Asymptotic Modularity of some Graph Classes *

Fabien de Montgolfier[1], Mauricio Soto[1], and Laurent Viennot[2]

[1] LIAFA, UMR 7089 CNRS - Université Paris Diderot.
[2] INRIA and Université Paris Diderot
fm@liafa.jussieu.fr mausoto@liafa.jussieu.fr Laurent.Viennot@inria.fr

Abstract. Modularity has been introduced as a quality measure for graph partitioning. It has received considerable attention in several disciplines, especially complex systems. In order to better understand this measure from a graph theoretical point of view, we study the modularity of a variety of graph classes. We first consider simple graph classes such as tori and hypercubes. We show that these regular graph families have asymptotic modularity 1 (that is the maximum possible). We extend this result to the general class of unit ball graphs of bounded growth metrics. Our most striking result concerns trees with bounded degree which also appear to have asymptotic modularity 1. This last result can be extended to graphs with constant average degree and to some power-law graphs.

1 Introduction

Graph partitioning (also known as graph clustering, see [13]) is a fundamental problem that recently became popular in the context of partitioning a network into communities. In that line of work, modularity [5, 11] has been introduced as a quality measure of a network partitioning. It has rapidly become popular for various applications from biological networks to social networks measurement (see, e.g. [6, 10, 12]) or modelling [7, 9]. It is now a standard measure for comparing the partitions obtained by various algorithms on practical networks.

From a graph-theoretic point of view, modularity has mainly been considered as a computational problem. Several heuristics have been proposed for computing a high modularity partition of a given network [4, 3, 1]. However, computing a maximum modularity partition is NP-complete [2].

This paper follows the approach of [2] with the goal to give some insight about what high modularity means from a graph-theoretic perspective. We are not interested in computing the modularity of a given graph extracted from "real" data. Instead, we analyze the modularity of *graph classes*. Given some well-known families of graphs, we wonder how good (or how bad) their members can be clustered. Our results are expressed as asymptotic modularity (limit when the graph size goes to infinity).

Informally, the modularity of a partition is, up to a normalization constant, the number of edges falling within clusters of the partition minus the expected

* Supported by the european project "EULER" (Grant No.258307) and by the INRIA project-team "GANG".

number in an equivalent network with edges placed at random. It is normalized so that it amounts to a number between -1 and 1. The modularity of a graph is the maximum modularity of a partition of the graph over all partitions. As a single cluster partition trivially leads to a modularity of 0, the modularity of a graph is always between 0 and 1.

Our contribution resides in the analysis of the asymptotic modularity of various classes of graphs. We show that grids, tori, and hypercubes have asymptotic modularity 1, and can thus be well clustered despite their very regular structure. We extend this result to unit ball graphs of bounded growth metrics, showing a lower bound of the modularity of such graphs depending on the growth constant of the metric. On the other hand, stars (who are trees of unbounded maximum degree) have modularity as low as 0. However, trees of bounded degree have asymptotic modularity 1. This result can be extended to any class of constant average degree d connected graphs, showing a lower bound of $\frac{2}{d}$ for their asymptotic modularity.

As a consequence, high modularity is not necessarily the sign of an intrinsic community structure. Grids for example can be clustered in various manners to obtain high modularity. Second the modularity obtained for a connected graph with average degree d should be compared to $\frac{2}{d}$. A value not significantly larger cannot be interpreted as the sign of intrinsically clustered data.

The paper is organized as follows. Section 2 gives the formal definition and some preliminary remarks. Section 3 introduces a *decomposable graphs* framework and is devoted to tori, and hypercubes. Section 4 introduces the class of unit ball graphs of a bounded growth metric (which generalizes grids for instance) and show a lower bound on its asymptotic modularity. Finally, Section 5 is devoted to trees with small maximum degree, and graphs with constant average degree, and power-law graphs.

2 Revisiting modularity

Given a graph $G = (V(G), E(G))$, we denote by n and m its number of vertices and edges respectively. Given a subset of vertices $S \subset V(G)$, the *size* $|S|$ of S is its number of vertices and its *volume* $\text{vol}(S) = \sum_{v \in S} \deg(v)$ is its degree sum. Let $E(S)$ denote the edge-set of the graph induced by S. $|E(S)|$ is thus the number of inner-edges in S. Similarly, $E(G)$ denotes the set of edges of G and $\text{vol}(G)$ denotes the volume of $V(G)$ (note that $\text{vol}(G) = 2m$).

A *clustering* is a partition of $V(G)$ into disjoint sets called *clusters*. Many quality measures can be defined for judging how good a clustering is. A popular one was introduced by Newman and Girvan [5, 11] and is called *modularity*.

Definition 1 ([5, 11]). *Let G be a graph and $\mathcal{C} = \{C_1, ...C_k\}$ a partition of $V(G)$ into k* clusters. *The* modularity *of the clustering $\mathcal{C}$ is defined as:*

$$\mathcal{Q}(\mathcal{C}) = \sum_{i=1}^{k} \left[\frac{|E(C_i)|}{|E(G)|} - \frac{\text{vol}(C_i)^2}{\text{vol}(G)^2} \right]$$

The modularity of graph G, denoted $\mathcal{Q}(G)$ is the maximum modularity among all possible clusterings of G.

In this definition both the *left term* (densities) and the *right term* (volumes) take values between 0 and 1 so $\mathcal{Q}(\mathcal{C})$ belongs to $[-1, 1]$ (and in fact to $[-1/2, 1[$, see [2]). The left term is the ratio of internal edges. The less clusters you have, the greater it will be. To counterbalance this tendency to have few clusters, the right term is the quality of the same clustering of a random graph with same degree sequence. Detailed motivation for the definition is given by Newman [10]. Computing the modularity of a given clustering clearly takes $O(n + m)$ time, but giving the modularity of a graph is an NP-hard problem [2].

According to Brandes *et al.* [2], for any graph G, we have $0 \leq \mathcal{Q}(G) < 1$. We can refine the result further:

Lemma 1. *If the maximum degree of Graph G is Δ then $\mathcal{Q}(G) \leq 1 - \frac{\Delta^2}{4m^2}$.*

Proof. Let x be a degree Δ vertex. If C denotes the cluster containing x then $\mathrm{vol}(C) \geq \Delta$. So the right term is at least $\Delta^2/4m^2$, and the left term at most 1.

Lemma 1 motivates to work on *asymptotic* modularity: no graph has modularity exactly 1, but the modularity of a *sequence* of graphs may have limit 1. A graph class has asymptotic modularity ℓ if any sequence of graph in the class taken with increasing number of vertices has a limit modularity of ℓ. We focus on $\ell = 1$, the ideal case. We now give some simple facts about modularity.

Lemma 2. *Given a cluster C, if $G[C]$ is not connected, then breaking C in $C_1 \uplus C_2$ by assigning each component of $G[C]$ to C_1 or C_2 improves modularity.*

Proof. We have $\mathrm{vol}(C)^2 \geq \mathrm{vol}(C_1)^2 + \mathrm{vol}(C_2)^2$ (since $x \mapsto x^2$ is superlinear) so right term of Definition 1 decreases. And the left term is the ratio of internal edge, so it remains unchanged as each edge of C goes within C_1 or C_2. (In fact $\mathrm{vol}(C)^2 > \mathrm{vol}(C_1)^2 + \mathrm{vol}(C_2)^2$ as soon as both C_1 and C_2 contain an edge.)

Corollary 1 (Brandes *et al.* [2]). *There exists a maximum modularity clustering where each cluster is connected.*

Lemma 3 (DasGupta and Devendra [?]). *If Clustering $\mathcal{C}$ contains k clusters then $\mathcal{Q}(\mathcal{C}) \leq 1 - 1/k$.*

As a consequence, note that a modularity close to 1 can only be obtained through a large number of clusters.

Proposition 1. *A star of m rays ($K_{1,m}$ graph) has modularity 0.*

Proof. Every clustering consists in a first cluster C_0 containing the center together with $a \leq m$ other vertices, plus other clusters that are stable sets. According to Corollary 1, in a clustering with maximal modularity, each one contains one vertex. Their number of edges is 0 and their volume is 1. Then $\mathcal{Q}(\mathcal{C}) =$

$$\sum_{i=1}^{k} \left[\frac{|E(C_i)|}{|E(G)|} - \frac{\mathrm{vol}(C_i)^2}{\mathrm{vol}(G)^2} \right] = \frac{a}{m} - \frac{(m+a)^2 + (m-a)}{4m^2} = \frac{(m-a)(a-m-1)}{4m^2}$$

Maximum is for $a = m$, *i.e* an unique cluster C_0. The modularity is then 0.

Consequently, trees may have modularity as low as 0. We shall now present, on the other hand, classes having modularity 1. We shall see further that bounded degree trees also have modularity 1 (Section 5.1).

3 Modularity of decomposable graphs

3.1 Decomposable graphs

To reach asymptotic modularity 1, it is necessary that the limit (when n goes to infinity) of the left term be 1 and that of the right term be 0. Furthermore, Lemma 3 implies that the number k of clusters must also tends to infinity. That leads us to the following definition:

Definition 2. *A graph G is* (k, c, e)-decomposable *if it can be split into k clusters such that each cluster has volume at most $c\frac{\mathrm{vol}(G)}{k}$, and the number of inter-cluster edges is at most $|E(G)| \times e$.*

Intuitively, a graph is decomposable if it can be split into k clusters of roughly balanced volume with few edges in-between clusters. The following lemma bounds the modularity of a decomposable graph.

Lemma 4. *If G is (k, c, e)-decomposable, then* $\mathcal{Q}(G) \geq 1 - e - \frac{c^2}{k}$.

Proof. Consider a clustering $\mathcal{C} = \{C_1, \ldots, C_k\}$ such that G is (k, c, e)-decomposable. Then $\sum_{i=1}^{k} |E(C_i)| \geq m - em$ and $\sum_{i=1}^{k} \mathrm{vol}(C_i)^2 \leq \sum_{i=1}^{k} (c\,\mathrm{vol}(G)/k)^2$ $\leq \frac{c^2}{k}\,\mathrm{vol}(G)^2$. By Definition 1 we get $\mathcal{Q}(\mathcal{C}) \geq \frac{m-em}{m} - \frac{\frac{c^2}{k}\,\mathrm{vol}(G)^2}{\mathrm{vol}(G)^2} \geq 1 - e - \frac{c^2}{k}$.

Let us apply this tool to computation of asymptotic modularity of some classes.

3.2 Multidimensional torus

Given a vector $\boldsymbol{p}$ of d numbers $p_1, \ldots, p_d$ let the d-dimensional torus $G_{\boldsymbol{p}} = C_{p_1} \times \cdots \times C_{p_d}$ (where C_a is a cycle of a vertices and $\times$ is the Cartesian product of graphs). We have $n = \prod_{i=1}^{d} p_i$ and $m = dn$.

Lemma 5. *d-dimensional tori are $(b, 2, \frac{2b}{dn^{1/d}})$-decomposable.*

Proof. Consider the largest dimension $p = \max_{i=1}^{d} p_i$ of the torus. We have $p \geq n^{1/d}$. Let $\mathcal{C}_b$ be the clustering where the torus is split into b slices according to the largest dimension. The cluster C_i, with $0 \leq i < b$, corresponds to nodes whose coordinate in the largest dimension falls in the interval $[i\,\lceil p/b \rceil, (i+1)\,\lceil p/b \rceil[$.

The nodes with same coordinate in the largest dimension form an $d-1$-hyperplane of size n/p. The number of edges outgoing a cluster is thus at most $2n/p$. So $e \leq \frac{2bn/p}{dn} \leq \frac{2b}{dn^{1/d}}$. Cluster C_i contains at most $n/p\,\lceil p/b \rceil \leq \frac{n}{b} + \frac{n}{p} \leq \frac{2n}{b}$ nodes as $b \leq p$. The degree of each node is $2d$ so $\mathrm{vol}(C_i) \leq \frac{4dn}{b}$. We have $\frac{\mathrm{vol}(C_i)}{\mathrm{vol}(G)} \leq \frac{4dn}{b}\frac{1}{2dn} \leq \frac{2}{b}$.

Theorem 1. *A d-dimensional torus with n nodes has modularity $1-O(n^{-1/2d})$.*

Proof. According to Lemma 4 $\mathcal{Q}(G) \geq 1 - \frac{2b}{dn^{1/d}} - \frac{2^2}{b}$ This value is maximal for $\frac{2n^{-1/d}}{d} = \frac{4}{b^2}$, i.e. $b = \sqrt{2d}n^{1/2d}$. For this value of b, we obtain $\mathcal{Q}(\mathcal{C}_b) \geq 1 - \frac{4\sqrt{2}}{\sqrt{d}} n^{-1/2d} \geq 1 - O(n^{-1/2d})$

A very similar proof (not presented here) can show the same bound applies to grids (with $c = 4$ instead of 2). Note that tori and grids have many natural cuts and there are several possible partitions with asymptotic modularity 1.

3.3 Hypercube

The d-dimensional hypercube has $n = 2^d$ vertices and $\text{vol}(G) = d2^d$. Its vertices are identified with the d-digits binary numbers. So we may say, for instance, that there is an edge uv iff u and v differ on one bit.

Lemma 6. *d-dimensional hypercubes are $(2^b, 1, \frac{b}{d})$-decomposable.*

Proof. Given an integer $b < d$, let the *b-prefix* clustering $\mathcal{C}_b$ of d-dimensional hypercube be the clustering with 2^b clusters such that cluster C_a contains vertices beginning with prefix a where a is a b bits binary string.

Among the d edges incident to a given vertex, b are external (go to another cluster, if vertices differ on a bit of number $i \leq b$) and $d - b$ are internal. So clearly $e = \frac{b}{d}$. Cluster C_i contains 2^{d-b} vertices and has thus volume $d2^{d-b}$. Then $\frac{\text{vol}(C_i)}{\text{vol}(G)} = 2^{-b} = 1/k$, i.e. $c = 1$.

Theorem 2. *A hypercube of dimension d has modularity $1 - O(\frac{\log\log n}{\log n})$.*

Proof. According to Lemma 4, $\mathcal{Q}(G) \geq 1 - \frac{b}{d} - \frac{1}{2^b}$. It reaches its maximum when $b = \log_2(d \ln 2)$ and: $\mathcal{Q}(G) \geq 1 - \frac{\log_2(d \ln 2)}{d} - \frac{1}{d \ln 2} = 1 - O(\frac{\log\log n}{\log n})$ since $d = \log_2 n$.

4 Unit Ball Graph of a Bounded Growth Metric

We now turn to the general class of unit ball graphs. Note that a grid can be defined as the unit ball graph of a regular mesh of points in a d-dimensional space. Varying the unit radius allows to play with the density of the grid. Varying the point positions yields non uniform grids. We restrict to the case where the metric induced by the points has bounded growth.

We are now interested in a cloud of points V from a metric space $\mathcal{E}$ (in the most usual cases, $\mathcal{E}$ is $\mathbb{R}^d$ or a d-dimensional $\mathbb{R}$-space, but we do not require that). We suppose there is a metric called *dist* (a function $\mathcal{E} \times \mathcal{E} \to \mathbb{R}^+$). Given

a length $R \in \mathbb{R}^+$, we define $E_R = \{uv | dist(u,v) \leq R\}$ We call *R-ball graph* or (since all balls have same radius) *unit ball graph* the graph $G_R = (V, E_R)$, and the *modularity of a clustering $\mathcal{C}$ of V* is then the modularity computed in G_R. In this section, we lower bound the asymptotic modularity of G_R provided that the radius R is taken in an appropriate range and that the metric has bounded growth.

4.1 Bounded growth metrics

The *bounded growth* property is a generalization of the Euclidean dimension. We let $B(u,r) = \{v \in V \mid dist(u,v) \leq r)$ denote the *ball* with center u and radius r.

Definition 3 (Grid Dimension [8] also known as *bounded growth* property). *Space V has* growth $\gamma > 0$ *if doubling the radius of any ball does not increase its volume by more than γ: $\forall x \in V, \forall r > 0, |B(x,2r)| \leq \gamma \cdot |B(x,r)|$*

For instance, the d-dimensional torus has growth 2^d. This definition also applies to continuous spaces. For instance a d-dimensional Euclidean space with the L_∞ norm has growth $\gamma = 2^d$.

4.2 *R*-nets

Now let us recall an algorithmic tool: the R-net. It is a covering of the space by points mutually at least R apart. Among its many uses, it allows to define a clustering where the radius of each cluster is bounded by R. The current section presents such clusterings, the next one proves that they have good modularity.

Definition 4 (R-net). *A subset U of V is a R-net if $\forall u, u' \in U$, $dist(u,u') > R$ (points are R apart), and $\forall v \in V\ \exists u \in U$ $dist(u,v) \leq R$ (U covers V)*

A greedy process can easily construct an R-net: take any vertex $u \in V$, put it in U, and recurse on $V - B(u,R)$. Notice we are not interested in minimizing $|U|$ nor maximizing $B(u,R)$.

Then an R-net U allows to define a clustering (denoted $\mathcal{C}_U$) of V. It consists in $|U|$ clusters. For each $u_i \in U$, we define cluster C_i as the points of V whose nearest neighbors in U is u_i (ties are arbitrarily broken). The *cover* part of the definition implies that for any $v \in C_i$, we have $dist(u_i, v) \leq R$. Cluster radius is thus at most R. Additionally, as a consequence of the nearest neighbor choice, a point $v \in B(u_i, R/2)$ cannot be in C_j with $j \neq i$ as $dist(u_i, u_j) > R$. Cluster C_i thus contains $B(u_i, R/2)$.

4.3 Modularity of an *R*-net clustering

Theorem 3. *Let V be a finite space of n points, together with Metric dist, and having growth at most γ, and $R \geq 0$ such that for all $v \in V$ we have $|B(x, R/2)| > 1$ and $|B(x, R/2)| = o(\sqrt{n})$. We have:*

$$\mathcal{Q}(G_R) \geq \frac{1}{2\gamma^3} - o(1)$$

Proof. Let $U = \{u_1, \ldots, u_k\}$ be an R-net of V and let $\mathcal{C}_U = \{C_i\}_{i\in\{1,\ldots,k\}}$ be the associated clustering as defined previously. Then, by construction, $B(u_i, R/2) \subseteq C_i \subseteq B(u_i, R)$. Let $b_i = |B(u_i, R/2)|$. The bounded growth hypothesis gives: $b_i \leq |C_i| \leq \gamma b_i$. Points from $B(u_i, R/2)$ are all mutually linked in G_R (from its definition) and form a clique of size b_i, included within C_i. Then we have:

$$\mathcal{Q}(\mathcal{C}_U) = \sum_{i=1}^{k} \left[\frac{|E(C_i)|}{m} - \frac{(\sum_{v\in C_i} d_v)^2}{4m^2} \right] \geq \sum_{i=1}^{k} \left[\frac{b_i(b_i-1)}{2m} - \frac{(\sum_{v\in C_i} d_v)^2}{4m^2} \right]$$

for every $v \in C_i$, its degree is $d_v = |B(v, R)| \leq |B(u_i, 2R)| \leq \gamma^2 b_i$, and thus:

$$\mathcal{Q}(\mathcal{C}_U) \geq \sum_{i=1}^{k} \left[\frac{b_i(b_i-1)}{2m} - \frac{(|C_i|\gamma^2 b_i)^2}{4m^2} \right] \geq \sum_{i=1}^{k} \left[\frac{b_i(b_i-1)}{2m} - \frac{\gamma^6 b_i^4}{4m^2} \right]$$

$$\text{As we have: } \sum_{i=1}^{k} b_i(b_i-1) \leq 2m = \sum_{u\in V(G)} d_u \leq \sum_{i=1}^{k} |C_i|\gamma^2 b_i \leq \gamma^3 \sum_{i=1}^{k} b_i^2,$$

$$\text{we get: } \mathcal{Q}(\mathcal{C}_U) \geq \frac{\sum_{i=1}^{k} b_i(b_i-1)}{\gamma^3 \sum_{i=1}^{k} b_i^2} - \frac{\gamma^6 \bar{b}^2 \sum_{i=1}^{k} b_i^2}{2m \sum_{i=1}^{k} b_i(b_i-1)}$$

where $\bar{b} = \max_i \{b_i\}$. The ball size hypothesis of the theorem implies $b_i \geq 2$ and $b_i = o(\sqrt{n})$. We have thence $\frac{\sum_{i=1}^{k} b_i(b_i-1)}{\sum_{i=1}^{k} b_i^2} = 1 - \frac{\sum_{i=1}^{k} b_i}{\sum_{i=1}^{k} b_i^2} \geq 1 - \frac{\sum_{i=1}^{k} b_i}{\sum_{i=1}^{k} 2b_i} = \frac{1}{2}$. Back to modularity: $\mathcal{Q}(\mathcal{C}_U) \geq \frac{1}{2\gamma^3} - \frac{\gamma^6 \bar{b}^2}{m}$ Finally, as $\bar{b} = o(\sqrt{n})$ and $m \geq \frac{n}{2}$, we get $\mathcal{Q}(\mathcal{C}_U) \geq \frac{1}{2\gamma^3} - o(1)$: the asymptotic modularity of the class is $\frac{1}{2\gamma^3}$

5 Constant average degree graphs

We now investigate some sparse classes of graphs, in the sense that the average degree is a constant d and does no go to infinity with n. Trees are such a class. We prove that they have asymptotic modularity 1 when maximum degree is low enough. We extend then the result to some constant average degree graph classes using tree spanners.

5.1 Trees of small maximum degree

Let T be a tree. The removal of any edge of T splits T into two parts. We are interested in the size of the smaller part. A *centroid edge* of T is an edge chosen to maximize the size of the smaller part (a centroid edge is thus incident with the unique or the two *centroid vertices* of the tree).

Let us now introduce a clustering tool, Algorithm *greedy-decompose*$_{\leq h}$. Given a forest F and a fixed integer $h \geq 1$, the algorithm works as follows. As long as F contains a tree T_0 with strictly more than h vertices, then it finds a centroid edge e of T_0 and removes it ($F := F - e$). The clustering $\mathcal{C}_h$ of T is the forest computed by *greedy-decompose*$_{\leq h}$ using T as initial forest.

Lemma 7. *If a tree has maximum degree at most Δ then* greedy-decompose$_{\leq h}$ *computes clusters of size $\frac{h}{\Delta} \leq |C_i| \leq h$.*

Proof. Clearly each cluster has size at most h. Let e be the centroid edge removed from a tree T_0 with n vertices. Let s be the size of the smallest part of $T_0 - e$ (we have $s \leq n/2$). Let x be the vertex incident with e and belonging to the largest part of $T_0 - e$. For every edge e' incident with x, the part of $T_0 - e'$ not containing x has at most s vertices (otherwise e is not a centroid edge). As x has degree at most Δ then $n \leq \Delta s + 1$ (each of the $\leq \Delta$ parts size is bounded by s, and $+1$ counts x itself). So $s \geq \frac{n-1}{\Delta}$. In the algorithm we split only trees with size $n > h$ so $s \geq \frac{h}{\Delta}$.

So the cluster sizes are roughly balanced (up to a ratio Δ) and parametrized with h. Indeed, as the sum of cluster sizes is n we have that *greedy-decompose*$_{\leq h}$ splits T into $\frac{n}{h} \leq k \leq \frac{\Delta n}{h}$ clusters.

Lemma 8. *For any tree T of degree bounded by Δ there exists k such that T is $(k, \Delta^2, \frac{k-1}{n-1})$-decomposable.*

Proof. Splitting a tree of n vertices into k connected clusters using *greedy-decompose*$_{\leq h}$ yields a fraction $\frac{k-1}{n-1}$ of external edges. As each vertex has degree at most Δ, for each cluster $\mathrm{vol}(C_i) \leq \Delta h$. As $h \leq \frac{\Delta n}{k}$ we get $\mathrm{vol}(C_i) \leq \Delta^2 \frac{\mathrm{vol}(T)}{k}$ (as $\mathrm{vol}(T) = 2(n-1) \geq n$ for $n \geq 2$).

Theorem 4. *Trees with max. degree $\Delta = o(\sqrt[5]{n})$ have asymptotic modularity 1.*

Proof. Let us find a good k. We have control over h and it gives k with $\frac{n}{h} \leq k \leq \frac{\Delta n}{h}$. Lemma 4 gives

$$\mathcal{Q}(T) \geq 1 - \frac{k-1}{n-1} - \frac{\Delta^4}{k} \geq 1 - \frac{(\Delta n/h) - 1}{n-1} - \frac{\Delta^4}{(n/h)} \geq 1 + \frac{1}{n-1} - \frac{\Delta n}{h(n-1)} - \frac{\Delta^4}{n} h$$

For maximizing $\mathcal{Q}$ we take derivative: $\frac{d\mathcal{Q}}{dh} = \frac{\Delta n}{h^2(n-1)} - \frac{\Delta^4}{n}$. Solving $\frac{d\mathcal{Q}}{dh} = 0$ gives $h = \frac{n}{\Delta\sqrt{\Delta(n-1)}}$. Using the clustering from *greedy-decompose*$_{\leq \frac{n}{\Delta\sqrt{\Delta(n-1)}}}$ we get

$$\begin{aligned}\mathcal{Q}(T) &\geq 1 + \frac{1}{n-1} - \frac{\Delta n}{(n-1)} \cdot \frac{\Delta\sqrt{\Delta(n-1)}}{n} - \frac{\Delta^4}{n} \cdot \frac{n}{\Delta\sqrt{\Delta(n-1)}} \\ &\geq 1 + \frac{1}{n-1} - \frac{2\Delta^{2.5}}{\sqrt{n-1}}. \text{ If } \Delta = o(\sqrt[5]{n}) \text{ then } \frac{2\Delta^{2.5}}{\sqrt{n-1}} = o(1).\end{aligned}$$

On the other hand, given a tree class, if there is a constant $c \geq 0$ such that there is a sequence T_i of trees of the class, each one containing a vertex x of degree cn then according to Lemma 1 $\mathcal{Q}(T_i) \leq 1 - \frac{cn^2}{4(n-1)^2} < 1 - \frac{c}{4}$. The asymptotic modularity of a tree class with maximal degree $\Omega(n)$ is strictly less than 1.

5.2 Graphs of average degree d

Let $d = 2m/n$ be the average degree of a graph. In this section we prove that a class of graphs with constant average degree d and bounded *maximum* degree have good asymptotic modularity (maximum degree is bounded by a function of n but may go to infinity). It is an extension to graphs of results from the previous section. Indeed, given a connected graph G of maximum degree Δ, we take a spanning tree T of G (thus having maximum degree at most Δ) and apply *greedy-decompose*$_{\leq h}$ on T. Lemma 7 remains true and we still have $\frac{n}{h} \leq k \leq \frac{\Delta n}{h}$ clusters. It is a clustering of G as T spans G. Each cluster is connected and has volume at most $\Delta|C_i| \leq \Delta h$.

Lemma 9. *For any connected graph G of maximum degree bounded by Δ and of average degree d there exists k such that G is $(k, \frac{\Delta^2}{d}, 1 - \frac{n-k}{m})$-decomposable.*

Proof. Among the m edges of G, $n-k$ belong to the clusters of T and are thus internal (we can not say anything about edges not in T). So we have $e \leq 1 - \frac{n-k}{m}$. On the other hand $k\frac{\mathrm{vol}(C_i)}{\mathrm{vol}(G)} \leq k\frac{\Delta h}{2m} \leq \frac{\Delta^2}{d}$ since $k \leq \frac{\Delta n}{h}$ and $d = \frac{2m}{n}$.

Theorem 5. *Connected graphs of average degree d and of maximum degree $\Delta = o(\sqrt[5]{d^3 n})$ have asymptotic modularity at least $2/d$.*

Proof. Using Lemma 4 we have: $\mathcal{Q}(G) \geq 1 - \left(1 - \frac{n-k}{m}\right) - \left(\frac{\Delta^2}{d}\right)^2 \cdot \frac{1}{k} \geq \frac{n}{m} - \frac{\Delta n}{hm} - \frac{\Delta^4 h}{nd^2}$ since $\frac{1}{k} \leq \frac{h}{n}$ and $k \leq \frac{\Delta n}{h}$. Finally: $\mathcal{Q}(G) \geq \frac{2}{d} - \frac{2\Delta}{dh} - \frac{\Delta^4}{nd^2}h$. For maximizing $\mathcal{Q}$ we take the derivative in h: $\mathcal{Q}' = \frac{2\Delta}{d}\frac{1}{h^2} - \frac{\Delta^4}{nd^2}$. It is zero for $h = \frac{\sqrt{2dn}}{\Delta^{1.5}}$. So taking clustering *greedy-decompose*$_{\frac{\sqrt{2dn}}{\Delta^{1.5}}}$ we have $\mathcal{Q}(G) \geq \frac{2}{d} - \frac{2\Delta}{d}\frac{\Delta^{1.5}}{\sqrt{2dn}} - \frac{\Delta^4}{nd^2}\frac{\sqrt{2dn}}{\Delta^{1.5}} \geq \frac{2}{d} - \frac{2^{1.5}\Delta^{2.5}}{d^{1.5}\sqrt{n}} \geq \frac{2}{d} - o(1)$ since $\Delta = o(\sqrt[5]{d^3 n})$.

Notice that for trees $d = \frac{2(n-1)}{n}$ has limit 2. This result thus generalizes the previous one.

5.3 Power-law graphs

Let us say a graph class has the *power-law* property of parameter α if for any graph the proportion of vertices of degree at least k is $O(k^{-\alpha})$. Note that our definition is much broader than the usual definition of power law graph.

Theorem 6. *A power-law connected graph class of parameter $\alpha > 5$ with constant average degree d has asymptotic modularity at least $\frac{2}{d}$.*

Proof. We first prove that the maximum degree of each graph is $\Delta = o(\sqrt[5]{n})$. The number of vertices of degree at least k is $nO(k^{-\alpha})$. Consider $k = n^\gamma$ for some $\gamma \in (\frac{1}{\alpha}, \frac{1}{5})$. Then we have $nO(k^{-\alpha}) = O(n^{1-\alpha\gamma}) = o(1)$. For n large enough, there are thus no vertex of degree at least n^γ. We thus have $\Delta = O(n^\gamma) = o(\sqrt[5]{n})$. We can finally apply the previous theorem since d is a constant.

6 Conclusion

Usually, people consider that a graph has a good modularity if there is some intrinsic clustering that induces the existence of edges, but is somehow blurred by adding or removing a few edges. Then a clustering algorithm has to "retrieve" this hidden structure. For instance, according to Newman and Girval [11], *"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."* We show however that some very regular graphs, like tori or hypercubes, where no "hidden naturally clustered structure" seems to exist, may also have a high quality clustering. That relativizes the use of modularity as a measurement of the "*clustering*" property of data: we think it should be seen only as the objective function of an algorithm.

References

1. V. Blondel, J.-L. Guillaume, et al. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008.
2. U. Brandes, D. Delling, et al. On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20:172–188, 2008. ISSN 1041-4347.
3. A. Clauset, M. E. J. Newman, et al. Finding community structure in very large networks. *Physical Review E*, 70:066111, 2004.
4. J. Duch and A. Arenas. Community detection in complex networks using extremal optimization. *Phys. Rev. E*, 72(2):027104, Aug 2005.
5. M. Girvan and M. Newman. Community structure in social and biological networks. *P.N.A.S*, 99(12):7821, 2002.
6. R. Guimera and L. A. N. Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, 2005.
7. R. Guimera, M. Sales-Pardo, et al. Modularity from fluctuations in random graphs and complex networks. *Phys. Rev. E*, 70:025101, 2004.
8. D. Karger and M. Ruhl. Finding nearest neighbors in growth-restricted metrics. In *Proceedings of STOC*, pp. 741–750. ACM, 2002.
9. N. Kashtan and U. Alon. Spontaneous evolution of modularity and network motifs. *P.N.A.S*, 102(39):13773, 2005.
10. M. E. Newman. Modularity and community structure in networks. *PNAS*, 103(23):8577–8582, 2006.
11. M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(066133), 2004.
12. J. Olesen, J. Bascompte, et al. The modularity of pollination networks. *Proceedings of the National Academy of Sciences*, 104(50):19891, 2007.
13. S. Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27 – 64, 2007.