

## Epidemic Spreading in Scale-Free Networks

Romualdo Pastor-Satorras<sup>1</sup> and Alessandro Vespignani<sup>2</sup>

<sup>1</sup>*Departament de Física i Enginyeria Nuclear, Universitat Politècnica de Catalunya, Campus Nord, Mòdul B4, 08034 Barcelona, Spain*

<sup>2</sup>*The Abdus Salam International Centre for Theoretical Physics (ICTP), P.O. Box 586, 34100 Trieste, Italy*  
(Received 20 October 2000)

The Internet has a very complex connectivity recently modeled by the class of scale-free networks. This feature, which appears to be very efficient for a communications network, favors at the same time the spreading of computer viruses. We analyze real data from computer virus infections and find the average lifetime and persistence of viral strains on the Internet. We define a dynamical model for the spreading of infections on scale-free networks, finding the absence of an epidemic threshold and its associated critical behavior. This new epidemiological framework rationalizes data of computer viruses and could help in the understanding of other spreading phenomena on communication and social networks.

DOI: 10.1103/PhysRevLett.86.3200

PACS numbers: 89.75.Hc, 05.50.+q, 05.70.Ln

Many social, biological, and communication systems can be properly described by complex networks whose nodes represent individuals or organizations, and links mimic the interactions among them [1,2]. Particularly interesting examples are the Internet [3,4] and the World Wide Web [5], which have been extensively studied because of their technological and economical relevance. These studies have revealed, among other facts, that the probability that a node of these networks has  $k$  connections follows a scale-free distribution  $P(k) \sim k^{-\gamma}$ , with an exponent  $\gamma$  that ranges between 2 and 3. The presence of nodes with a very large number of connections (local clustering) is indeed the key ingredient in the modeling of these networks with the recent introduction of scale-free (SF) graphs [6].

In view of the wide occurrence of complex networks in nature it is of great interest to inspect the effect of their features on epidemic and disease spreading [7], and more in general in the context of the nonequilibrium phase transitions typical of these phenomena [8]. The study of epidemics on these networks finds an immediate practical application in the understanding of computer virus spreading [9,10], and could also be relevant to the fields of epidemiology [11] and pollution control [12].

In this Letter, we analyze data from real computer virus epidemics, providing a statistical characterization that points out the importance of incorporating the peculiar topology of scale-free networks in the theoretical description of these infections. With this aim, we study by large scale simulations and analytical methods the susceptible-infected-susceptible (SIS) [11] model on SF graphs. We find the absence of an epidemic threshold and its associated critical behavior, which implies that SF networks are prone to the spreading and the persistence of infections at whatever spreading rate the epidemic agents possess. The absence of the epidemic threshold—a standard element in mathematical epidemiology [11]—radically changes many of the standard conclusions drawn in epidemic modeling. The present results are also relevant in the field

of absorbing-state phase transitions and catalytic reactions [8].

The analysis of computer viruses has been the subject of a continuous interest in the computer science community [10,13–15], mainly following approaches borrowed from biological epidemiology [11]. The standard model used in the study of computer virus infections is the SIS epidemiological model. Each node of the network represents an individual and each link is a connection along which the infection can spread to other systems. Individuals exist only in two discrete states, “healthy” or “infected.” At each time step, each susceptible (healthy) node is infected with rate  $\nu$  if it is connected to one or more infected nodes. At the same time, infected nodes are cured and become again susceptible with rate  $\delta$ , defining an effective spreading rate  $\lambda = \nu/\delta$  [16]. Without lack of generality, we can set  $\delta = 1$ . This model implicitly considers the presence of antivirus software, since all infected individuals eventually return to the susceptible state, and represents the case in which computer users do not become more alert with respect to viral infection once they have cleaned their computers which can again become infected [15]. The updating can be performed with both parallel and sequential dynamics [8]. In models with local connectivity (Euclidean lattices) and random graphs, the most significant result is the general prediction of a nonzero epidemic threshold  $\lambda_c$  [8,11]. If the value of  $\lambda$  is above the threshold,  $\lambda \geq \lambda_c$ , the infection spreads and becomes persistent. Below it,  $\lambda < \lambda_c$ , the infection dies out exponentially fast. The epidemic threshold is actually equivalent to a critical point in a nonequilibrium phase transition. In this case, the critical point separates an active phase with a stationary density of infected nodes from a phase with only healthy nodes and null activity. In particular, it is easy to recognize that the SIS model is a generalization of the contact process model that has been extensively studied in the context of absorbing-state phase transitions [8]. Statistical observations of virus incidents in the wild, on the other hand, indicate that all surviving viruses saturate to a very

low level of persistence, affecting just a tiny fraction of the total number of computers [10]. This fact is in striking contradiction with the theoretical predictions unless in the very unlikely chance that *all* computer viruses have an effective spreading rate tuned just infinitesimally above the threshold. This points out that the view obtained so far with the modeling of computer virus epidemics is very instructive but not completely adequate to represent the real phenomenon.

In order to gain further insight into the spreading properties of viruses in the wild, we have analyzed the prevalence data reported by the Virus Bulletin [17] from February 1996 to March 2000, covering a time window of 50 months. We have analyzed in particular the *surviving probability* of homogeneous groups of viruses, classified according to their infection mechanism [9]. We consider the total number of viruses of a given strain that are born and died within our observation window. Hence, we calculate the surviving probability  $P_s(t)$  of the strain as the fraction of viruses still alive at time  $t$  after their birth. Figure 1 shows that the surviving probability suffers a sharp drop in the first two months of a virus' life. This is a well-known feature [10,13] indicating that statistically only a small percentage of viruses gives rise to a significant outbreak in the computer community. Figure 1, on the other hand, shows for larger times a clean exponential tail,  $P_s(t) \sim \exp(-t/\tau)$ , where  $\tau$  represents the characteristic lifetime of the virus strain [18]. The numerical fit of the data yields  $\tau \approx 14$  months for boot and macroviruses and  $\tau \approx 6-9$  months for file viruses. The values of  $\tau$  are relatively independent of the observation window considered, i.e., the analysis of the viruses that are born and die in a time range of less than 50 months yields results compatible with the full data

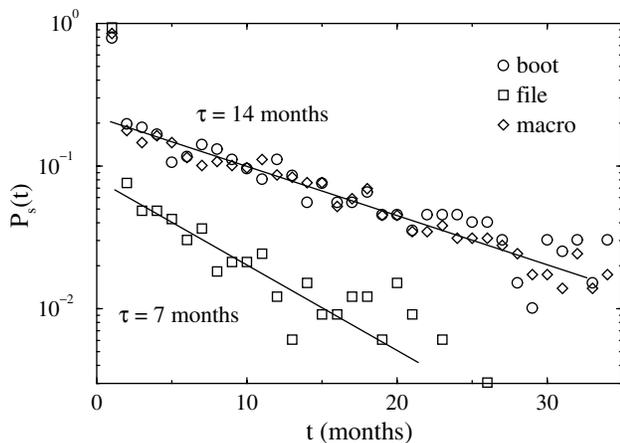


FIG. 1. Surviving probability for viruses in the wild. The 814 different viruses analyzed have been grouped in three main strains [9]: file viruses infect a computer when running an infected application; boot viruses also spread via infected applications, but copy themselves into the boot sector of the hard drive and are thus immune to a computer reboot; macroviruses infect data files and are thus platform independent. The presence of an exponential decay is evident in the plot, with characteristic time  $\tau$ .

set, with larger fluctuations, however, due to the smaller statistics. These characteristic times are impressively large if compared with the interval in which antivirus software is available on the market (usually within days or weeks after the first incident report) and corresponds to the occurrence of metastable endemic states. Such a long lifetime on the scale of the typical spread/recovery rates would suggest an effective spreading rate much larger than the epidemic threshold. On the other hand, this is again discordant with the always low prevalence levels of computer viruses.

The key point in understanding the puzzling properties exhibited by computer viruses resides in the capacity of many of them to propagate via data exchange with communication protocols (FTP, emails, etc.) [10]. Viruses will spread preferentially to computers which are highly connected to the outer world and thus are proportionally exchanging more data and information. It is thus rather intuitive to consider the Internet topology as the effective one on which the spreading occurs. The scale-free connectivity of the Internet implies that each node has a statistically significant probability of having a very large number of connections compared to the average connectivity  $\langle k \rangle$  of the network. That opposes conventional random networks (local or nonlocal) in which each node has approximately the same number of links  $k \approx \langle k \rangle$  [19]. It is then natural to foresee that scale-free properties should be included in a theory of epidemic spreading of computer viruses.

To address the effects of scale-free connectivity in epidemic spreading we study the SIS model on SF networks. As a prototypical example, we consider the graph generated by using the algorithm devised in Ref. [6]. We start from a small number  $m_0$  of disconnected nodes; every time step a new node is added, with  $m$  links that are connected to an old node  $i$  with  $k_i$  links according to the probability  $k_i / \sum_j k_j$ . After iterating this scheme a sufficient number of times, we obtain a network composed by  $N$  nodes with connectivity distribution  $P(k) \sim k^{-3}$  and average connectivity  $\langle k \rangle = 2m$ . In this work we take  $m = 3$ . We have performed numerical simulations on graphs with the number of nodes ranging from  $N = 10^3$  to  $N = 8.5 \times 10^6$  and studied the variation in time and the stationary properties of the density of infected nodes  $\rho$  in surviving infections; i.e., the virus prevalence. Initially we infect half of the nodes in the network, and iterate the rules of the SIS model with parallel updating. After an initial transient regime, the system stabilizes in a steady state with a constant average density of infected nodes. In this steady state, nodes are infected recurrently, without apparent periodicity. The prevalence is computed averaging over at least 100 different starting configurations, performed on at least 10 different realizations of the random networks.

The first arresting evidence from simulations is the *absence* of an epidemic threshold, i.e.,  $\lambda_c = 0$ . In Fig. 2 we show the virus prevalence in the steady state that decays with decreasing  $\lambda$  as  $\rho \sim \exp(-C/\lambda)$ , where  $C$  is a constant. This implies that for any finite value of  $\lambda$  the virus

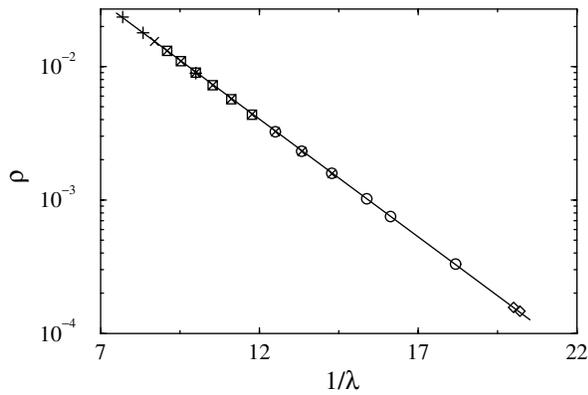


FIG. 2. Persistence  $\rho$  as a function of  $1/\lambda$  for different network sizes:  $N = 10^5$  (+),  $N = 5 \times 10^5$  ( $\square$ ),  $N = 10^6$  ( $\times$ ),  $N = 5 \times 10^6$  ( $\circ$ ), and  $N = 8.5 \times 10^6$  ( $\diamond$ ). The linear behavior on the semilogarithmic scale proves the stretched exponential behavior predicted for  $\rho$ . The full line is a fit to the form  $\rho \sim \exp(-C/\lambda)$ .

can pervade the system with a finite prevalence, in sufficiently large networks. In all networks with bounded connectivity the steady state prevalence is always null below the epidemic threshold; i.e., all infections die out. Further evidence to our results is given by the total absence of scaling of  $\rho$  with the number of nodes that is, on the contrary, typical of epidemic transitions in the proximity of a finite threshold [8]. This allows us to exclude the presence of any spurious results due to network finite size effects. The present result can be intuitively understood by noticing that for usual lattices, the higher the node's connectivity, the smaller the epidemic threshold. In a SF network the unbounded fluctuations in connectivity ( $\langle k^2 \rangle = \infty$ ) play the role of an infinite connectivity, annulling thus the threshold.

Finally, we analyze the spreading of infections starting from a localized virus source. We observe that the spreading growth in time has an algebraic form that is in agreement with real data that never found an exponential increase of a virus in the wild. Noteworthy, by applying the definition of surviving probability  $P_s(t)$  used to analyze real data, we recover in our model the same exponential behavior in time (see Fig. 3a). The characteristic lifetime depends on the spreading rate and the network sizes, allowing us to relate the average lifetime of a viral strain with an effective spreading rate and the Internet size [20]. At the same time, the divergence of lifetimes for larger networks points out that viruses live longer if the Internet expands.

We can also approach the system analytically by writing the single-site equation governing the time evolution of  $\rho(t)$ . In order to take into account connectivity fluctuations, we consider the relative density  $\rho_k(t)$  of infected nodes with given connectivity  $k$ ; i.e., the probability that a node with  $k$  links is infected. The dynamical mean-field (MF) reaction rate equations can be written as [8,21]

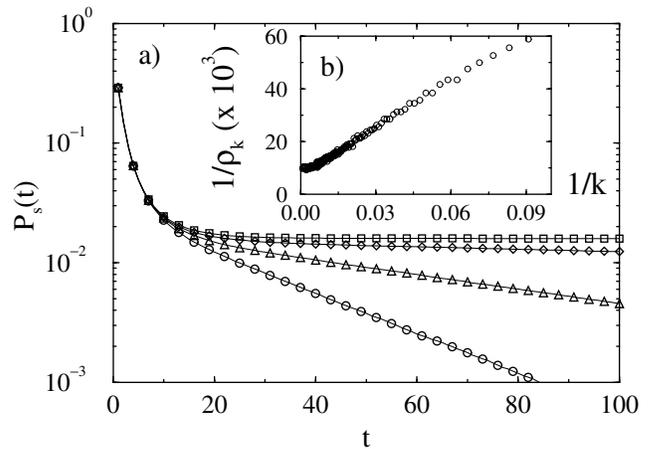


FIG. 3. (a) Surviving probability  $P_s(t)$  for a spreading rate  $\lambda = 0.065$  in scale-free networks of size  $N = 5 \times 10^5$  ( $\square$ ),  $N = 2.5 \times 10^4$  ( $\diamond$ ),  $N = 1.25 \times 10^4$  ( $\triangle$ ), and  $N = 6.25 \times 10^3$  ( $\circ$ ). The exponential behavior, following a sharp initial drop, is compatible with the data analysis of Fig. 1. (b) Relative density  $\rho_k$  versus  $k^{-1}$  in a SF network of size  $N = 5 \times 10^5$  and spreading rate  $\lambda = 0.1$ . The plot recovers the form predicted in Eq. (2).

$$\partial_t \rho_k(t) = -\rho_k(t) + \lambda k [1 - \rho_k(t)] \Theta(\lambda). \quad (1)$$

The creation term considers the probability that a node with  $k$  links is healthy  $[1 - \rho_k(t)]$  and gets the infection via a connected node. The probability of this event is proportional to the infection rate, the number of connections, and the probability  $\Theta(\lambda)$  that any given link points to an infected node. The MF character of this equation stems from the fact that we have neglected the density correlations among the different nodes. However, we have relaxed the homogeneity assumption on the node's connectivity usually implemented in regular networks. By imposing stationarity [ $\partial_t \rho_k(t) = 0$ ] we find the stationary densities

$$\rho_k = \frac{k \lambda \Theta(\lambda)}{1 + k \lambda \Theta(\lambda)}, \quad (2)$$

denoting that the higher the node connectivity, the higher the probability to be infected. This inhomogeneity must be taken into account in the self-consistent calculation of  $\Theta(\lambda)$ . Indeed, the probability that a link points to a node with  $s$  links is proportional to  $sP(s)$ . In other words, a randomly chosen link is more likely to be connected to a node with high connectivity, yielding

$$\Theta(\lambda) = \sum_k \frac{k P(k) \rho_k}{\sum_s s P(s)}. \quad (3)$$

Since  $\rho_k$  is on its turn function of  $\Theta(\lambda)$ , we obtain a consistency equation that allows us to find  $\Theta(\lambda)$  and  $\rho_k$ . Finally we can calculate the order parameter by evaluating the relation  $\rho = \sum_k P(k) \rho_k$  that expresses the average density of infected nodes in the system. In the SF model considered here, we have a connectivity distribution  $P(k) = 2m^2/k^{-3}$ , where  $k$  is approximated as a continuous variable [6]. In this case, integration of Eq. (3) allows

one to write  $\Theta(\lambda) = e^{-1/m\lambda}/\lambda m$ , at lowest order in  $\lambda$ . Averaging over  $\rho_k$ , this finally gives

$$\rho \approx 2e^{-1/m\lambda}. \quad (4)$$

This very intuitive calculation recovers the numerical findings and confirms the surprising absence of any epidemic threshold or critical point in the model; i.e.,  $\lambda_c = 0$ . Finally, as a further check of our analytical results, we have numerically computed in our model the relative densities  $\rho_k$ , recovering the predicted dependence upon  $k$  of Eq. (2) (see Fig. 3b). It is also worth remarking that the present framework can be generalized to networks with  $2 < \gamma \leq 3$ , recovering qualitatively the same results. Only for  $\gamma > 4$ , epidemics on SF networks have the same properties as on random networks. A detailed analysis of the various cases will be presented elsewhere [22].

The emerging picture for epidemic spreading in complex networks emphasizes the role of topology in epidemic modeling. In particular, the absence of epidemic threshold and critical behavior in a wide range of scale-free network provide an unexpected result that changes radically many standard conclusions on epidemic spreading. This indicates that infections can proliferate on these scale-free networks whatever spreading rates they may have. This very bad news is, however, balanced by the exponentially small prevalence for a wide range of spreading rates ( $\lambda \ll 1$ ). This point appears to be particularly relevant in the case of technological networks such as the Internet [4] that show scale-free connectivity with exponents  $\gamma \approx 2.5$ . For instance, the present picture qualitatively fits the observation from real data of computer virus spreading, and could solve the long-standing problem of the generalized low prevalence of computer viruses without assuming any global tuning of the spreading rates.

This work has been partially supported by the European Network Contract No. ERBFMRXCT980183. R.P.-S. also acknowledges support from the Grant No. CICYT PB97-0693. We thank S. Franz, M.-C. Miguel, R. V. Solé, M. Vergassola, S. Visintin, S. Zapperi, and R. Zecchina for helpful comments and discussions.

- 
- [1] See the special section on Complex systems [Science **284**, 79 (1999)]; S. Wasserman and K. Faust, *Social Network Analysis* (Cambridge University Press, Cambridge, 1994).  
 [2] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley, Proc. Natl. Acad. Sci. U.S.A. **97**, 11 149 (2000).

- [3] M. Faloutsos, P. Faloutsos, and C. Faloutsos, ACM SIGCOMM '99, Comput. Commun. Rev. **29**, 251 (1999).  
 [4] A. Medina, I. Matt, and J. Byers, Comput. Commun. Rev. **30**, 18 (2000); G. Caldarelli, R. Marchetti, and L. Pietronero, Europhys. Lett. **52**, 386 (2000).  
 [5] R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) **401**, 130 (1999).  
 [6] A.-L. Barabási and R. Albert, Science **286**, 509 (1999); A.-L. Barabási, R. Albert, and H. Jeong, Physica (Amsterdam) **272A**, 173 (1999).  
 [7] C. Moore and M. E. J. Newman, Phys. Rev. E **61**, 5678 (2000).  
 [8] J. Marro and R. Dickman, *Nonequilibrium Phase Transitions in Lattice Models* (Cambridge University Press, Cambridge, 1999).  
 [9] F. B. Cohen, *A Short Course on Computer Viruses* (John Wiley & Sons, New York, 1994).  
 [10] J. O. Kephart, G. B. Sorkin, D. M. Chess, and S. R. White, Sci. Am. **277**, No. 5, 56 (1997); S. R. White, in *Proceedings of the Virus Bulletin Conference, Munich, 1998*. Available on-line at <http://www.research.ibm.com/antivirus/SciPapers.htm>.  
 [11] N. T. J. Bailey, *The Mathematical Theory of Infectious Diseases* (Griffin, London, 1975), 2nd ed.; J. D. Murray, *Mathematical Biology* (Springer-Verlag, Berlin, 1993).  
 [12] M. K. Hill, *Understanding Environmental Pollution* (Cambridge University Press, Cambridge, 1997).  
 [13] J. O. Kephart, S. R. White, and D. M. Chess, IEEE Spectr. **30**, 20 (1993).  
 [14] W. H. Murray, Comput. Sec. **7**, 130 (1988).  
 [15] J. O. Kephart and S. R. White, in *Proceedings of the 1991 IEEE Computer Society Symposium on Research in Security and Privacy (SSP '91)* (IEEE, Washington, 1991), p. 343.  
 [16] It is also possible to define models in which the infection rate is proportional to the number of infected nearest neighbors [8]. In the small prevalence regime we are interested in, both prescriptions yield exactly the same behavior.  
 [17] Virus prevalence data publicly available at the web site <http://www.virusbtn.com/Prevalence/>.  
 [18] This is the usual way in which it is determined the survival probability in numerical simulations of spreading models; see Ref. [8].  
 [19] P. Erdős and P. Rényi, Publ. Math. Inst. Hung. Acad. Sci. **5**, 17 (1960); D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998); A. Barrat and M. Weigt, Eur. Phys. J. B **13**, 547 (2000).  
 [20] This characteristic scaling is often encountered at absorbing-state phase transitions in finite size systems [8]. In general,  $P_s(\infty)$  is finite only for infinite size networks.  
 [21] G. Szabó, Phys. Rev. E **62**, 7474 (2000).  
 [22] R. Pastor-Satorras and A. Vespignani (unpublished).