The real communication network behind the formal chart: community structure in organizations [*]

R. Guimerà [†] L. Danon [‡] A. Díaz-Guilera [§]

F. Giralt [¶] A. Arenas [‖]

September 21, 2004

[*] We thank L. A. N. Amaral, X. Guardiola, and J. Ottino for helpful comments and suggestions. We thank J. Tomas, O. Lorenzo, C. Llorach, J. Clavero and F. Salvador for collecting the e-mail data. This work has been supported by DGES of the Spanish Government (Grants No. PPQ2001-1519, No. BFM2000-0626, and No. BFM2003-08258) and EC-Fet Open project (No. IST-2001-33555). R.G. and L.D. acknowledge financial support from the Generalitat de Catalunya.

[†] Department of Chemical Engineering, Northwestern University, Evanston, IL, USA. e-mail: rguimera@northwestern.edu

[‡] Departament de Física Fonamental, Universitat de Barcelona, Barcelona, Catalonia, Spain. e-mail: ldanon@ffn.ub.es

[§] Corresponding author. Departament de Física Fonamental, Universitat de Barcelona, Marti i Franques 1, 08028, Barcelona, Catalonia, Spain. e-mail: albert.diaz@ub.edu. Phone: +34-93-402-1167. Fax: +34-93-402-1149

[¶] Departament d'Enginyeria Química, Universitat Rovira i Virgili, Tarragona, Catalonia, Spain. e-mail: fgiralt@etseq.urv.es

[‖] Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Tarragona, Catalonia, Spain. e-mail: aarenas@etse.urv.es

Abstract

The formal chart of an organization is intended to prescribe how employees interact. However, ties between individuals arise for personal, political, and cultural reasons. The characterization of the structure of such *informal networks* behind the formal chart is a key element for successful management. We surmise that the exchange of e-mails between individuals in organizations reveals how people interact and therefore provides a map of the real network structure behind the formal chart. We propose a methodology that allows to extract relevant information about the community structure of an organization from the network of e-mail interactions between its employees.

JEL: L29,M12,M54
Key words: email network, informal charts, complex networks

The real communication network behind the formal chart: community structure in organizations

1 Introduction

The network of interactions within an organization is considerably more complex than implied in the formal chart . Due to a variety of reasons such as resolving problems of an unexpected nature, personal and cultural similarities, and political motives, new connections are being made between individuals all the time. Understanding such informal networks and how they operate is necessary for successful management [Krackhardt and Hanson (1993), Mayo (1949), Morgan (1997)]. Traditionally, informal network study is performed in two steps [Krackhardt and Hanson (1993)]. First, employee questionnaires are used to survey the network. However, employees' answers often contain subjective elements such as "political" motives and the worry about offending colleagues. This effect can be minimized by the second step: cross-checking of the answers which is not free of subjectiveness either. A more significant limitation of the questionnaire based analysis is that time and effort costs make it prohibitively expensive to map the entire network even for medium sized organizations.

Rapid development of electronic communications provides a powerful alternative for studying informal networks. The exchange of e-mails between individuals in organizations is a good indicator of who is linked to who and it should contain information not found in the formal chart [Economist (2001), Ebel et al. (2002), Adamic and Adar (2002), Guimera et al. (2003)]. This is interesting not only from a managerial point of view, but also from fundamental and theoretical points of view [Radner (1993), Garicano (2000), Arenas et al. (2001), Guimera et al. (2002)].

However, extracting information from communication networks is not straightforward. For instance, by analyzing an e-mail network it is not possible to discriminate between different sorts of informal networks. Krackhardt and Hanson [Krackhardt and Hanson (1993)] stressed the differences between informal networks (advice network, trust network, etc.) and the importance of knowing them separately. In an e-mail network all the informal networks and even the formal chart contribute, interacting in a complex way. Nevertheless, the information obtained from communication network studies is still valuable. Another problem is that extraction of information from large and complex networks is not straightforward. Specific statistical techniques, developed recently in the field of statistical mechanics of complex networks need to be used [Watts and Strogatz (1998), Barabási and Albert (1999), Amaral et al. (2000), Albert and Barabási (2002), Dorogovtsev and Mendes (2002), Girvan and Newman (2002), Newman (2002)].

In this paper we describe a novel procedure to characterize the structure of networks, based on a recently proposed algorithm to identify *communities* in graphs [Girvan and Newman (2002), Guimera et al. (2003), Newman (2003)]. Our procedure allows to study quantitatively the hierarchical structure of nested communities in networks. Moreover we apply the procedure to a real network. From more than one million e-mails, we build and analyze the complex e-mail network of an organization with about 1,700 employees—the Universitat Rovira i Virgili, at Tarragona, Catalonia—and determine its community structure.

In the next section we describe how the network is built and study some of its statistical properties, such as the degree distribution [Barabási and Albert (1999), Amaral et al. (2000)], the clustering coefficient [Watts and Strogatz (1998)] and the assortativity [Newman (2002)]. Next, we describe how one can obtain insight into the community structure of the network and present this information in a useful way from a managerial point of view. Finally, we study the properties of the community structure. Surprisingly, we find that it shows emergent self-similar properties as occurs in other natural systems like, for example, river networks [Guimera et al. (2003), Gleiser and Danon (2003)].

2 Characterization of the e-mail network

Every time an e-mail is sent, some information is routinely registered in a server, including the addresses of the sender and the receiver. Using this information an e-mail network can be built with each address being represented by a node and each e-mail by a link between nodes. Since e-mail is directed, the links and therefore the entire network are directed.

At Universitat Rovira i Virgili (URV), there are three different servers that manage the e-mail accounts of all the staff (professors, technicians, managers, administrators, graduate students, etc.). The total number of users is approximately 1700[1]. To study this network, only e-mails sent within the university during the first three months of 2002 were considered. Out of a total of 1135818 e-mails registered by the servers, we restrict the analysis to e-mails with both sender and receiver belonging to the university. The resulting e-mail network is shown in figure 1. For the purpose of the present analysis, we disregard the fact that some links are much more active than others in the e-mail network. Such a consideration has been taken into account, for example, in [Caldarelli et al (2003)]. The analysis of other complex weighted networks has also revealed some interesting results [Barrat et al. (2004)].

2.1 Degree distribution

First, we consider the cumulative degree distribution, $P(k)$, that gives the probability of a certain node having more than k links[2]. Since e-mails can be

[1] A random code is assigned to each address, thus preserving the anonymity of the users.

[2] The cumulative distribution $P(k)$ is simply related to the probability density function $p(x)$ by $P(k) = \int_{-\infty}^{k} dx\, p(x)$. In particular, if $p(x)$ is a power law $p(x) \sim x^{-\alpha}$ then

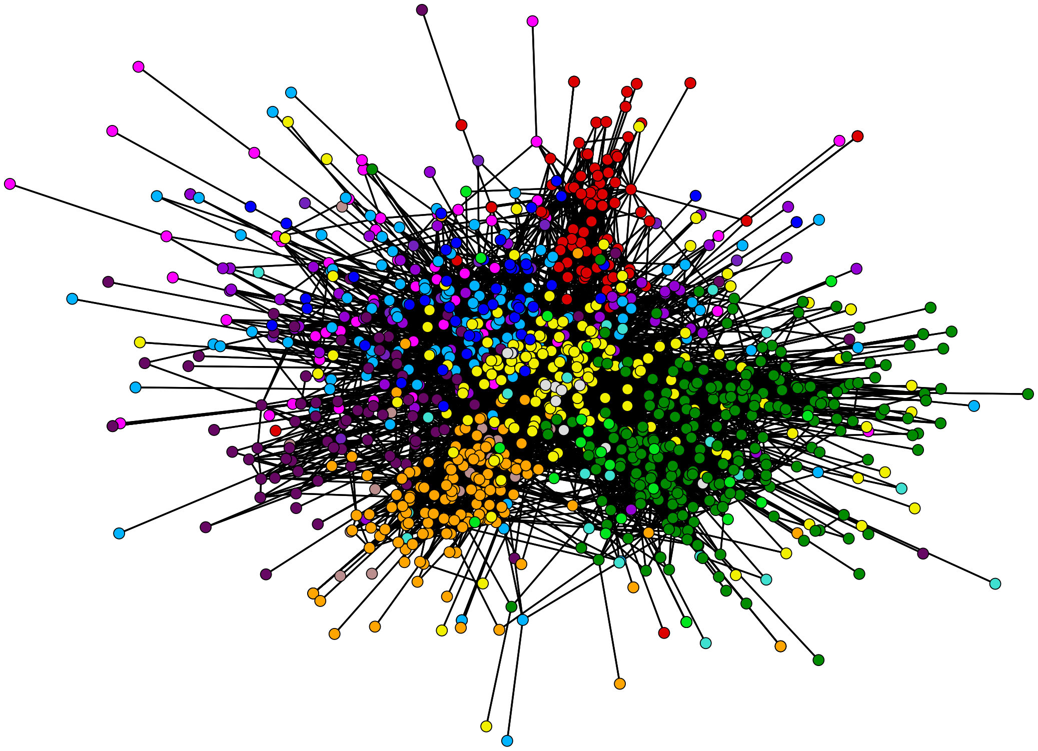
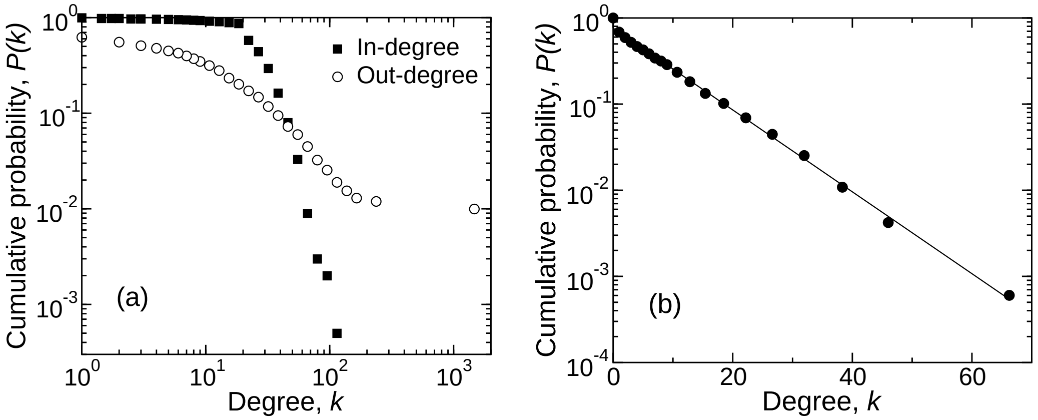


Figure 1: The e-mail network of URV. Each individual is represented by a colored node, with two individuals (A and B) being connected if A sends an e-mail to B and B replies, or vice versa. Each color corresponds to an individual's affiliation to a specific center within the university. E-mails sent to or received from outside are not considered and isolated nodes are not shown.

sent or received, these links are directed: incoming links account for received e-mails and outgoing links for sent e-mails. In principle, the in- and out-degree distributions are measured considering all the e-mails sent within the university: nodes are addresses and a directed link indicates that at least one e-mail has been sent from one address to another. The resulting degree distribution is shown in figure 2a. The asymmetry between the distribution of incoming (received) and outgoing (sent) e-mails is apparent from the plot. While the maximum in-degree (that is, the maximum number of users that are sending e-mails to the same address) is about 100, the maximum out-degree (that is, the maximum number of addresses that a given user is sending e-mails to) is more than 1000. Actually, the in-degree distribution decays very fast, while the out-degree distribution is highly skewed, since a few nodes send e-mails to more than 1000 different addresses.

The origin of the highly skewed out-degree distribution in figure 2a is related to the existence of e-mail lists, that is to the fact that some users send the same e-mail to a list of users (these lists can eventually include almost everyone in the university). No matter what the origin of the e-mails is, the skewness of the degree distribution will be crucial, for instance, when dealing

$P(k) \sim k^{-\alpha-1}$, and if $p(x)$ is an exponential $p(x) \sim \exp(-x/k^*)$ then $P(k) \sim \exp(-k/k^*)$.



(a) (b)

Figure 2: Degree distribution of the e-mail network of URV. (a) In- and out-degree distributions when all internal e-mails are considered. The out-degree distribution is highly skewed due to the presence of e-mail lists. (b) Total distribution when e-mails sent to more than 50 users are discarded and only bidirectional mails are considered. In this case, the distribution decays exponentially.

with virus propagation [Pastor-Satorras and Vespignani (2001)]. However, from our perspective bulk e-mails provide little or no information about how individuals or teams collaborate. To minimize the effect of *spam* e-mail: (i) we eliminate e-mails that are sent to more than 50 different recipients and (ii) we disregard links that are unidirectional, that is we consider only e-mails that represent a real communication link, where e-mails flow in both directions. With these two restrictions, the network becomes undirected (since all links are now bidirectional) and the degree distribution is exponential $P(k) \propto \exp(k/k^*)$, with $k^* \approx 10$ (figure 2b). This result is consistent with the proposal of Amaral and coworkers 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 [Amaral et al. (2000)]. It seems plausible that there exist limitations to maintain an arbitrarily large number of active social acquaintances. However, it is relatively easy to keep many *electronic* acquaintances *open* (although most of them are probably inactive from a social point of view) giving rise to heavily skewed degree distributions as happens in technology based social networks such as rough e-mail networks [Ebel et al. (2002)], the Instant Messaging Network [Smith (2002)] or the PGP encryption network [Guardiola et al. (2002)].

2.2 Assortativity

It has been stated that computer networks can be regarded in many cases as true social networks [Wellman (2001)]. To test whether the e-mail network indeed shows properties of a social network we use the measure recently proposed by Newman [Newman (2002)]. He reports that in social networks (for example co-authorship networks) highly connected nodes usually tend to be connected to other highly connected nodes, or in other words, there are positive correlations in the degree-degree correlation function. Conversely, technological and biological networks (for example the Internet or food webs) seem to show negative correlations. To classify networks according to this scheme, the Pearson correlation coefficient, r, is used: networks with $r > 0$—like social networks—are called assortative while networks with $r < 0$—like technological and biological networks— are called disassortative. The e-mail network of URV yields a value $r = 0.079$ suggesting that it is a weak assortative social network, even though this value is small to be considered conclusive.

2.3 Clustering coefficient

Next, we consider the clustering coefficient, C, of the network [Watts and Strogatz (1998)], which quantifies the *transitivity* of the network: if A is connected to B and C, the clustering coefficient gives the probability that B and C are also connected to one another. In the following analysis, we focus on the largest connected cluster of the URV e-mail network that contains 1133 nodes. The remaining nodes are isolated and will not be considered from now on. We find that the value of the clustering coefficient $C = 0.254$, which is approximately 30 times larger than the expected value for a random graph of the same size and average degree. Such a high value of C suggests a scenario where the network is comprised of several highly connected communities—with a lot of redundancy in the linking—which are loosely connected to other highly connected communities. In fact it has recently been shown that there is a close relation between highly clustered regions of a graph and the existence of communities [Eckmann and Moses (2002)]. In the next sections, we focus on the identification of such communities and on the characterization of their structure.

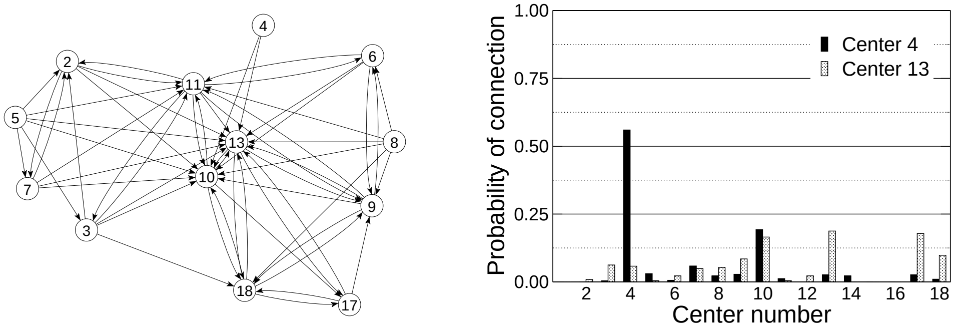
3 The hierarchical community structure

3.1 Interaction between formal communities within the organization

Any organization with more than a few employees is formally divided into *units* or *centers* that carry out different parts of the *production process*. In the case of a university, these centers are the colleges—for example, the School of Chemical Engineering—and the administrative and *support* units—such as the office of the President of the university or the library service, respectively.

Before we move to the identification and analysis of the real communities in the organization, we want to show that the e-mail network provides useful information about how formal centers actually interact with one another.

First, we focus on the average distance between centers. We take each node in the network and measure the number of steps across the e-mail network needed to reach any other node. Then we average over all the



(a) (b)

Figure 3: Interactions between centers from the e-mail network. (a) Inter-center relations from distances in the e-mail network. A directed link is established from A to B when the average distance between nodes in A and B is short (see text). Five small centers with less than 10 persons have been disregarded. (b) Probability of being connected to nodes that belong to other centers. Each bar represents the probability of a node in center 4 or 13, to be connected with a node in another center.

nodes in the same center and obtain average distances between centers. To visualize this information, we proceed as follows. First, we calculate the distance from one center A to all other centers, d_{AB}, d_{AC}, etc. Then we compute the average distance from A to the other centers $\langle d_A \rangle$. Finally, node A (that now represents a center, not an individual) is linked to another node B if $d_{AB} < \langle d_A \rangle$. In this case, the network is directed because, in general, $d_{AB} < \langle d_A \rangle$ does not imply $d_{BA} = d_{AB} < \langle d_B \rangle$. This is presented in figure 3 where numbers correspond to different centers in the university. Three central communities (10, 11 and 13) can be clearly identified and should correspond to central offices and administrative centers of the university. These three interact on the left with a group of four centers and on the right with another one formed by five centers. Significantly, only one direct link connects the centers on the left to those on the right. There is one center that is only connected to two of the central nodes and somehow isolated from the rest of the university. No further comments can be made here due to confidentiality constrains.

The second measure of interaction between centers is the probability of nodes in a center being connected to nodes in other centers. For each node in the network, we just regard its neighbors and, again, we average over all the nodes that belong to the same center. Two typical cases are shown in figure 3b.

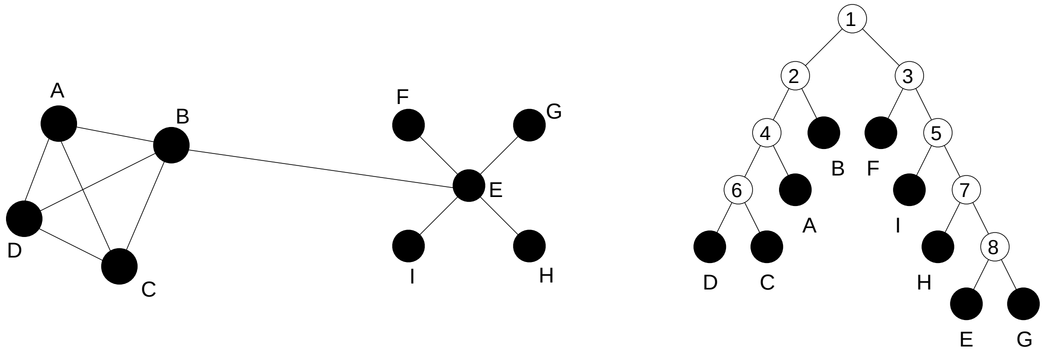
Center 13 is one of the three central nodes in figure 3a. As can be seen from figure 3b, individuals that belong to center 13 are connected with a reasonably high probability not only to other individuals in the same center but also to individuals belonging to most of the other centers. Conversely, individuals in center 4 are mostly connected to others from the same center and also to individuals in center 10, that has been already identified as a central management unit. Extreme cases of these two patterns could be considered pathological: groups with lots of outside connections and very few internal connections are said to show anomalous communication patterns, while groups with an extremely high fraction of internal connections but weakly connected to other groups are said to show imploded relationships [Krackhardt and Hanson (1993)].

3.2 Identification of *real* communities

As well as many other types of networks, social networks have "community structure", that is, the combination of regions with a high density of vertex-vertex connections and other regions which are very sparse. In an organization, such community structure should correspond, to some extend, to the formal chart. However, ties between individuals in an organization also arise due to personal, political and cultural reasons, giving rise to informal communities and to an *informal community structure*. The understanding of informal networks underlying the formal chart and of how they operate are key elements for successful management.

The traditional method for identifying communities in networks is hierarchical clustering [Jain and Dubes (1988)]. The idea is the following. Quantify first how closely connected is each pair of nodes in the network. Then create an empty network with all the nodes but no links between them, and start adding links between the nodes that are more closely connected. This procedure gives rise to a nested set of increasingly large components.

In this work we use a different community identification algorithm, proposed recently by Girvan and Newman (GN) [Girvan and Newman (2002)]. This new algorithm gives successful results even for networks in which hierarchical clustering methods fail [Girvan and Newman (2002)]. The algorithms works as follows. The betweenness of an edge is defined as the number of minimum paths connecting pairs of nodes that go through that edge [Wasserman and Faust (1994), Newman (2001)]. The GN algorithm is based on the idea that the edges which connect highly clustered communities have a higher edge betweenness—for example edge BE in figure 4a—and therefore cutting these edges should separate communities. Thus, the algorithm proceeds by identifying and removing the link with the highest betweenness in the network. This process is repeated (should it be necessary) until the 'parent' network splits, producing two separate 'offspring' networks. The offspring can be split further in the same way until they comprise of only one individual. In order to describe the entire splitting process, we generate a binary tree, in which bifurcations (white nodes in figure 4b) depict communities and leaves (black nodes) represent individuals. All the information about the community structure of the original network can be deduced from the topology of the binary tree constructed in this fashion.



(a) (b)

Figure 4: Community identification according to the GN algorithm. (a) A network containing two clearly defined communities connected by the link BE. This link will have the highest betweenness, since to get from any node in one community, to any node in the other, this link needs to be used. Therefore it will be the first link to be cut, splitting the network in two. The process of cutting this link corresponds to the bifurcation at the highest level of the binary tree in (b). Since there is no further community structure in the offspring networks, the rest of the nodes will be separated one by one, generating a binary tree with two branches corresponding to the two communities. For the community on the right, the most central node will be separated last. In general, branches of the binary tree correspond to communities of the original network and the tips of these branches correspond to the leaders of the communities.

3.3 Graphical representation of the hierarchical community structure

Consider again the network in figure 4a. At the beginning of the process, no links have been removed and the whole network is represented by node 1 in the binary tree of figure 4b. When edge BE is removed, the network splits in two groups: group 2, containing nodes A to D, and group 3, containing nodes E to I. After this first splitting, two completely separate communities are left, a very homogeneous one and a very centralized one. One can check that in both cases the algorithm will separate nodes one by one giving rise to two different branches in the binary tree. Actually, when communities with no further internal structure are found, they are disassembled in a very uneven way giving rise to branches. In other words, the impossible task of identifying communities from the original network is replaced by the easy task of identifying branches in the binary tree. When centralized network structures are treated, the central node(s) will appear at the end of the branch. This provides a method to identify which are the leaders of each community.

Figure 5a depicts the binary tree that results from the application of the GN algorithm to the e-mail network of URV. As in figure 1, each color corresponds to an individual's affiliation to a specific center within the university. Centers are in most of the cases faculties or colleges—for example the School of Engineering—and are usually comprised of departments—for example, the Department of Chemical Engineering or the Department of Mechanical Engineering. In turn, departments are divided into research teams—for instance, the group of Transport Phenomena or the group of Biotechnology in the Department of Chemical Engineering.

Instead of plotting the binary tree with the root at the top as in figure 4b, it is plotted optimizing the layout so that branches, that represent the real communities, are as clear as possible. Actually, the root is located at the position indicated with the arrow in the upper left region of the tree. From there downward, branches are separated at both sides until only yellow nodes (at the bottom) are left. Significantly, yellow nodes, that correspond to center 10, already appear at the center of figure 1 and at the center of figure 3a, suggesting that center 10 is an important management unit of the university. The branches obtained by the GN procedure (figure 5) are essentially of one color, indicating that we have correctly identified the centers of the university. This is especially true if one focuses on the ends of the branches since, as discussed above, these ends correspond to the most central nodes in the community. In regions close to the origin of the branches, the coexistence of colors corresponds to the boundary of a community. It is important to note that the GN algorithm is able to resolve not only at the level of centers, but is also able to differentiate groups (sub-branches) inside the centers, i.e., departments and even research teams.

For comparison, we also show the tree generated by the GN algorithm from a random graph of the same size and degree distribution as the e-mail network (figure 5c). The absence of community structure is apparent from

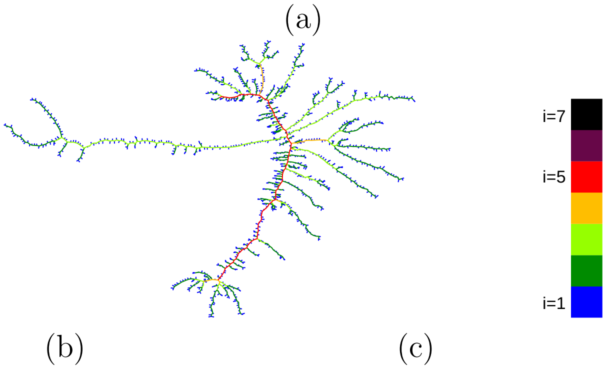


Figure 5: (a) Binary tree showing the result of applying the GN algorithm and our visualization technique to the e-mail network of URV. Each branch corresponds to a real community and the tips of the branches correspond to their leaders. The splitting procedure starts in the position indicated by an arrow at the top of the drawing and proceeds downward. The color of the nodes represents different centers within the university (five small centers containing less than 10 individuals are assigned the same color). Nodes of the same color (from the same center) tend to stick together meaning that individuals within the same center tend to communicate more, and that the algorithm is capable of resolving separate centers to a good degree of accuracy. (b) Same as before but without showing the nodes, so that the structure of the tree is clearly shown. Branches are colored according to their Horton-Strahler index (see text) (c) Binary tree showing the result of applying the GN algorithm to a random graph with the same size and degree distribution than the e-mail network. Again, colors correspond to Horton-Strahler indices.

the plot.

4 Emergent properties of the community structure

Next, we analyze the statistical properties of the community structure of the university. We will show that there are some self similar properties emerging in the network.

4.1 Community size distribution

The first quantity that will be considered is the community size distribution. Figure 6a represents a hypothetical tree generated by the community identification algorithm (for clarity, the tree is represented *upside down*). Black nodes represent the actual nodes of the original graph while white nodes are just graphical representations of groups that arise as a result of the splitting procedure. Indeed, 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

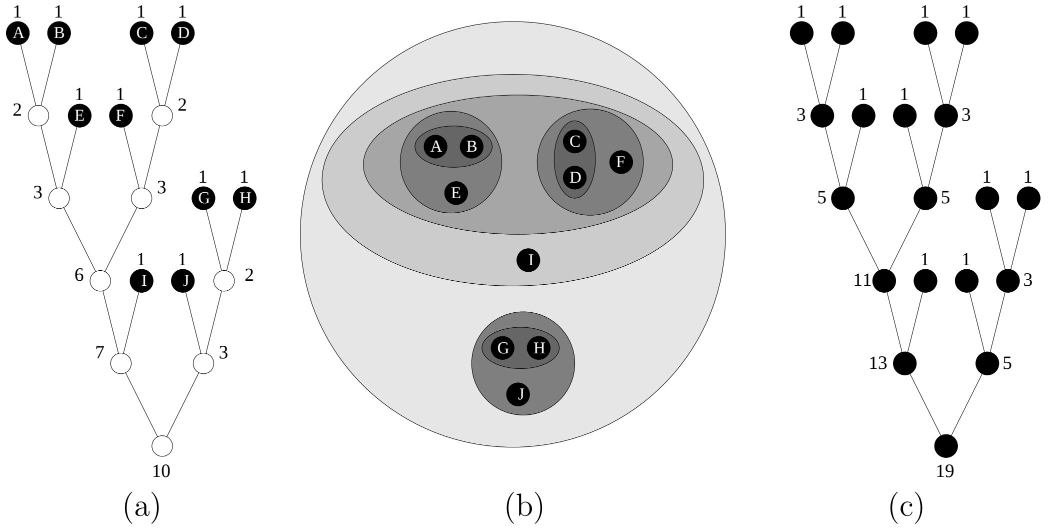


Figure 6: Community size distribution and analogy with river networks. (a) Calculation of community sizes from the community binary tree. (b) Representation of the hierarchical structure of nested communities. (c) Calculation of the drainage area distribution for a river network.

another community of size 3. These two groups together form a higher lever community of size 6. Following up to higher and higher levels, the community structure can be regarded as the set of nested groups depicted in figure 6b. A natural way of characterizing the community structure is to study the community size distribution. In figure 6a, for instance, there are three communities of size 2, three communities of size 3, one community of size 6, one community of size 7, and one community of size 10. Note that a single node belongs to different communities at different levels.

Figure 7 displays the heavily skewed cumulative distribution of community sizes, $P(s)$. It follows 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 $s \approx 1000$ the distribution shows a cutoff that corresponds to the size of the system (the whole network containing 1133 nodes). The power law of the community size distribution suggests that there is no characteristic community size in the network (up to size 100). To rule out the possibility that this behavior is due to the community identification algorithm we also considered the community size distribution for a random graph with the same size and degree distribution as the e-mail network. In this case (dotted line in figure 7), $P(s)$ shows a completely different behavior, with no communities of sizes between 10 and 600, as indicated by the plateau in figure 7. This corresponds to a situation in which all the branches (communities) are quite small (of size less than 10) with the backbone of the network formed by the union of all this small branches.

Some analytical approaches have been proposed in the literature to estimate the exponent of the size distribution of binary trees [De los Rios (2001)]. Although these approaches are interesting and could be adapted to calculate the exponents in our distributions, this is out of the

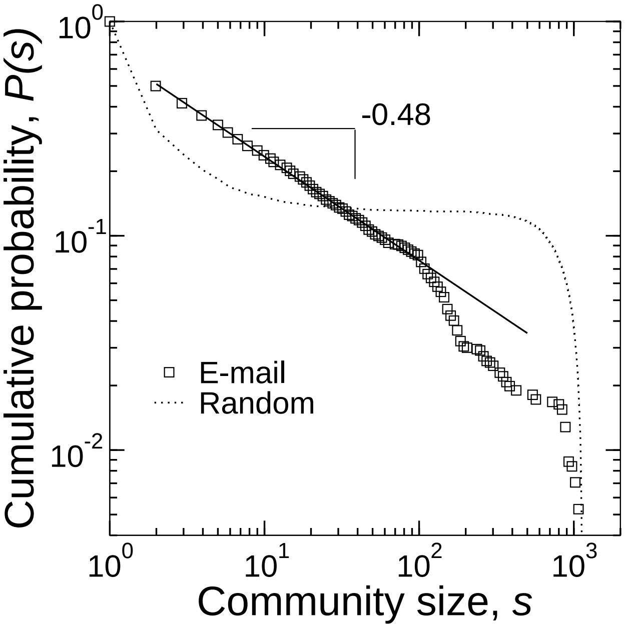


Figure 7: The distribution of community sizes, $P(s)$, showing a power law region with the exponent -0.48, followed by a sharp decrease at $s \approx 100$ and a cutoff corresponding to the size of the system at $s \approx 1000$. The distribution of community sizes in a random network is shown for comparison.

scope of the present paper.

4.2 Analogy with river networks

Figure 7 suggests a similarity between the distribution of community sizes and the distribution of drainage areas in river networks [Rinaldo et al. (1993), Rodriguez-Iturbe and Rinaldo (1996), Maritan et al. (1996), Banavar et al. (1999)]. This similarity can be understood by considering how this distribution is obtained from the community identification binary tree. Let us assign, as shown in figure 6a, a value of 1 to all the leaves in the binary tree or, in other words, to all the nodes that represent single nodes in the original network (black nodes of the binary tree). Then, the size of a community i, s_i, is simply the sum of the values s_{j_1} and s_{j_2} of the two communities (or individual nodes), j_1 and j_2, that are the offspring of i. Figure 6c shows how the drainage area of a given point in a river network is calculated. Consider that at any *node* of the river network there is a source of 1 unit of water (per unit time). Then, the amount of water that a given node drains is calculated exactly as the community size for the community binary tree, but adding the unit corresponding to the water *generated* at that point: $s_i = s_{j_1} + s_{j_2} + 1$. This quantity represents the amount of water that is generated upstream of a certain node. In this scenario, the community size distribution would be equivalent to the drainage area distribution of a river where water is generated only at the leaves of the branched structure.

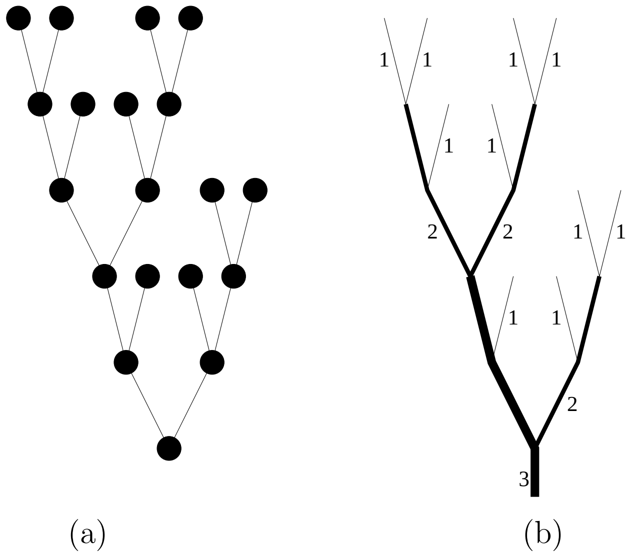


Figure 8: Calculation of the Horton-Strahler index. (a) Asymmetric binary tree (b) Corresponding Horton-Strahler indices of the leaves and branches. In this case there are $N_1 = 10$ branches with index 1, $N_2 = 3$ with index 2 and $N_3 = 1$ with index 3.

The similarity between the community size distribution of the e-mail network and the area distribution of a river network is striking (see, for instance, the data reported in [Maritan et al. (1996)] for the river Fella, in Italy). The exponent of the power law region is very close to the one obtained for the community size distribution: according to [Rinaldo et al. (1993)], $\alpha_{river} = 0.43 \pm 0.03$, while for the community size distribution we obtain $\alpha = 0.48$. Moreover, the behavior with first a sharp decay and then a final cutoff is also shared. River networks are known to evolve to a state where the total energy expenditure is minimized [Kramer and Marder (1992), Rinaldo et al. (1993), **?**]. The possibility that communities within networks might also spontaneously organize themselves into a form in which some quantity is optimized is very appealing and deserves further investigation.

4.3 Horton-Strahler index

The similarity between the community size distribution and the drainage area distribution of river networks prompts one question: is this similarity arising just by chance or are there other emergent properties shared by community trees and river networks? 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 [Horton (1945)], and later refined by Strahler [Strahler (1952)]. Consider the binary tree depicted in the left side of figure 8. The leaves of the tree are assigned a Strahler index $i = 1$. For any other branch that ramifies into two branches

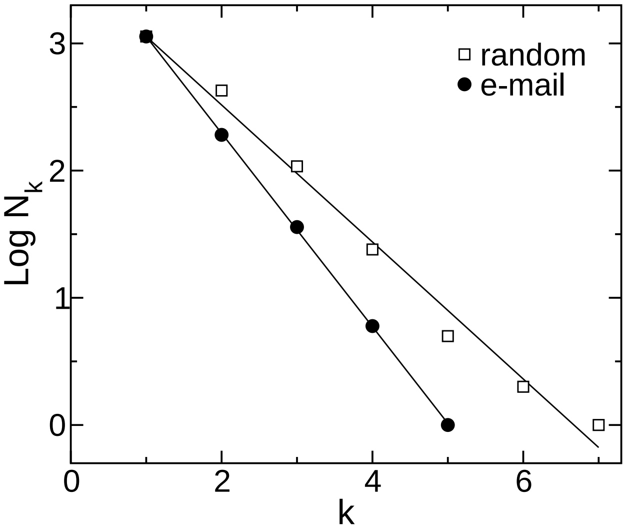


Figure 9: The number of branches with HS index i, as a function of i. From the definition of B_i, it is straightforward to show that, when topological self similarity holds, $N_i = N_1/B^{i-1}$. A fitting of this function to the points obtained for the e-mail community tree yields excellent agreement with $B = 5.76$. A much worse agreement is obtained for the community tree corresponding to the random network, with B_i fluctuating around 3.46.

with Strahler indices i_1 and i_2, the index is calculated as follows:

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

Therefore 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 (see 8b).

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 $B_i = N_i/N_{i+1}$ (by definition $B_i \geq 2$). 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 comprised of B sub-trees, which in turn are comprised of B smaller sub-trees with similar structures and so forth for all scales. River networks are found to be topologically self similar with $3 < B < 5$ [Halsey (1997)].

We find that the community tree as generated by the process described above is topologically self similar with $B_i \approx B = 5.76$ (see figure 9). The same analysis for the communities in a random graph shows that topological self similarity does not hold, since the values B_i are not constant; they fluctuate around a smaller value 3.46.

The HS index also turns out to be an excellent measure to assess the levels of complexity in organizations. First, let us consider the interpretation of the index in terms of communities within an organization. The index of a branch remains constant until another segment of the same magnitude is found. In other words, 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 (with $i = 2$), which in turn will join other groups to form a *second level* group ($i = 3$). Therefore, the index reflects the *level* of aggregation of communities. For example, in URV one could expect to find the following levels: individuals ($i = 1$), research teams ($i = 2$), departments ($i = 3$), faculties and colleges ($i = 4$), and the whole university ($i = 5$). Strikingly, the maximum HS index of the community tree is indeed 5, as shown in figure 9.

Figure 5b shows the community tree of the e-mail network with different colors for different HS indices. This helps to distinguish the individual, team and department levels within a branch. Actually, the *university level* is the "backbone" of the network along which the separation of communities occurs (from the top to the bottom of the figure). From this backbone, colleges, departments and some research teams separate, although it is worth noting that colleges or, in general, centers which are small and have no internal structure will be classified with a HS index corresponding to a department or even a team. Therefore, the HS index does not represent administrative hierarchy but organizational complexity. For comparison figure 5c shows in color the HS index for the binary tree of a random graph.

The fact that the community structure is topologically self-similar means that the organization is similar at different levels. In other words, it means that individuals form teams in a way that resembles very much the way in which teams join to form departments, to the way in which departments organize to form colleges, and to the way in which the different colleges join to form the whole university.

5 Conclusions

In this paper we have shown how to extract valuable information describing real complex networks behind the formal chart of an organization. We take advantage of the automatic registration of communication processes, in particular e-mails log files, to reconstruct the real network of interactions within the organization. This complex network is unraveled by the identification of the whole hierarchy of communities that individuals form at all levels within the organization. We propose a representation procedure that allows the identification of these communities by visual inspection and the determination of their level in the hierarchy using the Horton-Strahler index. To demonstrate the viability of the analysis, we study the e-mail network of the University Rovira i Virgili (Tarragona, Spain). From this analysis, we are able to identify the real organization of the individuals of the university into working teams, departments, faculties or colleges, and the whole university, as well as the interrelations between them.

We argue that this 'informal' organization of individuals into communities inside the university could be useful for management purposes, for example, to assess formal charts or to measure the degree of attainment over time of proposed organizational changes. The same analysis could be performed over the logs of phone calls, internal ordinary mail, faxes, etc.

From a theoretical point of view, the methodology identifies emerging properties in the community structure and we find a striking analogy with river networks. Although the study of this similarity should be performed over more and different organizations, we speculate that a common principle of optimization (of flow of information in our case or of flow of water in rivers) could be the underlying driving force in the formation and evolution of informal networks in organizations.

Note: After the submission of this paper, several other studies have stressed the importance of e-mail networks as a tool to understand the structure of social organizations[Tyler et al (2003), Eckmann et al (2003), Wu et al (2003), Caldarelli et al (2003)].

References

[Adamic and Adar (2002)] Adamic, L. A., Adar, E., 2002. Friends and neighbors on the web. Unpublished, http://citeseer.nj.nec.com/380967.html.

[Albert and Barabási (2002)] Albert, R., Barabási, A.-L., 2002. Statistical mechanics of complex networks. Reviews of Modern Physics 74, 47–97.

[Amaral et al. (2000)] Amaral, L. A. N., Scala, A., Barthelemy, M., Stanley, H. E., 2000. Classes of small-world networks. Proc. Nat. Acad. Sci. USA 97, 11149–11152.

[Arenas et al. (2001)] Arenas, A., Diaz-Guilera, A., Guimera, R., April 2001. Communication in networks with hierarchical branching. Phys. Rev. Lett. 86, 3196–3199.

[Banavar et al. (1999)] Banavar, J., Maritan, A., Rinaldo, A., 1999. Size and form in efficient transportation networks. Nature 399, 130.

[Barabási and Albert (1999)] Barabási, A.-L., Albert, R., 1999. Emergenge of scaling in random networs. Science 286, 509–512.

[Barrat et al. (2004)] Barrat, A., Barthelemy, M., Vespignani, A., 2004. Weighted evolving networks: coupling topology and weights dynamics. Phys. Rev. Lett. 92, 228701.

[Caldarelli et al (2003)] Caldarelli, G., Coccetti, F., de los Rios, P., 2003. Preferential Exchange: Strengthening Connections in Complex Networks. Unpublished arXiv:cond-mat/0312236.

[De los Rios (2001)] De los Rios, P., 2001. Power law size distribution of supercritical random trees. Europhys. Lett. 56 (6), 898–903.

[Dorogovtsev and Mendes (2002)] Dorogovtsev, S., Mendes, J. F. F., 2002. Evolution of networks. Advances in Physics 51, 1079.

[Ebel et al. (2002)] Ebel, H., Mielsch, L.-I., Bornholdt, S., 2002. Scale-free topology of e-mail networks. Phys. Rev. E 66, 035103.

[Eckmann and Moses (2002)] Eckmann, J.-P., Moses, E., 2002. Curvature of co-links uncovers hidden thematic layers in the world wide web. Proc. Nat. Acad. Sci. USA 99 (9), 5825–5829.

[Eckmann et al (2003)] Eckmann, J.-P., Moses, E., Sergi, D., 2003. Dialog in e-Mail Traffic. Unpublished arXiv:cond-mat/0304433.

[Economist (2001)] Economist, 2001. The big picture (network collaborations). The Economist January 4th.

[Garicano (2000)] Garicano, L., October 2000. Hierarchies and the organization of knowledge in production. Journal of Political Economy, 874–904.

[Girvan and Newman (2002)] Girvan, M., Newman, M. E. J., 2002. Community structure in social and biological networks. Proc. Nat. Ac. Sci. USA 99, 7821–7826.

[Gleiser and Danon (2003)] Gleiser, P.M., Danon, L., 2003. Community structure in Jazz. Advances in Complex Systems 6, 565–573.

[Guardiola et al. (2002)] Guardiola, X., Guimera, R., Arenas, A., Diaz-Guilera, A., Amaral, L. A. N., 2002. Micro- and macro-structure of trust networks. Unpublished arXiv:cond-mat/0206240.

[Guimera et al. (2002)] Guimera, R., Diaz-Guilera, A., Vega-Redondo, F., Cabrales, A., Arenas, A., 2002. Optimal network topologies for local search with congestion. Phys. Rev. Lett. 89, 248701.

[Guimera et al. (2003)] Guimera, R., Danon, L., Díaz-Guilera, A., Giralt, F., Arenas, A., 2003. Self-similar community structure in a network of human interactions. Phys. Rev. E 68, 065103(R)..

[Halsey (1997)] Halsey, T. C., July 1997. The branching structure of diffusion-limited aggregates. Europhysics Letters 39, 43–48.

[Horton (1945)] Horton, R. E., 1945. Erosional development of streams and their drainage basins: hydrophysical approach to quantitative morphology. Bulletin of the Geological Society of America 56, 275–370.

[Jain and Dubes (1988)] Jain, A. K., Dubes, R. C., 1988. Algorithms for clustering data. Englewood Cliffs: Prentice Hall.

[Krackhardt and Hanson (1993)] Krackhardt, D., Hanson, J. R., 1993. Informal networks: the company behind the chart. Harvard Bussiness Review 71, 104–113.

[Kramer and Marder (1992)] Kramer, S., Marder, M., 1992. Evolution of river networks. Phys. Rev. Lett. 68, 205–208.

[Maritan et al. (1996)] Maritan, A., Rinaldo, A., Rigon, R., Giacometti, A., Rodriguez-Iturbe, I., 1996. Scaling laws for river networks. Phys. Rev. E 53, 1510–1515.

[Mayo (1949)] Mayo, E., 1949. The social problems of an industrial civilization. Routhledge.

[Morgan (1997)] Morgan, G., 1997. Images of organization, 2nd Edition. London: SAGE Publications.

[Newman (2003)] Newman, M.E. J., 2004. Fast algorithm for detecting community structure in networks. Phys. Rev. E 69, 026113.

[Newman (2002)] Newman, M.E. J., 2002. Assortative mixing in networks. Phys. Rev. Lett. 89, 208701.

[Newman (2001)] Newman, M.E. J., 2001. Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. Phys. Rev. E 64, 016132.

[Pastor-Satorras and Vespignani (2001)] Pastor-Satorras, R., Vespignani, A., 2001. Epidemic spreading in scale-free networks. Phys. Rev. Lett. 86, 3200–3203.

[Radner (1993)] Radner, R., 1993. The organization of decentralized information processing. Econometrica 61, 1109–1146.

[Rinaldo et al. (1993)] Rinaldo, A., Rodriguez-Iturbe, I., Rigon, R., Ijjasz-Vasquez, E., Bras, R. L., 1993. Self-organized fractal river networks. Phys. Rev. Lett. 70, 822–825.

[Rodriguez-Iturbe and Rinaldo (1996)] Rodriguez-Iturbe, I., Rinaldo, A., 1996. Fractal river basins: chance and self-organization. Cambridge: Cambridge University Press.

[Smith (2002)] Smith, R., 2002. Instant messaging as a scale-free network. Unpublished arXiv:cond-mat/0206378.

[Strahler (1952)] Strahler, A. N., 1952. Dynamic basis of geomorphology. Bulletin of the Geological Society of America 63, 923–938.

[Tyler et al (2003)] Tyler, J.R., Wilkinson, D.R., Huberman,B.A., 2003. Email as Spectroscopy: Automated Discovery of Community Structure within Organizations. Unpublished arXiv:cond-mat/0303264.

[Wasserman and Faust (1994)] Wasserman, S., Faust, K., 1994. Social Network Analysis. Cambridge: Cambridge University Press.

[Watts and Strogatz (1998)] Watts, D. J., Strogatz, S. H., 1998. Collective dynamics of 'small-world' networks. Nature 393, 440.

[Wellman (2001)] Wellman, B., 2001. Computer networks as social networks. Science 293, 2031–2034.

[Wu et al (2003)] Wu, F., Huberman, B.A., Adamic, L.A., Tyler, J., 2003. Information Flow in Social Groups. Unpublished arXiv:cond-mat/0305305.