Structure multi-échelle de grands graphes de terrain

Thomas Aynaud — **Jean-Loup Guillaume**

LIP6 – CNRS – *Université Pierre et Marie Curie*
104 avenue du président Kennedy
75016 Paris, France

RÉSUMÉ. *La plupart des graphes de terrain peuvent être découpés en sous graphes denses appelés communautés et ces communautés sont généralement susceptibles d'être redécomposées. Cela permet de définir une hiérarchie de communautés imbriquées, appelée dendrogramme. Dans cet article nous analysons cette structure hiérarchique sur plusieurs graphes de terrain. Nous étudions tout d'abord la structure de cet arbre notamment sur les imbrications entre communautés. Puis, nous montrons que si le dendrogramme permet d'obtenir une information structurelle pertinente, la redécomposition excessive peut aboutir à des communautés vides de sens. Nous proposons plusieurs pistes pour éviter ce problème.*

ABSTRACT. *Most complex networks can be divided in dense sub-graphs called communities. These communities may also be divided recursively and this produce a hierarchical structure of communities, summarized in a tree called dendrogram. In this article we analyze this structure extracted from several complex networks. First we study the shape of the tree and how communities articulate themselves. Then we show that an excessive decomposition of communities can bring meaningless communities. We propose a couple of approaches to solve this problem.*

MOTS-CLÉS : *graphes de terrain, détection de communautés, structure multi-échelle, clustering hiérarchique*

KEYWORDS: *complex network, community detection, multi-scale, hierarchical clustering*

1re soumission à *Technique et science informatiques*, le 3 février 2009

Introduction

L'une des techniques éprouvée d'analyse des grands graphes de terrain est la décomposition en communautés. Intuitivement, on cherche des groupes de sommets de telle sorte que les nœuds d'un même groupe soient similaires, partagent quelque chose en commun. L'identification de ces groupes apporte un éclairage nouveau sur la structure du graphe et est importante dans de nombreux contextes. Elle pourrait, par exemple, être utilisée pour leur visualisation, pour développer une algorithmique tenant compte de leurs spécificités ou pour y faire de la fouille de données.

Une définition naturelle des communautés stipule qu'une communauté est dense, c'est-à-dire que les membres sont fortement connectés entre eux, et que, dans le même temps, ils sont peu liés à des membres en dehors de la communauté. Plus formellement, le problème de la détection de communautés revient à partitioner un graphe en sous groupes denses peu connectés entre eux. Ce problème est complexe (Brandes *et al.*, 2006) à résoudre de manière exacte et de nombreuses heuristiques existent pour y parvenir de façon plus ou moins satisfaisante (Blondel *et al.*, 2008; Fortunato *et al.*, 2008; Newman *et al.*, 2004; Pons *et al.*, 2006; Newman, 2006; Clauset *et al.*, 2004). Comparer différentes heuristiques n'est pas chose aisée et la qualité d'une partition est souvent évaluée par une fonction de qualité, la modularité. Cette fonction qui permettait à l'origine de comparer différentes partitions, et par là même différents algorithmes, est devenue une fonction objectif à maximiser. La complexité temporelle est aussi un critère très important dès lors que l'on souhaite calculer des communautés sur des graphes ayant des millions de sommets.

Quelle que soit la technique utilisée, on peut trouver plusieurs niveaux de hiérarchie : les communautés sont elles-mêmes composées de sous-communautés. Ainsi, si l'on considère un réseau social, on peut par exemple imaginer une communauté de toutes les relations d'une personne. Elle-même pourrait être découpée en un cercle familial, un cercle d'amis et un cercle de travail. Le cercle familial peut à nouveau sans doute se découper en famille du côté maternel et du côté paternel, etc.

L'objectif est alors d'obtenir la structure hiérarchique complète et non pas seulement la partition maximisant la modularité. Ceci peut être fait en appliquant à nouveau de manière récursive une technique de décomposition sur les communautés puis sur les sous-communautés obtenues, etc. Cet article a pour but de décrire cette structure hiérarchique.

Dans la suite de cet article, nous commençons par décrire plus précisément la méthode utilisée pour construire effectivement cet arbre en introduisant les concepts classiques. Ensuite, nous présentons plusieurs propriétés de l'arbre afin de mettre en évidence la structure multi-échelle. Enfin, nous montrons que la structure ainsi obtenue naïvement décompose des communautés lorsqu'il n'y a pas lieu et que, s'il est nécessaire d'étudier la structure hiérarchique, il ne faut pas vouloir trouver des sous-communautés là où il n'y en a pas.

1. Méthode de décomposition

La décomposition d'un graphe en communautés consiste à partitionner l'ensemble des nœuds du graphe de sorte que les nœuds soient regroupés en zones denses mais avec peu de liens vers l'extérieur. Pour évaluer la qualité d'une partition, on utilise généralement une fonction donnant un score à une partition et qui capture de manière formelle la définition informelle donnée plus haut. Ensuite, il reste à maximiser cette fonction pour trouver la meilleure partition. Il existe plusieurs fonctions de qualité, la plus utilisée étant la *modularité* définie dans (Girvan *et al.*, 2002). Pour une partition π de l'ensemble des nœuds, en notant L le nombre de liens du graphe, d_s le degré total d'une partie s et l_s le nombre de liens à l'intérieur d'une partie s, la modularité Q est définie par :

$$Q\left(\pi\right) = \sum_{s \in \pi} \left(\frac{l_s}{L} - \left(\frac{d_s}{2L} \right)^2 \right)$$

Cette grandeur peut être vue comme la somme, sur toutes les communautés, des différences entre la proportion de liens à l'intérieur de la communauté s (soit $\frac{l_s}{L}$) et la proportion de liens que devraient avoir une communauté dans un graphe aléatoire de même taille (soit $\left(\frac{d_s}{2L}\right)^2$). Une partition sera donc bonne s'il y a nettement plus de liens à l'intérieur des communautés que ce à quoi on s'attendait et, par conséquent, moins de liens hors des communautés. En pratique, trouver une partition ayant une forte modularité est un problème difficile (Brandes *et al.*, 2006) et de nombreuses heuristiques ont donc été proposées, voir par exemple (Fortunato *et al.*, 2008; Newman *et al.*, 2004; Pons *et al.*, 2006; Newman, 2006; Clauset *et al.*, 2004). Dans ce qui suit, nous avons utilisé un algorithme existant décrit dans (Blondel *et al.*, 2008) qui a l'avantage d'avoir une complexité temporelle faible tout en produisant des partitions de très bonne qualité.

Il a été montré que l'optimisation de la modularité a un effet non souhaité qui est de défavoriser les petites communautés, de tailles inférieures à $\sqrt{2L}$, lesquelles ont souvent (mais pas toujours) intérêt à fusionner pour améliorer la modularité (Fortunato *et al.*, 2007). Ce problème a reçu le nom de "résolution limite". Les communautés qui maximisent la modularité sont donc certainement des regroupements de communautés plus petites qui méritent d'être étudiées.

Afin d'étudier la structure multi-échelle des graphes de terrain et notamment des inclusions de communautés, nous avons donc utilisé une approche récursive consistant, à partir d'un graphe donné, à le décomposer en communautés, puis en sous-communautés, et ainsi de suite. Nous commençons par trouver une partition ayant la meilleure modularité possible et, ensuite, le sous-graphe correspondant à chaque communauté est extrait et la procédure de décomposition est appliquée à nouveau sur ce sous-graphe. Ce procédé est répété récursivement sur chaque communauté ou sous-communauté, jusqu'à obtenir des communautés qui ne peuvent plus être décomposées.

Ces communautés non décomposables ne sont pas forcément des sommets seuls, en effet certains graphes ne peuvent pas être décomposés en communautés sans obtenir une modularité plus faible que si on les gardait entiers, voir Figure 1 pour quelques exemples. Il faut aussi remarquer que, bien que certains graphes soient décomposables, vouloir décomposer à tout prix dès lors qu'il y a un gain de modularité peut amener à des communautés peu ou pas naturelles. Ceci sera évoqué plus bas.

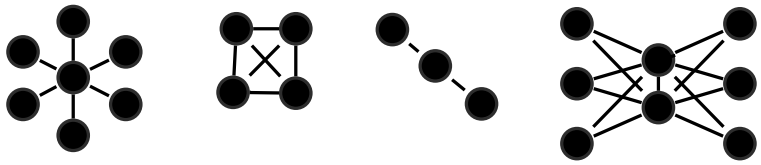

Figure 1. *Exemples de graphes pour lesquels toute tentative de décomposition en communautés fournit une modularité inférieure à celle du graphe considéré comme une seule communauté.*

Les résultats de ces décompositions successives peuvent être synthétisés dans un arbre, généralement appelé dendrogramme. Chaque sommet s de l'arbre correspond à une communauté calculée par l'algorithme, c'est-à-dire à un ensemble de sommets. Les fils d'un sommet s sont les sous-communautés de la communauté s telles que calculées par l'algorithme. On complète l'arbre avec un racine qui correspond au graphe entier et, de plus, les communautés non décomposables ont comme nœuds fils les sommets du graphe qui les composent. les seules feuilles de l'arbre sont donc les sommets du graphe original. On appelle niveau d'un sommet sa profondeur dans l'arbre. Ainsi la racine est au niveau 0, les communautés calculées initialement par l'algorithme sont au niveau 1, les sous-communautés au niveau 2, etc. La figure 2 présente un graphe et l'arbre obtenu avec la méthode de décomposition.

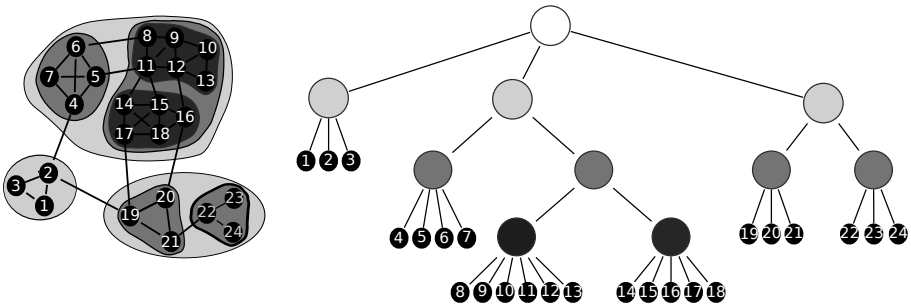

Figure 2. *Graphe avec représentation de la décomposition niveau par niveau et de l'arbre obtenu. Les feuilles de l'arbre sont les sommets du graphe initial.*

Les travaux présentés plus loin ont été réalisés sur quatre graphes de terrains. La

plupart des résultats étant similaires sur ces différents graphes, nous présentons les résultats sur le premier :

– *webndu* est un graphe du web pour lequel chaque nœud est une page web, et il y a un lien entre deux nœuds s'il y a un lien hypertexte de la première page vers la seconde. Le graphe considéré est une carte complète du domaine .nd.edu obtenue en 1999 et qui contient 325729 pages et plus d'un million de liens (Albert *et al.*, 1999). Initialement orienté, on l'a considéré comme un graphe non orienté ;

– *arxiv* est un graphe de cosignatures entre auteurs d'articles de recherche dans le domaine de la physique extrait de la base de données d'articles de recherche *arxiv.org* (Newman, 2000). A chaque auteur correspond un nœud du graphe, et il y a un lien entre deux auteurs s'ils ont cosigné un ou plusieurs articles [1] ;

– les deux derniers graphes sont extraits d'une trace d'échanges pair à pair sur le réseau eDonkey (Latapy *et al.*, 2008). Le premier graphe relie les clients du réseau pair à pair recherchant les mêmes fichiers, et l'autre les fichiers demandés par les mêmes personnes.

2. Analyse de l'arbre

L'arbre de décomposition s'avère en général peu profond. Parmi nos graphes de test, nous obtenons au maximum un arbre de profondeur 11, ce qui représente tout de même autant de niveau d'imbrication de communautés. Cette profondeur ne concerne cependant que très peu de nœuds. La figure 2 à gauche représente le nombre d'éléments (sommets internes à l'arbre et feuilles) se trouvant à une profondeur donnée. On constate une rapide augmentation du nombre d'éléments, laissant supposer que beaucoup des nœuds du graphe sont classés bien avant la dernière décomposition. Si des feuilles sont à profondeur i, c'est qu'elles sont la terminaison de l'imbrication de $i-1$ communautés, la dernière n'étant pas décomposable. Les arbres contiennent des feuilles dès le niveau 2 ce qui signifie que même parmi les communautés du premier niveau, celles qui maximisent la modularité, il en existe qui ne peuvent pas être re-décomposées. Approximativement 80% des feuilles sont à un niveau inférieur ou égal à 5, il en résulte qu'il y a en général moins de 4 niveaux d'imbrication de communautés pertinents.

La figure 4 présente la distribution du nombre de sous communautés, c'est-à-dire la distribution du nombre de fils pour tous les sommets de l'arbre. La distribution est très hétérogène et cette disparité s'explique en grande partie par le fait que la taille des communautés, dont la distribution est représentée sur la figure 5 à gauche, est aussi très hétérogène. Or, le nombre de sous communautés d'une communauté est lié à sa taille comme l'indique la figure 5 à droite qui représente la corrélation entre la taille d'une communauté et le nombre de sous communautés.

1. Les nœuds de degré 1 ont été supprimés dans ce graphe, ce qui explique la différence avec les données de l'article original, mais n'a pas d'incidence sur les calculs de communautés et donc sur les analyses faites plus bas.

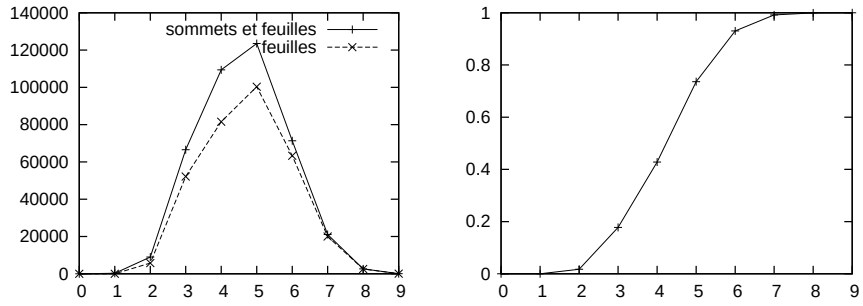

Figure 3. *A gauche, nombre d'éléments (feuilles et sommets internes) et nombre de feuilles par niveau dans l'arbre. A droite, proportion de feuilles à profondeur inférieure ou égale à i.*

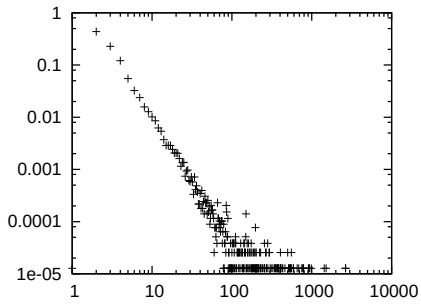

Figure 4. *Distribution du nombre de sous communautés.*

L'arbre de décomposition est donc déséquilibré, mais ce déséquilibre est le reflet de la structure du graphe décomposé. Il semble n'y avoir qu'un petit nombre de niveaux d'imbrication de communautés, nombre qui dépend de la taille de la première communauté.

3. La structure communautaire

Chaque niveau dans l'arbre définit aussi une partition du graphe initial. Ainsi, le premier niveau de l'arbre correspond aux communautés maximisant la modularité, le second niveau correspondant à l'ensemble des sous-communautés des précédentes, etc.

La figure 6 (gauche) présente la modularité des partitions niveau par niveau et

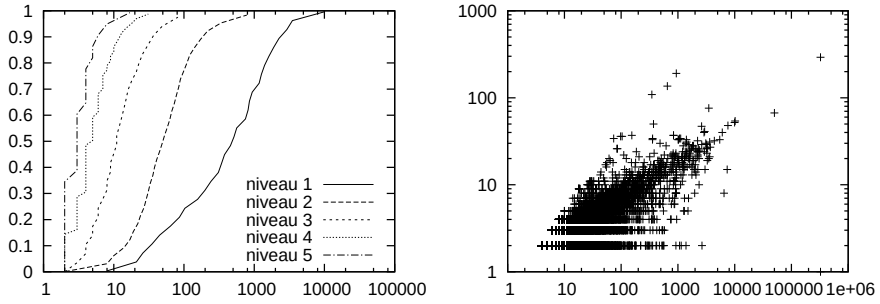

Figure 5. *A gauche, distribution cumulée des tailles des communautés aux différents niveaux. A droite, corrélation taille - nombre de sous communautés pour toutes les communautés de l'arbre. Les communautés qui ne peuvent pas être décomposées ont été exclues car leur taille est exactement égale à leur nombre de sous-communautés par définition.*

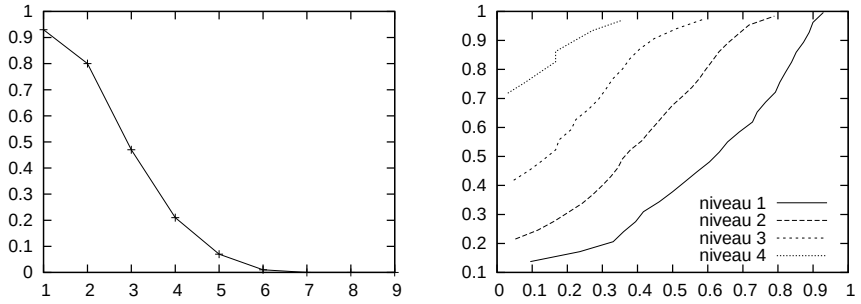

Figure 6. *A gauche, modularité de la décomposition niveau par niveau. A droite, distribution cumulée de la modularité aux différents niveaux.*

on constate qu'elle diminue régulièrement et assez vite passé le deuxième niveau, en raison notamment du problème de résolution limite évoqué plus haut qui défavorise les petites communautés. Mais en plus de cela, d'autres problèmes se greffent :

– les nœuds isolés, qui ne font plus partie d'un regroupement mais sont les composants d'une communauté non décomposable, ont un effet strictement négatif sur la modularité car il n'y a alors aucun lien interne et que des liens sortants ;

– la modularité est aussi connue pour ne pas donner de bons scores aux partitions mélangeant des parties de tailles très différentes (Fortunato *et al.*, 2007).

Cette modularité globale basse pour les partitions à partir du troisième niveau n'est donc pas forcément le signe qu'il n'y a plus de communautés significatives à partir de là, mais simplement que ces éventuelles communautés sont noyées dans la masse.

En analysant plus en détail les communautés à chaque niveau, on constate en effet que certaines communautés de tailles non négligeables ont effectivement une structure communautaire interne. En particulier, on peut considérer de manière indépendante chaque communauté de chaque niveau, puis associer un sous-graphe à chaque communauté que l'on peut ensuite décomposer pour calculer sa modularité. La figure 6 (droite) représente la distribution cumulée de la modularité niveau par niveau. Chaque courbe correspond à un niveau dans l'arbre. On y lit en abscisse les valeurs prises par la modularité et en ordonnée la proportion de communautés qui sont décomposées en une partition de modularité inférieure : si moins de 40% des communautés ont une modularité inférieure à $0,5$ au niveau 1, elles sont presque 70% au niveau 2 et plus de 90% au niveau 3. Ainsi, plus on s'enfonce dans l'arbre, plus la majorité des communautés ont une faible modularité, mais malgré tout, on trouve toujours quelques communautés avec une modularité élevée même dans les niveaux bas de l'arbre. Mais modularité élevée ne signifie pas nécessairement structure modulaire comme nous allons maintenant le voir.

Nous avons extrait la plus grande communauté, puis l'avons décomposée et avons à nouveau sélectionné la plus grande, et avons recommencé ainsi récursivement jusqu'à ce qu'on n'ait plus qu'une seule communauté. La modularité obtenue à chaque étape est représentée sur la figure 7 (gauche). On constate que la modularité reste élevée aux premiers niveaux mais décroît assez rapidement par la suite, alors que le sous-graphe associé reste de taille non négligeable comme indiqué sur la figure 7 (droite).

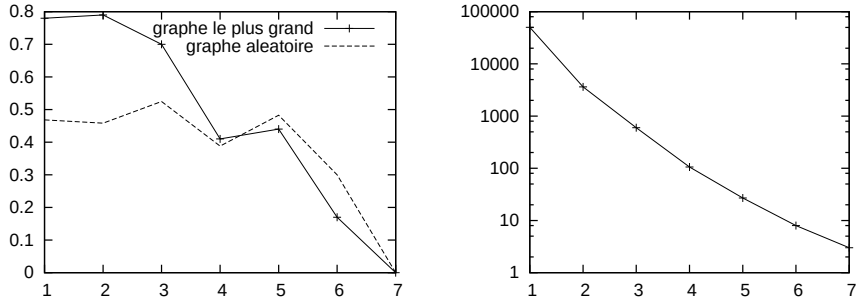

Figure 7. *A gauche, la modularité à l'intérieur de la plus grande communauté extraite récursivement. A droite, le nombre de nœuds dans ces communautés.*

A titre comparatif, la modularité moyenne sur des graphes aléatoires de même taille et de même distribution de degré et générés selon la méthode décrite dans (Viger *et al.*, 2005) est indiquée sur la figure 7. La modularité étant calculée en comparant

une proportion de liens internes à une communauté à ce que l'on aurait dans un graphe aléatoire de même taille, on espère que la modularité reste élevée et supérieure à celle d'un graphe aléatoire si le graphe est réellement modulaire. C'est effectivement le cas pour les premiers niveaux, ce qui montre qu'il y a bien une structure communautaire à plusieurs échelles. Par contre, ce n'est plus le cas pour les derniers niveaux où la modularité trouvée est similaire à ce que l'on trouve sur un graphe aléatoire, signe que l'information contenue dans ces sous-graphes n'est plus vraiment pertinente.

La comparaison peut aussi se faire directement grâce à des méthodes formelles permettant de prédire la modularité obtenue sur un graphe aléatoire (Reichardt *et al.*, 2006). Dans ce cas aussi, il est clair qu'un graphe qui n'est pas clairement plus modulaire qu'un graphe aléatoire de même taille n'est tout simplement pas modulaire.

Encore une fois, il semblerait donc que l'on identifie bien une sous structure communautaire. La modularité globale n'est pas très élevée, mais cette grandeur est limitée et n'est pas adaptée pour juger des communautés de tailles très variées. Il n'est néanmoins pas pertinent de prolonger trop profondément la détection de communautés car on décompose alors des graphes qui n'ont sans doute plus de structure communautaire.

4. Conclusion

Nous avons ici présenté quelques résultats sur la structure modulaire multi-échelle des grands graphes de terrain basés sur plusieurs graphes réels. Sur tous les graphes, les résultats sont similaires et mettent en évidence une structure multi-échelle dans laquelle les communautés peuvent effectivement être décomposées en sous-communautés qui ont du sens.

Cette structure multi-échelle est complexe et a de nombreuses propriétés très hétérogènes, notamment la taille des communautés ou le nombre de sous-communautés. De plus, dès les premiers niveaux, on trouve très rapidement des communautés non décomposables. L'étude de cet objet qui décrit le graphe de manière structurelle est donc à lui seul un problème intéressant.

Le problème de résolution limite exposé dans (Fortunato *et al.*, 2007) incite à ne pas considérer uniquement la partition d'un graphe qui maximise la modularité car cette partition est généralement peu favorable aux communautés de petites tailles. Il est donc nécessaire d'explorer la structure hiérarchique complète des graphes étudiés. Au contraire, nous avons montré que s'il est possible de redécomposer de manière récursive les communautés, on arrive rapidement soit à des communautés non décomposables, soit à des communautés qui ne sont plus modulaires et n'ont donc pas de raison d'être décomposées. Il faut donc trouver un juste équilibre entre les problèmes de résolution limite et de décomposition excessive.

La technique utilisée pour mettre en évidence le problème de la décomposition excessive, à savoir la comparaison à des graphes aléatoires de même taille, peut cer-

tainement être utilisée comme critère d'arrêt supplémentaire lors de la décomposition : tout graphe trop peu modulaire ne doit plus être décomposé.

Il ne faut pas non plus perdre de vue qu'une structure hiérarchique est limitée. Il n'y a pas de possibilité de communautés recouvrantes : un nœud fait partie d'une et une seule série de communautés imbriquées. Si l'on considère deux communautés voisines, celles-ci sont susceptibles de contenir chacune une partie d'une communauté à cheval entre elles, sans doute moins modulaire mais potentiellement intéressante. En extrayant une communauté, on perd toute cette information et on ne trouvera pas certaines sous communautés. Intégrer le recouvrement d'une manière ou d'une autre dans le dendrogramme permettrait d'affiner la compréhension que l'on peut avoir de la structure du graphe.

5. Bibliographie

Albert R., Jeong H., Barabasi A.-L., « The diameter of the world wide web », *Nature*, vol. 401, p. 130, 1999.

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. 2008, n° 10, p. P10008 (12pp), 2008.

Brandes U., Delling D., Gaertler M., Goerke R., Hoefer M., Nikoloski Z., Wagner D., « Maximizing Modularity is hard », *ArXiv Physics e-prints*, August, 2006.

Clauset A., Moore C., Newman M. E. J., « Finding community structure in very large networks », *Physical Review E*, vol. 70, n° 6, p. 066111-+, December, 2004.

Fortunato S., Barthélemy M., « Resolution limit in community detection », *Proceedings of the National Academy of Sciences*, vol. 104, n° 1, p. 36-41, 2007.

Fortunato S., Castellano C., « Community Structure in Graphs », *Encyclopedia of Complexity and System Science*, 2008.

Girvan M., Newman M. E. J., « Community structure in social and biological networks », *Proceedings of the National Academy of Science*, vol. 99, p. 7821-7826, June, 2002.

Latapy M., Aidouni F., Magnien C., « Capture à grande échelle de traffic eDonkey », *Algotel*, 2008.

Newman M. E. J., The Structure of Scientific Collaboration Networks, Working Papers n° 00-07-037, Santa Fe Institute, July, 2000.

Newman M. E. J., « Finding community structure in networks using the eigenvectors of matrices », *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, vol. 74, n° 3, p. 036104, 2006.

Newman M. E. J., Girvan M., « Finding and evaluating community structure in networks », *Phys. Rev. E*, vol. 69, n° 2, p. 026113, Feb, 2004.

Pons P., Latapy M., « Computing Communities in Large Networks Using Random Walks », *J. Graph Algorithms Appl.*, vol. 10, n° 2, p. 191-218, 2006.

Reichardt J., Bornholdt S., « When are networks truly modular ? », *Physica D Nonlinear Phenomena*, vol. 224, p. 20-26, December, 2006.

Viger F., Latapy M., « Efficient and simple generation of random simple connected graphs with prescribed degree sequence », *The Eleventh International Computing and Combinatorics Conference, Aug. 2005, kumming*, 2005.