Mesoscopic analysis of networks: applications to exploratory analysis and data clustering

Clara Granell, Sergio Gómez, and Alex Arenas[a)]
Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, 43007 Tarragona, Catalonia, Spain

(Dated: 7 February 2011)

We investigate the adaptation and performance of modularity-based algorithms, designed in the scope of complex networks, to analyze the mesoscopic structure of correlation matrices. Using a multi-resolution analysis we are able to describe the structure of the data in terms of clusters at different topological levels. We demonstrate the applicability of our findings in two different scenarios: to analyze the neural connectivity of the nematode *Caenorhabditis elegans*, and to automatically classify a typical benchmark of unsupervised clustering, the Iris data set, with considerable success.

PACS numbers: 89.75.Hc,89.75.Fb
Keywords: Clustering, networks, community structure, multiple resolution, modularity.

Facing the famous Salvador Dali's painting "Gala contemplating the Mediterranean sea which at twenty meters becomes a portrait of Abraham Lincoln", we have the best proof of how a complex system reveals different information when observed at different (in this case length) scales. We proposed a method[1] to unveil the equivalent phenomena in the description of complex networks from a topological perspective. By defining a parameter that controls the resistance of each node to belong to a group, we are able to analyze the community structure of the network at different topological scales. We apply the method to the exploratory analysis of the structural connectivity of the neuronal system of *C. elegans* and find a tentative classification of functional activity of groups of neurons at certain topological scales. We also have tested the method to automatically classify a typical benchmark of unsupervised data clustering, the Iris dataset. These results pave the way to the applicability of community detection algorithms in complex networks to the exploration and classification of real data sets.

I. INTRODUCTION

Complex networks are graphs representative of the intricate connections between elements in many natural and artificial systems[2–4], 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[5].

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[6]. 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[7] 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 expressed in terms of the weighted adjacency matrix w_{ij}, that represents the value of the weight in the link between nodes i and j, this weight would be 0 if no link existed, and the strengths $w_i = \sum_j w_{ij}$ as[8]

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

where the Kronecker delta function $\delta(C_i, C_j)$ takes the values, 1 if node i and j are into the same community, 0 otherwise, and the total strength $2w = \sum_i w_i$. 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 with the same number of nodes, and edges placed at random preserving the nodes' strength. The larger the modularity the best the partitioning is, cause more deviates from the null case. Note that the optimization of the modularity cannot be performed by exhaustive search since the number of different partitions is equal to the Bell[9] or exponential numbers, 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[10]. Several authors have attacked the problem, with considerable success, by proposing different optimization heuristics[11–16], see Fortunato[17] for a review.

Maximizing modularity one obtains the "best" parti-

[a)] Electronic mail: alex.arenas@urv.cat

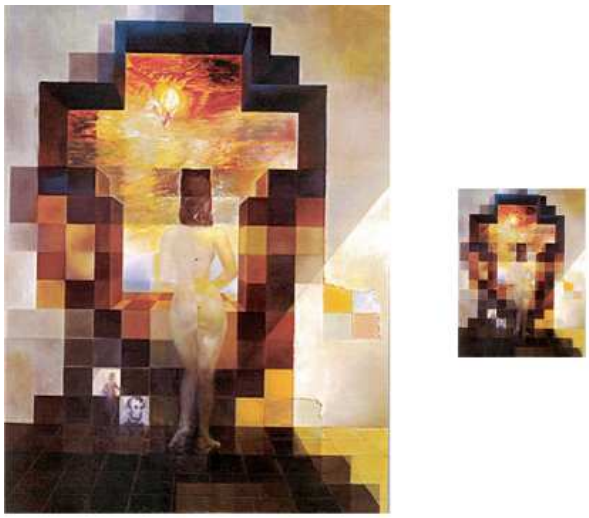

FIG. 1. *"Gala contemplating the Mediterranean sea which at twenty meters becomes a portrait of Abraham Lincoln"*, by Salvador Dali, 1974. Left, at closer distance, and right, at larger distance.

tion 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[7,18]. 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, that 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[19]. The problem seems then that modularity, as it has been prescribed, does not have access to these other levels of description, and then its direct interpretation must be cautiously used[20]. The reason for this is that the topological scale at which we have access by maximizing modularity has a topological resolution limit. The analogy with the observation of Dali's painting is clear, modularity is our tool to "observe" a complex network, and their limit is equivalent of a limit in the distance at which we observe the painting (Fig. 1). We proposed a method[1] 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. Our aim is to take advantage of this method to analyze real data sets in terms of clustering.

The paper is structured as follows: In the next section we overview the multiple resolution method. Once the method has been presented, we propose its application for exploratory analysis in the topology of the neural network of the nematode *C. elegans* in section III, and its application to data clustering in section IV. Finally we present the conclusions of the work in section V.

II. MULTIPLE RESOLUTION METHOD

In this section we provide the necessary tools to extend the multiple resolution method to the most general case of networks with weighted signed directed links.

A. General formulation of modularity

The generalization of modularity to any network, with weighted, directed and signed values of the weights[21] is as follows. Let us suppose that we have a weighted undirected complex network with weights w_{ij} as above. The relative strength p_i of a node

$$p_i = \frac{w_i}{2w}\,, \tag{2}$$

may be interpreted as the probability that this node makes links to other ones, if the network were random. This is precisely the approach taken by Newman and Girvan to define the modularity null case term, which reads

$$p_i p_j = \frac{w_i w_j}{(2w)^2}\,. \tag{3}$$

The introduction of negative weights destroys this probabilistic interpretation of p_i, since in this case the values of p_i are not guaranteed to be between zero and one. The problem is the implicit hypothesis that there is only one unique probability to link nodes, which involves both positive and negative weights. To solve this problem, we have to introduce two different probabilities to form links, one for positive and the other for negative links.

Let us formalize this approach. First, we separate the positive and negative weights:

$$w_{ij} = w^+_{ij} - w^-_{ij}\,, \tag{4}$$

where we use the notation

$$w^+_{ij} = \max\{0, w_{ij}\}\,, \tag{5}$$

$$w^-_{ij} = \max\{0, -w_{ij}\}\,. \tag{6}$$

These expressions are useful since in principle we do not know the sign of w_{ij}. The positive and negative strengths are given by

$$w^+_i = \sum_j w^+_{ij}\,, \tag{7}$$

$$w^-_i = \sum_j w^-_{ij}\,, \tag{8}$$

and the positive and negative total strengths by

$$2w^+ = \sum_i w^+_i = \sum_i \sum_j w^+_{ij}\,, \tag{9}$$

$$2w^- = \sum_i w^-_i = \sum_i \sum_j w^-_{ij}\,. \tag{10}$$

Consequently,

$$w_i = w_i^+ - w_i^- \tag{11}$$

and

$$2w = 2w^+ - 2w^- \,. \tag{12}$$

With these definitions at hand, the connection probabilities with positive and negative weights are respectively

$$p_i^+ = \frac{w_i^+}{2w^+} \,, \tag{13}$$

$$p_i^- = \frac{w_i^-}{2w^-} \,. \tag{14}$$

Now, there are two terms which contribute to modularity: the first one takes into account the deviation of actual positive weights against a null case random network given by probabilities p_i^+, and the other is its counterpart for negative weights. Thus, it is useful to define

$$Q^+ = \frac{1}{2w^+} \sum_i \sum_j \left(w_{ij}^+ - \frac{w_i^+ w_j^+}{2w^+} \right) \delta(C_i, C_j) \,, \tag{15}$$

$$Q^- = \frac{1}{2w^-} \sum_i \sum_j \left(w_{ij}^- - \frac{w_i^- w_j^-}{2w^-} \right) \delta(C_i, C_j) \,. \tag{16}$$

The total modularity must be a trade off between the tendency of positive weights to form communities and that of negative weights to destroy them. If we want that Q^+ and Q^- contribute to modularity proportionally to their respective positive and negative strengths, the final expression for modularity Q is

$$Q = \frac{2w^+}{2w^+ + 2w^-} Q^+ - \frac{2w^-}{2w^+ + 2w^-} Q^- \,. \tag{17}$$

An alternative equivalent form for modularity Q is

$$Q = \frac{1}{2w^+ + 2w^-} \sum_i \sum_j \left[w_{ij} - \left(\frac{w_i^+ w_j^+}{2w^+} - \frac{w_i^- w_j^-}{2w^-} \right) \right] \times \delta(C_i, C_j) \,. \tag{18}$$

The main properties of Eq. (18) are the following: without negative weights, the standard modularity is recovered; modularity is zero when all nodes are together in one community; and it is antisymmetric in the weights, i.e. $Q(C, \{w_{ij}\}) = -Q(C, \{-w_{ij}\})$.

The extension to directed networks[22] is simply obtained by the substitutions in Eq. (18) of

$$w_i^\pm \to w_i^{\pm,\text{out}} = \sum_k w_{ik}^\pm \,, \tag{19}$$

$$w_j^\pm \to w_j^{\pm,\text{in}} = \sum_k w_{kj}^\pm \,. \tag{20}$$

B. Mesocales analysis for weighted signed networks

The extension of the multiple resolution method[1] to the general case of weighted signed networks follows the same original idea. The method relies on the introduction of a magnitude r that we call *resistance*, represented by a self-link for each node, that stands for the opposition of a node to belong to a group, in the sense of modularity. We tune the resistance uniformly for all nodes because in this way the functional form of the strength distribution is preserved and does not distort the relative structural properties of nodes. More precisely, the formulation of modularity Q_r at different resolution scales tagged by r consists in substituting in Eq. (18)

$$w_{ij} \to w_{ij} + r\delta_{ij} \,, \tag{21}$$

$$w_i^\pm \to w_i^\pm + r^\pm \,, \tag{22}$$

$$2w^\pm \to 2w^\pm + Nr^\pm \,, \tag{23}$$

where

$$r = r^+ - r^- \,, \tag{24}$$

and

$$r^+ = \max\{0, r\} \,, \tag{25}$$

$$r^- = \max\{0, -r\} \,. \tag{26}$$

The topological scale determined by maximizing Q at which the detection of community structure has been attacked so far, corresponds to $r = 0$ (Newman's scale). For positive values of r, we have access to the substructure below $r = 0$, and for negative values of r we have access to the superstructures. For negative values of r, the resistance should be understood as an affinity of nodes to belong to the same group, and using Eq. (1) the formulation is still preserved but not the semantics in terms of probabilities. The main challenge in this new scenario is that the limiting cases of r that corresponds to the partition of individual nodes, and to the whole network as a unique module have to be computed using the new modularity formulation Eq. (18).

C. Resistance limiting cases for weighted signed networks

Here we present the mathematical proofs of the physical limiting cases of the resistance for weighted signed networks. Let us call $r_{\max}$ the limit of resistance for which all nodes are isolated in communities of size 1, and $r_{\min}$ the limit for which all nodes become members of a single group that represents the whole network. To determine $r_{\max}$ we look for a value of the resistance such that the increment in modularity when joining any pair of vertices in the same community is negative, and the contrary for $r_{\min}$. The idea is the following: if $r > 0$ and all the non-diagonal terms ($i \neq j$) of Eq. (18) are

negative,

$$w_{ij} \leq \frac{(w_i^+ + r)(w_j^+ + r)}{2w^+ + Nr} - \frac{w_i^- w_j^-}{2w^-} \,, \quad \forall i \neq j \,, \qquad (27)$$

then the maximum of Q_r is achieved with the partition which satisfies $\delta(C_i, C_j) = 0$ for all $i \neq j$, i.e. the partition in which all nodes are isolated. Eqs. (27) form a system of second order inequations in r. After some algebra, it can be shown that $r_{\max}$ is the lowest value of r for which the following set of inequalities per link (denoted ij) is satisfied:

$$\min_{r,ij}[Ar^2 + B_{ij}r + C_{ij} \leq 0] \qquad (28)$$

where

$$A = -2w^- \qquad (29)$$
$$B_{ij} = N(2w^- w_{ij} + w_i^- w_j^-) - 2w^-(w_i^+ + w_j^+) \qquad (30)$$
$$C_{ij} = 2w^- 2w^+ w_{ij} + 2w^+ w_i^- w_j^- - 2w^- w_i^+ w_j^+ \qquad (31)$$

Equivalently, if $r < 0$ and all the non-diagonal terms ($i \neq j$) of Eq. (18) are positive,

$$w_{ij} \geq \frac{w_i^+ w_j^+}{2w^+} - \frac{(w_i^- - r)(w_j^- - r)}{2w^- - Nr} \,, \quad \forall i \neq j \,, \qquad (32)$$

the maximum of Q_r is achieved with the partition which satisfies $\delta(C_i, C_j) = 1$ for all $i \neq j$, i.e. the partition in which all nodes are together in the same community. Thus, to determine a lower bound of $r_{\min}$ we look for the largest value of r satisfying

$$\max_{r,ij}[Ar^2 + B_{ij}r + C_{ij} \geq 0] \qquad (33)$$

where

$$A = 2w^+ \qquad (34)$$
$$B_{ij} = N(2w^+ w_{ij} - w_i^+ w_j^+) + 2w^+(w_i^- + w_j^-) \qquad (35)$$
$$C_{ij} = 2w^+ 2w^- w_{ij} - 2w^- w_i^+ w_j^+ + 2w^+ w_i^- w_j^- \qquad (36)$$

The value of r obtained from Eqs. (33) is only a lower bound of the exact $r_{\min}$, since these equations are only sufficient conditions for the existence of a unique communty holding all the nodes of the network (not all terms in Eq. (18) need to be positive in the $r_{\min}$ limit). On the other hand, Eqs. (28) are necessary and sufficient conditions, and thus the $r_{\max}$ found is the exact value.

The method to unveil the mesoscales of a complex network consists in to optimize Q_r for r in $[r_{\min}, r_{\max}]$. Different values of r will eventually reveal different optimal partitions (found by heuristic algorithms to detect community structure) that represent intermediate topological scales of the complex network. We have applied this method to study the mesoscales in synthetic structured networks and real complex networks.

D. Validation of the method in synthetic networks

In Fig. 2 we have screened the whole range of topological scales for three synthetic networks, representing the number of modules obtained at the optimal partition for Q_r and plotting in a matrix the superposition of scales found. More precisely, any graphical representation of the whole mesoscale should take into account, for every pair of nodes, the frequency of mesoscales at which they belong to the same community. Each mesoscale has a natural *length* defined by the range of resistances $[r_{\rm from}, r_{\rm to}]$ at which it is optimal:

$$\text{length} = \log(r_{\rm to} - r_{\min}) - \log(r_{\rm from} - r_{\min}) \ . \qquad (37)$$

Thus, the length frequency for a pair of nodes is the sum of the lengths corresponding to mesoscales in which they belong to the same community, normalized by the total length. The graphical representation of this table is the *frequency mesoscales matrix*. First we have computed the modular structure in a hierarchical scale-free network with 125 nodes, RB 125, proposed by Ravasz and Barabasi[23]. We clearly observe persistent structures in 5 and 25 communities respectively, that account for the subdivisions more significant in the process, showing two hierarchical levels for the structure.

Another network example used is the H 13-4 network[24], which corresponds to a homogeneous in degree network with two predefined hierarchical levels, being 256 the number of nodes, 13 the number of links of each node with the most internal community (formed by 16 nodes), 4 the number of links with the most external community (four groups of 64 nodes), and 1 more link with any other node at random in the network. Both hierarchical levels are revealed by the method as they correspond to the original construction of the network: the first hierarchical level consisting in 4 groups of 64 nodes, and the second level consisting in 16 groups of 16 nodes.

Finally, we have used the FB network proposed by Fortunato and Barthélemy[19] to demonstrate the resolution limit of modularity (at $r = 0$). It consists in two cliques of 20 nodes linked with two small cliques of 5 nodes. At $r = 0$ the best partition cannot separate the two small cliques. We observe that the partition searched by the authors, formed by the four cliques isolated in their own communities, is obtained by increasing the resolution r, showing that the resolution limit of modularity is overcome by the method.

The optimization of modularity in all these cases has been performed using existing heuristics found in the literature[1,14,16] and compiled in a free toolbox available at the authors' webpage[25].

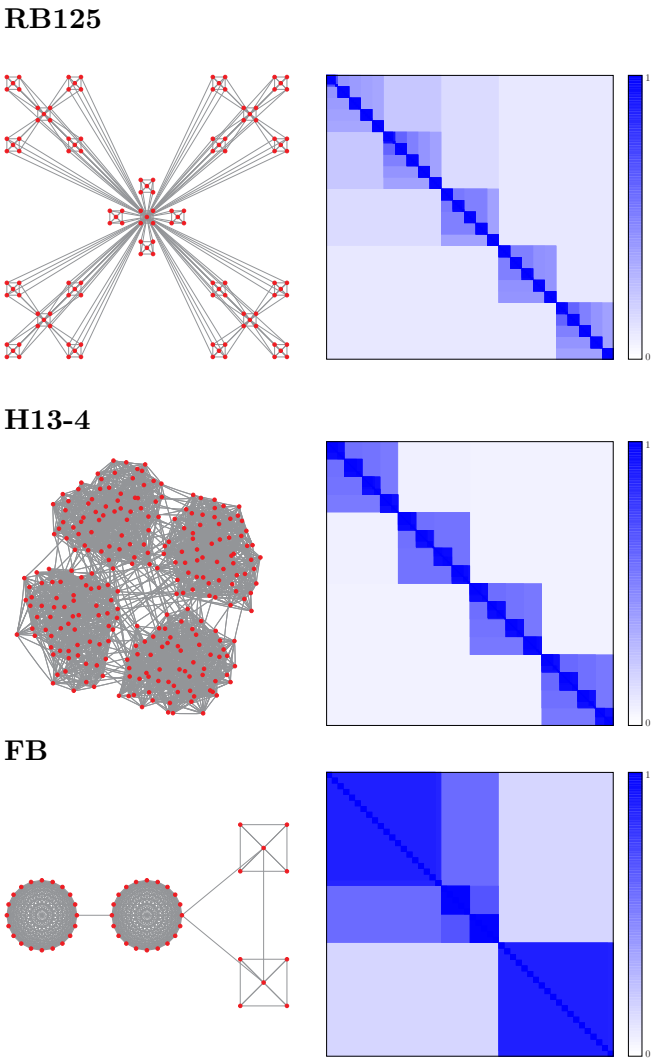

FIG. 2. Frequency mesoscales matrices in synthetic complex networks. We have computed the topological mesoscales for three synthetic networks. Left, we plot the networks and right we present their mesoscales matrices. The different color levels correspond to the superposition of the structures in r, which account for the persistence of the partitions revealed. See text for details.

III. APPLICATION TO EXPLORATORY DATA ANALYSIS

Exploratory data analysis stands for the approach to data analysis in which some rather general assumptions are used to reveal information of the data in a kind of inverse hypothesis testing. In our particular scenario, we will analyze the structure of the neural connectivity of the nematode *C. elegans*[26] using this approach. We do not pretend an exhaustive biological classification of all functionalities that are related to the topology but to show the applicability of the mesoscales analysis described before. A pretty exhaustive analysis of the same

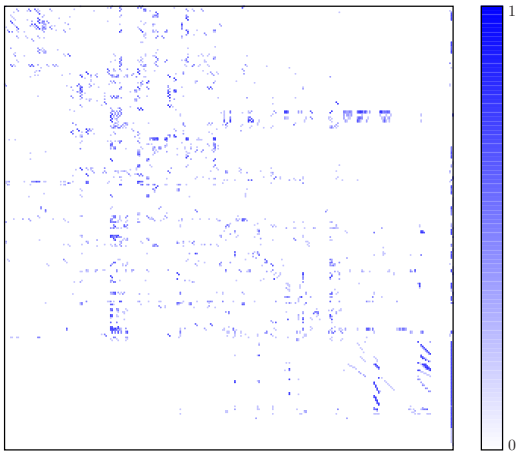

FIG. 3. Connectivity matrix of *C. elegans* neuronal network.

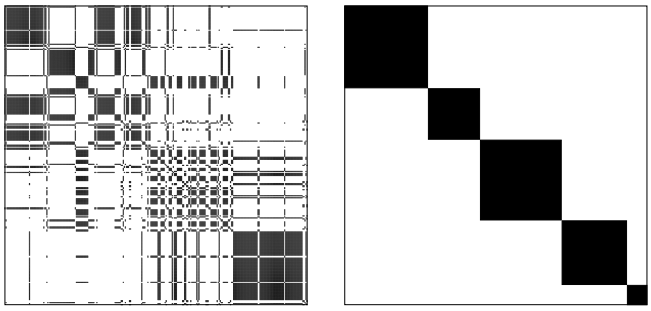

FIG. 4. Newman's scale of the *C. elegans* neuronal network. Left, original order, right, reordering by communities.

system has been recently presented[27] for the scale corresponding to $r = 0$. The whole nervous system of the nematode is composed by 302 neurons whose anatomical and connectivity description is completely known. The resulting network is represented as a weighted directed adjacency matrix, see Fig. 3. We will assume that those groups of nodes more persistent throughout the screening of the mesoscales of the topology have some functional role, and after we will look for this role in the current biological literature.

The original data[28] is a weighted and directed network, composed of 306 vertices (302 neurons + WE, WI, WM and WN) and 2359 arcs. We have discarded nine disconnected nodes from the network, the remaining 297 neurons form a single connected component and will be the subject of our analysis.

We have discretized the resistance range in 1000 non-uniform intervals, in such a way that the last resistance increment is ten times larger than the first one, and the size of the increments grow at a constant rate. The significant Newman's scale $r = 0$ has been added. The negative values of the resistance have been discarded, since we are interested only in sub-structure beyond the standard Newman's scale[29].

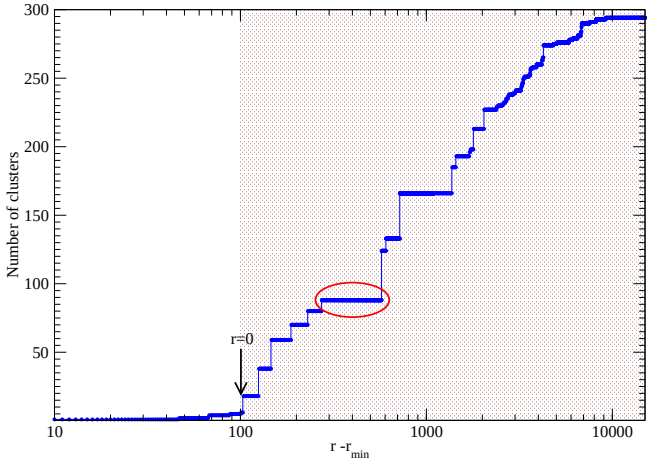

FIG. 5. Mesoscales of the C. elegans: number of clusters in the optimal partition at every value of the topological scale defined by the $\log(r - r_{\min})$, where $r_{\min}$ refers to the exact value, not its lower bound. Highlighted in circle, we represent the scale that most contributes to the frequency matrix.

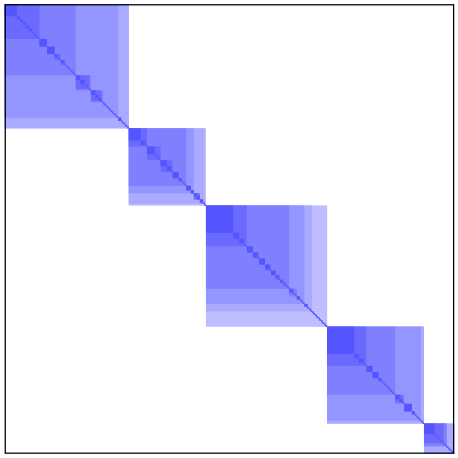

FIG. 6. Frequency matrix of *C. elegans* neuronal network thresholded at 0.6. We used a color scale (same as in Fig.3) to plot the persistence of neurons into the same groups, darker values corresponds to more persistent communities and, according to our hypothesis in the exploratory analysis, to specific functionalities

The order of the neurons in the matrix follows that in Watts and Strogatz[28] obtained from experimental data by White et al.[26]. The detection of the mesoscales in this neuronal system has been performed according to the method explained in the previous section. The best partition at $r = 0$ corresponding to the original Newman's scale provides with 5 communities. The representation of the obtained groups is depicted in Fig. 4 (left). This figure does not allow the observation of relevant information because the original order of the neurons in Fig. 3, however after ordering the neurons in the matrix by their communities, the representation shown in Fig. 4 emerges.

The coarse graining at $r = 0$ provides then with a large scale in this system, hence our interest has been specially focused in the sub-structural levels, not in supra-structural levels, that means that we have analyzed the mesoscale for $r \in [0, r_{\max}]$, see gray region of Fig. 5. We used the partition at $r = 0$ simply as a reference for sorting the neurons in the substructures found by the multiple resolution method.

Any trial of classification of the functional role of neurons of the *C. elegans* is extremely delicate because the multifunctional aspects they have. Many neurons participate in different synaptic pathways resulting in different functionalities. This property is also captured by our method that shows that at different scales the same neuron can appear in different groups, i.e. the method is not necessarily hierarchical. However, to extract information from the results obtained, we use an ensemble of the different partitions found by screening r, and construct a frequency mesoscales matrix, indicating the relative persistence of each neuron in a particular community. By fixing a threshold in the frequency value, we are able to unravel sub-structural scales that correspond to groups of neurons involved in different functionalities at different time scales.

The most interesting information is that provided at a large value of the frequency threshold, because in this case the substructures found will contain small groups of neurons whose activity response is topologically correlated, in particular the highlighted scales in Fig. 5 are the ones that most contribute to the frequency matrix. We have studied the ensemble frequency matrix at a threshold value of 0.6, Fig. 6, the lengths below the threshold are discarded, and the connected components of the graph defined by the remaining lengths are found. We have chosen this threshold fixing the sizes of the groups to be analyzed to be less than ten neurons. With this information at hand, and the wide description of each neuron found at the public database of *C. elegans*[30,31], we propose a tentative classification of some groups of neurons by functionality.

Our purpose, after identification of individual functionalities, has been to assign a specific action to the more persistent groups of neurons. The classification obtained (see appendix) does not pretend to be exact but to provide biologists with a useful information for future research.

IV. APPLICATION TO THE UNSUPERVISED CLASSIFICATION OF DATA

Unsupervised classification of data (or data clustering) stands for the process of grouping patterns of data according to their similarity. A pattern is a vector of features (usually understood as a point in a multidimensional space) that describes the item we wish to classify.

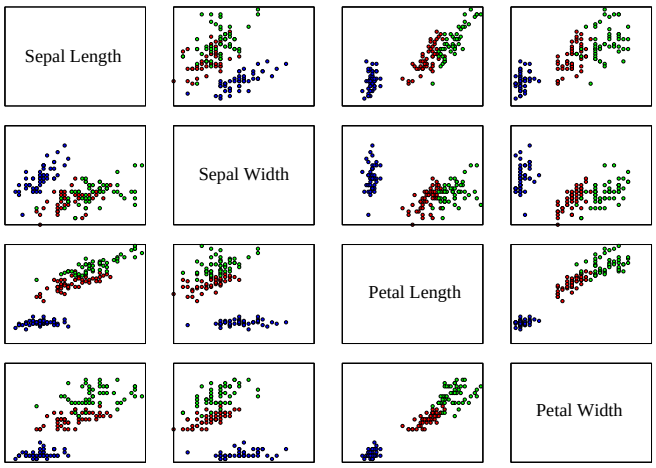

FIG. 7. Feature vectors for the Iris data set. Colors correspondence are: setosa-blue, versicolor-red, and virginica-green.

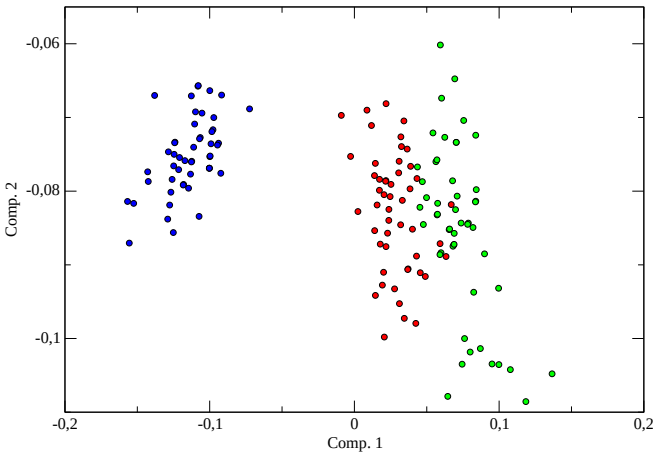

FIG. 8. Two principal components of the PCA analysis on the Iris dataset. Colors correspondence are: setosa-blue, versicolor-red, and virginica-green. The separation of pattern classes seems more clear in this projection.

The goal of the process of data clustering is to organize these patterns into groups, in such a way that patterns into the same group are more alike than with other patterns in other groups.

The problem of 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. Typical processes where data clustering is used are pattern analysis, decision-making, machine learning and image segmentation. These subjects have interesting applications as for example targeted marketing, biological taxonomy and detecting communities of interest in the World Wide Web[32].

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 dataset to a new one containing only the most relevant information. This first step is very important, as the result of the clustering often depends directly of the quality of it. 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 representation of the similarity[33]. Finally, the main step of the process, the grouping (or clustering) algorithm, which will decompose the similarity matrix and return the groups of data.

In our approach, the algorithm used to classify the similarity matrix is the multiple resolution algorithm based on modularity explained previously in this document. Given the nature of this algorithm, the result will not be a single partition into clusters, but a collection of different partitions. This fact deserves a reflection about how to evaluate the quality of the output obtained. If we make a screening between the minimum and maximum value of the resistance parameter to obtain every topological scale of resolution of the network, each one of these resolution levels will provide us with a partition of clusters. Then the question is, which one of these partitions is the right one? The answer is that every one of them is right, since what we are doing is analyzing the network at different levels of resolution, and all the information obtained though this process is found in the structure of the network. Having pointed that out, the problem of choosing the right partition is translated to that of choosing the more relevant partitions. The more relevant partitions in our scope are those that persist unchanged during larger intervals of values of the resistance parameter.

The dataset benchmark selected to perform the data clustering is the Iris flower dataset, presented by Sir Ronald Aylmer Fisher[34] in 1936. This dataset consists of 150 patterns corresponding to three different classes of flowers: Setosa, Versicolor and Virginica. Four features, the width and length of petal and sepal, form each pattern. Plots for the cross-variables and type of flowers are represented in Fig. 7. The unsupervised classification of this dataset is a major challenge in artificial intelligence and statistical theory, because of the patterns' organization, while one of the classes is linearly separable and then easily to classify by any elemental classification algorithm, the other two classes are not linearly separable and consequently far more difficult to classify.

Following the steps of data clustering explained above, we first performed a feature extraction/selection process. The idea here is simply to follow the workflow in any clustering problem, where the high dimensionality of the data and its redundancy is a main concern. In the partic-

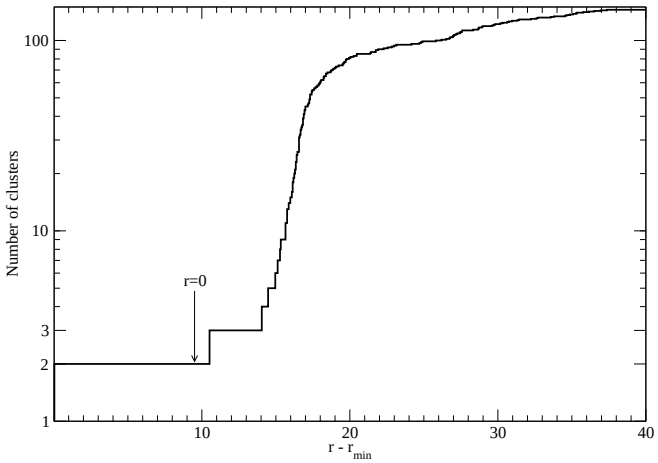

FIG. 9. Number of clusters as a function of the resolution parameter of the classification method (see text for details).

ular case we analyze, we can use all the original data with no computational stress, however we propose to address the feature extraction using PCA which will be the most common approach in many scenarios. We performed the 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. 8 a representation of these two components is shown. Based on these two variables, we propose to build up a similarity matrix as the euclidean distances between patterns components with respect to the center of mass of the data set 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 as a weighted network whose communities will, in principle, reproduce the right clustering of the data.

The results of the multiple resolution algorithm on the two main components of the Iris dataset is shown in the Fig. 9. It can be observed that the longest plateau in terms of the resistance interval values is that formed by those partitions that divide the dataset into two communities. This is not a surprising fact, as we know beforehand that one of the three classes of flowers is linearly separable, and then this partition makes totally sense, since there is one for the Setosa class and the other one containing the Versicolor and Virginica. However, the second longest plateau is the one formed by the three community partitions, and if we analyze the most resistant of them, we realize that it largely corresponds to the biological taxonomy of the flowers. To be specific, if we calculate the success as the number of correctly classified nodes divided by the total number of nodes, we achieve for the most resistant partition of three communities a 94,6% of success compared to the correct biological taxonomy.

Summarizing, we have presented a possible application of the multiple resolution method to the problem of data clustering. Our proposal has been proved competitive in success with other techniques used in the literature on the same benchmark[35], but as an essential difference we also provide information of grouping at different scales of resolution that are invisible to other algorithms. The methodology presented so far is plausible to be extensive to any data clustering problem expressed in terms of similarity matrices.

V. CONCLUSIONS

Scientists working on the field of complex networks have developed tools for the analysis of structural information embedded in the topological connectivity matrix. Specially interesting are the heuristic algorithms intended to find the community structure of networks, which remind the kind of problems of data clustering found in many disciplines.Here we have presented a possible application of community detection algorithms to help exploratory analysis and data clustering. In particular, we have used a previous methodology proposed by the authors that allows for a multiple resolution of topological scales in the substructure of networks.

The exploratory analysis of the neural connectivity of the nematode *C. elegans* has been presented. We found a tentative classification of groups of neurons presumably involved in specific tasks, according to the persistence of these groups in the topological analysis. We have also exposed the applicability of the method to the unsupervised classification of data, using the famous Iris dataset as a benchmark. The results are encouraging, we observe the full spectrum of clusters according to the organization of data, and the most persistent scales are those corresponding to well-known facts about its structure, a partition in two linearly separable groups, and a partition in three groups corresponding to the biological taxonomy. These results open the field of applicability of the theory of complex networks to other problems where the representation of data as a network allows the use of the technology developed so far.

Appendix A: Functional groups of C. elegans

Classification of functional groups of neurons resulting from the multiple resolution method. Using the database WormAtlas[30] and the results depicted in Fig. 6 we have identified nine groups of neurons of size lower than ten, whose functionality can be tentatively related to a specific action. The process to assign a tentative function to the groups of neurons has been done manually, reading the associated literature and using the worm-atlas database. We expose the list in Table I.

TABLE I. Temptative functionality of several significant groups of neurons found in the mesoscale.

Cluster of neurons	Tentative function
RIAL, RIAR, RMDR, RMDVR, SMDVR, RMDDL, SMDDR	Nose/Head orientation movement.
IL1DR, IL1VR, IL2DR, IL2VR, RIPR	Head-withdrawal reflex, more related to dorsal relaxation. When worms are touched on either the dorsal or ventral sides of their nose with an eyelash, they interrupt the normal pattern of foraging and undergo an aversive head-withdrawal reflex.
IL2, IL2R, OLQVL, OLQVR, RIH	Head-withdrawal reflex, more related to ventral relaxation.
ADLR, AIBR, ASEL, ASHR, AWCL, AWCR, AIAR, AIYL	Olfactory and thermosensation reflex.
ASGL, ASJL, ASKL, AIAL, PVQL	Chemotaxis to lysine reflex.
DB1, DB2, DD1, VB2, VD2, AS3, DA2, DA3, DA4, DA5	Backward sinusoidal movement of the worm, more related to touch stimulus.
AVAL, AVAR, AVBL, AVBR, AVDL, AVDR, AVEL, AVER, DA1, FLPL	Forward and Backward sinusoidal movement of the worm, more related to search for food in starving case, involve social feeding effect.
AVHL, AVHR, AVJL, AVFL, AVFR	Impossible to determine from the experimental data available. There is not any specific function known for any of these neurons.
AVKL, ACKR, PDEL, PDER, PVM, DVA, WN	The functionality of this group could be related to a relaxation state similar to a sleep state, with reduced motor activity, decreased sensory threshold, characteristic posture and easy reversibility, basically mediated by PDs neurons.

ACKNOWLEDGMENTS

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

[1] A. Arenas, A. Fernández, and S. Gómez, New J. Phys., **10**, 053039 (2008).
[2] S. H. Strogatz, Nature **410**, 268 (2001).
[3] C. M. Song, S. Havlin, and H. A. Makse, Nature **433**, 392 (2005).
[4] A.-L. Barabási, Science **308**, 639 (2005)
[5] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. USA **99**, 7821 (2002).
[6] R. Guimerà and L. A. N. Amaral, Nature **433**, 895 (2005).
[7] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[8] M. E. J. Newman, Phys. Rev. E **70**, 056131 (2004).
[9] E. T. Bell, Amer. Math. Monthly **41**, 411 (1934).
[10] U. Brandes, D. Delling, M. Gaertler, R. Goerke, M. Hoefer, Z. Nikoloski, and D. Wagner, IEEE Trans. Knowl. Data Eng., **20**, 172 (2008).
[11] M. E. J. Newman, Phys. Rev. E **69**, 066133 (2004).
[12] A. Clauset, M. E. J. Newman, and C. Moore, Phys. Rev. E **70**, 066111 (2004).
[13] R. Guimerà and L. A. N. Amaral, J. Stat. Mech., P02001 (2005).
[14] J. Duch and A. Arenas, Phys. Rev. E **72**, 027104 (2005).
[15] J. M. Pujol, J. Béjar, and J. Delgado, Phys. Rev. E **74**, 016107 (2006).
[16] M. E. J. Newman, Proc. Natl. Acad. Sci. USA **103**, 8577 (2006).
[17] S. Fortunato, Phys. Rep. **486**, 75 (2010).
[18] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas, J. Stat. Mech., P09008 (2005).
[19] S. Fortunato and M. Barthélemy, Proc. Natl. Acad. Sci. USA **104**, 36 (2007).
[20] B. H. Good, Y-Al. Montjoye, and A. Clauset, Phys. Rev. E **81**, 046106 (2010).
[21] S. Gómez, P. Jensen, and A. Arenas, Phys. Rev. E **80**, 016114 (2009).
[22] A. Arenas, J. Duch, A. Fernández, and S. Gómez, New J. Phys. **9**, 176 (2007).
[23] E. Ravasz and A.-L. Barabasi, Phys. Rev. E **67**, 026112 (2003).
[24] A. Arenas, A. Díaz-Guilera, and C. J. Perez-Vicente, Physica D **224** 27 (2006).
[25] http://deim.urv.cat/ sgomez/radatools.php Toolbox for community detection.
[26] J. G. White, E. Southgate, J. N. Thompson, and S. Brenner, Phil. Trans. Royal Soc. London. Series B **314**, 1 (1986).
[27] R.K. Pan, N. Chatterjee and S. Sinha, PLoS ONE **5**(2): e9240 (2010).
[28] D. J. Watts and S. H. Strogatz, Nature **393**, 440 (1998).
[29] A. Arenas, A. Fernández, and S. Gómez, Lect. Notes Comp. Sci., **5151**, 9 (2008).
[30] Z. F. Altun and d. H. Hall, (ed.s). WormAtlas. 2002-2006. http://www.wormatlas.org
[31] R. M. Durbin, *Studies on the Development and Organisation of the Nervous System of Caenorhabditis elegans*, PhD Thesis, University of Cambridge (1987).
[32] G. Gan, C. Ma, and J. Wu, *Data Clustering: Theory, Algorithms, and Applications*, ASA-SIAM Series on Statistics and Applied Probability, **20**, (2007).
[33] A. K. Jain, M. N. Murty, and P. J. Flynn, ACM Comp. Surv., **31**, 3 (1999).
[34] R. A., Fisher, Annals of Eugenics **7**, 179 (1936).
[35] S. K. Pal and J. Basak, IEEE Trans. Neural Nets. **11** (2), 366 (2000).