On the Complexity of Newman's Community Finding Approach for Biological and Social Networks[☆]

Bhaskar DasGupta[a,1,*], Devendra Desai[b]

[a] *Department of Computer Science, University of Illinois at Chicago, Chicago, IL 60607.*
Email: dasgupta@cs.uic.edu
[b] *Department of Computer Science, Rutgers University, Piscataway, NJ 08854.*
Email: devdesai@cs.rutgers.edu

Abstract

Given a graph of interactions, a *module* (also called a *community* or *cluster*) is a subset of nodes whose *fitness* is a function of the *statistical significance* of the pairwise interactions of nodes in the module. The topic of this paper is a *model-based* community finding approach, commonly referred to as *modularity clustering*, that was originally proposed by Newman [25] and has subsequently been extremely popular in practice (*e.g.*, see [1, 20, 28, 30, 32]). Various heuristic methods are currently employed for finding the optimal solution. However, as observed in [1], the exact computational complexity of this approach is still largely unknown.

To this end, we initiate a systematic study of the computational complexity of modularity clustering. Due to the specific quadratic nature of the modularity function, it is necessary to study its value on *sparse* graphs and *dense* graphs *separately*. Our main results include a $(1+\varepsilon)$-inapproximability for dense graphs and a logarithmic approximation for sparse graphs. We make use of several combinatorial properties of modularity to get these results. These are the first non-trivial approximability results beyond the NP-hardness results in [10].

Keywords: Community detection, Modularity clustering, Approximation

[☆] Results in this paper were also presented at the ICALP 2011 workshop on Graph algorithms and Applications, Zurich, Switzerland, July 3, 2011.

[*] Corresponding author.

[1] Supported by DIMACS special focus on Computational and Mathematical Epidemiology. Research partially done while the author was on Sabbatical leave at DIMACS.

algorithms, Approximation hardness, Social networks, Biological networks
2000 MSC: 05C85, 68Q17, 68Q25, 68W25, 68W40, 90C27, 90C35, 90C22, 90C35, 91C20

1. Introduction

Many systems of interaction in biology and social science are modeled as a graph of pairwise interaction of entities [2, 3]. An important problem for these types of graphs is to *partition* the nodes into so-called "communities" or "modules" of "statistically significant" interactions. Such partitions facilitate studying interesting properties of these graph in their applications, such as studying the behavioral patterns of an individual in a societal context, and serve as important components in computational analysis of these graph. In this paper we consider the *static* model of interaction in which the network interconnections do not change over time.

Simplistic definitions of modules, such as *cliques*, unfortunately do not apply well in the context of biological and social networks and therefore alternative definitions are most often used. In the "model-based" community finding approach, one first starts with an appropriate "global null model" $\mathcal{G}$ of a background random graph[2] and then attempts to place nodes in the same module if their interaction patterns are significantly stronger than that inferred from the null model. The null model $\mathcal{G}$ may provide, implicitly or explicitly, the probability $p_{i,j}$ of an edge between two nodes v_i and v_j. As an illustration, suppose that our input is an edge-weighted graph with all weights being positive and normalized between 0 and 1. Then, if $p_{i,j}$ differs significantly from $w_{i,j}$, the weight of the edge between nodes v_i and v_j, the edge may be considered to be *statistically significant*; thus, if $p_{i,j} \ll w_{i,j}$ then it is preferable that v_i and v_j should be placed in the same module whereas if $p_{i,j} \gg w_{i,j}$ then it is preferable that v_i and v_j should be placed in different modules. The standard $\{+,-\}$-correlation clustering that appears in the computer science literature extensively [8, 12, 33] can be placed in the above model-based clustering framework in the following manner: given the input graph G with each edge labeled as $+$ or $-$, let H be the graph consisting of

[2]Of course, any clustering measure that relies on a global null model suffers from the drawback that each node can get attached to any other node of the graph; for another possible drawback see [16]. The purpose of this paper is not to debate on the pros and cons of model-based clustering.

all edges labeled + in G, $p_{i,j} = 0$ (resp. $p_{i,j} = 1$) if the edge was labeled + or missing (resp., labeled −), the modularity of an edge is $a_{i,j} - p_{i,j}$ where $a_{i,j}$ is the $(i,j)^{\text{th}}$ entry in the adjacency matrix of H and the total modularity is a function of individual modularities of edges as induced by the clustering.

In this paper, we investigate a model-based clustering approach originally introduced by Newman and subsequently studied by Newman and others in several papers [25, 28, 30]. The null model in this approach is dependent on the degree distribution of the given graph. *Throughout the paper, by a set of communities (or clusters) we mean a partition $\mathcal{S}$ of the nodes of the graph and, except in Section 5.1, all graphs are undirected.*

1.1. The Basic Setup For Undirected Unweighted Graphs

The basic setup for undirected unweighted graphs as described below can easily be generalized to the case of edge-weighted undirected graphs (see Section 4.3) and edge-weighted directed graphs (see Section 5.1). Let $G = (V, E)$ denote the given input graph with $n = |V|$ nodes and $m = |E|$ edges, let d_v denote the degree of node $v \in V$, and let $A = [a_{u,v}]$ denote the *adjacency matrix* of G, *i.e.*, $a_{u,v} = 1$ if $\{u, v\} \in E$ and $a_{u,v} = 0$ otherwise. The null model $\mathcal{G}$ for modularity clustering is defined by the edge probability function $p_{u,v} = \frac{d_u d_v}{2m}$ for $u, v \in V$ *with* $u = v$ *being allowed*; note that the null model provides a random network such that the *expected degree* of a node v is precisely d_v. Intuitively, if $a_{u,v}$ differs significantly from $p_{u,v}$ then the connection (or, the lack of it) is a significant deviation from the null model. Based on this intuition, the *fitness* of the community formed by a subset of nodes $C \subseteq V$ is defined as[3]

$$\mathsf{M}(C) = \frac{1}{2m}\left(\sum_{u,v \in C}\left(a_{u,v} - \frac{d_u d_v}{2m}\right)\right) \tag{1}$$

Then, a partition $\mathcal{S} = \{C_1, C_2, \ldots, C_k\}$ of V has a *total modularity* of

$$\mathsf{M}(\mathcal{S}) = \sum_{C_i \in \mathcal{S}} \mathsf{M}(C_i) \tag{2}$$

Notice that each distinct pair of nodes u and v contribute *twice* to the inside term $a_{u,v} - \frac{d_u d_v}{2m}$ in Equation (1). The goal is to find a partition (modular

[3]The $1/(2m)$ factor is for *normalization purposes only* to make the *optimal* objective value to lie between 0 and 1.

clustering) $\mathcal{S}$ (with unspecified k) to *maximize* $\mathsf{M}(\mathcal{S})$. Note that by allowing u and v to be equal in the inside summation, we provide a *negative weight to every node*.

Let $\mathsf{OPT} = \max_{\mathcal{S}} \mathsf{M}(\mathcal{S})$ denote the *optimal* modularity value. It is easy to verify that $0 \leq \mathsf{OPT} < 1$.

1.2. Brief History of Modularity Clustering and Its Applications

The modularity clustering approach is extremely popular both in the context of biological networks [20, 32] as well as social networks [1, 25, 28, 30]. However, as observed in [1], not much was known about the computational complexity aspect modularity clustering beyond NP-completeness for dense graphs, though various heuristic methods have been proposed and empirically evaluated in publications such as [11, 15, 31] via methods such as finding minimum weighted cuts. For unweighted networks, it is known that $\mathsf{OPT} = 0$ if G is a clique, $\mathsf{OPT} = 1 - \frac{1}{k}$ if G is an union of k *disjoint* cliques each with n/k nodes, computing OPT is NP-complete for sufficiently dense graphs[4] and the above-mentioned NP-completeness result holds even if any solution is constrained to contain no more than two clusters [10].

1.3. Informal Summary of Our Results

Unless mentioned otherwise explicitly, all algorithmic results apply for edge-weighted graphs and all hardness results apply for unweighted graphs.

Hardness Results For *dense* graphs, namely for the complements of 3-regular graphs, Theorem 3.1 in Section 3.1 provides a $(1+\varepsilon)$-inapproximability of the modularity clustering problem irrespective of whether the number of clusters is pre-specified or the algorithm is allowed to select the best number of clusters[5]. The required approximation gap in our reduction is derived from the approximation gap of the maximum independent set problem for 3-regular graphs in [14]. The intuition behind our inapproximability result is that, for the type of dense graphs that is considered in our reduction, large-size cliques must be *properly* contained within the clusters. However, the gap preservation calculations need to be done extremely accurately to avoid

[4]The reduction roughly requires $d_v = \Omega(\sqrt{n})$ for every node v.

[5]The proof shows that ε is roughly 0.0006.

shrinking the inapproximability gap[6].

Lemma 2.1 in Section 2 shows, using probabilistic arguments, that small number of clusters well-approximate the optimal modularity value; in particular, partitioning into *just two* clusters already achieves at least *half* of the optimum. Thus, it behooves to look at the complexity of the problem when we have at most two clusters, which we refer to as the 2-*clustering problem*. Theorem 4.1 in Section 4 proves the NP-completeness of the 2-clustering problem for *sparse* graphs, namely for d-regular graphs with any fixed $d \geq 9$; the previous NP-completeness result for this case in [10] required the degree of every node to be large (roughly $\Omega(\sqrt{n})$). Notice that we cannot anymore use the idea of hiding a large-size clique since the graph does not have any cliques of size more than d and, for fixed d, one can indeed enumerate all these cliques in polynomial time. Instead, our reduction is from the *graph bisection* problem for 4-regular graphs. Intuitively, now an optimal solution for 2-clustering is constrained to have exactly the same number of nodes in each community to avoid any local improvement. The ideas in the reduction are motivated by the proof for this case in [10], but we have to do a more careful reduction and analysis to preserve both the low-degree and the regularity of the resulting graph.

Approximation Algorithms We first consider the case of sparse graphs. We show in Section 4.2 that a natural linear programming relaxation of modularity clustering has a large integrality gap, thereby ruling out this avenue for non-trivial approximations[7]. Theorem 4.5 in Section 4.3 provides a $O(\log d)$-approximation for most (unweighted) d-regular graph (*i.e.*, with $d \leq \frac{n}{2 \ln n}$), and an approximation that is logarithmic in the *maximum weighted degree* for weighted graphs provided maximum weighted degree[8] is no more than *about* $\sqrt[5]{n}$. It is easy to see that the modularity function is *neither monotone nor sub-modular*, thus we instead need to use semi-definite programming (SDP) techniques for *maximizing quadratic forms*. However, we face several technical hurdles in using SDP-based approximation algorithms

[6]For example, the inapproximability gap of Berman and Karpinski in [9] does not suffice for our purposes.

[7]Interestingly, the proof shows that d-regular expander graphs have small modularity values ($\approx 1/\sqrt{d}$).

[8]As noted in Section 4.3, we normalize all the weights such that their sum is *exactly* twice the number of edges.

for quadratic forms in [5, 6, 13]: the coefficient matrix has *negative diagonal entries* and the lower bounds (hence the approximation ratios) in [5, 6, 13] depend on the number of nodes and not on the degree. Thus, our proof proceeds in two steps. In the first step we obtain a lower bound on the optimal modularity value as a *function of the degree* or *the maximum weighted degree* using an *explicit* graph decomposition. In the second step, we show that the SDP-based method for quadratic forms can be used to obtain an approximation that is within a logarithmic factor of this lower bound in spite of the negative diagonal entries.

For *locally-dense* weighted graphs (*i.e.*, graphs in which every node has a weighted degree of $\Omega(n)$) we observe in Section 3.2 that one can get a solution within any *constant additive error* in polynomial time by a simple use of the *regularity lemma*. In view of our APX-hardness result for dense graphs described before, this is perhaps the best polynomial-time approximation one could hope for.

Directed weighted Graphs In Section 5.1 we show that all the hardness and approximation results for undirected weighted graphs can be extended to similar results for *directed* weighted graphs.

Alternative Objectives and Null Models There are two natural objections to Newman's modularity clustering: *approximate solutions provably tend to produce many trivial (single-node) clusters* and *the background null model could be different*[9]. Motivated by these observations, we consider two variations of the original modularity measure, one in which the modularity of the network is the *minimum* (instead of sum) of the modularities of individual clusters and the other in which the null model is the classical Erdös-Rényi random graph. Our results show that the minimum objective provides similar optimal modularity values as the original sum objective without allowing small clusters, and the Erdös-Rényi random graph null model is equivalent to Newman's modularity clustering in an *appropriately defined* regular graph.

[9]The idea of using alternative null models has been explored before by some researchers [19, 23]; in particular, Karrer and Newman [23] showed that the scale-free null model provided by linear preferential attachment do not provide a new null model. However, the focus in all these results was mainly to *empirically* compare null models using simple algorithms based on greedy approaches without provable approximation guarantees.

1.4. Comments on Our Results

Relationships to previous approximation algorithms for quadratic forms The special case of partitioning the nodes into *two clusters only* can be written down as maximizing a quadratic form. However, none of the existing approximability results for quadratic forms apply directly to our case. In particular, the $O(\log n)$-approximation in [5, 13] is not applicable since the diagonal entries of the resulting constraint matrix are negative[10], results such as in [21] do not apply since the constraint matrix is *not* necessarily a positive semi-definite matrix and the $O(1)$-approximations of [6] via Grothendieck's inequality do not apply since the quadratic form does *not* induce a bipartition of variables.

Possibility of logarithmic approximation without degree constraints Our logarithmic approximations require some bound on the maximum degree of the given graph. A natural question is of course if such degree bounds can be removed. Two observations regarding this are relevant:

✶ A technical difficulty that arises for this purpose is from the fact that the modularity value can be precisely 0 (such as when the given graph is K_n, $K_{n,n}$ or a graph obtained from K_n by removing polylog(n) edges) or arbitrarily close to 0 (such as when the given graph is the complement of small degree graph). Thus, at the very least, a non-trivial approximation without such degree bounds would require an efficient polynomial-time computable characterization of the topology of graphs whose modularity values can be arbitrarily small together with a special algorithmic approach to handle these graphs; approaches using quadratic forms or the regularity lemma do not suffice in this respect.

✶ The negative weights of the nodes start playing a more crucial role in the value of modularity when it is close to 0. As observed by other researchers before, negative diagonal entries in the coefficient matrix of the objective that shifts the objective value close to 0 are sometimes difficult for approximate.

Relationships to other clustering or partitioning methods Modularity clustering can be defined by several equivalent equations, which may *seem*

[10]The negative diagonal entries *are crucial* in the modularity measure [1, 26]. Moreover, they could be small or large depending on the graph, thus it is not possible to specify a priori bound on them.

to suggest at a first glance that combinatorially the problem may be either similar to (via Equations (1) and (2)) some form of *correlation clustering*, or (via Equation (5)) similar to *graph bisection* (for two clusters), or similar to *minimum ℓ-way cut/clique-partition* type of problem (for arbitrary number of clusters, depending on whether the graph is unweighted or weighted), or similar to (via Lemma 2.2) some type of *dense subgraph* problem. However, our results show both similarities and differences between modularity clustering and these problems. For example, our hardness result for dense graphs should be contrasted with other partitioning problems of similar nature, such as MAX-CUT, graph bisection, graph separation, minimum ℓ-way cut and some versions of correlation clustering, for which one can design a PTAS (*e.g.*, see [7, 8, 18]).

2. Basic Results on Partitioning into Fewer Clusters

In this section we show bounds on OPT as well as some useful properties of the solution if we restrict the number of clusters to some pre-specified value k; we will refer to this as the *k-clustering* problem. The objective function $\mathsf{M}(\mathcal{S})$ can be equivalently represented (via algebraic manipulation as observed in [10, 25, 28, 30]) as follows. Let m_i denote the number of edges whose both endpoints are in the cluster C_i, m_{ij} denote the number of edges one of whose endpoints is in C_i and the other in C_j and $D_i = \sum_{v \in C_i} d_v$ denote the sum of degrees of nodes in cluster C_i. Then,

$$\mathsf{M}(\mathcal{S}) = \sum_{C_i \in \mathcal{S}} \left(\frac{m_i}{m} - \left(\frac{D_i}{2m} \right)^2 \right) \tag{3}$$

Since $\sum_{v \in V} \left(a_{u,v} - \frac{d_u d_v}{2m} \right) = 0$ for any $u \in V$, we can alternatively express $\mathsf{M}(C)$ as

$$\mathsf{M}(C) = \frac{1}{2m} \left(\sum_{u \in C,\, v \notin C} \left(\frac{d_u d_v}{2m} - a_{u,v} \right) \right) \tag{4}$$

This, along with Equation (3), this gives us the following third equation of modularity (note that now each pair of clusters contributes to the sum in Equation (5) *exactly once*):

$$\mathsf{M}(\mathcal{S}) = \sum_{C_i, C_j \,:\, i<j} \left(\frac{D_i D_j}{2m^2} - \frac{m_{ij}}{m} \right) \tag{5}$$

Let OPT_k denote the modularity value of an optimal clustering when one is allowed at most k clusters.

The following two lemmas make use of the alternative formulations described above. The first lemma asserts, via a probabilistic argument, that the optimal value does not go down by too much in our restricted setting.

Lemma 2.1. *For any $k \geq 1$, $\left(1 - \frac{1}{k}\right) \mathsf{OPT} \leq \mathsf{OPT}_k \leq 1 - \frac{1}{k}$.*

Proof. The inequality $\mathsf{OPT}_k \leq 1 - \frac{1}{k}$ can be proved as follows. For any clustering $\mathcal{S}$ with at most k clusters, Equation (3) gives $\mathsf{M}(\mathcal{S}) = \sum_{i=1}^{k} \frac{m_i}{m} - \sum_{i=1}^{k} \left(\frac{D_i}{2m}\right)^2$. The first sum in this equation is upper-bounded by 1. Using Cauchy-Schwarz inequality, we get $k \sum_{i=1}^{k} D_i^2 \geq \left(\sum_{i=1}^{k} D_i\right)^2$, giving a lower-bound of $1/k$ for the second sum.

The inequality $\left(1 - \frac{1}{k}\right) \mathsf{OPT} \leq \mathsf{OPT}_k$ can be proved as follows. For $k = 1$, the statement is trivially true. Now consider $k > 1$. We will make use of Equation (5) for modularity values. Suppose that our optimal clustering $\mathcal{S}$ has more than k clusters. Denote each term in the summation of Equation (5) by M_{ij}, *i.e.*, $\mathsf{M}_{ij} = \frac{D_i D_j}{2m^2} - \frac{m_{ij}}{m}$; thus $\mathsf{OPT} = \mathsf{M}(\mathcal{S}) = \sum_{i<j} \mathsf{M}_{ij}$. We can randomly assign each of the clusters to one of k superclusters. Let I_{ij} be the indicator random variable of the event C_i and C_j are in different clusters and let $\mathcal{S}_k$ denote the random k-clustering. It is easy to see that any pair C_i and C_j will contribute M_{ij} to the final clustering if and only if they are not in the same supercluster. Therefore, $\mathsf{M}(\mathcal{S}_k) = \sum_{i<j} I_{ij} \mathsf{M}_{ij}$. Thus we get $\mathsf{OPT}_k \geq \mathbb{E}[\mathsf{M}(\mathcal{S}_k)] = \sum_{i<j} \mathbb{E}[I_{ij}] \mathsf{M}_{ij} = \sum_{i<j} \left(1 - \frac{1}{k}\right) \mathsf{M}_{ij} = \left(1 - \frac{1}{k}\right) \mathsf{OPT}$. □

The next lemma shows that the 2-clustering problem can also be alternatively viewed as a special kind of "subgraph selection" problem.

Lemma 2.2. *Let V_1 and V_2 be any partition of V. Then, $\mathsf{M}(V_1) = \mathsf{M}(V_2)$.*

Proof. Remember that, for any node u, $\sum_{v \in V} \left(a_{u,v} - \frac{d_u d_v}{2m}\right) = 0$. Thus,

$$0 = \sum_{u \in V_1} \sum_{v \in V} \left(a_{u,v} - \frac{d_u d_v}{2m}\right) = \mathsf{M}(V_1) + \sum_{u \in V_1} \sum_{v \in V_2} \left(a_{u,v} - \frac{d_u d_v}{2m}\right)$$

$$0 = \sum_{u \in V_2} \sum_{v \in V} \left(a_{u,v} - \frac{d_u d_v}{2m}\right) = \mathsf{M}(V_2) + \sum_{u \in V_2} \sum_{v \in V_1} \left(a_{u,v} - \frac{d_u d_v}{2m}\right)$$

and therefore $\mathsf{M}(V_1) = \mathsf{M}(V_2)$. □

3. Results for Dense Graphs

3.1. APX*-hardness*

This hardness result may be contrasted with the results in Section 3.2 where we show that the modularity value can be approximated to within any *constant additive error* for dense graphs using the regularity lemma. However, the APX-hard instances here have modularity values that are very close to 0 (around $1/n$), thus the constant additive error provides no guarantee on the approximation ratio.

Theorem 3.1. *It is* NP*-hard to approximate the k-clustering problem, for any k, on $(n-4)$-regular graphs within a factor of $1+\varepsilon$ for some constant $\varepsilon > 0$.*

Proof. We reduce the maximum-cardinality independent set problem for 3-regular graphs (3-MIS) to our problem. An instance of 3-MIS consists of a 3-regular graph $H = (V, E)$, and the goal is to find a *maximum cardinality* subset of nodes $V' \subset V$ such that every pair of nodes u and v in V' is *independent*, *i.e.*, $\{u, v\} \notin E$. For notational convenience, let $\delta_\ell = 94/194$ and $\delta_\mathrm{h} = 95/194$. The following inapproximability result is known for 3-MIS.

Theorem 3.2.[14] *For any language L in* NP*, there exists a polynomial-time reduction such that given an instance I of L produces an instance of H of* 3-MIS *with n nodes such that:*

- *if $I \in L$ then H has a maximum independent set of cardinality at least $\delta_\mathrm{h} n$;*
- *if $I \notin L$ then every maximum independent set of H is of cardinality at most $\delta_\ell\, n$.*

We start with an instance I of L and translate it to an instance H of 3-MIS as described in Theorem 3.2; we refer to such an instance of 3-MIS as a "hard" instance. Given a hard instance $H = (V, F)$ of 3-MIS with $|V| = n$ nodes and $|F| = \frac{3n}{2}$ edges such that a maximum independent set is of size either at most $\delta_\ell\, n$ or at least $\delta_\mathrm{h} n$, consider the complement $\overline{H} = (V, \overline{F})$ of H, *i.e.*, the graph with $\overline{F} = \{\, \{u, v\} \mid u, v \in V,\, u \neq v\} \setminus F$. Since H is 3-regular, $\overline{H}$ is $(n-4)$-regular. The input to our 2-clustering problem is this graph $\overline{H}$. For notational uniformity, we will denote the graph $\overline{H}$ by $G = (V, E)$ with $E = \overline{F}$. Note that $V' \subset V$ is an independent set of H if and only if V' is

a clique in G. Let $\mathbf{\Psi}$ and OPT denote the size of a maximum independent set of H and the optimal modularity value of G, respectively. We prove our claim by showing the following:

$$\textbf{(completeness)}\ \text{If } \mathbf{\Psi} \geq \delta_{\mathrm{h}} n \text{ then } \mathsf{OPT} \geq \frac{2(4\delta_{\mathrm{h}}^2 - \delta_{\mathrm{h}})}{(n-4)} > \frac{0.9388}{n-4}.$$

$$\textbf{(soundness)}\ \text{If } \mathbf{\Psi} \leq \delta_{\ell}\, n \text{ then } \mathsf{OPT} \leq \frac{4\delta_{\ell} - 1}{n-4} < \frac{0.9382}{n-4}.$$

For any subset $\emptyset \subset V' \subset V$ of nodes in G, let $m_{V'}$ be the number of edges in G with both end-points in V' and $D_{V'}$ be the sum of degrees of nodes in V' in the graph G, *i.e.*, $D_{V'} = \sum_{v \in V'} d_v$.

3.1.1. Proof of Completeness ($\mathbf{\Psi} \geq \delta_{\mathrm{h}} n$)

Lemma 3.3. *If* $\mathbf{\Psi} \geq \delta_{\mathrm{h}} n$ *then* $\mathsf{OPT} \geq \dfrac{2(4\delta_{\mathrm{h}}^2 - \delta_{\mathrm{h}})}{(n-4)}$.

Proof. Suppose H has a has an independent set V' with $|V'| = t\,n$ for some $t \geq \delta_{\mathrm{h}}$. Since V' is a clique of G, it follows that $2m_{V'} = tn(tn-1)$ and $D_{V'} = tn(n-4)$. Consider the solution $\mathcal{S} = \{V', V \setminus V'\}$ of 2-clustering on G. Using Lemma 2.2 and Equation (3) we get

$$\mathsf{M}(\mathcal{S}) = 2\,\mathsf{M}(V') = 2\left(\frac{m_{V'}}{m} - \left(\frac{D_{V'}}{2m}\right)^2\right)$$

$$= \frac{2\,t\,n\,(t\,n-1)}{n(n-4)} - 2\,t^2 = \frac{2(4\,t^2 - t)}{n-4} \geq \frac{2(4{\delta_{\mathrm{h}}}^2 - \delta_{\mathrm{h}})}{n-4} \qquad \square$$

3.1.2. Proof of Soundness ($\mathbf{\Psi} \leq \delta_{\ell}\, n$)

Case I: when an optimal solution has exactly 2 clusters.

Suppose that the optimal solution is $\mathcal{S} = \{V', V \setminus V'\}$ of 2-clustering on G with $|V'| = t\,n$ and $0 < t \leq 1/2$.

Lemma 3.4. *Let* $\alpha\, n$ *be the size (number of nodes) of a largest size clique in the node-induced subgraph* $G' = (V', E')$ *where* $E' = (V' \times V') \cap E$. *Then,* $\mathsf{M}(V') \leq \dfrac{4t^2 + 2\alpha - 3t}{n-4}$.

Proof. Since the size of the largest clique in G' is $\alpha\, n$, for each of the remaining $(t-\alpha)\, n$ nodes, they will not be connected to at least one node inside the clique. Hence, using Equation (3), we get

$$\mathsf{M}(V') = \frac{m_{V'}}{m} - \left(\frac{D_{V'}}{2m}\right)^2 \leq \frac{\frac{t\, n(t\, n-1)}{2} - (t-\alpha)\, n}{\frac{n(n-4)}{2}} - t^2 = \frac{4\, t^2 + 2\, \alpha - 3\, t}{n-4} \quad \square$$

Lemma 3.5. $\mathsf{M}(V') \leq \dfrac{2\delta_\ell - \frac{1}{2}}{n-4}$.

Proof. Using the previous lemma and the facts that $\alpha \leq \min\left\{t, \delta_\ell\right\}$ and $t \leq 1/2$, we have two cases:

Case 1: $t > \delta_\ell$. Then $\mathsf{M}(V') \leq \frac{4t^2+2\alpha-3t}{n-4}$. The function $f(t) = 4t^2 - 3t$ is increasing in the range $(\delta_\ell, 1/2]$ since $\delta_\ell > 3/8$ and $\frac{\partial f}{\partial t} = 8t - 3 > 0$ if $t > 3/8$. Thus, $\max_{\delta_\ell < t \leq 1/2} f(t) = f\,(1/2) = -1/2$, and thus $\mathsf{M}(V') \leq \frac{2\alpha-\frac{1}{2}}{n-4} \leq \frac{2\delta_\ell-\frac{1}{2}}{n-4}$.

Case 2: $t \leq \delta_\ell$. Since $\alpha \leq t$ and $4t^2 + 2\alpha - 3t$ is an increasing function of α, we have $\mathsf{M}(V') \leq \frac{4t^2+2t-3t}{n-4} = \frac{4t^2-t}{n-4}$. The function $f(t) = 4t^2 - t$ satisfies $f(0) = 0$ and

$$\frac{\partial f}{\partial t} = 8t - 1 \begin{cases} < 0 & \text{if } t < 1/8 \\ > 0 & \text{if } 1/8 < t \leq \delta_\ell \end{cases}$$

Thus, $\max_{0 < t \leq \delta_\ell} f(t) = f(\delta_\ell)$ and we have $\mathsf{M}(V') \leq \dfrac{4{\delta_\ell}^2 - \delta_\ell}{n-4} \leq \dfrac{2\delta_\ell - \frac{1}{2}}{n-4}$. $\square$

Finally, using Lemma 2.2, $\mathsf{M}(\mathcal{S}) = 2\, \mathsf{M}(V') \leq \dfrac{4\delta_\ell - 1}{n-4}$, completing the soundness proof for this case.

Case II: when an optimal solution has more than 2 clusters.

For convenience of calculations, we would like to drop the $\frac{1}{2m}$ scaling term from Equation (1). To this end, we define $\mathsf{M}^{\text{uns}}(C) = n(n-4)\, \mathsf{M}(C)$. Let $\mathcal{S} = \left\{V_1, V_2, \ldots, V_{m+1}\right\}$ be an optimal solution of modularity clustering that uses a *minimum* $m > 1$ number of clusters. Let $|V_i| = t_i\, n$, and suppose that $\emptyset \subset V_i' \subseteq V_i$ is a largest clique of size $\alpha_i\, n$ in the graph $(V_i, (V_i \times V_i) \cap E)$. Note that $0 < \alpha_i \leq \min\left\{\, t_i, \delta_\ell\right\}$ for all $1 \leq i \leq m+1$, $\sum_{i=1}^{m+1} t_i = 1$ and we need to show that $\mathsf{M}^{\text{uns}}(\mathcal{S}) \leq \left(4\delta_\ell - 1\right) n$. Let $\widehat{V_i}$ denote $V \setminus V_i$.

Lemma 3.6. $\mathsf{M}^{\text{uns}}(V_i) \leq (4t_i^2 - t_i)n$.

Proof. $\mathsf{M}^{\text{uns}}(V_i)$ is maximized when the nodes in V_i form a clique. Thus,

$$\mathsf{M}^{\text{uns}}(V_i) \leq \left(\frac{4}{n} - 1\right)(t_i n) + \left(\frac{4}{n}\right)(t_i n - 1)(t_i n) = \left(4t_i^2 - t_i\right) n \qquad \square$$

Corollary 3.7. *If* $|V_i| \leq n/4$ *then* $\mathsf{M}^{\text{uns}}(V_i) \leq 0$. *If* $|V_i| = \left(\frac{1}{4} + \delta\right) n > n/4$ *then* $\mathsf{M}^{\text{uns}}(V_i) \leq (4\delta^2 + \delta)\, n$.

Lemma 3.8. *Suppose that* $t_i = \frac{1}{2} + \delta > \frac{1}{2}$ *for some* $0 < \delta < 1/2$ *and* $\widehat{\alpha}_i$ *is the size of a largest clique in* $(\widehat{V}_i, (\widehat{V}_i \times \widehat{V}_i) \cap E)$. *Then,*

$$\mathsf{M}^{\text{uns}}(V_i) \leq \left(4\delta^2 - \delta - \frac{1}{2} + 2\widehat{\alpha}_i\right) n \leq \left(2\delta_\ell - \frac{1}{2}\right) n$$

Proof. Note that $\left|\widehat{V}_i\right| = \frac{1}{2} - \delta < 1/2$. Then by Lemma 2.2,

$$\begin{aligned}\mathsf{M}^{\text{uns}}(V_i) = \mathsf{M}^{\text{uns}}(\widehat{V}_i) &\leq \left(4\left(\frac{1}{2} - \delta\right)^2 + 2\widehat{\alpha}_i - 3\left(\frac{1}{2} - \delta\right)\right) n \\ &= \left(4\delta^2 - \delta - \frac{1}{2} + 2\widehat{\alpha}_i\right) n\end{aligned}$$

where the inequality follows from Lemma 3.4 if we replace V_i by $\widehat{V}_i$. Since $t_i \geq 1/2$, we have

$$4\delta^2 - \delta - \frac{1}{2} + 2\widehat{\alpha}_i = 4t_i^2 - 5t_i + 1 - 2\widehat{\alpha}_i \leq 4t_i^2 + 2\widehat{\alpha}_i - 3t_i$$

Since $\widehat{\alpha}_i \leq \delta_\ell < t_i$, the arguments in Lemma 3.5 can be directly applied on $4t_i^2 + 2\widehat{\alpha}_i - 3t_i$ to show that $\left(4\delta^2 - \delta - \frac{1}{2} + 2\widehat{\alpha}_i\right) n \leq \left(2\delta_\ell - \frac{1}{2}\right) n$. $\square$

Let us call a cluster V_i a *giant component* if $t_i > \delta_\ell$. Note that since $3\,\delta_\ell > 1$, we can have *at most two* giant components. We have therefore three cases depending on the number of giant components.

Case (i): $\mathcal{S}$ has no giant components Note that $\mathcal{S}$ can have at most three clusters containing *strictly* more than $n/4$ nodes.

If $\mathcal{S}$ contains no such cluster then by Corollary 3.7 $\mathsf{M}^{\text{uns}}(\mathcal{S}) \leq 0$.

If $\mathcal{S}$ contains exactly one such cluster, say V_1, then $\mathsf{M}^{\text{uns}}(\mathcal{S}) \leq \mathsf{M}^{\text{uns}}(V_1) \leq \left(2\delta_\ell - \frac{1}{2}\right)\, n < (4\delta_\ell - 1)\, n$ by Lemma 3.5 (if $t_i \leq 1/2$) or Lemma 3.8 (if $t_i > 1/2$).

If $\mathcal{S}$ contains exactly two such clusters, say V_1 and V_2, then again $\mathsf{M}^{\text{uns}}(\mathcal{S}) \leq \mathsf{M}^{\text{uns}}(V_1) + \mathsf{M}^{\text{uns}}(V_2) \leq 2\left(2\delta_\ell - \frac{1}{2}\right) n = \left(4\delta_\ell - 1\right) n$ by Lemma 3.5 and Lemma 3.8.

Otherwise, suppose that $\mathcal{S}$ contains *exactly three* such clusters, say V_1, V_2 and V_3. Let $t_i = \frac{1}{4} + \delta_i$ for $i = 1, 2, 3$. Then, $0 < \delta_1 + \delta_2 + \delta_3 < 1/4$. Using Corollary 3.7 we have:

$$\sum_{i=1}^{3} \mathsf{M}^{\text{uns}}(V_i) \leq \left(4\sum_{i=1}^{3}\delta_i^2 + \sum_{i=1}^{3}\delta_i\right) n < \left(4\left(\sum_{i=1}^{3}\delta_i\right)^2 + \frac{1}{4}\right) n$$
$$< \left(4\left(\frac{1}{4}\right)^2 + \frac{1}{4}\right) n = \frac{n}{2} < \left(4\delta_\ell - 1\right) n$$

Case (ii): $\mathcal{S}$ has one giant component Let V_1 be the giant component. Since $1 - t_1 < 1 - \delta_\ell < 3/4$, there are at most two other clusters with strictly more than $n/4$ nodes.

Subcase (ii-a): there is one other cluster with strictly more than $n/4$ nodes Let this cluster be V_2. By Corollary 3.7, $\sum_{j=3}^{m+1} \mathsf{M}^{\text{uns}}(V_j) \leq 0$. Note that $t_2 \leq \delta_\ell$. Now, by reusing the calculations of Lemma 3.5 and using Lemma 3.8 we get

$$\mathsf{M}^{\text{uns}}(\mathcal{S}) = \mathsf{M}^{\text{uns}}(V_1) + \mathsf{M}^{\text{uns}}(V_2) + \sum_{j=3}^{m+1} \mathsf{M}^{\text{uns}}(V_j) \leq \mathsf{M}^{\text{uns}}(V_1) + \mathsf{M}^{\text{uns}}(V_2)$$
$$\leq \underbrace{\left(2\delta_\ell - \frac{1}{2}\right) n}_{\substack{\text{by Lemma 3.8 if } t_1 > 1/2 \\ \text{by Lemma 3.5 if } t_1 \leq 1/2}} + \underbrace{\left(2\delta_\ell - \frac{1}{2}\right) n}_{\text{by Lemma 3.5 since } t_2 \leq \delta_\ell} = \left(4\delta_\ell - 1\right) n$$

Subcase (ii-b): there are two other clusters with strictly more than $n/4$ nodes Let these clusters be V_2 and V_3. Then, $\delta_\ell\, n < |V_1| < n/2$. By Corollary 3.7, $\sum_{j=4}^{m+1} \mathsf{M}^{\text{uns}}(V_j) \leq 0$. Let $t_2 = \frac{1}{4} + \delta_2$ and $t_3 = \frac{1}{4} + \delta_3$ with $0 < \delta_2 \leq \delta_3 < \frac{1}{2} - \delta_\ell < 2/100$. Thus,

$$\mathsf{M}^{\text{uns}}(\mathcal{S}) \leq \mathsf{M}^{\text{uns}}(V_1) + \mathsf{M}^{\text{uns}}(V_2) + \mathsf{M}^{\text{uns}}(V_3)$$
$$\leq \underbrace{\left(2\delta_\ell - \frac{1}{2}\right) n}_{\text{by Lemma 3.5 since } t_1 < 1/2} + \underbrace{\left(4\delta_2^2 + \delta_2\right) n}_{\text{by Corollary 3.7}} + \underbrace{\left(4\delta_3^2 + \delta_3\right) n}_{\text{by Corollary 3.7}}$$

Since $4\delta_2^2 + \delta_2 + 4\delta_3^2 + \delta_3 < 8(2/100)^2 + 2\,(2/100) < 2\delta_\ell - \frac{1}{2}$, we have $\mathsf{M}^{\text{uns}}(\mathcal{S}) \leq (4\delta_\ell - 1)\,n$.

Case (iii): $\mathcal{S}$ has two giant components Let V_1 and V_2 be the two giant components with $t_1 = \delta_\ell + \mu_1$ and $t_2 = \delta_\ell + \mu_2$ for some $0 < \mu_1 \leq \mu_2 < 1 - 2\delta_\ell$. Since $\left|\cup_{j=3}^{m+1} V_i\right| = (1 - t_1 - t_2)\,n \leq (1 - 2\delta_\ell)\,n < n/4$, by Corollary 3.7 $\sum_{j=3}^{m+1} \mathsf{M}^{\text{uns}}(V_j) \leq 0$. Now, by reusing the calculations in the proof of the case of $t > \delta_\ell$ of Lemma 3.5 and using Lemma 3.8 we get

$$\mathsf{M}^{\text{uns}}(\mathcal{S}) = \mathsf{M}^{\text{uns}}(V_1) + \mathsf{M}^{\text{uns}}(V_2) + \sum_{j=3}^{m+1} \mathsf{M}^{\text{uns}}(V_j)$$

$$\leq \underbrace{\left(2\delta_\ell - \frac{1}{2}\right) n}_{\substack{\text{by Lemma 3.8 if } t_1 > 1/2 \\ \text{by Lemma 3.5 if } t_1 \leq 1/2}} + \underbrace{\left(2\delta_\ell - \frac{1}{2}\right) n}_{\substack{\text{by Lemma 3.8 if } t_2 > 1/2 \\ \text{by Lemma 3.5 if } t_2 \leq 1/2}} = (4\delta_\ell - 1)\,n \qquad \square$$

3.2. Additive Approximations for Locally Dense Graphs

Using the algorithmic version of the regularity lemma in [18] we can show that if the given graph is dense then, for any given constant $\alpha > 0$, there is a polynomial-time algorithm that returns a solution of modularity value at least $\mathsf{OPT} - \alpha$.

Proposition 3.9 (constant addiditive error)**.** *Suppose that the given graph $G = (V, E)$ is dense,* i.e.*, $m = |E| = \delta n^2$ for some constant $0 < \delta < 1/2$. Then, for any given constant $0 < \alpha < 1$, there is a polynomial-time algorithm that returns a solution of value at least $\mathsf{OPT} - \alpha$.*

Proof. The ℓ-way cut problem is defined as follows. We are given an weighted graph $G = (V, E)$ with $w(u, v) \in \mathbb{R}$ being the weight of the edge $\{u, v\} \in E$. A valid solution is a partition of V to ℓ subsets $\mathcal{S} = \{S_1, S_2, \ldots, S_\ell\}$, and the goal is to *maximize* the sum of weights of those edges whose end-points are in different subsets, *i.e.*, maximize $w(\mathcal{S}) = \sum_{\{u,v\} \in E(\mathcal{S})} w(u, v)$, where $E(\mathcal{S}) = \{\,\{u, v\} \,|\, \forall\, 1 \leq j \leq \ell\colon\ |\{u, v\} \cap S_j| \neq 2\}$ is the set of all "inter-partition" edges. The following result was proved in [18].

Theorem 3.10. [18] *Given an weighted graph $G = (V, E)$ of n nodes and any constant $0 < \varepsilon < 1$ there is a polynomial-time algorithm A_ε which, computes a partition $\mathcal{S}_\varepsilon$ of V such that*

$$w(\mathcal{S}_\varepsilon) \geq w(\mathcal{S}^*) - \varepsilon n^2$$

where $\mathcal{S}^*$ is an optimal (maximum weight) partition.

Equation (4) can be used to assign edge weights to cast our modularity clustering problem as an ℓ-way cut problem in the following manner. Consider the complete graph on n nodes (K_n) and let $w_{u,v} = 2\,\delta\left(\frac{d_u d_v}{2m} - a_{u,v}\right)$ for the edge $\{u, v\}$ of K_n. Then, for a partition $\mathcal{S} = \{S_1, S_2, \ldots, S_\ell\}$ of the nodes of K_n,

$$w(\mathcal{S}) = \sum_{\{u,v\} \in E(\mathcal{S})} 2\,\delta \left(\frac{d_u d_v}{2m} - a_{u,v} \right) = 2\,m\,\delta\,\mathsf{M}(\mathcal{S}) = 2\,\delta^2 n^2\,\mathsf{M}(\mathcal{S})$$

Let APX_ε be the objective value of an approximate solution of the modularity clustering problem on the given graph obtained by using the ℓ-way partitioning of Theorem 3.10 with $\varepsilon = 2\,\alpha\,\delta^2$. Then,

$$2\delta^2 n^2 \mathsf{APX}_\varepsilon \geq 2\delta^2 n^2 \mathsf{OPT} - \varepsilon n^2 \;\equiv\; \mathsf{APX}_\varepsilon \geq \mathsf{OPT} - \alpha \qquad \square$$

4. Hardness and Approximation Algorithms for Sparse Graphs

4.1. NP*-hardness*

Brandes *et al.* [10] proved NP-hardness of the 2-clustering problem provided nodes with very large degrees are allowed in the input graph. Thus it is not a priori clear whether calculating modularity on very sparse graphs becomes easy and admits an *exact* polynomial-time algorithm. However, we rule out this possibility of exact solution. Our construction is similar to that in [10], but carefully replaces dense graphs with *nicely behaving* sparse graphs. We have to do a more careful analysis of the properties of an optimal 2-clustering so as to get the following result.

Theorem 4.1. *Computing* OPT_2 *is* NP*-complete even for d-regular graphs for any* constant $d \geq 9$.

Proof. The decision version 2BdRegModularity of our problem is as follows:

> *given a d-regular graph G and a number K, is there a clustering $\mathcal{S}$ of G into at most two clusters for which $\mathsf{M}(\mathcal{S}) \geq K$?*

Our reduction is from the minimum graph bisection problem for 4-regular graphs (MB4): *Given a 4-regular graph G with n nodes (with even n) and an integer c, is there a clustering into two clusters each of $n/2$ nodes such that it*

"cuts" at most c edges, i.e., *at most c edges have two end-points in different clusters*? MB4 is known to be NP-complete [24]. We reduce an instance G of MB4 to an instance of 2BdRegModularity in a manner similar to that in [10]. Every node in G is replaced by a copy of an n-node d-regular graph H such that the minimum cut (minimum number of edges in a cut) of H is at least d. Such a family of graphs can be constructed in the following recursive manner:

- For $d = 2$, the 2-regular graph, namely a simple cycle consisting of n nodes, has a minimum cut of 2 edges.
- For $d = 3$, consider two simple cycles $H_1 = (V_1, E_1)$ and $H_2 = (V_2, E_2)$, each consisting of $n/2$ nodes. Consider an arbitrary matching between the nodes of H_1 and H_2 and add the edges corresponding to this matching to obtain a 3-regular graph $H = (V, E)$. Consider an arbitrary subset of nodes $V' \subset V$ of H. Then,
  - If $V' \cap V_1 \neq \emptyset$ and $V' \cap V_2 \neq \emptyset$, then the number of cut edges is at least 4.
  - Otherwise, assume that $V' \cap V_1 = \emptyset$ (the other case is symmetric) and thus $\emptyset \subset V' \subseteq V_2$. If $V' = V_2$ then the number of cut edges is exactly $n/2 > 2$. Otherwise, the number of cut edges is at least 2 (corresponding to two edges of the cycle in H_2) plus 1 (corresponding to one of the matching edges added).
- For $d > 3$, a recursive construction of such graphs follows in a similar manner: take such a $(d-2)$-regular graph H on n nodes for which the inductive hypothesis applies and add a simple cycle to H all of whose edges are different from those in H. Consider a cut in this graph. By the induction hypothesis the cut contains at least $d-2$ edges of H and at least 2 additional edges of the new cycle added to H.

Let H_v denote the copy of H corresponding to the node $v \in G$. Delete two independent edges (*i.e.*, edges without any common end-points) in H_v. The four edges connected to v are now connected to the four endpoints of these deleted edges. This is done in order to make the final graph G' d-regular[11].

[11] This is one step that is different from the reduction in [10], where every node in G is replaced by a copy of K_n producing the final graph with non-constant degrees. Since G is 4-regular, we need $d > 8$.

Note that the number of nodes in the transformed graph G' is n^2, whereas the number of edges is $m = \frac{d\,n^2}{2}$. Since two edges are removed from H in the construction, the minimum cut in each modified copy of H is at least $d-2$. The correctness of the reduction follows by showing that MB4 has a solution with at most c cut edges if and only if $\mathsf{M}(\mathcal{S}^*) \geq \frac{1}{2} - \frac{c}{m}$.

Let $\mathcal{S}^*$ be an optimal clustering of G'.

Lemma 4.2. *$\mathcal{S}^*$ has exactly two clusters and $\mathsf{M}(\mathcal{S}^*) > 0$.*

Proof. It suffices to show a clustering $\mathcal{S} = \{C_1, C_2\}$ such that $\mathsf{M}(\mathcal{S}) > 0$. To this end, let $C_1 = \{H_v\}$ for some v, and let C_2 contain the rest. Then using Equation (5) and the fact that $d(n-1) > 4$, we get

$$\mathsf{M}(\mathcal{S}) = \frac{D_1(2m - D_1)}{2m^2} - \frac{4}{m} = \frac{dn(dn^2 - dn)}{\frac{d^2n^4}{2}} - \frac{4}{\frac{dn^2}{2}} = \frac{2d(n-1) - 8}{dn^2} > 0 \quad \square$$

The next lemma shows how to normalize a solution without decreasing the modularity value. Part **(a)** of the lemma states that $\mathcal{S}^*$ cannot have any copy of H split across clusters, whereas part **(b)** implies that any optimal clustering has to be a bisection of the graph.

Lemma 4.3. *It is possible to normalize an optimal solution $\mathcal{S}^*$ without decreasing the modularity value such that the following two conditions hold:*

(a) *For every $v \in G$, there exists a cluster $C \in \mathcal{S}^*$ such that $H_v \subseteq C$.*

(b) *Each cluster in $\mathcal{S}^*$ contains exactly $n/2$ copies of H.*

Proof. Suppose the set of nodes of G' is partitioned into three subsets A, B and C. Let $\mathcal{S}_1 = \{A \cup C, B\}$, and we want to transfer the nodes in C to the other cluster to form the clustering $\mathcal{S}_2 = \{A, B \cup C\}$. For any two disjoint subsets X and Y of nodes of G', let m_{XY} denote the number of edges one of whose endpoints is in X and the other in Y and $D_X = \sum_{v \in X} d_v$ denote the sum of degrees of nodes in X. Then, using Equation (3) or Equation (5), the gain in modularity $\Delta = \mathsf{M}(\mathcal{S}_2) - \mathsf{M}(\mathcal{S}_1)$ can be simplified and written as $\Delta = \dfrac{(D_A - D_B)D_C}{2m^2} + \dfrac{m_{BC} - m_{AC}}{m}$. Using the fact that G' is d-regular and substituting for m, we get

$$\frac{dn^4}{2}\Delta = d\,|C|\,\big(\,|A| - |B|\,\big) + n^2\,\big(m_{BC} - m_{AC}\big) \tag{6}$$

(a) Let us assume that there exists a $v \in G$ such that H_v is split across clusters in the optimal clustering $\mathcal{S}^* = \{C_1, C_2\}$. Without loss of generality, we can assume that $|C_1 \setminus H_v| \geq |C_2 \setminus H_v|$. We will transfer the part of H_v in C_1 from C_1 to C_2. Let $A = C_1 \setminus H_v$, $B = C_2$, $C = H_v \setminus C_2$, and $|C| = k$. Then the part of H_v in C_2 has a size of $n - k$. By our assumption,

$$|A| - |B| = |C_1 \setminus H_v| - |C_2| = |C_1 \setminus H_v| - |C_2 \setminus H_v| - |H_v \setminus C_2| \geq -(n-k)$$

Substituting this in Equation (6), we get

$$\frac{dn^4}{2}\Delta \geq d[-k(n-k)] + n^2(m_{BC} - m_{AC})$$

Now, since the original graph G was 4-regular, at most 4 extra inter-cluster edges will appear after the transfer. Thus, $m_{AC} \leq 4$. The term m_{BC} represents the number of edges between C_2 and $H_v \setminus C_1$, which is at least the number of edges between the two parts of H_v. Thus, m_{BC} is at least the number of edges in a minimum cut of H_v which is at least $d - 2$. This gives

$$\frac{dn^4}{2}\Delta \geq -dk(n-k) + n^2(d-2-4) \geq -d\frac{n^2}{4} + (d-6)n^2 = \frac{(3d-24)n^2}{4} > 0$$

where the second inequality is due to the fact that $k(n-k)$ is maximized when $k = n/2$, and the last inequality is satisfied when $d \geq 9$. Hence the modularity can be strictly improved by putting each copy of H completely in a cluster.

(b) By the previous part, each H_v is contained completely in one cluster of $\mathcal{S}^* = \{C_1, C_2\}$. Now assume that C_1 has more copies of H than C_2. Since n is even, this implies that C_1 has at least two more copies of H than C_2. We will create a new clustering by transferring a copy of H from C_1 to C_2. Then the gain in modularity after this transfer is given by Equation (6), where C denotes the transferred copy of H, $B = C_2$ and $A = C_1 \setminus C$. By our assumption, $|A| - |B| \geq |C|$. Therefore we can simplify the first term and get $\frac{dn^4}{2}\Delta \geq d|C|^2 + n^2(m_{BC} - m_{AC})$. Also, since the original graph G was 4-regular, at most 4 extra inter-cluster edges will appear after the transfer. Simplifying and substituting values, $\frac{dn^4}{2}\Delta \geq dn^2 - 4n^2 > 0$. Hence, the modularity can be strictly improved by balancing out the copies of H in both clusters. □

Armed with the above lemma, one can now prove the NP-completeness of our problem. We will use the above construction to reduce an instance

$\langle G, c\rangle$ of MB4 to an instance $\langle G', K\rangle$ of 2BdRegModularity with $K = \frac{1}{2} - \frac{c}{m}$. Now suppose $\mathcal{S}^* = \{C_1, C_2\}$ is an optimal 2-clustering of G'. Then, $\mathsf{M}(\mathcal{S}^*) = \frac{D_1 D_2}{2m^2} - \frac{m_{12}}{m}$. By Lemma 4.3**(b)**, $D_1 = D_2 = m$. Also, because of Lemma 4.3**(a)**, m_{12} only has edges from G, thus representing a bisection of G. Therefore, $m_{12} \leq c$ if and only if $\mathsf{M}(\mathcal{S}^*) \geq \frac{1}{2} - \frac{c}{m} = K$. □

4.2. Large Integrality Gap for an ILP *Formulation*

$$\begin{array}{l} \textit{maximize } \dfrac{\sum_{\{u,v\,:\,u\neq v\}} \left(a_{u,v} - \frac{d_u d_v}{2m}\right)(1 - x_{u,v})}{2m} - \displaystyle\sum_{v\in V} \frac{d_v^2}{2\,m} \\ \textit{subject to } \forall\, u \neq v \neq z\colon x_{u,z} \leq x_{u,v} + x_{v,z} \\ \qquad\qquad \forall\, u \neq v\colon 0 \leq x_{u,v} \leq 1 \end{array}$$

Figure 1: LP-relaxation of modularity clustering [1, 10, 12].

There is an integer linear programming (ILP) formulation of modularity clustering with arbitrarily many clusters as shown in Fig. 1: $x_{u,v} = 0$ if u and v belong to the *same* cluster and 1 otherwise, and the "triangle inequality" constraints $x_{u,z} \leq x_{u,v} + x_{v,z}$ ensure that if $\{u, v\}$ and $\{v, z\}$ belong to the same cluster then $\{u, z\}$ also belongs to the same cluster. Agarwal and Kempe [1] used such an LP-relaxation with several rounding schemes for empirical evaluations. However, as we show below, the worst case integrality gap of the LP-relaxation is at least about the square root of the degree of the graph, thereby ruling out logarithmic approximations via rounding such LP-relaxations.

Lemma 4.4. *For every $d > 3$ and for all sufficiently large n, there exists a d-regular graph with n nodes such that the integrality gap of the* LP*-relaxation in Fig. 1 is $\Omega(\sqrt{d}\,)$.*

Proof. Let OPT_f be the optimal objective value of the LP-relaxation. For *any* graph $G = (V, E)$, a valid fractional solution of the LP-relaxation is as follows: set $x_{u,v} = \frac{1}{2}$ for *every* $\{u, v\} \in E$ and set $x_{u,v} = 1$ otherwise. The value of this fractional solution is precisely $\frac{1}{2} - \sum_{v\in V} \frac{d_v^2}{2m}$. Thus, in particular, if G is a d-regular graph then $\mathsf{OPT}_f \geq \frac{1}{2} - \frac{1}{n}$.

On the other hand, suppose that G is a random d-regular graph and let λ be the second largest eigenvalue of the adjacency matrix A of G. It is

well-known that $\lambda < \beta\sqrt{d}$ for some positive constant β [17]. Consider an optimal solution $\emptyset \subset V' \subset V$ of 2-clustering of G with $0 < |V'| = \alpha n \leq n/2$ and let $\mathrm{cut}(V')$ denote the number of edges between V' and $V \setminus V'$. By the *expander mixing lemma*, we have

$$\left| \mathrm{cut}(V') - d\frac{(\alpha n) \times (1-\alpha)n}{n} \right| \leq \lambda\sqrt{(\alpha n)(1-\alpha)n}$$
$$\equiv \quad |\,\mathrm{cut}(V') - \alpha(1-\alpha)\,d\,n\,| \leq \lambda\sqrt{\alpha(1-\alpha)}\,n$$

which implies $\mathrm{cut}(V') \geq \alpha(1-\alpha)\,d\,n - \lambda\sqrt{\alpha(1-\alpha)}\,n > \alpha(1-\alpha)\,d\,n - \beta\sqrt{d}\,n$. Let $\mathrm{uncut}(V')$ denote the number of edges between pairs of nodes in V'. Then, $\mathrm{uncut}(V') = \frac{\alpha\,d\,n - \mathrm{cut}(V')}{2} < \frac{\alpha^2\,d\,n + \beta\sqrt{d}\,n}{2}$. Using this in Equation (3) (with $m = dn/2$) together with Lemma 2.1 and 2.2 shows

$$\mathsf{M}(V') = \frac{2 \times \mathrm{uncut}(V')}{d\,n} - \left(\frac{\alpha\,d\,n}{d\,n}\right)^2 < \frac{\beta}{\sqrt{d}}$$
$$\implies \mathsf{OPT} \leq 2\,\mathsf{OPT}_2 = 4\,\mathsf{M}(V') < \frac{4\,\beta}{\sqrt{d}} \implies \frac{\mathsf{OPT}_f}{\mathsf{OPT}} = \Omega(\sqrt{d}\,) \qquad \square$$

4.3. Logarithmic Approximation

Newman [27] extended the modularity measure to weighted graphs in the following manner. Let $G = (V, E, \ell)$ be the input weighted graph with $\ell : E \mapsto \mathbb{R}^+$ being the function mapping edges to *non-negative* real-valued weights. Now, if we redefine $d_u = \sum_{\{u,v\} \in E} \ell(u,v)$ as the "weighted" degree of the node u, $m = \sum_{u \in V} d_u$, and $A = [a_{u,v}]$ as the *weighted* adjacency matrix of G (*i.e.*, $a_{u,v} = \ell(u,v)$ if $\{u,v\} \in E$ and 0 otherwise), then Equation (1) applies to the weighted case also. The corresponding modification in Equation (3) can be obtained by redefining m_i as the *total weight* of edges whose both endpoints are in the cluster C_i, m_{ij} as the *total weight* of edges one of whose endpoints is in C_i and the other in C_j and $D_i = \sum_{v \in C_i} d_v$ as the sum of *weighted degrees* of nodes in cluster C_i. It is straightforward to see that Lemma 2.1 holds even for weighted graphs.

We denote the weighted degree, the maximum weighted degree and the average weighted degree of a node v by d_v, $d_{\max} = \max_{v \in V}\{d_v\}$ and $\Delta = \frac{\sum_{v \in V} d_v}{n}$, respectively, and, for convenience, *we normalize*[12] *all the weights such that $\sum_{v \in V} d_v$ is twice the number of edges of G.*

[12]It is easy to see that the modularity value of any clustering remains unchanged if all weights are scaled by the same factor.

Theorem 4.5.
(a) *There exists a polynomial time $O(\log d)$-approximation for d-regular graphs with $d < \frac{n}{2 \ln n}$.*
(b) *There exists a polynomial time $O(\log d_{\max})$-approximation for weighted graphs $d_{\max} < \frac{\sqrt[5]{n}}{16 \ln n}$.*

Proof. We begin with the approximation algorithm for regular graphs, which is somewhat easier to analyze, and later generalize the results for weighted graphs. A common theme for both the proofs is the following approach. By Lemma 2.1 $\mathsf{OPT}_2 \geq \mathsf{OPT}/2$, and thus it suffices to provide a logarithmic approximation for the 2-clustering problem on G. For notational convenience let $w_{u,v} = \dfrac{a_{u,v} - \frac{d_u d_v}{2m}}{2\,m}$. As observed in [29], letting $x_u \in \{-1, 1\}$ be the indicator variable denoting the partition that node $u \in V$ belongs to, Equation (2) can be rewritten for a 2-clustering as $\mathsf{M}(\mathcal{S}) = \sum_{u,v \in V} w_{u,v}\,(1 + x_u x_v) = \sum_{u,v \in V} w_{u,v} x_u x_v = \mathbf{x}^{\mathrm{T}} W \mathbf{x}$ where $\mathbf{x} \in \{-1, 1\}^n$ is a column vector of the indicator variables and $W = [w_{u,v}] \in \mathbb{R}^{n \times n}$ is the corresponding symmetric matrix. The following result is known on quadratic forms.

Theorem 4.6. [13] *Consider maximizing $\mathbf{x}^{\mathrm{T}} Z \mathbf{x}$ subject to $\mathbf{x} \in \{-1, 1\}^n$, where $Z = [z_{i,j}]$ is a $n \times n$ real matrix with $z_{i,i} \geq 0$. Then, for any $T > 1$, there exists a randomized approximation algorithm whose objective value κ satisfies $\mathbb{E}[\kappa] \geq \dfrac{\max_{\mathbf{x} \in \{-1,1\}^n} \mathbf{x}^{\mathrm{T}} Z \mathbf{x}}{T^2} - 8\,\mathrm{e}^{-T^2/2} \left(\sum_{i \neq j} |z_{i,j}| \right)$.*

The above approximation does not directly apply to the quadratic form for modularity clustering since the diagonal entries are negative for our case. Moreover, the lower bound on the optimal value of the quadratic form as used in [13] depends on n which we would like to avoid.

(a) The Case When the Input Graph is Regular.

The proof of the following lemma uses a result in [22] on the size of a maximum-cardinality matching of a regular graph. The above lemma is tight in the sense that there exist d-regular graphs for which $\mathsf{OPT} = O\left(1/\sqrt{d}\right)$ (the proof of Lemma 4.4 shows that d-regular expanders are one such class of graphs).

Lemma 4.7 (Lower Bound for OPT)**.** *If $n > 40d^9$ then $\mathsf{OPT} > \frac{0.26}{\sqrt{d}}$, else $\mathsf{OPT} > \frac{0.86}{d} - \frac{4}{n}$.*

Proof. Consider a maximum-cardinality matching $\{u_1, v_1\}, \ldots, \{u_k, v_k\}$ of G of size k. It is known [22] that for any $d > 2$,

$$k \geq \begin{cases} \min\left\{\dfrac{n(d^2+4)}{2d^2+2d+4}, \dfrac{n-1}{2}\right\}, & \text{if } d \text{ is odd} \\ \dfrac{(d^3-d^2-2)n-2d+2}{2(d^3-3d)}, & \text{otherwise} \end{cases}$$

which gives $k > 0.43\,n$ for any d. We create k clusters $\{V_1, V_2, \ldots, V_k\}$ where $V_i = \{u_i, v_i\}$ and for each remaining node $u \in V \setminus \left(\cup_{i=1}^{k} V_i\right)$ we create a cluster $\{u\}$ of one node. Using Equation (3), we have

$$M(\mathcal{S}) = \sum_{C_i} \left[\frac{m_i}{m} - \left(\frac{D_i}{2m}\right)^2\right] = \sum_{i=1}^{k}\left(\frac{2}{dn} - \frac{4}{n^2}\right) - \sum_{i=k+1}^{n}\frac{1}{n^2} > \frac{0.86}{d} - \frac{4}{n}$$

For fixed d and $n > 40d^9$, it was shown in [4] that every d-regular graph with n nodes has a bisection width of at most $\left(\frac{d}{2} - 0.13\sqrt{d}\right)\left(\frac{n}{2}\right)$. Consider the partition $\mathcal{S}$ of G into two clusters C_1 and C_2 corresponding to such a bisection with exactly $n/2$ nodes in each cluster. Then, $m = \frac{dn}{2}$, $D_1 = D_2 = m$, $m_1, m_2 > \left(\frac{d}{2} + 0.13 \times \sqrt{d}\right)\left(\frac{n}{4}\right)$ and using Equation (3) we get $\mathsf{M}(C_1) = \mathsf{M}(C_2) > \frac{0.13}{\sqrt{d}}$. Consequently, by Lemma 2.2 $\mathsf{M}(\mathcal{S}) > \frac{0.26}{\sqrt{d}}$. □

We now define the following quantities:

- $\mathsf{D} = \sum_{v \in V} |w_{v,v}|$.
- $W' = \left[w'_{u,v}\right]$ where $w'_{u,v} = \begin{cases} 0, & \text{if } u = v \\ w_{u,v}, & \text{otherwise} \end{cases}$.
- $\mathbf{W}'_{\text{total}} = \sum_{u,v \in V} |w'_{u,v}|$.

Thus, if $\mathsf{OPT}_2 = \max_{\mathbf{x} \in \{-1,1\}^n} \mathbf{x}^{\mathrm{T}} W \mathbf{x}$ and $\mathsf{OPT}'_2 = \max_{\mathbf{x} \in \{-1,1\}^n} \mathbf{x}^{\mathrm{T}} W' \mathbf{x}$ then $\mathsf{OPT}'_2 = \mathsf{OPT}_2 - \mathsf{D}$.

Lemma 4.8. $\mathbf{W}'_{\text{total}} < 2$.

Proof.

$$\mathbf{W}'_{\text{total}} < \sum_{u,v \in V} \left| w_{u,v} \right| = \sum_{w_{u,v} \geq 0} w_{u,v} - \sum_{w_{u,v} < 0} w_{u,v} = \underbrace{2 \left(\sum_{w_{u,v} \geq 0} w_{u,v} \right)}_{\text{since} \sum_{u,v \in V} w_{u,v} = \sum_{w_{u,v} \geq 0} w_{u,v} - \sum_{w_{u,v} < 0} w_{u,v} = 0} < \frac{\sum_{\{u,v\} \in E} a_{u,v}}{m} = 2$$

□

Next, we bound D by observing that, for any d, $\mathsf{D} = \frac{d^2 n}{4m^2} = \frac{1}{n}$. To complete the proof, we use the algorithm in Theorem 4.6 with $Z = W'$. Using Lemmas 2.1, 4.7 and 4.8 we get the desired approximation guarantees of Theorem 4.5 by choosing $T = \sqrt{4 \ln d}$ in the algorithm in Theorem 4.6. Then we have the following chain of implications for all sufficiently large d and n:

- $\mathsf{OPT}'_2 = \mathsf{OPT}_2 - \mathsf{D} \geq \frac{\mathsf{OPT}}{2} - \mathsf{D} > \frac{0.43}{d} - \frac{1}{n} > \frac{0.43}{d} - \frac{1}{2d \ln n} > \frac{0.4}{d}$.
- Thus, $\dfrac{\mathbf{W}'_{\text{total}}}{\mathsf{OPT}'_2} < \frac{2d}{0.4} = 5d$.
- Thus, $\mathbb{E}[\kappa] > \frac{\mathsf{OPT}'_2}{T^2} - 4\,\mathrm{e}^{-\frac{T^2}{2}} d\,\mathsf{OPT}'_2 = \frac{\mathsf{OPT}'_2}{4 \ln d} - \frac{4d}{d^2}\,\mathsf{OPT}'_2 > \frac{\mathsf{OPT}'_2}{4.1 \ln d}$.

Thus, the final modularity value achieved is at least

$$\begin{aligned}
\frac{\mathsf{OPT}'_2}{4.1 \ln d} - \mathsf{D} &= \frac{\mathsf{OPT}_2 - \mathsf{D}}{4.1 \ln d} - \mathsf{D} \\
&= \frac{\mathsf{OPT}_2}{4.1 \ln d} - \left(1 + \frac{1}{4.1 \ln d}\right) \left(\frac{0.4}{d} + \frac{1}{n}\right) \left(\frac{d}{d + 0.4n}\right) \\
&> \left(\frac{1}{4.1 \ln d} - \left(1 + \frac{1}{4.1 \ln d}\right) \left(\frac{1}{1 + 0.4\frac{n}{d}}\right)\right) \mathsf{OPT}_2 \\
> \left(\frac{1}{4.1 \ln d} - \left(1 + \frac{1}{4.1 \ln d}\right) \left(\frac{1}{1 + 0.8 \ln n}\right)\right) \mathsf{OPT}_2 &> \frac{\mathsf{OPT}_2}{4.2 \ln d} > \frac{\mathsf{OPT}}{8.4 \ln d}
\end{aligned}$$

(b) The Case When the Input Graph is Weighted

Since the given graph can be assumed to be connected, $\Delta \geq 1 - \frac{1}{n}$. We want to design an $O(\log d_{\max})$-approximation algorithm assuming $d_{\max} < \frac{\sqrt[5]{n}}{16 \ln n}$. Again, we first provide a lower bound for OPT.

(* $\mathcal{S}$ denotes the set of clusters *)
(* initialization *)
$\mathcal{S} = \emptyset$; $V'' = V$; $E'' = E' = \big\{ \{u,v\} \mid \{u,v\} \in E \,\&\, \ell(u,v) < 1/2 \big\}$; $\forall\, u \in V : C_u = \emptyset$
(* Algorithm *)
while the graph (V'', E'') contains at least one edge **do**
    pick a node $v \in V''$ that maximizes $L(v) = \sum_{\{u,v\} \in E''} \ell(u,v)$
    $C_v = \{v\} \cup \{u \mid \{u,v\} \in E''\}$; add the new cluster C_v to $\mathcal{S}$
    $V'' = V'' \setminus C_v$; $E'' = (V'' \times V'') \cap E'$
endwhile
for every $v \in V''$ **do**
    add the cluster $\{v\}$ to $\mathcal{S}$
endfor

Figure 2: Greedy algorithm for computing lower bounds for weighted graphs.

Lemma 4.9 (Lower bound on OPT for weighted graphs)**.** *If* $d_{\max} < \dfrac{\sqrt[5]{n}}{16 \ln n}$ *then* $\mathsf{OPT} > \dfrac{1}{8\, d_{\max}}$.

Proof. We execute the greedy algorithm on G' as shown in Fig. 2. Note that the graph $G' = (V, E')$ has a maximum weighted degree of precisely $d_{\max}$, The number of nodes adjacent to any node v in G' is at most $2\, d_v \leq 2\, d_{\max}$, and $\ell(E') = \sum_{\{u,v\} \in E'} \ell(u,v) = m - \sum_{\{u,v\} \in E \setminus E'} \ell(u,v) \geq m/2$.

Let $L(C_v) = \sum_{\substack{u,v \in C_v \\ u \neq v}} \ell(u,v)$. Since the weight of any edge in E' is at least $1/2$, it is easy to see that during each selection of cluster C_v, $L(C_v)$ is at least $1/d_{\max}$ times the total weight of edges whose one end-point was in C_v. Thus, $\sum_{C_v} L(C_v) \geq \dfrac{\ell(E')}{d_{\max}+1} \geq \dfrac{m}{2\,(d_{\max}+1)} = \dfrac{n\,\Delta}{2\,(d_{\max}+1)}$. Note that for all sufficiently large n,

$$w_{u,v} = \begin{cases} \dfrac{\ell_{u,v} - \frac{d_u d_v}{n\,\Delta}}{n\,\Delta} \geq \dfrac{\ell_{u,v} - \frac{(d_{\max})^2}{n\,\Delta}}{n\,\Delta} \geq \dfrac{\ell_{u,v}}{2\,n\,\Delta}\,, & \text{if } \{u,v\} \in E \\[2ex] \dfrac{-\,d_u d_v}{(n\,\Delta)^2} \geq \dfrac{-\,(d_{\max})^2}{n^2\Delta^2} \geq -\,\dfrac{1}{256\, n^{1.6} \ln^2 n\, \Delta^2}\,, & \text{otherwise.} \end{cases}$$

Thus, for all sufficiently large n, we have

$$\mathsf{M}(\mathcal{S}) = \sum_v \mathsf{M}(C_v) - \sum_{u \in V \setminus \left(\bigcup_v C_v\right)} w_{u,u} \geq \frac{\displaystyle\sum_{\substack{C_v \in \mathcal{S} \\ C_v \neq \emptyset}} \left(\sum_{\substack{u,v \in C_v \\ \{u,v\} \in E}} \ell_{u,v} \right)}{2\,n\,\Delta} - \frac{n\,(d_{\max})^2}{256\,n^{1.6}\,\ln^2 n\,\Delta^2}$$

$$\geq \frac{\sum_v L(C_v)}{2n\Delta} - \frac{n\,(d_{\max})^2}{512\,n^{1.6}\,\ln^2 n\,\Delta^2} - \frac{n\,(d_{\max})^2}{256\,n^{1.6}\,\ln^2 n\,\Delta^2} \geq \frac{\frac{n\Delta}{2(d_{\max}+1)}}{2n\Delta} - \frac{1}{512\,n^{1/5}\,\ln^4 n}$$

$$= \frac{1}{4\,(d_{\max}+1)} - \frac{1}{512\,n^{1/5}\,\ln^4 n} > \frac{1}{8\,d_{\max}} \qquad \square$$

Since $d_{\max} < \frac{\sqrt[5]{n}}{16 \ln n}$ and $\Delta \geq 1 - \frac{1}{n}$, $\mathsf{D} \leq \frac{n(d_{\max})^2}{2(n\,\Delta)^2} = \frac{1}{2n}\left(\frac{d_{\max}}{\Delta}\right)^2 \leq \frac{1}{512\,n^{3/5}\,\ln^2 n}$. Selecting $T = \sqrt{16 \ln d_{\max}}$ in Theorem 4.6, we have the following chain of implications:

- $\mathsf{OPT}'_2 = \mathsf{OPT}_2 - \mathsf{D} \geq \dfrac{\mathsf{OPT}}{2} - \mathsf{D} = \dfrac{1}{16\,d_{\max}} - \dfrac{1}{512\,n^{3/5}\,\ln^2 n} > \dfrac{1}{17\,d_{\max}}$.
- Thus, $\dfrac{\mathbf{W}'_{\text{total}}}{\mathsf{OPT}'_2} < 34\,d_{\max}$.
- Thus, $\mathbb{E}[\kappa] > \dfrac{\mathsf{OPT}'_2}{T^2} - 34\,\mathrm{e}^{-\frac{T^2}{2}}\,d_{\max}\,\mathsf{OPT}'_2 > \dfrac{\mathsf{OPT}'_2}{17\,\ln d_{\max}}$.

and thus the final modularity value achieved is at least

$$\frac{\mathsf{OPT}_2}{17 \ln d_{\max}} - \mathsf{D} = \frac{\mathsf{OPT}}{O(\ln d_{\max})} \qquad \square$$

5. Other Results

5.1. Modularity Clustering for Directed Weighted Graphs

Leicht and Newman [25] generalized the modularity measure to weighted directed graphs in the following manner. Let $G = (V, E, \ell)$ be the input *directed* graph with $\ell \colon E \mapsto \mathbb{R}^+$ being the function mapping edges to non-negative weights. For a node $v \in V$, let d_v^{in} and d_v^{out} denote the *weighted*

in-degree and the *weighted out-degree* of v, respectively. Let $m = \sum_{v \in V} d_v^{\text{in}} + \sum_{v \in V} d_v^{\text{out}}$ and let $A = [a_{u,v}]$ denote the weighted adjacency matrix of G, *i.e.*, $a_{u,v} = \ell(u,v)$ if $(u,v) \in E$ and $a_{u,v} = 0$ otherwise. Note that the matrix A is *not* necessarily symmetric now. Then, Equation (1) computing the modularity value of a cluster $C \subseteq V$ needs to be modified as

$$\mathsf{M}(C) = \frac{1}{m} \left(\sum_{u,v \in C} \left(a_{u,v} - \frac{d_u^{\text{out}}\, d_v^{\text{in}}}{m} \right) \right)$$

With some effort, we show that we can extend all our complexity results for undirected networks to directed networks. Let $\Delta = \dfrac{\sum_{v \in V} d_v^{\text{in}}}{n} = \dfrac{\sum_{v \in V} d_v^{\text{out}}}{n}$ denote the average weighted degree of nodes of G, and let $d_{\max}^{\text{in}} = \max\limits_{v \in V} d_v^{\text{in}}$ and $d_{\max}^{\text{out}} = \max\limits_{v \in V} d_v^{\text{out}}$ denote the maximum weighted in-degree and maximum weighted out-degree, respectively, of nodes in G. For convenience, we normalize all the weights such that $\sum_{v \in V} d_v^{\text{in}} + \sum_{v \in V} d_v^{\text{out}}$ is *exactly* twice the number of directed edges of G. Since the given graph can be assumed to be weakly-connected, $\Delta \geq 1 - \frac{1}{n}$.

Theorem 5.1.[13]
(a) *Computing* OPT_2 *is* NP*-complete even if every node* v *has* $d_v^{\text{in}} = d_v^{\text{out}} = d$*, for any fixed* $d \geq 9$*.*
(b) *It is* NP*-hard to approximate the* k*-clustering problem, for any* k*, within a factor of* $1 + \varepsilon$ *for some constant* $\varepsilon > 0$ *even if every node of the given directed graph has* $d_v^{\text{in}} = d_v^{\text{out}} = n - 4$*.*
(c) *There is an* $O(\log d)$ *approximation algorithm for* unweighted *directed graphs if the in-degree and out-degree of all nodes is exactly the same, say* d*, and* $d \leq \frac{n}{100 \ln n}$*.*
(d) *There is an* $O\left(\log\left(d_{\max}^{\text{in}} + d_{\max}^{\text{out}}\right)\right)$*-approximation algorithm for weighted graphs provided* $\max\left\{ d_{\max}^{\text{in}}, d_{\max}^{\text{out}} \right\} \leq \frac{\sqrt[5]{n}}{64 \ln n}$*.*

Proof. Remember that

$$\mathsf{M}(C) = \frac{1}{m} \left(\sum_{u,\, v \in C} \left(a_{u,v} - \frac{d_u^{\text{out}} d_v^{\text{in}}}{m} \right) \right) \tag{7}$$

[13]We made no serious attempts to optimize various constants in this theorem.

The corresponding modification in Equation (3) is

$$\mathsf{M}(\mathcal{S}) = \sum_{C_i \in \mathcal{S}} \left(\frac{m_i}{m} - \left(\frac{D_i^{\text{in}} \times D_i^{\text{out}}}{m^2} \right) \right) \tag{8}$$

where $D_i^{\text{in}} = \sum_{v \in C_i} d_v^{\text{in}}$, $D_i^{\text{out}} = \sum_{v \in C_i} d_v^{\text{out}}$ and m_i as the total weight of edges whose both endpoints are in the cluster C_i. Finally, since $\sum_{v \in V} \left(a_{u,v} - \frac{d_u^{\text{out}} d_v^{\text{in}}}{m} \right)$ $= \sum_{v \in V} \left(a_{u,v} - \frac{d_u^{\text{in}} d_v^{\text{out}}}{m} \right) = 0$ for any $u \in V$, we can alternatively express $\mathsf{M}(C)$ as $\mathsf{M}(C) = \frac{1}{m} \left(\sum_{u \in C,\, v \notin C} \left(\frac{d_u^{\text{out}} d_v^{\text{in}}}{m} - a_{u,v} \right) \right)$. Thus, Equation (5) now becomes

$$\mathsf{M}(\mathcal{S}) = \sum_{C_i, C_j} \left(\frac{D_i^{\text{out}} D_j^{\text{in}}}{m^2} - \frac{m_{ij}}{m} \right) \tag{9}$$

where m_{ij} as the total weight of the edges directed from C_i to C_j.

(a) & (b) These two results follow by the following easy observation. Consider a given undirected unweighted graph G with n nodes and m edges, and let $\widetilde{G}$ be the directed graph obtained by replacing each edge $\{u, v\}$ of G by two directed edges (u, v) and (v, u), each of weight 1; thus $\widetilde{m} = \sum_{v \in V} d_v^{\text{in}} + \sum_{v \in V} d_v^{\text{out}} = 4m$. Let $\widetilde{A} = [\widetilde{a}_{u,v}]$ be the adjacency matrix of $\widetilde{G}$, and $\widetilde{d}_v^{\text{in}}$ and $\widetilde{d}_v^{\text{out}}$ be the in-degree and out-degree of the node v in $\widetilde{G}$. Then, it is easy to see that every clustering of G of modularity value x translates to a corresponding clustering of $\widetilde{G}$ of the same modularity value and vice versa.

(c) & (d) It is easy to see that the proof of Lemma 2.1 works for directed networks as well by using Equation (9) instead of Equation (5) in the proof. Thus again it suffices to approximate OPT_2.

Let $W = [w_{u,v}] \in \mathbb{R}^{n \times n}$ be the matrix whose entries are defined by $w_{u,v} = \dfrac{a_{u,v} - \dfrac{d_u^{\text{out}} d_v^{\text{in}}}{m}}{2m}$. Then, letting $x_u \in \{-1, 1\}$ be the indicator variable denoting in which partition the node $u \in V$ belongs, Equation (7) can be rewritten

for a 2-clustering of directed networks as

$$\mathsf{M}(\mathcal{S}) = \sum_{u,v \in V} w_{u,v} \left(1 + x_u x_v\right) = \sum_{u,v \in V} w_{u,v} x_u x_v$$
$$= \mathbf{x}^{\mathrm{T}} W \mathbf{x} = \mathbf{x}^{\mathrm{T}} \left(\frac{W + W^{\mathrm{T}}}{2}\right) \mathbf{x} = \mathbf{x}^{\mathrm{T}} W' \mathbf{x}$$

where $W' = \frac{W+W^{\mathrm{T}}}{2} = [w'_{u,v}]$ is a *symmetric* matrix. Note that $w'_{u,v} = \dfrac{\delta_{u,v} - \dfrac{d_u^{\mathrm{out}} d_v^{\mathrm{in}} + d_u^{\mathrm{in}} d_v^{\mathrm{out}}}{2m}}{2m}$ where $\delta_{u,v}$ is given by:

$$\delta_{u,v} = \delta_{v,u} = \begin{cases} 1, & \text{if both } (u,v) \in E \text{ and } (v,u) \in E \\ 0, & \text{if both } (u,v) \notin E \text{ and } (v,u) \notin E \\ 1/2, & \text{otherwise.} \end{cases}$$

Let $\widehat{W} = [\widehat{w_{u,v}}]$ be the real symmetric matrix defined by $\widehat{w_{u,v}} = \begin{cases} 0, & \text{if } u = v \\ w'_{u,v}, & \text{otherwise.} \end{cases}$ As in the proof of Theorem 4.5, it follows that $\sum_{u,v \in V} \widehat{w_{u,v}} < 2$. For notational convenience, define $\mathsf{D} = \mathrm{trace}\left(\widehat{W} - W'\right) = \sum_{u \in V} w'_{u,u}$ and $\mathsf{OPT}'_2 = \max_{\mathbf{x} \in \{0,1\}^n} \mathbf{x}^{\mathrm{T}} \widehat{W} \mathbf{x}$.

(c) G is an unweighted directed graph with $d_v^{\mathrm{in}} = d_v^{\mathrm{out}} = d$ for every node v, and $d \leq \frac{n}{5 \ln n}$.

The proof of Theorem 4.5 on the quadratic form $\max_{\mathbf{x} \in \{0,1\}^n} \mathbf{x}^{\mathrm{T}} \widehat{W} \mathbf{x}$ gives an approximation factor of $\gamma \ln d$, for some constant $\gamma > 0$, for our directed network provided we can show that

- $\dfrac{\mathsf{OPT}'_2}{\gamma \ln d} - \mathsf{D} = \Omega\left(\dfrac{\mathsf{OPT}'_2}{\gamma \ln d}\right)$, and
- $\mathsf{OPT}_2 = \Omega\left(d^{-c}\right)$ for some constant $c > 0$.

Let H be the undirected graph obtained from the given graph G by ignoring the direction of the edges and removing parallel edges (if any); every node in H has a degree between d and $2d$. Greedily pick a *maximal* matching in H, each time selecting an edge and deleting all (at most $4d - 1$) edges that have a common end-point with the picked edge. Such a matching contains at least

$\frac{(nd)/2}{4d} = \frac{n}{8}$ edges, each of weight at least $\frac{1}{4m} - \frac{8d^2}{4m^2} = \frac{1}{8dn} - \frac{1}{2n^2}$ in G. Consider the clustering of G where each edge in the matching is a separate cluster of two nodes, and each of the remaining nodes is a separate cluster of one node. The modularity value of this solution is *at least*

$$\left(\frac{1}{8dn} - \frac{1}{2n^2}\right)\frac{n}{8} - \operatorname{trace}\left(W' - \widehat{W}\right) \geq \frac{1}{64d} - \frac{1}{16n} - \frac{1}{2n}$$

Thus, $\mathsf{OPT}'_2 \geq \frac{1}{128\,d} - \frac{9}{32\,n} = \Omega\left(d^{-1}\right)$. Moreover, since $d \leq \frac{n}{100 \ln n}$ we have

$$\frac{\mathsf{OPT}'_2}{\ln d} - \mathsf{D} = \frac{\mathsf{OPT}'_2}{\ln d} - \frac{1}{2n} = \Omega\left(\frac{\mathsf{OPT}'_2}{\ln d}\right)$$

(d) $\max\left\{d^{\text{in}}_{\max}, d^{\text{out}}_{\max}\right\} < \frac{\sqrt[5]{n}}{64 \ln n}$.

Let $G'' = (V, E'')$ be the undirected weighted graph obtained from G whose adjacency matrix is $W'' = \left[w''_{u,v}\right]$ with $w''_{u,v} = \begin{cases} w'_{u,v} - \frac{1}{2}, & \text{if } \delta_{u,v} = 1 \\ w'_{u,v}, & \text{otherwise.} \end{cases}$

Since $w''_{u,v} \geq w'_{u,v}$, it suffices to show an approximation for $\max_{\mathbf{x}\in\{0,1\}^n} \mathbf{x}^{\mathrm{T}} W'' \mathbf{x}$. The algorithm in the proof of Theorem 4.7(b) with $W = W''$ can now be appropriately modified to obtain the desired approximation if one identified the quantity $d_{\max}$ in that proof with $d^{\text{in}}_{\max} + d^{\text{out}}_{\max}$. □

5.2. Alternative Modularity Measure: the max-min *Objective*

Exact or approximate solutions to the modularity measure may produce many *trivial* clusters of single nodes. For example, the following proposition shows that for a large class of graphs there exists a clustering in which every cluster except one consists of a single node gives a modularity value that has a modularity value of *at least* 25% of the optimal.

Proposition 5.2. *There exists a clustering for a graph G in which every cluster except one consists of a single node and whose modularity value is at least 25% of the optimal if*

- *G is d-regular with $d < \frac{n}{2 \ln n}$, or*
- *G is an undirected weighted graph with $d_{\max} < \frac{\sqrt[5]{n}}{16 \ln n}$.*

Proof. Let $\{V', V \setminus V'\}$ be an optimal 2-clustering of G. By Lemma 2.1, $\mathsf{OPT}_2 \geq \mathsf{OPT}/2$. By Lemma 2.2 $\mathsf{M}(V') = \mathsf{OPT}_2/2 = \mathsf{OPT}/4$. Suppose that we replace the cluster $V \setminus V'$ by $|V \setminus V'|$ trivial clusters each of a single node, and let C be this new clustering If G is d-regular, then $\mathsf{M}(C) = \mathsf{M}(V') - \mathsf{D} = \frac{\mathsf{OPT}}{4} - \frac{1}{n}$. By Lemma 4.7, $\mathsf{OPT} > \frac{0.86}{d} - \frac{4}{n}$, and thus $\mathsf{M}(C) = \frac{\mathsf{OPT}}{4} - \mathrm{o}(1)$. Similarly, for the case when G is undirected weighted with $d_{\max} < \frac{\sqrt[5]{n}}{16 \ln n}$, the proof of Theorem 4.5 shows that $\mathsf{D} \leq \frac{1}{512\, n^{3/5} \ln^2 n}$, and thus $\mathsf{M}(C) = \mathsf{M}(V') - \mathsf{D} \geq \frac{\mathsf{OPT}}{4} - \frac{1}{512\, n^{3/5} \ln^2 n}$. By Lemma 4.9 $\mathsf{OPT} > \frac{1}{8\, d_{\max}}$, and thus again $\mathsf{M}(C) = \frac{\mathsf{OPT}}{4} - \mathrm{o}(1)$. □

We investigate one alternative to overcome such a shortcoming: define the modularity of the network as the *minimum* of the modularities of individual clusters. Equation (2) now becomes

$$\mathsf{M}^{\mathsf{max\text{-}min}}(\mathcal{S}) = \min_{C_i \in \mathcal{S}} \mathsf{M}(C_i)$$

We will add the superscript "max-min" to differentiate the relevant quantities for this objective from the usual summation objective discussed before, *e.g.*, we will use $\mathsf{OPT}^{\mathsf{max\text{-}min}}$ instead of OPT. In a nutshell, our results in the following lemma show that the max-min objective indeed avoids generating trivial clusters (Lemma 5.3(a)), and the optimal objective value for max-min objective is precisely scaled by a factor of 2 from that of the SUM objective, thereby keeping the overall quantitative measure the same (Lemma 5.3(b)).

Lemma 5.3. *Let G be a weighted undirected graph with m edges and maximum degree $d_{\max}$. Then, the following claims hold:*

(a) *No optimal solution for* max-min *objective has a cluster with fewer than* $\frac{4\, m\, \mathsf{OPT}^{\mathsf{max\text{-}min}}}{d_{\max}}$ *nodes.*

(b) $\mathsf{OPT}^{\mathsf{max\text{-}min}} = \frac{\mathsf{OPT}_2}{2}$.

Proof.
(a) Since only an edge with positive weight can increase the modularity of a cluster, it is easy to check that a cluster with y nodes can have a modularity value of at most $\frac{y\, d_{\max}}{4\, m}$.

(b) Consider an optimal clustering $\mathcal{S} = \{V_1, V_2, \ldots, V_k\}$ with a *minimum* number k of clusters such that $\mathsf{OPT}^{\mathsf{max\text{-}min}} = \mathsf{M}^{\mathsf{max\text{-}min}}(\mathcal{S}) = \min_{1 \leq i \leq k} \{\mathsf{M}(V_i)\} >$

0. First, consider the case when $k > 3$. We will show that for some non-empty subset T of $\{V_1, V_2, \ldots, V_k\}$ we must have $\mathsf{M}(\cup_{V_j \in T} V_j) \geq \mathsf{M}^{\mathsf{max\text{-}min}}(\mathcal{S})$; this contradicts the minimality of k in our choice of of the optimal cluster. Note that $\mathsf{M}(\mathcal{S}) = \sum_{i=1}^{k} \mathsf{M}(V_i) \geq k \cdot \mathsf{M}^{\mathsf{max\text{-}min}}(\mathcal{S})$. We will make use of Equation (1) of modularity of a cluster. Let $\mathsf{M}(\widetilde{\mathcal{S}}) = \frac{1}{2m}\left(\sum_{\substack{u \in V_i,\, v \in V_j \\ i \neq j}} \left(a_{u,v} - \frac{d_u d_v}{2m}\right)\right)$. Then, $\mathsf{M}(\widetilde{\mathcal{S}}) = -\mathsf{M}(\mathcal{S})$. Consider a subset T obtained by randomly and uniformly selecting each V_i with a probability of $1/2$. Note that each pair of nodes u and v belonging to the same cluster is selected with a probability of $1/2$, whereas each pair of nodes belonging to different clusters is selected with a probability of $1/4$. Thus,

$$\begin{aligned}\mathbb{E}\left[\,\mathsf{M}(\cup_{V_j \in T} V_j)\,\right] &= \frac{\mathsf{M}(\mathcal{S})}{2} + \frac{\mathsf{M}(\widetilde{\mathcal{S}})}{4} = \frac{\mathsf{M}(\mathcal{S})}{4} \\ &\geq \left(\frac{k}{4}\right)\mathsf{M}^{\mathsf{max\text{-}min}}(\mathcal{S}) \geq \mathsf{M}^{\mathsf{max\text{-}min}}(\mathcal{S})\end{aligned}$$

and therefore there exists such a subset T with the properties as claimed.

Otherwise, consider the case when $k = 3$. Let $\mathsf{M}_{i,j} = \dfrac{\sum_{\substack{u \in V_i \\ v \in V_j}} \left(a_{u,v} - \frac{d_u d_v}{2m}\right)}{2m}$ for $i < j$. Without loss of generality, let $\mathsf{M}(V_1) = a$, $\mathsf{M}(V_2) = a + b$ and $\mathsf{M}(V_3) = a + c$ for some $a > 0$ and $b \geq c \geq 0$; thus, $\mathsf{M}^{\mathsf{max\text{-}min}}(\mathcal{S}) = a$. Consider the three 2-clusterings of G: $C_1 = \big(V_1 \cup V_2, V_3\big)$, $C_2 = \big(V_2 \cup V_3, V_1\big)$ and $C_3 = \big(V_1 \cup V_3, V_2\big)$. Since none of these three 2-clusterings should be an optimal solution, we must have

$$\begin{aligned}&\mathsf{M}^{\mathsf{max\text{-}min}}(C_1) - \mathsf{M}^{\mathsf{max\text{-}min}}(\mathcal{S}) < 0 \\ &\qquad\qquad \equiv \min\big\{2a + b + \mathsf{M}_{1,2},\, a + c\big\} < a \equiv \mathsf{M}_{1,2} < -(a + b)\end{aligned}$$

$$\begin{aligned}&\mathsf{M}^{\mathsf{max\text{-}min}}(C_2) - \mathsf{M}^{\mathsf{max\text{-}min}}(\mathcal{S}) < 0 \\ &\qquad\qquad \equiv \min\big\{2a + b + c + \mathsf{M}_{2,3},\, a\big\} < a \equiv M_{2,3} < -(a + b + c)\end{aligned}$$

$$\begin{aligned}&\mathsf{M}^{\mathsf{max\text{-}min}}(C_3) - \mathsf{M}^{\mathsf{max\text{-}min}}(\mathcal{S}) < 0 \\ &\qquad\qquad \equiv \min\big\{2a + c + \mathsf{M}_{1,3},\, a\big\} < a \equiv M_{1,3} < -(a + c)\end{aligned}$$

Thus, we have $\mathsf{M}(V_1) + \mathsf{M}(V_2) + \mathsf{M}(V_3) = 3a + b + c = -\mathsf{M}_{1,2} - \mathsf{M}_{2,3} - \mathsf{M}_{1,3} > 3a + 2b + 2c$ which implies $b + c < 0$, contradicting $b \geq c \geq 0$.

Thus, we have shown there is an optimal solution for our max-min objective with no more than two clusters. Obviously, if $\mathsf{OPT}^{\text{max-min}} > 0$ then an optimal solution cannot consist of a single cluster. Let V_1, V_2 be the two clusters in this case. By Lemma 2.2, we have $\mathsf{M}(V_1) = \mathsf{M}(V_2)$ which implies $\mathsf{OPT}^{\text{max-min}} = \frac{\mathsf{OPT}_2}{2}$. □

5.3. Alternative Null Model: Erdös-Rényi Random Graphs

A theoretically appealing choice for alternative null models is the classical Erdös-Rényi random graph model $G(n, p)$, namely each possible edge $\{u, v\}$ is selected in G uniformly and randomly with a probability of p for some fixed $0 < p < 1$. To summarize, our results in this section show that the new modularity measure is precisely Newman's modularity measure on an appropriately defined regular graph, and thus our previous results on regular graphs can be applied to this case.

We will add the superscript "ER" to differentiate the relevant quantities for this objective from the usual summation objective discussed before, *e.g.*, we will use $\mathsf{OPT}^{\mathsf{ER}}$ instead of OPT. For simplicity, we consider the case of *unweighted graphs only*. Let $G = (V, E)$ be the given unweighted input graph with $m = n\,\Delta$ number of edges. Select $p = \frac{2\,\Delta}{n-1}$ such that the null model has the *same* number of edges *in expectation* as the given graph G. Equation (1) then becomes

$$\mathsf{M}^{\mathrm{ER}}(C) = \frac{\sum_{u,v \in C} (a_{u,v} - p)}{2m}$$

Let n be sufficiently large such that $p \approx (2\Delta)/n$. It can then be seen that $\mathsf{M}^{\mathsf{ER}}(C)$ is precisely the same as $\mathsf{M}(C)$ on a (2Δ)-regular graph. Thus, our previous results on regular graphs can be generalized to this case in the following manner:

- Computing $\mathsf{OPT}^{\mathsf{ER}}$ is NP-complete for graphs with $\Delta \geq 18$.
- If $\Delta < \frac{n}{4 \ln n}$ then the problem admits a $O(\log \Delta)$-approximation.

Acknowledgements We thank Mario Szegedy for suggestion to investigate the problem for dense graphs and other useful discussions, Geetha Jagannathan and Alantha Newman for useful discussions, and Mark Newman for

explaining the significance of negative self-loops in his modularity measure and pointing out references [23, 25, 27].

References

[1] G. Agarwal and D. Kempe, *Modularity-Maximizing Graph Communities via Mathematical Programming*, European Physics Journal B, 66/3, 2008.

[2] R. Albert. *Scale-free networks in cell biology*, Journal of Cell Science, 118, 4947-4957, 2005.

[3] R. Albert, H. Jeong and A-L. Barabási. *The Diameter of the World-Wide Web*, Nature, 401, 130-131, 1999.

[4] N. Alon. *On the edge expansion of graphs*, Combinatorics, Probability and Computing, 6, 145-152, 1997.

[5] N. Alon, K. Makarychev, Y. Makarychev, and A. Naor. *Quadratic forms on graphs*, Proceedings of the thirty-seventh annual ACM symposium on Theory of computing, 486-493, 2005.

[6] N. Alon and A. Naor. *Approximating the cut-norm via Grothendieck's inequality*, SIAM Journal of Computing, 35(4), 787-803, 2006.

[7] S. Arora, D. Karger and M. Karpinski. *Polynomial Time Approximation Schemes for Dense Instances of* NP*-Hard Problems*, Journal of Computer & System Sciences, 58 (1), 193-210, 1999.

[8] N. Bansal, A. Blum, and S. Chawla. *Correlation clustering*, Machine Learning, 56 (1-3), 89-113, 2004.

[9] P. Berman and M. Karpinski. *On some tighter inapproximability results*, Proceedings of the twenty-sixth International Colloquium on Automata, Languages, and Programming, 200-209, 1999.

[10] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski and D. Wagner, *On Modularity Clustering*, IEEE Transaction on Knowledge and Data Engineering, 20 (2), 172-188, 2007.

[11] S. Cafieri, P. Hansen, L. Liberti. *Improving heuristics for network modularity maximization using an exact algorithm*, Proceedings of the third international workshop on model-based metaheuristics, 130-139, 2010.

[12] M. Charikar, V. Guruswami, and A. Wirth. *Clustering with qualitative information*, Proceedings of the forty-fourth Annual IEEE Symposium on Foundations of Computer Science, 524-533, 2003.

[13] M. Charikar and A. Wirth, *Maximizing quadratic programs: extending Grothendieck's inequality*, Proceedings of the forty-fifth Annual IEEE Symposium on Foundations of Computer Science, 54-60, 2004.

[14] M. Chlebík and J. Chlebíková. *Complexity of approximating bounded variants of optimization problems*, Theoretical Computer Science, 354 (3), 320-338, 2006.

[15] H. N. Djidjev. *A scalable multilevel algorithm for graph clustering and community structure detection*, in Algorithms and Models for the Web-Graph, W. Aiello, A. Broder, J. Janssen, and E. Milios (Eds.), LNCS 4936, 117-128, Springer-Verlag, 2007.

[16] S. Fortunato and M. Barthélemy. *Resolution limit in community detection*, Proceedings of the National Academy of Sciences, 104 (1), 36-41, 2007.

[17] J. Friedman, J. Kahn and E. Szemerédi. *On the Second Eigenvalue in Random Regular Graphs*, Proceedings of the twenty-first annual ACM symposium on Theory of computing, 587-598, 1989.

[18] A. Frieze and R. Kannan, *The regularity lemma and approximation schemes for dense problems*, Proceedings of the thirty-seventh Annual IEEE Symposium on Foundations of Computer Science, 12-20, 1996.

[19] M. Gaertler, R. Görke and D. Wagner. *Significance-Driven Graph Clustering*, Algorithmic Aspects in Information and Management, LNCS, 4508, 11-26, Springer Verlag, 2007.

[20] R. Guimer‘a, M. Sales-Pardo and L. A. N. Amaral. *Classes of complex networks defined by role-to-role connectivity profiles*, Nature Physics, 3, 63-69, 2007.

[21] E. Hazan and S. Kale. *Approximating Quadratic Programs with Positive Semidefinite Constraints*, Computer, 1, 1-3, 2008.

[22] M. A. Henning and A. Yeo, *Tight Lower Bounds on the Size of a Maximum Matching in a Regular Graph*, Graphs and Combinatorics, 23 (6), 647-657, 2007.

[23] B. Karrer and M. E. J. Newman. *Random graph models for directed acyclic networks*, Physical Review E, 80, 046110, 2009.

[24] D. Kefeng, Z. Ping and Z. Huisha. *Graph Separation of* 4*-regular Graphs is* NP*-complete*, Journal of Mathematical Study, 32 (2), 1999.

[25] E. A. Leicht and M. E. J. Newman, *Community Structure in Directed Networks*, Physical Review Letters, 100, 118703, 2008.

[26] M. E. J. Newman, personal communication, 2009.

[27] M. E. J. Newman. *Analysis of weighted networks*, Physical Review E, 70, 056131, 2004.

[28] M. E. J. Newman, *Modularity and community structure in networks*, Proceedings of the National Academy of Sciences, 103 (23), 8577-8582, 2006.

[29] M. E. J. Newman. *Finding community structure in networks using the eigenvectors of matrices*, Physical Review E, 74, 036104, 2006.

[30] M. E. J. Newman and M. Girvan, *Finding and evaluating community structure in networks*, Physical Review E, 69, 026113, 2004.

[31] A. Noack and R. Rotta. *Multi-level Algorithms for Modularity Clustering*, Proceedings of the eighth International Symposium on Experimental Algorithms, 257-268, 2009.

[32] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai and A.-L. Barabási. *Hierarchical Organization of Modularity in Metabolic Networks*, Science, 297 (5586), 1551-1555, 2002.

[33] C. Swamy. *Correlation clustering: maximizing agreements via semidefinite programming*, Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms, 526-527, 2004.