International Journal of Bifurcation and Chaos

UNSUPERVISED CLUSTERING ANALYSIS: A MULTISCALE COMPLEX NETWORKS APPROACH

CLARA GRANELL
Department d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili
Av. Països Catalans 26, 43007 Tarragona, Catalonia, Spain
clara.granell@urv.cat

SERGIO GÓMEZ
Department d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili
Av. Països Catalans 26, 43007 Tarragona, Catalonia, Spain
sergio.gomez@urv.cat

ALEX ARENAS
Department d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili
Av. Països Catalans 26, 43007 Tarragona, Catalonia, Spain
alexandre.arenas@urv.cat

Received (to be inserted by publisher)

Unsupervised clustering, also known as natural clustering, stands for the classification of data according to their similarities. Here we study this problem from the perspective of complex networks. Mapping the description of data similarities to graphs, we propose to extend two multiresolution modularity based algorithms to the finding of modules (clusters) in general data sets producing a multiscales' solution. We show the performance of these reported algorithms to the classification of a standard benchmark of data clustering and compare their performance.

Keywords: Clustering, networks, community structure, multiple resolution, modularity.

1. Introduction

The problem of unsupervised data clustering consists in classifying elements so that two data points belonging to the same cluster are more similar between them than with elements in a different cluster. An element, or pattern, is a vector of features (usually understood as a point in a multidimensional space) that describes the item we wish to classify. The goal of the process of data clustering is to organize these patterns finding a partition of the sample according to the natural classes that are present in it. Data clustering has been the subject of interest in many disciplines where the mining of raw information is crucial to understand some phenomenon or gain insight into a system. It has applications in several fields such as pattern recognition, astronomic classification, biological taxonomy, marketing, and more [Gan *et al.*, 2008].

The methodology used to obtain the clusters from the raw data is as follows: First of all, a representation of the patterns has to be chosen, and also a feature selection or extraction is performed. Feature selection means choosing, from all the available features, those that will make easier the process of clustering, leaving the redundant, correlated and less informative features out of the analysis. On the other hand,

feature extraction consists in transforming the original data set to a new one containing only the most relevant information. This preprocessing of the data is very important, as the result of the clustering often depends directly on the quality of this first step. Secondly, the similarity or dissimilarity between each pair of patterns has to be computed, which is often done by defining a measure of distance. The result of this step is the similarity matrix which, using the mapping to complex networks, can be understood as a graph where each node is a pattern and the links are the similarities between them. Finally, the main step of the process, the grouping (or clustering) algorithm, which will decompose the similarity matrix and return the groups of data [Jain *et al.*, 1999; Xu & Wunsch, 2005].

The problem of clustering is inherently ill-posed, i.e. any data set can be clustered in drastically different ways, with no clear criterion for preferring one clustering over another. In particular, in the case of unsupervised approaches, a satisfactory clustering of data depends on the desired resolution which determines the number of clusters and their size. For example, k-means clustering fixes a priori the number of groups (k), which implies indeed a certain resolution of the clustering. Other algorithms such as hierarchical clustering [Kaufman & Rousseeuw, 2005] group the patterns extending the measure of distance between them to distances between clusters of patterns. This process generates a complete dendrogram. Cutting the dendrogram at different heights we obtain different partitions of the data, all them hierarchically nested. In this situation the following question arises: To what resolution should one look at the data to find a scientific meaning in the classification? We claim that the answer to this question is totally dependent on the final purpose of the classification process, and that the concept of best solution should be reconsidered. Different partitions will be representative of properties of the data at different scales and then all of them are worth to be studied.

In this work we perform a comparison between two different multiresolution algorithms, used in the field of complex networks to detect community structure, applied to the problem of data clustering. We also compare our results with a hierarchical clustering (HC) algorithm. In contrast with hierarchical clustering, the multiresolution methods are not necessarily hierarchical. The first algorithm is the multiresolution static screening of the topology of the network, based on the introduction of a control parameter in the resolution of modularity [Arenas *et al.*, 2008] (AFG method), proposed by the authors. The second one is a multiresolution dynamic screening of the network structure using a method, inspired in the Potts model, proposed by Reichardt and Bornholdt [Reichardt & Bornholdt, 2004] (RB method). Both algorithms show to be competitive with classical clustering methods in the classification of the Iris data set.

2. Data clustering preprocessing

Here we briefly review the two stages of the data clustering process before performing the clustering, for an extended revision see [Hall *et al.*, 2009]. Basically, it consists in two stages concerning the data representation and the definition of similarity measures between data points.

The first stage of data clustering is to represent the data to which we perform the clustering analysis. These data are usually obtained experimentally, and our first task is to prepare them properly to give the best possible result when applying the clustering algorithm. A good representation of the patterns will result in a simple and easy clustering process, while a poor representation can lead to complex groups whose structure is difficult or impossible to ascertain. It is worth then to invest some time analyzing the original data to see if one can make a proper pretreatment.

Given that any clustering process will try to find regularities among the data, a good pretreatment should facilitate the process by filtering noisy or redundant information, and reducing the data dimensionality to simplify its computational handling. Usually data are represented as vectors of features, being those categorical or numerical. Without loss of generality, in what follows we will assume that the clustering is intended on vectors of numerical features.

One of the techniques to preprocess the data is feature selection. It will be necessary to apply a feature selection algorithm when some of the features are correlated with each other. In this case, these variables provide redundancy into the system and can introduce a bias towards the final classification based on differences in other not-correlated features. Another scenario where this is useful consists in cases where we have an excessive number of variables and a discriminatory elimination could enhance the handling of

them. Among the different methods for feature selection, we have for example, forward selection/backward elimination: In forward selection, we grow subsets of features depending on the classification obtained, while in backward elimination, we start with all the variables and we eliminate those less promising also according with the classification obtained. Other technique is the decision tree, where we consider the problem of variable selection as a decision problem. Once this analogy is assumed, the decision consists in finding out which subset of variables is more appropriate. As in any decision problem based on trees, the result of the selection will depend on the utility functions used. Another alternative is the naive Bayes classifier, which is a simple probabilistic classifier based on the application of Bayes' theorem. In the context of variable selection this method can involve certain assumptions about dependence or independence of variables and compute their conditional probabilities. Finally, it is worth mentioning the neural networks approaches, e.g. self-organized Kohonen maps, in which the c-plane map of variables is analyzed in order to determine which of those variables can offer better differentiation groups.

There may be some cases in which all features are significant and the elimination of any of them would cause a significant loss of information. In these situations, a feature extraction method is more adequate than any feature selection. A feature extraction algorithm is a method that takes as input the original features and mixes and/or merges them producing a set of new categories that can be filtered and analyzed in the same way as the original data. Examples of feature extraction algorithms are: Principal Component Analysis (PCA) [Jolliffe, 2005], a method aimed to perform a linear transformation of the data converting a set of correlated variables into a new set of less correlated variables called principal components. The first principal component recovers the maximum variance, the second component retrieves the second highest variance and so on, until all have described the variability of the original data. Algebraically, the process involves finding a basis of orthogonal vectors (the principal components) in the n-dimensional space of the original variables, such that the length of the components provides information on the volume and distribution of the data in different directions of the space. In this way, not using all the original features but only the main components can capture most of the information in a reduced set of variables. Other alternatives apart from PCA include nonlinear projections such as self-supervised backpropagation in neural networks or Independent Component Analysis (ICA).

The second stage of the process of clustering is to calculate the similarity (or dissimilarity) between patterns according to a similarity measure, which is usually based on a distance function. The representation of these similarities form a square matrix of size $N \times N$, where N is the number of patterns we have. The similarity matrix can be understood as a complete weighted graph where each node is regarded as one of the patterns and the weight of the link between them informs about their similarity. Please note that if the similarity measure used is not symmetric, then the graph should be directed. Once the similarity matrix is obtained, one can apply graph based community analysis algorithms to perform the clustering of the data.

3. The complex networks approach

Complex networks are graphs representative of the intricate connections between elements in many natural and artificial systems [Strogatz, 2001; Song *et al.*, 2005; Barabási, 2005], whose description in terms of statistical properties has been largely developed in the curse for a universal classification of them. However, when the networks are locally analyzed some characteristics that become partially hidden in the statistical description emerge. The most relevant perhaps is the discovery in many of them of *community structure*, meaning the existence of densely (or strongly) connected groups of nodes, with sparse (or weak) connections between them [Girvan & Newman, 2002].

The study of the community structure helps to elucidate the organization of the networks and, eventually, could be related to the functionality of groups of nodes [Guimerà & Amaral, 2005b]. The most successful solutions to the community detection problem, in terms of accuracy, are those based in the optimization of a quality function called *modularity*, proposed by Newman and Girvan [Newman & Girvan, 2004], that allows the comparison of different partitioning of the network. Given a network partitioned into communities, being C_i the community to which node i is assigned, the mathematical definition of

modularity is

$$Q = \frac{1}{2w} \sum_i \sum_j \left(w_{ij} - \frac{w_i w_j}{2w} \right) \delta(C_i, C_j) \,, \tag{1}$$

where w_{ij} is the weight of the link between nodes i and j (zero if no link exists), $w_i = \sum_j w_{ij}$ is the strength of node i and $2w = \sum_i w_i$ is the total strength of the network [Newman, 2004a]. The Kronecker delta function $\delta(C_i, C_j)$ takes the value 1 if node i and j are into the same community and 0 otherwise. The modularity of a given partition is then the probability of having edges falling within groups in the network minus the expected probability in an equivalent (null case) network. This null case network has the same number of nodes as the original network and the edges placed at random preserving the nodes' strengths. Having a large modularity value means a higher deviation from the null case, and therefore a better partitioning of the network. Note that the optimization of the modularity cannot be performed by exhaustive search since the number of different partitions is equal to the Bell or exponential numbers [Bell, 1934], which grow at least exponentially in the number of nodes N. Indeed, optimization of modularity is a NP-hard (Non-deterministic Polynomial-time hard) problem [Brandes *et al.*, 2008]. Several authors have attacked the problem, with considerable success, by proposing different optimization heuristics [Newman, 2004b; Clauset *et al.*, 2004; Guimerà & Amaral, 2005a; Duch & Arenas, 2005; Pujol *et al.*, 2006; Newman, 2006], see [Fortunato, 2010] for a review.

Maximizing modularity one obtains the "best" partition of the network into communities. This partition represents an intermediate topological scale of organization, or mesoscale, that in many cases has been shown to coincide with known information about subdivisions in the network [Newman & Girvan, 2004; Danon *et al.*, 2005]. However, recently, it has been pointed out that the optimization of the modularity has a characteristic scale related to the number of links in the network, which delimits the resolution beyond which no separation into smaller groups can be obtained when optimizing modularity, even though these smaller partitions, and then different levels of description, are plausible to exist from direct observation [Fortunato & Barthélemy, 2007]. The problem seems then that modularity, as it has been prescribed, does not have access to these other levels of description. The reason for this is that the topological scale at which we have access by maximizing modularity has a topological resolution limit.

We proposed a method that allows the full screening of the topological structure at any resolution level using the original formulation and semantics of modularity, overcoming then the resolution limit [Arenas *et al.*, 2008]. Our aim is to take advantage of this method to analyze real data sets in terms of clustering. In contrast with the solution proposed in [Angelini *et al.*, 2007], in which the clustering is found using modularity at the standard Newman scale, our approach uses the multiple resolution method that optimizes modularity at each level of resolution. The mathematical form of our prescription is given by

$$Q_{\mathrm{AFG}}(r) = Q[w_{ij} \leftarrow w_{ij} + r\delta_{ij}] \,, \tag{2}$$

where r (resistance) is the parameter controlling the resolution of the partitions we want to find, and $w_{ij} + r\delta_{ij}$ is the new weights' matrix after adding a self-loop with value r to each node. When r is zero, we recover the standard modularity Q. The definition of Q_{AFG} does preserve the original semantics of modularity. A recent approach proposed by [Pons & Latapy, 2011] uses also a multi-hierarchical technique based on optimizing a class of quality functions that extend the concept of Newman's modularity parameterizing both terms of equation (1). We do not use this approach in the current work although it could be compared in the same way we compare with the Reichardt and Bornholdt method explained below.

In [Reichardt & Bornholdt, 2004], the authors propose a method in which the graph is understood as a dynamical system of q-states interacting spins (usually known as Potts' model), and the states of the spins represent the community to which a node is assigned. In this scenario, the partition in modules is equivalent to the ground state of the mentioned dynamical system. Indeed, the authors made a very interesting connection between the statistical mechanics of the Potts model and modularity. Moreover, although the finding of the resolution limit was discovered later, the Reichardt and Bornholdt (RB) method already solved this problem by the tuning of a parameter, as pointed out in [Kumpula *et al.*, 2007]. The result is that the ground state of the system corresponding to the minimum of its Hamiltonian can be

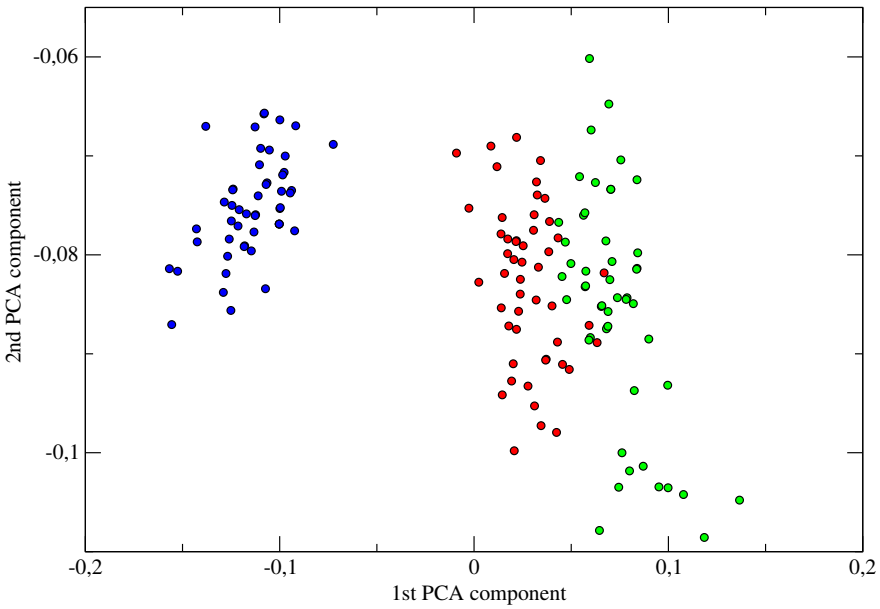

Fig. 1. Two principal components of the PCA analysis on the Iris data set. Colors correspondence is: *setosa*-blue, *versicolor*-red, and *virginica*-green. While *setosa* is clearly linearly separable, the other two species are not.

written as

$$Q_{\mathrm{RB}}(\gamma) = \frac{1}{2w} \sum_i \sum_j \left(w_{ij} - \gamma \frac{w_i w_j}{2w} \right) \delta(C_i, C_j) \,, \tag{3}$$

where γ is the resolution control parameter in this case. Note that the original Q corresponds to $\gamma = 1$ and other values of this parameter represent different quality functions as the weight of the null model term is modified.

To screen the whole spectrum of resolution levels of the topological structure of any given network, we must determine the values of $r_{\min}$ and $r_{\max}$ for the AFG model, and the $\gamma_{\min}$ and $\gamma_{\max}$ for the RB model. Assigning the minimum value to the parameter, the network will appear as a unique module, while at the maximum value the network will be divided in as many modules as nodes. The mathematical determination of these limits is discussed in the Appendix for the most general case of directed and signed networks. The screening of the mesoscales is done by optimizing both modularities $Q_{\mathrm{AFG}}(r)$ and $Q_{\mathrm{RB}}(\gamma)$, for the different values of r and γ respectively.

4. Results

To show the ability of multiresolution community detection methods to solve the problem of unsupervised data clustering, we have chosen to study the classical benchmark of the Iris data set. This dataset, presented by Sir R.A. Fisher in 1936, consists of 50 samples from each of three species of Iris flowers (*Iris setosa*, *Iris versicolor* and *Iris virginica*). We know the petal length, petal width, sepal length and sepal width from each sample. For the moment, we will ignore the species information and we will cluster the data using only the raw measurements as in [Fisher, 1936]. When this is done, a comparison between the real classification and the obtained clusters can be made, in order to evaluate its quality.

Following the steps of data clustering explained above, we first perform a principal component analysis of the four features that form each pattern, and choose to work with the two principal components corresponding to the largest part of the data variance. In Fig. 1 a representation of these two components is shown. Based on these two variables, we build up a similarity matrix from the euclidean distances between patterns components with respect to the average distance in this space. For any pair of flowers i and j, we define the similarity $s_{ij} = \bar{d} - \|x^i - x^j\|)$, where $\bar{d}$ stands for the average distance of the set, and $\|\cdot\|$ is the euclidean distance between the feature vectors of each flower. The resulting similarity matrix is interpreted

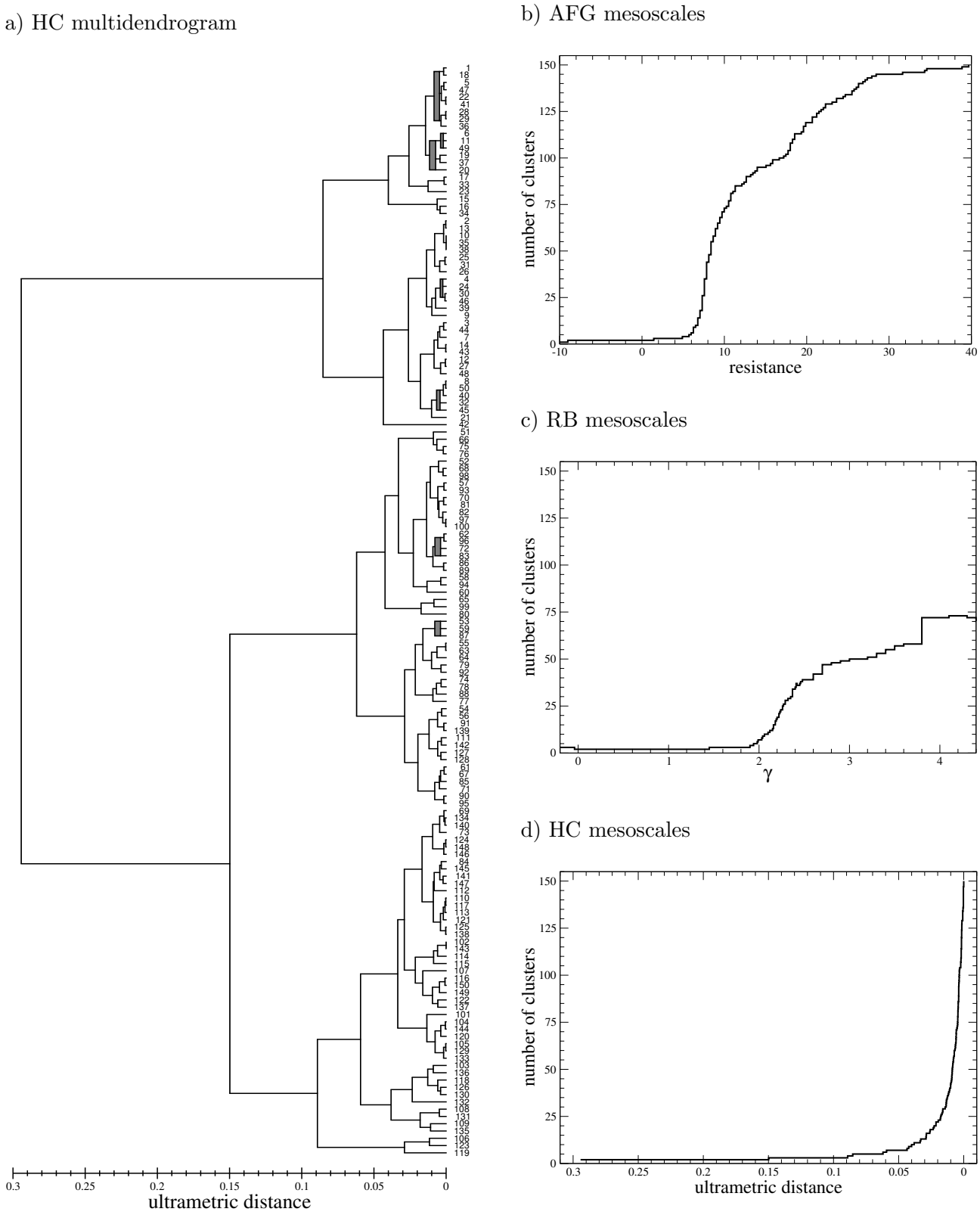

Fig. 2. Mesoscales of the Iris data set, showing the number of clusters as a function of the resolution parameter: a) HC complete linkage multidendrogram; b) AFG mesoscales; c) RB mesoscales; d) HC mesoscales from the previous multidendrogram.

as a weighted network whose communities will, in principle, reproduce the right clustering of the data. Note that this matrix has positive and negative links, and that modularity should account for this signed values, see Appendix.

We present the comparison of the results obtained using the algorithms described above, and also compare with the solution obtained applying a classical hierarchical clustering technique, see Fig. 2.

In particular, we constructed the hierarchical clustering using complete linkage, where the distance between groups is defined as the distance between the most distant pair of individuals, one from each group. In other words, the distance between two clusters is given by the value of the longest link between the clusters. At each stage of hierarchical clustering, the clusters at minimum distance are merged. Moreover, instead of using the standard pair-group hierarchical clustering approach, we take advantage of a recent development by some of the authors [Fernández & Gómez, 2008] that allows to solve the non-uniqueness problem when there are tied distances during the agglomeration process (code available at [Gómez & Fernández, 2010]). The result, known as a *multidendrogram*, is presented in Fig. 2a. We plot the tag number of each specimen at the leaves of the tree. The analysis of the multidendrogram can be performed as follows: starting from the root of the tree, we can compute the distances between different partitions of the data and analyze each of them separately.

The comparison between the three methods can be done by computing the multiple scales of the topology in terms of community structure, screening the values of r in the AFG method, the values of γ in the RB method, and the distances in the dendrogram. In Fig. 2b we present the whole mesoscale for the AFG method, we observe the persistence of the partition in two and three clusters as the most representatives of the mesoscale. In Fig. 2c we present a portion of the mesoscale for the RB method, again the last observation holds for this method, however, the variations of γ do not ensure a monotonic behavior of the number of clusters as a function of γ (see Appendix for details). Finally, we plot the mesoscale in terms of distances in the dendrogram, see Fig. 2d. The hierarchical clustering approach defines also two main resolution levels corresponding to two and three clusters partitions, respectively. The fact that the partition that divides the data in two communities is always the most relevant in any of the used methods corresponds to the true partition of the Iris data set in two linearly separable sets.

We define two measures to make the comparison between the different methods, centering our attention in the most relevant partitions in terms of the scale length, see Fig. 3. The first measure is the success, which is the percentage of correctly classified nodes when comparing the partition obtained with the original classification made by biologists using more features of the flowers. In this case and for the partition in three clusters, both HC and AFG methods achieve a $94,67\%$ of success, corresponding to a mismatch of eight flowers in total. The RB method obtains a success of $90,67\%$ in this case. The second measure we contemplate is the Jaccard index presented in [Jaccard, 1912], which is the fraction of pairs of patterns in the same cluster in one partition which are also in the same cluster in the other partition. The larger the fraction of same cluster co-ocurrences, the better the quality of the agreement. In Fig. 3(right) we observe that the best classification in three clusters is performed by the AFG method by a slight difference (0.8194 the AFG method versus 0.8180 the HC).

5. Conclusions

This article has presented the adaptation and performance of two multiresolution methods for the detection of the community structure in networks to the problem of unsupervised data clustering. A multiresolution method, in contrast to those that find communities at a fixed scale, is a method that allows finding partitions at different levels of resolution. We focused on the determination of groups in the similarity matrix using modularity as the quality function. We have analytically computed the two limiting cases for the AFG method, from a unique cluster to the classification of every data point as a single cluster. The different partitions of the set were obtained, which correspond to the different scales of resolution. We also discussed about the problem of finding the best partition from the mesoscale. As all the partitions reflect the structure of the data, we established that the concept of choosing the "best" partition should be translated to choosing the most persistent partition for values of the resistance parameter. The results obtained on the classical Iris data set are competitive with classical unsupervised clustering techniques. These results

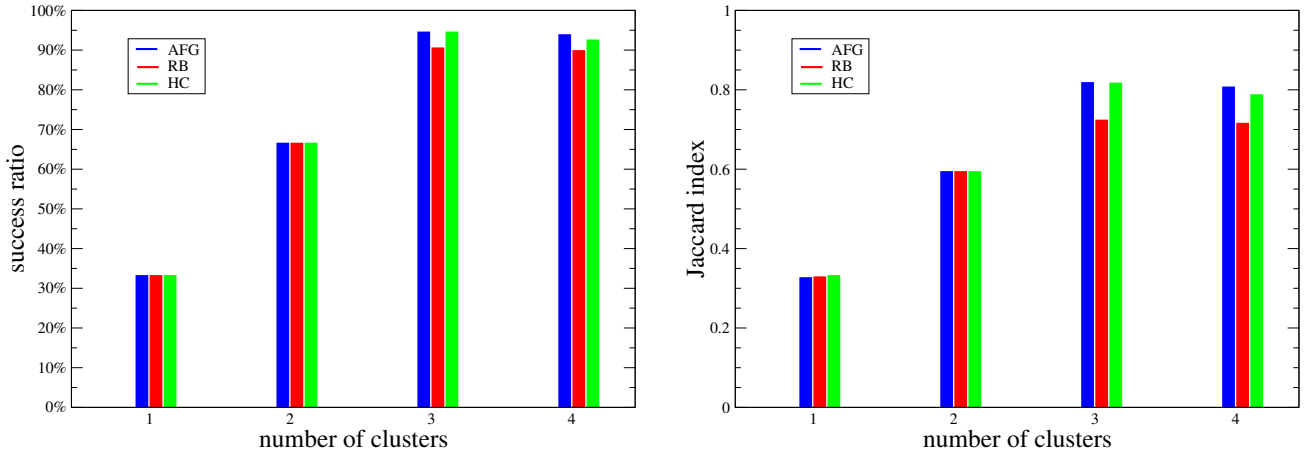

Fig. 3. Comparison between the three methods used in the classification of the Iris data set. Two measures are used: the success ratio (left) and the Jaccard index (right). Only the partitions with highest performance and less than five clusters are shown.

are encouraging, and point out that the mapping of clustering problems to networks' structural analysis is a field worth to be explored.

Acknowledgments

We acknowledge support from the Spanish Ministry of Science and Innovation FIS2009-13730-C02-02 and the Generalitat de Catalunya SGR-00838-2009.

Appendix A Determination of AFG mesoscales boundaries

The generalization of modularity Eq. (1) for undirected weighted signed networks (see [Gómez *et al.*, 2009]) is

$$Q = \frac{1}{2w^+ + 2w^-} \sum_i \sum_j \left[w_{ij} - \left(\frac{w_i^+ w_j^+}{2w^+} - \frac{w_i^- w_j^-}{2w^-} \right) \right] \times \delta(C_i, C_j) \,. \tag{A.1}$$

where

$$w_i^+ = \sum_{j, w_{ij}>0} w_{ij} \,, \tag{A.2}$$

$$w_i^- = \sum_{j, w_{ij}<0} |w_{ij}| \,, \tag{A.3}$$

are the positive and negative strengths of node i, and

$$2w^+ = \sum_i w_i^+ \,, \tag{A.4}$$

$$2w^- = \sum_i w_i^- \,, \tag{A.5}$$

are the positive and negative total strengths respectively. Please note that these four strengths are defined to be non-negative.

To simplify the notation, we make use of the modularity matrix

$$B_{ij} = w_{ij} - \left(\frac{w_i^+ w_j^+}{2w^+} - \frac{w_i^- w_j^-}{2w^-} \right) \,, \tag{A.6}$$

therefore

$$Q = \frac{1}{2w^+ + 2w^-} \sum_{i=1}^{N} \sum_{j=1}^{N} B_{ij} \delta(C_i, C_j) \,. \tag{A.7}$$

Following [Arenas *et al.*, 2008], the analysis of the mesoscale is performed with the addition of a common self-loop to all the nodes in the network. The boundaries of the mesoscale are the *macroscale*, a partition in which all nodes belong to the same community, and the *microscale*, a partition in which each node is isolated in its own community. The determination of these boundaries is equivalent to finding two values of the self-loops, $r_{\min}$ and $r_{\max}$, for which the maximum of modularity $Q_{\mathrm{AFG}}(r)$ is achieved at the macroscale and microscale respectively. The solution is as follows: if all the non-diagonal terms of the modularity matrix are positive or zero, modularity is optimized at the macroscale, and if they are negative, it is optimized at the microscale. Diagonal terms are irrelevant since $\delta(C_i, C_i) = 1$ for all nodes.

If we introduce a positive self-loop r^+, the modularity matrix becomes

$$B_{ij}^{\mathrm{AFG}}(r^+) = w_{ij} + r^+ \delta_{ij} - \left(\frac{(w_i^+ + r^+)(w_j^+ + r^+)}{2w^+ + Nr^+} - \frac{w_i^- w_j^-}{2w^-} \right) , \tag{A.8}$$

and with a negative self-loop $-r^-$

$$B_{ij}^{\mathrm{AFG}}(-r^-) = w_{ij} - r^- \delta_{ij} - \left(\frac{w_i^+ w_j^+}{2w^+} - \frac{(w_i^- + r^-)(w_j^- + r^-)}{2w^- + Nr^-} \right) . \tag{A.9}$$

The existence of $r_{\max}$ is straightforward, since $B_{ij}^{\mathrm{AFG}}(r^+) \sim -r^+ < 0$ for large enough r^+ and $i \neq j$. Its determination is just an exercise of solving the system of inequations $B_{ij}^{\mathrm{AFG}}(r^+) \leq 0$ for $i < j$, and taking the smallest solution as $r_{\max}$. More precisely,

$$r_{\max} = \max_{\substack{i<j \\ D_{ij}^2 \geq 4E_{ij}}} \left(-\frac{D_{ij}}{2} + \frac{1}{2}\sqrt{D_{ij}^2 - 4E_{ij}} \right) , \tag{A.10}$$

where

$$D_{ij} = w_i^+ + w_j^+ - N \left(w_{ij} + \frac{w_i^- w_j^-}{2w^-} \right) , \tag{A.11}$$

$$E_{ij} = w_i^+ w_j^+ - 2w^+ \left(w_{ij} + \frac{w_i^- w_j^-}{2w^-} \right) . \tag{A.12}$$

In the same way, $B_{ij}^{\mathrm{AFG}}(-r^-) \sim r^- > 0$ proves the existence of $r_{\min}$, and it is calculated by solving $B_{ij}^{\mathrm{AFG}}(-r^-) \geq 0$ for $i < j$, and taking the largest solution as $r_{\min}$, i.e.

$$r_{\min} = - \max_{\substack{i<j \\ D_{ij}^2 \geq 4E_{ij}}} \left(-\frac{D_{ij}}{2} + \frac{1}{2}\sqrt{D_{ij}^2 - 4E_{ij}} \right) , \tag{A.13}$$

where

$$D_{ij} = w_i^- + w_j^- + N \left(w_{ij} - \frac{w_i^+ w_j^+}{2w^+} \right) , \tag{A.14}$$

$$E_{ij} = w_i^- w_j^- + 2w^- \left(w_{ij} - \frac{w_i^+ w_j^+}{2w^+} \right) . \tag{A.15}$$

Please note that while the value of $r_{\max}$ is the exact point at which the macroscale is obtained, the value of $r_{\min}$ is only a lower bound of the microscale. This lower bound may be improved using a bisection method between the value of $r_{\min}$ given by Eq. (A.13) and $r = 0$.

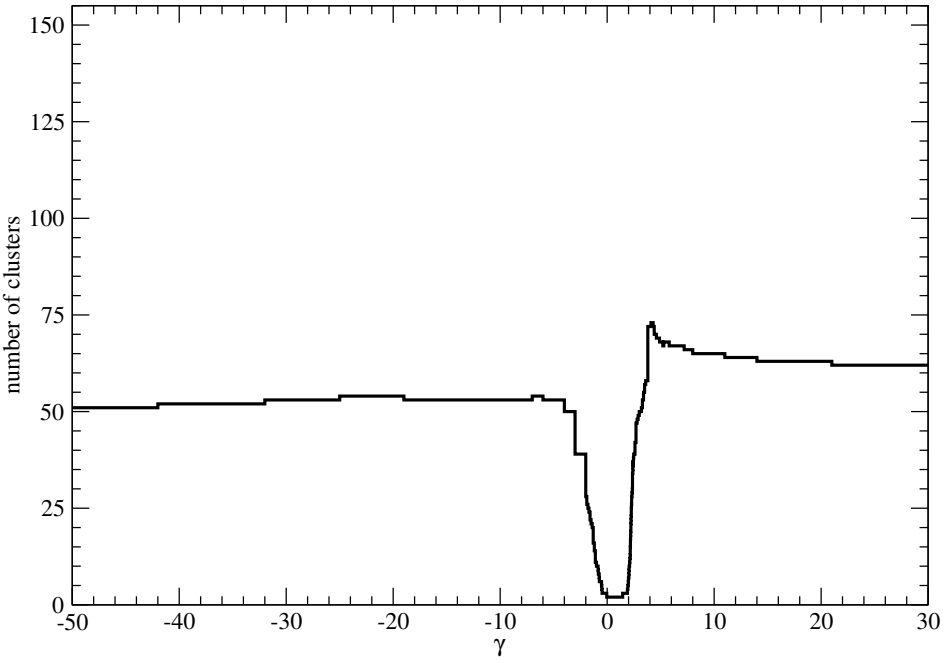

Fig. B.1. Expanded Iris data set RB mesoscales analysis.

When the network is directed, the analysis of the AFG mesoscale is exactly the same, but with the substitutions

$$w_i^{\pm} \to w_i^{\pm,\text{out}} = \sum_{k,\pm w_{ik}>0} |w_{ik}|\,, \tag{A.16}$$

$$w_j^{\pm} \to w_j^{\pm,\text{in}} = \sum_{k,\pm w_{kj}>0} |w_{kj}|\,, \tag{A.17}$$

$$D_{ij} \to \frac{1}{2}(D_{ij} + D_{ji})\,, \tag{A.18}$$

$$E_{ij} \to \frac{1}{2}(E_{ij} + E_{ji})\,. \tag{A.19}$$

The code for the determination of the AFG mesoscales is available at [Gómez *et al.*, 2010]).

Appendix B Boundaries of RB mesoscales

In the RB formulation of mesoscales, a parameter γ is introduced in front of the null-case term to weight its relative importance against the real network, i.e.

$$B_{ij}^{\text{RB}}(\gamma) = w_{ij} - \gamma \left(\frac{w_i^+ w_j^+}{2w^+} - \frac{w_i^- w_j^-}{2w^-} \right)\,. \tag{B.1}$$

It is also possible to have different parameters for the positive and negative null-case terms as in [Traag & Bruggeman, 2009], however this leads to a bidimensional analysis of the mesoscales, which is almost unaffordable for most real networks. Thus, we will focus on the single-parameter RB modularity matrix Eq. (B.1).

Without negative weights, the macroscale is recovered at $\gamma_{\min} = 0$, and the microscale at the $\gamma_{\max}$ which makes all modularity terms negative. The existence of $\gamma_{\max}$ is guaranteed by the fact that all null-case terms are positive. However, the addition of negative weights makes it possible to have both positive and negative null-case terms, which does not allow to ensure the recovery of macro and microscale. Therefore, RB signed modularity may not cover the whole mesoscale. This is experimentally confirmed in Fig. B.1 for

the Iris data set, where a larger interval of the γ parameter has been analyzed. While Fig. 2c only shows the useful part of the mesoscales range, where the number of clusters goes from 2 to 73 ($\gamma \in [0.0, 4.2]$), in Fig. B.1 it is shown the inability of RB to find the macroscale (microscale) for lower (larger) values of γ.

References

Angelini, L., Marinazzo, D., Pellicoro, M. & Stramaglia, S. [2007] "Natural clustering: the modularity approach," *J. Stat. Mech.* , L08001.

Arenas, A., Fernández, A. & Gómez, S. [2008] "Analysis of the structure of complex networks at different resolution levels," *New J. Phys.* **10**, 053039.

Barabási, A.-L. [2005] "Network theory – the emergence of the creative enterprise," *Science* **308**, 639.

Bell, E. T. [1934] "Exponential numbers," *Amer. Math. Monthly* **41**, 411.

Brandes, U., Delling, D., Gaertler, M., Goerke, R., Hoefer, M., Nikoloski, Z. & Wagner, D. [2008] "On modularity clustering," *IEEE Trans. Knowl. Data Eng.* **20**, 172.

Clauset, A., Newman, M. E. J. & Moore, C. [2004] "Finding community structure in very large networks," *Phys. Rev. E* **70**, 066111.

Danon, L., Díaz-Guilera, A., Duch, J. & Arenas, A. [2005] "Comparing community structure identification," *J. Stat. Mech.* , P09008.

Duch, J. & Arenas, A. [2005] "Community identification using extremal optimization," *Phys. Rev. E* **72**, 027104.

Fernández, A. & Gómez, S. [2008] "Solving non-uniqueness in agglomerative hierarchical clustering using multidendrograms," *Journal of Classification* **25**, 43.

Fisher, R. A. [1936] "The use of multiple measurements in taxonomic problems," *Annals of Eugenics* **7**, 179.

Fortunato, S. [2010] "Community detection in graphs," *Phys. Rep.* **486**, 75.

Fortunato, S. & Barthélemy, M. [2007] "Resolution limit in community detection," *Proc. Natl. Acad. Sci. USA* **104**, 36.

Gan, G., Ma, C. & Wu, J. [2008] *Data Clustering: Theory, Algorithms, and Applications*, Series on Statistics and Applied Probability, Vol. 76 (ASA-SIAM).

Girvan, M. & Newman, M. E. J. [2002] "Community structure in social and biological networks," *Proc. Natl. Acad. Sci. USA* **99**, 7821.

Gómez, S. & Fernández, A. [2010] URL http://deim.urv.cat/~sgomez/multidendrograms.php.

Gómez, S., Fernández, A., Borge-Holthoefer, J. & Arenas, A. [2010] URL http://deim.urv.cat/~sgomez/radatools.php.

Gómez, S., Jensen, P. & Arenas, A. [2009] "Analysis of community structure in networks of correlated data," *Phys. Rev. E* **80**, 016114.

Guimerà, R. & Amaral, L. A. N. [2005a] "Cartography of complex networks: modules and universal roles," *J. Stat. Mech.* , P02001.

Guimerà, R. & Amaral, L. A. N. [2005b] "Functional cartography of complex metabolic networks," *Nature* **433**, 895.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. [2009] "The weka data mining software: an update," *ACM SIGKDD* **11**, 10.

Jaccard, P. [1912] "The distribution of flora in the alpine zone," *The New Phytologist* **11**, 37.

Jain, A. K., Murty, M. N. & Flynn, P. J. [1999] "Data clustering: A review," *ACM Comp. Surv.* **31**, 264.

Jolliffe, I. [2005] *Principal Component Analysis* (Wiley).

Kaufman, L. & Rousseeuw, P. J. [2005] *Finding Groups in Data: An Introduction to Cluster Analysis* (Wiley).

Kumpula, J. M., Saramaki, J., Kaski, K. & Kertesz, J. [2007] "Limited resolution and multiresolution methods in complex network community detection," *Fluctuation Noise Letters* **7**, 209.

Newman, M. E. J. [2004a] "Analysis of weighted networks," *Phys. Rev. E* **70**, 056131.

Newman, M. E. J. [2004b] "Fast algorithm for detecting community structure in networks," *Phys. Rev. E* **69**, 066133.

Newman, M. E. J. [2006] "Modularity and community structure in networks," *Proc. Natl. Acad. Sci. USA* **103**, 8577.

Newman, M. E. J. & Girvan, M. [2004] "Finding and evaluating community structure in networks," *Phys. Rev. E* **69**, 026113.

Pons, P. & Latapy, M. [2011] "Post-processing hierarchical community structures: Quality improvements and multi-scale view," *Theoretical Computer Science* **412**, 892.

Pujol, J. M., Béjar, J. & Delgado, J. [2006] "Clustering algorithm for determining community structure in large networks," *Phys. Rev. E* **74**, 016107.

Reichardt, J. & Bornholdt, S. [2004] "Detecting fuzzy community structures in complex networks with a potts model," *Phys. Rev. Lett.* **93**, 218701.

Song, C. M., Havlin, S. & Makse, H. A. [2005] "Self-similarity of complex networks," *Nature* **433**, 392.

Strogatz, S. H. [2001] "Exploring complex networks," *Nature* **410**, 268.

Traag, V. A. & Bruggeman, J. [2009] "Community detection in networks with positive and negative links," *Phys. Rev. E* **80**, 036115.

Xu, R. & Wunsch, D. [2005] "Survey of clustering algorithms," *IEEE Trans. Neural Networks* **16**, 645.