# Communautés : Arrêtons de ne compter que les arêtes

Adrien Friggeri, Guillaume Chelius, Eric Fleury

# *Communautés : Arrêtons de ne compter que les arêtes*

Adrien Friggeri[1] and Guillaume Chelius[2] and Eric Fleury[1]

[1] *ENS de Lyon, INRIA D-NET, LIP UMR CNRS 5668, Université de Lyon*
[2] *INRIA D-NET, ENS de Lyon, LIP UMR CNRS 5668, Université de Lyon*

## 1 Introduction

Si un intérêt croissant s'est focalisé sur la détection de communautés dans les réseaux dits de « terrain » – aussi nommés grands graphes ou réseaux complexes [8], la notion même de communauté [2, 4] reste sans définition théorique. Le seul consensus général rejoint la notion intuitive qu'une communauté doit être un groupe de sommets *relativement dense*, qui a donc *plus de liens en son sein que vers le reste du graphe*. Malheureusement, ces notions admises de densité interne à un groupe par rapport aux nombre de liens qui le relie au reste n'ont jamais débouché sur une formalisation. En laissant de coté cette problématique épineuse de la notion de qualité d'une communauté, on a préféré considérer la qualité d'une partition, glissement impliquant que si une partition est jugée de qualité, alors chaque ensemble la constituant l'est aussi, ce qui n'est pas évident dans le cas général. La fonction de qualité la plus usitée est la « *Q-modularité* » [10]. Cette métrique compare la densité de liens à l'intérieure d'une communauté par rapport à sa densité estimée si les liens étaient tirés au hasard (*null model*). D'un point de vue algorithmique, le problème de partitionner un graphe en communautés se réduit à un problème d'optimisation où il s'agit de trouver une partition de l'ensemble des nœuds maximisant la fonction de qualité choisie. En général, ce problème est *NP*-difficile (c'est le cas pour la $Q$-modularité) et on a recours à diverses heuristiques [1, 5].

Plus récemment, un nouvel intérêt est né dans la détection de communautés *recouvrantes* qui permettent à un sommet d'appartenir à plusieurs communautés, ce qui intuitivement reflète mieux la réalité des choses dans beaucoup de réseaux de terrain : un article peut avoir un intérêt pour plusieurs communautés scientifiques, une personne peut appartenir à plusieurs groupes d'amis sur Facebook (famille, collègues, amis de lycée, de Fac...). Pour des raisons historiques, une grande majorité des méthodes proposées pour détecter des communautés recouvrantes s'est inspirée de l'existant et a ou bien adapté la définition de la modularité [9, 12], ou bien tiré parti de l'instabilité des partitions [13] ou encore tenté de découvrir des cliques [11]. Cependant, le passage d'une partition à un recouvrement a des conséquences sur le lien entre qualité globale d'un ensemble de communautés et qualité intrinsèque de chaque communauté : étant donné qu'il est possible d'ajouter arbitrairement ces communautés de bonne qualité dans le cas recouvrant, il est donc possible d'augmenter artificiellement la qualité globale tout en conservant des communautés de qualité moindre.

Dans cet article, nous souhaitons revenir sur la question de la définition d'une communauté en tant qu'ensemble de sommets $U$ sans avoir à en juger la qualité au regard des autres communautés, recouvrantes ou non. Ce qui importe c'est uniquement l'ensemble $U$ considéré et le graphe sous-jacent et ce indépendamment de tout découpage global. À dessein, nous introduisons la « *cohésion* » qui repose sur la relation forte qui existe entre des triplets de sommets lorsqu'ils forment un triangle ou au contraire sur la non présence de triangle traduisant la présence de lien faible [†]. La notion de communauté découle de cette mesure confinée à un sous-ensemble de sommets plongé dans son graphe d'origine : une communauté est un ensemble de sommets offrant une forte cohésion. Après avoir introduit la métrique de cohésion dans la section 2, nous illustrons dans la section 3 son application sur la découverte de communautés egocentrées dans des réseaux sociaux en utilisant un algorithme – similaire à celui exposé dans [3] mais se basant sur la cohésion – et donnons dans le section 4 quelques résultats sur l'application de ce calcul d'egomunautés.

†. Notion de *weak tie* introduite par A. RAPOPORT en 1957 et reprise par M.S. GRANOVETTER en 1973 dans [7]

## 2 Cohésion

Pour un graphe $G=(V,E)$, l'évaluation de la qualité d'un ensemble de communautés $\mathcal{S}=\{C_1,C_2,\ldots,C_k\}$ où $\bigcup C_i = V$ soulève la question des *frontières* et celle du *contenu*. La première se positionne au niveau globale et se doit d'évaluer si l'ensemble considéré $\mathcal{S}$ de communautés est cohérent en tant que tout. La seconde est relative à la qualité intrinsèque de chaque communauté prise séparément et indépendamment du reste de l'ensemble. Vouloir que $\mathcal{S}$ forme une partition (cas non recouvrant) implique que $\forall i \neq j, C_i \cap C_j = \emptyset$ : la notion d'appartenance à une communauté devient une relation d'équivalence sur les sommets. La transitivité de la relation induit alors que les deux propriétés (frontière/contenu) sont liées dans le cas non recouvrant. Dans le cas recouvrant, ces deux questions deviennent totalement découplées, puisque l'on est en mesure de modifier une communauté sans affecter les autres. Nous allons ici nous concentrer sur la question du contenu et proposer une mesure de cohésion qui pour un ensemble de sommets $U \subseteq V$ donne sa propension à être un ensemble *communautaire*. Nous souhaitons que la cohésion reflète les 3 propriétés suivantes pour une communauté : *(i)* la qualité d'une communauté ne doit pas dépendre de l'existence ou non d'autres communautés ; *(ii)* un sommet lointain n'a pas d'influence sur elle ; *(iii)* l'information circule plus facilement au sein de la communauté que vers l'extérieur de cette dernière. Cette dernière propriété n'est pas uniquement due à la densité d'arêtes mais reprend l'argumentation exposée par GRANOVETTER sur la notion de lien faible [‡]. Ce que nous affirmons comme important dans la définition d'une communauté, c'est qu'il ne suffit pas de compter le nombre d'arêtes internes et externes mais qu'il faut absolument prendre en compte les triangles, *i.e.*, la structure forte qui fait le liant et permet la cohésion au sein d'une communauté, et que les arêtes qui ne font pas partie d'un triangle ont vocation à jouer le rôle de *pont local* [§] entre des sommets n'étant *a priori* pas dans la même communauté.

**Définition 1** *La cohésion d'un ensemble $U \subseteq V$ est définie par $\mathcal{C}(G,U) = \frac{\triangle_{in}(G,U)}{\binom{|U|}{3}} \frac{\triangle_{in}(G,U)}{\triangle_{in}(G,U)+\triangle_{out}(G,U)}$ où $\triangle_{in}(G,U)$ est le nombre de triangles de G dans U et $\triangle_{out}(G,U)$ est le nombre de triangles pointant vers l'extérieur de U (*i.e.*, ayant deux sommets dans U et le troisième dans $V \backslash U$). Le premier facteur est la densité en triangles de U et le second facteur est la proportion de triangles ne pointant pas vers l'extérieur. Intuitivement, une communauté est un ensemble qui a une forte densité de triangles et qui en* coupe *peu.*

**Quelques propriétés [6]** Si on note $G_\triangle = (V, E_\triangle)$ le graphe sans lien faible alors $\forall U \subseteq V$, $\mathcal{C}(G,U) = \mathcal{C}(G_\triangle, U)$. On démontre aussi que si $U \subseteq V$ et $U' \subseteq V$ sont deux sous-ensembles déconnectés alors si $\mathcal{C}(U) < \mathcal{C}(U \cup U')$ on a $\mathcal{C}(U') \geq \mathcal{C}(U \cup U')$. On vérifie de façon analytique que si on fait varier la densité de liens d'un ensemble en interne et vers l'extérieur, la cohésion donne un meilleur score aux ensembles denses ayant une faible densité de liens vers l'extérieur, et reste ainsi en accord avec certaines vues classiques. Une dernière propriété montre qu'une large clique n'englobe pas forcément une plus petite en terme de score de cohésion si leur recouvrement est inférieur à un seuil fixé par la taille de la plus petite. Ces différentes propriétés [6] tendent à démontrer que la cohésion permet d'évaluer la qualité d'un ensemble *per se*.

## 3 Egomunautés

Nous présentons dans cette section une application de la mesure de cohésion afin de découvrir des *egomunautés*. Revenons brièvement sur les raisons de vouloir structurer un graphe en communautés. La première raison invoquée par NEWMAN de partitionner un réseau est une meilleure compréhension structurelle et intrinsèque du réseau pour en avoir une meilleur représentation. Dans le cas recouvrant, cet objectif doit se faire en étroite relation avec la sémantique même du réseau car potentiellement on peut obtenir $O(2^{2^n})$ communautés recouvrantes ce qui est loin de simplifier la problématique de visualisation. Nous prenons ici un double parti pris : *(i)* on s'intéresse aux réseaux dits *sociaux* car la cohésion trouve ses origines dans une analyse sociologique ; *(ii)* on place l'utilisateur au centre et non tout le graphe lui même. L'objectif n'est donc pas de structurer tout un réseau comme Facebook mais de proposer à l'utilisateur un moyen de structurer son voisinage. Cette structuration egocentrée de ses *amis* peut servir à mieux gérer la diffusion d'information : on n'affiche pas les mêmes photos à ses amis, ses collègues, sa famille, etc.

‡. « [...] social systems lacking in weak ties will be fragmented and incoherent. New ideas will spread slowly and subgroups separated by ethnicity, geography, or other characteristics will have difficulty reaching a *modus vivendi*. »

§. Ces arêtes sont des *weak ties* dans la terminologie employée par GRANOVETTER

**Algorithm 1** Algorithme glouton d'egomunautés.

**Require:** $G = (V, E)$ un graphe, $u \in V$. On note Γ(u) le voisinage de $u$.
  $\mathcal{E} \leftarrow \emptyset$
  $\mathcal{V} \leftarrow \Gamma(u)$
  **while** $\mathcal{V} \neq \emptyset$ **do**
    soit $v$ le sommet de plus fort degré dans $\mathcal{V}$
    initialiser ε nouvelle egomunauté contenant $\{u, v\}$
    soit $\mathcal{S}$ l'ensemble des voisins $v'$ de $u$ et $v$ tels que $\mathcal{C}(G, \varepsilon \cup \{v'\}) > \mathcal{C}(G, \varepsilon)$
    **while** $\mathcal{S} \neq \emptyset$ **do**
      ajouter à ε le sommet $v \in \mathcal{S}$ de plus fort $\triangle_{\text{in}}(\varepsilon \cup \{v\})$, (en cas d'ex aequo, prendre le plus fort $\triangle_{\text{out}}(\varepsilon \cup \{v\})$
      ajouter à $\mathcal{S}$ les voisins $v'$ de $v$ tels que $\mathcal{C}(G, \varepsilon \cup \{v'\}) > \mathcal{C}(G, \varepsilon)$
    **end while**
    retirer de $\mathcal{V}$ les sommets présent dans ε
    ajouter ε à $\mathcal{E}$
  **end while**
  **return** $\mathcal{E}$

L'idée sous-jacente de l'algorithme glouton est que chaque voisin de $u$ va être ajouté à un moment à au moins une egomunauté. Il est donc possible de partir de n'importe quel sommet *a priori* comme graine initiale. Néanmoins, en choisissant un sommet de plus fort degré (donc formant le plus de triangles avec $u$), on crée *de facto* un ensemble de sommets avec un faible $\triangle_{\text{in}}$ et un fort $\triangle_{\text{out}}$. Le choix glouton de la phase d'expansion se fait en prenant le parti de maximiser $\triangle_{\text{in}}$ tant que la cohésion croît. On ne cherche pas maximiser directement la cohésion car cela peut générer des cas où un sommet est choisi parce qu'il fait décroître $\triangle_{\text{out}}$ ce qui limite le choix de candidats à la prochaine étape. Notre stratégie peut se résumer en une croissance d'une egomunauté par l'ajout de noeuds internes et seulement ensuite par l'ajout des *frontières*. Le calcul de la cohésion peut être couteux à chaque étape à cause de l'énumération des triangles pour une egomunauté $e$ mais heureusement il est possible d'ajuster localement sa valeur pour l'ajout d'un sommet : $\mathcal{C}(G, e \cup \{v\}) = \frac{(\triangle_{\text{in}}(e)+I_v)^2}{(\binom{|e|+1}{3})(\triangle_{\text{in}}(e)+\triangle_{\text{out}}(e)+O_v)}$ où $I_v = \triangle_{\text{in}}(e \cup \{v\}) - \triangle_{\text{in}}(e)$ (resp. $O_v = \triangle_{\text{out}}(e \cup \{v\}) - \triangle_{\text{out}}(e)$) est le nombre de triangles internes (resp. externes) qui sont ajoutés par l'inclusion de $v$ à $e$.

# 4 Fellows : résultats & perspectives

Une expérimentation de grande envergure[¶] a été lancée sur Facebook le 8 février 2011 pour évaluer la qualité de la mesure de cohésion. En utilisant l'algorithme présenté ci-dessus, les *amis* d'un participant sont regroupés en groupes *cohérents* et lui sont ensuite présentés pour qu'il émette un jugement subjectif – une note de 1 à 4 étoiles – sur la qualité de chacun de ceux ci. Sur une période de deux mois, 2180 utilisateurs ont participé à l'expérience, générant un total de 57790 egomunautés. Il est important de noter que toutes les egomunautés calculées ont été présentées, indépendamment de leur cohésion – *i.e.* même des egomunautés de faible cohésion ont été présentées aux participants. On obtient que dans 31% des cas, les egomunautés calculées sont jugées très pertinentes, dans 22% des cas elles sont jugées pertinentes, dans 22% peu pertinentes et dans 25% non pertinentes. Il est important de rappeler ici que le but de l'expérience ne réside pas dans l'évaluation de l'algorithme mais dans celle de la mesure sous-jacente. Sur la Figure 1 on a représenté la moyenne $\bar{R} \in [0, 3]$ des notes obtenues par les egomunautés de cohésion $\mathcal{C}$. La corrélation de Spearman $\rho = 0.89$ avec un $p$-value égale à $1.7 \times 10^{-35}$, indique une corrélation monotone forte entre ces deux valeurs, et donc que la note moyenne croît lorsque la cohésion croît, et réciproquement. Si la corrélation n'est clairement pas affine, la corrélation de Pearson entre $\ln(\mathcal{C})$ et $\ln(\bar{R})$ vaut en revanche $r = 0.97$ ($p$-value $1.1 \times 10^{-61}$), ce qui indique une corrélation forte entre les deux log. Ainsi, $\mathcal{C} = a\bar{R}^k$ avec une erreur standard σ, ici, $a = 3$, $k = 0.34$ et $\sigma$ =0.0086. L'expérience Fellows a donc permis de valider la cohésion en tant que mesure de l'aspect communautaire d'un groupe de gens en exhibant une corrélation importante entre la métrique et la perception subjective des participants.

Nos travaux futurs concernent l'application aux réseaux pondérés (avoir un poids sur les échanges entre

¶. http://fellows-exp.com

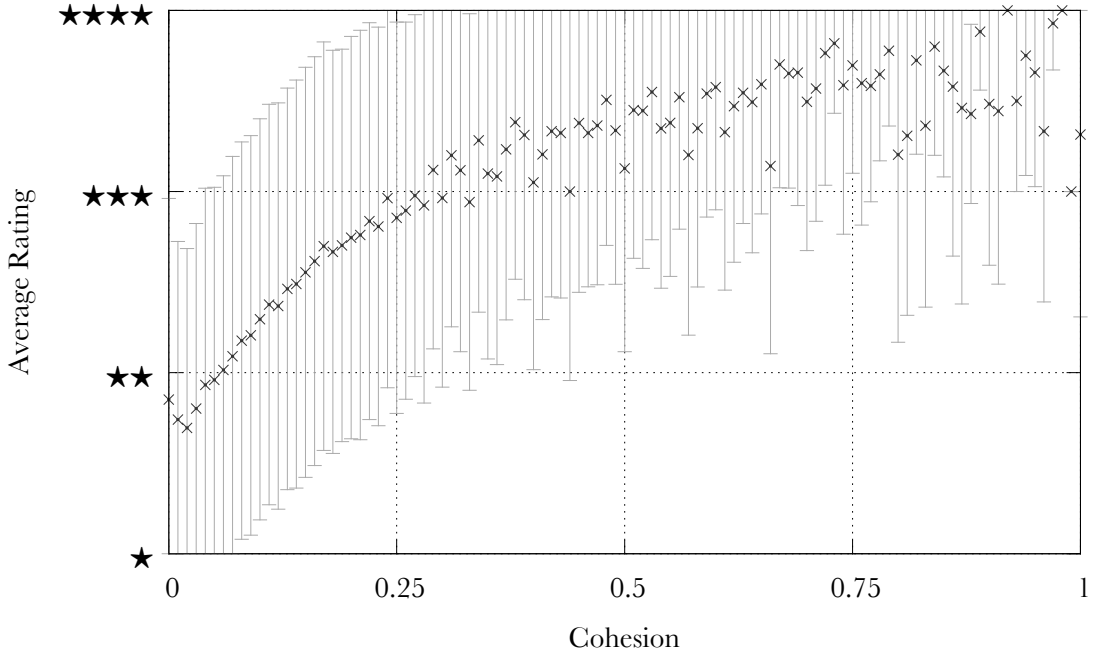


**FIGURE 1:** Note moyenne obtenue par valeur de cohésion arrondie à 1% dans l'expérience Fellows

amis permet de renforcer encore plus la notion de lien) et l'utilisation de la cohésion pour calculer des communautés recouvrantes non egocentrées. Nous travaillons par ailleurs sur l'exploitation des egomunautés pour inférer certaines méta-informations liées à l'individu au centre.

# Références

[1] V. D. Blondel, J. L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics : Theory and Experiment*, 2008.

[2] C. Castellano, F. Cecconi, V. Loreto, D. Parisi, and F. Radicchi. Self-contained algorithms to detect communities in networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 38(2) :311–319, 2004.

[3] A. Clauset. Finding local community structure in networks. *Physical Review E*, 72(2), aug 2005.

[4] G. Flake, S. Lawrence, C. Giles, and F. Coetzee. Self-organization of the web and identification of communities. *Communities*, 35(3) :66–71, 2002.

[5] S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5) :75–174, jan 2010.

[6] A. Friggeri, G. Chelius, and E. Fleury. Egomunities, Exploring Socially Cohesive Person-based Communities. Research Report RR-7535, INRIA, 02 2011.

[7] M. Granovetter. The Strength of Weak Ties. *Amer. J. of Sociology*, 78(6) :1360–1380, 1973.

[8] M. Latapy. *Grands graphes de terrain*. Habilitation à diriger des recherches, UPMC, 2007.

[9] T. Nepusz, A. Petróczi, L. Négyessy, and F. Bazsó. Fuzzy communities and the concept of bridgeness in complex networks. *Physical Review E*, 77(1) :16107, 2008.

[10] M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2) :26113, 2004.

[11] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043) :814–818, 2005.

[12] H. Shen, X. Cheng, and J. Guo. Quantifying and identifying the overlapping community structure in networks. *Journal of Statistical Mechanics : Theory and Experiment*, 2009 :P07042, 2009.

[13] Q. Wang and E. Fleury. Uncovering Overlapping Community Structure. In *Complex Networks*, 2010.