

Bad communities with high modularity

Ath. Kehagias^a and L. Pitsoulis

Faculty of Engineering, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece

Received 27 February 2013 / Received in final form 10 May 2013

Published online 24 July 2013 – © EDP Sciences, Società Italiana di Fisica, Springer-Verlag 2013

Abstract. In this paper we discuss some problematic aspects of Newman and Girvan’s modularity function Q_N . Given a graph G , the modularity of G can be written as $Q_N = Q_f - Q_0$, where Q_f is the intracluster edge fraction of G and Q_0 is the expected intracluster edge fraction of the null model, i.e., a randomly connected graph with same expected degree distribution as G . It follows that the maximization of Q_N must accomodate two factors pulling in opposite directions: Q_f favors a small number of clusters and Q_0 favors many balanced (i.e., with approximately equal degrees) clusters. In certain cases the Q_0 term can cause overestimation of the true cluster number; this is the opposite of the well-known underestimation effect caused by the “resolution limit” of modularity. We illustrate the overestimation effect by constructing families of graphs with a “natural” community structure which, however, does not maximize modularity. In fact, we show there exist graphs G with a “natural clustering” \mathbf{V} of G and another, balanced clustering \mathbf{U} of G such that (i) the pair (G, \mathbf{U}) has higher modularity than (G, \mathbf{V}) and (ii) \mathbf{V} and \mathbf{U} are arbitrarily different.

1 Introduction

This paper describes some problems which may arise in using Newman and Girvan’s modularity function Q_N [1] for community detection. Modularity is one of the most popular quality functions in the community detection literature. It is not only used to evaluate the community structure of a graph, but also to perform community detection by modularity maximization. However, it is well-known that modularity maximization can, in certain cases, yield the “wrong” community decomposition. Previous work on this aspect has focused on the modularity resolution limit [2], which causes underestimation of the true number of communities. Some researchers have also studied the opposite effect, namely overestimation of the true community number. For example, it is reported in references [3,4] that sparse graphs tend to cluster into more modules than predicted by certain statistical mechanics models of community structure. This property of sparse graphs can lead modularity maximization to overestimate the number of communities (see, e.g., [5] where the overestimation effect is studied in connection to both modularity maximization and the Infomap framework [6]).

In this paper we focus on the overestimation of the number of communities by modularity maximization and provide some precise results in this direction.

More specifically, the paper is organized as follows. In Section 2, we present our nomenclature and notation; let us stress from the beginning that we will use “cluster” as a synonym of “community” and “clustering” to denote both a partition of the nodes of a graph and the activity of creating such a partition. In Section 3, we present an in-

terpretation of Q_N which, as far as we know, has not been discussed previously. It is well-known that the modularity of a graph G can be written in the form $Q_N = Q_f - Q_0$, where Q_f is the intracluster edge fraction of G and Q_0 is the expected intracluster edge fraction of the null model, i.e., a graph G' which has the same expected degree distribution as G but randomly distributed edges. As explained in Section 3.2, maximization of Q_f favors clusterings with a small number of clusters and few edges across clusters. On the other hand, as explained in Section 3.3, minimization of Q_0 favors clusterings with a large number of clusters and each cluster having approximately equal degree (we call these “balanced clusterings”). Cluster number selection is performed by balancing these two opposite effects in the maximization of Q_N .

In Section 4.1, we exploit the behavior of Q_0 and construct examples in which modularity maximization yields arbitrarily inaccurate clusterings. More specifically, we construct a class of graphs G_{K,N_1,N_2} (where K, N_1, N_2 are parameters of the graph) with the following properties.

1. Each graph G_{K,N_1,N_2} has a “natural” clustering \mathbf{V}_{K,N_1,N_2} (which, however, does not maximize modularity).
2. We can find graphs G_{K,N_1,N_2} and clusterings $\mathbf{U}_{K,N_1,N_2,J}$ such that, by appropriate selection of K, N_1, N_2 and J , the following hold¹:

¹ Note that in the following remarks we are talking about the modularity $Q_N(\mathbf{V}, G)$ of a clustering/graph pair (\mathbf{V}, G) . Indeed, as will be seen in Section 3, the computation of $Q_N(\mathbf{V}, G)$ involves the adjacency matrix of the graph and the classes assigned by the clustering, i.e., $Q_N(\mathbf{V}, G)$ depends on both \mathbf{V} and G .

^a e-mail: kehagiat@auth.gr

- the pair $(G_{K,N_1,N_2}, \mathbf{U}_{K,N_1,N_2,J})$ has higher modularity than the pair $(G_{K,N_1,N_2}, \mathbf{V}_{K,N_1,N_2})$;
- the modularity of $(G_{K,N_1,N_2}, \mathbf{U}_{K,N_1,N_2,J})$ can become (by appropriate selection of J) arbitrarily close to one;
- the Jaccard similarity between clusterings \mathbf{V}_{K,N_1,N_2} and $\mathbf{U}_{K,N_1,N_2,J}$ can become (by appropriate selection of J) arbitrarily close to zero (hence \mathbf{V}_{K,N_1,N_2} and $\mathbf{U}_{K,N_1,N_2,J}$ are arbitrarily different in the Jaccard sense).

We prove similar results for another class of graphs in Section 4.2.

Finally, in Section 5 we discuss the implications of our results and (previously published) related work by other authors and propose some future research directions.

2 Preliminaries

1. A graph G is a pair (V, E) , where V is the node set (we will always assume $V = \{1, 2, \dots, n\}$; hence the number of nodes is $n = |V|$) and $E \subseteq \{\{u, v\} : u, v \in V\}$ is the edge set (and $m = |E|$ is the number of edges). In this paper we will deal with finite graphs without multiple edges and loops.
2. The adjacency matrix of G is an $n \times n$ matrix A with $A_{u,v} = 1$ if $\{u, v\} \in E$ and 0 otherwise. There is a one-to-one correspondence between a graph G and its adjacency matrix A .
3. A clustering of $G = (V, E)$ is a partition $\mathbf{V} = \{V_1, \dots, V_K\}$ of V . The clusters are the node sets V_1, \dots, V_K , which satisfy $\cup_{k=1}^K V_k = V$ and $\forall k, l : V_k \cap V_l = \emptyset$. The size of the clustering is K , the number of clusters. Given a graph $G = (V, E)$, we denote by \mathcal{V} the set of all clusterings of V and by \mathcal{V}_K the set of clusterings of size K . Sometimes we call V_k a community; this is simply a synonym of “cluster”.
4. Given a clustering $\mathbf{V} = \{V_1, \dots, V_K\}$ of the graph $G = (V, E)$, we define the following edge sets ($k = 1, \dots, K$):

$$E_k = \{\{u, v\} : u, v \in V_k \text{ and } \{u, v\} \in E\},$$

i.e., E_k is the set of edges with both ends being nodes of V_k . The edges contained in $\cup_{k=1}^K E_k$ are the intracenter edges; the remaining edges, i.e., the ones contained in $E - \cup_{k=1}^K E_k$ are the extracenter edges.

5. The degree function $\deg(\cdot) : V \rightarrow \mathbb{Z}$ is defined as follows: for any $v \in V$, $\deg(v) = |\{\{v, w\} : \{v, w\} \in E\}|$ is the number of edges incident on v ; we also define, for any $U \subseteq V$, $\deg(U) = \sum_{v \in U} \deg(v)$, i.e., the sum of degrees of the nodes contained in U .
6. The Jaccard similarity index is defined as follows. Given any two clusterings $\mathbf{W}_1, \mathbf{W}_2$ define

- a_{11} = “num. of node pairs $\{u, v\}$ in same cluster under \mathbf{W}_1 and same cluster under \mathbf{W}_2 ”;
- a_{10} = “num. of node pairs $\{u, v\}$ in same cluster under \mathbf{W}_1 and different cluster under \mathbf{W}_2 ”;
- a_{01} = “num. of node pairs $\{u, v\}$ in different cluster under \mathbf{W}_1 and same cluster under \mathbf{W}_2 ”.

Then the Jaccard similarity index $S(\mathbf{W}_1, \mathbf{W}_2)$ is defined by:

$$S(\mathbf{W}_1, \mathbf{W}_2) = \frac{a_{11}}{a_{10} + a_{01} + a_{11}}.$$

$S(\mathbf{W}_1, \mathbf{W}_2)$ takes values in $[0, 1]$; values close to 1 show that $\mathbf{W}_1, \mathbf{W}_2$ are very similar; values close to 0 that they are very different.

3 An interpretation of modularity

3.1 Modularity

Given a graph $G = (V, E)$ with adjacency matrix A , we denote the modularity of a clustering \mathbf{V} by $Q_N(\mathbf{V}, G)$ and, following [1], we define it by:

$$Q_N(\mathbf{V}, G) = \frac{1}{2m} \sum_{i,j \in V} \left(A_{ij} - \frac{\deg(i) \deg(j)}{2m} \right) \Delta(i, j), \quad (1)$$

where $\Delta(i, j)$ equals one if i and j belong to the same cluster and zero otherwise. Our notation emphasizes that $Q_N(\mathbf{V}, G)$ is a function of both the graph and the clustering (cmp. to footnote 1). In other words, the value $Q_N(\mathbf{V}, G)$ characterizes the pair (\mathbf{V}, G) , not just the graph G .

The motivation for introducing modularity originates in the fact that $Q_N(\mathbf{V}, G)$ measures the fraction of intracenter edges in G minus the expected value of the same quantity in a graph G' with the same clusters but random connections between the nodes; G' is often called the null model². $Q_N(\mathbf{V}, G)$ can be either positive or negative, with positive values indicating the possible presence of community structure. Thus, one can search for community structure by looking for the partitions of a graph that have positive, and preferably large, values of the modularity. This is the justification of graph clustering by modularity maximization, as presented in reference [7], from which we have paraphrased most of the above remarks. The way we understand Newman’s argument is that, by definition, \mathbf{V} is a better clustering of G than \mathbf{V}' iff $Q_N(\mathbf{V}, G) > Q_N(\mathbf{V}', G)$ and the overall best clustering of G is $\mathbf{V}^* = \arg \max_{\mathbf{V}} Q_N(\mathbf{V}, G)$. Furthermore, according to the above reasoning, a large value $Q_N(\mathbf{V}^*, G)$ should indicate both (i) that \mathbf{V}^* is a good clustering of G and (ii) G has strong community structure. Hence modularity is a clustering quality function (CQF) in the sense of reference [8].

There are reasons to doubt the above conclusions. For example, it is not clear exactly what is a “large $Q_N(\mathbf{V}, G)$ value”. While it is known [9] that $-\frac{1}{2} \leq Q_N(\mathbf{V}, G) \leq 1$ for every pair (\mathbf{V}, G) , examples appear in the community detection literature [7] of graphs which have strong (intuitively perceived) community structure and yet their maximum modularity is closer to zero than to one. On the

² Note that the intracenter edge fraction of both G and G' is computed with respect to \mathbf{V} .

other hand, graphs exist which do not have an intuitively obvious modular structure and yet can achieve high modularity values. For example, in reference [10] it is shown that trees and treelike networks can achieve high modularity values, despite the fact that trees are sparsely connected and are not generally considered to possess modular structure. In reference [11] it is shown that graph classes such as tori and hypercubes (which do not have any obvious modular structure) can asymptotically achieve the maximum possible modularity value, namely one.

An additional shortcoming of modularity maximization is its tendency to either underestimate or overestimate the number of clusters in a graph; this fact has been widely reported in the literature; we will also discuss it in Section 3.3.

A frequently proposed explanation for the shortcomings of modularity is that the use of the null model is not well justified [8]. In Section 3.3 we will consider an alternative, complementary explanation. But first we will examine another CQF.

3.2 Intracluster edge fraction

A popular characterization of a graph community is that “there must be more edges ‘inside’ the community than edges linking vertices of the community with the rest of the graph” [8, Section III-B.1]. Variations of this principle have been stated by several authors³.

A *prima facie* reasonable way to quantify the principle is through the intracluster edge fraction, denoted by $Q_f(\mathbf{V}, G)$ and defined by:

$$Q_f(\mathbf{V}, G) = \frac{\sum_{k=1}^K |E_k|}{m}. \quad (2)$$

For every G and \mathbf{V} , $Q_f(\mathbf{V}, G) \in [0, 1]$. A high (i.e., close to 1) value of $Q_f(\mathbf{V}, G)$ indicates that the pair (\mathbf{V}, G) has many intracluster and few extracluster edges.

Unfortunately, a high $Q_f(\mathbf{V}, G)$ value does not guarantee either that G has strong community structure or that \mathbf{V} is a good clustering of G . Indeed we can always achieve the maximum value $Q_f(\mathbf{V}, G) = 1$ by taking $\mathbf{V} = \{V\}$ (i.e., the unique clustering of size one) but this tells us nothing about the “true” community structure of G . This observation can be generalized. First define the following function:

$$F_G(K) = \max_{\mathbf{V} \in \mathcal{V}_K} Q_f(\mathbf{V}, G). \quad (3)$$

In words, for a given graph G , $F_G(K)$ is the maximum intracluster edge fraction achieved by clusterings of size K . Now we can prove the following.

Theorem 3.1. *For any graph $G = (V, E)$, $F_G(K)$ is a nonincreasing function of K .*

³ An extreme statement of this idea appears in reference [12]: “a community network $G_0 = (V, E_0)$ [is] a graph G_0 that is a disjoint union of complete subgraphs”.

Proof. There exists a single clustering of size one, namely $\mathbf{V}^{(1)} = \{V\}$. Denote the set of intracluster edges by $E_1^{(1)}$; obviously $E_1^{(1)} = E$ (i.e., all edges are intracluster). Hence

$$F_G(1) = \frac{|E_1^{(1)}|}{|E|} = 1.$$

Let $\mathbf{V}^{(K)} = \{V_1^{(K)}, V_2^{(K)}, \dots, V_K^{(K)}\}$ be the optimal clustering of size K ; the intracluster edge sets are $E_1^{(K)}, \dots, E_K^{(K)}$. Create a clustering \mathbf{V}' of size $K-1$ by merging $V_{K-1}^{(K)}$ and $V_K^{(K)}$. In other words

$$\mathbf{V}' = \{V_1^{(K)}, V_2^{(K)}, \dots, V_{K-2}^{(K)}, V_{K-1}^{(K)} \cup V_K^{(K)}\}.$$

Under \mathbf{V}' the intracluster edges are

$$E'_1 = E_1^{(K)}, \dots, E'_{K-2} = E_{K-2}^{(K)}, E'_{K-1}.$$

We have

$$E_{K-1}^{(K)} \cup E_K^{(K)} \subseteq E'_{K-1}$$

and

$$|E_{K-1}^{(K)}| + |E_K^{(K)}| \leq |E'_{K-1}|.$$

Hence

$$\begin{aligned} F_G(K) &= Q_f(\mathbf{V}^{(K)}, G) = \frac{\sum_{k=1}^K |E_k^{(K)}|}{|E|} \\ &\leq \frac{\sum_{k=1}^{K-2} |E_k^{(K)}|}{|E|} + \frac{|E'_{K-1}|}{|E|} = Q_f(\mathbf{V}', G). \end{aligned}$$

But

$$Q_f(\mathbf{V}', G) \leq \max_{\mathbf{V} \in \mathcal{V}_{K-1}} Q_f(\mathbf{V}, G) = F_G(K-1).$$

It follows that

$$0 \leq F_G(n) \leq \dots \leq F_G(2) \leq F_G(1) = 1$$

and the proof is complete.

Hence, for any G , $Q_f(\mathbf{V}, G)$ is maximized at $K = 1$ and this gives us no information about the actual community structure of G . In other words, Theorem 3.1 implies that Q_f maximization cannot determine the optimal number of clusters. On the other hand, if K is given in advance (as a parameter) then $\mathbf{V}^{(K)} = \arg \max_{\mathbf{V} \in \mathcal{V}_K} Q_f(\mathbf{V}, G)$ is a reasonable candidate for the best clustering of size K . This has sometimes been phrased as a criticism of community detection by Q_f maximization. For instance, in reference [8] is stated that “Algorithms for graph partitioning are not good for community detection, because it is necessary to provide as input the number of groups”. However, this criticism is valid only to the extent that other algorithms exist which can obtain the true number of groups (clusters). For example, an alleged advantage of modularity is that its maximization yields the correct number of clusters; let us now discuss this claim.

3.3 Modularity as augmented intracluster edge fraction

The claim that modularity maximization can determine the true number of clusters has been put in doubt by the discovery of the modularity resolution limit. As explained in references [2,13] and several other papers, there exist graphs G for which the clustering obtained by maximizing modularity has fewer clusters than the “intuitively correct” clustering of G . In other words, modularity maximization can underestimate the number of clusters. In addition, modularity maximization can overestimate the number of clusters. Some explanations of this fact have been presented in the literature (see, e.g., [5]). We will now present an intuitive and (to the best of our knowledge) new explanation of cluster number overestimation, which will form the basis of some precise results presented in Section 4.

It is easy to convert (1) to the following (well-known) equivalent form:

$$Q_N(\mathbf{V}, G) = \sum_{k=1}^K \frac{|E_k|}{m} - \sum_{k=1}^K \left(\frac{\deg(V_k)}{2m} \right)^2. \quad (4)$$

Defining

$$Q_0(\mathbf{V}, G) = \sum_{k=1}^K \left(\frac{\deg(V_k)}{2m} \right)^2 \quad (5)$$

we can rewrite (4) as:

$$Q_N(\mathbf{V}, G) = Q_f(\mathbf{V}, G) - Q_0(\mathbf{V}, G). \quad (6)$$

Hence Newman and Girvan’s modularity is the difference of $Q_f(\mathbf{V}, G)$ and the auxiliary function $Q_0(\mathbf{V}, G)$. As already mentioned, the introduction of $Q_0(\mathbf{V}, G)$ is usually motivated by appeal to the null model [1]; we will now present an alternative, complementary view.

Suppose momentarily that K is given and we want to minimize $Q_0(\mathbf{V}, G)$ with respect to $\mathbf{V} = \{V_1, \dots, V_K\}$. For simplicity of notation, define $p_k = \frac{\deg(V_k)}{2m}$; then

$$Q_0(\mathbf{V}, G) = \sum_{k=1}^K \left(\frac{\deg(V_k)}{2m} \right)^2 = \sum_{k=1}^K p_k^2$$

and we also have

$$\sum_{k=1}^K p_k = \sum_{k=1}^K \frac{\deg(V_k)}{2m} = 1.$$

Hence we want to solve the following problem: given K , minimize $\sum_{k=1}^K p_k^2$ subject to:

$$0 \leq p_k \leq 1 \text{ and } \sum_{k=1}^K p_k = 1. \quad (7)$$

Of course there are additional constraints on the p_k ’s: each of them must be obtained by summing the degrees of V_k , which is a set of nodes of the given graph G . However,

assume for the time being that the p_k ’s are continuously valued and must only satisfy the constraints of (7) (these assumptions will be removed a little later). Under these assumptions, the solution to (7) is $p_k = \frac{1}{K}$ for all k ; the minimum thus achieved is $\frac{1}{K}$.

Next consider the problem: minimize $\sum_{k=1}^K p_k^2$ subject to:

$$K \in \{1, \dots, n\}, 0 \leq p_k \leq 1 \text{ and } \sum_{k=1}^K p_k = 1. \quad (8)$$

We can solve (8) by first solving (7) separately for each $K \in \{1, \dots, n\}$ and then looking for the overall minimum; we see that this is $\frac{1}{n}$ and is achieved at $K = n$ and $p_k = \frac{1}{n}$ for all k .

Going back to the minimization of $Q_0(\mathbf{V}, G)$ we note that, in general, the overall minimum $\sum_{k=1}^K p_k^2 = \frac{1}{n}$ will only be achieved under very special circumstances. Namely, if all nodes of G have equal degree, then

$$\min_{\mathbf{V} \in \mathcal{V}} Q_0(\mathbf{V}, G) = Q_0(\mathbf{V}^*, G) = \frac{1}{n}$$

where $\mathbf{V}^* = \{V_1, \dots, V_n\}$ and $V_i = \{i\}$ for $i \in \{1, \dots, n\}$. But even when the nodes of G do not have equal degrees, it seems intuitively obvious that small values of $Q_0(\mathbf{V}, G)$ are achieved by clusterings \mathbf{V} which have many clusters (large K) and distribute nodes between clusters so that $p_k = \frac{\deg(V_k)}{2m}$ is approximately the same for all $k \in \{1, \dots, K\}$. In Section 4, we will see precise examples which justify this intuition.

Let us now apply the above observations to modularity maximization. Since (i) $Q_N = Q_f - Q_0$, (ii) Q_f achieves its maximum at $K = 1$ and (iii) we expect Q_0 to achieve its minimum at or near $K = n$, we conclude that the following factors will influence the outcome of modularity maximization: the Q_f term pulls K towards small values and the Q_0 towards large ones; in addition the Q_f term favors clusterings which correspond to the “natural” community structure of G (i.e., there exist few extracluster edges) while the Q_0 favors “balanced” clusterings (i.e., each cluster has more or less the same degree). The final outcome depends on (among other factors) the relative magnitudes of Q_f and Q_0 . These observations agree with previously published remarks, e.g., that “the existing modularity optimization method does not perform well in the presence of unbalanced community structures” [14] and “for modularity’s null model graphs, the modularity maximum corresponds to an equipartition of the graph” [8]. In particular, the issue of cluster number underestimation (modularity resolution limit) has been discussed in, for example, [2,13], while overestimation has been discussed in references [5,15,16]. We will present our own analysis of cluster number overestimation in Section 4.

Let us note, in concluding this section, that one method used to address the modularity resolution limit is to introduce a modified modularity function. This function is often written in the form

$$Q(\mathbf{V}, G; \gamma) = Q_f(\mathbf{V}, G) - \gamma Q_0(\mathbf{V}, G)$$

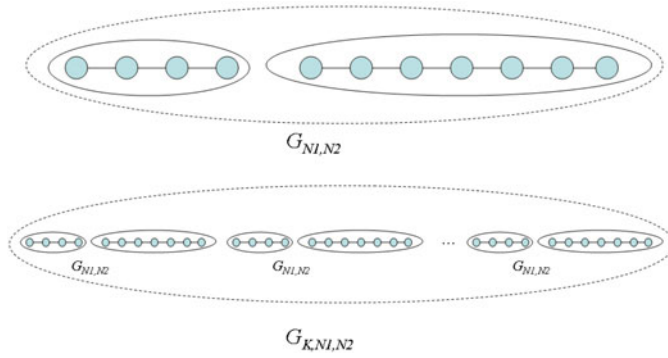


Fig. 1. Graph family G_{K,N_1,N_2} .

where γ is a “tuning parameter” (see Refs. [3,17–20] and also [21,22]). With $\gamma = 1$, $Q(\mathbf{V}, G; 1) = Q_N(\mathbf{V}, G)$, the original Newman and Girvan’s modularity. If this underestimates (resp. overestimates) the “true” number of clusters, formation of more (resp. fewer) clusters can be encouraged by increasing (resp. decreasing) γ and hence the influence of the $Q_0(\mathbf{V}, G)$ term on the maximization problem. However, it seems that no “universally correct” value of γ exists; in other words, the resolution limit can occur for any γ value [19,20].

4 Bad clusterings with high modularity

In this section we construct graphs admitting (i) a “natural” clustering and (ii) a sequence of “arbitrarily bad” clusterings which achieve higher modularity than the natural one. In fact, as we will see, the arbitrarily bad clusterings can achieve modularity arbitrarily close to one and they can be “arbitrarily different” from the natural clustering (we will presently explain precisely what we mean by the terms “natural”, “arbitrarily bad” and “arbitrarily different”). These results indicate that, at least in certain cases, modularity is not a good CQF.

4.1 First example

To establish the abovementioned results, we will construct a family of graphs G_{K,N_1,N_2} (where K, N_1, N_2 are parameters) such that the graph G_{K,N_1,N_2} has an easily recognized “natural” clustering \mathbf{V}_{K,N_1,N_2} (for every K, N_1, N_2).

We define G_{K,N_1,N_2} as follows. First, for any N_1, N_2 we define the disconnected graph G_{N_1,N_2} to be the union of a path of N_1 nodes and a path of N_2 nodes; second, we let the disconnected graph G_{K,N_1,N_2} be the union of K copies of G_{N_1,N_2} . The construction is illustrated in Figure 1.

We claim that the natural clustering of G_{K,N_1,N_2} is $\mathbf{V}_{K,N_1,N_2} = \{V_{K,N_1,N_2,1}, V_{K,N_1,N_2,2}, \dots, V_{K,N_1,N_2,2K}\}$, where $V_{K,N_1,N_2,k}$ is the node set of the k th connected component of G (with $k \in \{1, 2, \dots, 2K\}$, see Fig. 1). At the risk of belaboring the obvious, we note that, if $u \in V_{K,N_1,N_2,i}$ and $v \in V_{K,N_1,N_2,j}$ (with $i \neq j$) then there exists no path connecting u and v ; hence they should never

be put in the same cluster. So the biggest possible clusters are the $V_{K,N_1,N_2,i}$ ’s. On the other hand, there is no justification for splitting some $V_{K,N_1,N_2,i}$ at any particular edge, since all edges (except the border edges) have the same connectivity pattern, i.e., the i th edge connects nodes i and $i+1$. Hence \mathbf{V}_{K,N_1,N_2} is the “intuitively best” (i.e., the “natural”) clustering of G_{K,N_1,N_2} .

Lemma 4.1. *For every $K, N_1, N_2 \in \mathbb{N}$ with $N_1, N_2 \geq 3$ and $J \leq n = K(N_1 + N_2)$ we have*

$$Q_N(\mathbf{V}_{K,N_1,N_2}, G_{K,N_1,N_2}) = 1 - \frac{(N_1 - 1)^2 + (N_2 - 1)^2}{K(N_1 + N_2 - 2)^2}. \quad (9)$$

Proof. We fix K, N_1, N_2 and, for brevity, we write G for G_{K,N_1,N_2} and \mathbf{V} for \mathbf{V}_{K,N_1,N_2} . We have

$$Q_N(\mathbf{V}, G) = \frac{\sum_{k=1}^{2K} |E_k|}{m} - \frac{\sum_{k=1}^{2K} (\deg(V_k))^2}{(2m)^2}.$$

Under \mathbf{V} , G has no extracenter edges hence we have

$$\frac{\sum_{k=1}^{2K} |E_k|}{m} = 1. \quad (10)$$

We can separate \mathbf{V} into two subsets of clusters: $\mathbf{V}' = \{V_1, V_3, \dots, V_{2K-1}\}$ contains the clusters with N_1 nodes and $\mathbf{V}'' = \{V_2, V_4, \dots, V_{2K}\}$ contains the clusters with N_2 nodes. Each $V_k \in \mathbf{V}'$ has $N_1 - 2$ “inner nodes” of degree 2 and two “border nodes” of degree 1; similarly, each $V_k \in \mathbf{V}''$ has $N_2 - 2$ inner nodes and 2 border nodes. Hence

$$\begin{aligned} \forall V_k \in \mathbf{V}' : \deg(V_k) &= 2(N_1 - 2) + 2 = 2(N_1 - 1) \\ \forall V_k \in \mathbf{V}'' : \deg(V_k) &= 2(N_2 - 2) + 2 = 2(N_2 - 1). \end{aligned}$$

The total number of edges is

$$\begin{aligned} m &= \frac{\sum_{V_k \in \mathbf{V}} \deg(V_k)}{2} \\ &= \frac{\sum_{V_k \in \mathbf{V}'} \deg(V_k) + \sum_{V_k \in \mathbf{V}''} \deg(V_k)}{2} \\ &= K(N_1 + N_2 - 2). \end{aligned}$$

Also,

$$\begin{aligned} \frac{\sum_{k=1}^{2K} (\deg(V_k))^2}{(2m)^2} &= \frac{\sum_{V_k \in \mathbf{V}'} (2(N_1 - 1))^2}{(2K(N_1 + N_2 - 2))^2} \\ &\quad + \frac{\sum_{V_k \in \mathbf{V}''} (2(N_2 - 1))^2}{(2K(N_1 + N_2 - 2))^2} \\ &= \frac{K(N_1 - 1)^2 + K(N_2 - 1)^2}{K^2(N_1 + N_2 - 2)^2} \\ &= \frac{(N_1 - 1)^2 + (N_2 - 1)^2}{K(N_1 + N_2 - 2)^2}. \quad (11) \end{aligned}$$

Combining (10) and (11) we get (9).

Let us now introduce the “bad clusterings”. For every triple (K, N_1, N_2) , we define a sequence $\{\mathbf{U}_{K, N_1, N_2, J}\}_{J=1}^n$ of clusterings of G_{K, N_1, N_2} . For a fixed J , let $L = \lfloor \frac{n}{J} \rfloor$; writing for brevity \mathbf{U}_J in place of $\mathbf{U}_{K, N_1, N_2, J}$, we let $\mathbf{U}_J = \{U_1, \dots, U_J, U_{J+1}\}$ consist of the following $J+1$ clusters:

$$\begin{aligned} U_1 &= \{1, \dots, L\}, & U_2 &= \{L+1, \dots, 2L\}, \dots, \\ U_J &= \{(J-1)L+1, \dots, JL\}, \\ U_{J+1} &= \{JL+1, \dots, n\}; \end{aligned}$$

if $n = JL$ then $U_{J+1} = \emptyset$. In other words, \mathbf{U}_J contains J clusters each containing the same number of nodes (namely $L = \lfloor \frac{n}{J} \rfloor$) and perhaps an additional cluster (with fewer than L nodes). Obviously \mathbf{U}_J is a “well balanced” clustering.

Lemma 4.2. *For every $K, N_1, N_2, J \in \mathbb{N}$ with $N_1, N_2 \geq 3$ we have*

$$Q_N(\mathbf{U}_{K, N_1, N_2, J}, G_{K, N_1, N_2}) \geq 1 - \frac{1}{K(N_1 + N_2 - 2)}J - \frac{2(N_1 + N_2)^2}{(N_1 + N_2 - 2)^2}J^{-1}. \quad (12)$$

Proof. We write G for G_{K, N_1, N_2} and \mathbf{U}_J for $\mathbf{U}_{K, N_1, N_2, J}$. We have

$$Q_N(\mathbf{U}_J, G) = \frac{\sum_{k=1}^{J+1} |E_k|}{m} - \frac{\sum_{k=1}^{J+1} (\deg(U_k))^2}{(2m)^2}.$$

Consider first $\frac{\sum_{k=1}^{J+1} |E_k|}{m}$. A little thought shows that \mathbf{U}_J has at most $J+1$ clusters and J extracenter edges. Hence

$$\begin{aligned} \forall J : \frac{\sum_{k=1}^{J+1} |E_k|}{m} &\geq \frac{m-J}{m} = 1 - \frac{J}{m} \\ &= 1 - \frac{1}{K(N_1 + N_2 - 2)}J. \end{aligned} \quad (13)$$

Consider now $\frac{\sum_{k=1}^{J+1} (\deg(U_k))^2}{(2m)^2}$. Each U_k has no more than $\frac{n}{J} = \frac{K(N_1 + N_2)}{J}$ nodes and each node has degree at most 2. Hence

$$\begin{aligned} \forall J : \frac{\sum_{k=1}^{J+1} (\deg(U_k))^2}{(2m)^2} &\leq \frac{(J+1) \left(2 \frac{K(N_1 + N_2)}{J} \right)^2}{4K^2(N_1 + N_2 - 2)^2} \\ &\leq \frac{2(N_1 + N_2)^2}{(N_1 + N_2 - 2)^2}J^{-1} \end{aligned} \quad (14)$$

(since $\forall J \in \mathbb{N} : \frac{J+1}{J} \leq 2$). Combining (13) and (14) we get (12).

To ensure that

$$Q_N(\mathbf{U}_{K, N_1, N_2, J}, G_{K, N_1, N_2}) > Q_N(\mathbf{V}_{K, N_1, N_2}, G_{K, N_1, N_2})$$

(i.e., that the natural clustering \mathbf{V}_{K, N_1, N_2} has lower modularity than $\mathbf{U}_{K, N_1, N_2, J}$) it suffices to select K, N_1, N_2, J

appropriately and use Lemmas 4.1 and 4.2. A sufficient condition, obtained from (9) and (12), is

$$1 - \frac{1}{K(N_1 + N_2 - 2)}J - \frac{2(N_1 + N_2)^2}{(N_1 + N_2 - 2)^2}J^{-1} > 1 - \frac{(N_1 - 1)^2 + (N_2 - 1)^2}{K(N_1 + N_2 - 2)^2}. \quad (15)$$

Inspecting (15), we see that one way to satisfy it is by fixing N_1 and letting J be “sufficiently larger” than K and N_2 “sufficiently larger” than J . This is the main idea used in the proof of the following theorem.

Theorem 4.3. *For every $K \in \mathbb{N}$ and $\varepsilon \in (0, \frac{1}{2K})$ there exist $N_1, N_2, J \in \mathbb{N}$ (depending on ε and K) such that*

$$Q_N(\mathbf{V}_{K, N_1, N_2}, G_{K, N_1, N_2}) < 1 - \frac{1}{2K} < 1 - \varepsilon < Q_N(\mathbf{U}_{K, N_1, N_2, J}, G_{K, N_1, N_2}), \quad (16)$$

$$S(\mathbf{V}_{K, N_1, N_2}, \mathbf{U}_{K, N_1, N_2, J}) < \varepsilon. \quad (17)$$

Proof. Take any K and let $N_1 = 3$, $J = xK$, $N_2 = x^2K$ (with $x \in \mathbb{N}$). To prove (16) note that

$$Q_N(\mathbf{V}_{K, N_1, N_2}, G_{K, N_1, N_2}) = 1 - \frac{4 + (x^2K - 1)^2}{K(1 + x^2K)^2}$$

and

$$Q_N(\mathbf{U}_{K, N_1, N_2, J}, G_{K, N_1, N_2}) \geq 1 - \frac{x}{(1 + x^2K)} - \frac{2(3 + x^2K)^2}{(1 + x^2K)^2 xK}.$$

Define $z = \frac{1}{x}$; then we have $x = \frac{1}{z}$ and

$$\begin{aligned} Q_N(\mathbf{V}_{K, N_1, N_2}, G_{K, N_1, N_2}) &= 1 - \frac{4 + (x^2K - 1)^2}{K(1 + x^2K)^2} \\ &= 1 - \frac{4 + ((1/z)^2 K - 1)^2}{K(1 + (1/z)^2 K)^2}. \end{aligned} \quad (18)$$

We can simplify the final $Q_N(\mathbf{V}_{K, N_1, N_2}, G_{K, N_1, N_2})$ expression of (18) and write it as the following function:

$$f_1(z) = \frac{K^3 - K^2 + 2(K + K^2)z^2 + (K - 5)z^4}{K(z^2 + K)^2}.$$

Now, $1 - \frac{4 + ((1/z)^2 K - 1)^2}{K(1 + (1/z)^2 K)^2}$ has a removable singularity at $z_0 = 0$, but for every other $z \in \mathbb{R}$ it is identical to $f_1(z)$. We can expand $f_1(z)$ in a Taylor series around $z_0 = 0$ which will also hold for $Q_N(\mathbf{V}_{K, N_1, N_2}, G_{K, N_1, N_2})$. Hence around $z_0 = 0$ we have

$$Q_N(\mathbf{V}_{K, N_1, N_2}, G_{K, N_1, N_2}) = 1 - \frac{1}{K} + r_1(z),$$

where $r_1(z) = a_2 z^2 + a_3 z^3 + \dots$ and, from the Taylor series remainder theorem, there exists a constant A such that, for z close to zero, we have

$$|r_1(z)| < Az^2.$$

Then, for large finite x (and, in particular, for $x > \sqrt{2KA}$) we have

$$\begin{aligned} Q_N(\mathbf{V}_{K,N_1,N_2}, G_{K,N_1,N_2}) &= 1 - \frac{4 + (x^2 K - 1)^2}{K(1 + x^2 K)^2} \\ &< 1 - \frac{1}{K} + \frac{A}{x^2} < 1 - \frac{1}{2K}. \end{aligned} \quad (19)$$

Similarly (with $z = \frac{1}{x}$) we have

$$\begin{aligned} Q_N(\mathbf{U}_{K,N_1,N_2,J}, G_{K,N_1,N_2}) &= 1 - \frac{xK}{K(1 + x^2 K)} - \frac{2(3 + x^2 K)^2}{(1 + x^2 K)^2 xK} \\ &= 1 - \frac{(1/z)}{(1 + (1/z)^2 K)} - \frac{2(3 + (1/z)^2 K)^2}{(1 + (1/z)^2 K)^2 (1/z) K}. \end{aligned} \quad (20)$$

Again, we can rewrite the final $Q_N(\mathbf{U}_{K,N_1,N_2,J}, G_{K,N_1,N_2})$ expression of (20) as:

$$f_2(z) = \frac{K^3 - 3K^2 z + 2K^2 z^2 - 13Kz^3 + Kz^4 - 18z^5}{K(z^2 + K)^2}$$

and $1 - \frac{(1/z)}{(1 + (1/z)^2 K)} - \frac{2(3 + (1/z)^2 K)^2}{(1 + (1/z)^2 K)^2 (1/z) K}$ has a removable singularity at $z_0 = 0$, but for every other $z \in \mathbb{R}$ it is identical to $f_2(z)$. Hence we can expand $f_2(z)$ in a Taylor series around $z_0 = 0$, which will also hold for $Q_N(\mathbf{U}_{K,N_1,N_2,J}, G_{K,N_1,N_2})$. Around $z_0 = 0$ we have

$$Q_N(\mathbf{U}_{K,N_1,N_2,J}, G_{K,N_1,N_2}) = 1 - \frac{3}{K}z + r_2(z)$$

where $r_2(z) = b_3 z^3 + b_4 z^4 + \dots$ and there exists a constant B such that, for z close to zero, we have

$$|r_2(z)| < Bz^3 < Bz^2;$$

this in turn implies that

$$r_2(z) > -Bz^2.$$

Then, for large x (and, in particular, for $x > KB$) we have

$$\begin{aligned} Q_N(\mathbf{U}_{K,N_1,N_2,J}, G_{K,N_1,N_2}) &= 1 - \frac{xK}{K(1 + x^2 K)} - \frac{2(3 + x^2 K)^2}{(1 + x^2 K)^2 xK} \\ &> 1 - \frac{3}{Kx} - \frac{B}{x^2} > 1 - \frac{4}{Kx}. \end{aligned} \quad (21)$$

For any $\varepsilon \in (0, \frac{1}{2K})$, choose any x such that

$$x > \max\left(\frac{4}{K\varepsilon}, \sqrt{2KA}, KB\right);$$

then we have $\frac{1}{2K} > \varepsilon > \frac{4}{Kx}$ which, combined with (19) and (21), gives

$$\begin{aligned} Q_N(\mathbf{U}_{K,N_1,N_2,J}, G_{K,N_1,N_2}) &> 1 - \frac{4}{Kx} > 1 - \varepsilon > 1 - \frac{1}{2K} \\ &> Q_N(\mathbf{V}_{K,N_1,N_2}, G_{K,N_1,N_2}). \end{aligned}$$

In short, we can satisfy (16) for every $K \in \mathbb{N}$ and every $\varepsilon \in (0, \frac{1}{2K})$, by taking x “sufficiently large” and $N_1 = 3$, $J = xK$, $N_2 = x^2 K$.

We now turn to (17). Let b (resp. c) be the number of node pairs in the same cluster under $\mathbf{U}_{K,N_1,N_2,J}$ (resp. under \mathbf{V}_{K,N_1,N_2}). We obviously have $b = a_{01} + a_{11} \geq a_{11}$ and $a_{10} + a_{01} + a_{11} \geq a_{10} + a_{11} = c > 0$. Hence

$$S(\mathbf{U}_{K,N_1,N_2,J}, \mathbf{V}_{K,N_1,N_2}) = \frac{a_{11}}{a_{10} + a_{01} + a_{11}} \leq \frac{b}{c}.$$

We first obtain an upper bound for b . Since each U_j contains no more than $L = \frac{n}{J}$ nodes, the number of node pairs that can be formed in U_j is no more than $\frac{(\frac{n}{J})(\frac{n}{J}-1)}{2} < \frac{n^2/2}{J^2}$. Also, $n = K(N_1 + N_2)$ so, for big N_2 , $\frac{n^2/2}{J^2} < \frac{(2KN_2)^2}{J^2}$. There are at most $J + 1$ clusters, so we have

$$\begin{aligned} b &< (J + 1) \frac{(2KN_2)^2}{J^2} \\ &= (xK + 1) \frac{(2Kx^2 K)^2}{(xK)^2} = 4K^3 x^3 + 4K^2 x^2. \end{aligned}$$

Next we compute c . In \mathbf{V}_{K,N_1,N_2} there exist K clusters of $N_1 = 3$ nodes and each cluster has $\frac{N_1(N_1-1)}{2} = 3$ node pairs; there also exist K clusters of N_2 nodes and each cluster has $\frac{N_2(N_2-1)}{2}$ node pairs. We have

$$\begin{aligned} c &= 3K + K \frac{N_2(N_2-1)}{2} \\ &= 3K + K \frac{x^2 K(x^2 K - 1)}{2} \\ &= \frac{1}{2} K^3 x^4 - \frac{1}{2} K^2 x^2 + 3K. \end{aligned}$$

And so we have

$$\begin{aligned} 0 &\leq S(\mathbf{U}_{K,N_1,N_2,J}, \mathbf{V}_{K,N_1,N_2}) < \frac{4K^3 x^3 + 4K^2 x^2}{\frac{1}{2} K^3 x^4 - \frac{1}{2} K^2 x^2 + 3K} \\ &\Rightarrow 0 \leq \lim_{x \rightarrow \infty} S(\mathbf{U}_{K,N_1,N_2,J}, \mathbf{V}_{K,N_1,N_2}) \\ &\leq \lim_{x \rightarrow \infty} \frac{4K^3 x^3 + 4K^2 x^2}{\frac{1}{2} K^3 x^4 - \frac{1}{2} K^2 x^2 + 3K} = 0. \end{aligned}$$

Hence, for every $\varepsilon > 0$ and x sufficiently large, (17) is satisfied.

We see from (16) that we can always find a clustering $\mathbf{U}_{K,N_1,N_2,J}$ which achieves higher modularity than the natural clustering \mathbf{V}_{K,N_1,N_2} and, in fact, greater than $1-\varepsilon$, where ε can get arbitrarily small independently of K . On the other hand, $Q_N(\mathbf{V}_{K,N_1,N_2}, G_{K,N_1,N_2})$ is no greater than $1 - \frac{1}{2K}$; for small K values this can be appreciably less than one. In other words, we can choose K so that G_{K,N_1,N_2} does not have very high “natural modularity” but its “artificial modularity” (the one achieved by the pair $(\mathbf{U}_{K,N_1,N_2,J}, G_{K,N_1,N_2})$) can be arbitrarily close to one.

We see from (17) that, with respect to the Jaccard similarity criterion, $\mathbf{U}_{K,N_1,N_2,J}$ and \mathbf{V}_{K,N_1,N_2} are very different. We could have reached a similar conclusion in a simpler manner. Recall that the number of clusters of $\mathbf{U}_{K,N_1,N_2,J}$ is at least $J = xK$ and we can choose x arbitrarily large; on the other hand, \mathbf{V}_{K,N_1,N_2} has $2K$ clusters. Intuitively, $\mathbf{U}_{K,N_1,N_2,J}$ must be very different from \mathbf{V}_{K,N_1,N_2} , since the ratio of their cluster number is $\frac{x}{2}$ and x can become arbitrarily large (of course the Jaccard similarity index captures this fact in a more precise manner).

Let $\mathbf{V}^* = \arg \max_{\mathbf{V} \in \mathcal{V}} Q_N(\mathbf{V}, G_{K,N_1,N_2})$. While it is conceivable that \mathbf{V}^* is more similar (in the Jaccard sense) to \mathbf{V}_{K,N_1,N_2} than to some $\mathbf{U}_{K,N_1,N_2,J}$, this seems unlikely. In light of the remarks of Section 3.3, it is more likely that \mathbf{V}^* will have many more clusters than \mathbf{V} . In other words, it appears that, for the graphs G_{K,N_1,N_2} , modularity maximization leads to an overestimation of the number of clusters, i.e., we have a case of modularity “over-resolution”.

The bounds utilized in Lemmas 4.1 and 4.2, and Theorem 4.3 are quite conservative. In many cases the inequality

$$Q_N(\mathbf{V}_{K,N_1,N_2}, G_{K,N_1,N_2}) < Q_N(\mathbf{U}_{K,N_1,N_2,J}, G_{K,N_1,N_2}) \quad (22)$$

is attained even when the abovementioned bounds are not satisfied. This can be seen in Table 1, which has been compiled by taking fixed $K = 3$, $N_1 = 3$ and using several x values (recall that $J = xK$, $N_2 = x^2K$). The first six entries of each column list the quantities used in the proof of Theorem 4.3 and, for “sufficiently large” x , should form an increasing sequence, in accordance to the inequalities (15), (16) and (19)–(21). This is indeed the case for $x = 8$ and $x = 10$; on the other hand, for $x = 6$ one inequality is violated (between the third and fourth row) but (22) still holds.

From rows 2 and 7 we see that $Q_N(\mathbf{V}_{K,N_1,N_2}, G_{K,N_1,N_2})$ is a decreasing and $Q_N(\mathbf{U}_{K,N_1,N_2,J}, G_{K,N_1,N_2})$ an increasing function of x . From row 8 we see that the the Jaccard similarity is a decreasing function of x . These observations verify straightforward conclusions which can be drawn from the proof of Theorem 4.3.

4.2 Second example

It might be argued that the results of Section 4.1 are only possible because we have used the disconnected graphs G_{K,N_1,N_2} . This is not the case. In this section we will illustrate the same issues using the family of connected graphs

Table 1. Several quantities appearing in the proof of Theorem 4.3. In each column and for rows 2 to 7, for large enough x , the value of each row must be no less than that of the previous one.

x	6	8	10
$Q_N(\mathbf{V}_{K,N_1,N_2}, G_{K,N_1,N_2})$	0.678	0.673	0.671
$1 - \frac{(N_1-1)^2 + (N_2-1)^2}{K(N_1+N_2-2)^2}$	0.678	0.673	0.671
$1 - \frac{1}{2K}$	0.833	0.833	0.833
$1 - \frac{4}{Kx}$	0.777	0.833	0.866
$1 - \frac{J}{K(N_1+N_2-2)} - \frac{2(N_1+N_2)^2}{J(N_1+N_2-2)^2}$	0.829	0.873	0.899
$Q_N(\mathbf{U}_{K,N_1,N_2,J}, G_{K,N_1,N_2})$	0.891	0.917	0.934
$S(\mathbf{U}_{K,N_1,N_2,J}, \mathbf{V}_{K,N_1,N_2})$	0.154	0.119	0.096

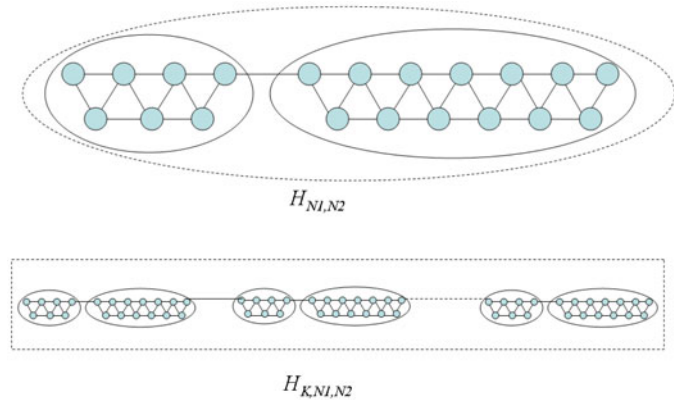


Fig. 2. Graph family H_{K,N_1,N_2} .

H_{K,N_1,N_2} illustrated in Figure 2. We start with connected H_{N_1,N_2} graphs, each of which is a path of $N_1 + N_2$ nodes, with extra edges added between the first N_1 (resp. the second N_2) nodes at distance two of each other. Then we construct the H_{K,N_1,N_2} graphs by joining in series K H_{N_1,N_2} subgraphs.

We will use the same clusterings \mathbf{V}_{K,N_1,N_2} and clustering sequences $\{\mathbf{U}_{K,N_1,N_2,J}\}_{J=1}^n$ as in Section 4.1. Once again, for reasons similar to the ones discussed in Section 4.1, we claim that \mathbf{V}_{K,N_1,N_2} is the natural clustering of H_{K,N_1,N_2} . Namely, cluster boundaries should occur across edges incident on the most weakly connected nodes; this shows that the $V_{K,N_1,N_2,k}$ clusters must be preserved; any partition of $V_{K,N_1,N_2,k}$ into finer clusters cannot be justified, since all of its edges have the same connectivity pattern. Hence \mathbf{V}_{K,N_1,N_2} is the “intuitively best” (i.e., the “natural”) clustering of H_{K,N_1,N_2} .

Once again, we obtain (in three steps) a result similar to Theorem 4.3. First we need two lemmas.

Lemma 4.4. For every $K, N_1, N_2 \in \mathbb{N}$ with $N_1, N_2 \geq 5$ we have

$$Q_N(\mathbf{V}_{K,N_1,N_2}, H_{K,N_1,N_2}) < 1 - \frac{K \left((4N_1 - 8)^2 + (4N_2 - 8)^2 \right)}{(4K(N_1 + N_2 - 2))^2}. \quad (23)$$

Proof. We fix K, N_1, N_2 and, for brevity, we write H for H_{K, N_1, N_2} and \mathbf{V} for \mathbf{V}_{K, N_1, N_2} ; \mathbf{V}' and \mathbf{V}'' have the same meaning as previously. In \mathbf{V}_{K, N_1, N_2} there exist $2K - 1$ extracuster edges, so we have

$$\frac{\sum_{k=1}^{2K} |E_k|}{m} < 1. \quad (24)$$

For each $V_k \in \mathbf{V}'$, there are two border nodes on the left, two border nodes on the right and $N_1 - 4$ inner nodes. Each of the inner nodes has degree 4; each of the border nodes has degree 3, except for the first and last node of the graph, which have degree 2. Hence for each $V_k \in \mathbf{V}'$ we have the bounds

$$\begin{aligned} (N_1 - 4)4 + 4 \times 2 &= 4N_1 - 8 < \deg(V_k) \\ &< 4N_1 - 4 = (N_1 - 4)4 + 4 \times 3. \end{aligned}$$

Similarly, for each $V_k \in \mathbf{V}''$ we have the bounds

$$\begin{aligned} (N_2 - 4)4 + 4 \times 2 &= 4N_2 - 8 < \deg(V_k) \\ &< 4N_2 - 4 = (N_2 - 4)4 + 4 \times 3. \end{aligned}$$

The total number of edges is $m = \frac{\sum_{k=1}^{2K} \deg(V_k)}{2}$ and we have

$$\begin{aligned} \frac{K(4N_1 - 8 + 4N_2 - 8)}{2} &< \frac{\sum_{k=1}^{2K} \deg(V_k)}{2} \\ &< \frac{K(4N_1 - 4 + 4N_2 - 4)}{2} \\ \Rightarrow 2K(N_1 + N_2 - 4) &< m < 2K(N_1 + N_2 - 2). \end{aligned} \quad (25)$$

In addition we have

$$\begin{aligned} K\left((4N_1 - 8)^2 + (4N_2 - 8)^2\right) &< \sum_{k=1}^{2K} (\deg(V_k))^2 \\ &< K\left((4N_1 - 4)^2 + (4N_2 - 4)^2\right). \end{aligned} \quad (26)$$

Combining (25) and (26) we get

$$\frac{\sum_{k=1}^{2K} (\deg(V_k))^2}{(2m)^2} > \frac{K\left((4N_1 - 8)^2 + (4N_2 - 8)^2\right)}{(4K(N_1 + N_2 - 2))^2}. \quad (27)$$

Combining (24) and (27) we get the required bound.

Lemma 4.5. For every $K, N_1, N_2, J \in \mathbb{N}$ with $N_1, N_2 \geq 5$ and $J \leq n = K(N_1 + N_2)$ we have

$$\begin{aligned} Q_N(\mathbf{U}_{K, N_1, N_2, J}, H_{K, N_1, N_2}) \\ > 1 - \frac{3}{2K(N_1 + N_2 - 4)}J - \frac{2(N_1 + N_2)^2}{(N_1 + N_2 - 4)^2}J^{-1}. \end{aligned} \quad (28)$$

Proof. Extracuster edges in \mathbf{U}_J can only occur between successive clusters⁴ U_k, U_{k+1} ; between any such pair there

exist at most three such edges; hence \mathbf{U}_J cannot have more than $3J$ extracuster edges. Consequently

$$\frac{\sum_{k=1}^{J+1} |E_k|}{m} \geq \frac{m - 3J}{m} = 1 - \frac{3J}{m} > 1 - \frac{3J}{2K(N_1 + N_2 - 4)}. \quad (29)$$

Each U_k has at most $\frac{n}{J} = \frac{K(N_1 + N_2)}{J}$ nodes and each node has degree at most 4. Hence

$$\begin{aligned} \frac{\sum_{k=1}^{J+1} (\deg(U_k))^2}{(2m)^2} &\leq \frac{(J+1) \left(4 \frac{K(N_1 + N_2)}{J}\right)^2}{(4K(N_1 + N_2 - 4))^2} \\ &\leq \frac{2(N_1 + N_2)^2}{(N_1 + N_2 - 4)^2}J^{-1}. \end{aligned} \quad (30)$$

Combining (29) and (30) we get the bound (28).

To ensure that

$$Q_N(\mathbf{U}_{K, N_1, N_2, J}, H_{K, N_1, N_2}) > Q_N(\mathbf{V}_{K, N_1, N_2}, H_{K, N_1, N_2})$$

it suffices to choose appropriate K, N_1, N_2, J and use Lemmas 4.4 and 4.5. A sufficient condition, obtained from (23) and (28), is

$$\begin{aligned} 1 - \frac{3}{2K(N_1 + N_2 - 4)}J - \frac{2(N_1 + N_2)^2}{(N_1 + N_2 - 4)^2}J^{-1} \\ > 1 - \frac{K\left((4N_1 - 8)^2 + (4N_2 - 8)^2\right)}{(4K(N_1 + N_2 - 2))^2}. \end{aligned} \quad (31)$$

Now we can prove the following.

Theorem 4.6. For every $K \in \mathbb{N}$ and $\varepsilon \in (0, \frac{1}{2K})$ there exist $N_1, N_2, J \in \mathbb{N}$ (depending on ε, K) such that

$$\begin{aligned} Q_N(\mathbf{V}_{K, N_1, N_2}, H_{K, N_1, N_2}) \\ < 1 - \frac{1}{2K} < 1 - \varepsilon < Q_N(\mathbf{U}_{K, N_1, N_2, J}, H_{K, N_1, N_2}) \end{aligned} \quad (32)$$

$$S(\mathbf{V}_{K, N_1, N_2}, \mathbf{U}_{K, N_1, N_2, J}) < \varepsilon. \quad (33)$$

Proof. Take any K . Letting $N_1 = 6, J = xK, N_2 = x^2K$ we have

$$\begin{aligned} Q_N(\mathbf{V}_{K, N_1, N_2}, H_{K, N_1, N_2}) \\ < 1 - \frac{K(16^2 + (4x^2K - 8)^2)}{(4K(4 + x^2K))^2}, \\ Q_N(\mathbf{U}_{K, N_1, N_2, J}, H_{K, N_1, N_2}) \\ > 1 - \frac{3x}{2(2 + x^2K)} - \frac{2(6 + x^2K)^2}{xK(2 + x^2K)^2}. \end{aligned}$$

Defining $z = \frac{1}{x}$ we have $x = \frac{1}{z}$ and

$$\begin{aligned} 1 - \frac{K(16^2 + (4x^2K - 8)^2)}{(4K(4 + x^2K))^2} \\ = 1 - \frac{K(16^2 + (4(1/z)^2K - 8)^2)}{(4K(4 + (1/z)^2K))^2}. \end{aligned} \quad (34)$$

⁴ There is an exception when $J = n$, but in this case too (29) holds.

Similarly to the proof of Theorem 4.3, there is a function $f_3(z)$ which, for every $z \neq z_0 = 0$, is equal to the right part of (34) and around z_0 has the Taylor expansion

$$f_3(z) = 1 - \frac{1}{K} + r_3(z)$$

where $r_3(z) = c_2 z^2 + c_3 z^3 + \dots$. Furthermore, there exists a constant C such that, for z close to zero, we have

$$|r_3(z)| < C z^2.$$

Then, for large x (and in particular for $x > \sqrt{2KC}$) we have

$$Q_N(\mathbf{V}_{K,N_1,N_2}, H_{K,N_1,N_2}) < 1 - \frac{1}{K} + \frac{C}{x^2} < 1 - \frac{1}{2K}. \quad (35)$$

Similarly, with $z = 1/x$, we have

$$\begin{aligned} & Q_N(\mathbf{U}_{K,N_1,N_2,J}, H_{K,N_1,N_2}) \\ & > 1 - \frac{3xK}{2K(2+x^2K)} - \frac{2(6+x^2K)^2}{xK(2+x^2K)^2} \\ & = 1 - \frac{3(1/z)K}{2K(2+(1/z)^2K)} - \frac{2(6+(1/z)^2K)^2}{(1/z)K(2+(1/z)^2K)^2} \\ & = f_4(z). \end{aligned} \quad (36)$$

Once again, there is a function $f_4(z)$ which, for every $z \neq z_0 = 0$, is equal to the right part of (36) and around z_0 has the Taylor expansion

$$f_4(z) = 1 - \frac{7}{2K}z + r_4(z)$$

where $r_4(z) = d_3 z^3 + d_4 z^4 + \dots$. And there exists a constant D such that, for z close to zero, we have

$$|r_4(z)| < D z^3 < D z^2, \quad r_4(z) > -D z^2.$$

Then, for large x (and, in particular, for $x > 2KD$) we have

$$\begin{aligned} & Q_N(\mathbf{U}_{K,N_1,N_2,J}, H_{K,N_1,N_2}) \\ & > 1 - \frac{3xK}{2K(2+x^2K)} - \frac{2(6+x^2K)^2}{xK(2+x^2K)^2} \\ & > 1 - \frac{7}{2Kx} - \frac{D}{x^2} > 1 - \frac{4}{Kx}. \end{aligned} \quad (37)$$

For any $\varepsilon \in (0, \frac{1}{2K})$ choose any x such that

$$x > \max\left(\frac{4}{K\varepsilon}, \sqrt{2KC}, 2KD\right);$$

then we have $\frac{1}{2K} > \varepsilon > \frac{4}{Kx}$ which, combined with (35) and (37), yields

$$\begin{aligned} & Q_N(\mathbf{U}_{K,N_1,N_2,J}, G_{K,N_1,N_2}) > 1 - \frac{4}{Kx} \\ & > 1 - \varepsilon > 1 - \frac{1}{2K} > Q_N(\mathbf{V}_{K,N_1,N_2}, G_{K,N_1,N_2}). \end{aligned}$$

Table 2. Several quantities appearing in the proof of Theorem 4.6. In each column and for rows 2 to 7, for large enough x , the value of each row must be no less than that of the previous one.

x	6	8	10
$Q_N(\mathbf{V}_{K,N_1,N_2}, G_{K,N_1,N_2})$	0.687	0.678	0.674
$1 - \frac{K((4N_1-8)^2 + (4N_2-8)^2)}{(4K(N_1+N_2-2))^2}$	0.701	0.686	0.679
$1 - \frac{1}{2K}$	0.833	0.833	0.833
$1 - \frac{4}{Kx}$	0.777	0.833	0.866
$1 - \frac{3J}{2K(N_1+N_2-4)} - \frac{2(N_1+N_2)^2}{J(N_1+N_2-4)^2}$	0.798	0.851	0.881
$Q_N(\mathbf{U}_{K,N_1,N_2,J}, G_{K,N_1,N_2})$	0.874	0.898	0.918
$S(\mathbf{U}_{K,N_1,N_2,J}, \mathbf{V}_{K,N_1,N_2})$	0.169	0.118	0.095

In short, we can satisfy (32) for every $K \in \mathbb{N}$ and every $\varepsilon \in (0, \frac{1}{2K})$, by taking x “sufficiently large” and $N_1 = 6$, $J = xK$, $N_2 = x^2K$.

Finally, (33) is exactly the same as (17) and has already been proved.

Similarly to Section 4.1, the bounds utilized in Lemmas 4.4 and 4.5 and Theorem 4.6 are conservative and the inequality

$$Q_N(\mathbf{V}_{K,N_1,N_2}, H_{K,N_1,N_2}) < Q_N(\mathbf{U}_{K,N_1,N_2,J}, H_{K,N_1,N_2}) \quad (38)$$

can be satisfied even when the bounds are violated. This can be seen in Table 2, which is analogous to Table 1 of Section 4.1. We have used $K = 3$, $N_1 = 6$ and several x values. The first six entries of each column list the quantities used in the proof of Theorem 4.3 and, for “sufficiently large” x , should form an increasing sequence. This is the case for $x = 8$ and $x = 10$; for $x = 6$ the sequence is not increasing but (38) holds.

From rows 2 and 7, we see that $Q_N(\mathbf{V}_{K,N_1,N_2}, G_{K,N_1,N_2})$ is decreasing with x and $Q_N(\mathbf{U}_{K,N_1,N_2,J}, G_{K,N_1,N_2})$ is increasing; from row 8 we see that the Jacard similarity is decreasing with x .

5 Discussion and related work

Theorems 4.3 and 4.6 cast doubt on the efficacy of Newman and Girvan’s modularity Q_N as “an objective metric for choosing the number of communities” [1]. Our results are related to those of other authors who have shown that clusterings which achieve high modularity values may be found on very regular graphs (such as tori or hypercubes [11]) or trees and treelike graphs [10], despite the fact that none of these graphs has a “natural community structure”. However, in this paper we have shown that even in graphs which do have a “natural community structure”, high modularity values can be achieved by partitions which do not respect this natural structure. In this sense, our results are more closely connected to those of [5], where it is shown that modularity maximization

can lead to cluster number overestimation when applied to graphs with a natural community structure (rings of rings).

The common characteristic of all the abovementioned graphs (as well as the ones used in the current paper) is local and relatively sparse connectivity. Perhaps the counterintuitive behavior of modularity maximization in such graphs is due to the fact that “modularity, while ostensibly rewarding densely inter-connected groups, can actually be optimized solely through the discovery of bottlenecks” [10]. We believe a more complete explanation requires additional study of the properties of modularity and we pose this as a future research problem. In our opinion, a useful step in this direction will be an axiomatic foundation of the properties that a “reasonable” quality function must possess; we defer the development of such an axiomatic system to a future publication.

Another direction which we believe warrants further research is the evaluation of cluster number selection criteria. The Newman-Girvan modularity was initially introduced with exactly this goal in mind (Ref. [1], Section 4) but we have seen that it can lead to arbitrarily wrong estimates. Perhaps alternative criteria can be found by revisiting the “classic” clustering literature, where cluster number selection has been recognized as “a fundamental, and largely unsolved, problem in cluster analysis” [23] and consequently a large number of cluster number selection criteria have been developed and tested (see for instance [24,25]). The adaptation of such criteria to the community detection problem will not be a trivial problem.

Finally, we believe that our results can be refined. For example, perhaps the bounds of Lemmas 4.2 and 4.4 can be made tighter (or even exact expressions can be obtained) by splitting each component subgraph into a fixed number of clusters. Or tighter results can be used by arranging the subgraphs in a cycle (rather than path) configuration, in which case all nodes will have the same degrees⁵. Once again, we defer the study of these questions to the future.

This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program “Education and Lifelong Learning” of the National Strategic Reference Framework (NSRF) – Research Funding Program: THALIS – UOA (MIS 375891).

References

1. M.E.J. Newman, M. Girvan, Phys. Rev. E **69**, 026113 (2004)
2. S. Fortunato, M. Barthélemy, Proc. Natl. Acad. Sci. **104**, 36 (2007)
3. J. Reichardt, S. Bornholdt, Phys. Rev. E **74**, 016110 (2006)
4. J. Reichardt, S. Bornholdt, Phys. Rev. E **76**, 015102 (2007)
5. M.T. Schaub, J.-C. Delvenne, S.N. Yaliraki, M. Barahona, PloS one **7**, e32210 (2012)
6. M. Rosvall, C.T. Bergstrom, Proc. Natl. Acad. Sci. **105**, 1118 (2008)
7. M.E.J. Newman, Proc. Natl. Acad. Sci. **103**, 8577 (2006)
8. S. Fortunato, Phys. Rep. **486**, 75 (2010)
9. U. Brandes, D. Dellinger, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, D. Wagner, On finding graph clusterings with maximum modularity, *Graph-Theoretic Concepts in Computer Science* (Springer, 2007), pp. 121–132
10. J.P. Bagrow, Phys. Rev. E **85**, 066118 (2012)
11. F. de Montgolfier, M. Soto, L. Viennot, Asymptotic modularity of some graph classes, *Algorithms and Computation* (Springer, 2011), pp. 435–444
12. W.Y.C. Chen, A.W.M. Dress, W.Q. Yu, IET systems biology **1**, 286 (2007)
13. B.H. Good, Y.A. de Montjoye, A. Clauset, Phys. Rev. E **81**, 046106 (2010)
14. S. Zhang, H. Zhao, Phys. Rev. E **85**, 066114 (2012)
15. E. Le Martelot, C. Hankin, in *Proceedings of the 2011 International Conference on Knowledge Discovery and Information Retrieval (KDIR 2011)*, 2011, pp. 216–225.
16. M. Rosvall, C.T. Bergstrom, Proc. Natl. Acad. Sci. **104**, 7327 (2007)
17. J.I. Alvarez-Hamelin, B.M. Gastón, J.R. Busch, [arXiv:1008.3443](https://arxiv.org/abs/1008.3443) (2010)
18. G. Krings, V.D. Blondel, [arXiv:1103.5569](https://arxiv.org/abs/1103.5569) (2011)
19. V.A. Traag, P. Van Dooren, Y. Nesterov, Phys. Rev. E **84**, 016114 (2011)
20. J. Xiang, K. Hu, [arXiv:1108.4244](https://arxiv.org/abs/1108.4244) (2011)
21. J.M. Kumpula, J. Saramäki, K. Kaski, J. Kertesz, Eur. Phys. J. B **56**, 41 (2007)
22. Z. Li, S. Zhang, R.S. Wang, X.S. Zhang, L. Chen, Phys. Rev. E **77**, 036109 (2008)
23. C.A. Sugar, G.M. James, J. Am. Stat. Assoc. **98**, 750 (2003)
24. M. Halkidi, Y. Batistakis, M. Vazirgiannis, ACM Sigmod Record **31**, 40 (2002)
25. M. Halkidi, Y. Batistakis, M. Vazirgiannis, ACM Sigmod Record **31**, 19 (2002)

⁵ We are grateful to the anonymous referee who suggested these improvements.