PHYSICAL REVIEW E **68**, 065103(R) (2003)

Self-similar community structure in a network of human interactions

R. Guimerà,[1,2] L. Danon,[3,4] A. Díaz-Guilera,[3,1] F. Giralt,[1] and A. Arenas[4]
[1]*Departament d'Enginyeria Química, Universitat Rovira i Virgili, 43007 Tarragona, Catalunya, Spain*
[2]*Department of Chemical Engineering, Northwestern University, Evanston, Illinois 60208, USA*
[3]*Departament de Física Fonamental, Universitat de Barcelona, 08028 Barcelona, Catalunya, Spain*
[4]*Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, 43007 Tarragona, Catalunya, Spain*
(Received 14 January 2003; revised manuscript received 9 May 2003; published 17 December 2003)

We propose a procedure for analyzing and characterizing complex networks. We apply this to the social network as constructed from email communications within a medium sized university with about 1700 employees. Email networks provide an accurate and nonintrusive description of the flow of information within human organizations. Our results reveal the self-organization of the network into a state where the distribution of community sizes is self-similar. This suggests that a universal mechanism, responsible for emergence of scaling in other self-organized complex systems, as, for instance, river networks, could also be the underlying driving force in the formation and evolution of social networks.

DOI: 10.1103/PhysRevE.68.065103 PACS number(s): 89.75.Fb, 89.75.Da, 89.75.Hc

Signatures of complex systems appear in disciplines as diverse as biology, chemistry, economy, and computer science, to name just a few. More specifically, the study of the complex networks of interactions in such systems has received a lot of attention from the statistical physics community [1–5]. The structure of these complex networks is a reflection of the dynamics of their formation and evolution, and can be partially characterized using statistical observables such as the average distance between nodes [1], the clustering coefficient [1], and the degree distribution [2,3]. Even though these measures are very useful in some situations, often they are not sufficient to describe key features of networks. In the specific field of social sciences, a more detailed description of human interactions is crucial to understand the formation and evolution of complex social networks.

In this paper we describe a procedure to characterize the structure of networks, based on a recently proposed algorithm to identify *communities* in graphs [6]. Our procedure allows one to study quantitatively the hierarchical structure of nested communities in networks. Moreover we apply the procedure to a real social network. We define and analyze the complex email network of an organization with about 1700 employees and determine its community structure. Our results reveal that this network self-organizes into a self-similar structure, suggesting that some universal mechanism could be the underlying driving force in the formation and evolution of social networks, as happens in other complex systems [7,8].

Apart from work related reasons, ties between individuals in any organization arise, without external influence, due to personal, political, and cultural reasons, among others. The rapid development of electronic communications provides a powerful tool to analyze the informal self-organized social network arising as a result of the formation of such ties. Indeed, every time an email is sent, the addresses of the sender and the receiver are routinely registered in a server. Therefore, an *email network* can be built regarding each email address as a node and linking two nodes if there is an email communication between them. We take as a case study the email network of University at Rovira i Virgili (URV) in Tarragona, Spain, containing 1669 users including faculty, researchers, technicians, managers, administrators, and graduate students.

Bulk emails provide little or no information about how individuals or teams interact, so to minimize their effect: (i) we eliminate emails that are sent to more than 50 different recipients and (ii) we disregard links that are unidirectional, that is, we consider that two nodes A and B are connected only if A has sent an email to B and B has also sent an email to A. With these restrictions, the network is an undirected graph [26].

The cumulative degree distribution $P(k)$ of the email network—representing the probability that a node has k or more links to other nodes—is exponential

$$P(k) \propto \exp(-k/k^*) \tag{1}$$

for $k \geqslant 2$, with $k^* = 9.2$. This result is in contrast with recent findings indicating that some technology based social networks—such as rough email networks [9], the instant messaging network [10], or the PGP (pretty good privacy) encryption network [11]—which show heavily skewed degree distributions, but is consistent with the proposal of Amaral and co-workers that the truncation of the scale-free behavior in real world networks is due to the existence of limitations or costs in the establishment of connections [3,12]. Indeed, it seems plausible that there are costs to maintaining active social acquaintances and therefore active communications. However, it is relatively easy to keep many *electronic* acquaintances *open*, although most of them are probably inactive from a social point of view.

Out of the total 1669 nodes, 1133 belong to the giant component. The rest are isolated or, at most, connected by pairs. In the following, we focus on this giant component that can be characterized by statistical properties such as its clustering coefficient $C = 0.254$ and its average shortest path length $d = 3.606$ [1]. For comparison, we construct a random network with exactly the same exponential degree distribution as the email network following the procedure proposed in Ref. [13] (from now on we will call it random exponential

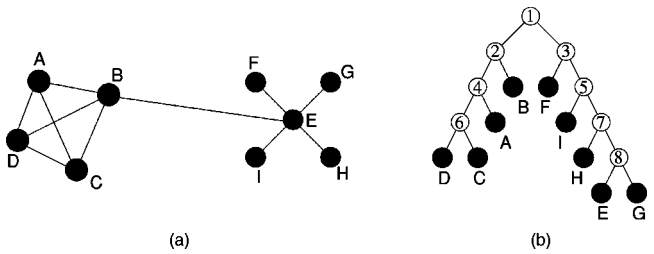

FIG. 1. Community identification according to the GN algorithm. (a) A simple network with two communities. (b) Binary tree generated by the GN algorithm. Each branch in the binary tree corresponds to a community in the original network and central nodes in a community, such as *E*, appear as the tips of the branches.

network). The clustering of the exponential network is approximately ten times smaller $C=0.028$ while the average shortest path length is very similar $d=3.317$, as happens in small world networks [1].

To understand the structure of the social network of the organization, we are interested in determining how individuals interact and form groups that, in turn, interact with each other giving rise to higher order groups, that is, groups of groups. In other words, we want to unravel the *community structure* of the network. To do so we use the algorithm proposed recently by Girvan and Newman (GN) [6] to identify communities in complex networks (see Fig. 1).

The GN algorithm proceeds as follows [6]. The betweenness of an edge is defined as the number of minimum paths connecting pairs of nodes that go through that edge [14,15]. The key idea is that the edges that connect highly clustered communities have a higher edge betweenness—edge *BE* in Fig. 1(a)—and therefore cutting these edges should separate communities. The algorithm identifies and removes the link with the highest betweenness in the network. After every removal, the betweenness of the edges is recalculated and the process is repeated until the ''parent'' network splits, producing two separate ''offspring'' networks. The offspring can be split recursively in the same way until they comprise only one individual.

In order to describe the entire splitting process, we generate a binary tree in which bifurcations (white nodes) depict communities and leaves (black nodes) represent individual addresses of the email network [Fig. 1(b)]. At the beginning of the process, the network in Fig. 1(a) is a single entity, represented by node 1 in the tree. After the removal of the edge *BE*, the network is split into two subnetworks, 2 and 3, containing nodes $A-D$ and $E-I$, respectively. Since the two offspring networks have no further internal community structure all the links within each have the same betweenness. In this case, one of them will be selected at random for removal. Iterating the link removal procedure, nodes will be separated randomly one by one by the GN algorithm, in such a way that each community will appear as a branch in the binary tree. It is important to note that central nodes, such as node *E*, will be separated last. This particular characteristic of the GN algorithm can be used with managerial purposes to detect those persons that act like hubs in the organization.

The community binary tree for URV is shown in Fig. 2. Each color in Fig. 2(a) corresponds to one center of the university, that is, to a department or college, or to management units such as the office of the Rector of the university. Two properties of the tree are worth noting. First, a clear branching structure emerges, with branches essentially containing nodes of the same color. This shows that the identification of communities is successful, despite the complexity of the interactions in the original email network. Second, the branching structure is far from simple. Indeed, each branch is formed, in general, by a system of nested smaller subbranches that give rise to a complicated structure that visually resembles some self-similar systems in nature such as river networks [16] or diffusion-limited aggregates [17]. For comparison, we also show the tree generated by the GN algorithm from the random exponential network [Fig. 2(c)]. In contrast to the tree for the URV email network, the branching structure is almost trivial with most of the branches containing only one or two nodes. This is the expected result for a network that does not have any sort of community structure.

Once the binary tree has been obtained, we look for a quantitative characterization of the community structure. First, we consider the cumulative community size distribution $P(s)$, that is, the probability of a community having a size larger or equal to *s*. Each node *i* of the binary tree represents a community—or a single email address. Its community size s_i is just the summation of the sizes of its two offspring j_1 and j_2: $s_i=s_{j_1}+s_{j_2}$. Figure 3(a) shows how to compute the sizes of all the communities in a simple binary tree and the corresponding probability distribution $P(s)$, that is, the probability that a community has size larger or equal to *s*. The community size distribution for the email network is presented in Fig. 3(d). The distribution is heavily skewed, following a power law behavior $P(s)\propto s^{-\alpha}$ with $\alpha=0.48$ between $s=2$ and $s\approx 100$. Beyond this value, the distribution shows a sharp decay and, at $s\approx 1000$ a cutoff that corresponds to the size of the system. The power law of the community size distribution suggests that there is no characteristic community size in the network (up to $s\approx 100$). To rule out the possibility that this behavior is due to the community identification algorithm, we also consider the community size distribution for a random exponential network and for a hierarchical network as proposed by Ravasz and Barabasi (RB) [18]. While the community size distribution of the random exponential network is completely different—with essentially no communities of sizes between 2 and 100— the behavior of the RB model is similar to the scaling presented by the email network. Therefore, it seems that the self-replicating structure of RB networks, which is implicit by construction, is a reasonable first approximation to the structure of the email network.

The characterization of the community binary tree using the cumulative size distribution has its analogy in the river network literature [16,20,21]. The equivalent measure is the distribution of drainage areas, which represents the amount of water that is generated upstream of a given point [see Fig. 3(b)]. The drainage area of a given point is the number of nodes upstream of it plus one. For a point *i* with offspring j_1 and j_2, the drainage area s_i is therefore $s_i=s_{j_1}+s_{j_2}+1$. The similitude between the community size distribution of the

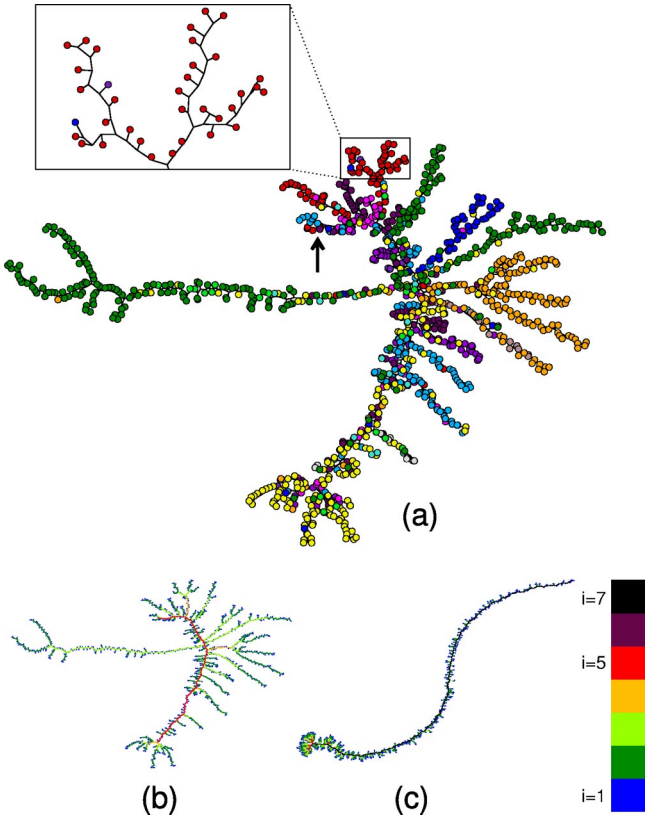

FIG. 2. (Color online) Communities in the email network of URV. (a) Binary tree showing the result of applying the GN algorithm to the email network of URV. The position indicated by the arrow represents the root of the tree [equivalent to node 1 in Fig. 1(b)] and branches are depicted so that they can be clearly differentiated. In particular, only the leaves of the tree, corresponding to email addresses, are shown, as in the zoomed detail. Colors depict different centers. (b) Same as before but without showing the leaves. Branches are now colored according to their Horton-Strahler index (see text). (c) Same as (b) for a random network. The lack of community structure is reflected in the absence of branches, in contrast with the intricate self-similar structure of (b).

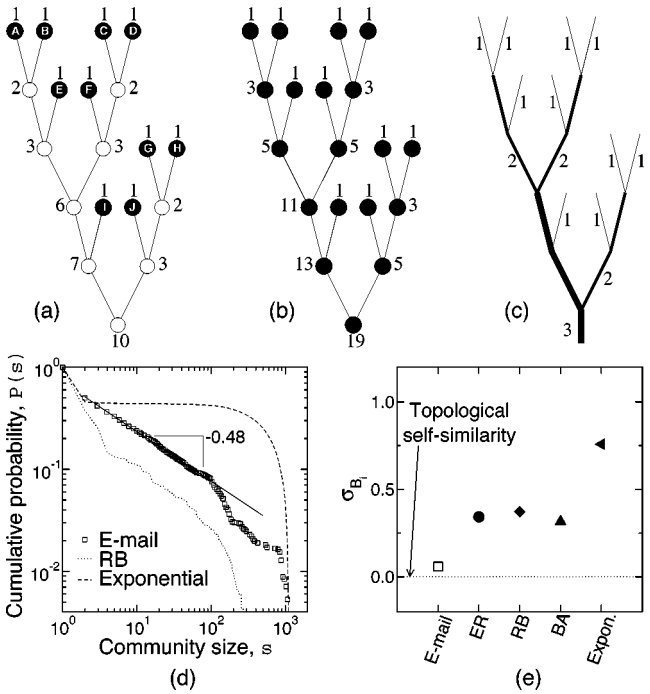

FIG. 3. Self-similarity in the community structure. (a) Calculation of the community size distribution for a binary tree generated by the community identification algorithm. Black nodes represent the actual nodes of the original graph while white nodes are just graphical representations of communities that arise as a result of the splitting procedure. Nodes *A* and *B* belong to a community of size 2, and together with *E* form a community of size 3. Similarly, *C*, *D*, and *F* form another community of size 3. These two groups together form a higher level community of size 6. Note that a single node belongs to different communities, i.e., different hierarchical levels. (b) Calculation of the drainage area distribution for a river network. (c) Calculation of the Horton-Strahler index. In this case, there are ten branches with index 1, three branches with index 2, and one branch with index 3. (d) Comparison between the distribution $P(s)$ of community sizes in the email network, in the random exponential network, and in the hierarchical network model proposed by Ravasz and Barabasi (RB), with $n=4$ and 5 levels [18]. (e) The standard deviation of the bifurcation ratios B_i for the email network, an Erdos-Renyi (ER) random graph with the same number of nodes and links [19], a hierarchical RB network [18], a scale-free network as proposed by Barabasi and Albert (BA) with the same size as the email network and $m=5$ [2], and the random exponential network. The community tree of the email network is topologically self-similar with $B=5.8$. Topological self-similarity does not hold for the other networks.

current email network in Fig. 3(d) and the area distribution of the Fella river network in Italy reported in Fig. 2 of Ref. [21] is striking. The exponent $\alpha=0.45$ for the power law region of this river and the average exponent for several rivers $\alpha_{river}=0.43\pm0.03$ reported by Refs. [20,21], respectively, are very close to the current $\alpha=0.48$. Moreover, the behavior shown in Fig. 3(d) with first a sharp decay and then a final cutoff is also shared by river networks, which are known to evolve to a state where the total energy expenditure is minimized [20,22,23]. The possibility that communities within organizations might also spontaneously self-organize into a form in which some quantity is optimized is very appealing and deserves further investigation.

To further understand this point, it is pertinent to ask whether there are other emergent properties shared by both. To answer this question we consider a standard measure for categorizing binary trees: the Horton-Strahler (HS) index, originally introduced for the study of river networks by Horton [24], and later refined by Strahler [25]. Consider the binary tree depicted in Fig. 3(c). The leaves of the tree are assigned a HS index $i=1$. For any other branch that ramifies into two branches with HS indices i_1 and i_2, the index is calculated as follows:

$$i=\begin{cases} i_1+1 & \text{if } i_1=i_2 \\ \max(i_1,i_2) & \text{if } i_1\neq i_2. \end{cases} \tag{2}$$

Note that the index of a branch changes when it meets a branch with higher index, or when it meets a branch with the same value and both of them join forming a branch with

higher index. In terms of communities, the interpretation of the HS index is the following. The index of a community changes when it joins a community of the same index. Consider, for instance, the lowest levels: individuals ($i=1$) join to form a group (or team, with $i=2$), which in turn will join other groups to form a *second level* group (or department, $i=3$). Therefore, the index reflects the *level* of aggregation of communities. The number of branches N_i with index i can be determined once the HS index of each branch is known. The bifurcation ratios B_i are then defined by

$$B_i = \frac{N_i}{N_{i+1}} \tag{3}$$

(by definition $B_i \geqslant 2$).

When $B_i \approx B$ for all i, the structure is said to be topologically self-similar, because the overall tree can be viewed as being composed of B subtrees, which in turn are composed of B smaller subtrees with similar structures and so forth for all scales [17]. River networks are found to be topologically self-similar with $3<B<5$ [17].

As a measure of topological self-similarity one can calculate the standard deviation σ_{B_i} of the bifurcation ratios B_i, which tends to 0 when topologically self-similarity holds. In Fig. 3(e), we compare σ_{B_i} of the email network with that of several model networks. We find that the community tree of the email network is topologically self-similar with $B \approx 5.8$ and $\sigma_B \approx 0.05$. All other network models significantly deviate from topological self-similarity. In particular, the hierarchical RB model [18], which has a similar scaling behavior as the email network [Fig. 3(d)], does not show topological self-similarity. The lack of topological self-similarity in this case is related, paradoxically, to scale-free connectivity distribution of the RM model, which makes the *central units* different from the peripheral ones.

By revealing the structure of the email network, the proposed methodology leads us to realize that community structure is self-similar. Self-similarity is a fingerprint of the replication of the structure at different levels of the social network, and could be the result of a trade-off between the need for cooperation and the costs of keeping active connections. Moreover, the emergence of scaling, as well as the similarity with river networks, raises important questions about the mechanisms underlying the interactions between individuals. As pointed out in a recent paper [8], the scaling properties of river networks are ubiquitous. By using the same argument, one can expect that the scaling behavior we obtain should be observable in any human social network. At the same time, the similarity with river networks suggests that a common principle of optimization—of flow of information in organizations or of flow of water in rivers—could be the underlying *driving force* in the formation and evolution of social networks.

We thank L. A. N. Amaral, M. Buchanan, X. Guardiola, and J. Ottino for helpful comments and suggestions. We also thank J. Tomas, O. Lorenzo, C. Llorach, J. Clavero, and F. Salvador, for collecting the email data. This work was supported by DGES of the Spanish Government (Grant Nos. PPQ2001-1519, BFM2000-0626, BEC2000-1029, and BEC2001-0980) and EC-Fet Open Project (Grant No. IST-2001-33555). R.G. and L.D. also acknowledge financial support from the Generalitat de Catalunya.

[1] D.J. Watts *et al.*, Nature (London) **393**, 440 (1998).
[2] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).
[3] L.A.N. Amaral, A. Scala, M. Barthelemy, and H.E. Stanley, Proc. Natl. Acad. Sci. U.S.A. **97**, 11 149 (2000).
[4] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[5] S. Dorogovtsev *et al.*, Adv. Phys. **51**, 1079 (2002).
[6] M. Girvan and M.E.J. Newman, Proc. Natl. Acad. Sci. U.S.A. **99**, 7821 (2002).
[7] J. Banavar *et al.*, Nature (London) **399**, 130 (1999).
[8] M. Buchanan, Nature (London) **419**, 787 (2002).
[9] H. Ebel *et al.*, Phys. Rev. E **66**, 035103 (2002).
[10] R. Smith (unpublished); R. Smith, e-print cond-mat/0206378.
[11] X. Guardiola *et al.* (unpublished); X. Guardiola e-print cond-mat/0206240.
[12] M.E.J. Newman *et al.*, Phys. Rev. E **66**, 035101(R) (2002).
[13] M.E.J. Newman *et al.*, Phys. Rev. E **64**, 026118 (2001).
[14] S. Wasserman and K. Faust, *Social Network Analysis* (Cambridge University Press, Cambridge, 1994).
[15] M.E.J. Newman, Phys. Rev. E **64**, 016132 (2001).
[16] I. Rodriguez-Iturbe and A. Rinaldo, *Fractal River Basins: Chance and Self-organization* (Cambridge University Press, Cambridge, 1996).
[17] T.C. Halsey, Europhys. Lett. **39**, 43 (1997).
[18] E. Ravasz and A.-L. Barabasi, Phys. Rev. E **67**, 026112 (2003).
[19] B. Bollobas, *Random Graphs*, 2nd ed. (Cambridge University Press, Cambridge, 2001).
[20] A. Rinaldo *et al.*, Phys. Rev. Lett. **70**, 822 (1993).
[21] A. Maritan *et al.*, Phys. Rev. E **53**, 1510 (1996).
[22] S. Kramer and M. Marder, Phys. Rev. Lett. **68**, 205 (1992).
[23] K. Sinclair and R.C. Ball, Phys. Rev. Lett. **76**, 3360 (1996).
[24] R.E. Horton, Bull. Geol. Soc. Am. **56**, 275 (1945).
[25] A.N. Strahler, Bull. Geol. Soc. Am. **63**, 923 (1952).
[26] We consider that the network is also unweighted, disregarding the fact that some links correspond to a single interchange of emails and others to the interchange of many emails.