PHYSICAL REVIEW E **66**, 035103(R) (2002)

Scale-free topology of e-mail networks

Holger Ebel,* Lutz-Ingo Mielsch, and Stefan Bornholdt†
Institut für Theoretische Physik, Universität Kiel, Leibnizstraße 15, D-24098 Kiel, Germany

(Received 12 February 2002; published 30 September 2002)

We study the topology of e-mail networks with e-mail addresses as nodes and e-mails as links using data from server log files. The resulting network exhibits a scale-free link distribution and pronounced small-world behavior, as observed in other social networks. These observations imply that the spreading of e-mail viruses is greatly facilitated in real e-mail networks compared to random architectures.

DOI: 10.1103/PhysRevE.66.035103 PACS number(s): 89.75.Hc, 89.20.Hh, 05.65.+b

Complex networks such as the World Wide Web or social networks often do not have an engineered architecture but instead are self-organized by the actions of a large number of individuals. From these local interactions nontrivial global phenomena can emerge as, for example, small-world properties [1] or a scale-free distribution of the degree [2]. In small-world networks short paths between almost any two sites exist even though nodes are highly clustered like in a regular lattice. Scale-free networks are characterized by a power-law distribution of a node's degree, defined as the number of its next neighbors, meaning that structure and dynamics of the network are strongly affected by nodes with a great number of connections. These global properties have considerable implications on the behavior of the network under error or attack [3], when random or highly connected nodes are destroyed, as well as on the spreading of information or epidemics [4–6]. The highly connected "hub" nodes of a scale-free network and the short paths in a strongly clustered small world greatly facilitate the propagation of an infection over the whole network, which has to be taken into account for designing effective vaccination strategies [7–9]. Here we report that networks composed of persons connected by exchanged e-mails show both the characteristics of small-world networks and scale-free networks.

Most of the scaling exponents reported so far for the degree distributions of computer and social networks lie in the range of -2.0 to -3.4 [10]. One exception is the social network of co-authorships in high energy physics, for which Newman found an exceptionally small scaling exponent of -1.2 [11]. Similar to our work are studies of networks of phone calls made during one day. These phone-call networks show scale-free behavior of the degree distribution as well, with an exponent of -2.1 [12,13].

The scale-free e-mail network. The e-mail network studied here is constructed from log files of the e-mail server at Kiel University, recording the source and destination of every e-mail from or to a student account over a period of 112 days [25]. The nodes of the e-mail network correspond to e-mail addresses which are connected by a link if an e-mail has been exchanged between them. The resulting network consists of $N=59\,812$ nodes (including 5165 student accounts) with a mean degree of $\langle k\rangle=2.88$ and contains several separated clusters with less than 150 nodes and one giant component of 56 969 nodes (mean degree $\langle k_{\mathrm{giant}}\rangle=2.96$). The distribution of the degree k obeys a power law

$$n(k)\propto k^{-1.81}, \tag{1}$$

with exponential cutoff (Fig. 1).

Let us briefly discuss how our result on e-mail networks may be influenced by the measurement process. The sampling of the network has been restricted to one distinct e-mail server. Therefore, only the degrees of accounts at this server are known exactly. Here, these internal accounts correspond to e-mail addresses of local students, whereas the external nodes are given by all other e-mail addresses. We resolve the degree distribution of internal accounts only (Fig. 2), and find that it can be approximated by a power-law $n_{\mathrm{int}}(k)\propto k^{-1.32}$ as well (mean degree $\langle k_{\mathrm{int}}\rangle=14.86$). Since the degrees of external nodes typically are underestimated, this exponent is smaller than for the whole network. For the same reason, there are more nodes with small degree in the distribution of the whole network (Fig. 1) than in the distribution restricted to internal nodes (Fig. 2). Note that the cutoff of both distributions is about the same. Therefore, external sources addressing almost all internal nodes (e.g. advertisement or spam) do not bias the degree statistics.

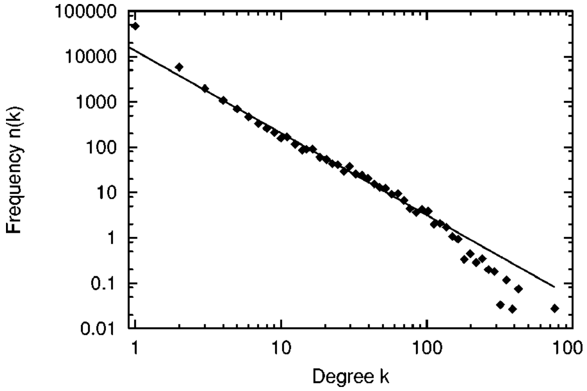

FIG. 1. Degree distribution of the e-mail network. The double-logarithmic plot of the number of e-mail addresses with which a node exchanged e-mails exhibits a power law with exponent -1.81 ± 0.10 over two decades. This distribution is used to calculate estimates for the clustering coefficient and the average shortest path length for the entire network (see text).

*Corresponding author.
Electronic address: ebel@theo-physik.uni-kiel.de

†Present address: Interdisziplinäres Zentrum für Bioinformatik, Universität Leipzig, Kreuzstraße 7b, D-04103 Leipzig, Germany. Electronic address: bornholdt@izbi.uni-leipzig.de

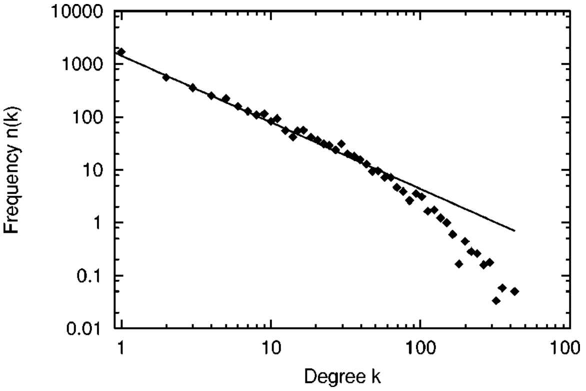

FIG. 2. Degree distribution of the student accounts in the e-mail network. The degree distribution of the subset of student e-mail addresses with completely known degree can be approximated by a power law as well with exponent -1.32 ± 0.18. This exponent is smaller than for the whole network since the degree of external nodes is underestimated by the measurement.

Small-world properties. In addition to its scale-free degree statistics, the e-mail network shows the properties of a "small world" [1], i.e. a high probability that two neighbors of one node are connected themselves (clustering) and a small average length ℓ of the shortest path between two nodes. The clustering is measured by the clustering coefficient C of a network which is defined in the following way: The clustering coefficient C_ν of a node ν is given by the ratio of existing links E_ν between its k_ν first neighbors to the potential number of such ties $\frac{1}{2}k_\nu(k_\nu-1)$. By averaging C_ν over all nodes one arrives at the clustering coefficient C of the network

$$C=\langle C_\nu\rangle_\nu=\left\langle\frac{2E_\nu}{k_\nu(k_\nu-1)}\right\rangle_\nu. \tag{2}$$

A similar definition for the clustering coefficient is provided by the fraction of fully connected "triples," with a triple being a connected subgraph which contains exactly three nodes [14,15]

$$C_\Delta=\frac{3\times(\text{number of fully connected triples})}{(\text{number of triples})}. \tag{3}$$

Note that even though definition (3) translates to reversing the order of averaging and division in Eq. (2) these two definitions are not equivalent. When calculating the clustering coefficient one has to ask how it is influenced by the finite size of the sample. Due to the measurement process, neighbors of external nodes are only partly known and connections between external nodes cannot be determined at all. Applying definitions (2) and (3) to the whole sample results in $C=3.44\times10^{-2}$ and $C_\Delta=3.15\times10^{-3}$. We compare these values to the clustering coefficient of random networks with constant probability p that two nodes are connected, leading to $C_{\text{rand}}=p$. Both values, C and C_Δ, are much larger than the clustering in such a random network of identical size $C_{\text{rand}}=4.82\times10^{-5}$. The fraction of the clustering contributed by internal nodes is much smaller than the portion of the external nodes because the measurement neglects links between external nodes which dominate most of internal nodes' neighborhoods. Since many of the external nodes have a large number of internal nodes as neighbors, which are only sparsely connected to other internal nodes, definition (3) results in a clustering coefficient C_Δ smaller than C. It is even smaller than $C'=1.87\times10^{-2}$, the clustering coefficient of a network of identical size with the same degree distribution but randomly assigned links [16]. Another way of determining clustering is to compute it for the completely known subgraph of internal nodes. There, the clustering is by more than one order of magnitude larger than expected for a random graph or a network with identical degree distribution but random connections ($C=8.09\times10^{-2}$, $C_\Delta=1.54\times10^{-1}$, $C_{\text{rand}}=2.30\times10^{-3}$, $C'=3.45\times10^{-3}$). In particular, the probability that two internal nodes who share an external neighbor are neighbors themselves is more than two orders of magnitude larger compared with random connections. Altogether, taking into account the limitations of the measurement process, it can be concluded that high clustering is a characteristic property of the e-mail network.

The mean shortest path length in the giant component was determined to $\ell=4.95\pm0.03$ with the Dijkstra algorithm [17]. It is larger than the mean shortest path length in a network with the same degree distribution $\ell'=3.43$ [16] since more links are consumed for forming local clusters [26]. It is still smaller than the path length of a random network $\ell_{\text{rand}}=10.10$ (where each pair of nodes is connected with a constant probability leading to the same mean degree [15,16]) because of the highly connected "hubs" present in a scale-free network.

The e-mail network as a directed network. To further investigate the emergence of the scale-free degree distribution, we study the e-mail network as a directed graph, where an e-mail corresponds to a directed link pointing from the sender to the receiver. Although the e-mail network has to be treated as an undirected graph in the context of virus spreading (see below), it seems reasonable that the sending and receiving of e-mails are governed by different processes. Again, the analysis is done for the distributions of all nodes and of internal nodes only, where for the latter, the in- and out-degree can be determined exactly. The distribution of the in-degree i, i.e. a node's number of different nodes it has received e-mails from, are very similar for all nodes and for internal nodes, respectively (Fig. 3). They can both be approximated by a power law $n(i)\propto i^{-1.49}$ over about two orders of magnitude. Deviations of the two distributions for in-degrees $i<6$ are due to the underestimation of the degree of external nodes. One explanation for an in-degree exponent of about -1.5 is the assumption of stochastic multiplicative growth as in the model of Huberman and Adamic [18,19]. They proposed that the number of links a node receives at a time step is a random fraction of the number of links it already has received. The treatment of the out-degree is more difficult. For the whole network, the distribution of out-degree j, i.e. a node's number of links to other nodes, shows pronounced scale-free behavior $n(j)\propto j^{-2.03}$ (Fig. 4). However, the corresponding distribution for internal nodes is broad but does not show scale-free behavior over a sufficient

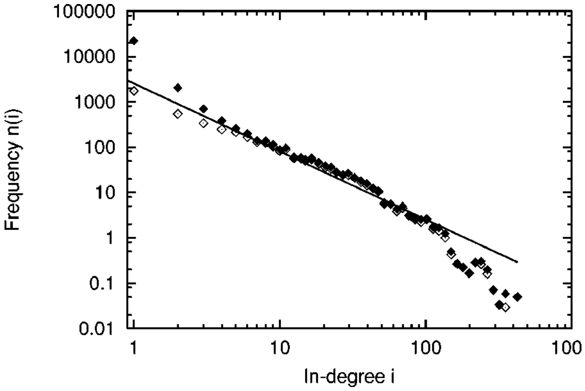

FIG. 3. In-degree distributions for the e-mail network. The double-logarithmic plot of the in-degree distributions for all nodes (filled diamonds, $\langle k_{\mathrm{in}}\rangle = 1.62$) and for student nodes only (open diamonds, $\langle k_{\mathrm{in}}\rangle_{\mathrm{int}} = 13.06$) shows a power-law distribution with an exponent of -1.49 ± 0.18. Note that again the in-degree of external nodes is underestimated by the measurement process.

range. This may be caused by the limited size of the sample but may also point to the systematic error caused by students possibly using different (external) accounts for sending e-mails. The out-degree scaling exponent of the whole network lies in a quite common range for communication and social networks, as, e.g., the movie actors' network or the phone call network [10], where the principle of *preferential attachment* can be used for modeling [2]. It applies to the assumption that the probability p_j that a link originates in the set of nodes with out-degree j is proportional to the number of links already starting in this set $f[j]$:

$$p_j \propto j f[j]. \tag{4}$$

This corresponds to Simon's general model for such copy and growth processes [20,21]. Let us briefly apply this model to the e-mail network. From our data we estimated the ratio of the growth rate of nodes to the growth rate of links to $\alpha = 0.597$ nodes per links which is sufficient to calculate the scaling exponent γ [21]:

$$\gamma = 1 + \frac{1}{1-\alpha}. \tag{5}$$

Thus, the preferential linking model leads to a steep exponent of -3.48 not in accordance with observation. On the other hand, a model based only on *transitive linking* [16], i.e. on the assumption that two nodes are more likely to be linked if they have a common neighbor, can in principle reproduce the small-world properties and a broad degree distribution but leads to a too high clustering and does not yield a power-law degree distribution for this particular network. From this we conjecture that including both preferential and transitive linking may consistently model the e-mail network.

Spreading of e-mail viruses. What are the implications of the above results for the spreading of e-mail viruses? The occurrence of e-mail viruses has become a well-known phenomenon in today's communication experience. An e-mail virus or e-mail worm is a program attached to an e-mail which, when opened by the recipient, causes the recipient's e-mail program to remail numerous infected e-mails to e-mail addresses found in the address book or in stored e-mails. In terms of a directed network, where links point from the sender of an e-mail to its receiver, an e-mail virus can follow a directed link, e.g. by taking e-mail addresses from the address book, as well as propagate in the reverse direction, for instance by using the senders' addresses of stored e-mails. Hence, for the propagation of e-mail viruses the network is undirected. This is different for chain e-mails, where each recipient is asked to forward the chain e-mail to other addresses. E-mail viruses can cause serious damage to computer networks by destroying data at infected computers or by overloading e-mail servers and other infrastructure. In May 2000, for instance, the e-mail worm "I love you" infected more than 500 000 individual systems worldwide [22] and obstructed 21% of the computer workplaces in Germany [23].

In scale-free networks, the threshold for the propagation rate above which an infection of the network spreads and becomes persistent is very much lower than in other disordered networks, or even vanishes [5]. This means that the self-organized structure of the e-mail network facilitates the spreading of computer viruses, as well as of any other information. In addition, the e-mail network is quite robust in case of "failures" of random nodes if, for instance, some participant does not answer e-mails for a while or uses antivirus software. However, it is sensitive to the loss of highly connected participants because of the power-law degree statistics [3]. Hence uniformly applied immunization of nodes is less likely to eradicate infections until almost all participants are immunized, whereas targeting prevention efforts at the highly connected sites significantly suppresses epidemic outbreaks and prevalence [7–9].

These observations suggest helpful and advantageous applications, but also point to the inherent dangers of e-mail

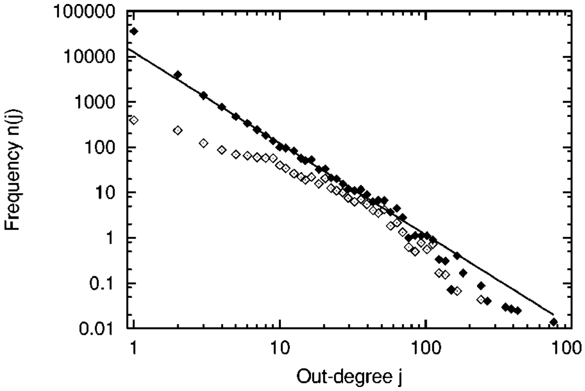

FIG. 4. Out-degree distributions for the e-mail network. For the out-degree, only the distribution of internal and external nodes (filled diamonds, $\langle k_{\mathrm{out}}\rangle = 1.62$) exhibits a pronounced power law over two decades with exponent -2.03 ± 0.12. The distribution of the out degree of internal nodes (open diamonds, $\langle k_{\mathrm{out}}\rangle_{\mathrm{int}} = 12.39$) is broad as well but cannot be identified with a scale-free regime which may be due to the limited size of the sample.

networks. The security of e-mail communication can be improved by identifying highly connected hub addresses and monitoring them for viruses more strictly, e.g., in corporate e-mail networks to prevent the damaging and costly spreading of e-mail viruses. In a different application, making use of the high clustering, commercial e-mail providers can identify communities of users more easily [24] and focus marketing more efficiently. In general, communication by e-mail can be interfered with as well as utilized more extensively due to the nontrivial topological features of the e-mail network that we found here. Exploring the web of e-mails does not only extend our knowledge of social and communication networks but it also shows how vulnerable and exploitable these systems can be.

In conclusion, we have shown that an e-mail network, where nodes are given by e-mail addresses and links by exchanged messages, exhibits both small-world properties and scale-free behavior. The e-mail network is studied in terms of undirected, as well as directed networks. Spreading of e-mail viruses is considered, based on the appropriate viewpoint of an undirected graph. The scale-free nature of the e-mail network strongly eases persistence and propagation of e-mail viruses but also points to effective countermeasures.

We thank A.-L. Barabási, J. Davidsen, S. N. Dorogovtsev, M. E. J. Newman, and A. Vespignani for useful discussions and comments. H.E. acknowledges support by the Studienstiftung des deutschen Volkes (German National Merit Foundation).

[1] D.J. Watts and S.H. Strogatz, Nature (London) **393**, 440 (1998).

[2] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).

[3] R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) **406**, 378 (2000).

[4] R. Pastor-Satorras and A. Vespignani, Phys. Rev. Lett. **86**, 3200 (2001).

[5] R. Pastor-Satorras and A. Vespignani, Phys. Rev. E **63**, 066117 (2001).

[6] S. Mossa, M. Barthélémy, H.E. Stanley, and L.A.N. Amaral, Phys. Rev. Lett. **88**, 138701 (2002).

[7] M.E.J. Newman, e-print cond-mat/0201433.

[8] R. Pastor-Satorras and A. Vespignani, Phys. Rev. E **65**, 036104 (2002).

[9] Z. Dezső and A.-L. Barabási, Phys. Rev. E **65**, 055103(R) (2002).

[10] R. Albert and A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002).

[11] M.E.J. Newman, Phys. Rev. E **64**, 016131 (2001).

[12] J. Abello, P.M. Pardalos, and M.G.C. Resende, *DIMACS Series in Discrete Mathematics and Theoretical Computer Science* (American Mathematical Society, Providence, 1999), Vol. 50, pp. 119–130.

[13] W. Aiello, F.R.K. Chung, and L. Lu, in *32nd ACM Symposium on Theory of Computing* (ACM, New York, 2000), pp. 171–180.

[14] A. Barrat and M. Weigt, Eur. Phys. J. B **13**, 547 (2000).

[15] M.E.J. Newman, S.H. Strogatz, and D.J. Watts, Phys. Rev. E **64**, 026118 (2001).

[16] J. Davidsen, H. Ebel, and S. Bornholdt, Phys. Rev. Lett. **88**, 128701 (2002).

[17] *Handbook of Discrete and Combinatorial Mathematics*, edited by K.H. Rosen (CRC Press, Boca Raton, 2000).

[18] B.A. Huberman and L.A. Adamic, Nature (London) **401**, 131 (1999).

[19] L.A. Adamic and B.A. Huberman, Science **287**, 2115a (2000).

[20] H.A. Simon, Biometrika **42**, 425 (1955).

[21] S. Bornholdt and H. Ebel, Phys. Rev. E **64**, 035104(R) (2001).

[22] CERT Coordination Center, Carnegie Mellon University, http://www.cert.org/advisories/CA-2000-04.html

[23] *Erster periodischer Sicherheitsbericht*, edited by Bundesministerium des Inneren (Federal Ministry of the Interior of Germany), http://www.bmi.bund.de/Downloads/27.pdf

[24] J. Kleinberg and S. Lawrence, Science **294**, 1849 (2001).

[25] Three e-mail addresses has been excluded as artifacts, since they reached a large quantity of the students only because all students used the same server (e.g. by service e-mails from the university computer center). Therefore their large degree was caused solely by sampling log files from only one e-mail server.

[26] To be specific, ℓ has to be calculated using the degree distribution of the giant component. Since this distribution does not differ significantly from the degree distribution of the whole network, employing the latter yields the same result within numerical accuracy used here.