Maximisation de la modularité sous condition de fusion

Olivier Gach[*,**] — **Jin-Kao Hao**[**]

[*] *LIUM*
Université du Maine
Av. O. Messiaen
72085 Le Mans, France
olivier.gach@univ-lemans.fr

[**] *LERIA*
Université d'Angers
2 Boulevard Lavoisier
49045 Angers Cedex 01, France
hao@info.univ-angers.fr

RÉSUMÉ. *La recherche de communautés est un problème important dans le domaine des réseaux complexes. La modularité est probablement la mesure de qualité de partitionnement la plus populaire pour la détection de communautés. Malheureusement, elle présente un défaut, nommé limite de résolution, qui tend à faire disparaître les petites communautés d'autant plus que le graphe est grand. L'algorithme de Louvain, parmi d'autres, procède par fusion de communautés pour optimiser la modularité. Nous présentons une condition de fusion qui a pour objet d'empêcher les fusions incorrectes. L'évaluation expérimentale sur 20 graphes générés montre que cette technique, appliquée à l'algorithme de Louvain, réduit fortement les effets de la limite de résolution.*

ABSTRACT. *Community detection is an important issue in the field of complex networks. Modularity is probably the most popular measure of clustering quality for community detection. Unfortunately, it presents a defect, called resolution limit, which tends to make disappear the small communities especially as the graph is large. The algorithm of Louvain proceeds by merge of communities to optimize the modularity. We present a merge condition which aims to prevent wrong merges. The experimental evaluation on 20 generated graphs shows that this technique, applied to the algorithm of Louvain, strongly reduces the effects of the resolution limit. The merge condition can also be used in other algorithms based on community fusion.*

MOTS-CLÉS : *heuristique, détection de communautés, graphes de terrain, réseaux complexes, partitionnement de graphe, modularité, optimisation combinatoire, limite de résolution*

KEYWORDS: *heuristic, community detection, complex networks, community fusion, modularity, resolution limit*

1. Introduction

Certains réseaux formés par l'activité humaine ou observés dans la nature présentent des caractéristiques topologiques non triviales qui les distinguent fondamentalement de réseaux plus simples, en particulier ceux constitués aléatoirement par des modèles mathématiques. De par leurs propriétés communes, ces réseaux font l'objet d'études indépendamment de leur origine depuis une dizaine d'années, sous la dénomination de *réseaux complexes* ou *graphes de terrain*. Une de ces propriétés remarquable est la présence de zones de forte densité décrites comme des groupes, modules ou communautés selon les applications (par exemple, groupe d'individus au sens sociologique [MOR 34] ou ensemble de protéines interagissant entre elles [ITO 01] en biologie). De manière informelle, une communauté est définie comme un ensemble d'éléments du réseau fortement connectés à l'intérieur du groupe et faiblement connectés à l'extérieur. L'étude de la structure communautaire sucite une intense activité de recherche pour visualiser et comprendre la dynamique d'un réseau à différentes échelles.

La modélisation par un graphe fournit un support à l'étude des réseaux complexes, en identifiant un élément du réseau à un nœud et un lien entre deux éléments à une arête, si l'on considère un graphe non orienté. Dans la plus simple des formulations, chaque nœud n'appartient qu'à une seule communauté, un découpage en communauté étant une partition de l'ensemble des nœuds. Dans un graphe $G = (V, E)$, muni d'un ensemble de n nœuds V et d'un ensemble de m arêtes E, nous notons une partition en communautés $\mathcal{P} = \{C_1, C_2, ..., C_k\}$. Les communautés peuvent être définies localement, par un critère d'existence par exemple, ou globalement par une fonction de qualité de partition. Ce problème proche du partitionnement de graphe, s'en distingue par la méconnaissance a priori du nombre de communautés et de la taille de celles-ci.

Il existe une grande variété de mesures de qualité de partitionnement. La plus utilisée est la modularité, introduite par Newman en 2004 [NEW 04]. Dans une communauté C, les arêtes ayant leurs deux extrémités dans la communauté sont dites internes et leur nombre est noté $l(C)$. La somme des degrés des nœuds appartenant à C est nommée degré de communauté et noté $d(C)$. A partir de ces définitions, la modularité $Q(\mathcal{P})$ s'écrit :

$$Q(\mathcal{P}) = \sum_{C \in \mathcal{P}} \left(\frac{l(C)}{m} - \left(\frac{d(C)}{2m} \right)^2 \right) \quad (1)$$

Cette mesure globale ouvre la voie aux algorithmes d'optimisation pour la détection de communautés. Le problème étant NP-difficile [BRA 08], les heuristiques deviennent indispensables dès que le graphe atteint une centaine de nœuds. Parmi les nombreuses méthodes d'optimisation de Q, les approches fondées sur une recherche locale par déplacement de nœud et fusion de communautés sont les plus efficaces et rapides. Une d'entre elles, l'algorithme de Louvain [BLO 08], qui utilise une technique multi-niveau, offre un compromis intéressant entre l'optimisation et la rapidité de calcul, avec une complexité constatée sur des graphes peu denses en $O(m)$.

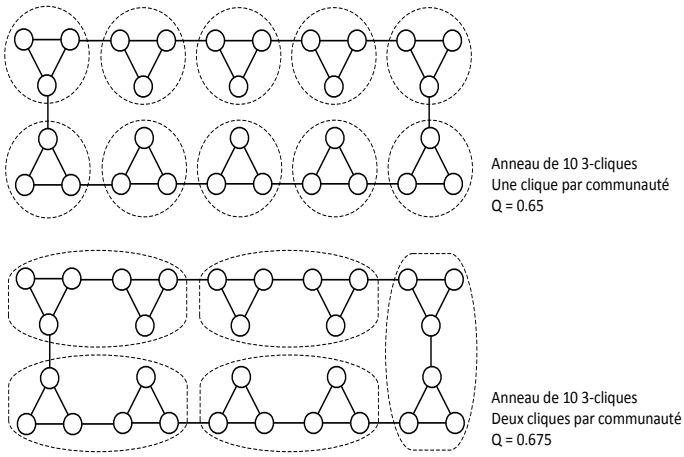

Figure 1. *Illustration de la limite de résolution. À partir d'un anneau de dix 3-cliques reliés deux à deux par un seul lien, la partition naturelle associe une communauté à chaque clique avec une modularité d'environ 0.65. La modularité peut être améliorée en regroupant les cliques par deux, ce qui pourtant constitue manifestement une partition erronée.*

Malheureusement, la modularité souffre d'un défaut qui s'aggrave avec l'augmentation de la taille du graphe, nommé limite de résolution et établi par Fortunato et Barthemely en 2007 [FOR 07]. L'optimisation de la modularité conduit, dans certaines conditions, à faire disparaître les petites communautés. La démonstration des auteurs portent sur des cas extrêmes de cliques connectées entre elles de façon minimale, par exemple un anneau de cliques reliées par paires par une seule arête (voir la Figure 1). Dans ce cas, si le graphe est suffisamment grand, les cliques de taille inférieure ou égale à $\sqrt{m}$ sont fusionnées avec d'autres cliques alors qu'elles devraient être identifiées comme des communautés. Ce phénomène est aussi observé dans des situations plus réalistes, par exemple avec des graphes et des partitions générés aléatoirement [LAN 08, LAN 09].

L'algorithme de Louvain procède en deux phases : une phase de constitution d'une partition avec la procédure VM (*vertex mover*, [SCH 08]), puis une phase de regroupements des communautés aux niveaux supérieurs, assimilables à des fusions de communautés. En exécutant l'algorithme avec des graphes générés par le modèle LFR [LAN 08], le nombre de communautés obtenu à l'issu de VM est plus grand que le nombre de communautés de la solution, alors qu'il devient plus petit après la phase de fusion de Louvain. Il en va de même pour les petits communautés qui grossissent et deviennent trop grandes après cette seconde phase. C'est la manifestation la plus évidente de la limite de résolution. Même si intrinsèquement l'optimisation de la

modularité engendre ce défaut, nous pouvons conjecturer que certaines fusions sont erronées alors que d'autres sont justifiées.

Notre idée est d'établir un critère local de fusion de communautés. A partir de deux communautés C et C' et du sous-graphe $C \cup C'$ qu'elles engendrent, ce critère est vrai si C et C' doivent être fusionnées et faux sinon. Les graphes LFR nous permettent d'apprécier la performance d'un tel critère car nous pouvons vérifier d'après la partition solution si deux communautés doivent ou non être fusionnées. Notre idée intuitive est de fonder le critère de fusion sur la modularité calculée pour la partition $\{C, C'\}$ dans le sous-graphe $C \cup C'$. L'idée sous-jacente est que deux communautés doivent être réunies si elles ne constituent pas séparément deux modules forts, c'est-à-dire si leur modularité est faible. Malheureusement, selon nos tests préliminaires, il n'existe pas de corrélation significative entre la valeur de modularité pour la partition (C,C') et la décision juste de fusionner C et C'. En revanche, la modularité paramétrique Q_λ [REI 06] révèle une corrélation d'autant plus forte que le paramètre λ, qui désigne la résolution à laquelle le graphe est partitionné, est proche de 0.25.

2. Formulation

2.1. *Formule générale*

Soit deux communautés C et C' de la partition $\mathcal{P}$ définie dans le graphe $G = (V, E)$. Notons $G_{C,C'} = (C \cup C', E_{C,C'})$ le sous-graphe de G engendré par le sous-ensemble de nœuds $C \cup C'$. $E_{C,C'}$ est l'ensemble des arêtes de ce sous-graphe et son cardinal est noté e pour simplifier les formules. Le degré d'un nœud v dans un sous-graphe S quelconque de G est noté $d(v, S)$ et se définit comme le nombre d'arêtes de S incidentes à v. Le degré de communauté $d(C, S)$ est alors défini comme la somme des degrés $d(v, S)$ des nœuds v appartenant C, soit $d(C, S) = \sum_{v \in C} d(v, S)$. Cette définition préalable permet d'écrire la modularité paramétrique de la partition $\{C, C'\}$ dans le graphe $G_{C,C'}$:

$$Q_\lambda(C, C') = \frac{l(C) + l(C')}{e} - \lambda \frac{d(C, G_{C,C'})^2 + d(C', G_{C,C'})^2}{4e^2} \qquad (2)$$

Pour un paramètre θ compris entre 0 et 1, le critère de fusion $M_{\lambda,\theta}(C, C')$ que nous proposons, s'écrit :

$$Q_\lambda(C, C') < \theta \qquad (3)$$

Cela signifie que les communautés C et C' sont fusionnées si la modularité de paramètre λ de la partition $\{C, C'\}$ dans le sous-graphe $C \cup C'$ est strictement inférieure au seuil θ.

Pour caractériser les valeurs de θ, nous devons tout d'abord reformuler le critère en réduisant le nombre de variables.

2.2. *Reformulation*

Notons $d_{out}(C)$ le degré externe de la communauté C, c'est-à-dire le nombre d'arêtes ayant une extrémité dans C et l'autre en dehors. Nous avons $d(C) = 2l(C) + d_{out}(C)$ car en calculant la somme des degrés des nœuds de C, les arêtes internes sont comptées deux fois. Dans le contexte du sous-graphe $G_{C,C'}$, les arêtes comptées dans $d_{out}(C)$ relient nécessairement un nœud de C à un nœud de C', soit $d_{out}(C) = d_{out}(C') = l(C, C')$, le nombre des arêtes reliant C et C'. Nous en déduisons trois relations importantes pour la suite, dans le contexte de deux communautés C et C' engendrant un sous-graphe $G_{C,C'}$:

$$\begin{cases} d(C, G_{C,C'}) = 2l(C) + l(C, C') \\ d(C', G_{C,C'}) = 2l(C') + l(C, C') \\ l(C) + l(C') + l(C, C') = e \end{cases} \tag{4}$$

Comme $d(C, G_{C,C'}) = e + l(C) - l(C')$ et $d(C', G_{C,C'}) = e - l(C) + l(C')$, la modularité se réécrit :

$$\begin{aligned} Q_\lambda(C, C') &= \frac{l(C) + l(C')}{e} - \frac{\lambda}{4e^2}\left((e + l(C) - l(C'))^2 + (e - l(C) + l(C'))^2\right) \\ &= \frac{l(C) + l(C')}{e} - \frac{\lambda}{2e^2}(l(C) - l(C'))^2 - \frac{\lambda}{2} \end{aligned} \tag{5}$$

Dans cette forme, l'expression de la modularité ne conserve que trois variables : $l(C)$, $l(C')$ et $e = |E_{C,C'}|$.

2.3. *Expression de* θ

Pour borner le paramètre θ, nous considérons des cas extrêmes, en partant du principe que e, le nombre d'arêtes du sous graphe engendré par $C \cup C'$, est constant et en faisant varier $l(C)$ ou $l(C')$ entre 0 et e, le maximum possible.

2.3.1. *Borne supérieur de* θ

Supposons que les communautés C et C' ne soient pas connectées, c'est-à-dire que $l(C, C') = 0$. La troisième égalité dans (4) donne $l(C)+l(C') = e$ soit $l(C)-l(C') = 2l(C) - e$. La modularité écrite en 5 devient :

$$\begin{aligned} Q_\lambda(C, C') &= 1 - \frac{\lambda}{2e^2}(2l(C) - e)^2 - \frac{\lambda}{2} \\ &= \frac{-2\lambda l(C)^2}{e^2} + \frac{2\lambda l(C)}{e} + 1 - \lambda \end{aligned} \tag{6}$$

La valeur de e étant donnée, la modularité peut être vue comme une fonction de la variable $l(C)$ définie sur $\{0, ..., e\}$. Cette fonction est croissante sur l'intervalle

$[0...e/2]$ et décroissante sur $[e/2...e]$. Pour $l(C) = 0$ et $l(C) = e$, $Q_\lambda(C, C') = 1 - \lambda$. Nous avons donc $Q_\lambda(C, C') \geq 1 - \lambda$. Le critère de fusion, qui s'écrit $Q_\lambda(C, C') < \theta$ doit toujours être négatif car, dans la situation d'hypothèse ($l(C, C') = 0$), C et C' ne doivent jamais être fusionnées. Il faut donc choisir θ de façon à ce que $Q_\lambda(C, C') \geq \theta$, quelque soit C et C' non connectées. Cela est vérifié si $\theta \leq 1 - \lambda$ car dans ce cas $Q_\lambda(C, C') \geq 1 - \lambda \geq \theta$.

2.3.2. *Borne inférieure de θ*

A l'inverse ici, considérons que C et C' sont connectées, donc que $l(C, C') > 0$. Cette hypothèse nous assure que les communautés peuvent être fusionnées. Considérons de plus que la communauté C' contient seulement un nœud et donc aucune arête interne : $l(C') = 0$. Il en résulte que $e = l(C) + l(C, C')$ et, selon la première hypothèse, $l(C) < e$. La modularité utilisée dans le critère de fusion devient :

$$Q_\lambda(C, C') \quad = \quad -\frac{\lambda}{2e^2} l(C)^2 + \frac{l(C)}{e} - \frac{\lambda}{2} \tag{7}$$

Comme précédemment, la modularité peut être vue comme une fonction de la variable $l(C)$ définie sur $\{0, ..., e\}$. Dans l'intervalle $[0...e]$, cette fonction est strictement croissante avec pour borne $-e/\lambda$ pour $l(C) = 0$ et $1 - \lambda$ pour $l(C) = e$. Donc, dans l'intervalle de variation de $l(C)$, comme de plus par hypothèse $l(C) < e$, nous avons $Q_\lambda(C, C') < 1 - \lambda$. Dans la situation d'hypothèse, où une communauté C' ne contient qu'un nœud, il est souhaitable par principe de n'interdire aucune fusion avec C' de façon à ce que toutes les situations de fusion de C' avec d'autres communautés voisines soient examinées pour trouver une fusion de gain de modularité maximum. Donc, le critère de fusion doit être positif et il faut choisir θ de sorte que $Q_\lambda(C, C') < \theta$ ce qui est vérifié si $\theta \geq 1 - \lambda$ car dans ce cas $Q_\lambda(C, C') < 1 - \lambda \leq \theta$.

La première hypothèse de connexion de C et C' fait que, dans le cas de non connexion, c'est-à-dire si $l(C, C') = 0$, soit $l(C') = e$, la fusion est rejetée. Ce n'est pas préjudiciable car il est inutile que le critère autorise une fusion qui, par principe, n'est pas viable.

2.3.3. *Valeur de θ*

Nous avons ainsi borné le paramètre θ selon deux cas particuliers où la réponse du critère de fusion est imposée. Il en résulte que la seule valeur acceptable de θ est $1 - \lambda$, soit un critère de fusion $M_\lambda(C, C')$ qui s'écrit simplement :

$$Q_\lambda(C, C') < 1 - \lambda, \text{ avec } \lambda > 0 \tag{8}$$

3. Étude générale

Le critère de fusion étant posé avec un seul paramètre λ, nous souhaitons étudier dans quelles conditions et pour quelles valeurs de λ, ce critère opère correctement

et en particulier atténue la limite de résolution. L'idée générale est qu'il fonctionne comme un filtre préalable avant l'examen du gain de modularité obtenu par la fusion de deux communautés.

Nous aurons besoin pour la suite d'une écriture du critère de fusion sans e au dénominateur :

$$\frac{l(C)+l(C')}{e} - \frac{\lambda}{2e^2}(l(C)-l(C'))^2 - \frac{\lambda}{2} < 1-\lambda$$
$$2e(l(C)+l(C')) - \lambda(l(C)-l(C'))^2 + (\lambda-2)e^2 < 0 \quad (9)$$

3.1. *Valeurs de λ admissibles*

Interrogeons nous sur la valeur minimale de la modularité paramétrique $Q_\lambda(C,C')$, d'après sa forme générale écrite en 5. Posons $l(C)-l(C')=h$, h variant de $-e$ à e. La modularité paramétrique devient :

$$Q_\lambda(C,C') \quad = \quad \frac{2l(C)}{e} - \frac{h}{e} - \frac{\lambda}{2e^2}h^2 - \frac{\lambda}{2} \quad (10)$$

Pour h constant, cette fonction de $l(C)$ est croissante, donc son minimum est atteint pour $l(C)=0$ et vaut $-h/e-\lambda h^2/(2e^2)-\lambda/2$. Il y a ainsi un minimum pour chaque valeur de h possible. Sous la condition que $l(C)=0$, h peut varier de $-e$ à 0. La fonction $f(h)=-h/e-\lambda h^2/(2e^2)-\lambda/2$ est toujours négative, puisque $\lambda \geq 0$, et a la forme d'un polynôme du second degré, avec un coefficient de monôme de plus haut degré négatif. Donc sur l'intervalle $[-e...0]$ son minimum est soit $f(-e)$, soit $f(0)$, c'est-à-dire soit $1-\lambda$, soit $-\lambda/2$. En comparant ces deux valeurs, le minimum des deux, c'est-à-dire le minimum absolu de $Q_\lambda(C,C')$ est $-\lambda/2$ si $\lambda<2$ et $1-\lambda$ sinon. Dans ce dernier cas, le critère est toujours faux car la modularité paramétrique doit être strictement inférieure à $1-\lambda$. Ainsi, le critère de fusion est opérant uniquement si $\lambda<2$.

3.2. *Communautés mal formées*

Le critère le plus faible d'existence d'une communauté [HU 08] exige que celle-ci ait au moins deux fois plus d'arêtes internes que de liens partagés avec n'importe quelle autre communauté. Dans la situation réduite à deux communautés C et C', cette condition nécessite que $2l(C)$ et $2l(C')$ soit strictement supérieur à $l(C,C')$.

Plaçons nous du point de vue d'une seule des deux communautés, par exemple C, la situation étant symétrique. Selon le critère d'existence de communauté, deux cas peuvent se présenter.

1) La communauté C est bien formée : $2l(C) > l(C,C')$. La condition de fusion doit opérer pleinement sans préjuger de sa valeur. A la limite, lorsque $l(C,C')=1$, pouvons-nous exiger que la fusion soit interdite (voir la section 5).

2) La communauté C est mal formée : $2l(C) \leq l(C, C')$. Dans ce cas la prévention de la limite de résolution n'est pas utile car la fusion est a priori légitime. Il faut systématiquement autoriser la fusion pour reporter la décision de fusion effective sur la recherche du gain de modularité maximum.

Recherchons dans quelle condition le critère de fusion est toujours vrai si $2l(C) \leq l(C, C')$. Notons $l(C, C') = hl(C)$ avec $h \geq 2$. Comme $e = (1 + h)l(C) + l(C')$, la condition de fusion devient :

$$\begin{aligned}
&[(2\lambda - 2)h + 4\lambda]l(C)l(C') + [(\lambda - 2)h^2 + (2\lambda - 2)h]l(C)^2 < 0 \\
&[(2\lambda - 2)h + 4\lambda]l(C') + [(\lambda - 2)h^2 + (2\lambda - 2)h]l(C) < 0 \\
&[(2\lambda - 2)h + 4\lambda]l(C') < [(2 - \lambda)h^2 + (2 - 2\lambda)h]l(C) \qquad (11)
\end{aligned}$$

La condition doit être vraie quelque soit les valeurs de $l(C)$ et $l(C')$ qui sont positives. Cela est rendu possible si les deux conditions suivantes sont réunies :

$$\begin{cases} (2\lambda - 2)h + 4\lambda < 0 \\ (2 - \lambda)h + 2 - 2\lambda > 0 \end{cases} \qquad (12)$$

La première inéquation est équivalente à $h < 4\lambda/(2 - 2\lambda)$ si $\lambda \geq 1$ ou $h > 4\lambda/(2 - 2\lambda)$ si $\lambda < 1$. La première est impossible car si $\lambda \geq 1$ alors $4\lambda/(2 - 2\lambda)$ est négatif ce qui oblige h à être négatif. Il reste $h > 4\lambda/(2 - 2\lambda)$. Selon le même raisonnement que précédemment, comme $h \geq 2$, cette inéquation est rendue toujours vraie en prenant λ de sorte que $2 > 4\lambda/(2 - 2\lambda)$, soit $\lambda < 1/2$ (compatible avec la condition initiale $\lambda < 1$).

Le second inéquation est équivalente à $h > (2\lambda - 2)/(2 - \lambda)$ si $\lambda \leq 2$ (l'autre alternative est impossible car toujours inférieur à 2). Par le même procédé, on obtient $\lambda < 3/2$.

Les deux conditions devant être vérifiées, la plus restrictive prime et il est donc nécessaire que $\lambda < 1/2$ pour que toute communauté mal formée passe le filtre du critère de fusion.

4. Validation expérimentale

Nous avons établi que la condition de fusion $Q_\lambda(C, C') < 1 - \lambda$ est efficace si $0 < \lambda < 1/2$. La borne inférieure est incontestable car la condition serait toujours vraie et donc non discriminante si λ était négatif. En revanche la borne supérieure est soumise à une hypothèse qu'il est préférable de vérifier expérimentalement. Nous avons donc choisi de tester les valeurs de λ de 0 à 1 par pas de 0.1. La valeur $\lambda = 0$ représente l'absence de condition puisque Q_0 est toujours inférieur ou égal à 1. La valeur $\lambda = 1$ correspond à un critère fondé sur la modularité standard $Q_1 = Q$.

Nos tests consistent à appliquer la condition de fusion à l'algorithme de Louvain sur 20 graphes artificiels LFR de caractéristiques variées : n de 80 à 100000, m de

200 à presque un million, *mixing parameter* μ (donnant la moyenne des parts d'arêtes externes de nœud) de 0.3 (communautés bien définies) à 0.85 (communautés très faiblement définies), degrés et tailles variables de communautés, différentes situations représentées (graphes peu denses, graphes très denses, distributions étendues de tailles de communautés, etc). Le critère de fusion intervient dans la second phase de l'algorithme de Louvain pour interdire ou autoriser l'examen d'une fusion qui sera effectivement effectuée si son gain de modularité est maximum. La validation de la condition de fusion par les graphes générés est pertinente car la partition de référence permet de mesurer, par les similarités NMI (information mutuelle normalisée [DAN 05]) et NID (distance d'information normalisée [VIN 10]), la justesse de la partition trouvée pour chaque graphe. Les partitions sont d'autant plus proches que la valeur NMI ou NID est proche de 1. En complément de ces mesures globales, nous nous attachons à observer trois indicateurs structurels : le nombre de communautés k, la taille de la plus petite communauté $min|C_i|$ et la taille de la plus grande communauté $max|C_i|$. Pour chaque graphe de test, nous avons généré 100 instances avec un ordre des nœuds aléatoire. Pour tous les tests présentés, la grandeur (NMI, NID, k, $max|C_i|$ et $min|C_i|$) associée à un graphe est la moyenne de ces valeurs sur les 100 instances.

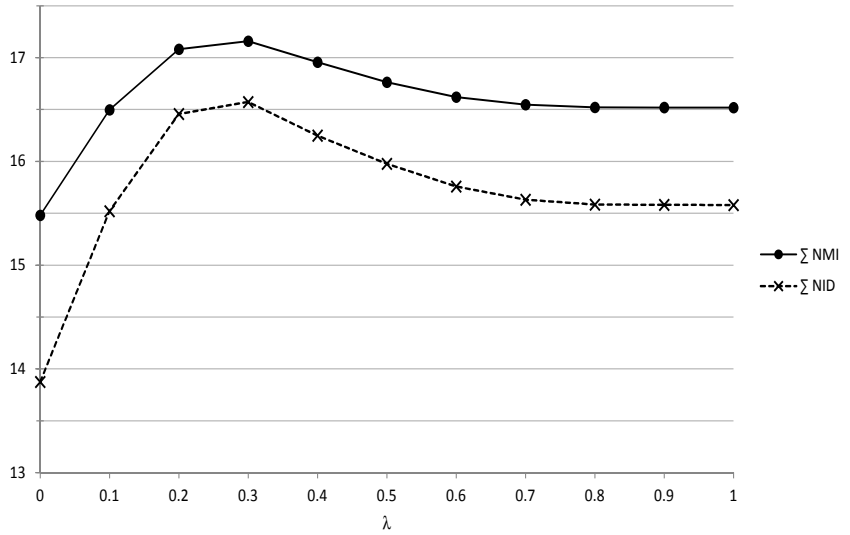

Figure 2. *Pertinence de la condition de fusion. Les similarités NMI et NID sont mesurées entre les partitions trouvées par l'algorithme de Louvain avec condition de fusion (paramètre λ en abscisse) et les partitions de référence. Les valeurs de NMI et NID, entre 0 et 1, sont sommées pour les 20 graphes testés, soit une amplitude de 0 à 20 en ordonnée.*

La figure 2 représente la somme de NMI et de NID pour les 20 graphes de test, selon le paramètre λ. Les deux grandeurs augmentent fortement de $\lambda = 0$ à $\lambda = 0.3$ pour diminuer ensuite moins rapidement mais continument jusqu'à $\lambda = 1$. La condi-

tion de fusion améliore nettement l'algorithme de Louvain ($\lambda = 0$) quant à la justesse des partitions produites, comparées aux solutions de graphe artificiel. D'après les résultats détaillés par graphe, non publiés, entre $\lambda = 0$ et $\lambda = 0.3$ la NMI est augmentée pour tous les graphes, avec une moyenne d'augmentation absolue de 0.08. L'augmentation relative moyenne est de 11.5%. C'est une amélioration très significative compte tenu du caractère logarithmique de la NMI.

Sur la figure 3, nous observons qu'à l'issue de la première phase VM, le nombre de communautés k est trop grand par rapport à la solution car des fusions manquent. En revanche, la second phase de l'algorithme de Louvain ($\lambda = 0$) procède à un trop grand nombre de fusions de sorte que k devient trop petit. Les meilleurs résultats sont obtenus dans l'intervalle $0.2 < \lambda < 0.4$. Nous avons constaté, dans le détail par graphe non publié ici, que l'essentiel des erreurs avec le meilleur paramètre $\lambda = 0.3$ sont dues à seulement un tiers des graphes.

Observons enfin dans la même figure la plus grande et la plus petite communauté qui sont des indicateurs structurels importants car il peuvent révéler de manière significative la limite de résolution. On constate qu'avec Louvain ($\lambda = 0$), la taille de plus petite communauté est huit fois plus grande que la taille attendue, ce qui traduit bien la disparition des petites communautés. La valeur la plus juste, proche du ratio 1, est obtenue pour $\lambda = 0.2$. Ce ratio diminue pour des valeurs de λ plus grande, aggravant l'erreur. Un autre enseignement, tout aussi important, est fourni par le ratio de plus grande communauté qui dénote une erreur importante (valeur trois fois plus grande) quelque soit le paramètre λ et même dès la première phase VM. Ainsi, l'heuristique VM souffre de la limite de résolution en constituant des communautés trop grandes, erreur qui ne peut pas être réparée par la seconde phase de fusion. Là encore, indiquons que cette erreur ne se manifeste que pour seulement un tiers des graphes.

5. Limite de résolution

Cherchons à quelle condition le critère de fusion est faux dans le cas de deux communautés C et C' reliées par seulement une arête. A partir de l'écriture 9 du critère de fusion, avec la condition $l(C, C') = 1$, soit $e = l(C) + l(C') + 1$ et en posant $l(C') = hl(C)$, nous obtenons la condition de fusion fausse suivante :

$$(4\lambda l(C) + 2\lambda - 2)h + 4\lambda l(C)^2 + 4(\lambda - 1)l(C) + \lambda - 2 \geq 0 \qquad (13)$$

Plaçons nous dans le cas où C est la plus petite communauté sans perte de généralité (inversion de C' et C dans le cas contraire), donc si $h \geq 0$. L'expression du critère est de la forme $Ah+B$ avec $A = 4\lambda l(C)+2\lambda-2$ et $B = 4\lambda l(C)^2+4(\lambda-1)l(C)+\lambda-2$. Dans l'hypothèse où $h = 0$, le signe de B est le signe de $2(l(C)+1)(2\lambda l(C)+\lambda-2)$ par factorisation, soit le signe de $(2\lambda l(C)+\lambda-2)$ puisque $l(C)+1$ est positif. Donc, dans l'hypothèse où h est nul, le critère de fusion est faux si $(2\lambda l(C) + \lambda - 2) \geq 0$, soit $l(C) \geq 1/\lambda - 1/2$. Si maintenant h augmente (h est positif), la fonction $Ah + B$ doit être croissante pour qu'elle reste positive, donc A doit être positif, soit $l(C) \geq$

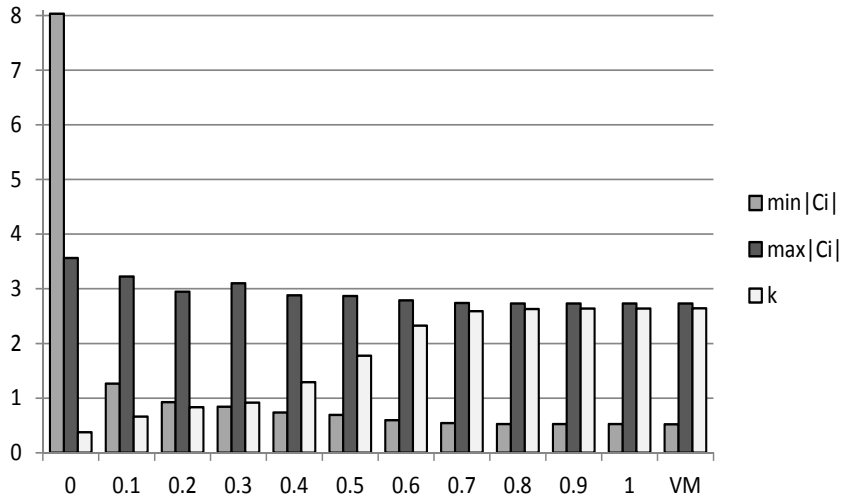

Figure 3. *Erreurs sur le nombre de communautés, la taille de la plus petite et de la plus grande communauté. Pour chacune de ces grandeurs, le graphique représente la moyenne, sur l'ensemble des graphes, du rapport entre la valeur trouvée et la valeur de la solution. Ainsi la moyenne de 1 constitue un comportement idéal de l'algorithme qui trouve la valeur attendue. Ces ratios sont représentés par rapport au paramètre λ placé en abscisse et pour l'algorithme VM (première phase de Louvain).*

$1/(2\lambda) - 1/2$. Des deux conditions, $l(C) \geq 1/\lambda - 1/2$ et $l(C) \geq 1/(2\lambda) - 1/2$, la première est plus stricte car $\lambda > 0$.

Le critère de fusion ne s'appuie que sur le nombre d'arêtes de la communauté C et non sur sa densité. Pour imposer que la fusion soit interdite et que donc le critère soit faux, si les deux communautés à fusionner sont reliées par seulement une arête, il faut que la plus petite communauté C vérifie $l(C) \geq 1/\lambda - 1/2$. Si λ s'approche de zéro, la condition n'est vraie que pour des valeurs de $l(C)$ de plus en plus grandes, ce qui n'est pas souhaitable. Si l'on souhaite interdire les fusions pour toute communauté telle que $l(C) >= K$, il faut $\lambda \geq 2/(2K + 1)$. Une valeur raisonnable est $K = 3$ qui exige $\lambda \geq 2/7 \approx 0.285$. Dans ce cas, les communautés qui ont une arête externe et au moins trois arêtes internes ne seront jamais fusionnées. Il est important de noter que C n'est pas nécessairement une clique. Nous touchons là un point important concernant l'implantation de cette condition de fusion dans un algorithme. La limite de résolution est atténuée spécialement dans le cas des communautés denses reliées par une seule arête à ne pas fusionner. Mais la fusion de deux communautés peu denses reliées par une seule arête sera tout aussi bien bloquée de façon probablement indue. Il est donc indispensable que les communautés soumises à ce critère de fusion soit "bien formées" avec une densité aussi forte que possible. En cela, l'application du critère à un algorithme qui procède d'abord par déplacement de nœuds (procédure VM) puis

par fusion, comme Louvain, est recommandée. Alors qu'un algorithme comme Fast-Greedy [CLA 04] qui fusionne dès le départ, avec des communautés minuscules et faibles, ne gagnera sans doute pas à utiliser la condition de fusion.

6. Conclusion

Nous avons présenté un critère discriminant les fusions de communautés pertinentes, à utiliser dans un algorithme de détection de communautés dans les réseaux complexes. Cette condition s'appuie sur la modularité de paramètre λ calculée pour chaque paire de communautés à fusionner. Elle a une portée générale et peut s'appliquer à n'importe quel algorithme qui procède à des regroupements de communautés en optimisant la modularité. Son objectif est de réduire l'impact de la limite de résolution, défaut reconnu de l'optimisation de la modularité, qui se traduit dans ce type d'algorithme par des fusions incorrectes conduisant à des communautés trop grandes. Une application à l'algorithme de Louvain sur un échantillon d'une vingtaine de graphes générés montre son efficacité pour un paramètre λ aux alentours de 0.3. Pour cette valeur de λ, la mesure moyenne de similarité, égale à 0.774 sans la condition de fusion passe à 0.858 avec, l'idéal étant de 1. L'erreur relative moyenne sur le nombre de communautés passe avec la condition de fusion de -62% à -16%. Cette technique n'est efficace qu'à la condition que les communautés examinées pour la fusion soient préalablement bien constituées avec une densité significative. L'algorithme de Louvain s'y prête particulièrement bien.

Remerciements

Ce travail est partiellement financé par la Région Pays de la Loire dans le cadre des projets "LigeRO" (2009-2013) et "Radapop" (2009-2013). Nous remercions les rapporteurs de notre articles pour leurs commentaires et questions qui ont aidé à améliorer sa présentation.

7. Bibliographie

[BLO 08] BLONDEL V. D., GUILLAUME J.-L., LAMBIOTTE R., LEFEBVRE E., « Fast unfolding of communities in large networks », *Journal of Statistical Mechanics : Theory and Experiment*, vol. 10, 2008, p. 8-+.

[BRA 08] BRANDES U., DELLING D., GAERTLER M., GORKE R., HOEFER M., NIKOLOSKI Z., WAGNER D., « On Modularity Clustering », 2008.

[CLA 04] CLAUSET A., NEWMAN M. E. J., MOORE C., « Finding community structure in very large networks », *Phys. Rev. E*, vol. 70, n° 6, 2004, page 066111, American Physical Society.

[DAN 05] DANON L., DÍAZ-GUILERA A., DUCH J., ARENAS A., « Comparing community structure identification », *Journal of Statistical Mechanics : Theory and Experiment*, vol. 2005, n° 09, 2005, page P09008, Institute of Physics Publishing.

[FOR 07] FORTUNATO S., BARTHÉLEMY M., « Resolution limit in community detection », *Proceedings of the National Academy of Sciences*, vol. 104, n° 1, 2007, p. 36–41.

[HU 08] HU Y., CHEN H., ZHANG P., LI M., DI Z., FAN Y., « Comparative definition of community and corresponding identifying algorithm », *Physical Review E*, vol. 78, n° 2, 2008, page 026121, APS.

[ITO 01] ITO T., CHIBA T., OZAWA R., YOSHIDA M., HATTORI M., SAKAKI Y., « A comprehensive two-hybrid analysis to explore the yeast protein interactome », *Proceedings of the National Academy of Sciences*, vol. 98, n° 8, 2001, p. 4569–4574, National Acad Sciences.

[LAN 08] LANCICHINETTI A., FORTUNATO S., RADICCHI F., « Benchmark graphs for testing community detection algorithms », *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, vol. 78, n° 4, 2008, APS.

[LAN 09] LANCICHINETTI A., FORTUNATO S., « Community detection algorithms : a comparative analysis. », *Physical Review E - Statistical, Nonlinear and Soft Matter Physics*, vol. 80, n° 5 Pt 2, 2009, page 056117, APS.

[MOR 34] MORENO J. L., *Who Shall Survive ? A New Approach to the Problem of Human Interrelations*, vol. 58, Nervous and Mental Disease Publishing, 1934.

[NEW 04] NEWMAN M. E. J., GIRVAN M., « Finding and evaluating community structure in networks », *Phys. Rev. E*, vol. 69, n° 2, 2004, page 026113, American Physical Society.

[REI 06] REICHARDT J., BORNHOLDT S., « Statistical mechanics of community detection », *Physical Review E*, vol. 74, n° 1, 2006, p. 1–14, APS.

[SCH 08] SCHUETZ P., CAFLISCH A., « Efficient modularity optimization by multistep greedy algorithm and vertex mover refinement », *Phys. Rev. E*, vol. 77, n° 4, 2008, page 046112, American Physical Society.

[VIN 10] VINH N. X., EPPS J., BAILEY J., « Information theoretic measures for clusterings comparison : Variants, properties, normalization and correction for chance », *The Journal of Machine Learning Research*, vol. 11, 2010, p. 2837–2854, JMLR. org.