Network community structure and loop coefficient method

I. Vragović and E. Louis
Departamento de Física Aplicada, Instituto Universitario de Materiales and Unidad Asociada del Consejo Superior de Investigaciones Científicas, Universidad de Alicante, San Vicente del Raspeig, Alicante 03690, Spain

(Received 17 October 2005; revised manuscript received 10 April 2006; published 11 July 2006)

A modular structure, in which groups of tightly connected nodes could be resolved as separate entities, is a property that can be found in many complex networks. In this paper, we propose a algorithm for identifying communities in networks. It is based on a local measure, so-called *loop coefficient* that is a generalization of the clustering coefficient. Nodes with a large loop coefficient tend to be core inner community nodes, while other vertices are usually peripheral sites at the borders of communities. Our method gives satisfactory results for both artificial and real-world graphs, if they have a relatively pronounced modular structure. This type of algorithm could open a way of interpreting the role of nodes in communities in terms of the local loop coefficient, and could be used as a complement to other methods.

DOI: 10.1103/PhysRevE.74.016105 PACS number(s): 89.75.Hc, 05.10.−a, 87.23.Ge, 89.20.Hh

I. INTRODUCTION

In recent years, there has been an increasing interest in analyzing the *community structure* in networks [1,2]. Very often complex networks can be divided into modules with a large number of internal edges, interconnected by a smaller number of external links. Such a purely topological decomposition is usually accompanied by different functional roles, such as in the case of related web sites [3,4], animal communities [5,6] or biochemical networks, and electronic circuits [7–10].

Approaches to network division can be classified into spectral analysis methods [11–17], maximization of modularity [18–23], agglomerative methods [24–29], and link removal methods [30–33]. A detailed review of various approaches can be found in Refs. [1,2].

Some of the methods focus on local topological properties. The algorithm of Radicchi *et al.* [32] uses a local measure based on counting the number of short loops that run through each edge. In most cases, an intermodular link would not be a part of too many short loops. Focusing on triangles, Radicchi *et al.* introduced the edge clustering coefficient that represents the fraction of realized triangles over potentially possible triangles going through a particular node. A network is decomposed by cutting edges with the smallest value of the edge clustering coefficient, which is recalculated after each removal step. The disadvantage of this method is that it cannot be applied to networks with a small number of triangles. In order to overcome this shortcomming, Radicchi *et al.* used measures based on squares and bigger loops, slightly improving the performance of the basic algorithm.

Eckmann and Moses proposed a concept of a local curvature, which is equivalent to the standard clustering coefficient and depends on the average distance between the first neighbors of a reference node [34,35]. This method is based on the assumption that nodes within a high curvature region belong to the same community. The disadvantage of this method is the same as in the previous case: it cannot be applied to networks with low clustering.

II. THE METHOD

In this paper we propose a method based on a local quantity which takes into account the smallest loops running through a particular node. It is based on the shortest distance between the first neighbors (j and k) of one site (i), when this site is cutoff, denoted as $d_{jk/i}$. In principle, it is a modification of the local efficiency introduced by Latora and Marchiori [36]. In our version, however, the paths $d_{jk/i}$ are allowed to

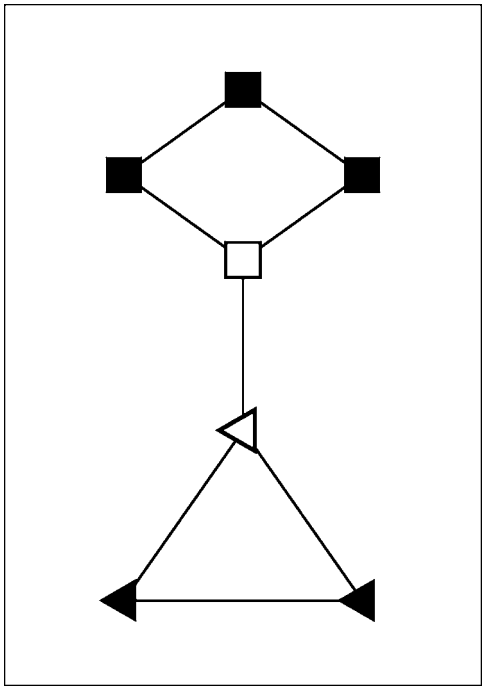

FIG. 1. A simple network consisting of two modules. Core and peripheral sites are represented by black and white symbols, respectively. The clustering coefficient: $C(\blacktriangle)=1$, triangle core; $C(\triangle)=1/3$, triangle peripheral; and $C(\blacksquare,\square)=0$, for all square nodes. The loop coefficient: $D(\blacktriangle)=1$, triangle core; $D(\triangle)=1/3$, triangle peripheral; $D(\blacksquare)=1/2$, square core; and $D(\square)=1/6$, square peripheral. The clustering coefficient of a peripheral node in the clustered module is smaller than for the core nodes. However, the clustering coefficient cannot distinguish between the peripheral and core nodes in the declustered module, as it is equal to zero for all sites. On the other hand, the loop coefficient of the peripheral sites is smaller than that of the core sites, no matter whether triangle loops are present or not.

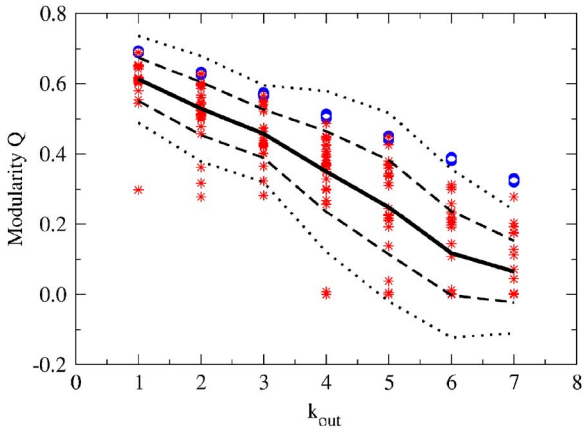

FIG. 2. (Color online) Comparison of the modularity of the division into four initial modules (blue circles) with the modularity of the division obtained by our method (red stars) for the computer-generated graphs described in the text. For each k_{out} we plotted the modularity of 40 different network realizations. Full black line is the average performance of our method, while black dashed and dotted lines are intervals of one and two standard deviations around the average, respectively.

go also through nodes that are not the first neighbors of i Ref. [37]. The modified local efficiency of an arbitrary node i reads

$$D(i) = \frac{1}{k_i(k_i-1)} \sum_{j \neq k \in \Gamma_i} \sum \frac{1}{d_{jk/i}}, \tag{1}$$

where Γ_i refers to a subgraph of the first neighbors. In the cases of $k_i=1$ and $k_i=0$, we take $D(i)=0$. This quantity is a counterpart of the clustering coefficient $C(i)$ defined as the ratio of the existing number of links between the neighbors of a site i and its maximum possible number $k_i(k_i-1)/2$

$$C(i) = \frac{1}{k_i(k_i-1)} \sum_{j \neq k \in \Gamma_i} \sum n_{jk}, \tag{2}$$

where $n_{jk}=1$ if there is a link between j and k, and $n_{jk}=0$ otherwise.

Contrary to clustering or quadrilateral coefficients [38,39], the modified local efficiency does not depend exclusively on the presence of triangles, or squares or any other particular kind of loops, but only on loops in general [37]. It simply takes into account any of the smallest loop going through a reference site i and its pair of neighbors (j,k). In the rest of the paper we will refer to $D(i)$ as *loop coefficient*. A similar measure (*cyclic coefficient*) was recently introduced in Ref. [40], and defined as the inverse of the full length of the shortest loop $S^i_{lk}=d_{lk/i}+2$, instead of the shortest path $d_{lk/i}$. As the smallest possible loops are triangles, the cyclic coefficient takes values between 0 and 1/3. On the other hand, the loop coefficient varies from 0 to 1 [37]. Moreover, if only the shortest paths $d_{lk/i}=1$ are taken into account, the loop coefficient $D(i)$ coincides with the standard clustering coefficient $C(i)$.

We note that core nodes, defined as nodes fully surrounded by sites belonging to the same community, usually have large fraction of closely connected neighbors (not "necessarily" directly linked) and thus large values of the loop coefficient. On the other hand, peripheral vertices which are also linked to the nodes of the other modules tend to have smaller loop coefficients. The reasoning is the same as in the case of Radicchi *et al.* [32]. If one of the first neighbors of a reference node belongs to another module, then it is not closely connected with other first neighbors of that peripheral node thus leading to both smaller loop coefficient of the analyzed node and smaller edge clustering coefficient of the intermodular link. On the basis of the value of the loop coefficient, we intend to interpret the role of each node classifying them as core inner nodes or peripheral sites, see Fig. 1. We note, however, that our method cannot be applied to trees that have no loops.

We start by calculating the loop coefficient of all sites, a calculation that is performed only once. After sorting the nodes we obtain something like a *relief*, with *hills* representing core nodes of distinctive modules and *valleys* corresponding to peripheral sites connected by intermodular links. In the first step we identify the nodes with the largest loop coefficient and start building modules around them. Their first neighbors are interpreted as the first layer of peripheral sites, so that they are put into the same group as the corresponding highly clustered initial node. If several initial nodes share the same peripheral sites, they are grouped together representing one single module, no matter if directly linked or not. They resemble a *plateau* with a common *valley*.

In the next steps, when smaller values of the loop coefficient are considered, a new module is created if an analyzed node is not a member of any existing group, i.e., if it is surrounded by nodes of even smaller loop coefficient. Creating a new module, we follow the initialization procedure described above. However, in most cases the analyzed node will already be attached to some of the core sites with larger loop coefficient. Again, the first neighbors of a reference site that still do not belong to any module are put into the current group.

We also check the first neighbors of a given node that are already peripheral members of other modules. We determine the number of links pointing from a reference site i to the inner nodes of its group $\mathrm{in}(i)$, and the number of external links to other modules $\mathrm{out}(i,m)$, where m counts the external modules. This is done in order to evaluate the current decomposition against the criteria of the community definition [32]. In a strong sense, a community is a subgraph when $\mathrm{in}(i) > \mathrm{out}(i,m)$ for every i, i.e., each node has more internal than external connections. As such a criterion is rarely fulfilled for peripheral nodes, another definition has been proposed [32]. In a weak sense, a community is a subgraph when $\Sigma_i \mathrm{in}(i) > \Sigma_i \Sigma_m \mathrm{out}(i,m)$, focusing on the total number of internal and external links. In our case, we need to analyze only peripheral sites, as all inner nodes have only internal links. We further weaken the community criteria, assuming that a node is correctly grouped if the number of the internal links $\mathrm{in}(i)$ is larger than the biggest number of external links, i.e., $\mathrm{in}(i) > \max[\mathrm{out}(i,m)]$, no matter if $\Sigma_m \mathrm{out}(i,m)$ is larger than $\mathrm{in}(i)$ even for each of the peripheral nodes, or even if the number of internal links is eventually smaller than the total

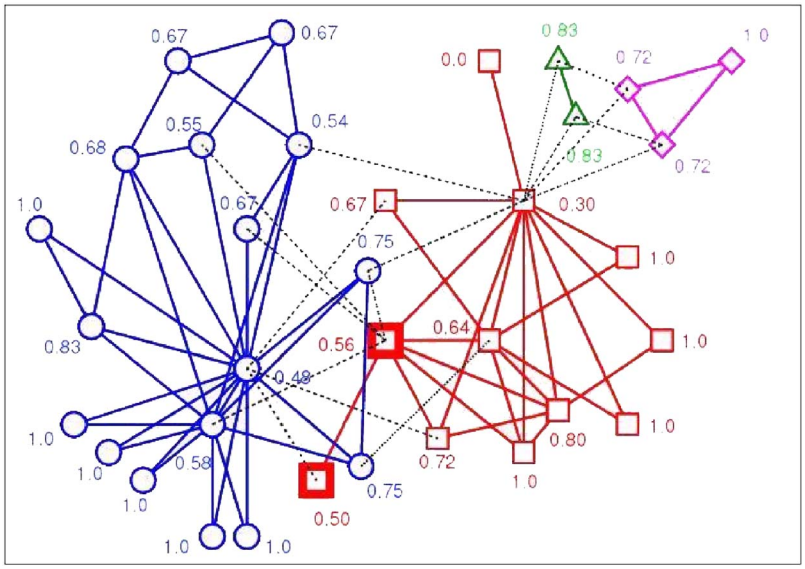

FIG. 3. (Color online) Decomposition of the karate club network of Zachary [1,30,41,42,44]. Nodes and links are arranged in the same way as in Fig. 4(a) of Ref. [30]. Members of the administrator's faction are represented by circles, while others belonging to the instructor's faction are drawn as squares, diamonds, or triangles. Internal links are plotted as full lines, and external intermodular links as dashed lines. The numbers attached to each node are the values of the loop coefficient. The member number 3 of Ref. [30] (upper bold square), whose membership was incorrectly determined by Girvan-Newman algorithm, is put into the correct group. One node is misclassified (node number 10 of Ref. [30], here lower bold square), because it has only one connection with both administrator's and instructor's factions. Diamond nodes (numbered as 6, 7, and 17 in Ref. [30]) are identified as a separate submodule, connected with the main part of the instructor's faction by an intermodular layer of triangle nodes (numbered as 5 and 11 in Ref. [30]). Several steps that led to this decomposition are shown in Fig. 4.

number of external links [i.e., $\Sigma_i \mathrm{in}(i) < \Sigma_i \Sigma_m \mathrm{out}(i,m)$].

A correction is made if a peripheral node is connected to one or more external groups with $\mathrm{out}(i,m) \geq \mathrm{in}(i)$. In such a case we change the membership of the analyzed site and of all its nearest neighbors, whose membership was determined in the current step. We put them into the external group m corresponding to the largest $\mathrm{out}(i,m)$. The membership of other nearest neighbors that were classified in previous steps is not changed. If there are more than one external group with the same $\mathrm{out}(i,m)$ [that could be equal to $\mathrm{in}(i)$], we identify the right group by looking for the external (or internal) neighbor of the analyzed node i within these groups that has the largest loop coefficient. Finally, after dealing with nodes with the smallest loop coefficient, we do the last correction step, changing only the membership of nodes whose $\max[\mathrm{out}(i,m)]$ is eventually larger than $\mathrm{in}(i)$, not altering their (already grouped) neighbors.

Concerning the computational time, the most demanding step is the determination of the loop coefficient for all nodes. As each node must be temporarily cutoff from the network and the new shortest paths between its first neighbors are allowed to pass through any other node, the computational time is N times longer than that needed to determine the shortest paths between nodes (where N denotes the number of vertices). This procedure could be shortened by limiting the length of the shortest paths taken into account ($\max[d_{lk/i}]$), so that larger separations do not contribute to the loop coefficient ($1/d_{lk/i} \to 0$ for $d_{lk/i} > \max[d_{lk/i}]$). For $\max[d_{lk/i}]=1$, we get the standard clustering coefficient. In this approximation, most of the core nodes could be determined quite fast as sites with the largest clustering coefficient. However, it would significantly reduce the efficiency of the method, especially in the case of sparse graphs with a small amount of triangles.

III. THE IMPLEMENTATION

A. Computer-generated networks

We have applied our method to several widely used examples. First, we considered computer-generated networks of known modular structure [30,41,42]. We created four regular rings of 32 vertices and 256 edges, so that each node has 16 links. Then, connections in each module were randomized by pair-wise rewiring in order to get initial random modules. In the next step, the four randomized modules were interconnected by pair-wise rewiring, so that the degree of each node $k=k_{in}+k_{out}$ was kept constant. We examined the performance of our algorithm on computer-generated graphs, when the number of external links per node was varied from $k_{out}=1$ to $k_{out}=7$. For each value of k_{out} we generated 40 different networks. The measured quantity was the modularity [18]

$$Q = \sum_i (e_{ii} - a_i^2), \tag{3}$$

representing the difference between the fraction of the links that fall within the communities (e_{ii}) and the expected value of the same number of links distributed randomly, regardless the community structure ($a_i=\Sigma_j e_{ij}$, with e_{ij} being the fraction of edges connecting nodes of group i with the nodes of group

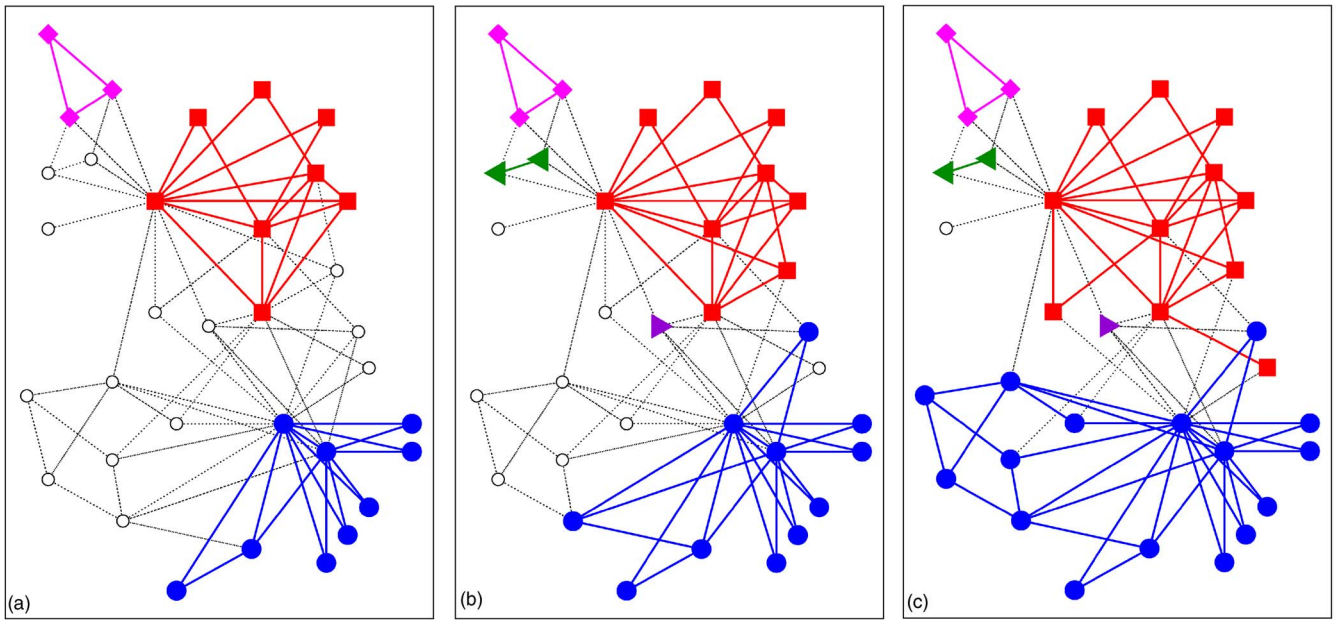

FIG. 4. (Color online) Decomposition of the karate club network of Zachary [1,30,41,42,44]. After the first step (a) the nodes with the largest loop coefficient of $D(i)=1$ (the values of the loop coefficient are given in Fig. 3) are identified as initial nodes of three modules. Their nearest neighbors are put into the corresponding modules as peripheral sites. After four steps [(b) with $D(i)=0.749$], an intermodular layer (triangle up, green) and one isolated node (triangle down, purple) appear as new distinct groups. The isolated node is assigned correctly to the circle module in the final correction step. In the following steps, both circle- and square-modules grow. (c) The division after the tenth step [with $D(i)=0.558$]. Compare with the final decomposition after the last correction step, given in Fig. 3.

j) [18]. Systems with $Q > Q_R$, where Q_R is a modularity of a corresponding random graph with the same number of nodes and links, seem to have a significant community structure [43].

For each k_{out}, we compare the modularity of the division into four initial modules with the modularity of the division obtained with our loop coefficient method (not necessarily into four modules). As the number of external links is increased, the initial modules are gradually falling apart and the network is turning into a single global module. The modularity of both divisions is decreasing, see Fig. 2. For smaller number of external edges, the algorithm gave very good results with all nodes correctly grouped. However, the resulting modularity appears to be lower than the reference one by 0.1–0.2. The cause of this discrepancy lies in the sensitivity of our method, as the initial random modules are divided into even smaller submodules that are strongly interconnected.

For $k_{out} > 4$, the initial modules become even more interconnected and our algorithm completely fails to recognize them. We could only detect larger supermodules; a merger of two or more initial groups or their parts. Having only two or three groups, the modularity of some divisions drops sharply (see Fig. 2 for $k_{out}=5$; 6 or 7). For $k_{out}=7$, our algorithm hardly recognizes the initial modular structure, putting in most cases the whole network into a single module ($Q=0$).

B. Real-world networks

We tested our method on two real-world networks: Zachary's karate club [44] and the largest component of the Santa Fe Institute collaboration network [1,30,41,42,45].

In the case of the karate club of Zachary [1,30,41,42,44], the behavior of 34 members were observed during two years in order to determine the network of friendships. After a disagreement between the administrator of the club and the instructor, the club was split into two groups. Applying our algorithm on an unweighted network, we identified a modular structure somewhat richer than the one usually reported in the literature [1,30,41,42]. We correctly identified two main groups, with a small submodule and an intermodular layer within the instructor's faction. One peripheral node was misclassified (lower bold square in Fig. 3). As it has one link to both main groups, the algorithm interpreted it as a peripheral node of the instructor's faction because its neighbor within that group has a larger loop coefficient than the neighbor in the *correct* administrator's faction. Finally, in order to check the quality of our division and to compare it with the division previously obtained by Girvan-Newman method [30], we calculated the modularity Q [see Eq. (3)]. The modularity of our division into four groups is $Q=0.378$, while merging submodules of the instructor's faction into one larger module we get the same modular structure as reported in [18], with $Q=0.381$. (See 4.)

The largest component of the Santa Fe Institute collaboration network [1,30,41,42] consists of 118 scientists that are connected if they coauthored one or more articles (links are unweighted). The aim of any algorithm would be to identify the communities that correspond to the particular fields of research. The rough division resulting from our method is shown in Fig. 5. We successfully resolved the group working on the Structure of RNA and three groups of the Statistical

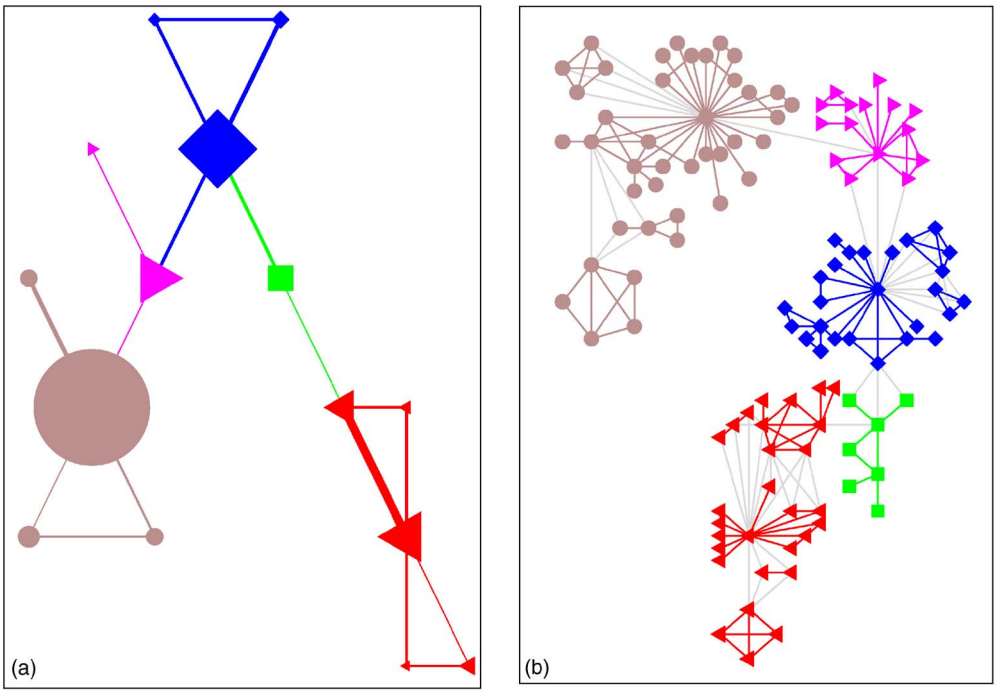

FIG. 5. (Color online) (a) Rough decomposition of the largest component of the Santa Fe Institute collaboration network into five main groups, having internal submodular structures. The modularity of this rough decomposition is $Q=0.7$. The sizes of the symbols represent relative sizes of the groups, while the thicknesses of the connections correspond to the number of intermodular links. (b) A detailed decomposition of the largest component of the Santa Fe Institute collaboration network with the structure of each of the five groups divided into 15 submodules. The modularity of this detailed decomposition is $Q=0.659$. The color of the internal links is that of the corresponding submodules, while the external links are plotted grey. Color and symbol codes for different disciplinary groups: Structure of RNA (red triangles up, 29 members), Statistical Physics (green squares with 7 members, blue diamonds with 26 members, and magenta triangles down with 15 members), Mathematical Ecology and agent-based models (brown circles, 41 members).

Physics; see Fig. 5 and compare with Ref. [30]. The modularity of the division into five large groups is $Q=0.7$. Most of these communities could be further divided into a main cluster surrounded by smaller submodules. A detailed decomposition [see Fig. 5 (b)] shows that these small submodules are usually fully connected graphs, representing a group of coauthors sharing a single publication. The modularity of the division into 15 small subgroups is $Q=0.659$. Finally, our algorithm could not make a clear distinction between the communities of the Mathematical Ecology and the agent-based models (both depicted with brown circles) that could be resolved by Girvan-Newman method [30]. We could only distinguish two interconnected small submodules of sizes 5 and 4 that are linked to the main group of the Mathematical Ecology by a common coauthor.

IV. CONCLUSION

We have introduced a algorithm for identifying communities in networks that is based on a local property, the *loop coefficient*. The method takes into account the smallest loops of any size, and actually is a generalization of the widely used clustering coefficient. Therefore, our method could be applied to any kind of networks with loops, regardless of the number of triangles. However, it is not suitable for analyzing trees, without cycles. Our concept of community is based on the core nodes having the largest values of the loop coefficient, surrounded by pheripheral nodes that are the origin of intermodular links.

Applying our algorithm to both computer generated and real-world networks with well defined community structures, we got reasonable results that are in accordance with the findings obtained by other methods. We were able to identify both the rough division into the main communities and the detailed division into smaller submodules and intermodular layers. When the number of intermodular links in computer generated systems is high, our method fails to make a distinction between the initial groups, assigning the majority of nodes to a single supermodule.

The proposed algorithm gives a way of identifying the core nodes using a new local measure, i.e., the loop coefficient. We hope that it could be used as a complement of other methods, especially in combination with approaches focused more on the boundaries of the modules.

ACKNOWLEDGMENTS

We are very thankful to Mark E. J. Newman for sharing the data on the network structures. Financial support by Fet Open Project COSIN IST-2001-33555 and the University of Alicante is gratefully acknowledged.

[1] M. E. J. Newman, Eur. Phys. J. B **38**, 321–330 (2004).
[2] L. Danon, J. Duch, A. Arenas, and A. D.- Guilera, e-print cond-mat/0505245 (2005).
[3] D. Gibson, J. Kleinberg, and P. Raghavan, *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia* (Association of Computing Machinery, New York, 1998).
[4] G. W. Flake, S. R. Lawrence, C. L. Giles, and F. M. Coetzee, IEEE Comput. Graphics Appl. **35**, 66 (2002).
[5] D. Lusseau, Proc. R. Soc. London, Ser. B **270**, 186 (2003).
[6] D. Lusseau and M. E. J. Newman, Proc. R. Soc. London, Ser. B **271**, 477 (2004).
[7] P. Holme, M. Huss, and H. Jeong, Bioinformatics **19**, 532 (2003).
[8] P. Holme and M. Huss, e-print q-bio.MN/0309011 (2003).
[9] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, Science **298**, 824 (2002).
[10] S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, Nat. Genet. **31**, 64 (2002).
[11] M. Fielder, Czech. Math. J. **23**, 298 (1973).
[12] A. Pothen, H. Simon, and K.-P. Lion, SIAM J. Matrix Anal. Appl. **11**, 430 (1990).
[13] B. W. Kernigham and S. Lin, Bell Syst. Tech. J. **49**, 291 (1970).
[14] B. Bollobas, *Modern Graph Theory* (Springer, New York, 1998).
[15] L. Donetti and M. A. Muñoz, J. Stat. Mech.: Theory Exp. P10012 (2004).
[16] A. Capocci, V. D. P. Servedio, G. Caldarelli, and F. Colaiori, Lect. Notes Comput. Sci. **3243**, 181 (2004).
[17] F. Wu and B. A. Huberman, Eur. Phys. J. B **38**, 331 (2004).
[18] M. E. J. Newman, Phys. Rev. E **69**, 066133 (2004).
[19] A. Clauset, M. E. J. Newman, and C. Moore, Phys. Rev. E **70**, 066111 (2004).
[20] C. P. Massen and J. P. K. Doye, Phys. Rev. E **71**, 046101 (2005).
[21] J. Duch and A. Arenas, Phys. Rev. E **72**, 027104 (2005).
[22] S. Boettcher and A. G. Percus, Phys. Rev. Lett. **86**, 5211–5214 (2001).
[23] S. Boettcher and A. G. Percus, Phys. Rev. E **64**, 026114 (2001).
[24] J. Scott, *Social Network Analysis: A Handbook* (Sage, London, 2000).
[25] R. S. Burt, Social Forces **55**, 93 (1976).
[26] S. Wasserman and K. Faust, *Social Network Analysis* (Cambridge University Press, Cambridge, 1994).
[27] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms, and Applications* (Prentice Hall, Upper Saddle River, NJ, 1993).
[28] J. Hopcroft, O. Khan, B. Kulis, and B. Selman, Proc. Natl. Acad. Sci. U.S.A. **101**, 5249 (2004).
[29] J. P. Bagrow and E. M. Bollt, Phys. Rev. E **72**, 046108 (2005).
[30] M. Girvan and M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002).
[31] J. R. Tyler, D. M. Wilkinson, and B. A. Huberman, in *Proceedings of the First International Conference on Communities and Technologies*, edited by M. Huysman, E. Wenger, and V. Wulf (Kluwer, Dordrecht, 2003).
[32] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, Proc. Natl. Acad. Sci. U.S.A. **101** (9), 2658 (2004).
[33] S. Fortunato, V. Latora, and M. Marchiori, Phys. Rev. E **70**, 056104 (2004).
[34] J.-P. Eckmann and E. Moses, Proc. Natl. Acad. Sci. U.S.A. **99**, 5825 (2002).
[35] J.-P. Eckmann, E. Moses, and D. Sergi, Proc. Natl. Acad. Sci. U.S.A. **101**, 14333 (2004).
[36] V. Latora and M. Marchiori, Phys. Rev. Lett. **87**, 198701 (2001).
[37] I. Vragović, E. Louis, and A. Díaz-Guilera, Phys. Rev. E **71**, 036122 (2005).
[38] T. Petermann and P. De Los Rios, Phys. Rev. E **69**, 066116 (2004).
[39] G. Caldarelli, R. Pastor-Satorras, and A. Vespignani, Eur. Phys. J. B **38**, 183–186 (2004).
[40] H.-J. Kim and J. M. Kim, Phys. Rev. E **72**, 036109 (2005).
[41] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).
[42] H. Zhou, Phys. Rev. E **67**, 061901 (2003).
[43] J. Reichardt and S. Bornholdt, e-print cond-mat/0603718 (2006).
[44] W. W. Zachary, J. Anthropol. Res. **33**, 452–473 (1977).
[45] H. Zhou, Phys. Rev. E **67**, 041908 (2003).