Community detection in complex networks using extremal optimization

Jordi Duch and Alex Arenas
Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, 43007 Tarragona, Spain

(Received 17 January 2005; revised manuscript received 27 June 2005; published 24 August 2005)

We propose a method to find the community structure in complex networks based on an extremal optimization of the value of modularity. The method outperforms the optimal modularity found by the existing algorithms in the literature giving a better understanding of the community structure. We present the results of the algorithm for computer-simulated and real networks and compare them with other approaches. The efficiency and accuracy of the method make it feasible to be used for the accurate identification of community structure in large complex networks.

DOI: 10.1103/PhysRevE.72.027104 PACS number(s): 89.75.Hc, 87.23.Ge

The description of the structure of complex networks has been one focus of attention of the physicists' community in recent years. The levels of description range from the microscopic (degree, clustering coefficient, centrality measures, etc., of individual nodes) to the macroscopic description in terms of statistical properties of the whole network (degree distribution, total clustering coefficient, degree-degree correlations, etc.) [1–4]. Between these two extremes there is a "mesoscopic" description of networks that tries to explain its community structure. The general notion of community structure in complex networks was first pointed out in the physics literature by Girvan and Newman [5], and refers to the fact that nodes in many real networks appear to group in subgraphs in which the density of internal connections is larger than the connections with the rest of the nodes in the network.

The community structure has been empirically found in many real technological, biological, and social networks [6–12] and its emergence seems to be at the heart of the network formation process [13].

The existing methods intended to devise the community structure in complex networks have been recently reviewed in [10]. All these methods require a definition of community that imposes the limit up to which a group should be considered a community. However, the concept of community itself is qualitative: nodes must be more connected within their community than with the rest of the network, and its quantification is still a subject of debate. Some quantitative definitions that came from sociology have been used in recent studies [14], but in general, the physics community has widely accepted a measure for the community structure based on the concept of modularity Q introduced by Newman and Girvan [15],

$$Q = \sum_r (e_{rr} - a_r^2) \tag{1}$$

where e_{rr} are the fraction of links that connect two nodes inside the community r, a_r the fraction of links that have one or both vertices inside of the community r, and the sum extends to all communities r in a given network. Note that this measure provides a way to determine if a certain mesoscopic description of the graph in terms of communities is more or less accurate. The larger the values of Q the more accurate is a partition into communities.

The search for the optimal (largest) modularity value seems to be a NP-hard problem due to the fact that the space of possible partitions grows faster than any power of the system size. For this reason, a heuristic search strategy is mandatory to restrict the search space while preserving the optimization goal [16,17]. Indeed, it is possible to relate the current optimization problem for Q with classical problems in statistical physics, e.g., the spin-glass problem of finding the ground-state energy [18], or the ground-state energy of a Potts model [27,30], where algorithms inspired in natural optimization processes such as simulated annealing [19] and genetic algorithms [20] have been successfully used.

In this Brief Report, we propose a divisive algorithm that optimizes the modularity Q using a heuristic search based on the extremal optimization (EO) algorithm proposed by Boettcher and Percus [21,22]. This algorithm is inspired in turn by the evolution model of Bak and Sneppen [23], and basically operates by optimizing a global variable by improving extremal local variables that involve coevolutionary avalanches. The performance of EO algorithms has been shown to overcome the efficiency of classical simulated annealing and genetic algorithms providing competitive accuracy [24,25].

In our case, the global variable to optimize is Q as defined in Eq. (1). Thus, the definition of the local variables used in the extremal optimization problem should be related to the contribution of individual nodes i to the summation in Eq. (1) given a certain partition into communities,

$$q_i = \kappa_{r(i)} - k_i a_{r(i)}, \tag{2}$$

where $\kappa_{r(i)}$ is the number of links that a node i belonging to a community r has with nodes in the same community, and k_i is the degree of node i. Note that $Q=(1/2L)\Sigma_i q_i$ where i refers to all nodes in the network given a certain partition into communities and L is the total number of links in the network. Equation (2) provides a measure that depends on the node degree, and its normalization involve all the links in the network after summation. Rescaling the local variable q_i by the degree of node i we obtain a proper definition for the contribution of node i to the modularity, relative to its own

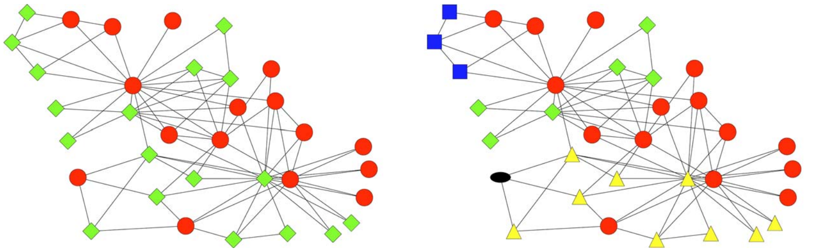

FIG. 1. (Color online) Left: Random initialization of the Zachary network into two partitions, circles (red) and squares (green). Right: Five different communities identified as connected components in each partition. Each symbol (color) defines a different connected component.

degree and normalized in the interval $[-1,1]$.

$$\lambda_i = \frac{q_i}{k_i} = \frac{\kappa_{r(i)}}{k_i} - a_{r(i)}. \tag{3}$$

Keeping in mind this definition of λ_i we can compare the relative contributions of individual nodes to the community structure. We will consider λ_i as the local variable involved in the extremal optimization process that characterizes an individual node; from now on we will refer to λ_i as the fitness of node i using the common jargon in extremal optimization problems.

The heuristic search we propose to find the optimal modularity value evolves as follows.

(1) Initially, we split the nodes of the whole graph into two random partitions having the same number of nodes in each one. This splitting creates an initial community division, where communities are understood as connected components in each partition.

(2) At each time step, the system self-organizes by moving the node with the lower fitness (extremal) from one partition to the other. In principle, each movement implies the recalculation of the fitness of many nodes because the right-hand side of Eq. (3) involves the pseudoglobal magnitude $a_{r(i)}$.

(3) The process is repeated until an "optimal state" with a maximum value of Q is reached. After that, we delete all the links between both partitions and proceed recursively with every resultant connected component. The process finishes when the modularity Q cannot be improved.[1]

Note that this process is not a bipartitioning of the graph as known in computer science [22], because the number of nodes in each partition is dependent on the evolution process and not restricted to be the same at the end of the process; and, more importantly, each partition could contain different connected components (communities) that when the partitions are disconnected result in several subgraphs.

Let us illustrate the above-mentioned heuristics in a simple case. We will apply it to the well-known Zachary karate club network [26]. Initially we split the nodes in two random partitions (see Fig. 1 left). Note that the number of initial communities (connected components in each partition) in this case is five (see Fig. 1 right). After that, the self-organization process starts: the node with the "worst fitness" is selected and moved from its partition to the other partition, and this movement provokes an avalanche of changes in the fitness of the rest of nodes. We calculate the new value for the modularity Q, and again repeat the process until no changes could improve it (see Fig. 2).

The application of the algorithm to the Zachary network provides the optimal modularity value after three recursive iterations. The network is decomposed in four communities and the value for the modularity is 0.419, greater than the value 0.381 reported by Newman [16], the value 0.406 reported by Reichardt *et al.* [27], and the value 0.412 reported by Donetti *et al.* [28] using different optimization methods.

The extremal optimization approach presented here has several technical implementation details that are relevant for our purposes. In the original EO algorithm, the node selected is always the node with the worst λ_j value. This is a deterministic and fast way to solve the problem, but it presents some drawbacks: the final result strongly depends on the initialization and there is no possibility to escape from local maxima. Instead, we use a probabilistic selection called τ-EO [21], in which the nodes are ranked according to their fitness values, and then the node of rank q is selected according to the following probability distribution:

$$P(q) \propto q^{-\tau}. \tag{4}$$

This solution is less sensitive to different initializations and allows escape from local maxima. The exponent τ has been tuned around the optimal values obtained for random networks of size N that approach the scaling $\tau \sim 1 + 1/\ln(N)$ [21]. The use of this technique also implies the determination of the number of self-organization steps αN needed to decide that the maximum value has little chance to be improved. In practice, we keep track at each step of the last maximum value obtained for Q; if this maximum is not improved in αN steps we stop the search. Usually α is empirically determined by balancing accuracy and efficiency in the algorithm; we use $\alpha = 1$ allowing as many steps as nodes to improve the current maximum value of Q. The computational cost involved in the whole process is $O(N^2 \ln^2 N)$ where the factor $N \ln N$ is the cost associated with the ranking process; however, it can be substantially reduced using heap data structures [29] for the ranking selection process up to $O(N)$. The total cost of the algorithm can then be improved up to $O(N^2 \ln N)$.

To test the performance of the algorithm we use first computer-generated graphs with a known community structure [5]. These graphs have 128 vertices grouped in four communities of 32 vertices. Each vertex has on average z_{in}

[1] The value of Q always refers to the whole network, i.e., it is the sum over all the communities. At a certain moment more subdivisions into communities will necessarily decrease Q because the limit of decomposition is one community per node, whose value of Q is negative.

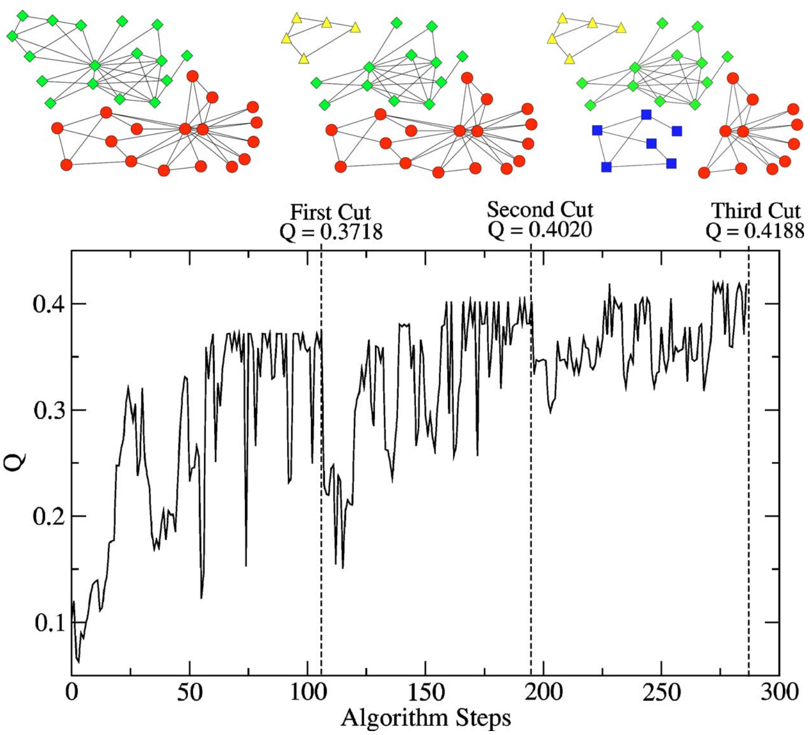

FIG. 2. (Color online) Top: Network after edge removal at each recursive cut. Bottom: Evolution of the Q value at each step of the adaptation process. Separation bars indicate recursive divisions of the graph performed at maximum Q.

edges to vertices in the same community and z_{out} edges to vertices in other communities, keeping an average degree $z_{in}+z_{out}=16$. We generate several graphs using z_{out} values between 0 and 10, and compare the results of our algorithm with those obtained using the heuristics proposed by Newman [16]. This shows the capabilities of each algorithm in identifying the communities when these are more fuzzy inside the whole network. Using the Girvan-Newman algorithm, which is the reference algorithm for community identification, the communities are well detected until values of $z_{out}=6$. In contrast, our algorithm detects the communities up to $z_{out}=8$, where the community structure still persists but is much more difficult to reveal (see Fig. 3). In this particular case 50% of the links are within the community and 50% are links with nodes outside the community. This result, which could seem contradictory, is not. Note that the 50% of links with nodes outside the community are, in average, equally distributed among the rest of the communities; their contribution to the definition of community is reduced by the number of communities in the rest of the network, in our case 3. For this reason it is expected to find community structure even in these cases.

For values higher than 8, the average maximum modularity rapidly approaches the limit $Q=0.208$ (see inset of Fig. 3), the expected modularity for a random network with the same number of links and nodes, as has been shown in [30].

We have also analyzed the community structure of several real networks: the jazz musicians network [31], a university e-mail network [13], the *C. elegans* metabolic network [32], a network of users of the pretty good privacy (PGP) algorithm for secure information transactions [33], and finally the relations between authors that shared a paper in cond-mat [34].

In Table I we present the results for the maximum modularity achieved by our algorithm compared to the modularity obtained using [16]. The difference in maximum modularity is up to 15% depending on the network considered. These differences result in a best determination of the unknown community structure of the whole network. The partition into communities is clearly different for large networks, as the different number of communities found using the two algorithms shows.

Note that since the core of the algorithm is stochastic, different runs could yield in principle different partitions. We have performed 100 runs of the algorithm for the e-mail network and for a random network with the same number of links and nodes to check the consistency of the proposed

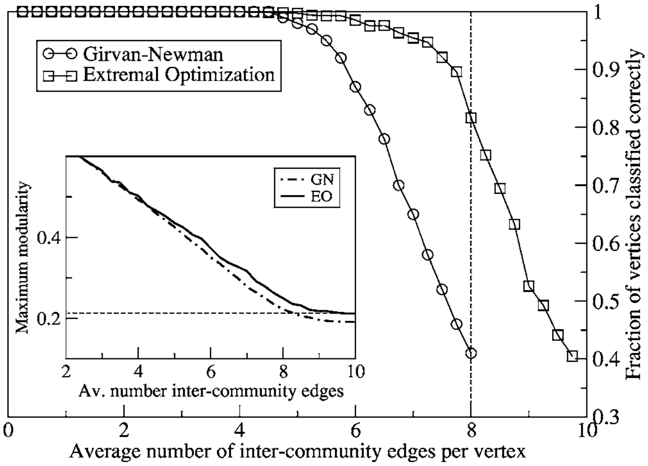

FIG. 3. Fraction of nodes correctly classified using computer-generated graphs described in the text. Each point is an average over 100 different networks. The fraction of nodes correctly classified we represent follows the definition proposed by Newman in [16]. Inset: Average of the maximum modularity obtained at each point.

TABLE I. Maximum modularity obtained using the algorithm [16] Q_N and the extremal optimization algorithm Q_{EO} for different complex networks. Also included is the number of communities found at the configuration with maximum modularity.

Network	Size	Q_N	No. coms_N	Q_{EO}	No. coms_{EO}
Zachary	34	0.3810	2	0.4188	4
Jazz	198	0.4379	4	0.4452	5
C. elegans	453	0.4001	10	0.4342	12
E-mail	1133	0.4796	13	0.5738	15
PGP	10680	0.7329	80	0.8459	365
Cond-Mat	27519	0.6683	302	0.6790	647

method. In Fig. 4 we present the results for the fraction of times two nodes are classified in the same partition. The community structure is clearly revealed for the e-mail network while for the random network this structure is nonexistent. Recently, Guimerà and Amaral obtained similar results by applying simulated annealing to find the community structure in the context of metabolic networks [35].

Summarizing, we have presented an extremal-optimization-based algorithm that optimizes the modularity and allows an accurate identification of community structure in complex networks. The results outperform all previous algorithms we are aware of in the literature.

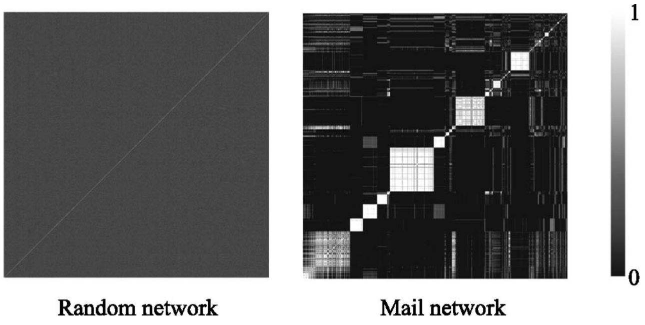

FIG. 4. Fraction of nodes classified in the same partition over 100 realizations of the algorithm. The color of the position (i,j) corresponds to the fraction of times that nodes i and j belong to the same partition.

We thank M. Boguña, L. Danon, A. Diaz-Guilera, and R. Guimerà for helpful comments and suggestions. We also thank M. E. J. Newman for providing us the cond-mat network. This work has been supported by DGES of the Spanish Government Grant No. BFM-2003-08258 and EC-FET Open Project No. IST-2001-33555.

[1] S. Strogatz, Nature (London) **410**, 268 (2001).
[2] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[3] S. Dorogovtsev and J. Mendes, Adv. Phys. **51**, 1079 (2002).
[4] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[5] M. Girvan and M. Newman, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002).
[6] J.-P. Eckmann and E. Moses, Proc. Natl. Acad. Sci. U.S.A. **99**, 5825 (2002).
[7] K. A. Eriksen, I. Simonsen, S. Maslov, and K. Sneppen, Phys. Rev. Lett. **90**, 148701 (2003).
[8] P. Holme, M. Huss, and H. Jeong, Bioinformatics **19**, 532 (2003).
[9] A. Arenas, L. Danon, A. Diaz-Guilera, P. M. Gleiser, and R. Guimerà, Eur. Phys. J. B **38**, 373 (2004).
[10] M. E. J. Newman, Eur. Phys. J. B **38**, 321 (2004).
[11] C. P. Massen and J. P. K. Doye, Phys. Rev. E **71**, 046101 (2005).
[12] M. Latapi and P. Pons, e-print cond-mat/0412368.
[13] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas, Phys. Rev. E **68**, 065103(R) (2003).
[14] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, Proc. Natl. Acad. Sci. U.S.A. **101**, 2658 (2004).
[15] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[16] M. E. J. Newman, Phys. Rev. E **69**, 066133 (2004).
[17] A. Clauset, M. E. J. Newman, and C. Moore, Phys. Rev. E **70**, 066111 (2004).
[18] D. Sherrington and S. Kirkpatrick, Phys. Rev. Lett. **35**, 1792 (1975).
[19] S. Kirkpatrick, C. Gilatt, and M. Vecchi, Science **220**, 671 (1983).
[20] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning* (Addison-Wesley, Reading, MA, 1989).
[21] S. Boettcher and A. G. Percus, Phys. Rev. Lett. **86**, 5211 (2001).
[22] S. Boettcher and A. G. Percus, Phys. Rev. E **64**, 026114 (2001).
[23] P. Bak and K. Sneppen, Phys. Rev. Lett. **71**, 4083 (1993).
[24] S. Boettcher and A. G. Percus, Artif. Intell. **119**, 275 (2000).
[25] S. Boettcher and P. Sibani, e-print cond-mat/0406543.
[26] W. W. Zachary, J. Anthropol. Res. **33**, 452 (1977).
[27] J. Reichardt and S. Bornholdt, Phys. Rev. Lett. **93**, 218701 (2004).
[28] L. Donetti and M. Munoz, J. Stat. Mech.: Theory Exp. **2004**, 10012 (2004).
[29] A. V. Aho, J. D. Ullman, and J. E. Hopcroft, *Data Structures and Algorithms* (Addison-Wesley, Reading, MA, 1983).
[30] R. Guimerá, M. Sales-Pardo, and L. A. N. Amaral, Phys. Rev. E **70**, 025101(R) (2004).
[31] P. Gleiser and L. Danon, Adv. Complex Syst. **6**, 565 (2003).
[32] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, Nature (London) **407**, 651 (2000).
[33] X. Guardiola, R. Guimera, A. Arenas, A. Diaz-Guilera, and L. A. N. Amaral, e-print cond-mat/0206240.
[34] M. E. J. Newman, Phys. Rev. E **64**, 016131 (2001).
[35] R. Guimera and L. Amaral, Nature (London) **433**, 895 (2005).