Modularité asymptotique de quelques classes de graphes

Fabien de Montgolfier, Mauricio Soto, Laurent Viennot

Modularité asymptotique de quelques classes de graphes

Fabien de Montgolfier, Mauricio Soto et Laurent Viennot

fm@liafa.jussieu.fr mausoto@liafa.jussieu.fr Laurent.Viennot@inria.fr
Équipe-Projet GANG entre l'INRIA Paris-Rocquencourt et le LIAFA, UMR 7089 CNRS - Université Paris Diderot.

De nombreuses disciplines scientifiques font appel au *clustering* pour l'analyse de leurs réseaux d'interaction. Pami les très nombreux algorithmes existants, toute une famille utilise la *modularité* de Newman-Girvan comme objectif à maximiser, et cette valeur est devenue un paramètre de graphe standard. Dans cette étude nous prenons à rebours l'approche empirique usuelle et nous posons la question théorique de la modularité de *classes* de graphes. Nous montrons que des classes très régulières et sans "*clusters*" naturels (grilles, hypercubes,...) ont une modularité asymptotiquement 1 (le maximum possible), soit bien plus que les valeurs usuelles des données "bien clusterisées". Nous montrons que sous réserve d'une condition sur le degré maximum, les arbres ont aussi modularité asymptotique 1. Résultat qui nous permet de fournir une borne inférieure pour la modularité des graphes connexes peu denses et de certains *power-law graphs*, qui ont modularité asymptotique au moins *2/degré moyen*, ainsi qu'un algorithme garantissant cette performance.

1 Introduction

Le *clustering*, action qui consiste à répartir les sommets d'un graphe (ou d'un réseau) en *clusters* (ou communautés), est un outil incontournable pour l'analyse de nombreux réseaux d'interaction apparaissant en sociologie, biologie, télécommunications, etc. Le but recherché est en général une forte densité de liens internes aux clusters, avec le moins de liens possible entre clusters. Il s'agit donc d'un problème fondamental pour décomposer de manière hiérarchique un réseau. Il s'avère aussi critique pour la visualisation de grands graphes. Dans ce cadre, il apparait important de quantifier si un réseau admet ou pas un clustering naturel, ou dit autrement, une structure communautaire. Pour une liste concrète d'exemples de réseaux où cette question apparaît cruciale, voir par exemple [4].

Dans l'approche classique, les clusters forment une partition, et parmi les nombreuses définitions de la *qualité* d'un clustering, c'est-à-dire de la fonction objectif d'un algorithme, la plus fréquemment utilisée (du moins à ce qu'il nous semble) est la **modularité** de Newman et Girvan [2, 5]. Informellement, la modularité d'un clustering est le nombre d'arêtes intracluster, comparé au nombre moyen d'arêtes intracluster que l'on aurait après un brassage aléatoire du graphe (couper toutes les arêtes en deux puis les reconnecter suivant une permutation aléatoire, ce qui préserve les degrés). Ce facteur est normalisé entre -1 et 1 afin que, expliquent Newman et Girvan, la qualité moyenne soit de 0, et « *values approaching Q = 1, which is the maximum, indicate strong community structure. In practice, values for such networks typically fall in the range from about 0.3 to 0.7. Higher values are rare* » [5]. La plus forte valeur apparaissant dans leurs articles est 0.86.

Les gens s'intéressent en fait à la modularité d'un graphe, qui est celle de son meilleur clustering. Ce problème est NP-complet [1] même si de nombreux algorithmes existent (dont l'énumération nous ferait dépasser les quatre pages de format et n'est pas notre but). La modularité a donc surtout été étudiée du point de vue algorithmique. Nous nous intéressons ici au comportement asymptotique de la modularité, quand la taille des graphes tend vers l'infini.

Nous montrons que, de façon surprenante, des classes de graphes très régulières (grilles, tores, hypercubes...) ont une modularité asymptotiquement 1, le maximum. C'est le cas également des arbres de degré borné, ce qui nous permet de borner inférieurement la modularité des graphes connexes peu denses, incluant certains *power-law graphs*, qui ont modularité asymptotiquement au moins *2/degré moyen*. Notre preuve, étant constructive, fournit un algorithme garantissant cette performance pour un graphe donné. Nous en concluons qu'avoir une modularité élevée n'est pas, pour un graphe, révélateur d'une nature "intrinsèquement

découpée en clusters". Il existe par exemple beaucoup de clustering différents des grilles de très forte modularité. En tous cas la modularité d'un graphe peu dense devrait toujours être comparée avec *2/degré moyen*, valeur minimum garantie qui ne signifie rien de particulier.

Une version de ces résulats a été publiée dans ISAAC 2011 [3], dont ce papier constitue un résumé francophone, et on pourra y trouver des preuves plus détaillées. Mais nous publions ici de meilleures bornes.

2 Modularité de clustering, de graphe, ou de classe

On utilise les notations standard : graphe $G = (V,E)$ de n sommets et m arêtes, etc. On note $m(A)$ le nombre d'arêtes entre deux sommets de $A \subset V$ et $\text{vol}(A) = \sum_{x \in A} deg(x)$. Un **clustering** est une partition $\mathcal{C}$ de V en *clusters* $C_1 ... C_k$. La **modularité** de ce clustering [5] est :

$$Q(\mathcal{C}) = \sum_{i=1}^{k} \frac{m(C_i)}{m(G)} - \sum_{i=1}^{k} \frac{\text{vol}(C_i)^2}{\text{vol}(G)^2}$$

On appellera "terme de gauche" et "terme de droite" chaque opérande de la soustraction principale. La **modularité** $Q(G)$ **d'un graphe** est la modularité maximum d'un clustering de G. Quelques faits remarquables, la pluspart venant de Brandes & al [1] :

- Étant normalisée la modularité est dans l'intervalle $[-1,1]$
- Le clustering trivial $\mathcal{C} = \{V\}$ a modularité 0, et donc pour tout graphe $Q(G) \geqslant 0$
- Certains graphes (cliques K_n, étoiles $K_{1,n}$...) ont modularité 0 (le clustering trivial est le meilleur)
- Un clustering en k clusters a une modularité *au plus* $1 - 1/k$
- Ce qui implique qu'aucun graphe fini n'a modularité 1 mais motive la définition suivante :

Soit $\mathcal{G}$ une classe de graphes. On dit que la **modularité asymptotique** de la classe est ℓ (resp. au moins ℓ) si pour toute suite $\{G_i\}_i$ de graphes de cette classe de taille strictement croissante, $\lim_{i \to \infty} Q(G) = \ell$ (resp. $\forall \varepsilon > 0 \; \exists N \; t.q. \; n(G) \geqslant N \implies Q(G) \geqslant \ell - \varepsilon$).

3 Classes régulières mais de modularité asymptotiquement 1

On a vu que certaines classes (étoiles et cliques) atteignent le minimum $\ell = 0$. Montrons que la modularité asymptotique maximum $\ell = 1$ peut elle aussi être atteinte par d'autres classes. L'analyse de la définition de la modularité montre que pour cela il faut que

- Le "terme de gauche" (somme des $m(C_i)/m(G)$ dans la def.) tende vers 1, soit $\lim \frac{|aretes \; intracluster|}{|aretes|} = 1$
- Le "terme de droite" tende vers 0. Une façon d'y parvenir est que le clustering soit en k clusters avec
  - k tend vers l'infini quand la taille des graphes tend vers l'infini, et
  - les clusters ont "presque" même volume : $\exists c$ constante t.q. tout cluster a volume au plus $c\frac{\text{vol}(G)}{k}$

On peut exhiber un clustering qui vérifie ces propriétés pour les graphes suivants.

3.1 Grilles, tores et hypertores

On considère un tore G $a \times a$, c'est-à-dire un graphe 4-régulier produit cartésien de deux cycles de a sommets. Il possède $n = a^2$ sommets et $m = 2a^2$ arêtes. Soit b un diviseur de a et considérons une partition (un pavage) en $k = b^2$ carrés (sous-grilles induites du tore, chacune de dimension $(a/b) \times (a/b)$) de G. On l'appellera le **clustering carré** $\mathcal{C}_b$ de G.

Le nombres d'arêtes "externes" (entre clusters) est de $2ab$ car on a tranché le tore b fois verticalement et b fois horizontalement. Le terme de gauche vaut donc $\frac{2a^2-2ab}{2a^2} = 1 - \frac{b}{a}$. Les clusters ont tous même volume $4(a/b)^2$ donc $\frac{vol(C_i)^2}{\text{vol}(G)^2} = \frac{(4(a/b)^2)^2}{(4a^2)^2} = 1/b^4$ et le terme de droite vaut k fois cela donc $1/b^2$. Donc $Q(\mathcal{C}_b) = 1 - \frac{b}{a} - \frac{1}{b^2}$. En prenant $b = \sqrt[6]{n}$ il vient $Q(\mathcal{C}_b) = 1 - \frac{2}{\sqrt[3]{n}}$. **Les tores ont donc modularité asymptotique 1**. Pour simplifier nous avons supposé que le tore est carré $a \times a$ et que b divise a. Ces conditions ne sont pas nécéssaires ; et le résultat vaut aussi pour des grilles. Il se généralise en dimension d :

Théoreme 1 *Soit G un hypertore de dimension d de côtés $p_1 \times ... \times p_d$ (produit cartésien de d cycles).*

$$Q(G) = 1 - O(n^{-1/2d})$$

L'idée est de prendre le plus long côté (qui vaut au moins $\sqrt[d]{n}$) et de couper en b tranches suivant ce côté (hyperplans de dimensions $d-1$) qui ont toutes même largeur (sauf une éventuellement). Un calcul [3] montre que choisir $b = \sqrt{2d}n^{1/2d}$ fournit le résultat annoncée.

3.2 Hypercubes

On considère l'hypercube de dimension d qui a $n = 2^d$ sommets, chacun étant identifié à un nombre binaire à d chiffres. Ainsi il y a une arête entre deux sommets ssi ils diffèrent en un et un seul chiffre.

On construit le **clustering préfixe** $\mathcal{C}_p$ de la façon suivante, pour $p < d$: pour tout nombre a de p bits le cluster C_a est l'ensemble des sommets commençant par le préfixe a. Ainsi

- Le ratio arêtes intracluster/arêtes est exactement $1 - p/d$ puisque chaque sommet a degré d dont p arêtes allant vers un autre cluster. En prenant $p = o(d)$ le terme de gauche tendra donc vers 1
- Il y a $k = 2^p$ clusters, chacun de volume $d2^{d-p}$. Donc $\frac{\text{vol}(C_i)}{\text{vol}(G)} = 2^{-p} = 1/k$. Le terme de droite est donc $k(1/k)^2 = 1/k$ et tend vers 0. En prenant $p = \log_2(d)$ on a $Q(\mathcal{C}_p) = 1 - \frac{p}{d} - \frac{1}{2^p} = 1 - O(\frac{\log\log n}{\log n})$

Les hypercubes ont donc modularité asymptotique 1.

4 Arbres

Les deux classes précédentes sont très petites ; leur intérêt est qu'elles sont très régulières, ce qui montre que la modularité maximale n'est pas atteinte via des clusters "naturellement" pré-existant dans le graphe (zones plus denses, etc). Mais étudions maintenant une classe bien plus utile : les arbres. Elle contient les étoiles, qui ont modularité asymptotique 0. Mais bornons le degré maximum du graphe, en demandant qu'il ne dépende pas linéairement de n. On considère la classe des arbres où, pour toute suite de graphes de taille croissante la limite du degré maximum sur le nombre de sommet est 0.

Théoreme 2 *La classe des arbres de degré maximum $\Delta = o(n)$ a modularité asymptotique de 1*

Démonstration : L'idée est de construire le **clustering centroïdal** $\mathcal{C}_{\leqslant h}$ qui est le suivant. Tout d'abord définissons une **arête centroïde** d'un arbre comme une arête dont la suppression coupe l'arbre en deux composantes de taille aussi proche que possible (idéalement, deux composantes de même taille). Le clustering centroïdal d'un arbre T, étant donné h, et partant d'un clustering trivial $\mathcal{C} = \{V\}$, construit une forêt en répétant simplement l'opération suivante : *tant qu'un arbre de la forêt a strictement plus que h sommets, couper cet arbre en deux en enlevant une arête centroïde*. Chaque arbre obtenu est alors un cluster.

Lemme 1 *Si T est un arbre de degré maximum Δ alors les clusters de $\mathcal{C}_{\leqslant h}$ ont une taille comprise entre h/Δ et h. Le nombre k de ces clusters est compris entre n/h et $\Delta n/h$.*

La démonstration en est facile. Or, comme on part d'un arbre et que les clusters sont connexes on a $k-1$ arêtes entre clusters exactement. Le terme de gauche vaut donc $1 - \frac{k-1}{n-1}$. Un arbre de x sommets a volume $2x-2$. Le volume de T est donc $2n-2$ tandis que celui d'un cluster C_i, qui a au plus h sommet, est $\text{vol}(Ci) \leqslant (2h-2)+(k-1)$: en effet il y a $k-1$ arêtes "hors cluster" qui peuvent augmenter le volume. Le terme de droite est donc $\sum_{i=1}^{k} \frac{\text{vol}(C_i)^2}{\text{vol}(T)^2} \leqslant \sum_{i=1}^{k} \frac{(2h+k-3)\,\text{vol}(C_i)}{\text{vol}(T)^2} \leqslant \frac{2h+k-3}{2n-2}$. En bornant k par $\Delta n/h$ (lemme 1) et en choisissant $h = \sqrt{\Delta n}$ on obtient : $Q(\mathcal{C}_{\leqslant h}) \geqslant 1 - \frac{k-1}{n-1} - \frac{2h+k-3}{2n-2} \geqslant 1 - \frac{5\sqrt{\Delta n}-5}{2n-2}$. Comme $\Delta = o(n)$ le terme soustrait est $o(1)$ et on obtient le résultat annoncé.

5 Graphes peu denses

On a donc une dichotomie entre les arbres de degré maximum linéaire en n, dont certains ont modularité 0, et les autres, de modularité asymptotiquement 1. Mais le théorème 2 sert aussi de base pour le résultat très général suivant :

Théoreme 3 *La classe $\mathcal{G}_d$ des graphes connexes de degré moyen d et de degré maximal $\Delta = o(d\sqrt{n})$ a modularité asymptotique d'au moins $2/d$.*

Démonstration : Pour $G \in \mathcal{G}_d$, on prend un **arbre couvrant** T de G (quelconque !) et on lui applique l'algorithme de clustering h-centroïdal. Les clusters de T que l'on obtient sont alors transformés en clusters de G en rajoutant les arêtes manquantes. Le lemme 1 s'applique toujours. On suppose le pire : qu'aucune de ces arêtes de $G-T$ n'est entre deux sommets du même cluster. Le terme de gauche vaut alors au moins le nombre d'arêtes intracluster de l'arbre T sur le nombre total d'arêtes donc $\frac{(n-1)-(k-1)}{m}$, que l'on peut aussi écrire $\frac{2}{d} - \frac{k}{m}$ car $2m = dn$. Pour le terme de droite, on borne maintenant le volume d'un cluster par Δh, degré max fois taille max. Ce terme est donc borné par $\sum_{i=1}^{k} \frac{\Delta h \,\mathrm{vol}(C_i)}{\mathrm{vol}(G)^2}$ avec cette fois $\mathrm{vol}(G) = 2m$ ce qui donne $\frac{\Delta h}{2m}$ comme borne sup du terme de droite. Donc, toujours en bornant k par $\Delta n/h$; et en prenant $h = \sqrt{n}$; et avec $2m = nd$; on a $\mathcal{Q} \geqslant \frac{2}{d} - \frac{k}{m} - \frac{\Delta h}{2m} \geqslant \frac{2}{d} - \frac{3\Delta}{d\sqrt{n}}$. Comme $\Delta = o(d\sqrt{n})$ on a le résultat annoncé.

La preuve est constructive et fournit donc un algorithme garantissant une modularité supérieure à $\frac{2}{d} - \frac{3\Delta}{d\sqrt{n}}$.

6 Power-law graphs

La classe des power-law graphs de paramètre α est la classe des graphes vérifiant que, pour tout graphe, la proportion de sommets de degré au moins k varie en $O(k^{-\alpha})$. Notons qu'il n'y a pas consensus sur la définition exacte mais c'est l'idée générale (on prend souvent aussi degré *exactement k* ce qui ajoute 1 à α). Supposons que $\alpha > 2$. Prenons $\gamma \in [\frac{1}{\alpha}, \frac{1}{2}]$: pour $k = dn^{\gamma}$, il y a $nO(dk^{-\alpha}) = O(dn^{1-\alpha\gamma}) = o(1)$ sommets ayant au moins ce degré, donc $\Delta < dn^{\gamma}$ pour d borné et n assez grand. On applique le théorème précédent :

Théoreme 4 *Pour $\alpha > 2$ la classe des power-law graphs connexes de paramètre α et de degré moyen d a modularité asymptotique au moins $2/d$.*

Et justement, les graphes "de terrain" sont en général des power-law graphs de paramètre proche de 2 (entre 1 et 2 empiriquement). Bien entendu leur degré maximum "réel" peut varier. Mais ce résultat est néanmoins très significatif quant au fait que les graphes étudiés au cas par cas dans la littérature ont une modularité importante. Rappelons que notre construction prend un arbre couvrant et suppose "hors clustering" toutes les autres arêtes : en pratique, grâce en particulier au *coefficient de clustering* † élevé, on aura bien mieux.

7 Conclusion

L'idée est généralement répandue qu'après application d'un algorithme de clustering, obtenir une modularité élevée dit quelque chose de fort sur la donnée : qu'elle aurait une structure naturellement découpée en clusters, mais masquée par le bruit de quelques arêtes en trop ou manquantes, et que l'algorithme aurait juste retrouvée cette « *strong community structure* » [5] cachée. Puisque des graphes très réguliers ont modularité bien plus élevée (un tore d'un million de sommets atteint déjà 0,98) nous pensons au contraire que la modularité ne devrait être utilisée que comme objectif à maximiser pour un algorithme, sans lui faire dire quoi que ce soit sur la donnée. En tous cas il faudrait toujours comparer une modularité prétendument élevée avec la borne inférieure en *2/degré moyen* , avant d'avancer qu'un graphe "contient des clusters".

Références

[1] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner. On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20 :172–188, 2008.

[2] M. Girvan and M. Newman. Community structure in social and biological networks. *P.N.A.S*, 99(12) :7821, 2002.

[3] F. d. Montgolfier, M. Soto, and L. Viennot. Asymptotic modularity of some graph classes. In *22nd International Symposium on Algorithms And Computation (ISAAC)*, pages 435–444, 2011.

[4] M. Newman. The structure and function of complex networks. *SIAM review*, pages 167–256, 2003.

[5] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(066133), 2004.

†. Utilisation du même anglicisme mais avec un sens différent : c'est la probabilité que deux sommet ayant un voisin commun soient voisins. Notion popularisée par Watts et Strogatz et formant l'un des axiomes des *graphes petits monde*.