Available online at www.sciencedirect.com

SCIENCE DIRECT®

Physica A 358 (2005) 593–604

PHYSICA A

www.elsevier.com/locate/physa

Detection of community structures in networks via global optimization[☆]

A. Medus, G. Acuña, C.O. Dorso[*]

Departamento de Física-Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Pabellón 1 Ciudad Universitaria, 1428 Buenos Aires, Argentina

Received 12 November 2004
Available online 2 June 2005

Abstract

We present an analysis of communality structure in networks based on the application of simulated annealing techniques. In this case we use as "cost function" the already introduced modularity Q (1), which is based on the relative number of links within a commune against the number of links that would correspond in case the links were distributed randomly. We compare the results of our approach against other methodologies based on betweenness analysis and show that in all cases a better community structure can be attained.
© 2005 Elsevier B.V. All rights reserved.

Keywords: Networks; Communality; Betweenness

1. Introduction

The analysis of different properties of networks have recently attracted the attention of researchers in different areas. Among them we recall the emergence of Small world effect [1], degree distribution of the nodes [2], etc.

In this communication we will focus in the recently addressed problem of detecting "community structures" in networks. This means that for many networks in nature,

[☆] Partially supported by the University of Buenos Aires via Grant X139.

[*] Corresponding author.
E-mail address: codorso@df.uba.ar (C.O. Dorso).

doi:10.1016/j.physa.2005.04.022

the nodes composing them, can be divided into groups such that the connections within each group are dense, while connections between groups are sparse. The analysis presented in this communication has roots both in the field of networks and the problem of phase transitions in small systems. For the first we recall that the problem of communality has been investigated by Newman and Girvan [3] (hereafter referred as I) who have performed an already extensive set of studies in this field [4]. In their work both a parameter for measuring the merit of a given partition of a graph in communities has been given, which we will use in our approach, and a method for calculating the community structures was devised. We will critically reanalyze the properties of such a method. On the other hand we recall a series of works in which one of us has investigated the formation of cluster in equilibrated [5] and out of equilibrium fragmentation [6,7]. In this case also a parameter for the analysis of the merit of a given grouping of particles in clusters was given and a method based on global optimization (using simulated annealing) of such a quantity was given [8].

In Section 2 we briefly review fundamental definitions in the field of networks and we describe the parameter Q which will be used to quantify the merit of a given partition of the graph under analysis into communities. In Section 3 we describe the community recognition algorithms used in this work. In Section 4 we study the properties of a simple graph and unveil the main properties of the algorithms described in the previous section. In Section 5 we present the analysis of more complex networks already analyzed in the literature. Finally, conclusions are presented.

2. Community structures in networks

A network $\mathbf{N}$ is defined by a set of nodes $\{n\}$ $(n_1, n_2, \ldots, n_n)$, and a set of links $\{l\}$ $(l_{12}, l_{14}, \ldots, l_{km})$. 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, on the other hand a network will be defined as weighted if l_{ij} can attain values $0, 1, 2, 3\ldots$ thus indicating that the relation between nodes is also characterized by a given strength. In most of this work we will focus on unweighted networks. We will assume that for every node n_i there exists at least another node n_j such that l_{ij} is different from 0, moreover we will consider networks such that for every conceivable pair of nodes there will be a path (i.e., a sequence of links $\{l_{ij} l_{jk} l_{km} \ldots\}$) 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., $l_{ij} = l_{ji}$. Further on, 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, $N_{l_{\max}}$ given by $N_{l_{\max}} = n_n(n_n - 1)/2$, with n_n the total number 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 the length, or number of links that are to be traversed, when we move from i to j along the minimum path joining them.

Given the network **N** we will define a partition **P** as a given grouping of the nodes in subsets p_i $(1 \leqslant i \leqslant g)$, 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 partition **P** comprising g subsets, a matrix **e** (of dimension $g * g$) is defined such that the corresponding component e_{ij} is the fraction of edges in the original network that connect nodes in subset i with nodes in subset j. In *I* the *modularity Q* is defined as

$$Q = \sum_i e_{ii} - \sum_{ijk} e_{ij} e_{ki} = Tr\mathbf{e} - |\mathbf{e}^2| \, . \tag{1}$$

Q stands for the difference between the relative quantity of links within subsets and the expected relative number of links that would result from a random placement of links when no attention is given to the community structure of the network under consideration [9].

If the network under consideration has no community structure, Q 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 will be.

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

3. Community recognition algorithms

In this section we review the algorithm presented in I based on edge removal (hereafter referred as edge removal (ER)), and describe our approach based on simulated annealing (hereafter referred as SA).

3.1. Community recognition via edge removal

In a recent work, Newman and Girvan [3] 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

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

with $\sum_{paths}$ the sum over all the path joining the n_n nodes. α_{no} is the degeneracy of the path between nodes n and o, and $\sum_{l_{km} \varepsilon path_{no}}$ 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 the 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, it 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 $2, 3, \ldots, N'$ subsets can be obtained. The best one, according to the discussion in the previous section, is the one that maximizes the magnitude Q.

3.2. Simulated annealing analysis

In this section we present a methodology to study the community structure in networks based on the search for that partition that maximizes the value of Q. This is accomplished by resorting to a SA [10] calculation in the space of the partitions of the network under analysis. SA 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 SA (see [10] for details) 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 $q_{ij} = \min(1, \exp(-\beta'(Q_j - Q_i))$ 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 $(Q_j - Q_i)$ stands for $(Q_{initial} - Q_{final})$.

The Markov Chain is implemented in the following way:

We start out from an arbitrary initial partition $\{p\}_0$ of the n_n nodes in which the initial number of subsets is taken at random between $m_0 = 2$ and $m_0 = (n_n - 1)$ and then the nodes that compose the network are randomly assigned to any to the corresponding p_i subsets. In order to allow the procedure to increase the number of subsets in the partition of the system, the algorithm is implemented in such a way that at every step there is an empty subset present. Starting from this configurations we randomly choose a node n_i which belongs to the subset p_k of the partition $\{p\}_0$. We then choose at random another subset p_l of the partition $\{p\}_0$. If p_l is not empty we check if there is a link between the chosen node n_i and any of the nodes belonging to p_l. If no such a link exists, the possibility of transferring n_i from p_k to p_l is discarded and the selection procedure is repeated. If, on the other hand, there exists at least one link l_{im} between the node n_i and node n_m in p_l, or p_l is empty, we perform the Metropolis acceptance analysis:

(i) We calculate the value of Q for the partition $\{p\}_0$.
(ii) The transference of n_i from subset p_k to subset p_l is proposed, thus giving a new partition $\{p\}'$.
(iii) The value of $Q'$for $\{p\}$' is calculated.
(iv) The new partition is accepted with a probability $q = \min(1, \exp(-\beta\Delta Q))$ and the transference of node n_i is performed accordingly. If p_l was empty the total number of subsets, m_k, is increased by 1.
(v) Steps (ii)–(iii) are repeated with and exponentially increasing value of β until no new configurations are accepted within a fixed number of steps. The pseudo inverse temperature β is changed according to the rule $\beta' = \alpha\beta$, with $\alpha > 1$.
(vi) Whenever a partition with a higher Q is visited during the development of the Markov Chain, it is recorded. If the asymptotic partition is worse (lower Q) than the recorded one we consider this one as the result of the calculation.

This steps are repeated until the rate of acceptance of particle transfer drops to 0.

In order to improve the performance of this algorithm we have implemented steps of the Markov Chain in which multiple transferences of nodes between partitions are performed before the acceptance criterion is checked over the resulting partition. This procedure is used in order to avoid trapping configurations (those in which the removal of a single node is highly improbable but the configuration as a whole is not a maximum of Q).

It should be noted at this time that the resulting methodology is the same as the one developed by one of us for the analysis of fragmentation of highly excited liquid drops [8] and for the case of phase transitions in small constrained systems [6] (see these references for further details, for a comparison between different fragment recognition algorithms see [11]).

4. 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 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 have to choose one at random (as proposed by [3]) and this edge is removed. In our example we choose $l_{12,13}$ obtaining the graph displayed in Fig. 1B.

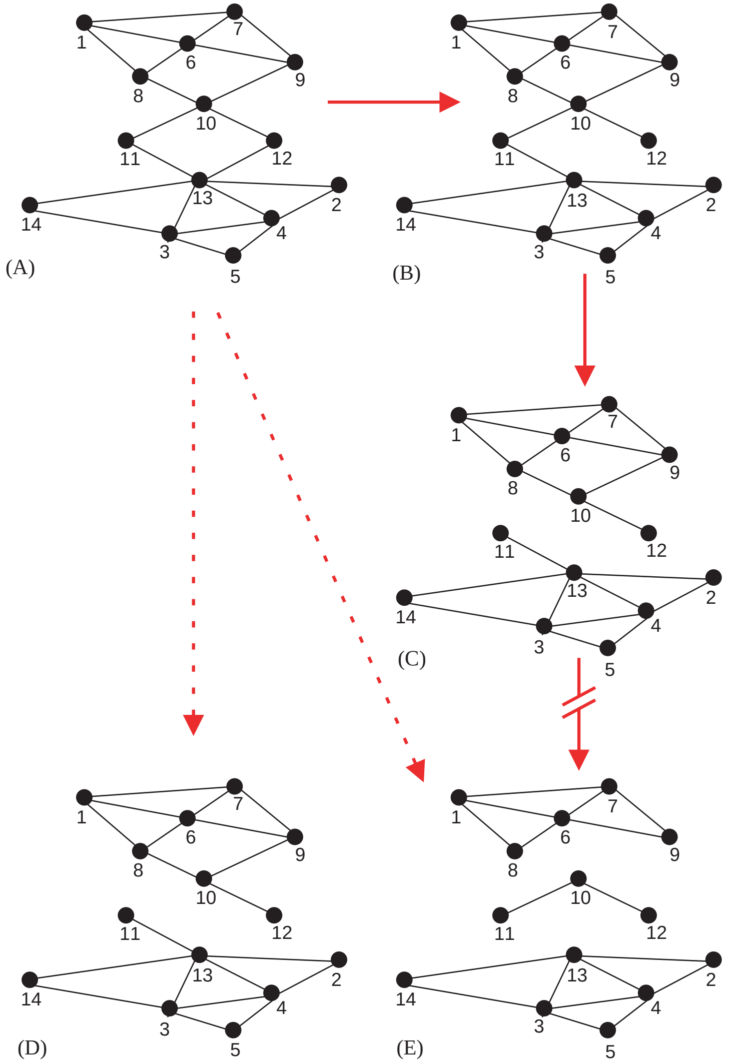

Fig. 1. Development of community structures in terms of the ER and SA analysis. Full arrows denote steps in the ER approach. Dotted arrows denote results from SA methodology. Starting from network A by applying ER methodology we first get to network B and, after the second removal of a link, to network 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.

(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 does not exists in the ER methodology.

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. A decision is made at a given stage of the analysis based only on the betweenness of the links with total disregard to the possible value of Q at the end of the calculation. Once a link is removed it is not rebuild into the system at any other stage of the calculation. On the other hand, when we analyze the sequence of results obtained with SA when we impose the condition of having $2, 3, 4, \ldots$ 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.

It is interesting to note that in a recent paper [12] it has been proposed that all links that share the highest value of betweenness are to be removed. If such a recipe is applied for our test case, there would be no route to a three communities solution. In fact the first structure that appears gives four communities in which two nodes $(11, 12)$ are isolated.

5. Examples

Once we have gained insight into the properties of the different approaches analyzed in this paper, we have found it appropriate to reanalyze some examples present in the literature and compare the results already published with the ones obtained using our methodology. We will analyze the Zachary's Karate Club network [3] and the relationship network of the characters of the novel *Les Misérables* by Victor Hugo, as compiled by Knuth [13] and analyzed in [3].

5.1. Zachary's karate club network

The main reason for the election of this network is that it is a classic social network and it has been analyzed by means of ER algorithms in some previous works [3]. This network was constructed by Wayne Zachary [14], who dedicated two years to the observation of social interactions between the members of a karate club. The data collected by Zachary made it possible to build the corresponding adjacency matrix that characterize relations between the members of the club.

In Figs. 2 and 3 we show the best partition obtained, in terms of the modularity Q, for the Zachary's non-weighted network by means of the SA and ER algorithms, respectively [15]. For the case of the SA algorithm, we achieve the largest modularity ($Q = 0.42$) for a structure of four communities (Fig. 2).

On the other hand the best community structure recognized with the ER approach [3] corresponds to five communities with a modularity value of $Q = 0.37$ (see Fig. 3).

In the actual case analyzed by Zachary, a dispute arose between the club's director and the karate teacher, and as a result the club split in two smaller clubs, one centered around the director and the other around the karate teacher. We performed the analysis of Q for this two-community split using both algorithms. We started the

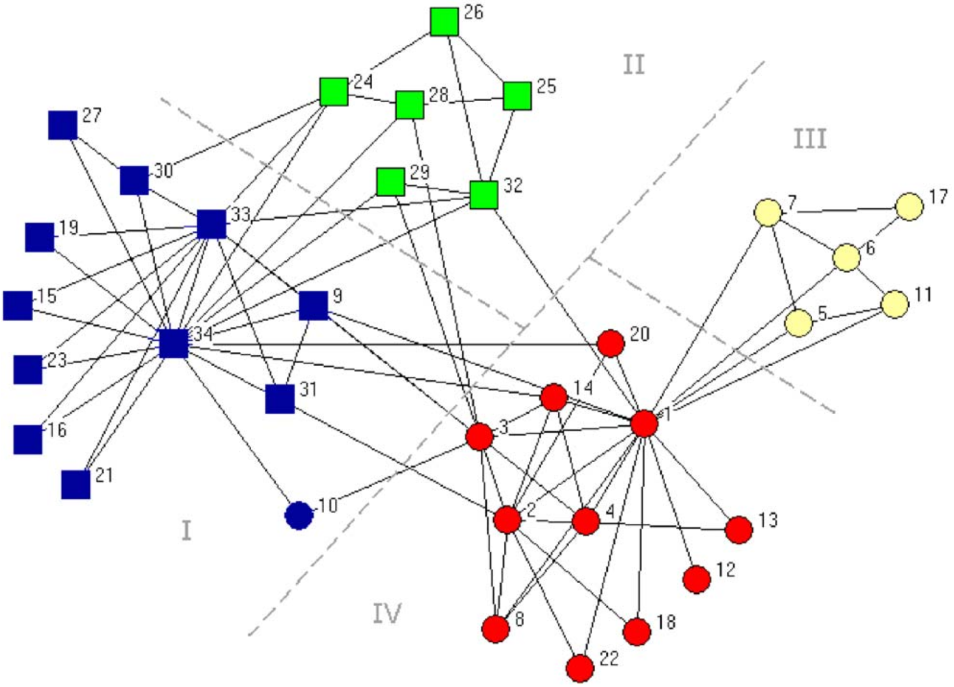

Fig. 2. Community structures for the Zachary network according to SA approach. In this figure, squares and circles denote the members of the two subsets according to observations by Zachary. Broken lines denote the partitions obtained according to SA approach.

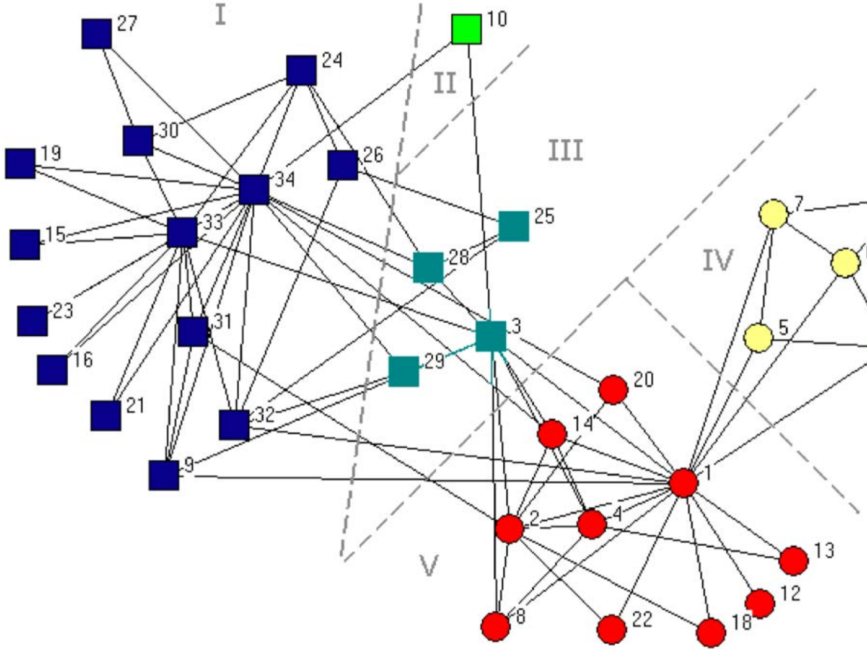

Fig. 3. Community structures for the Zachary network according to ER approach.

calculation using the network structure previous to the incident. The results are shown in Figs. 2 and 3 through circles and squares. Only the node 10 is misclassified by SA approach in comparison with the actual two-communities division observed by Zachary. However, when we run our algorithm for this problem, taking into account that the network can be transformed into a weighted network, with the weight of the links given by the "affinity degree" among the members, (see [14]) we obtain the actual two-community split shown in Fig. 4. Here we want to emphasize the fact that our SA algorithm can by applied either to weighted or unweighted networks without modification. This is so, because all the information about weights is contained in the adjacency matrix M and using this information, changes in the value of Q due to nodes regrouping, can be straightforwardly calculated (Fig. 4).

5.2. *Les Misérables network*

In the following example we analyze the network of interactions between major characters in the novel *Les Misérables*, by Victor Hugo, using the list of character appearances by scene as compiled by Knuth [13]. For this case, a link between two characters (nodes) represents the simultaneous appearance of both characters in one or more chapters.

Fig. 5 shows the community structure achieved by our SA algorithm. The best community structure has a modularity of $Q = 0.546$ and corresponds to 5

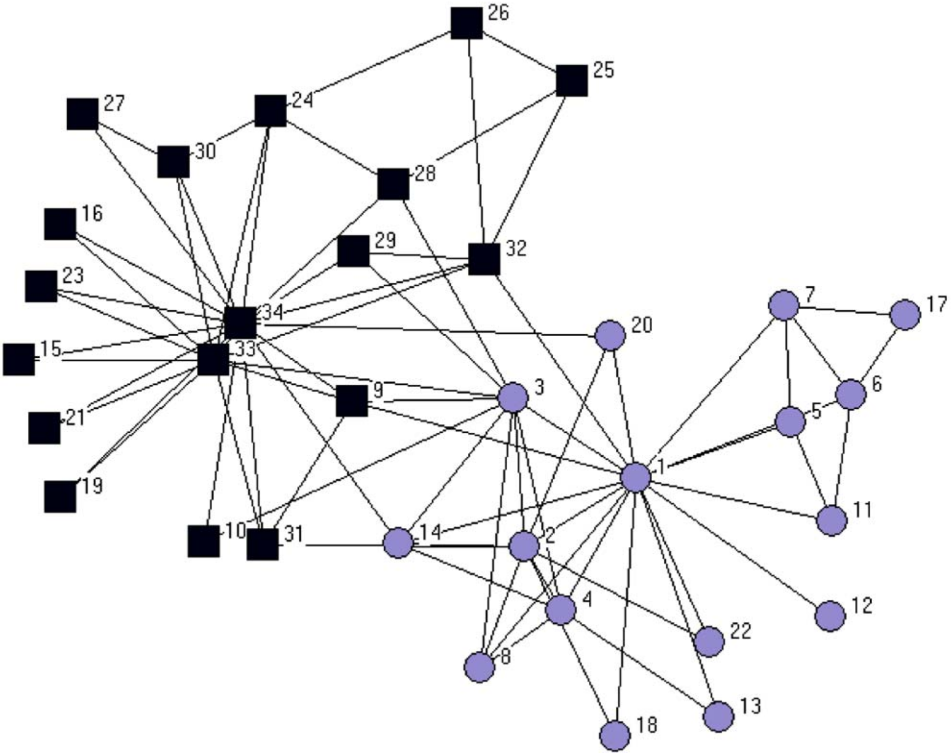

Fig. 4. Actual community structures as recorded by Zachary. Once again squares and circles denote the members of each subset.

communities, two of which are centered on the protagonists Jean Valjean and his persecutor, the police officer Javert; as can be seen in the figure. The other communities are centered on Marius, Fantine and bishop Myriel (here we want to note that the original data collected by Knuth, which we use in our calculations, has some mistakes, like the inclusion of Jondrette as a individual character, while Jondrette is only a pseudonymous of Thenardier). In the course of the analysis we find 3 isolated nodes, this mean that they have not links with other nodes, and for this reason they have been excluded from this figure.

When we run the ER algorithm for the same network, we obtain a structure with 11 communities and $Q = 0.538$, smaller than the obtained with our algorithm.

6. Conclusions

In this paper we have presented an analysis of communality structures in graphs based on a process of global optimization on the cost function Q. As stated in Section 1 (see [3,4,9,12]) the higher the value of Q the better the communality

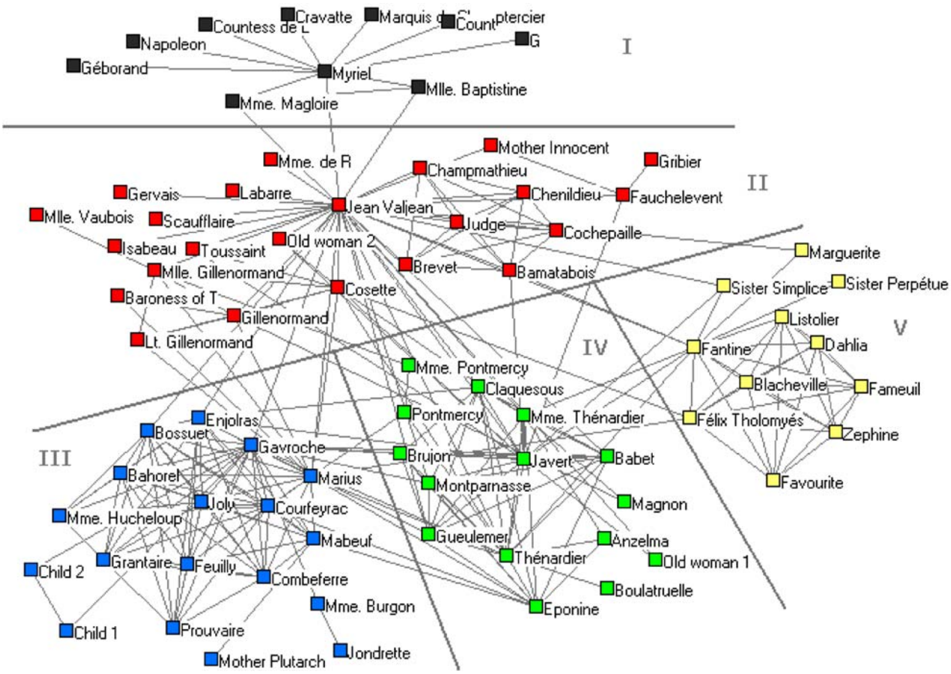

Fig. 5. Community structures for the "Les Misérables" network. Lines denote the best partition as obtained according to SA approach.

structure detected. We have compared our results with other calculations in the literature based on the Edge Removal approach as described above. In this case edges are removed according to their *betweenness* which is a measure of the importance of a link.

It has been shown via the analysis of a very simple graph that the process involved in the ER approach might fail due to a couple of reasons:

(a) Edge removal is performed in a local and irreversible way and it is completely blind to further developments of the graphs communality properties. Being the ER an irreversible process it might render subsequent (with better communality structure) partitions of the graph unreachable.

(b) When there is degeneracy (the highest betweenness corresponds to more than one link) one is to choose a link at random in order to cut it. Moreover if, according to a recent publication [12], one removes all of them at once, less accurate solutions are obtained.

All this properties have become apparent in the analysis of the simple graph. We have further applied our approach to other graphs already analyzed in the literature (Zachary's Karate club, Les Misérables) and in all cases the results obtained with the SA approach are better (larger Q) that the results obtained using ER methodology.

So far we have talked about accuracy, another issue relevant for this kind of analysis is speed, SA approaches are intrinsically slow, so, by the time being, we are restricted to not too big graphs (we are currently working on different schemes in order to improve the efficiency of the calculation). As an example, for the Zachary case SA is 10 times slower than ER.

Acknowledgements

C.O. Dorso is a member of the ''Carrera del Investigador'' CONICET Argentina. One of us (C.O.D.) thanks bibliographical support from S. Serebrinsky and V. Molinero. We thank H. Solari and M. Ison for careful reading of the manuscript.

References

[1] D.J. Watts, S.H. Strogatz, Nature 393 (1998) 440.
[2] R. Albert, A.L. Barabási, Rev. Mod. Phys. 74 (2002) 47.
[3] M.E.J. Newman, M. Girvan, Phys. Rev. E 69 (2004) 026113.
[4] M.E.J. Newman, Phys. Rev. E 69 (2004) 066133;
M.E.J. Newman, Phys. Rev. E 68 (2003) 026121.
[5] J. Lopez, C.O. Dorso, Phase Transformations in Nuclear Matter, World Scientific, 2000.
[6] A. Chernomoretz, P. Balenzuela, C.O. Dorso, Nucl. Phys. A 723 (2003) 229.
[7] A. Chernomoretz, M. Ison, S. Ortiz, C.O. Dorso, Phys. Rev. C 64 (2001) 024606.
[8] C.O. Dorso, J. Randrup, Phys. Lett. B 301 (1993) 328.
[9] M.E.J. Newman, Phys. Rev. E 70 (2004) 056131.
[10] P.J.M. van Laarhovween, E.H.L. Aarts, Simulated Annealing: Theory and Applications, Reisel, Dordrecht, 1987.
[11] A. Strachan, C.O. Dorso, Phys. Rev. C 56 (1997) 995.
[12] M. Girvan, M.E.J. Newman, Proc. Natl. Acad. Sci. USA 99 (2002) 7821.
[13] D.E. Knuth, The Stanford GraphBase: A Platform for Combinatorial Computing, Addison-Wesley, Reading, MA, 1993.
[14] W.W. Zachary, J. Antropol. Res. 33 (1977) 45.
[15] All networks have been drawn using the social network analysis progran NETDRAW, www.analytictech.com.