain. Nous proposons une nouvelle méthode de détection de communautés dans un réseau d'information appelée 2Mod-Louvain, basée sur l'optimisation en parallèle de la modularité de Newman et de la modularité que nous introduisons. Enfin, dans la section 4, à travers des expérimentations, nous évaluons les performances de cette méthode et sa robustesse à des dégradations des données.

2 Critère de modularité basée sur l'inertie

On considère V, un ensemble d'éléments représentés dans un espace vectoriel plongé dans $\mathbb{R}^{|T|}$. Chaque élément $v \in V$ est décrit par un vecteur d'attributs qui, par souci de simplification des notations, est aussi noté v :

$$v = (v_1, \ldots, v_{|T|}) \tag{1}$$

On suppose de plus qu'une masse m_v égale à 1 est associée à chaque élément v de V. La somme de ces masses est égale à N, le nombre d'éléments de V.

$I(V)$ désigne l'inertie de V par rapport à son centre de gravité g ou simplement comme inertie interne de V ou moment centré d'ordre 2 et défini de la façon suivante :

$$I(V) = \sum_{v \in V} m_v \left\| v - g \right\|^2 \tag{2}$$

L'inertie $I(V, v)$ de V par rapport à un élément v est la somme des carrés des distances entre les éléments de V et v.

$$I(V, v) = \sum_{v' \in V} m_{v'} \left\| v' - v \right\|^2 \tag{3}$$

Étant donnée une partition $\mathscr{P} = \{C_1, \ldots, C_r\}$ en r classes disjointes de V, $Q_{inertie}(\mathscr{P})$, le critère de mesure de la qualité de la partition $\mathscr{P}$, que nous introduisons, est défini par :

$$Q_{inertie}(\mathscr{P}) = \sum_{(v,v') \in V \cdot V} \left[\left(\frac{I(V,v) \cdot I(V,v')}{(2N \cdot I(V))^2} - \frac{\left\| v - v' \right\|^2}{2N \cdot I(V)} \right) \cdot \delta \left(c_v, c_{v'} \right) \right] \tag{4}$$

où $c(v)$ est la classe de l'élément v et δ est la fonction de Kronecker qui vaut 1 si ses arguments sont égaux et 0 sinon.

Ainsi, alors que la modularité considère la force du lien et vise à regrouper les éléments les plus fortement liés, notre critère exploite la norme entre les éléments et vise à regrouper ceux qui sont les moins dissemblables ; ce qui apparait, après normalisation, dans le second terme de l'équation :

$$\frac{\left\| v - v' \right\|^2}{2N \cdot I(V)} \tag{5}$$

Le critère $Q_{inertie}$ compare la norme de chaque paire d'éléments (v, v') issus d'une même communauté (contrôlé par la fonction de Kronecker), avec une valeur attendue $d^2(v, v')$ définie par :

$$d^2(v, v') = I(V, v) \cdot I(V, v') \tag{6}$$

Il s'agit donc de comparer une fonction du carré de la distance entre v et v', à une fonction du carré des distances de chacun de ces éléments v et v' aux autres éléments de V. Si la valeur attendue est plus grande que la valeur réelle, alors les deux éléments sont de bons candidats à appartenir à une même classe. C'est la raison pour laquelle, dans le critère global $Q_{inertie}(\mathscr{P})$, la valeur observée est soustraite de la valeur attendue alors que dans le cas de la modularité c'est la valeur attendue qui est retranchée de la force du lien observé.

Le critère $Q_{inertie}(\mathscr{P})$ que nous proposons varie entre -1 et 1. En effet, la partie gauche de la soustraction (6), comprenant les produits d'inerties pour toutes les paires de sommets, vaudra au plus 1. De même, la partie droite du critère $Q_{inertie}(\mathscr{P})$ (équation 5) ne pourra pas dépasser 1. Les deux parties étant strictement positives, le critère, contraint par les valeurs de la fonction de Kronecker, varie entre -1 et 1.

Ce critère présente plusieurs propriétés intéressantes. Premièrement, il conserve la même valeur, quelle que soit la transformation affine que l'on applique aux attributs, autrement dit l'ajout d'une constante et/ou la multiplication par un scalaire des vecteurs associés aux éléments à classer n'a pas d'incidence sur la valeur du critère. Enfin l'ordre des attributs n'a aucune incidence sur le résultat.

En revanche, ce critère présente aussi certaines limites. Il est indéfini si les vecteurs numériques sont identiques, car l'inertie totale est alors nulle. Ceci n'est pas réellement un inconvénient lors de la détection de communautés dans un réseau d'information car dans ce cas, les attributs n'apportant aucune information, la détection des communautés sera basée uniquement sur les données relationnelles. De plus, comme la modularité de Newman et Girvan, on peut s'attendre à ce que ce critère présente une limite de résolution. Il convient donc de s'interroger sur la façon d'y remédier. Une adaptation des travaux d'Arenas *et al.* et Reichardt *et al.* visant à pallier cet effet pourrait être envisagée, via l'introduction d'un paramètre permettant d'ajuster le comportement du critère (Arenas et al. (2008); Reichardt et Bornholdt (2006)).

3 Méthode 2Mod-Louvain

Comme nous l'avons indiqué en introduction, une des applications immédiates du critère $Q_{inertie}$ est la détection de communautés dans un réseau d'information représenté par un graphe avec attributs $G = (V, E)$ où V est l'ensemble des sommets, $E \subset V \times V$ est l'ensemble des arêtes et où chaque sommet $v \in V$ est associé à un vecteur $v = (v_1, \ldots, v_j, \ldots, v_T)$ à valeurs réelles (Zhou et al. (2009)).

Dans cette section, en tirant parti du critère de modularité basée sur l'inertie $Q_{inertie}$, nous proposons une méthode, appelée 2Mod-Louvain, dédiée à la détection de communautés dans ce type de réseaux. Cette méthode basée sur le principe d'exploration de la méthode de Louvain, exploite conjointement le critère $Q_{inertie}$ et la modularité de Newman $Q_{NG}(\mathscr{P})$ puisqu'elle consiste à optimiser le critère global $QQ^{+}(\mathscr{P})$ défini par :

$$QQ^{+}(\mathscr{P}) = Q_{NG}(\mathscr{P}) + Q_{inertie}(\mathscr{P}) \tag{7}$$

avec :

$$Q_{NG}(\mathscr{P}) = \frac{1}{2m}\Sigma_{vv'}\left[\left(A_{vv'} - \frac{k_v \cdot k_{v'}}{2m}\right)\delta(c_v, c_{v'})\right] \tag{8}$$

où (v, v') prend toutes les valeurs de $V \times V$, k_v est le degré du sommet v, A désigne la matrice d'adjacence associée à G, m est la somme des poids de toutes les arêtes du graphe et δ est la fonction de Kronecker. Il n'est pas utile de normaliser les deux critères $Q_{NG}(\mathscr{P})$ et $Q_{inertie}(\mathscr{P})$ car leurs bornes sont identiques comme indiqué dans la section précédente.

La méthode 2Mod-Louvain est détaillée dans l'algorithme 1, qui comporte deux étapes. La première est une phase itérative qui vise à déplacer un sommet de sa classe vers celle d'un de ses voisins dans le graphe si ce changement induit un gain de la modularité globale $QQ^{+}(\mathscr{P})$. La seconde est une phase de fusion qui consiste à construire un nouveau graphe dont les sommets correspondent aux communautés obtenues à l'issue de la phase précédente. Cette seconde phase fait intervenir deux procédures *Fusion_Matrice_Adjacence* et *Fusion_Matrice_Inertie*. La procédure *Fusion_Matrice_Adjacence* est identique à celle mise en œuvre dans la méthode de Louvain (Aynaud et al. (2010)). La procédure *Fusion_Matrice_Inertie* est décrite dans la section suivante.

Algorithme 1 : 2Mod-Louvain

Entrées : Un réseau d'information G_0
Sorties : Une partition $\mathscr{P}_{res}$

$\mathscr{P} \leftarrow$ partition discrète des sommets de V_0;
$\mathcal{A} \leftarrow$ matrice d'adjacence de G_0;
$\mathcal{D} \leftarrow$ matrice des carrés des distances euclidiennes entre les sommets de V_0 calculées sur leurs attributs ;
$G \leftarrow G_0$;
répéter
    fin $\leftarrow$ faux;
    QQ^+_antérieur $\leftarrow QQ^+(\mathscr{P})$;
    tant que *il y a des sommets déplacés* **faire**
        pour tous les *sommet* u *de* G **faire**
            $B \leftarrow$ communauté voisine maximisant le gain de QQ^+;
            si *le placement de* u *dans* B *induit un gain strictement positif* **alors**
                Placer u dans la communauté B;
                Mettre à jour la partition $\mathscr{P}$ suite au transfert de u dans B;
    si $QQ^+(\mathscr{P}) > QQ^+$*_antérieur* **alors**
        $G, \mathcal{A} \leftarrow$ Fusion_Matrice_Adjacence($\mathcal{A}, \mathscr{P}$) ;
        $\mathcal{D} \leftarrow$ Fusion_Matrice_Inertie($\mathcal{D}, \mathscr{P}$) ;
    sinon
        fin $\leftarrow$ vrai ;
jusqu'à *fin* ;
$\mathscr{P}_{res} \leftarrow \mathscr{P}$ **partition des sommets de** V_0 ;

3.1 Synthèse des informations de distance dans la deuxième phase

Si le graphe G considéré au début de la phase itérative comporte $|V|$ sommets alors la matrice $\mathcal{D}$ est une matrice carrée symétrique de taille $|V| \times |V|$ dont chaque terme $\mathcal{D}[a,b]$ correspond au carré des distances entre les vecteurs descriptifs des sommets v_a et v_b de V. A l'issue de la phase itérative, on obtient une partition $\mathscr{P}'$ de V en k communautés, dont chaque classe va correspondre à un sommet de V' dans le nouveau graphe G'. La matrice $\mathcal{D}'$ associée au graphe G' sera définie par :

$$\mathcal{D}'[x,y] = \sum_{(v_a,v_b)\in V\times V} \mathcal{D}[v_a,v_b] \cdot \delta(\tau(v_a),x) \cdot \delta(\tau(v_b),y) \tag{9}$$

où la fonction τ indique pour chaque sommet v de V par quel sommet v', correspondant à sa classe d'affectation, il est représenté dans V'.

3.2 Optimisation de l'algorithme durant la phase itérative par calcul incrémental du gain de modularité

Un des avantages de la méthode de Louvain est de limiter les calculs à ceux nécessaires pour connaître la classe dans laquelle il est le plus avantageux d'affecter le sommet étudié (Aynaud et al. (2010)). De même, dans la méthode 2Mod-Louvain, le calcul du gain de modularité basée sur l'inertie peut être limité au calcul du gain induit par le déplacement d'un sommet de sa classe vers celle d'un de ses voisins. Nous détaillons ci-après les optimisations pouvant être opérées par ce calcul local de la modularité basée sur l'inertie et par conséquent du critère global QQ^{+}.

Considérons deux partitions, $\mathscr{P}$ la partition d'origine et $\mathscr{P}'$ la partition induite par un transfert d'un sommet u de sa classe d'origine A vers sa classe d'affectation B.

$$\mathscr{P} = (A, B, C_1, .., C_r) \tag{10}$$

$$\mathscr{P}' = (A \setminus \{u\}, B \cup \{u\}, C_1, .., C_r) \tag{11}$$

Par la suite, $A \setminus \{u\}$ désigne la classe A privée du sommet u et $B \cup \{u\}$ la classe B augmentée du sommet u. Dans un souci de simplification des notations dans la suite nous notons le terme $D[v, v']$ de la matrice abusivement $\mathcal{D}_{vv'}$. La modularité basée sur l'inertie de la partition $\mathscr{P}$ vaut :

$$Q_{inertie}(\mathscr{P}) = \sum_{C \in \mathscr{P}} \frac{1}{2N \cdot I(V)} \sum_{v,v' \in C} \left[\frac{I(V,v) \cdot I(V,v')}{2N \cdot I(V)} - \mathcal{D}_{vv'} \right] \tag{12}$$

$$\begin{aligned}
=& \frac{1}{2N \cdot I(V)} \sum_{v,v' \in A} \left[\frac{I(V,v) \cdot I(V,v')}{2N \cdot I(V)} - \mathcal{D}_{vv'} \right] \\
&+ \frac{1}{2N \cdot I(V)} \sum_{v,v' \in B} \left[\frac{I(V,v) \cdot I(V,v')}{2N \cdot I(V)} - \mathcal{D}_{vv'} \right] \\
&+ \frac{1}{2N \cdot I(V)} \sum_{C \neq A,B} \sum_{v,v' \in C} \left[\frac{I(V,v) \cdot I(V,v')}{2N \cdot I(V)} - \mathcal{D}_{vv'} \right]
\end{aligned} \tag{13}$$

$$\begin{aligned}
=& \frac{1}{2N \cdot I(V)} \sum_{v,v' \in A \setminus \{u\}} \left[\frac{I(V,v) \cdot I(V,v')}{2N \cdot I(V)} - \mathcal{D}_{vv'} \right] \\
&+ \frac{1}{2N \cdot I(V)} \left[\frac{I(V,u)^2}{2N \cdot I(V)} - \mathcal{D}_{uu} \right] \\
&+ \frac{1}{N \cdot I(V)} \sum_{v \in A \setminus \{u\}} \left[\frac{I(V,u) \cdot I(V,v)}{2N \cdot I(V)} - \mathcal{D}_{uv} \right] \\
&+ \frac{1}{2N \cdot I(V)} \sum_{v,v' \in B} \left[\frac{I(V,v) \cdot I(V,v')}{2N \cdot I(V)} - \mathcal{D}_{vv'} \right] \\
&+ \frac{1}{2N \cdot I(V)} \sum_{C \neq A,B} \sum_{v,v' \in C} \left[\frac{I(V,v) \cdot I(V,v')}{2N \cdot I(V)} - \mathcal{D}_{vv'} \right]
\end{aligned} \tag{14}$$

La modularité de la partition $\mathscr{P}'$ vaut quant à elle :

$$Q_{inertie}(\mathscr{P}') = \sum_{C\in\mathscr{P}} \frac{1}{2N\cdot I(V)} \sum_{v,v'\in C} \left[\frac{I(V,v)\cdot I(V,v')}{2N\cdot I(V)} - \mathcal{D}_{vv'}\right] \tag{15}$$

$$\begin{aligned}
=&\frac{1}{2N\cdot I(V)} \sum_{v,v'\in A\setminus u} \left[\frac{I(V,v)\cdot I(V,v')}{2N\cdot I(V)} - \mathcal{D}_{vv'}\right] \\
&+ \frac{1}{2N\cdot I(V)} \sum_{v,v'\in B\cup u} \left[\frac{I(V,v)\cdot I(V,v')}{2N\cdot I(V)} - \mathcal{D}_{vv'}\right] \\
&+ \frac{1}{2N\cdot I(V)} \sum_{C\neq A\setminus\{u\}, B\cup\{u\}} \sum_{v,v'\in C} \left[\frac{I(V,v)\cdot I(V,v')}{2N\cdot I(V)} - \mathcal{D}_{vv'}\right]
\end{aligned} \tag{16}$$

$$\begin{aligned}
=&\frac{1}{2N\cdot I(V)} \sum_{v,v'\in A\setminus u} \left[\frac{I(V,v)\cdot I(V,v')}{2N\cdot I(V)} - \mathcal{D}_{vv'}\right] \\
&+ \frac{1}{2N\cdot I(V)} \sum_{v,v'\in B} \left[\frac{I(V,v)\cdot I(V,v')}{2N\cdot I(V)} - \mathcal{D}_{vv'}\right] \\
&+ \frac{1}{2N\cdot I(V)} \left[\frac{I(V,u)^2}{2N\cdot I(V)} - \mathcal{D}(u,u)\right] \\
&+ \frac{1}{N\cdot I(V)} \sum_{v\in B} \left[\frac{I(V,u)\cdot I(V,v)}{2N\cdot I(V)} - \mathcal{D}_{uv}\right] \\
&+ \frac{1}{2N\cdot I(V)} \sum_{C\neq A\setminus\{u\}, B\cup\{u\}} \sum_{v,v'\in C} \left[\frac{I(V,v)\cdot I(V,v')}{2N\cdot I(V)} - \mathcal{D}_{vv'}\right]
\end{aligned} \tag{17}$$

Le gain de modularité lors du passage de $\mathscr{P}$ à $\mathscr{P}'$ a donc pour valeur :

$$\begin{aligned}
\Delta Q_{inertie} =& Q_{inertie}(\mathscr{P}') - Q_{inertie}(\mathscr{P}) \\
=&\frac{1}{2N\cdot I(V)} \sum_{v,v'\in A\setminus\{u\}} \left[\frac{I(V,v)\cdot I(V,v')}{2N} - \mathcal{D}_{vv'}\right] \\
&+ \frac{1}{N\cdot I(V)} \sum_{v\in B} \left[\frac{I(V,u)\cdot I(V,v)}{2N\cdot I(V)} - \mathcal{D}_{uv}\right] \\
&+ \frac{1}{2N\cdot I(V)} \sum_{v,v'\in B} \left[\frac{I(V,v)\cdot I(V,v')}{2N\cdot I(V)} - \mathcal{D}_{vv'}\right] \\
&+ \frac{1}{2N\cdot I(V)} \left[\frac{I(V,u)^2}{2N\cdot I(V)} - \mathcal{D}_{uu}\right] \\
&+ \frac{1}{2N\cdot I(V)} \sum_{C\neq A\setminus\{u\}, B\cup\{u\}} \sum_{v,v'\in C} \left[\frac{I(V,v)\cdot I(V,v')}{2N\cdot I(V)} - \mathcal{D}_{vv'}\right]
\end{aligned} \tag{18}$$

$$
\begin{aligned}
&- \left[\frac{1}{2N \cdot I(V)} \sum_{v,v' \in A \setminus \{u\}} \left[\frac{I(V,v) \cdot I(V,v')}{2N \cdot I(V)} - \mathcal{D}_{vv'} \right] \right. \\
&+ \frac{1}{2N \cdot I(V)} \sum_{v,v' \in B} \left[\frac{I(V,v) \cdot I(V,v')}{2N \cdot I(V)} - \mathcal{D}_{vv'} \right] \\
&+ \frac{1}{2N \cdot I(V)} \left[\frac{I(V,u)^2}{2N \cdot I(V)} - \mathcal{D}_{uu} \right] \\
&+ \frac{1}{N \cdot I(V)} \sum_{v \in B} \left[\frac{I(V,u) \cdot I(V,v)}{2N \cdot I(V)} - \mathcal{D}_{uv} \right] \\
&\left. + \frac{1}{2N \cdot I(V)} \sum_{C \neq A \setminus \{u\}, B \cup \{u\}} \sum_{v,v' \in C} \left[\frac{I(V,v) \cdot I(V,v')}{2N \cdot I(V)} - \mathcal{D}_{vv'} \right] \right] \qquad (19) \\
=&\frac{1}{N \cdot I(V)} \sum_{v \in B} \left[\frac{I(V,u) \cdot I(V,v)}{2N \cdot I(V)} - \mathcal{D}_{uv} \right] \\
&- \frac{1}{N \cdot I(V)} \sum_{v \in A \setminus \{u\}} \left[\frac{I(V,u) \cdot I(V,v)}{2N \cdot I(V)} - \mathcal{D}_{uv} \right] \qquad (20)
\end{aligned}
$$

De plus, on peut remarquer que la variation de modularité induite par la suppression de u de sa classe d'origine sera la même quelle que soit sa classe d'affectation. Par conséquent le calcul de variation de modularité peut être effectué en considérant uniquement la différence induite par l'insertion de u dans sa nouvelle communauté d'affectation, décrite par le premier terme de l'équation 20.

Ces calculs nous permettent de montrer que notre critère bénéficie lui aussi de la possibilité d'être calculé de façon incrémentale. Le gain de modularité basée sur l'inertie repose uniquement sur des informations locales relatives au sommet déplacé et à sa distance avec les autres sommets.

4 Évaluation de la méthode *2Mod-Louvain* sur des réseaux artificiels

Dans cette section, nous nous proposons d'évaluer la méthode 2Mod-Louvain qui optimise le critère global QQ^+ basé à la fois sur la modularité de Newman et la modularité par rapport à l'inertie. Pour ce faire, nous étudions sa robustesse sur des réseaux artificiels vis-à-vis d'une dégradation de la structure de communautés définie par rapport aux relations, ou des classes définies par rapport aux attributs, ou encore d'une augmentation de la taille du réseau d'information ou d'une variation de la densité des liens.

On construit un réseau de référence R comportant 168 liens et 99 sommets uniformément répartis entre 3 catégories à l'aide du modèle proposé par Dang en fixant à 2 le nombre d'arêtes introduites avec chaque nouveau sommet (Dang (2012)). De plus, chaque sommet est décrit par une valeur d'attribut réelle qui suit une loi normale d'écart-type 7, centrée autour d'une valeur propre à sa classe d'origine. Ainsi la première classe a un centre $\mu_1 = 10$, la deuxième

un centre $\mu_2 = 40$ et la troisième un centre $\mu_3 = 70$. La classe d'origine du sommet sert de vérité terrain pour l'évaluation. Le tableau 1 montre la répartition des arêtes entre les classes dans le graphe R.

	Classe 1	Classe 2	Classe 3
Classe 1	55		
Classe 2	2	53	
Classe 3	1	7	50

TAB. 1 – *Répartition des extrémités des liens du graphe R*

À partir de ce réseau de référence R, nous avons construit quatre familles de réseaux :

– R.1.x dans lesquels l'information relationnelle est dégradée par rapport à R, en remplaçant un certain pourcentage p d'arêtes intraclasses par des arêtes interclasses avec $p = 0.25$ pour R1.1 et $p = 0.5$ pour R1.2.
– R.2.x dans lesquels les valeurs des attributs sont moins représentatives de chaque classe, en passant l'écart de 7 à 10 pour le réseau R2.1 et à 12 pour R2.2
– R.3.x qui comportent plus de sommets que R, en considérant 999 sommets pour R3.1 et 5001 sommets pour R3.2
– R.4.x qui comportent plus d'arêtes que R en passant respectivement à 5 le nombre d'arêtes introduites par nouveau sommet dans R4.1 et à 10 dans R4.2

Les résultats de 2Mod-Louvain sont comparés à ceux de la méthode de Louvain et des K-means en fixant le paramètre à 3 ainsi qu'à ToTeM, une autre méthode de détection de communautés adaptée aux graphes avec attributs (Combe et al. (2013)). Les tableaux 2 et 3 présentent respectivement les taux de bien classés et d'information mutuelle normalisée (NMI).

	Louvain		K-means	ToTeM		2Mod-Louvain	
	TBC	#cl.	TBC	TBC	#cl.	TBC	#cl.
Graphe de référence							
R	84%	4	96%	97%	3	**98%**	3
Dégradation de l'information relationnelle							
$degr_{rel} = 0,25$	33%	8	N/A*	18%	30	**78%**	5
$degr_{rel} = 0,5$	23%	9	N/A*	14%	36	**63%**	6
Étalement des distributions							
$\sigma = 10$	N/A*		90%	95%	3	**96%**	3
$\sigma = 12$	N/A*		87%	20%	26	**98%**	3
Augmentation de la taille du réseau							
$\|V\| = 999$	50%	11	**97%**	**97%**	3	84%	4
$\|V\| = 5001$	40%	12	**98%**	0,5%	1 518	85%	4
Augmentation du nombre d'arêtes							
$\|E\| = 315$	**96%**	3	N/A*	95%	3	94%	3
$\|E\| = 508$	97%	3	N/A*	**98%**	3	**98%**	3

* La dégradation de l'information relationnelle et le changement de densité n'influencent pas les résultats des K-means ; la dégradation de l'information des attributs n'influence pas les résultats de la méthode de Louvain.

TAB. 2 – *Évaluation selon le taux de bien classés*

NMI	Louvain	K-means	ToTeM	2mod-Louvain
Graphe de référence				
R	0,784	0,883	0,861	**0,930**
Dégradation de l'information relationnelle				
R.1.1	0,220		0,489	**0,603**
R.1.2	0,118		**0,377**	0,353
Dégradation des attributs				
R.2.1		0,721	0,819	**0,885**
R.2.2		0,637	0,567	**0,930**
Augmentation de la taille du réseau				
R.3.1	0,597	**0,880**	0,854	0,800
R.3.2	0,586	**0,892**	0,376	0,774
Augmentation du nombre d'arêtes				
R.4.1	**0,848**		0,807	0,816
R.4.2	0,876		**0,917**	**0,917**

TAB. 3 – *Évaluation selon la NMI*

En exploitant l'information vectorielle en plus de l'information relationnelle, la méthode 2Mod-Louvain gagne en robustesse par rapport à la méthode de Louvain face à une dégradation de l'information relationnelle et par rapport à la méthode des K-Means en cas de dégradation de l'information vectorielle. De plus, lorsque la taille du réseau augmente, la méthode proposée permet de parer à la multiplication des classes qui survient alors avec la méthode de Louvain (4 classes contre 10). Les K-means conservent de bons résultats dans le cas où la taille du réseau évolue, car l'information des attributs demeure de bonne qualité mais ils sont globalement avantagés par rapport aux autres méthodes du fait que le nombre de classe est fourni en paramètre. Ainsi, l'utilisation simultanée des deux types d'information à travers des mesures de modularité adaptées permet de détecter plus efficacement des communautés dans des réseaux d'information. Enfin, la méthode 2Mod-Louvain produit aussi des résultats meilleurs ou du même ordre que Totem, une autre méthode utilisant les deux types d'information.

5 Conclusion

Dans cet article, nous avons étudié le problème du partitionnement de graphes avec nœuds et arêtes valués. En nous inspirant de la modularité de Newman et Girvan conçue pour la détection de communautés dans un réseau social, nous avons introduit une mesure de modularité basée sur l'inertie. Cette mesure est adaptée pour évaluer la qualité d'une partition d'éléments représentés dans un espace vectoriel réel. Nous avons également présenté 2Mod-Louvain, un algorithme qui combine notre critère de modularité basée sur l'inertie avec la modularité de Newman et Girvan pour détecter des communautés dans des réseaux d'information. Nous avons démontré formellement que ce nouvel algorithme peut être optimisé dans sa phase itérative, lui permettant d'être aussi efficient que l'algorithme de Louvain lors d'un passage à l'échelle. Comme le montrent nos expérimentations, en exploitant de manière conjointe les données relationnelles et vectorielles, la méthode 2Mod-Louvain détecte plus efficacement

les communautés que des méthodes utilisant un seul type de données et elle est robuste face à une dégradation des relations, des attributs, une augmentation de la densité des relations ou de la taille du réseau. Nous continuons les expérimentations et nos recherches afin de préciser encore plus finement les situations où 2Mod-Louvain est particulièrement efficace.

Références

Arenas, A., A. Fernández, et S. Gómez (2008). Analysis of the structure of complex networks at different resolution levels. *New Journal of Physics 10*(5), 053039.

Aynaud, T., V. Blondel, J.-L. Guillaume, et R. Lambiotte (2010). Optimisation locale multi-niveaux de la modularité. In Charles-Edmond Bichot et P. Siarry (Eds.), *Partitionnement de graphe : optimisation et applications*, Chapter 13, pp. 389–422. Hermes-Lavoisier.

Combe, D., C. Largeron, E. Egyed-Zsigmond, et M. Géry (2013). Totem : une méthode de détection de communautés adaptées aux réseaux d'information. In *EGC*, pp. 305–310.

Dang, T. A. (2012). *Analysis of communities in social networks*. Ph. D. thesis, Université Paris 13.

Davies, D. et D. Bouldin (1979). A cluster separation measure. *Pattern Analysis and Machine Intelligence* (2), 224—-227.

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*.

Good, B., Y. D. Montjoye, et A. Clauset (2010). Performance of modularity maximization in practical contexts. *Physical Review E 81*(4), 046106.

Guimera, R., M. Sales-Pardo, et L. Amaral (2004). Modularity from fluctuations in random graphs and complex networks. *Physical Review E 70*(2), 025101.

Hubert, L. et P. Arabie (1985). Comparing partitions. *Journal of Classification 2*(1), 193–218.

Lancichinetti, A. et S. Fortunato (2011). Limits of modularity maximization in community detection. *Physical Review E84, 066122*.

Newman, M. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America 103*(23), 8577–8582.

Reichardt, J. et S. Bornholdt (2006). Statistical mechanics of community detection. *Physical Review E 74*(1).

Rosenberg, A. et J. Hirschberg (2007). V-measure : A conditional entropy-based external cluster evaluation measure. *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning 1*(June), 410–420.

Rousseeuw, P. (1987). Silhouettes : A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics 20*(1), 53–65.

Vinh, N., J. Epps, et J. Bailey (2010). Information theoretic measures for clusterings comparison : Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research 9999*, 2837–2854.

Yang, J. et J. Leskovec (2012). Defining and Evaluating Network Communities based on Ground-truth. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*,

pp. 3. ACM.

Zhou, Y., H. Cheng, et J. X. Yu (2009). Graph clustering based on structural/attribute similarities. *Proceedings of the VLDB Endowment 2*(1), 718–729.

Summary

The modularity was introduced by Newman to estimate the quality of a graph vertex partition but this measure doesn't consider the values describing the nodes in the graph. In this article, we introduce a new complementary modularity measure, based on the inertia and specially conceived to evaluate the quality of a partition over vector space elements and consequently. We propose 2Mod-Louvain, a method using our inertia based quality criteria combined with Newman's modularity in order to detect communities in information networks. Our experiments show that combining the relational information and the vector information when partitioning a network detects communities more efficiently than methods using only one type of information. Our method is in addition, more robust to data degradation.