Caractérisation de la structure communautaire d'un grand réseau social

Ivan Keller — Emmanuel Viennet

Université Paris 13, Sorbonne Paris Cité
Laboratoire de Traitement et Transport de l'Information (L2TI)
F-93430, Villetaneuse, France
keller.ivan@gmail.com, emmanuel.viennet@univ-paris13.fr

RÉSUMÉ. *Comprendre la structure communautaire enfouie dans les réseaux complexes est un problème difficile tant du point de vue théorique que pratique. Malgré le nombre important d'algorithmes de détection de communautés proposés dans la dernière décennie, il n'existe toujours pas de méthode satisfaisante pour déterminer si un réseau donné possède ou non une structure communautaire. Dans cet article, nous proposons un nouveau critère basé sur l'étude de la formation de communautés consensuelles obtenues en exécutant plusieurs fois un algorithme non déterministe. En testant sur des graphes synthétiques dont on connaît la structure communautaire et sur plusieurs graphes réels, nous montrons que ces graphes peuvent être classés en plusieurs catégories en fonction de la dynamique du processus de formation des communautés consensuelles, selon la présence plus ou moins marqué d'une structure modulaire.*

ABSTRACT. *Understanding the community structure in graphs arising from complex network is an important and difficult problem, both from theoretical and practical points of views. Although a lot of community detection algorithms have been proposed in the last decade, there is still no satisfactory way to determine if a given network possess or not a community structure, that is, its nodes can be partitioned in well separated clusters. In this paper, we propose a new criterium based on the study of the formation of consensual communities obtained by running several times a non deterministic algorithm. By testing on synthetic benchmarks (with known structure) and on several real world networks, we show that the graphs can be categorized in several classes according to the dynamic of the consensual communities formation process. This result is promising to derive new approaches to characterize the modular structure of graphs.*

MOTS-CLÉS : *réseaux complexes, communautés, classification de consensus*

KEYWORDS: *complex networks, community structure, consensus clustering*

1. Introduction

La détection de communautés dans les grands réseaux de terrain (réseaux sociaux ou autres) a fait l'objet d'intenses recherches ces dernières années ([NEW 04, FOR 10]) et des algorithmes efficaces ont été proposés : plusieurs sont basés sur des heuristiques de maximisation gloutonne d'un critère de qualité comme la modularité, le problème général étant NP-complet.

Plus récemment, il a été montré que la modularité posait plusieurs problèmes empêchant de considérer sa valeur sur une partition optimale d'un graphe comme un critère de caractérisation de la présence d'une structure communautaire dans ce graphe : les méthodes de détection de communautés implémentant l'optimisation de ce critère ne permettent pas la détection de communautés en dessous d'une certaine taille relative à l'ordre du graphe (problème de limite de résolution ([FOR 07, LAN 11]) ; aussi, dans des graphes creux à la topologie similaire, le maximum de la modularité augmente avec l'ordre du graphe (([GOO 10]) ; de plus, il est possible de trouver des partitions de modularité élevé dans des graphes ne présentant pas de structure communautaire ([GUI 04]) ; enfin, de nombreuses solutions – des partitions différentes ayant toutes des valeurs proches de la modularité maximale – peuvent être produites par l'algorithme, révélant la présence d'un plateau de maxima locaux dans le paysage de la modularité ([GOO 10]). Cette inconsistance dans les résultats des algorithmes a incité certaines équipes à chercher des communautés consensuelles (ou "cœurs de communautés"), groupes de sommets fréquemment classés ensemble dans différentes solutions ([SEI 12a, LAN 12, KWA 11]). Ces "cœurs" stables sont ce que l'on appelle des "formes fortes" dans le contexte de la classification classique ([DID 71]).

L'application de ces méthodes aux grands réseaux (plusieurs dizaines de millions de nœuds) nécessite certes d'important temps de calcul mais pourrait nous éclairer sur leur structure. En effet, sur certains graphes aléatoires, on observe dans la formation des cœurs plusieurs régimes et en particulier un phénomène de transition de phase dont la présence pourrait indiquer s'il existe ou non une structure en communautés. On retrouve sur des graphes réels les mêmes phénomènes. En se basant sur l'étude de ces cœurs, notre étude propose un critère caractérisant la présence ou non d'une structure de communautés dans des graphes réels.

2. Partitions consensuelles et cœurs de communautés

Dans cette étude nous considérons des communautés non recouvrantes dans les graphes : un sommet appartient toujours à une communauté et à une seule. Les communautés d'un graphe constituent donc une *partition* de l'ensemble de ses sommets. Nous sommes conscients que cette approche est restrictive tant l'idée de communautés se chevauchant semble naturelle dans bien des cas comme celui des réseaux sociaux. Néanmoins, la difficulté accrue du problème de la détection de communautés recouvrantes empêche d'y faire des avancées aussi décisives que celles qu'a connu celui de la recherche de partitions de communautés non recouvrantes.

En analyse des données, la recherche d'une partition consensuelle (on parle aussi "d'agrégation de classifications" et en anglais de *consensus clustering* ([LEC 07]) parmi plusieurs partitions d'un ensemble est un problème classique quoique non dépourvu de difficultés. Une solution consiste à considérer la matrice de consensus dont les termes indiquent, pour chaque paire d'éléments de l'ensemble, la fréquence d'appartenance de ces deux éléments à la même partie dans les différentes partitions considérées.

L'application de cette idée dans notre contexte permet de définir de la manière suivante les cœurs de communautés : un algorithme non-déterministe de partitionnement, ici l'algorithme de Louvain ([BLO 08]), est exécuté un certain nombre de fois sur le graphe étudié, produisant autant de partitions. En notant n le nombre de sommets du graphe, la matrice de consensus C est la matrice $n \times n$ dont les éléments C_{ij} correspondent pour chaque paire de sommet (i, j) du graphe au nombre de fois où les deux sommets i et j ont été classifiés ensemble, divisé par le nombre total de partitions afin d'obtenir une fréquence. C peut être considérée comme la matrice d'adjacence d'un graphe de consensus ayant les mêmes sommets que le graphe d'origine et dont les liens expriment les fréquences de co-occurrence des paires de sommets dans les partions considérées. Dans le graphe de consensus, les sommets souvent classifiés ensemble parmi les partitions se retrouvent liés par des liens fort tandis que ceux qui sont rarement identifiés dans la même communauté sont liés par des liens faibles. On se donne alors un seuil $\alpha \in [0; 1]$ et on supprime les liens inférieurs à ce seuil dans le graphe de consensus. Cela à pour effet de déconnecter les communautés entre elles.

Les *cœurs de communautés* sont définis comme étant les composantes connexes de ce graphe de consensus seuillé. À cette étape, plusieurs voies d'études se présentent : nous pouvons réitérer la procédure sur ce nouveau graphe dont les liens sont pondérés par leur fréquence de co-occurence dans des communauté et ce, jusqu'à convergence. Cela revient à augmenter itérativement la densité des liens intra-communautaires, ce qui renforce la cohésion au sein des communautés, tout en diminuant celle des liens inter-communautaires, ce qui renforce la séparation des communautés, et ainsi facilite la tâche à l'algorithme de détection de communautés dont les résultats deviennent alors consistants. C'est ce qu'ont fait récemment [LAN 12] ainsi que [KWA 11]. A l'inverse, nous pouvons considérer les cœurs pour toutes les valeurs disponibles de du seuil α et s'intéresser à leur formation lorsqu'il varie.

3. Caractérisation de la structure communautaire d'un graphe

A la suite des travaux de M. Seifi ([SEI 12a, SEI 12b]), nous nous intéressons aux caractéristiques des cœurs lorsqu'on fait varier le seuil α défini ci-dessus. Lorsque qu'il augmente, de plus en plus de liens sont supprimés dans le graphe de consensus et les cœurs de communautés se multiplient en se fragmentant. On observe que la vitesse de fragmentation des cœurs lorsque α varie va dépendre de la présence plus ou moins marquée de communautés. Afin de comprendre le phénomène et d'essayer d'en tirer des conclusions exploitables, nous avons procédé à diverses expériences sur

différents types de graphes artificiels ou réels, comportant ou ne comportant pas de communautés avérées.

Nous nous sommes particulièrement intéressés au modèle de graphes synthétiques du benchmark LFR de Lancichinetti-Fortunato-Radicchi ([LAN 08]) qui permet de construire des graphes dont les communautés sont plus ou moins marquées selon un paramètre μ que l'on contrôle. Nous avons généré plusieurs de ces graphes en ne faisant varier que ce paramètre de manière à rendre progressivement incertaine la netteté des communautés. Pour chacun des graphes nous analysons la formation des cœurs en fonction de α en mesurant le nombre de cœurs et en visualisant leurs tailles. Lorsque μ atteint une certaine valeur, les communautés deviennent difficile à détecter et les partitions retournées par l'algorithme de détection de communautés diffèrent de plus en plus. La partition consensuelle devient très instable selon le seuil α choisi et on observe alors une transition de phase dans la formation des cœurs : sur un court intervalle de valeurs pour α, le nombre de cœurs explose en passant d'un état trivial composé d'un cœur géant à un état où il y a de très nombreux petits cœurs dont beaucoup de cœurs singletons ne comportant qu'un sommet.

Nous avons appliqué la méthode décrite ci-dessus à huit graphes de terrain issus de données provenant de divers domaines et utilisés dans la littérature. Pour certains de ces graphes "réels" (par opposition au caractère artificiel des graphes du modèle LFR), on dispose d'information a priori sur la structure communautaire. Les observations sur la formation des cœurs pour les graphes du modèle du benchmark LFR restent valables lorsqu'on considère des graphes réels et semblent refléter la présence plus ou moins nette de communautés détectables.

Nous proposons un indice basé sur la vitesse de fragmentation des cœurs de communautés lorsque le seuil α varie pour caractériser les différents régimes observés. D'après nos expériences, il pourrait nous indiquer si un graphe présente ou non une structure communautaire.

Nous avons effectué des mesures préliminaires de la sensibilité de l'indice à d'autres paramètres comme l'ordre du graphe, la taille et le nombre de communautés. Ces expériences semblent indiquer une meilleure robustesse que d'autres indices consistant à mesurer la variabilité des partitions avec des indices classiques de similarité.

Nous travaillons actuellement sur l'utilisation de ces résultats pour la caractérisation de grands réseaux réels pour lesquels l'existence d'une structure communautaire est incertaine, la seule valeur de la modularité ne permettant pas une caractérisation suffisante. En particulier, nous traiterons le cas du réseau social Skyrock comportant plusieurs millions de sommets auquel nous avons eu accès dans le cadre du projet ANR CEDRES/ExDEUSS).

Remerciements

Ce travail a été financé par les projets ANR ExDEUSS et DGCIS CEDRES, du pôle Cap Digital.

4. Bibliographie

[BLO 08] BLONDEL V., GUILLAUME J., LAMBIOTTE R., MECH E., « Fast unfolding of communities in large networks », *Journal of Statistical Mechanics : Theory and Experiment*, vol. P10008, 2008, p. 1742-5468.

[DID 71] DIDAY E., « La méthode des nuées dynamiques », *Statistiques Appliquées*, vol. 2, n° 19, 1971, p. 19–34.

[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, National Academy of Sciences.

[FOR 10] FORTUNATO S., « Community detection in graphs », *Physics Reports*, vol. 486, n° 3–5, 2010, p. 75–174.

[GOO 10] GOOD B., MONTJOYE Y. D., CLAUSET A., « Performance of modularity maximization in practical contexts », *Physical Review E*, vol. 81, n° 4, 2010, page 046106, APS.

[GUI 04] GUIMERA R., SALES-PARDO M., AMARAL L., « Modularity from fluctuations in random graphs and complex networks », *Physical Review E*, vol. 70, n° 2, 2004, page 025101, APS.

[KWA 11] KWAK H., EOM Y.-H., Y. Y. C., JEONG H., MOON S., « Consistent community identification in complex networks », *J. Korean Phys. Soc.*, vol. 59, 2011, p. 3128–3132.

[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, page 046110, APS.

[LAN 11] LANCICHINETTI A., FORTUNATO S., « Limits of modularity maximization in community detection », *Physical Review E*, vol. 84, n° 6, 2011, page 066122, APS.

[LAN 12] LANCICHINETTI A., FORTUNATO S., « Consensus clustering in complex networks », *Scientific Reports*, vol. 2, 2012, page 336, Macmillan Publishers Limited.

[LEC 07] LECLERC B., « Consensus from Frequent Groupings », BRITO P., CUCUMEL G., BERTRAND P., CARVALHO F., Eds., *Selected Contributions in Data Analysis and Classification*, Studies in Classification, Data Analysis, and Knowledge Organization, p. 317-324, Springer Berlin Heidelberg, 2007.

[NEW 04] NEWMAN M. E. J., GIRVAN M., « Finding and evaluating community structure in networks », *Physical Review E*, vol. 69, n° 2, 2004, p. 026113+.

[SEI 12a] SEIFI M., GUILLAUME J.-L., JUNIER I., ROUQUIER J.-B., ISKROV S., « Stable community cores in complex networks », *3rd Workshop on Complex Networks (CompleNet 2012)*, 2012.

[SEI 12b] SEIFI M., « Cœurs stables de communautés dans les graphes de terrain », PhD thesis, Université Paris 6, mars 2012.