Adding Cohesion Constraints to Models for Modularity Maximization in Networks

Sonia Cafieri, Alberto Costa, Pierre Hansen

Les Cahiers du GERAD ISSN: 0711–2440

Adding Cohesion Constraints to Models for Modularity Maximization in Networks

S. Cafieri, A. Costa,
P. Hansen

G–2014–11

March 2014

Adding Cohesion Constraints to Models for Modularity Maximization in Networks

Sonia Cafieri

ENAC, MAIAA, F-31055 Toulouse, France
and Université de Toulouse, IMT
F-31400 Toulouse, France

sonia.cafieri@enac.fr

Alberto Costa

Singapore University of Technology and Design
138682 Singapore

costa@sutd.edu.sg

Pierre Hansen

GERAD & HEC Montréal
Montréal (Québec) Canada, H3T 2A7

pierre.hansen@gerad.ca

March 2014

Les Cahiers du GERAD

G–2014–11

Abstract: Finding communities in complex networks is a topic of much current research and has applications in many domains. On the one hand, criteria for doing so have been proposed, the most studied of which is modularity. On the other hand, properties to be satisfied by each community of a partition have been suggested. It has recently been observed that one of the best known such properties, i.e., Radicchi et al.'s weak condition [F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi, Proc. Natl. Acad. Sci. USA, 101, 2658 (2004)] was not satisfied by one or more communities in a partition which maximizes (approximately) some of the best known criteria. It was therefore proposed by Wang et al. [J-G. Wang, L. Wang, Y-Q. Qui, Y. Wang, X-S. Zhang, Lect. Notes in Oper. Res., 11, 142 (2009)] to merge both approaches by maximizing a criterion subject to the weak condition. We consider the effect of adding five cohesion conditions, one at a time, to a modularity maximization problem. We solve the problems exactly. Strong, semi-strong, and almost-strong cohesion conditions appear to be too restrictive and the extra-weak condition too lax. The weak cohesion condition is verified by some but not all modularity maximizing partitions of real-world problems considered. Imposition of this condition on those partitions for which some communities do not verify it reduces modularity moderately but sometimes changes the optimal number of communities and their composition.

Résumé : La détermination de communautés dans les réseaux complexes fait couramment l'objet de nombreuses recherches et a des applications dans de nombreux domaines. D'une part, des critères ont été proposés; le plus étudié d'entre eux est la modularité. D'autre part, des propriétés qui doivent être satisfaites par chacune des communautés d'une partition de l'ensemble des objets considérés ont été proposées. Il a été récemment observé que l'une des plus connues de ces propriétés, c'est-à-dire la condition faible de Radicchi et al. [F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi, Proc. Natl. Acad. Sci. USA, 101, 2658 (2004)] n'était pas satisfaite par une ou plusieurs communautés de la partition qui maximise (approximativement) certains des critères les plus connus. Il fut donc proposé par Wang et al. [J-G. Wang, L. Wang, Y-Q. Qui, Y. Wang, X-S. Zhang, Lect. Notes in Oper. Res., 11, 142 (2009)] de fusionner les deux approches en maximisant un critère sous la condition faible. Nous considérons l'effet de l'addition de cinq conditions de cohésion, une à la fois, dans un problème de maximisation de la modularité. Nous résolvons ce problème exactemment. Les conditions forte et semi-forte sont trop restrictives et la condition extra-faible trop peu. La condition de cohésion faible est vérifiée par certains mais pas toutes les partitions d'une série de problèmes de la littérature que nous considérons. L'imposition de ces conditions aux partitions pour lesquelles certaines classes ne satisfont pas la condition de cohésion faible diminue faiblement la valeur de modularité mais peut changer, parfois fortement, les conditions d'appartenance des entités aux classes.

Acknowledgments: S. Cafieri has been supported by the French National Research Agency (ANR) through grant ANR 12-JS02-009-01 ATOMIC. A. Costa has been supported by IDC Grant IDG21300102.

1 Introduction

Networks are often used for representation and study of complex systems with applications in many domains. A network, or graph, $G = (V, E)$ consists of a set of vertices V and a set of edges E. Vertices are represented by points and associated with entities of the system under study. Edges are represented by lines joining pairs of vertices and are associated with relations between the entities. The shape of these lines does not matter, but only their presence or absence. Newman recently gave a detailed introduction to networks in [1].

In many complex systems there are sets of entities which share some common characteristics and/or are likely to have some common function. Henceforth, they will be called *communities* or *clusters* or *modules*. In the associated network, these communities correspond to sets of vertices for which the number of inner edges, that is edges joining two vertices of the community, is larger than the (possibly weighted) number of outer edges, that is the number of edges joining two vertices one of which belongs to the community and the other not. Detection of communities is a currently central and much studied problem in the theory and application of network science, with a vast literature, see [2] for an in-depth survey and references therein. Usually, one seeks a partition of the given set of entities into disjoint communities, that is each entity must belong to one and only one community. Sometimes this last condition is relaxed: the aim is to find a covering of the set of entities, that is some entities may belong to several communities. We do not consider overlapping communities in this paper. The quality of the partition obtained can be judged in several ways:

- Some heuristics do not involve a criterion to be optimized. An example is Girvan and Newman's [3] edge removal heuristic in which edges with maximum betweeness are iteratively removed, yielding partitions into an increasing number of communities. Then the quality of the results obtained can only be judged a posteriori, usually based upon some substantive information.
- A variety of criteria to be minimized or maximized have been suggested by several authors. The best known of them is *modularity*, introduced by Girvan and Newman [3]. Modularity of a community is defined as the difference between the number of inner edges and the expected number of edges in a random configuration model which keeps the distribution of vertex degrees unchanged. A large number of heuristics [4–13] provide in moderate computing time a near optimal solution, and a few exact algorithms provide an optimal solution [14–16]. Strengths and weaknesses of modularity are discussed in [2, 17–19]. Another, more recent, criterion is *modularity density* [20].
- Instead of considering an objective function, one may specify conditions to be satisfied by each community of a partition. The first two such conditions, a *strong* and a *weak* one, were proposed by Radicchi et al. in [21]. Two further conditions, a *semi-strong* and a *extra-weak* one, were suggested by Hu et al., in [22]. The *almost-strong* condition was introduced in [23]. Precise definitions and some properties of these five conditions will be given in the next section. We call these five conditions *cohesion conditions*.

Several authors have checked if some of these conditions are violated by some community of an optimal modularity maximizing partition [24, 25]. Moreover, in [26] it is proposed to add the weak constraint to models for modularity maximization and for modularity density maximization.

The aims of the present paper are: (i) to study to what extent optimal partitions for modularity maximization of real world problems satisfy the five conditions mentioned above and (ii) to examine the effect of imposing these constraints one at a time in modularity maximization models. These models are solved exactly. Observe that using exact optimization methods does allow separation of the effect of imposing a constraint from the possible error due to the use of a heuristic (other reasons which justify using exact optimization together with heuristics for their mutual improvement are given in the introduction of [16]).

The paper is organized as follows. Definitions of five cohesion conditions are given in Sect. 2 and some of their properties are discussed. It is then examined in Sect. 3 to what extent the optimal modularity maximizing partitions of 11 real-world problems from the literature do satisfy these conditions. Modifications to models for modularity maximization due to adding each of the five cohesion constraints are studied in Sect. 4. Computational results are presented in Sect. 5 which also includes a detailed discussion of the effect of imposing the weak condition on the optimal partitions of three well known problems. Conclusions are drawn in Sect. 6.

2 Cohesion Conditions

As mentioned above, an important approach to communities detection in network is based on the satisfaction of reasonable *a priori* conditions to have a community. Radicchi et al. [21] proposed two such conditions defining communities in a strong and a weak sense, respectively. Hu et al. [22] propose two further conditions defining communities in (what we call) a semi-strong and in an extra-weak sense, respectively. A further condition, defininining communities in an almost-strong sense, was introduced in [23]. In this section, we give mathematical definitions of these five conditions and discuss some of the properties of corresponding communities.

Let $G = (V, E)$ be a network with vertex set V and edge set E. Recall that the degree k_i of a vertex v_i belonging to V is the number of its neighbors (or adjacent vertices). Let $S \subseteq V$ be a subset of vertices. Then the degree k_i can be separated into two components $k_i^{in}(S)$ and $k_i^{out}(S)$, that is the number of neighbors of v_i inside S and the number of neighbors of v_i outside S. Let M be the number of communities of a partition $S_1, S_2, \ldots, S_M$ of V and let $A = (A_{ij})$ be the adjacency matrix of G, where $A_{ij} = 1$ if an edge joins vertices v_i and v_j and $A_{ij} = 0$ otherwise.

- **Strong Cohesion Condition (SCC):** [21]
  A set of vertices S forms a community in the *strong sense* if and only if every one of its vertices has more neighbors within the community than outside:

$$\forall v_i \in S \quad k_i^{in}(S) > k_i^{out}(S). \tag{1}$$

- **Almost-Strong Cohesion Condition (ASCC):** [23]
  A set of vertices S forms a community in the *almost-strong sense* if and only if every one of its vertices with degree different from 2 has more neighbors within the community than outside, and every vertex having degree 2 has at least one neighbor in the same community:

$$\forall v_i \in S \,|\, k_i \neq 2 \quad k_i^{in}(S) > k_i^{out}(S) \tag{2}$$
$$\forall v_i \in S \,|\, k_i = 2 \quad k_i^{in}(S) > 0. \tag{3}$$

- **Semi-Strong Cohesion Condition (SSCC):** [22]
  A set of vertices S forms a community in the *semi-strong sense* if and only if every one of its vertices has more neighbors within the community than the maximum number of neighbors within any other community:

$$\forall v_i \in S \quad k_i^{in}(S) > \max_{t=1,2,\ldots,M,\, S \neq S_t} \sum_{v_j \in S_t} A_{ij}. \tag{4}$$

  Note that we use a strict inequality in the last formula instead of a non-strict one, as proposed by Hu et al. [22]. This is done for two reasons: first, to express all five conditions in an uniform way and, second, because in this form the strong cohesion condition implies the semi-strong one. It is not the case otherwise as shown on Fig. 1(a). Indeed, the partition into two communities $C_1 = \{1, 2, 3\}$ and $C_2 = \{4, 5, 6\}$ satisfies the semi-strong condition with equality but not the strong condition.
- **Weak Cohesion Condition (WCC):** [21]
  A set of vertices S forms a community in the *weak sense* if and only if the sum of internal degrees within S is larger than the sum of external degrees, that is the number of edges joining S to the rest of the network $V \setminus S$:

$$\sum_{v_i \in S} k_i^{in}(S) > \sum_{v_i \in S} k_i^{out}(S). \tag{5}$$

  This is equivalent to the condition that the number of edges within S is at least half the number of edges in the cut of S. As already observed in [21] the strong cohesion condition implies the weak cohesion condition, but the converse is not true.

- **Extra-Weak Cohesion Condition (EWCC):** [22]
  A set of vertices S forms a community in the *extra-weak sense* if and only if the sum of internal degrees within S is larger than the maximum number of edges joining a vertex of S to a vertex in some other community in the rest of the network:

$$\sum_{v_i \in S} k_i^{in}(S) > \max_{t=1,2,\ldots,M,\, S \neq S_t} \sum_{v_i \in S} \sum_{v_j \in S_t} A_{ij}. \tag{6}$$

  Again, we use a strict inequality in the last formula instead of a non-strict one as suggested in [22].

Let us now consider implications between cohesion conditions. Clearly, the strong cohesion condition implies the weak cohesion condition as inequality (5) is equal to the sum of the inequalities (1) for all vertices $v_i \in S$. The strong cohesion condition implies the semi-strong cohesion condition as they have the same left-hand side and $\max_{t=1,2,\ldots,M} \sum_{v_j \in S_t} A_{ij} \leq k_i^{out}(S)\, \forall v_i \in S$. Note that both conditions are identical when the number of communities M is equal to 2. Similarly, the weak cohesion condition does imply the extra-weak condition as left-hand sides are the same and $\max_{t=1,2,\ldots,M,\, S \neq S_t} \sum_{v_i \in S} \sum_{v_j \in S_t} A_{ij} \leq \sum_{v_i \in S} k_i^{out}(S)$. Again, the right-hand sides are equal when $M = 2$. The strong conditions imply the almost-strong as the latter is equivalent to the former, except for vertices of degree 2 for which the almost-strong condition is simply a weakened version of the strong (the strict inequality being replaced by a non-strict one). Finally, the semi-strong cohesion condition implies the extra-weak cohesion condition as summing the left-hand side of (4) gives the left-hand side of (6) and summing the right hand-side of (4) gives $\sum_{v_i \in S} \max_{t=1,2,\ldots,M,\, S \neq S_t} \sum_{v_j \in S_t} A_{ij} = \max_{t=1,2,\ldots,M,\, S \neq S_t} \sum_{v_i \in S} \sum_{v_j \in S_t} A_{ij}$, that is the right hand side of (6). These implications are represented as follows:

$$\begin{array}{ccccc} ASCC & \Leftarrow & SCC & \Rightarrow & SSCC \\ & & \Downarrow & & \Downarrow \\ & & WCC & \Rightarrow & EWCC \end{array}$$

These implications are asymmetric, i.e., the reverse implication does not hold. Consider for example the graph on Fig. 1(a). This graph with 6 vertices and 9 edges admits a partition into two communities $C_1 = \{1, 2, 3\}$ and $C_2 = \{4, 5, 6\}$ satisfying the conditions WCC and EWCC, but not the conditions SCC and SSCC. Moreover, no other partition in two communities satisfies one or the other of these conditions. So, the implications $SCC \Rightarrow WCC$ and $SSCC \Rightarrow EWCC$ are asymmetric. Let us next consider the graph of Fig. 1(b). This graph with 9 vertices and 14 edges admits a partition into three communities $C_1 = \{1, 2\}$, $C_2 = \{3, 4, 5\}$ and $C_3 = \{6, 7, 8, 9\}$ satisfying the condition EWCC, but not the condition WCC.

3 Cohesion Conditions in modularity maximization

We next check if the optimal solutions obtained by modularity maximization for a set of real-world problems do satisfy or not, and to which degree, the five cohesion conditions described above. To that effect, we use

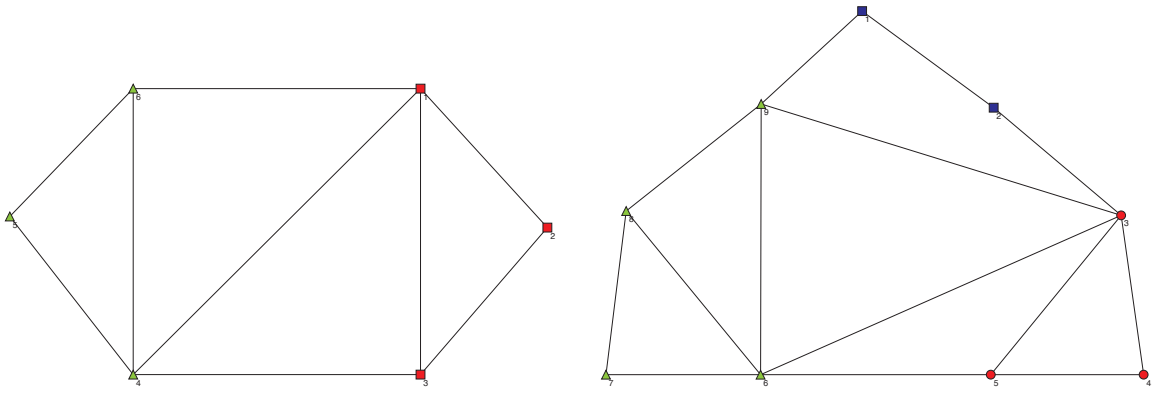

Figure 1: (Color online) (a) Partition satisfying the conditions WCC and EWCC, but not the conditions SCC and SSCC. (b) Partition satisfying the condition EWCC but not the condition WCC.

the optimal solutions of 11 test problems whose maximum modularity partitions have been obtained exactly using the methods presented in [14–16]. The five cohesion conditions have been coded and checked for each cluster of each such partition. Results are summarized in Table 1, which shows that:

- The strong cohesion condition (SCC) appears to be difficult to satisfy: while it holds for at least one of the communities of each optimal partition for all datasets but one, it is satisfied by all communities of one partition only. On average, the strong condition holds for 37.93% of the communities taken together.
- The almost-strong cohesion condition (ASCC) is also difficult to satisfy. It is satisfied by zero to three more clusters than the strong condition. On average, this condition is satisfied by 53.45% of the communities.
- The semi-strong cohesion condition (SSCC) again appears to be difficult to satisfy. Indeed, once more for one only of the problems this condition holds for all communities in the optimal partition. The average number of communities which satisfy the semi-strong condition is 51.72% of them.
- The weak cohesion condition (WCC) appears to be much easier to satisfy than the former three. Indeed, it is satisfied by all communities of 8 out of 11 of the optimal partitions. When this is not the case, that is for `dolphins`, `p53 protein`, and `political books`, only one community does not satisfy the weak cohesion condition. The average number of communities which satisfy this condition is 94.83% of them. The difference between the maximum modularity partitions obtained without and with the weak cohesion condition will be further studied in Sect. 5.
- The extra-weak cohesion condition (EWCC) appears to be easy to satisfy. Indeed, it it verified by all but one of the optimal partitions, that is by `political books`, and in that case it does not hold for one community only. The average number of communities which satisfy the extra-weak condition is 98.28% of them.

In the next section we proceed to add the five types of cohesion constraints, one at a time, to existing models for modularity maximization.

Table 1: Networks from real-world problems: network dimension (n = number of vertices, m = number of edges), number of clusters (M) found by an exact algorithm for modularity maximization and number of clusters verifying the strong condition (M_strong), the almost-strong condition ($M\_almost–strong$), the semi-strong condition ($M\_semi–strong$), the weak condition (M_weak) and the extra-weak condition ($M\_extra–weak$). Percentage of communities satisfying each of these conditions are given in the last line.

dataset	n	m	M	M_strong	$M\_almost–strong$	$M\_semi–strong$	M_weak	$M\_extra–weak$
`strike` [27]	24	38	4	2	3	2	4	4
`karate` [28]	34	78	4	1	2	2	4	4
`Korea1` [29]	35	69	5	2	2	3	5	5
`Korea2` [29]	35	84	5	3	4	3	5	5
`sawmill` [30]	36	62	4	4	4	4	4	4
`dolphins small` [31]	40	70	6	3	6	3	6	6
`graph` [32]	60	114	7	0	2	3	7	7
`dolphins` [31]	62	159	5	2	2	3	4	5
`Les Misérables` [33, 34]	77	254	6	2	2	3	6	6
`p53 protein` [35]	104	226	7	1	2	2	6	7
`political books` [36]	105	441	5	2	2	2	4	4
percentage of communities satisfying the condition				37.93%	53.45%	51.72%	94.83%	98.28%

4 Adding cohesion constraints to modularity maximization models

Exact modularity maximization without further constraints has been studied in [14–16]. Exact methods are applied using two different formulations. The first one is a reformulation of the problem as a clique partitioning problem [37], the second one is a mixed-integer quadratic optimization problem [14, 16].

Adding cohesion constraints is more or less difficult according to the approach selected. In clique partitioning assignment of entities to communities is not explicitly considered: it only appears as a consequence of the optimal solution, so adding cohesion constraints does not seem to be easy in that framework. Contrarywise, the quadratic mixed-integer formulation already uses variables to denote assignment of vertices or of edges to communities. So, in that case cohesion constraints can be added easily.

We now describe these new constraints in more detail. Before that we recall the main elements in the model for modularity maximization of [14]. The objective function is the modularity Q of the network under study, that is the sum of modularities of its communities:

$$Q = \sum_{s=1}^{M} \left[\frac{m_s}{m} - \left(\frac{D_s}{2m} \right)^2 \right], \tag{7}$$

where m_s denotes the number of edges in community s, that is the subgraph induced by S_s, D_s denotes the sum of degrees k_i of the vertices of community s, and M is the number of communities which is not a priori known. Binary variables are then used to identify the communities to which each vertex and each edge belongs:

$$X_{rs} = \begin{cases} 1 & \text{if edge } r \text{ belongs to community } s \\ 0 & \text{otherwise} \end{cases}$$

for $r \in E$ and $s = 1, 2, \dots M$ and

$$Y_{is} = \begin{cases} 1 & \text{if vertex } v_i \text{ belongs to community } s \\ 0 & \text{otherwise} \end{cases}$$

for $v_i \in V$.

Constraints involving X and Y variables express that an edge belongs to a community only if both its end vertices belong to that community:

$$\forall r = \{v_i, v_j\} \in E, \forall s \in \{1, \dots, M\} \quad X_{rs} \leq Y_{is}$$
$$\forall r = \{v_i, v_j\} \in E, \forall s \in \{1, \dots, M\} \quad X_{rs} \leq Y_{js}.$$

In fact, due to maximization, this edge will belong to this community if the two conditions given above are satisfied (hence it is not necessary to add the constraint $X_{rs} \geq Y_{is} + Y_{js} - 1$ and $X_{rs} \geq 0$). Further constraints are presented in [14] and allow bounding the cardinality of each community from below or above, using binary variables F_s such that F_s is equal to 1 if the community s is nonempty, 0 otherwise. Finally, symmetry breaking constraints are used, as:

$$\sum_{v_i \in V | i \leq M,\ s=1,\dots,i} Y_{is} = 1, \tag{8}$$

and, following Plastria [38]:

$$\forall s \in \{3, \dots, M-1\} \forall i \in \{s, \dots, n\} \quad \sum_{j=2,\dots,i-1} \sum_{\ell \in \{1,\dots,s-1\}} Y_{j\ell} - \sum_{\ell \in \{1,\dots,s\}} Y_{i\ell} \leq i - 3. \tag{9}$$

The resulting model, derived from that presented in [14], is the following:

$$\max \quad \sum_{s=1}^{M} \left[\frac{m_s}{m} - \left(\frac{D_s}{2m} \right)^2 \right] \tag{10}$$

$$\text{s.t.} \quad \forall r = \{v_i, v_j\} \in E, \forall s \in \{1, \dots, M\} \quad X_{rs} \leq Y_{is} \tag{11}$$

$$\forall r = \{v_i, v_j\} \in E, \forall s \in \{1, \dots, M\} \quad X_{rs} \leq Y_{js} \tag{12}$$

$$\forall s \in \{1, \dots, M\} \quad m_s = \sum_{r \in E} X_{rs} \tag{13}$$

$$\forall v_i \in V \quad \sum_{s=1}^{M} Y_{is} = 1 \tag{14}$$

$$\forall s \in \{1,\dots,M\} \quad D_s = \sum_{v_i \in V} k_i Y_{is} \tag{15}$$

$$\forall s \in \{2,\dots,M\} \quad F_s \leq F_{s-1} \tag{16}$$

$$\forall s \in \{1,\dots,M\} \quad \sum_{r \in E} X_{rs} \geq F_s \tag{17}$$

$$\forall s \in \{1,\dots,M\} \quad \sum_{r \in E} X_{rs} \leq (m-1) F_s \tag{18}$$

$$\sum_{v_i \in V | i \leq M,\ s=1,\dots,i} Y_{is} = 1 \tag{19}$$

$$\forall s \in \{3,\dots,M-1\}\, \forall i \in \{s,\dots,n\} \quad \sum_{j=2,\dots,i-1} \sum_{\ell \in \{1,\dots,s-1\}} Y_{j\ell} - \sum_{\ell \in \{1,\dots,s\}} Y_{i\ell} \leq i-3 \tag{20}$$

$$\forall v_i \in V, \forall s \in \{1,\dots,M\} \quad Y_{is} \in \{0,1\} \tag{21}$$

$$\forall s \in \{1,\dots,M\} \quad F_s \in \{0,1\} \tag{22}$$

$$\forall r \in E, \forall s \in \{1,\dots,M\} \quad X_{rs} \in \mathbb{R}. \tag{23}$$

To this basic model we add cohesion constraints as follows. Notice that, for each cohesion condition, the constant 1 in the right-hand side transforms the strict inequality into a non-strict one.

- SCC:

$$\forall s \in \{1,\dots,M\}, v_i \in V \quad \sum_{v_j \in V : j \neq i} A_{ij} Y_{js} \geq Y_{is} \left(\lfloor \frac{k_i}{2} \rfloor + 1 \right). \tag{24}$$

  Indeed, it follows from the definition of SCC that:

$$\forall s \in \{1,\dots,M\}, v_i \in V \quad \sum_{v_j \in V : j \neq i} A_{ij} Y_{js} \geq k_i - \sum_{v_j \in V : j \neq i} A_{ij} Y_{js} + 1, \tag{25}$$

  which expresses the fact that the in-degree of vertex v_i is strictly greater than the out-degree, i.e., than the degree minus the in-degree. It is valid for unweighted graphs where all coefficients A_{ij} are integer. A few algebraic manipulations lead to:

$$\forall s \in \{1,\dots,M\}, v_i \in V \quad \sum_{v_j \in V : j \neq i} A_{ij} Y_{js} \geq \lfloor \frac{k_i}{2} \rfloor - (1 - Y_{is}) \lfloor \frac{k_i}{2} \rfloor + Y_{is}, \tag{26}$$

  as it is easily checked for both cases $Y_{is} = 1$ and $Y_{is} = 0$. Formula (24) follows.
- ASCC:
  The almost-strong cohesion condition can be expressed in terms of the variables Y as follows:

$$\forall s \in \{1,\dots,M\}, v_i \in V \,|\, k_i \neq 2 \quad \sum_{v_j \in V : j \neq i} A_{ij} Y_{js} \geq Y_{is} \left(\lfloor \frac{k_i}{2} \rfloor + 1 \right) \tag{27}$$

$$\forall s \in \{1,\dots,M\}, v_i \in V \,|\, k_i = 2 \quad \sum_{v_j \in V : j \neq i} A_{ij} Y_{js} \geq Y_{is}. \tag{28}$$

- SSCC:
  The semi-strong cohesion condition can be expressed in terms of the variables Y as follows:

$$\forall s,t \in \{1,\dots,M\} \,|\, s \neq t, v_i \in V \quad \sum_{j \in V : j \neq i} A_{ij} Y_{js} \geq \sum_{v_j \in V : j \neq i} A_{ij} Y_{jt} + 1 - (1 - Y_{is})(k_i + 1). \tag{29}$$

  Indeed, consider the following two cases: (i) $Y_{is} = 1$, that is vertex v_i belongs to community s. Then the left-hand side term of (29) is equal to the in-degree of v_i, the first term of the right-hand side

represents the part of the out-degree of v_i corresponding to edges with extremities in community s and $t \neq s$. As the last term disappears, the condition expresses that this partial out-degree must be strictly smaller than the in-degree of v_i. As similar conditions hold for all other communities, it is clear that such a relation holds for the community for which the partial out-degree of v_i is largest. (ii) $Y_{is} = 0$. Then the right-hand side of (29) is non-positive and the condition is verified.

- WCC:
  The weak cohesion condition can be written as follows in terms of variables X and Y:

$$\forall s \in \{1, \ldots, M\} \quad 4 \sum_{r \in E} X_{rs} \geq \sum_{v_i \in V} k_i Y_{is} + 1. \tag{30}$$

  Indeed, the sum of in-degrees for community s may be written as $2 \sum_{r \in E} X_{rs}$ (where the factor 2 is due to edges having both vertices in the community) and must be greater than the sum of out-degrees of community s, that is the sum of all the degrees minus the sum of in-degrees for vertices of that community: $\sum_{v_i \in V} k_i Y_{is} - 2 \sum_{r \in E} X_{rs}$.
- EWCC:
  The extra-weak cohesion condition was proposed in [22] without a mathematical expression. It can be written as follows:

$$\forall s, t \in \{1, \ldots, M\} \,|\, s \neq t, \quad 2 \sum_{r \in E} X_{rs} \geq \sum_{r = \{v_i, v_j\} \in E} (Y_{is} Y_{jt} + Y_{js} Y_{it}) + 1. \tag{31}$$

  The left-hand side of (31) is equal to twice the number of edges in community s. The right-hand side is equal to the number of edges with an end vertex in s and the other in community t. This expression can be linearized introducing $\forall r = \{v_i, v_j\} \in E$ non-negative variables $Z_r = Y_{is} Y_{jt}$ and $Z'_r = Y_{js} Y_{it}$:

$$\forall s, t \in \{1, \ldots, M\} \,|\, s \neq t, \quad 2 \sum_{r \in E} X_{rs} \geq \sum_{r \in E} (Z_r + Z'_r) + 1, \tag{32}$$

  and adding linearization constraints $\forall s, t \in \{1, \ldots, M\} \,|\, s \neq t$:

$$\begin{aligned}
Z_r &\leq Y_{is} && (33)\\
Z_r &\leq Y_{jt} && (34)\\
Z_r &\geq Y_{is} + Y_{jt} - 1 && (35)\\
Z'_r &\leq Y_{js} && (36)\\
Z'_r &\leq Y_{it} && (37)\\
Z'_r &\geq Y_{js} + Y_{it} - 1. && (38)
\end{aligned}$$

5 Results

We first compare the results of the simulated annealing heuristic of Wang et al. [26] with those of our exact algorithms [16]. Three problems are common to both lists of problems solved in [16] and [26]: `karate`, `dolphins`, `political books`. In [26], results are presented for direct optimization and for considering the weak cohesion condition through a constrained model. In both cases, the value of modularity Q is given together with the number of communities of the optimal partition found and the number of such communities which satisfy the weak cohesion condition. As simulated annealing is a probabilistic optimization heuristic, results depend on the seed used in its random number generator. The Q values are given with limited precision, i.e., only 2 significant digits. Comparing first the results of a direct optimization with the optimal partitions found in [16], we observe that:

- the number of communities and the proposed solutions differ from the optimal ones in several cases: `dolphins` (4 instead of 5), `political books` (4 instead of 5);
- the optimal solution is found for the smallest problem, that is `karate`.

Turning now to optimization subject to weak cohesion constraints, again the same three problems are common to the list of those solved by the algorithms of the present paper (see Table 2) and the constrained optimization one of [26]. Results are as follows. For `karate`, the number of communities found in [26] is optimal but the solution is not as its Q value is 0.40 and the optimal solution is 0.41979 with the weak condition satisfied by all communities. For `dolphins` and `political books`, the number of communities found is optimal, while the modularity value is 0.52 in [26] versus 0.526799 with the algorithm of the present paper for `dolphins`, and 0.53 versus 0.526938 for `political books` (so, up to the precision of two digits, the solution found in [26] is optimal).

Table 2: Modularity maximization with weak and extra-weak cohesion conditions. Network dimension (n = number of vertices, m = number of edges), number of clusters found (M, M_w, M_{ew}) and corresponding modularity value (Q, Q_w, Q_{ew}) for the standard modularity maximization problem and the modularity maximization problem with weak and extra-weak constraints.

network			modularity maximization		weak		extra-weak	
dataset	n	m	M	Q	M_w	Q_w	M_{ew}	Q_{ew}
`strike`	24	38	4	0.561981	4	0.561981	4	0.561981
`karate`	34	78	4	0.41979	4	0.41979	4	0.41979
`Korea1`	35	69	5	0.477736	5	0.477736	5	0.477736
`Korea2`	35	84	5	0.450822	5	0.450822	5	0.450822
`sawmill`	36	62	4	0.550078	4	0.550078	4	0.550078
`dolphins small`	40	70	4	0.620714	4	0.620714	4	0.620714
`graph`	60	114	7	0.502655	7	0.502655	7	0.502655
`dolphins`	62	159	5	0.528519	**4**	**0.526799**	5	0.528519
`Les Misérables`	77	254	6	0.560008	6	0.560008	6	0.560008
`p53 protein`	104	226	7	0.535134	**6**	**0.534488**	7	0.535134
`political books`	105	441	5	0.527237	**4**	**0.526938**	**4**	**0.526938**
average			5.090909	0.521334	4.818182	0.521092	5	0.521307

In Table 2 we report the results obtained by unconstrained modularity maximization and modularity maximization subject to the weak constraint and the extra-weak constraint respectively. It appears that the only three cases in which results obtained with the weak constraints differ from the unconstrained case are `dolphins`, `p53 protein` and `political books` (represented in bold in Table 2).

We now discuss these three cases in more detail.

- `dolphins`: there are 5 communities with a modularity of 0.528519 in the unconstrained solution while there are 4 communities with a modularity of 0.526799 in the constrained solution. Detailed memberships of both solutions are given in Tables 3 and 4. It appears that the smallest community, that is C_3, is split among three of the four other communities: dolphin 40 goes to community C_2^w, dolphins 4, 9, and 60 go to C_3^w, and dolphin 37 to C_4^w. However, these three assignments entail further changes: dolphins 21 and 45 move from C_1 to C_4^w while dolphins 54 and 62 move from C_5 to C_1^w. Thus, nine dolphins change community when the weak condition constraint is added.
- `p53 protein`: there are 7 communities with a modularity of 0.535134 in the unconstrained solution while there are 6 communities with a modularity of 0.534488 in the constrained solution. Both partitions are depicted in Fig. 3. It appears that a small cluster with five entities in blue in the center of the figure does not satisfy the weak cohesion condition. This community is split, four entities going to the cluster below it and one entity to the cluster above.
- `political books`: there are 5 communities with a modularity of 0.527237 in the unconstrained solution while there are 4 communities with a modularity of 0.526938 in the constrained solution. Both partitions are depicted in Fig. 4. It appears that the smallest community of the unconstrained solution, that is the community containing books 49, 50, and 58 depicted in orange on the right side of the figure, does not satisfy the weak cohesion condition. All of these three books move to the community above it.

Table 3: Partition obtained with the standard modularity maximization model on `dolphins` dataset.

C_1	C_2	C_3	C_4	C_5
1, 3, 11, 21 29, 31, 43, 45, 48	2, 6, 7, 8, 10 14, 18, 20, 23 26, 27, 28, 32 33, 42, 49, 55 57, 58, 61	4, 9, 37, 40 60	5, 12, 16, 19 22, 24, 25, 30 36, 46, 52, 56	13, 15, 17, 34 35, 38, 39, 41 44, 47, 50, 51 53, 54, 59, 62

Table 4: Partition obtained with the modularity maximization model and the weak cohesion constraint on `dolphins` dataset.

C_1^w	C_2^w	C_3^w	C_4^w
1, 3, 11, 29 31, 43, 48, 54 62	2, 6, 7, 8, 10 14, 18, 20, 23 26, 27, 28, 32 33, 40, 42, 49 55, 57, 58, 61	4, 5, 9, 12, 16 19, 22, 24, 25 30, 36, 46, 52 56, 60	13, 15, 17, 21 34, 35, 37, 38 39, 41, 44, 45 47, 50, 51, 53 59

It appears from Table 2 that the only case in which the modularity maximization constrained with the extra-weak constraint differs from the unconstrained solution is `political books`, for which the constrained optimal partition contains 4 communities.

The average reduction in modularity when adding extra-weak constraints is of 0.000027, i.e., it is very moderate. Reduction in modularity when imposing the weak condition is slightly larger, i.e., 0.000242.

The effect of imposing the other cohesion conditions is much larger, as shown in Table 5. For the strong and semi-strong conditions, average modularity is reduced from 0.521334 to 0.354608; for the almost-strong conditions from 0.521334 to 0.486862. We also observe drastic reductions in the average number of communities, that is reduced from 5.090909 to 2.727273 for the strong and semi-strong conditions, from 5.090909 to 3.90909 for the almost-strong condition. The effect of imposing the strong and the semi-strong conditions are the same for the considered datasets, for which the same optimal partitions are obtained. The effect of introducing almost-strong constraints instead of strong ones is less drastic, as already observed in [23].

Table 5: Modularity maximization with strong, almost-strong, and semi-strong condition. Network dimension (n = number of vertices, m = number of edges), number of clusters found (M_s, M_{as}, M_{ss}) and corresponding modularity value (Q_s, Q_{as}, Q_{ss}) for the standard modularity maximization problem and the the modularity maximization problem with strong, almost-strong and semi-strong constraints.

network			modularity maximization		strong		almost–strong		semi–strong	
dataset	n	m	M	Q	M_s	Q_s	M_{as}	Q_{as}	M_{ss}	Q_{ss}
`strike`	24	38	4	0.561981	2	0.257271	3	0.54813	2	0.257271
`karate`	34	78	4	0.41979	2	0.132807	4	0.402038	2	0.132807
`Korea1`	35	69	5	0.477736	4	0.383638	4	0.383638	4	0.383638
`Korea2`	35	84	5	0.450822	3	0.424036	4	0.432469	3	0.424036
`sawmill`	36	62	4	0.550078	4	0.550078	4	0.550078	4	0.550078
`dolphins small`	40	70	4	0.620714	3	0.573571	4	0.620714	3	0.573571
`graph`	60	114	7	0.502655	1	0	4	0.438135	1	0
`dolphins`	62	159	5	0.528519	2	0.359242	3	0.480598	2	0.359242
`Les Misérables`	77	254	6	0.560008	4	0.437868	6	0.52921	4	0.437868
`p53 protein`	104	226	7	0.535134	2	0.284204	4	0.472502	2	0.284204
`political books`	105	441	5	0.527237	3	0.497969	3	0.497969	3	0.497969
average			5.090909	0.521334	2.727273	0.354608	3.909091	0.486862	2.727273	0.354608

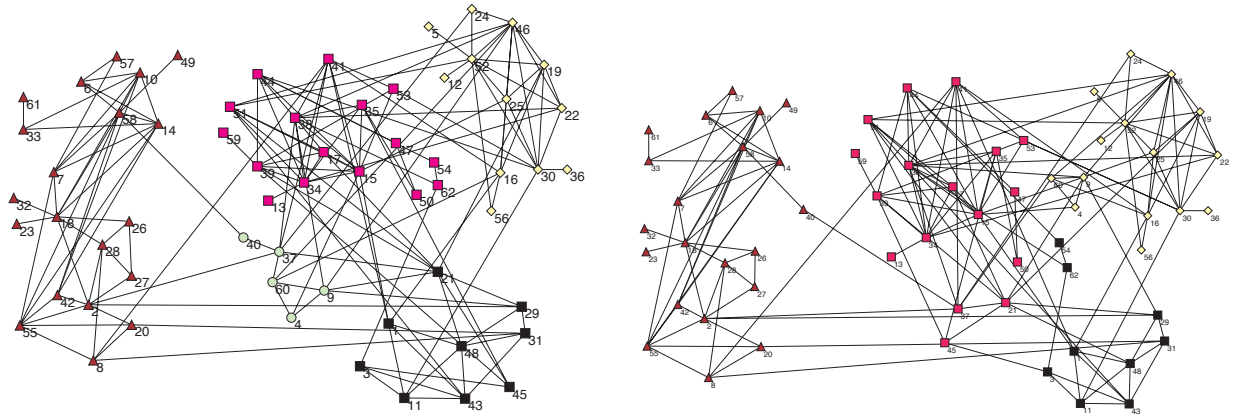

Figure 2: (Color online) Partitions obtained on the `dolphins` network using the original modularity maximization model (left) and the modularity maximization model with the weak cohesion constraint (right).

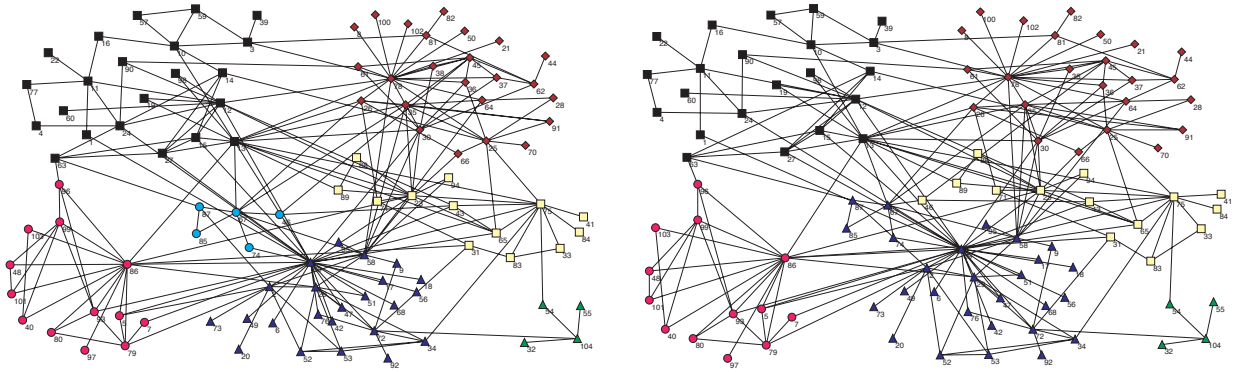

Figure 3: (Color online) Partitions obtained on the `p53` network using the original modularity maximization model (left) and the modularity maximization model with the weak cohesion constraint (right).

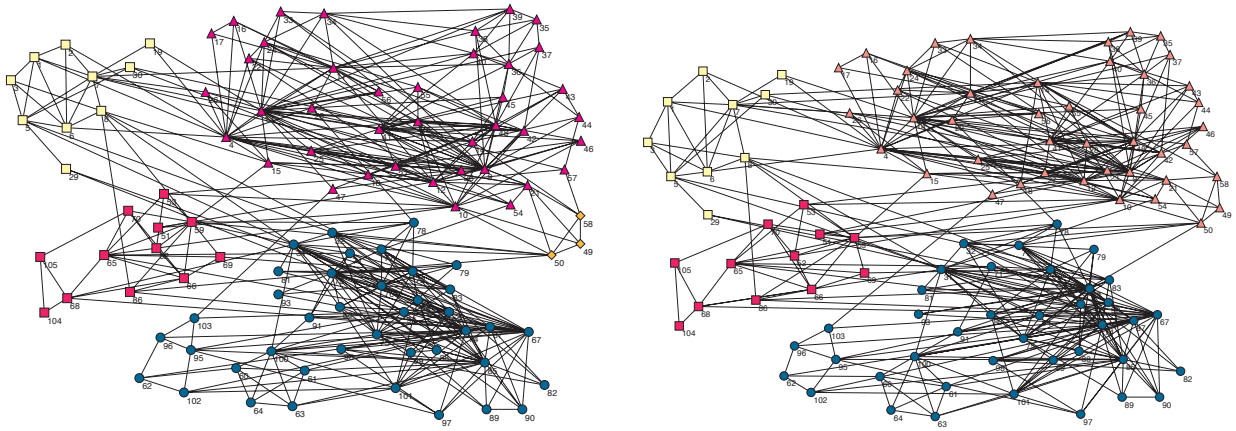

Figure 4: (Color online) Partitions obtained on the `political books` network using the original modularity maximization model (left) and the modularity maximization model with the weak cohesion constraint (right).

6 Conclusions

The effect of adding five kinds of cohesion conditions to a modularity maximization problem has been studied. After reviewing the strong, almost-strong, semi-strong, weak, and extra-weak cohesion conditions, we check whether they were satisfied by all communities in maximum modularity partitions of some known real world problems. It appears that the strong and semi-strong cohesion conditions are very strict. Indeed, all of the communities of an optimal partition satisfy these conditions for one dataset only. The almost-strong cohesion condition is slightly less strict, but still quite stringent. The weak cohesion condition is more intuitive and satisfied by all communities of 8 out of the 11 optimal partitions. The extra-weak cohesion condition appears to be very lax as there is only a single community of only one of the optimal unconstrained partitions which does not satisfy it.

We then show how these five cohesion constraints can be added, one at a time, to a quadratic convex mixed-integer optimization formulation proposed for modularity maximization. For three real world datasets, we discuss in detail the modifications in the optimal solutions due to the imposition of weak cohesion constraints.

Adding cohesion constraints to a modularity maximization model appears to be feasible and in particular for the weak condition yields intuitive and appealing results. While the present work is based on the model of Xu et al. [14], we will try to explore the addition of cohesion constraints to other models for modularity maximization.

References

[1] M. E. J. Newman, *Networks: An introduction* (Oxford University Press, Oxford, 2010).

[2] S. Fortunato, Physics Reports **486**, 75 (2010).

[3] M. Girvan and M. Newman, Proceedings of the National Academy of Sciences of the USA **99**, 7821 (2002).

[4] R. Guimerà and A. N. Amaral, Nature **433**, 895 (2005).

[5] A. Medus, G. Acuna, and C. Dorso, Physica A **358**, 593 (2005).

[6] S. Lehmann and L. Hansen, The European Physical Journal B **60**, 83 (2007).

[7] M. Tasgin, A. Herdagdelen, and H. Bingol, arXiv:0711.0491 (2007).

[8] J. Duch and A. Arenas, Physical Review E **72**, 027104 (2005).

[9] G. Agarwal and D. Kempe, The European Physical Journal B **66**, 409 (2008).

[10] A. Noack and R. Rotta, Lecture Notes in Computer Science **5526**, 257 (2009).

[11] P. Schuetz and A. Caflisch, Physical Review E **77**, 046112 (2008).

[12] S. Cafieri, P. Hansen, and L. Liberti, Physical Review E **83**, 056105 (2011).

[13] S. Cafieri, A. Costa, and P. Hansen, Annals of Operations Research (2012), http://dx.doi.org/10.1007/s10479-012-1286-z.

[14] G. Xu, S. Tsoka, and L. Papageorgiou, The European Physical Journal B **60**, 231 (2007).

[15] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner, IEEE Transactions on Knowledge and Data Engineering **20**, 172 (2008).

[16] D. Aloise, S. Cafieri, G. Caporossi, P. Hansen, S. Perron, and L. Liberti, Physical Review E **82**, 046112 (2010).

[17] S. Fortunato and M. Barthelemy, Proceedings of the National Academy of Sciences of the USA **104**, 36 (2007).

[18] C. P. Massen and J. P. K. Doye, Physical Review E **71**, 046101 (2005).

[19] S. Cafieri, P. Hansen, and L. Liberti, Physical Review E **81**, 046102 (2010).

[20] Z. Li, S. Zhang, R.-S. Wang, X.-S. Zhang, and L. Chen, Physical Review E **77**, 036109 (2008).

[21] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, Proceedings of the National Academy of Sciences of the USA **101**, 2658 (2004).

[22] Y. Hu, H. Chen, P. Zhang, M. Li, Z. Di, and Y. Fan, Physical Review E **78**, 026121 (2008).

[23] S. Cafieri, G. Caporossi, P. Hansen, S. Perron, and A. Costa, Physical Review E **85**, 046113 (2012).

[24] X.-S. Zhang, R. S. Wang, Y. Wang, J. Wang, Y. Qiu, L. Wang, and L. Chen, Europhysics Letters **87**, 38002 (2009).

[25] X.-S. Zhang, Z. Li, R.-S. Wang, and Y. Wang, Journal of Combinatorial Optimization **23**, 425 (2012).

[26] J.-G. Wang, L. Wang, Y.-Q. Qiu, Y. Wang, and X.-S. Zhang, Lecture Notes in Operation Research, The Third International Symposium on Optimization and Systems Biology (OSB09), 142 (2009).

[27] J. H. Michael, Forest Products Journal **47**, 41 (1997).

[28] W. Zachary, Journal of Anthropological Research **33**, 452 (1977).

[29] E. Rogers and D. Kincaid, *Communication networks: toward a new paradigm for research* (Free Press, 1981).

[30] J. H. Michael and J. G. Massey, Forest Products Journal **47**, 25 (1997).

[31] D. Lusseau, K. Schneider, O. Boisseau, P. Haase, E. Slooten, and S. Dawson, Behavioral Ecology and Sociobiology **54**, 396 (2003).

[32] http://vlado.fmf.uni-lj.si/pub/networks/data/DIC/TG/glossTG.pdf.

[33] D. Knuth, *The Stanford GraphBase: A Platform for Combinatorial Computing* (Addison-Wesley, Reading, MA, 1993).

[34] V. Hugo, *Les Misérables* (Gallimard, Bibliotheque de la Pleiade, Paris, 1951).

[35] L. Dartnell, E. Simeonidis, M. Hubank, S. Tsoka, I. Bogle, and L. Papageorgiou, FEBS Letters **579**, 3037 (2005).

[36] V. Krebs, http://www.orgnet.com/ (unpublished).

[37] M. Grötschel and Y. Wakabayashi, Mathematical Programming **45**, 59 (1989).

[38] F. Plastria, European Journal of Operational Research **140**, 338 (2002).