Graph Clustering by Maximizing Statistical Association Measures

Julien Ah-Pine

Graph Clustering by Maximizing Statistical Association Measures

Julien Ah-Pine

University of Lyon, ERIC Lab
5, avenue Pierre Mendès France
69676 Bron Cedex, France
julien.ah-pine@eric.univ-lyon2.fr

Abstract. We are interested in objective functions for clustering undirected and unweighted graphs. Our goal is to define alternatives to the popular modularity measure. To this end, we propose to adapt statistical association coefficients, which traditionally measure the proximity between partitions, for graph clustering. Our approach relies on the representation of statistical association measures in a relational formulation which uses the adjacency matrices of the equivalence relations underlying the partitions. We show that graph clustering can then be solved by fitting the graph with an equivalence relation *via* the maximization of a statistical association coefficient. We underline the connections between the proposed framework and the modularity model. Our theoretical work comes with an empirical study on computer-generated graphs. Our results show that the proposed methods can recover the community structure of a graph similarly or better than the modularity.

Keywords: Graph clustering, Community detection, Statistical association measures, Modularity.

1 Introduction

Many real-world problems can be designed using graphs where entities of the studied system are represented as nodes and their relationships as edges between nodes. In many domains such as biology, ecology, social network analysis ..., graph theory tools are employed as means for representing complex systems. In this context, graph clustering consists in partitioning nodes into groups such that vertices belonging to the same group are better interconnected to each other than to vertices outside of the group. Discovering such clusters can lead to new and important insights. In biology for example, clustering a protein-protein interaction network helps to find proteins with the same biological function. Another example is in social network analysis, where graph clustering leads to the detection of community structures [1]. Such knowledge can help to better understand the social system and its related phenomenons.

There exist many graphs clustering techniques. We particularly focus on methods that optimize an objective function. The benefit criterion aims at

reflecting the quality of a clustering. In this context, density-based objective functions are well-known approaches. In these cases, clusters are defined as subgraphs with high densities of edges. The modularity measure proposed by Newman and Girvan in [2] is a popular density-based objective function. It assumes that two nodes belong to the same community if the number of edges between them is greater than the expected number of edges under a null random model.

We address the graph clustering task from a viewpoint different from the one underlying the modularity. We suppose that an undirected and unweighted graph can be seen as a perturbed equivalence relation and finding groups of nodes can be interpreted as fitting the graph with a partition. To this end, we need to quantify the proximity between two partitions. In the statistical literature there are numerous coefficients addressing this exact problem. These criteria are known as statistical association measures (SAM) between categorical variables or partitions. Our proposal is thus to fit a given graph with a partition by maximizing a SAM. However, using such measures in this context is not straightforward. Indeed, these coefficients are typically defined by using contingency tables over the set of categories of the two partitions. Yet, the contingency table between a given graph (which is not an equivalence relation) and a partition does not exist. To overcome this drawback, we review the research works of Marcotorchino who showed in [3, 4], that many SAM can be equivalently expressed through the adjacency matrices of the equivalence relations underlying the categorical variables. Based on this approach, we show how we can convert SAM to define new density-based quality functions for graph clustering.

In section 2, we recall some density-based objective functions for graph clustering. We particularly emphasize the modularity concept. Then, in section 3, we introduce our framework. We recall SAM both in their contingency and their relational formulations. Then we show how these measures can lead to graph clustering methods. Moreover, we study the relationships between the modularity and SAM. Next, in section 4, we empirically examine the behaviors of the proposed objective functions on artificial graphs and we compare their results with the ones provided by the modularity. We finally conclude and sketch some future works in section 5.

2 Related work : modularity optimization

There are several types of density-based benefit functions for graph clustering [5, 6]. One first family is based on graph cuts measures which iteratively split the set of nodes of a graph into two, providing that the density of edges between the two clusters is low. To apply such methods, one can generally use any max-flow/min-cut algorithm such as the Ford-Fulkerson one. Another method is spectral clustering which computes the Fiedler eigenvector of the Laplacian of the graph. Edges cuts criteria and the aforementioned algorithms are particularly used to tackle graph partitioning problems. These tasks are slightly distinct from graph clustering problems. In graph partitioning, the number of clusters and their sizes are known and one has to recover the correct partition given

these pieces of information. In contrast, in graph clustering, we do not assume any information about the number nor the shape of the communities.

In order to better deal with the graph clustering task, Newman and Girvan proposed the modularity concept [2]. Their approach has the advantage to better formalize the concept of community and to avoid setting the number of clusters manually. This measure is denoted by Q and it can be expressed as follows [7] : $Q =$ Number of edges within communities $-$ Expected number of such edges.

More formally, let us assume that we are given a graph with n vertices and m edges. Its adjacency matrix is denoted by A. Let us denote by P the pairwise matrix of general term P_{ij} which is the expected number of edges between nodes i and j. Since we are concerned with undirected and unweighted graph, P_{ij} can be interpreted as the probability to have an edge between i and j. The modularity can be formulated by the equation below [7] :

$$Q(A,\delta) = \frac{1}{2m}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(A_{ij} - P_{ij}\right)\delta(g_i,g_j) \ . \tag{1}$$

where g_i is the cluster of i and $\delta(g_i,g_j) = 1$ if $g_i = g_j$ and 0 otherwise.

From this general formulation, Newman adopted different assumptions which led to the definition of a specific null random model. His hypothesis are the following ones [7] : (i) since the graph is undirected then P should satisfy the relation, $\forall i,j : P_{ij} = P_{ji}$; (ii) Q should be null when all vertices are in a single group and thus[1] $\sum_{i,j} A_{ij} = \sum_{i,j} P_{ij} = 2m$; (iii) the degrees distribution of the random model should be approximately the same as the one of the given graph which leads to the following constraint, $\forall i : \sum_j P_{ij} = k_i$ where $k_i = \sum_j A_{ij}$ is the observed degree of node i; (iv) edges should be placed at random meaning that the probability of observing an edge between i and j should be independent from the probability of observing an edge involving i and the probability of observing and edge involving j. Under these assumptions, the simplest null random model is when $\forall i,j : P_{ij} = k_i k_j/2m$ [7]. Accordingly, the following modularity formulation is the one which is commonly used in the literature :

$$Q(A,\delta) = \frac{1}{2m}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(g_i,g_j) \ . \tag{2}$$

It is worthwhile to mention that apart from (2), other coefficients relying on the modularity concept could be designed from (1). In that perspective, Newman suggested that the assumptions (i) and (ii) are fundamental and should be considered as axioms of the modularity framework unlike (iii) and (iv) [7].

Adopting (2), one can optimally solve the graph clustering problem *via* modularity maximization with the following integer linear program (see for e.g. [8]) :

$$\max_{Y} \frac{1}{2m}\sum_{i,j}\left(A_{ij} - \frac{k_i k_j}{2m}\right)(1 - Y_{ij}) \text{ subject to : } \begin{cases} Y_{ik} \leq Y_{ij} + Y_{jk} & \forall i,j,k \\ Y_{ij} \in \{0,1\} & \forall i,j \end{cases} . \tag{3}$$

[1] In order to lighten the notations we write $\sum_{i,j}$ as a shortcut for $\sum_{i=1}^{n}\sum_{j=1}^{n}$.

where $Y_{ij} = 1$ if i and j are not in the same cluster and 0 otherwise.

However, optimizing the modularity (and any other objective functions) over the set of possible partitions is an NP-hard problem. As a result, many research works have been devoted to approximately maximize the modularity with different strategies and heuristics [5, 6].

The application of modularity to the graph clustering task has demonstrated good performances both on artificial and real-world networks. However, some recent works have highlighted certain limits of this method [9]. In particular, optimizing the modularity tends to split large groups while small communities below a certain threshold are not correctly detected.

In the next section, we introduce new quality functions for graph clustering which provide alternatives to the modularity criterion given in (2).

3 The proposed approach : statistical association measures (SAM) optimization

Density-based techniques typically rely on the definition of a community and use heuristics to discover sub-graphs satisfying this definition. From our viewpoint, since clustering a given graph consists in detecting a hidden community structure, we can interpret the graph as an equivalence relation perturbed by noise. Thereby, we argue that graph clustering can be thought of as recovering the real community structure and this can be achieved by fitting the graph with a partition. This approach assumes there is a way to assess the proximity between the graph and a partition. In what follows, we introduce some statistical association measures which aim at measuring the similarity between two partitions by means of contingency tables. Then, we recall the relational formulation of these coefficients due to Marcotorchino. Using the latter expressions of SAM, we show how to use these coefficients as benefit functions for the graph clustering task. In that perspective, we underline some theoretical links between SAM and the modularity concept in order to bring into light some conceptual similarities between these two frameworks in the context of graph clustering.

3.1 SAM and their relational representation

Let us assume two categorical variables V^k and V^l with respectively p_k and p_l categories. Note that a categorical variable infers a set of disjoint groups of items which in turn can be interpreted as a partition or a clustering or an equivalence relation[2]. In categorical data analysis, in order to analyse the relationship between two categorical variables, we use the contingency table of dimensions $(p_k \times p_l)$ denoted by $\mathbf{N}$ whose general term is defined by : $\mathbf{N}_{uv}$ = Number of items belonging to both category u of V^k and category v of V^l.

Then, a core concept in categorical data analysis is the deviation from the statistical independence situation which occurs when for all pairs of categories

[2] Therefore, we will use these different terms interchangeably.

(u, v), the probability of jointly observing u and v equals the product between the probability of observing u and the probability of observing v. Using $\mathbf{N}$, this principle translates into the following formula[3] : $\forall(u, v) : \mathbf{N}_{uv}/n = (\mathbf{N}_{u.}\mathbf{N}_{.v})/n^2$ where $\mathbf{N}_{u.} = \sum_v \mathbf{N}_{uv}$. In this context, the greater the difference between $\mathbf{N}_{uv}$ and $(\mathbf{N}_{u.}\mathbf{N}_{.v})/n$ (for all pairs (u, v)), the stronger the relationship between the categorical variables.

Accordingly, we propose to study the following coefficients :

$$B(V^k, V^l) = \sum_{u=1}^{p_k}\sum_{v=1}^{p_l}\left(\mathbf{N}_{uv} - \frac{\mathbf{N}_{u.}\mathbf{N}_{.v}}{n}\right)^2 \quad . \tag{4}$$

$$E(V^k, V^l) = \sum_{u=1}^{p_k}\sum_{v=1}^{p_l}\left(\mathbf{N}_{uv}^2 - \frac{\mathbf{N}_{u.}^2\mathbf{N}_{.v}^2}{n^2}\right) \quad . \tag{5}$$

$$J(V^k, V^l) = \frac{1}{n}\sum_{u=1}^{p_k}\sum_{v=1}^{p_l}\left(\mathbf{N}_{uv}\left(\mathbf{N}_{uv} - \frac{\mathbf{N}_{u.}\mathbf{N}_{.v}}{n}\right)\right) \quad . \tag{6}$$

$$LM(V^k, V^l) = \sum_{u=1}^{p_k}\sum_{v=1}^{p_l}\frac{\mathbf{N}_{uv}^2}{\mathbf{N}_{u.}} - \frac{1}{n}\sum_{v=1}^{p_l}\mathbf{N}_{.v}^2 \quad . \tag{7}$$

The SAM B, E, J and LM are respectively the Belson [10], Marcotorchino's square independence deviation [3], the Jordan[4] [11] and the Light-Margolin [12] criteria. They are all null in case of statistical independence. B and LM can only have positive values while E and J can be either positive or negative [3]. Given V^k, these coefficients achieve their maxima when V^l is exactly the same partition as V^k [3].

The contingency representation is the usual way to introduce SAM. However, there exists an equivalent representation of these coefficients which emphasizes the relational nature of categorical variables. Indeed, as we mentioned beforehand, categorical variables are equivalence relations and such algebraic structures can be represented by graphs. This point of view was adopted by Marcotorchino and enabled him to formulate SAM with adjacency matrices[5] [3, 4]. Let us denote by C^k the adjacency matrix[6] associated to V^k. Its general term is defined by $C_{ij}^k = 1$ if i and j belong to the same category and 0 otherwise. Marcotorchino provided correspondence formulas between the contingency table $\mathbf{N}$ on the one hand and the relational representations C^k and C^l on the other hand [3, 4]. Here are some of these transformation formulas : (i) $\sum_{u=1}^{p_k}\sum_{v=1}^{p_l}\mathbf{N}_{uv}^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}C_{ij}^kC_{ij}^l$; (ii) $\sum_u \mathbf{N}_{u.}^2 = \sum_{i,j}C_{ij}^k$; (iii) $\sum_{u,v}\mathbf{N}_{uv}\mathbf{N}_{u.}\mathbf{N}_{.v} = \sum_{i,j}((C_{i.}^k +$

[3] In order to lighten the notations we write $\sum_u$ as a shortcut for $\sum_{u=1}^{p_k}$.

[4] It is actually an interpretation of Jordan's measure given by Marcotorchino in [3].

[5] The study of association and aggregation of binary relations using graph theory and mathematical programming led to the Relational Analysis method developed by Marcotorchino and which has many applications in statistics, data-mining and multiple-criteria decision making (see for e.g. [13] and references therein).

[6] Also called relational matrix in the Relational Analysis framework.

$C^k_{.j})/2)C^l_{ij}$; (iv) $\sum_{u,v}(\mathbf{N}^2_{uv}/\mathbf{N}_{u.}) = \sum_{i,j}(2C^k_{ij}/(C^k_{i.} + C^k_{.j}))C^l_{ij}$ where $C^k_{i.} = \sum_j C^k_{ij}$ and $C^k_{i.} = C^k_{.i}$ since C^k is symmetric.

Applying these correspondence formulas enables the following expressions of SAM in terms of C^k and C^l :

$$B(C^k, C^l) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(C^k_{ij} - \frac{C^k_{i.}}{n} - \frac{C^k_{.j}}{n} + \frac{C^k_{..}}{n^2}\right)C^l_{ij} \ . \tag{8}$$

$$E(C^k, C^l) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(C^k_{ij} - \frac{C^k_{..}}{n^2}\right)C^l_{ij} \ . \tag{9}$$

$$J(C^k, C^l) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(C^k_{ij} - \frac{1}{2}\left(\frac{C^k_{i.}}{n} + \frac{C^k_{.j}}{n}\right)\right)C^l_{ij} \ . \tag{10}$$

$$LM(C^k, C^l) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{2C^k_{ij}}{C^k_{i.} + C^k_{.j}} - \frac{1}{n}\right)C^l_{ij} \ . \tag{11}$$

It is noteworthy that the different formulations of the statistical independence deviation with contingency tables in (4), (5), (6) and (7), translate into different types of deviation concepts in the relational representation (8), (9), (10) and (11). Such properties were examined in [14] and led to the formalization of the central tendency deviation principle in cluster analysis. Indeed, one can observe the following central tendencies : in (9) $C^k_{..}/n^2$ is the mean average over all the terms of C^k; in (10) $(C^k_{i.} + C^k_{.j})/2n$ is the arithmetic mean of $C^k_{i.}/n$ and $C^k_{.j}/n$ and in (11) $1/n$ is the mean average over all terms of the matrix of general term $2C^k_{ij}/(C^k_{i.} + C^k_{.j})$ (which is equivalent to $C^k_{ij}/C^k_{i.}$). Regarding (8), the central tendency concept is of geometrical nature. Since C^k is a dot product matrix (or Gram matrix) the transformation of C^k_{ij} into $C^k_{ij} - C^k_{i.}/n - C^k_{.j}/n + C^k_{..}/n^2$ is known as the double centering (or Torgerson) transformation. This operation results in dots products between vectors centered with respect to the mean vector.

Now that we have provided the expression of SAM using the graph relations underlying partitions, we show in the next paragraph how to employ such criteria for clustering graphs.

3.2 Graph clustering by maximizing SAM

We interpret a given undirected and unweighted graph as a perturbed equivalence relation and clustering the graph can be seen as attempting to recover the real partition. To solve the graph clustering task, we thus propose to fit the graph encoded by its adjacency matrix A with an equivalence relation by maximizing one of the SAM introduced previously. In other words, we want to find the partition that is the most similar to A according to a given SAM. More formally, we introduce the following benefit functions for clustering graphs :

$$B(A, X) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(A_{ij} - \left(\frac{A_{i.}}{n} + \frac{A_{.j}}{n} - \frac{A_{..}}{n^2}\right)\right)X_{ij} \ . \tag{12}$$

$$E(A,X) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(A_{ij} - \frac{A_{..}}{n^2}\right)X_{ij} \ . \tag{13}$$

$$J(A,X) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(A_{ij} - \frac{1}{2}\left(\frac{A_{i.}}{n} + \frac{A_{.j}}{n}\right)\right)X_{ij} \ . \tag{14}$$

$$LM(A,X) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{2A_{ij}}{A_{i.}+A_{.j}} - \frac{1}{n}\right)X_{ij} \ . \tag{15}$$

where X is the adjacency matrix of the partition we want to recover and whose general term is $X_{ij} = 1$ if nodes i and j are in the same cluster and 0 otherwise.

X represents an equivalence relation which, from an algebraic standpoint, is a binary relation with the following properties : (i) reflexivity ($X_{ii} = 1, \forall i$); (ii) symmetry ($X_{ij} = 1 \Leftrightarrow X_{ji} = 1, \forall i,j$) and (iii) transitivity ($X_{ij} = 1 \wedge X_{jk} = 1 \Rightarrow X_{ik} = 1, \forall i,j,k$). Marcotorchino and Michaud showed that these relational properties can be formulated as linear constraints through X [15] and this finding allowed them to model the clustering problem as an integer linear program :

$$\max_X \Delta(A,X) \text{ subject to : } \begin{cases} X_{ii} = 1 & \forall i \\ X_{ij} - X_{ji} = 0 & \forall i,j \\ X_{ij} + X_{jk} - X_{ik} \leq 1 & \forall i,j,k \\ X_{ij} \in \{0,1\} & \forall i,j \end{cases} \ . \tag{16}$$

where, in our case, $\Delta(A,X)$ is either (12) or (13) or (14) or (15).

It is important to mention that this integer linear program allowed Marcotorchino to design the maximal association model for clustering data described by categorical variables. In Marcotorchino's works, A was considered as an equivalence relation [15] or the sum over several equivalence relations [16]. Our proposal can thus be understood as the extension of the maximal association framework to graph clustering by considering A to be a general adjacency matrix without any particular property (except being undirected).

Before moving to the section dedicated to the experiments, we establish some interesting relationships between the modularity framework and our proposal based on SAM.

3.3 Some relationships between modularity and SAM

Firstly, using the notations introduced previously, the standard modularity defined in (2) can be reformulated as below :

$$Q(A,X) = \frac{1}{A_{..}}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(A_{ij} - \frac{A_{i.}A_{.j}}{A_{..}}\right)X_{ij} \ . \tag{17}$$

In addition to the correspondence formulas given in paragraph 3.1, let us introduce the following identity : (v) $\sum_v(\sum_u \mathbf{N}_{u.}\mathbf{N}_{uv})^2 = \sum_{i,j} C^k_{i.}C^k_{.j}C^l_{ij}$ [4]. From this equation, if we identify C^k and C^l to A and X respectively and if we assume

A and X to be partitions associated to two categorical variables V^k and V^l, we can easily show that the modularity given in (17) can be expressed by means of a contingency table as follows :

$$Q(V^k, V^l) = \frac{1}{\sum_{u=1}^{p_k} \mathbf{N}_{u.}^2} \left(\sum_{u=1}^{p_k} \sum_{v=1}^{p_l} \mathbf{N}_{uv}^2 - \frac{1}{\sum_{u=1}^{p_k} \mathbf{N}_{u.}^2} \left(\sum_{v=1}^{p_l} \left(\sum_{u=1}^{p_k} \mathbf{N}_{u.} \mathbf{N}_{uv} \right)^2 \right) \right) . \tag{18}$$

Then, one can easily check that $Q(V^k, V^l)$ is null in case of statistical independence between V^k and V^l. This outcome shows the potential application of the modularity measure in categorical data analysis.

Let us now place the SAM in the context of the modularity concept developed by Newman. Let us formally introduce the following central tendencies : $\mu_{ij}^Q = A_{i.}A_{.j}/A_{..}$; $\mu_{ij}^B = A_{i.}/n + A_{.j}/n - A_{..}/n^2$; $\mu_{ij}^E = A_{..}/n^2$; $\mu_{ij}^J = A_{i.}/(2n) + A_{.j}/(2n)$ and $\mu_{ij}^{LM} = 1/n$. In that case, (17), (12), (13), (14) and (15) can all be reformulated as : $\alpha \sum_{i=1}^n \sum_{j=1}^n (A_{ij} - \mu_{ij}^Z)$ with $Z \in \{Q, B, E, J, LM\}$, $\alpha = 1$ when $Z \in \{B, E, LM\}$, $\alpha = 1/A_{..}$ when $Z = Q$, $\alpha = 1/n$ when $Z = J$ and by substituting A_{ij} with $\widehat{A}_{ij} = 2A_{ij}/(A_{i.} + A_{.j})$ when $Z = LM$. This expression of SAM better underlines the connections between the general modularity framework given in (1) and cluster analysis methods based on the central tendency deviation principle [14]. Furthermore, one can easily check that Newman's axioms we recalled in section 2 are both satisfied by all SAM under study except the LM method : (i) $\forall i, j : \mu_{ij}^Z = \mu_{ji}^Z$ for $Z \in \{B, E, J, LM\}$; (ii) $\sum_{i,j} A_{ij} = \sum_{i,j} \mu_{ij}^Z$ for $Z \in \{B, E, J\}$. As a result, B, E and J fit in the modularity model. However, hypothesis (iii) and (iv) are not satisfied by any SAM under study except B for which we have (iii) $\forall i : \sum_j \mu_{ij}^B = k_i = A_{i.}$.

In such a context, it is also interesting to notice that the SAM E given in (13) corresponds to another suggested modularity model which assumes a Bernoulli distribution for P_{ij} in (1) and which boils down to the following constant[7], $P_{ij} = A_{..}/n^2$, $\forall i, j$ [7].

After having introduced the proposed objective functions and some properties about the relationships between the modularity concept and SAM, we examine in the next section if our proposals lead to interesting graph clustering methods from an empirical standpoint. Another goal of these experiments is to enable us to initiate a comparison between the modularity framework and SAM based optimization with regard to the hypothesis underlying each method.

4 Experiments

Our experiments are based on computer-generated graphs of different sizes. We give below the details about the algorithm we used to maximize the different density-based objective functions presented previously. We explain the tool we

[7] Note that in the case of E, we assume that the graph is reflexive unlike Q. In the latter case, the constant is $A_{..}/(n(n-1))$.

employed and the parameters we set to generate the artificial graphs. We then analyse the quality of the graph clustering results obtained with the different techniques.

4.1 Greedy optimization by agglomerative hierarchical clustering

The optimization problems given in (16) and (3) are NP-hard[8], and thus, numerous heuristics attempting to provide sub-optimal solutions have been proposed (see for e.g. the surveys [5, 6, 1]). In our experiments, we used the greedy optimization algorithm proposed by Newman in [17] in order to maximize the modularity criterion given in (2).

This heuristic is based on a simple agglomerative hierarchical clustering strategy. It starts with n distinct clusters and at each iteration it merges the two clusters that allow the best improvement of the modularity value. The merging process goes on until there is no pair of clusters whose fusion enables increasing the modularity value. This heuristic has the advantage to avoid fixing the number of clusters as a parameter.

This algorithm can be applied to other kinds of quality measures and in order to provide a fair comparison between the different objective functions, we thus used this technique to maximize (12), (13), (14) and (15) as well.

4.2 LFR benchmark graphs

The computer-generated graphs we analyzed in our experiments rely on the LFR benchmarks proposed by Lancichinetti, Fortunato and Radicchi in [18, 19]. These benchmarks aim at providing the research community with graphs whose properties reflect real-world cases. Indeed, observed complex networks are characterized by heterogeneous distributions both for node degrees and cluster sizes. As a consequence, the authors developed a model that generates graphs which satisfy these features. They also implemented a freely available tool[9] that we used to conduct our empirical work.

Their approach is based on the planted l-partition model in which node degrees follow a power law distribution with exponent τ_1 and clusters size a power law distribution with exponent τ_2. Overall the parameters of their model are : (i) n the number of nodes; (ii) the average degree of nodes; (iii) the maximum degree of nodes; (iv) τ_1; (v) τ_2 and (vi) $\mu \in [0, 1]$ the mixing parameter. The latter parameter μ is the one that allows gradually monitoring the presence or the absence of a community structure in the graph. It represents the percentage of edges that a node shares with vertices that do not belong to its group. Therefore, as μ grows, the community structure progressively degrades and the limit case

[8] Note that the constraints in (16) and in (3) are equivalent : the former models the properties of an equivalence relation while the latter models the properties of the complementary of an equivalence relation which is a distance relation satisfying the triangle inequality constraint.

[9] `http://santo.fortunato.googlepages.com/inthepress2`

$\mu = 1$ corresponds to the situation where edges are totally placed randomly. Typically, we assume that there is a strong community structure within the graph as long as $\mu < 0.5$.

4.3 Experiments settings and results

We studied graphs of different sizes : 500, 1000 and 2000 nodes. We used the same parameters values as in [19] : the average degree was set to 20; the maximum degree to 50; $\tau_1 = 2$; $\tau_2 = 1$ and we vary μ from 0 to 0.7 with a 0.1 step.

The LFR benchmark graphs also provide a built-in community structure which allowed us to compare the clustering results with the real partition. To assess the proximity between the found partition and the ground-truth, the normalized mutual information (NMI) measure was used. Let us denote by V^k the clustering output of our algorithm with p_k clusters and by V^l the real clustering with p_l groups. Let $P(V^k = u, V^l = v) = P(u, v) = \mathbf{N}_{uv}/n$ be the probability of jointly observing u and v and let $P(u) = \mathbf{N}_{u.}/n$ and $P(v) = \mathbf{N}_{.v}/n$ be the probability of observing u and v respectively. The mutual information measure between V^k and V^l denoted by $I(V^k, V^l)$ is defined as follows : $I(V^k, V^l) = \sum_{u=1}^{p_k} \sum_{v=1}^{p_l} P(u, v) \log(P(u, v)/(P(u)P(v)))$. Its normalized version denoted by $NMI(V^k, V^l)$ is then given by $NMI(V^k, V^l) = 2I(V^k, V^l)/(H(V^k) + H(V^l))$ where $H(V^k) = -\sum_u P(u) \log(P(u))$ is the entropy of V^k. The NMI measure ranges from 0 and 1. It equals 1 when $V^k = V^l$ and it is null when V^k and V^l are statistically independent. Note that we used the NMI coefficient to assess our clustering models because this measure is often used in the graph clustering literature (see for e.g. [19]). In that way, we can also position our contributions with respect to other papers and graph clustering techniques.

To have a better estimation of the performances, we generated 5 different graphs for each distinct parameter setting and we took the median value. The experimental results obtained for NMI measures are shown in the first row of Fig. 1. We also computed the number of clusters found by each method in order to examine if the different techniques are able to recover the right number of clusters. In this case, we also took the median over the 5 trials. These results are shown in the second row of Fig. 1.

The first row of Figure 1 allows us to compare the quality of the clustering outputs found by the different methods. We claim the following outcomes : (i) as expected, all the methods have their quality diminishing as μ grows; (ii) the LM coefficient dominates the modularity Q and other methods and this superiority seems to grow with the size of the graphs; (ii) B and J perform similarly than Q whatever the size of the graphs; (iii) E is the less good approach and it particularly performs the worst when $\mu \in [0.3, 0.7]$.

Concerning the number of clusters found, we can make the following observations from the second row of Fig. 1 : (i) the number of real communities provided by the LFR benchmarks is stable and varies around 20, 40 and 80 for graphs of size 500, 1000, 2000 respectively; (ii) except for E with graphs of size 500, all the techniques produce less clusters than the correct number of groups; (iii) as μ grows the number of clusters decreases for all the methods and beyond a certain

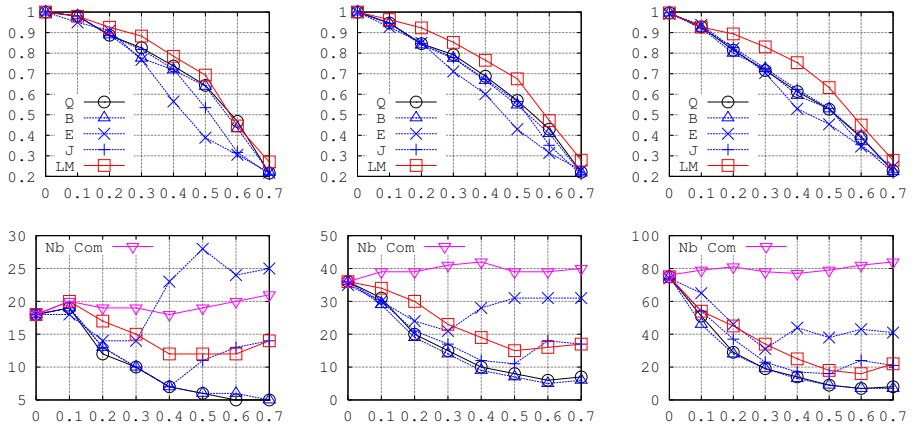

Fig. 1. First row : NMI values (vertical axis) *versus* mixing parameters μ (horizontal axis). Second row : Number of clusters found (vertical axis) *versus* mixing parameters μ (horizontal axis). We show the curves for each objective function. From left to right, the plots correspond to graphs with 500, 1000 and 2000 nodes respectively.

point (approximately 0.6 for Q, B, LM; 0.5 for J and 0.3 for E), it tends to either grow or stabilize; (iv) the LM criterion tends to produce more clusters than Q, B and J.

Overall, B and J are comparable to Q while LM is clearly a better objective function than the other ones. One reason that could explain the superiority of LM is the fact that it implicitly transforms the binary matrix A into a non negative one, $\widehat{A}$ whose general term is $\widehat{A}_{ij} = 2A_{ij}/(A_{i.} + A_{.j})$. Then its related central tendency scheme, $\mu_{ij}^{LM} = 1/n$, gives the same value for all pairs of nodes (i, j). Such an approach is indeed different from the other quality functions we examined, since they all keep the binary matrix but what changes from one function to the other is the underlying central tendency scheme μ^Z, $Z \in \{Q, B, J, E\}$ as we underlined in paragraph 3.3.

Moreover, our experimental results invite us to further analyze the hypothesis underlying the different objective functions. Regarding Newman's assumptions for the modularity given in (2), it is interesting to notice that B and J perform similarly than Q despite the fact they do not satisfy all the hypothesis assumed by the latter criterion. More importantly, LM violates most of Newman's hypothesis but outperforms all other methods including Q.

5 Conclusion

We have proposed new objective functions for clustering undirected and unweighted graphs. Our method consists in maximizing SAM represented in their relational representation, in order to fit the given graph with a partition. Our empirical study on artificial graphs has shown encouraging results. Most of the

proposed SAM perform equivalently or better than the modularity criterion. In particular, the Light-Margolin coefficient dominates the latter approach. As for future work, we plan to develop the analysis provided in paragraph 3.3 and leverage the empirical results presented previously by further comparing the modularity concept and SAM from a theoretical viewpoint. We also plan to extend our experiments on larger graphs both for computer-generated cases and for real-world networks in order to further validate our findings.

References

1. Papadopoulos, S., Kompatsiaris, Y., Vakali, A., Spyridonos, P.: Community detection in social media. Data Min. Knowl. Discov. **24**(3) (May 2012) 515–554
2. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**(2) (February 2004) 026113
3. Marcotorchino, J.F.: Utilisation des comparaisons par paires en statistique des contingences partie I (1984) Technical Report F069, IBM.
4. Marcotorchino, J.F.: Utilisation des comparaisons par paires en statistique des contingences partie II (1984) Technical Report F071, IBM.
5. Schaeffer, S.E.: Graph clustering. Computer Science Review **1**(1) (2007) 27 – 64
6. Fortunato, S.: Community detection in graphs. Physics Reports **486**(35) (2010) 75 – 174
7. Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Physical Review E **74**(3) (2006) 036104
8. Agarwal, G., Kempe, D.: Modularity-maximizing graph communities via mathematical programming. The European Physical Journal B **66**(3) (2008) 409–418
9. Lancichinetti, A., Fortunato, S.: Limits of modularity maximization in community detection. Phys Rev E Stat Nonlin Soft Matter Phys **84**(6 Pt 2) (2011) 066122
10. Belson, W.: Matching and prediction on the principle of biological classification. Applied statistics **7** (1959)
11. Jordan, C.: Les coefficients d'intensité relative de Korosy. Revue de la société hongroise de statistique **5** (1927)
12. Light, R.J., Margolin, B.H.: An analysis of variance for categorical data. Journal of the American Statistical Association **66**(335) (1971) pp. 534–544
13. Ah-Pine, J., Marcotorchino, J.F.: Overview of the relational analysis approach in data-mining and multi-criteria decision making. In Usmani, Z.U.H., ed.: Web Intelligence and Intelligent Agents. (2010)
14. Ah-Pine, J.: Cluster analysis based on the central tendency deviation principle. In: Proceedings of the 5th International Conference on Advanced Data Mining and Applications. ADMA '09, Springer-Verlag (2009) 5–18
15. Marcotorchino, J.F., Michaud, P.: Heuristic approach of the similarity aggregation problem. Methods of operations research **43** (1981) 395–404
16. Marcotorchino, J.F.: Maximal association theory as a tool of research. In: Classification as a tool of research. North-Holland, Amsterdam (1986) 275–288
17. Newman, M.: Fast algorithm for detecting community structure in networks. Physical Review E **69** (September 2003)
18. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. Phys. Rev. E **78**(4) (October 2008) 046110
19. Fortunato, S., Lancichinetti, A.: Community detection algorithms: a comparative analysis: invited presentation, extended abstract. In Stea, G., Mairesse, J., Mendes, J., eds.: VALUETOOLS, ACM (2009) 27