# Information flow dynamics and timing patterns in the arrival of email viruses

Larry S. Liebovitch[1,2] and Ira B. Schwartz[1]

[1] *Naval Research Laboratory, Code 6792, Plasma Physics Division, Washington, DC 20375, USA*

[2] *Center for Complex Systems and Brain Sciences, Center for Molecular Biology and Biotechnology, Department of Psychology, Florida Atlantic University, Boca Raton, Florida 33431, USA*

(Received 25 March 2003; published 3 July 2003)

Analysis of the timing of the arrival of email viruses at different computers provides a way of probing the structural and dynamical properties of the Internet. We found that the intervals $t$ between the arrival of four different strains of email viruses have a power law distribution proportional to $t^{-d}$, where $1.5 \leqslant d \leqslant 3.2$ and that there are positive correlations between these intervals. Salient features of the data were reproduced with a model having subnetwork units of different size where the structural components and the dynamical components all have power law scaling relationships with the size of the units. This is an assumption, that we hope will encourage empirical evaluation of these relationships.

DOI: 10.1103/PhysRevE.68.017101 PACS number(s): 89.75.Da, 89.75.Fb, 89.75.Hc

Many social, biological, engineering, and communication systems may be modeled as complex networks. Because of the widespread connectivity of the Internet, much attention has been given to the organization and transmission of information on large finite networks, particularly with respect to virus attacks [1,2]. One such example is the small-world network [3], where a regular network is combined with a few random interconnections. Similar models have generated statistical cluster analysis based on percolation [4], scaling laws [5], and control of information [6]. Recent work on Internet connectivity has shown that with the right scaling law, the Internet is robust when computer attacks remove certain nodes [7]. Moreover, percolation theories have been used to model the propagation of an epidemic probabilistically [8]. By examining properties of epidemic outbreaks on networks, theories of control have been modeled as well [9–11]. Although epidemic modeling on networks has generated probability models of control, most of the previous scaling law models do not consider the dynamics of the rate of information sent from different cluster sizes, which we argue is important in understanding rates of transmission laws [12]. In this paper, we show from data and a model that the arrival times of email viruses at different computers depend on both the structural and dynamical properties of the Internet.

Data for email virus receipt have been collected by a provider in the UK [13]. The provider is a monitoring node (MN) that monitors the emails passing from Internet service providers to their client computers. Their software detects emails infected with viruses and deletes the viruses. It maintains records of the arrival dates and times as well as assigning a unique integer to each IP address. The number of arrival times analyzed for four common viruses [14,15] AnnaKournikova, Magistr.b, Klez.e, and Sircam.a, respectively, are 20 883, 153 518, 413 182, and 781 626. These were recorded, respectively, over 57.249 days (starting 12 February 2001 13:21 UT), 288.875 days (starting 4 September 2001 12:49 UT), 154.629 days (starting 16 January 2002 18:47 UT), and 338.109 days (starting 17 July 2001 7:27 UT). For each virus, we determined the probability density function and the Hurst rescaled range analysis of the times between the reported arrivals of the virus.

The probability density function (PDF) $P(t)=N(t,t+dt)/(N_T dt)$, where $N(t,t+dt)$ is the number of times $t$ between the arrival of viruses within the interval $(t,t+dt]$ and $N_T$ is the total number of times. The usual method to evaluate the PDF is to form a histogram of $N(t,t+dt)$ with a fixed bin size $dt$. The problems with the method are the following: (1) If the bin size $dt$ is chosen to be small, then there are few times in the bins at large $t$; (2) if the bin size $dt$ is chosen to be large to capture more times in the bins at large $t$, then good resolution at small $t$ is lost. We overcome these limitations by using a multihistogram method that combines PDFs generated from histograms of bins of different size, and is more accurate in forming the PDFs from the known test data of several different functional forms including single exponential and power law distributions [16].

Figure 1 illustrates the PDF $P(t)$ of two of the four viruses, which are all approximately straight lines on a plot of ln($P$) vs ln($t$), where $t$ is the time interval (in days) between the arrival of the viruses. Thus, the PDF has the power law form that $P(t)$ is proportional to $t^{-d}$. The exponents $d$ determined from the slope of the best least squares fit of ln($P$) vs ln($t$) were 1.51 for AnnaKournikova, 3.19 for Magistr.b, 2.40 for Klez.e, and 2.69 for Sircam.a.

The Hurst rescaled range method analyzes the dependency of the range of the fluctuations as a function of the window size over which they are measured [17,18]. The range $R$ in each window is measured as the difference between the maxima and minima of the running sum of the data values minus the mean and is then divided by the standard deviation $S$ in that window. The size of the windows is called the lag, $\tau$. This method is good at determining if there are long range self-similar correlations present in the data from a plot of ln($R/S$) vs ln($\tau$).

As shown in Fig. 2, the Hurst rescaled range plots of ln($R/S$) vs ln($\tau$), the viruses can be fit with straight lines, with $H>0.5$, indicating that there are significant correlations in the times between the arrival of the viruses. The values of $H$ determined from the slope of the best least squares fit of ln($R/S$) vs ln($\tau$) were 0.80 for AnnaKournikova, 0.80 for Magistr.b, 0.82 for Klez.e, and 0.86 for Sircam.a. However, there are also significant deviations from the simplest

1063-651X/2003/68(1)/017101(4)/$20.00

©2003 The American Physical Society

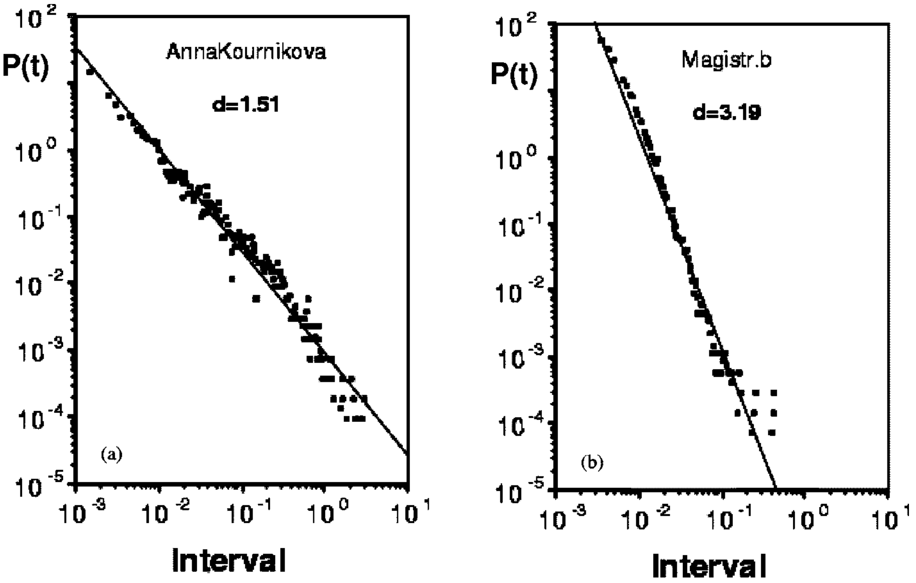

FIG. 1. The probability density function $P(t)$ of the intervals $t$ (in days) between the arrivals of the email viruses is a power law distribution, as illustrated here for two of the four email viruses.

straight line fit. Since $H>0.5$, there are persistent, positive correlations between the arrival times of the viruses but the deviations from linearity mean that these correlations are not so simple as exactly self-similar ones.

Values of the slope $H>0.5$ on the plots of ln($R/S$) vs ln($\tau$) can also result from short term as well as long term correlations [19]. However, since $H>0.5$, whether caused by short term or long term processes, these are persistent, positive correlations. It is surprising, and perhaps counterintuitive, that viruses transmitted by different independent sources arrive at their receiving computers strongly correlated in time.

It is known that both the structural and dynamical properties of the Internet display power law scaling relationships. For example, in structure, the number of nodes that require $k$ links to reach another node is proportional to $k^{-v}$, where $v$ characterizes the scaling pattern of spatial connectivity [20,21]. In dynamics, the distribution of the transfer times for files, the number of files with transfer time $t$, is proportional to $t^{-w}$, where $w$ characterizes the scaling pattern of temporal behavior [23,22]. To combine structural and dynamical properties into one model, we assume that subnetworks of the Internet are grouped into the units of size $k$. The viruses sent from any of the $k$ computers within each unit will first pass through its local area network and then through the gateway which connects that unit to the rest of the Internet.

The MN receives emails from these units on their way toward the receiving computers of the Internet service provider. As shown in Fig. 3, we picture emails transmitted from different numbers of units of different size $k$. We assume that there are $n(k)$ units of size $k$. We further assume that only one unit transmits at a time [25]. During each transmission, it sends $e(k)$ viruses each separated by a constant time $t$. We assume that the structural properties, $n(k)$, and the dynami-

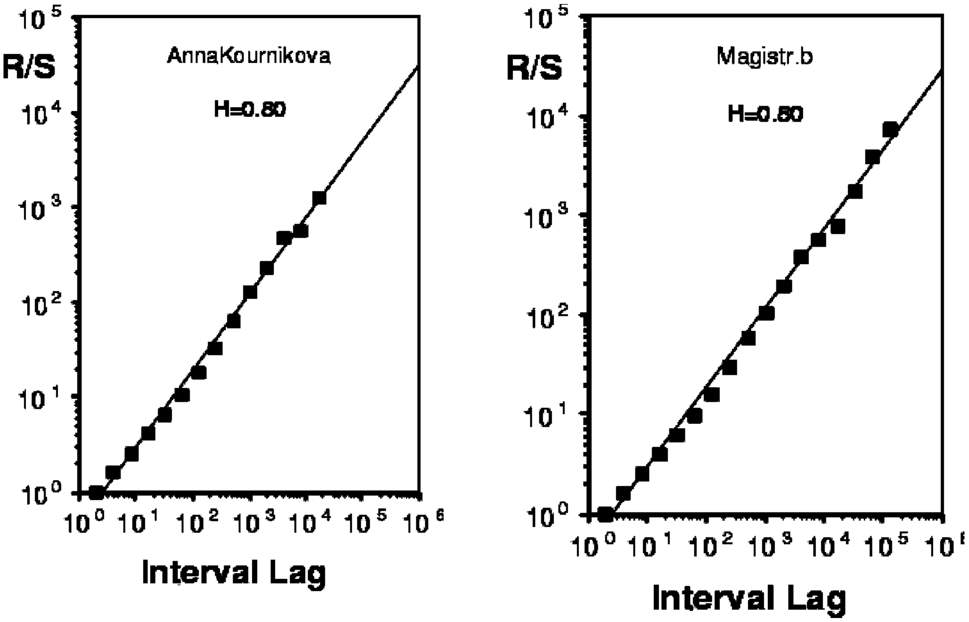

FIG. 2. The Hurst rescaled range analysis of the intervals $t$ between the arrivals of each of two of four email viruses. The range $R$, normalized by the standard deviation $S$, within a window is plotted vs the size of that window ($\tau$). The slope of ln($R/S$) vs ln($\tau$), the Hurst exponent $H$, is $0.80<H<0.86$.

cal properties, $e(k)$ and $t(k)$, all have power law scaling relationships with the number of computers $k$ in each unit. Thus, $n(k)$ is proportional to $k^{-a}$, $e(k)$ is proportional to $k^{b}$, and $t(k)$ is proportional to $k^{-c}$.

The relative number of viruses received from units of size $k$ is $n(k)e(k) = k^{b-a}$. We now use the relationship between the time between the arrival of the viruses, namely $t(k)$ proportional to $k^{-c}$, to determine the relative number of viruses received from units of size $k$ in terms of the time $t$. Namely, since $k$ is proportional to $t^{-1/c}$, then $n(k)$ is proportional to $t^{a/c}$ and $e(k)$ is proportional to $t^{-b/c}$. Thus, the relative number of viruses received from units of size $k$, $n(k)e(k) = t^{a/c-b/c}$.

The number of times that the time, $t$, between the arrival of the viruses is in the range $(t, t+dt]$ depends on the relative number of viruses $n(k)e(k)$. If $n(k)e(k)$ is proportional to $t^{-r}$, then $\ln[n(k)e(k)]$ is proportional to $-r\ln(t)$, and its associated distribution with respect to $\ln(t)$, $P_x(x) = \exp(-rx)$ where $x = \ln(t)$. The relationship between $P_x(x)$ and $P(t)$ is given by $P_x(x)dx = P(t)dt$. Thus, $P(t) = P_x(x)|dx/dt|$. Since $dx/dt = 1/t$ and $P_x(x = \ln(t)) = t^{-r}$, $P(t)$ is proportional to $t^{-(1+r)}$. Since, $n(k)e(k) = t^{-r}$, where $r = -a/c + b/c$, this means that $P(t)$ is proportional to $t^{-(1-a/c+b/c)}$. The PDF of all four viruses has the form that the $P(t)$ is proportional to $t^{-d}$. Hence $d = 1 - a/c + b/c$.

This relationship for the PDF is confirmed in the results of numerical simulations computed in MATLAB shown in Fig. 4. We simulated a network with units ranging in size from 1 to 1000 elements over 200 equally spaced logarithmic steps. First, a unit was chosen at random with probability propor-

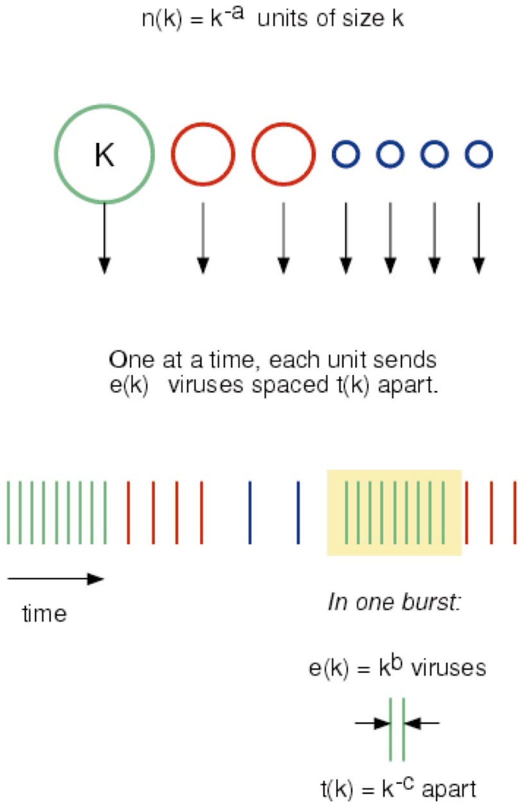
n(k) = k-a units of size k
K
One at a time, each unit sends
e(k) viruses spaced t(k) apart.
time
In one burst:
e(k) = kb viruses
t(k) = k-c apart

FIG. 3. (Color) We model the email viruses as transmitted from the units of $k$ computers. There are $n(k) = k^{-a}$ units of size $k$. At each event, one unit sends $e(k) = k^{b}$ viruses separated by a time $t(k) = k^{-c}$.

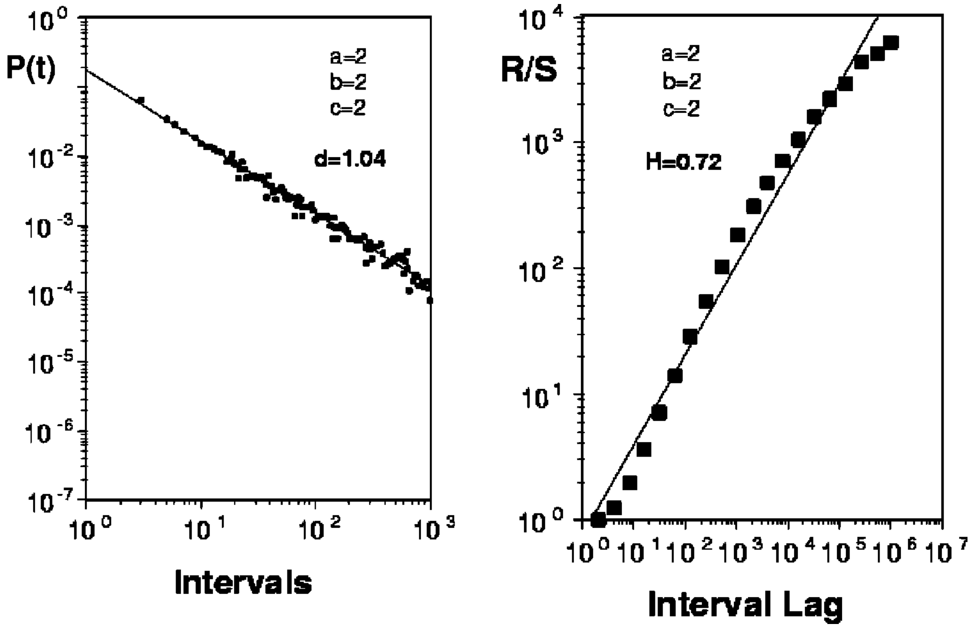

FIG. 4. (a) The probability density function $P(t)$ of the intervals $t$ (in days) between the arrivals of viruses computed from a numerical simulation of the model where the parameters $a=2$, $b=2$, and $c=2$. The PDF has a power law distribution $t^{-d}$, like the data in Fig. 1. The value of $d$ computed from this numerical simulation was 1.04, compared to 1 determined analytically from this model. The value of $d$ depends on the parameters $a$, $b$, and $c$. (b) The Hurst rescaled range analysis of the intervals $t$ between the arrivals of viruses computed from a numerical simulation of the model where the parameters $a=2$, $b=2$, and $c=2$. The slope of $\ln(R/S)$ vs $\ln(\tau)$, the Hurst exponent, is $H=0.72$.

tional to $n(k)$. It then generated $e(k)$ viruses at $t(k)$ intervals between the viruses. A total of at least 1 048 576 ($1M$) virus arrival times were computed. The multihistogram method was then used to compute $P(t)$ and the slope $d$ determined from best least squares fit of ln($P$) vs ln($t$). Numerical simulations were computed for the cases where $a=b=c=2$; and $a=0$ and $b=c=2$. The numerical results $d=1.04$ and $d=2.04$ compare favorably with the analytic results that $d=1$ and $d=2$, respectively.

Some interesting conclusions can be drawn from the relationship $d=1-a/c+b/c$, which relates the structural properties ($a$) and dynamical properties ($b$, $c$, and $d$) of the Internet. First, the relative number of viruses received from all the units of size $k$ is proportional to $k^{b-a}$. The exponent $(b-a)=c(d-1)>0$ when $d>1$. That is, when $d>1$ relatively more viruses are received from the larger units than from the smaller units. When $d<1$, the situation is reversed and and relatively more viruses are received from the smaller units. Since $d>1$ for all four viruses studied here, more viruses are received from the larger units. If a virus was found whose $P(t)$ was proportional to $t^{-d}$, where $d<1$, this would indicate that more viruses were being received from smaller units instead. Thus, the critical value of $d=1$ may be useful in diagnosing different transmission scenarios.

Second, the time between the receipt of viruses from a unit of size $k$ is proportional to $t^{-c}$ where $c=(b-a)/(d-1)$. Since $d>1$ for all the four viruses studied here, this means that when $b>a$, the rate of receiving viruses is larger from the larger units, and when $b<a$, the rate of receiving viruses is larger from the smaller units.

We computed the Hurst rescaled range numerically from this model for the cases where $a=b=c=2$; and $a=0$ and $b=c=2$. The ln($R/S$) vs ln($\tau$) plot computed from the numerical simulations have overall slopes of $H=0.74$ and $H=0.72$. Since $H>0.5$, there are persistent, positive correlations between the arrival times of the viruses. It is interesting that these time correlations, which are seen in the experimental data, are also generated by this simple model. The model also deviates from linearity in the ln($R/S$) vs ln($\tau$) plot. The asymptotically horizontal line at small lags is due to the fact that the times between the arrivals of the consecutive viruses from the same unit are equally spaced and so as the range $R$ and the standard deviation $S$ both approach zero, the limit of $R/S$ approaches 1. It is also interesting that some of these same trends, although less pronounced, are seen in the experimental data, namely, the ln($R/S$) plot first falls below the linear trend and then meets it or rises above it.

The analysis of the statistical properties of the arrival times of email viruses provides a way to probe the interactions between the structural and dynamical properties of the Internet. These properties can be simulated with a simple model with units of different sizes where the number of units and the numbers and rates at which they send viruses scale as a power law of the size of the units. This model well reproduces the distribution of the times $t$ between the arrival of viruses, $P(t)$, and approximately reproduces some of the correlation properties present in the experimental data. It also provides some insight into the fact that the PDF is proportional to $t^{-d}$. For the virus data analyzed here, where $1.5\leqslant d\leqslant 3.2$, the model implies that more viruses are received from larger units. However, if the data from a virus had $d<1$, then that would mean that more viruses were received from the smaller units.

This work was supported by the Office of Naval Research. We thank Julie Tarr for tracking down the data.

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

[2] Y. Moreno *et al.*, Eur. Phys. J. B **26**, 521 (2002).

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

[4] M. Newman and D. Watts, Phys. Rev. E **60**, 7332 (1999).

[5] A. Barabasi *et al.*, Physica A **272**, 173 (1999).

[6] S. Pandit and R. Amritkar, Phys. Rev. E **60**, R1119 (1999).

[7] R. Cohen *et al.*, Phys. Rev. Lett. **85**, 4626 (2000).

[8] C. Moore and M. Newman, Phys. Rev. E **61**, 5678 (2000).

[9] L. Billings, W.M. Spears, and I.B. Schwartz, Phys. Lett. A **297**, 261 (2002).

[10] Z. Dezso and A. Barabasi, Phys. Rev. E **65**, 055103(R) (2002).

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

[12] S. Staniford, V. Paxson, and N. Weaver, in *Proceedings of the 11th USENIX Security Symposium* (Advanced Computing Systems Association, San Francisco, 2002)

[13] MessageLabs Ltd. (Head Office), Gloucester, GL3 4AB United Kingdom.

[14] Computer Associates Virus Information Center, http://www3.ca.com/Virus

[15] Network Associates Virus Information Library, http://vil.nai.com/vil

[16] L.S. Liebovitch *et al.*, Phys. Rev. E **59**, 3312 (1999).

[17] B. B. Mandelbrot, *The Fractal Geometry of Nature* (Freeman, New York, 1982).

[18] J. Feder, *Fractals* (Plenum, New York, 1988).

[19] G. Rangarajan and M. Ding, Phys. Rev. E **61**, 4991 (2000).

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

[21] M. Faloutsos *et al.*, Comput. Common Rev. **29**, 251 (1999).

[22] A. B. Downey, ACM SIGCOMM Internet Measurement Workshop (2001).

[23] M.E. Corvella and Z. Bestavros, IEEE/ACM Trans. Netw. **5**, 835 (1997).

[24] W. Willinger *et al.*, IEEE/ACM Trans. Netw. **5**, 71 (1997).

[25] We assume that only one unit transmits viruses at a time so that the arrivals of the viruses from different units do not overlap, which would greatly complicate the computation of $P(t)$. This assumption may also have some reality in how the Internet functions and the model that it allows us to compute may also serve as a first useful step towards more complex models. We also note that our assumption here of consecutively active, nonoverlapping sources is the opposite of the limiting case analyzed by Willinger *et al.*, [24] who studied a large number of uncorrelated, overlapping sources.