Scaling in the Structure of Directory Trees in a Computer Cluster

Konstantin Klemm,[1] Víctor M. Eguíluz,[2] and Maxi San Miguel[2]

[1]*Department of Bioinformatics, University Leipzig, Härtelstrasse 16-18, 04103 Leipzig, Germany*
[2]*Instituto Mediterráneo de Estudios Avanzados IMEDEA (CSIC-UIB), E07122 Palma de Mallorca, Spain*

(Received 6 December 2004; published 14 September 2005)

We describe the topological structure and the underlying organization principles of the directories created by users of a computer cluster. Users create trees with a scale-free degree distribution whose properties are reproduced by a growth and preferential attachment mechanism with a single parameter. The degree distribution has a nonuniversal exponent associated with different values of the parameter. However, the distribution of branch sizes has a universal exponent analytically obtained from the model.

DOI: 10.1103/PhysRevLett.95.128701 PACS numbers: 89.20.Ff, 05.65.+b, 89.75.Da, 89.75.Hc

The processes of storing and retrieving information are rapidly gaining importance in science as well as in society as a whole [1–4]. A considerable effort is being undertaken, first to characterize and describe how publicly available information, for example, in the World Wide Web (WWW), is organized, and second to design efficient methods to access this information. It seems clear that to design methods for accessing information we first need to know how information is stored or organized as it is being produced.

Within this general framework a crucial step in building general knowledge on these processes is the understanding of how each of us organizes knowledge and information produced by ourselves. To be specific, we pose the question of general organizational principles in the managing of our own electronic files. To answer this question, we analyze the structure and organization of the files stored in a computer cluster by the users of the computer facilities at a research institute. Within the general study of complex networks, we are here looking at *trees* and we report a first observation of the scale-free property in trees. It is important to point out that we are not studying a single large tree, but rather we are considering a forest of many trees, each of them being the result of an individual construction. We are then able to consider samples of organizational schemes of many different sizes, since each user has created a structure with a different number of directories. This allows the study of different samples of the same reality. We also note that contrary to other networks, such as the WWW or food webs, the structures considered here are not the outcome of a collective action but the creation of a single individual. Our research gives information about the management of information at the individual level.

Two *a priori* possible answers to the question posed are that we follow a *random* process of file storing or that, on the contrary, we implement a carefully *planned* structure as we do when organizing the sections and chapters of a Ph.D. thesis or a scientific paper. What we find is the signature of a complex system halfway between these two possibilities, but still with well defined patterns of organization. In this Letter, we report an extensive characterization of individual user computer directory trees, calculating a number of quantitative measures. These include degree distributions, average distance between directories [5–8], distribution of branch sizes in the tree [9], and allometric scaling exponents [10,11]. Our data turn out to be well described by a directory attachment model for constructing the tree. The model depends on a single parameter q that interpolates between random placement of new directories and the agglomeration into a star structure. The trees of the different users are described by different values of the parameter q: diversity in individual behavior here boils down to a different value of a parameter.

Data analysis.—The data material under consideration is taken from the computer facilities of the Cross-disciplinary Physics Department of IMEDEA (Mediterranean Institute for Advanced Studies). The personal accounts of the 63 users running Linux and UNIX have been considered. The users include academic staff, postdoctoral researchers, graduate students, and longtime visitors. Each user is able to choose freely his or her own organizational scheme without specific software. The nodes in the directory tree of a given user are all directories (file folders) stored in the user's computer account. There is a direct link between nodes i and j if directory i is a subdirectory of directory j or vice versa. We consider the trees as rooted with the home directory as the root. Some of the users establish additional connections (so-called symbolic links) between directories or files. We assume here that the context indicated by these connections is much weaker than for genuine links between directory and parent. Therefore we neglect all symbolic links. In the pure tree structures, we analyze the distributions of degree and of branch sizes as well as the allometric scaling.

A local measure of the importance of a given node i is the nodal degree k_i counting the number of nodes directly connected to i. In a tree of N nodes the average degree is always $\langle k \rangle = 2 - 2/N$. The distribution of the degree, however, varies strongly across different types of structures. The distribution is narrow in simple chains and

binary trees, while it is broadest for a star (having $N-1$ nodes with degree $k=1$ and one center node with degree $k=N-1$). The degree distributions of the observed directory trees [Fig. 1(a)] lie in between these two extremes. The probability of finding a node with degree k decays as a power law $k^{-\gamma}$ with a cutoff at the maximum degree $k_{\max}$ due to finite size. There is no indication of an upper bound on the degree that would limit the scaling at large k. Given trees generated by different users, the observed values of γ do not coincide in general. The degree exponent is not universal. Often the degree distribution is also quantified by the branching ratio, given by the average of the degree k taken over all nodes with degree $k \geq 2$, i.e., neglecting the leaves. In the present case of scale-free distributions, however, the exponent γ is used as the leading observable.

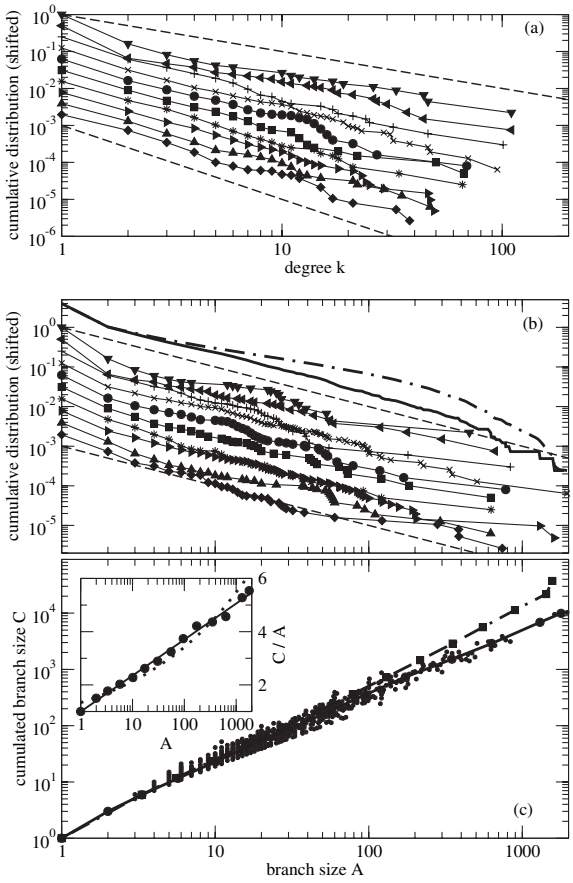

FIG. 1. Scaling in the distributions of branching ratio (degree) and sizes of the branches (subtrees). (a) Cumulative degree distributions for the ten largest trees. The dashed lines have slopes -1 and -2 indicating degree exponents $2<\gamma<3$. In the whole data set, however, exponents $\gamma>3$ have been observed as well. (b) Cumulative distributions of branch size plotted as in (a). The dashed lines have slope -1 corresponding to branch size exponent $\tau=2$. The overall cumulative distribution of the sizes of the 16 452 branches in all 63 directory trees (thick solid curve) and the surrogate data from randomized trees (dot-dashed curve) are shown as well. (c) Allometric scaling: Each data point (small circle) shows cumulative branch size C (sum of sizes of all subbranches) against the size A of the branch itself. Logarithmic binning is applied to the original data (large circles) and the surrogate data from randomized trees (squares). The inset shows the binned original data rescaled with A (circles) and best fits for logarithm (solid line) and power law (dotted curve). The surrogate data in (b) and (c) are taken from 6300 trees, 100 trees obtained from each original tree by independent random rewiring. Rewiring is performed by iteratively swapping two randomly chosen node disjoint subtrees that do not contain the root. This standard network randomization procedure [23], here applied to rooted trees, conserves the degree distribution.

An alternative characterization of the trees is obtained by iterative decomposition into subtrees rather than single nodes. Here we consider the *branch structure* of the trees. For each node i, a branch S_i is the subtree rooted at the node i and all nodes below i. In the directory trees, a branch S_i is the tree formed by a directory i, all its subdirectories, the subdirectories of these, and so forth. A branch S_i is again a rooted tree with node i as the root. Calculating the sizes $A_i=|S_i|$ of all branches for each tree, we find the statistics in Fig. 1(b). The distribution of branch sizes decays as a power law $A^{-\tau}$. The exponent $\tau=2$ appears to be universal. The scaling of branch size A is a property independent of the scaling of the degree k. When the trees are randomized under conserving degrees of all nodes, the functional form of the branch size distribution changes and obtains a scaling region with a larger exponent $\tau>2$.

In order to capture also the correlations between branch sizes, we perform *allometric scaling* analysis [11]. For each branch S_i we calculate the quantity $C_i=\sum_{j\in S_i}A_j$; i.e., we sum up all the sizes of all branches contained in S_i, including S_i itself. Figure 1(c) shows the data point (A_i, C_i) for each branch i in the 63 trees. We find that the growth of C with A is superlinear. The observations made for the degree and branch size distributions and the allometric scaling in the directory trees also hold for the file trees. The latter are constructed by including the files stored in the directories as additional leaves attached to them.

Modeling.—Let us now consider a possible mechanism for the emergence of the common properties of directory trees. Networks with a scale-free degree distribution can be generated by growth and preferential attachment [12]. Here we implement such a growth process for trees. In each construction step a new node joins the tree by establishing a link to one of the N existing nodes. A node with degree k is chosen as the parent of the new node with probability

$$\Pi(k)=q\frac{k-1}{N}+(1-q)\frac{1}{N} \tag{1}$$

from the set of N nodes in the tree. The growth is controlled by a single parameter $q\in]0,1]$. Using the formal equivalence with the network model by Dorogovtsev *et al.* [13], we find that the degree distribution is asymptotically scale-free with exponent $\gamma=2+a=1+1/q$ [rewrite Eq. (1) as $\Pi(k)\propto k^{\text{in}}+a$ with the number of links $k^{\text{in}}=k-1$

received after creation of the node and the "initial attractiveness" $a = 1/q - 1$ [13]].

In the context of directory trees, Eq. (1) has an intuitive interpretation. The first term describes creation of directories with functions similar to existing ones. Here the probability of being the parent of a new directory is proportional to the number $k - 1$ of existing subdirectories. The second term is a fully random placement where all directories have equal probability of becoming the parent of new directories. The two processes occur at rates q and $1 - q$, respectively. The validity of this dynamical picture may be assessed by observing the dynamics of the directory trees. The present work, however, is restricted to the comparison between trees generated by the model and our data set of snapshots of empirical trees.

The evolution of branch sizes is described by the probability

$$\tilde{\Pi}(A) = q\frac{A-1}{N} + (1-q)\frac{A}{N} = \frac{A-q}{N} \qquad (2)$$

that the next node is attached to one of the nodes of a given branch of size A, thereby incrementing A. From a continuous rate equation approach [12] we obtain $A_i(N) = (1 - q)N/i + q$ as the expected size of branch S_i in a tree of size N. The index i is the time step of creation of the branch as a single node with $A = 1$. The linear growth of A with N implies that the branch size distribution of the model decays asymptotically as $A^{-\tau}$ with universal (q-independent) exponent $\tau = 2$, in agreement with the data.

For an estimate of the allometric scaling, first note the general property $C_i = A_i + \sum_{j\in S_i} d_{ij}$, where the chemical distance d_{ij} is the number of nodes contained in the direct path between nodes i and j. Adding a new node j^* to branch S_i, the expected distance $\langle d_{ij^*}\rangle$ from node i is $C_i/A_i - 1$ for preferential and C_i/A_i for homogeneous attachment. Thus on average C grows as $dC/dA = 1 + C/A - q$, where the finite difference has been approximated by the derivative and the index i is suppressed. For the initial condition $C(1) = 1$ we obtain the solution $C(A) = A[(1 - q)\ln A + 1]$. The allometric scaling of the model trees is linear with logarithmic correction. In order to compare with the observed trees, we replot the binned data as $(A_i, C_i/A_i)$ in the inset of Fig. 1(c). The data are captured well by a logarithmic dependence (best fit $C/A = 0.59\ln A + 0.99$, correlation coefficient $r = 0.997$) in good agreement with the prediction of the model.

In order to provide a more stringent check of the validity of the model [Eq. (1)], we first project the trees into a space of four observables, namely, the second, third, and fourth moments of the degree distribution and the average chemical distance between nodes. For a given value x of an observable and given tree size N we estimate the most likely parameter value q_x by weighting all possible values $q \in [0, 1]$ with the probability that they produce x up to a small error. Figure 2 shows the results and gives details of the method in the caption. For almost all trees there is excellent agreement between the four parameter estimates based on different observables. Note that for different one-parameter models the agreement between parameter estimates cannot be obtained. For instance, random (maximum entropy) ensembles of trees with given branching ratio have very narrow degree distributions. Thus estimates of the branching ratio based on different moments of the broad empirical degree distribution do not agree. Random ensembles with given scale-free degree distributions still fail because their average distances are larger than in the empirical trees. These "failing" examples illustrate that the agreement between the model and the empirical trees due to parameter estimates is not trivial. We thus have strong evidence that the proposed growth mechanism [Eq. (1)] produces statistically the same structures as seen in the directory trees. The parameter q is the appro-

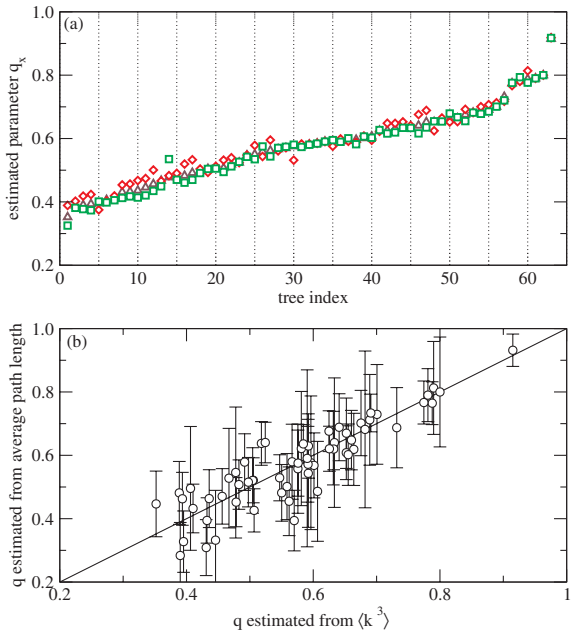

FIG. 2 (color online). Estimating the model parameter q from the empirical trees. (a) Independent estimates of q from the moments $\langle k^n\rangle$ of the degree distribution coincide for each tree. Estimates are plotted for $n = 2$ (diamonds), $n = 3$ (triangles), and $n = 4$ (squares). Tree index reflects the ordering of the trees with respect to the estimated q. (b) Comparing the q values estimated from the average path length and the third moment of the degree for each tree. For all estimates in (a) and (b) the following method is used. Given an empirical tree of size N with observable x_{emp}, 10^5 parameter values $q \in [0, 1]$ are drawn equally distributed. For each value drawn an artificial tree of size N is generated by the model. The tree is accepted if its value x_{model} of the considered observable does not differ by more than 10% from the empirical value x_{emp}. We take $\langle q\rangle$ as the average over parameter values of all accepted trees. The range of an error bar in (b) indicates the standard deviation of q across the accepted trees.

priate quantity to consider because differences between trees can be captured mainly by variation of q.

Discussion.—The structure of directory trees has been characterized from a statistical point of view. Our main result is the striking structural similarity between trees created by independent users in the absence of regulations. Users create trees with a broad, scale-free degree distribution with a nonuniversal exponent. The distribution of branch sizes, however, scales with a universal exponent $\tau \approx 2$. The allometric scaling is linear with a logarithmic correction. Branch structure and allometric scaling are significantly different in random surrogate trees with the same degree distribution. The statistical properties of the empirical trees are reproduced by a model that generates trees by adding nodes iteratively. The model has a single parameter q controlling the tendency to accumulate many subdirectories in the same parent directory. By varying q, the degree exponent can be tuned in the empirically observed range. The exponent $\tau = 2$ and the allometric scaling $C \sim A \ln A$ have been derived analytically and are independent of the parameter q. The validity of the model has been evidenced further by determining the most likely value of the parameter q. For a given tree, estimates based on different moments of the degree distribution as well as the diameter coincide, while estimates vary across trees. Consequently, directory trees can be distinguished by their specific value of the growth parameter q.

A generally interesting question is to decide about universal properties and universality classes of different natural and artificial or man-made complex networks. The branch distribution exponent $\tau \simeq 2$ that we find for our directory trees is in agreement with the one reported for the Internet [14,15] and for the communities of scientific collaborations [16,17]. However, a different class is formed by river networks [18–20], informal networks in organizations [9], and jazz musician networks [17], where the corresponding exponent gives a value $\tau \sim 1.45$ [15]. These examples seem to belong to the class of efficient networks obtained from an optimization principle in which transportation costs are minimized [10]. For the class of efficient networks, one can prove [10,21,22] that allometric scaling is given by a power law dependence $C \sim A^{\eta}$, with a universal exponent $\eta = (D + 1)/D$, where D is the embedding dimension. At difference with the prediction from efficiency, we find $C \sim A \ln A$ for the directory trees as reproduced by our growth model. This result is also compatible with effective (apparent) exponents observed in food webs [11].

We have shown that directory trees as individually man-made but not designed objects are an interesting direction of further research into hierarchical networks. Finding common statistical features of directory trees offers improved insight into how people naturally structure information.

We acknowledge financial support from MEC (Spain) through project CONOCE2 (FIS2004-00953), Deutsche Forschungsgemeinschaft (DFG), and Deutscher Akademischer Austausch Dienst (DAAD).

[1] R. M. Shiffrin and K. Börner, Proc. Natl. Acad. Sci. U.S.A. **101**, 5183 (2004).
[2] S. Lawrence and C. L. Giles, Nature (London) **400**, 107 (1999).
[3] R. F. I. Cancho and R. V. Solé, Proc. R. Soc. B **268**, 2261 (2001).
[4] M. Sigman and G. A. Cecchi, Proc. Natl. Acad. Sci. U.S.A. **99**, 1742 (2002).
[5] S. H. Strogatz, Nature (London) **410**, 268 (2001).
[6] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).
[7] S. N. Dorogovtsev and J. F. F. Mendes, Adv. Phys. **51**, 1079 (2002).
[8] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[9] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas, Phys. Rev. E **68**, 065103(R) (2003).
[10] J. R. Banavar, A. Maritan, and A. Rinaldo, Nature (London) **399**, 130 (1999).
[11] D. Garlaschelli, G. Caldarelli, and L. Pietronero, Nature (London) **423**, 165 (2003).
[12] A. L. Barabási and R. Albert, Science **286**, 509 (1999).
[13] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, Phys. Rev. Lett. **85**, 4633 (2000).
[14] G. Caldarelli, R. Marchetti, and L. Pietronero, Europhys. Lett. **52**, 386 (2000).
[15] P. De Los Rios, Europhys. Lett. **56**, 898 (2001).
[16] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, Proc. Natl. Acad. Sci. U.S.A. **101**, 2658 (2004).
[17] A. Arenas, L. Danon, A. Díaz-Guilera, P. Gleiser, and R. Guimera, Eur. Phys. J. B **38**, 373 (2004).
[18] I. Rodríguez-Iturbe and A. Rinaldo, *Fractal River Basins: Chance and Self-organization* (Cambridge University Press, New York, 1996).
[19] A. Rinaldo, I. Rodriguez-Iturbe, R. Rigon, E. Ijjasz-Vazquez, and R. L. Bras, Phys. Rev. Lett. **70**, 822 (1993).
[20] A. Maritan, A. Rinaldo, R. Rigon, A. Giacometti, and I. Rodriguez-Iturbe, Phys. Rev. E **53**, 1510 (1996).
[21] G. B. West, J. H. Brown, and B. J. Enquist, Science **276**, 122 (1997).
[22] J. R. Banavar, J. Damuth, A. Maritan, and A. Rinaldo, Proc. Natl. Acad. Sci. U.S.A. **99**, 10 506 (2002).
[23] S. Maslov and K. Sneppen, Science **296**, 910 (2002).