Community detection in networks

C.O. Dorso* and A.D. Medus
Departamento de Física, Facultad de Ciencias Exactas y Naturales,
Universidad de Buenos Aires, Pabellón 1, Ciudad Universitaria,
Ciudad Autónoma de Buenos Aires (1428), Argentina
(Dated: March 30, 2009)

The problem of community detection is relevant in many disciplines of science. A community is usually defined, in a qualitative way, as a subset of nodes of a network which are more connected among themselves than to the rest of the network. In this article we introduce a new method for community detection in complex networks. We define new merit factors based on the weak and strong community definitions formulated by Radicchi et al (*Proc. Nat. Acad. Sci. USA* **101**, 2658-2663 (2004)) and we show that this local definitions properly describe the communities observed experimentally in two typical social networks.

I. INTRODUCTION

The study of networks (a set of nodes interconnected by links) has become a ubiquitous topic in many branches of science. This is because many systems of interest can be represented in this way, as for example, Internet, the WWW, food webs, neural networks, communication networks , social networks etc. Many different properties have been revealed as: small world effect, high network transitivity, power law degree distributions, etc.(for a recent review on these topics see [Boccaletti *et al.*, 2006])

In this work we focus on one of these properties, the so called community structure. Community structure is defined, in a qualitative way, as the possibility of recognizing within the networks, subsets of nodes which are more connected among themselves that to the rest of the network.

If we can detect such structures we will get information of practical importance. Such groups in the WWW might correspond to sets of web pages on related topics, in the case of social networks they would indicate groups that share interests, problems etc. In a metabolic network it might help to identify groups of nodes which perform different functions.

It is quite interesting that based only on such a qualitative characterization of the communities, many methods of detecting them have been developed. Among these methods one has become the most popular, the one developed by Newman and Girvan [Newman & Girvan, 2004] (hereafter referred as I, which we analyze in section II)

Only recently quantitative definitions of community have been put forward by Radicchi et al. which capture the qualitative one. In [Radicchi *et al.*, 2004] the authors have defined two kinds of communities, the ones in strong sense and the ones in weak sense (see section V).

In this communication we will explore the problem of the detection of communities. In section II we analyze the properties of the method proposed by Newman and Girvan which is based on the maximization of a merit factor named modularity (Q_N). In section III we will examine the divisive approach for the maximization of Q_N and we will show that it is not capable of finding the optimal partition due to the inability of such an approach to explore the complete set of partitions. We show that using a strategy in the spirit of simulated annealing this problem can be circumvented. In Section IV we examine the meaning of the communities obtained using the maximization of Q_N . In section V we will review the quantitative definitions of Raddicchi et. al. and, based on these definitions, we will propose new merit factor to calculate the goodness of a given partition of a network in communes. In section VI we will show the results of our method with two examples , well known in the literature, the Zachary Karate club network and the Bottle nose dolphins networks. Finally conclusions are drawn.

II. MODULARITY (Q_N)

A. Community structures in networks

A network **G** is defined by a set of nodes $\{n\}$, (n_1, n_2,...,n_n.), and a set of links $\{l\}$, ($l_{12}, l_{14}..., l_{km}$). A link l_{ij} denotes a relation between node n_i and node n_j. Depending on the possible values of l_{ij} the resulting network can be of two types. If l_{ij} can only have the values 1 or 0 we will call the network unweighted, otherwise it will be referred as weighted. In this work we will focus on unweighted networks. We will consider networks such that for every conceivable pair of nodes there is a path (i.e. a sequence of links $\{l_{ij}l_{jk}l_{km}...\}$) joining them, in such a case we say that we are dealing with connected networks. We will consider that the links are undirected i.e. $l_{ij} = l_{ji}$. We will focus on sparse networks for which the number of links in $\{l\}$, N_l , is much less than the maximum possible number of links, $N_{l_{\max}}$ with $N_{l_{\max}} = n_n(n_n - 1)/2$, with n_n the total number

*Electronic address: codorso@df.uba.ar; Also member ofCarrera del Investigador Científico de CONICET

of nodes in $\{n\}$. The associated adjacency matrix M is defined as $m_{ij} = l_{ij}$.

The distance between two nodes d_{ij} will be defined as number of links that are to be traversed, when we move from i to j along the minimum path joining them.

Given the network **G** we will define a partition **P** as a given grouping of the nodes in subsets p_i $(1 \leq i \leq g)$, while keeping the structure of the adjacency matrix unaltered.

Following I we will quantify the degree of communality of a given partition **P** in the following way:

Given a m-subgraphs partition $\{C_j\}_{1\leq j\leq m}$ of the graph G, where $\bigcup_{j=1}^{m} C_j = G$, the mathematical expression of Q_N is :

$$Q_N = \sum_{i=1}^{m} \left[\frac{l_i}{L} - \left(\frac{d_i}{2L} \right)^2 \right] \tag{1}$$

where l_i denotes the total number of internal links for subgraph $C_i \subset G$, $d_i = \sum_{j\in C_i} k_j$, and $L = \frac{1}{2}\sum_{j\in G} k_j$ is the total number of links in G.

The term l_i/L in Eq 1 denotes the actual fraction of internal links in subgraph C_i, while $d_i/2L$ can be interpreted as the probability of a link to be connected to some node in subgraph C_i. Then, $(d_i/2L)^2$ constitutes the expected fraction of links within subgraph C_i when all nodes in G are randomly connected, keeping the degree of the nodes fixed. This last ideal random picture is used to compare with the actual one because it is assumed that corresponds to a situation with no communities (although it was shown in [Guimerà *et al.*, 2004] that random networks may have a community structure).

It is then proposed that, if the network under consideration has no community structure, Q_N equals 0. On the other hand, if the network under consideration does have a community structure, the closer the chosen partition is to the actual community structure of the network, the larger the modularity Q_N will be.

In this way, the search of community structures in networks is reduced to finding the partition **P** which maximizes the modularity Q_N.

We should notice that this merit factor implies, in turn, a community definition (which does not necessarily corresponds to the intuitive one stated above): a subgraph C_j will be a community if the actual number of links that connects nodes in C_j is bigger than the expected one when all nodes in the network are randomly connected, this is to say, when $l_i/L - (d_i/2L)^2 > 0$. Clearly this last condition depends on the global parameter L Then, we say that the community definition associated with Q_N is *non-local*

III. COMMUNITY RECOGNITION ALGORITHMS

In this section we review the algorithm presented in I, based on edge removal (hereafter referred as ER), and describe our approach based on Simulated Annealing (hereafter referred as SA) for the maximization of Q_N .

A. Community recognition via edge removal

Newman and Girvan [Newman & Girvan, 2004] have proposed to study the structure of the network by analyzing the effect of the removal of links with highest betweenness. The betweenness b_{ij} of a given link l_{ij} is :

$$b_{ij} = \sum_{paths} \alpha_{no}^{-1} \sum_{l_{km}\epsilon path_{no}} \delta(l_{ij} - l_{km}) \tag{2}$$

with $\sum_{paths}$ the sum over all paths joining the n_n nodes, α_{no} is the degeneracy of the path between nodes n and o, and $\sum_{l_{km}\epsilon path_{no}}$is the sum over all the links l_{km} that form the path under consideration. In this way the link with highest betweenness is the one that appears most often when we study all the components of all the minimum paths between all pairs of nodes.

According to this prescription:

i)One calculates the betweenness of all the links in the network. ii) The one with the highest betweenness is removed.

The process is continued until a disjoint cluster is obtained. Afterwards,the same procedure is applied to each of the resulting subgraphs.

Special care is to be taken when the highest betweenness is degenerate. Because it is not possible to foresee which will be the optimum cut, we should select at random the link to be removed.

In this way, partitions with $2, 3, ..., N^{'}$ subsets can be obtained. The best one, according to the discussion in the previous section, is the one that maximizes the magnitude Q_N.

B. Simulated Annealing Analysis

We now describe a methodology for the calculation of the maximum modularity resorting to a Simulated Annealing [Dorso & Randrup, 1993] calculation in the space of the partitions of the network under analysis. Simulated Annealing is a generalization of the well known Metropolis Monte Carlo (MMC) procedure. MMC consists in the realization of a Markov Chain in the space of the configurations of the system according to certain transition probabilities chosen in such a way that the asymptotic frequency of each state satisfies the Boltzmann distribution $exp(-\beta E_i)/Z$ with $\beta = (1/kT)$ where T is the Temperature of the system , E_i the energy of state i and Z the canonical partition function. The transition probability q_{ij} reads:

$$q_{ij} = \min(1, \exp(-\beta(E_j - E_i))$$

In Simulated Annealing the same procedure is employed but instead of using the Temperature of the system we use a pseudo Temperature, τ, which controls the behavior of the transition probability and instead of the energy, the observable that we want to maximize. The pseudo temperature τ is monotonously lowered until an extremum of the relevant observable is attained. In our case the Markov Chain is performed in the space of the partitions of the network under consideration. The transition probabilities read, in our case, $q_{ij} = \min(1, \exp(-\beta'(Q_j - Q_i))$ with $\beta' = 1/\tau$ and E_k has been replaced by Q_k, the modularity of partition k. Moreover, because we are looking for the maximum of the modularity $(Q_{Nj} - Q_{Ni})$ stands for $(Q_{Ninitial} - Q_{Nfinal})$.

IV. CASE STUDY

In order to check the properties of the two approaches above mentioned, we have found it helpful to analyze the following simple undirected graph Fig.1A). The advantage of dealing with such a small and simple graph is that the calculations can be performed by hand and the properties of the recognition algorithms can be easily understood.

In Fig.1) we show the comparison between the results obtained with the above mentioned algorithms (see figure captions for details).

We first analyze what happens when we apply the edge removal (ER) approach:

1)We search for the links with highest betweenness, in this case there is degeneration and links $l_{10,11}$, $l_{10,12}$, $l_{12,13}$, $l_{11,13}$, stand on an equal footing. We then choose one at random and remove it. In our example we choose $l_{12,13}$ obtaining the graph displayed in Fig.1B).

2) We repeat step 1) and we find that the edge with highest betweenness is $l_{10,11}$. It should be noticed that as a consequence of removing this link the graph breaks up in two pieces Fig.1C). The value of Q is in this case $Q = 0.409$.

As we continue in this way we will obtain that the next breaking of the network takes place when removing link $l_{10,12}$. By removing this link we obtain 3 clusters with a modularity value of $Q = 0.405$. Notice that the removal of $l_{11,13}$ is equivalent to removing $l_{10,12}$, giving a different graph with the same value of Q.

We now apply the SA approach. i)If no restriction on the number of partitions is imposed, we obtain the result displayed in Fig.1E). In this case the original network is broken into 3 subsets with a modularity value of $Q = 0.446$, ii)If, on the other hand, we restrict the number of partitions to two, we obtain the result displayed in Fig.1D) , which is the same graph as the one obtained using ER for two subsets (of course the equivalent configuration resulting from the removal of $l_{10,12}$ and $l_{11,13}$ can be obtained as well)

It is relevant to notice that the best result according to SA cannot be reached using ER, because in order to get the graph displayed in Fig.1E), from the previous step in the calculation (Fig.1C), the link $l_{10,11}$ must be reconstructed, but this step is not allowed in the ER methodology.

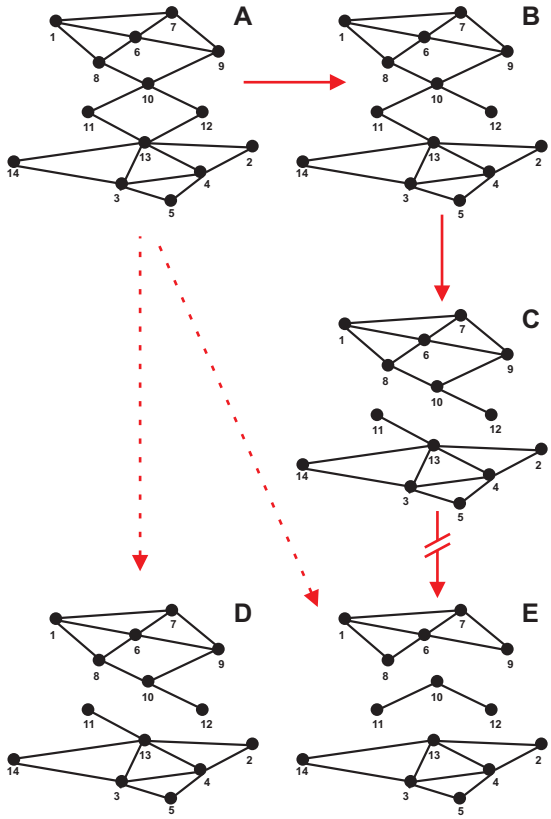

FIG. 1: Development of community structures in terms of the ER and SA analysis. Full arrows denote steps in the ER approach. Dotted arrows denote results from SA methodology. Starting from network A by applying ER methodology we first get to network B and, after the second removal of a link, to network C. On the other hand, starting from the same initial network the SA will give network D if we impose the constraint that the final configuration should display two communes. If we do not impose any constraint the result according to SA will be network E. It is important to notice that network E is unreachable from network C. This is the main drawback of the ER approach.

From this analysis it is clear that the SA algorithm is able to find a better (as measured by the quantity Q) solution to the communality analysis than the ER criterion.

The reason why the ER approach fails to reach the best result is because this methodology is local and irreversible. On the other hand, when we analyze the sequence of results obtained with SA when we impose the condition of having 2, 3, 4,... partitions, we see that in going from two partitions to three partitions the link $l_{10,11}$ appears again. This is no problem in SA because we are working with different groupings of the nodes and all the information about the links is conserved at all times.

V. QUANTITATIVE COMMUNITY DEFINITIONS

In order to formalize the qualitative definition of community stated in the Introduction, we consider a graph G containing N nodes, with k_i the degree of node $i \in G$. If C is a subgraph of G with k_i^{in} and k_i^{out} the number of links of node $i \in C$ that connect it to nodes inside and outside of C respectively. There are two quantitative community definitions introduced by Radicchi et al [Radicchi *et al.*, 2004]:

i) Community in strong sense: *C is a community in strong sense if:*

$$k_i^{in} > k_i^{out} \quad \forall i \in C \tag{3}$$

ii) Community in weak sense: *C is a community in weak sense if:*

$$\sum_{i \in C} k_i^{in} > \sum_{i \in C} k_i^{out} \tag{4}$$

In words: a subgraph $C \subset G$ will be a community in the strong sense if each of its nodes has more links connecting it with nodes in C than those that connect it with other nodes not belonging to C. In the similar way, $C \subset G$ will be a community in the weak sense if the sum of the numbers of links that interconnect nodes inside C is larger than the sum of all links that connect nodes in C with nodes not belonging to C. These community definitions are simple, intuitive and *local*: given a subgraph $C \subset G$ we can decide if it constitutes a community, in either strong or weak sense, without knowledge of the entire structure of G.

A. Merit factors for the weak and strong community definitions.

Given a graph G and a m-subgraphs partition $\{C_j\}_{1 \leq j \leq m}$, where each subgraph $C_j \subset G$ constitutes a community according to any of the local definitions mentioned in the previous section, we want to define a quantity that measures the "quality" of each of the resulting communities. In the context of the above mentioned local framework, this quantity must only depend on the local characteristics of the subgraph C_j. Following the weak and strong definitions of community, the more internal links a community has, with respect to the external ones, the "stronger" it will be. If $k_i = k_i^{in} + k_i^{out}$ is the degree of node $i \in C_j$, where k_i^{in} and k_i^{ext} are the number of internal and external links for node i, we define the *"community strength"* (S) that measures the normalized difference between the number of internal and external links for nodes in C_j:

$$S(C_j) = \sum_{i \in C_j} \frac{k_i^{in} - k_i^{out}}{2L(C_j)} \tag{5}$$

were $L(C_j) = \frac{1}{2}\sum_{i \in C_j} k_i$. Then, $-1 \leq S(C_j) \leq 1$, and it achieves its maximum value 1 when $k_i^{out} = 0$, $\forall i \in C_j$, i.e., when the subgraph C_j is isolated. C_j will be a community in weak sense if $S(C_j) > 0$.

The definition of Eq. 5 is valid for unweighted networks, but it can be extended to the case with weighted links. In such a case, we must interpret k_i as the sum of the weights of the links that connect to node i, for both k_i^{in} and k_i^{out}.

We now introduce the merit factor Q_W for the weak community definition as the sum of $S(C_j)$ over all subgraphs $C_j \subset G$:

$$Q_W = \sum_{j=1}^{m} S(C_j) = \sum_{j=1}^{m} \sum_{i \in C_j} \frac{k_i^{in} - k_i^{out}}{2L(C_j)} \tag{6}$$

As in the case of Q_N: the bigger Q_W is, the better the m-subgraphs partition $\{C_j\}_{1 \leq j \leq m}$ of G will be, in the sense of weak community definition. Then, it is possible to implement the optimization algorithms developed for Q_N for this new merit factor Q_W. Because, each subgraph $C_j \subset \{C_j\}_{1 \leq j \leq m}$ must satisfy the weak community definition we include an extra constraint into the optimization process:

$$S(C_j) > 0 \quad \forall C_j \subset \{C_j\}_{1 \leq j \leq m} \tag{7}$$

Now, our definition of optimal partition can be stated in the following way:

Definition: *the optimal m-subgraphs partition $\{C_j\}_{1 \leq j \leq m}$ of a graph G in the weak sense is that one with maximal merit factor $Q_W = \sum_j S(C_j)$, such that $S(C_j) > 0$, $\forall C_j \subset \{C_j\}_{1 \leq j \leq m}$.*

In the same spirit we now define a merit factor Q_S for the strong community definition:

$$Q_S = \sum_{j=1}^{m} S(C_j) = \sum_{j=1}^{m} \sum_{i \in C_j} \frac{k_i^{in} - k_i^{out}}{2L(C_j)} \tag{8}$$

with the constraint

$$(k_i^{in} - k_i^{out}) > 0 \ ::: \ \forall i \in C_j \tag{9}$$

Definition: *the optimal m-subgraphs partition $\{C_j\}_{1 \leq j \leq m}$ of a graph G in the strong sense is that one with maximal merit factor $Q_S = \sum_j S(C_j)$, such that $(k_i^{in} - k_i^{out}) > 0$, $\forall\ i \in C_j \subset \{C_j\}_{1 \leq j \leq m}$.*

In next section we will show some application examples of this new merit factors in networks partition problems.

VI. EXAMPLES

A. Zachary's karate club network.

In all examples presented in this section we have used a optimization algorithm based in simulated annealing,

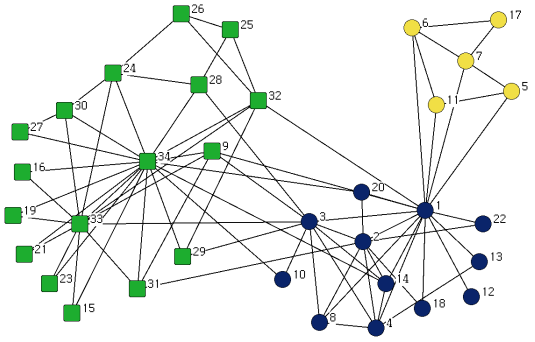

FIG. 2: Best partition for Zachary network (color online). Squares and circles denote the two communities obtained with our approach when the number of communities are fixed in two. This partition perfectly corresponds to the one consigned by Zachary in [Zachary, 1977]

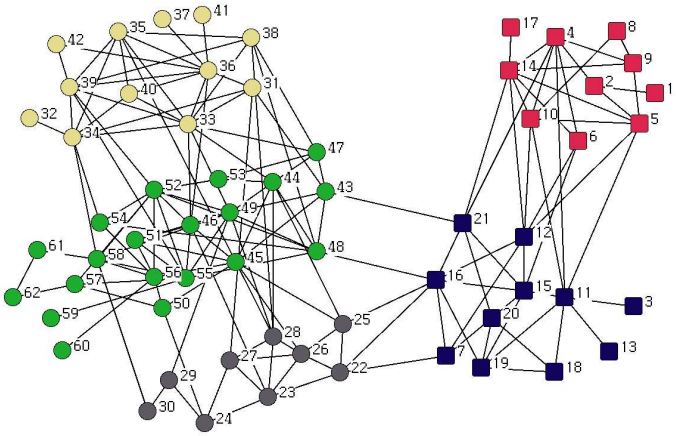

FIG. 3: Bottlenose dolphin network. This network has a size of 62 nodes and it is known from direct observation that it has two communities. In this figure squares and tringles denote the communities detected by our Strong Community approach and the colors (shades of gray) show the results of the weak community approach. Notice that the optimization according to Q_w merely subdivides the communities obtained through Q_s optimization.

described in previous sections, but for our new merit factors.

The Zachary's Karate Club network [Zachary, 1977], has turned into an unavoidable example in publications about community structure. This network represents the relationships between members of a karate club at a University in the 1970s, and it has been shown that it has a strong community structure in previous studies [Newman & Girvan, 2004 ;Medus *et al.*, 2005]. Applying the optimization algorithm for the weak community definition merit factor Q_W, we obtained, for the unweighted version of Zachary network, a partition in three communities of 17 (C_1), 12(C_2) and 5 (C_3) nodes with $Q_W = 1.792$ (Fig. 2). When the number of communities was restricted to two, we obtained the actual partition in two communities of 17 nodes each one observed by Zachary, with $Q_W = 1.487$ (circles and squares in Fig. 2).

With this analysis we can know, in addition, the strength $S(C_j)$ of each community C_j in the network. For the best partition of Zachary network in three communities of 17 (C_1), 12 (C_2) and 5 (C_3) nodes, we have: $S(C_1) = 0.744$, $S(C_2) = 0.548$ and $S(C_3) = 0.0.5$, with C_1 as the strongest community. On the other hand, the partition into two communities, is composed by two strong communities of 17 nodes each one, with $S(C) = 0.744$. .

When we performed the community analysis using the strong community merit factor Q_S, we obtained two communities: C_1 with 29 nodes ($S(C_1) = 0.943$) and C_2 with 5 nodes ($S(C_2) = 0.5$). In Fig. 2 can be observed that node 10 has one internal and one external link and this situation can not be allowed in strong community definition. For this reason, the communities with 17 and 12 nodes are joined together.

B. The bottlenose dolphins network

Another social network which has attracted considerable interest is the one corresponding to the bottlenose dolphins network, which has been fully analyzed in [Lusseau, 2003] . This social network is composed by 62 nodes and it is known to consist of two communities of sizes 41 and 21 nodes each. Following the approach proposed in this work we first analyze this network applying the Q_N analysis in our Simulated annealing approach. The result of this analysis is the partition of the network in four communes composed by 21, 16, 13, and 12 nodes each. When we perform the optimization of the Weak community definition we obtain five communities of 20, 12, 11, 10, 9 nodes each. Finally when the dolphin network is analyzed in terms of the Strong community definition we obtain the actual partition in two communities of 41 and 21 nodes each. These last two results are displayed in Fig. 3. In this figure we show the two communities according to the Strong community definition as triangles (41 nodes community) and as squares (21 nodes community). The corresponding analysis according to the Weak community definition further divides the previous two communities and are denoted by the different shades of gray (see caption for details). It is interesting to notice that all the communities detected by the optimization of Q_N are communities in the weak sense but the resulting partition is suboptimal.

VII. CONCLUSIONS

In this work we have proposed new merit factors to recognize communities in networks. These merit factors are more realistic that the ones currently in use in the literature because they strictly adhere to what a community is expected to be, i.e., a subset of nodes which are more connected among themselves than to the rest of the network under consideration.

We started by putting forward this qualitative definition of a community and then we reviewed the meaning of the quite popular measure of the quality of a given partition known as the modularity Q_N. As we have discussed above, the community definition associated to this quantity is non-local and does not necessarily corresponds to the aforementioned qualitative definition. One of the consequences of the non-local character intrinsic to this quantity is the limit resolution problem as stated in [Fortunato & Barthélemy, 2007].

In order to recognize communities in networks that strictly adhere to the qualitative definition, we have used (following [Radicchi *et al.*, 2004]) two local community definitions: weak and strong . In order to use this definitions to recognize communities we have developed a criteria to quantify the strength of a community (S). Afterwards, we have defined two merit factors associated with S which we named Q_S,and Q_W. As with Q_N the problem of recognizing communities in a network is mapped onto a optimization problem, i.e., the communities in a network are the elements of the partition which maximizes Q_S,or Q_W. We have performed the optimization of these merit factors on some standard networks by implementing an algorithm in the spirit of simulated annealing. The limit resolution intrinsic to the Q_N definition is not present in our approach.

It is worth noticing at this point that the solution to the detection of communities in the strong sense is also a solution in the weak sense but not necessarily optimal. On the other hand, the converse is generally not true.

The strong community definition tends to give larger communities because of its inability to deal with nodes that are equally shared by two highly connected subgraphs, but on the other hand has the nice property that it is the only one that gives no partition for symmetric string networks and , for example, solves exactly the bottle nose dolphins network.

We finally note that the main purpose of this work was to show the properties of our new definitions of the strength of communities, according to which the community recognition problem is transformed into a optimization one. The method used to solve the resulting optimization problem (simulated Annealing in partition space) is quite powerful but intrinsically slow. If very big networks, comprising millions of nodes, are to be analyzed with our definitions, new, faster, approaches are to be devised.

Acknowledgments

C.O.Dorso acknowledges partial support from CONICET through grant PIP5969.

References

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., & Hwang, U. [2006] "Complex networks: Structure and dynamics," *Physics Reports***424**,175-308.

Dorso, C.O. & Randrup, J. [1993] "Early recognition of clusters in molecular dynamics," *Phys. Lett. B* **301**,328-332.

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

Fortunato, S., & M. Barthélemy, M. [2007] "Resolution limit in community detection," *Proc. Nat. Acad. Sci. USA* **104**, 36-41.

Guimerà, R., Sales Pardo, M. & Amaral, L.A.N. [2004] "Modularity from fluctuations in random graphs and complex networks," it Phys. Rev. E **70**, 025101.

Karrer, B., Levina, E., & Newman, M.E.J. [2008] "Robustness of community structure in networks," *Phys.Rev. E* **77**, 046119.

Lusseau, D. [2003] "The emergent properties of a dolphin social network," *Proc. of the Royal Society B:Biological Sciences* **270** S186.

Medus, A., Acuña, G. & Dorso C. O. [2005] "Detection of community structures in networks via global optimization," *Physica A* **358**, 593-604.

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

Radicchi,F., Castellano, C., Cecconi, F., Loreto, V. & Parisi, D. [2004] "Defining and identifying communities in networks," *Proc. Nat. Acad. Sci. USA* **101**, 2658-2663.

Zachary, W.W. [1997] "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research* **33**, 452-473.

All figures have been drawn using NetDraw http://www.analytictech.com.