Measuring and Modeling Computer Virus Prevalence

Jeffrey O. Kephart and Steve R. White
High Integrity Computing Laboratory
IBM Thomas J. Watson Research Center
P.O. Box 704, Yorktown Heights, NY 10598

Abstract

In an effort to understand the current extent of the computer virus problem and predict its future course, we have conducted a statistical analysis of computer virus incidents in a large, stable sample population of PCs and developed new epidemiological models of computer virus spread. Only a small fraction of all known viruses have appeared in real incidents, partly because many viruses are below the theoretical epidemic threshold. The observed sub-exponential rate of viral spread can be explained by models of localized software exchange. A surprisingly small fraction of machines in well-protected business environments are infected. This may be explained by a model in which, once a machine is found to be infected, neighboring machines are checked for viruses. This "kill signal" idea could be implemented in networks to greatly reduce the threat of viral spread. A similar principle has been incorporated into a cost-effective anti-virus policy for organizations which works quite well in practice.

1 Introduction

Rational anti-virus policies must be based upon accurate information about computer virus prevalence and a solid understanding of the factors which govern it. These two essential ingredients have been sadly lacking.

A few years ago, many people severely underestimated the magnitude of the computer virus problem — even claiming that viruses were a myth. In 1992, the opposite myth of Michelangelic Armageddon was promulgated by the media.

The frenzy over the Michelangelo virus was a dramatic illustration of the general unavailability of information on virus prevalence. Estimates of the number of computers infected by Michelangelo ranged over three orders of magnitude (to as high as 5 million worldwide! [1]), contributing greatly to widespread concern and handsome profits for anti-virus software vendors.

Recently, Certus [2] and Dataquest [3] have attempted to measure the extent of the computer virus problem by surveying hundreds of business, government, and educational organizations in the United States. They made some interesting discoveries — for example, the minimal extent to which most organizations are armed against computer viruses. Unfortunately, however, a number of fundamental conceptual and methodological problems prevented them from getting a clear picture of the prevalence of computer viruses. The substantial overestimates of the number of Michelangelo infections can almost certainly be traced to an understandable misinterpretation of some of the Dataquest results by the media and by some prominent people in the anti-virus industry [4, 5].

Given that the current prevalence of computer viruses has been subject to tremendous exaggeration in both directions, it should hardly be surprising that predictions of their future prevalence have been subject to exaggeration as well. In March, 1990, a well-publicized claim was made that viruses would increase in number exponentially, and that 8 million PCs would be infected by March, 1992 [2]. Based on this theory, it was concluded that virus scanning was ineffectual, and that the only solutions were either broad usage of restricted function computers or a massive campaign to strictly control the execution of all software on all of the world's PCs.

Our own observations of real-world virus incidents on a large, stable population of PCs and our theoretical modeling of computer virus spread reveal a much more realistic picture of the situation, and provide a much different, less drastic set of recommendations for dealing with the problem.

In Section 2, we review briefly some of our previous theoretical work. This provides the context for two new epidemiological models described in Section 3. Both of these models serve two purposes: they may help to explain some of the observations presented in Section 4, and they lead to prescriptions for novel anti-virus technologies and policies. In Section 4, we present some of our real-world virus statistics, interpreting them in the light of our theoretical results. We summarize our findings and discuss future directions in Section 5.

2 Epidemiological Models

In our modeling of computer virus spread [6, 7], we have borrowed some important concepts and simplifications from the well-established field of mathematical epidemiology [8] [1].

In particular, we ignore the details of infection within an individual (in our case, a computer system,

[1] The reader can consult reference [6] for a critique of other attempts at modeling virus spread, a brief review of mathematical epidemiology, and a more extensive list of references than is provided here.

along with all associated storage media), considering it to be in one of a small number of discrete states, such as *infected* or *susceptible*. Furthermore, we ignore the details of how disease is transmitted among individuals. We assume that, from time to time, individuals have "adequate contacts" with one another, resulting in transmission of the disease if one individual is infected and the other is susceptible. The details of what constitutes adequate contact vary from one disease (or computer virus) to another, but we simply assume that the total rate of adequate contacts between one individual and the rest of society is β. We also assume that there is some death rate δ at which the individual is cured of the infection [2].

For computer viruses, the rate of adequate contact β is influenced by anything that promotes or hinders viral replication, including mechanisms by which the virus infects programs, the rate of software transfer among computers, and precautions taken by users such as the use of a write-protect tab or integrity maintenance systems. The death rate δ is influenced by intrinsic characteristics of the virus which might disguise or reveal its presence, user awareness and vigilance, and detection (and subsequent removal) of the virus by anti-virus software.

In addition to borrowing ideas from mathematical epidemiology, we have extended it by incorporating topological effects which turn out to be quite important [6, 7]. In the homogeneous mixing assumption, every individual in the population is assumed to be equally likely to infect or to be infected by every other individual. Our work has shown that this approximation works well when each individual has many randomized contacts with others. However, if the number of contacts that a typical individual has with others is fairly small and/or the pattern of contacts is more or less localized, the approximation fails terribly. We suspect that the majority of today's computer populations are characterized by a degree of sparsity and locality that invalidates the homogeneous mixing approximation.

Figure 1 exemplifies a situation in which individuals (represented by nodes in the graph) are connected in both a sparse and a local manner. It can be thought of as representing a likely scenario in which workers within one group exchange software frequently among themselves, somewhat less frequently with other members of their department, and even less frequently with users in other companies, universities, or countries. The resulting topology contains random hierarchically-nested clusters with occasional cross-links. It is said to be *sparse* because each individual has adequate contacts (represented by edges of the graph) with just a few others. In other words, the average degree of the nodes in the graph is some small constant independent of the size of the graph. It is said to be *local* because, if nodes B and C are neighbors of (*i.e.* connected to) A, the probability for B and C to be neighbors is significantly enhanced over what it would be in a random graph.

By analyzing and simulating viral spread on a variety of topological structures, we have reached the following conclusions [3]:

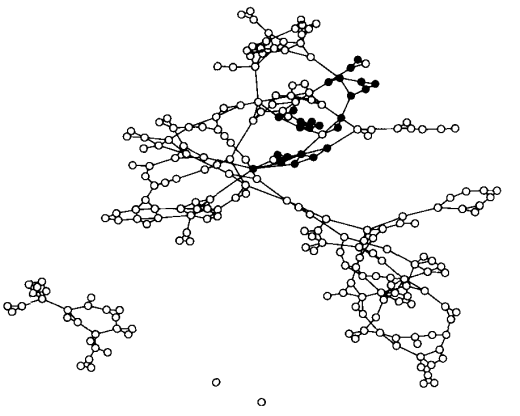

Figure 1: Snapshot of viral-spread simulation running on sparsely-connected, hierarchically-clustered topology. Each individual, represented by a node, has adequate contact with an average of three others. White and black nodes represent uninfected and infected individuals, respectively. The pattern of exchange is fairly localized, and therefore so is the pattern of infection.

1. In homogeneous systems (fully-connected graphs), an epidemic threshold occurs when $\rho \equiv \frac{\delta}{\beta} = 1$. When $\beta > \delta$ ($\rho < 1$), the system is above the "epidemic threshold", and an epidemic occurs with probability $1 - \rho$. If it does occur, the number of infections increases exponentially ($\sim e^{(\beta-\delta)t}$), eventually saturating at an equilibrium of $N(1-\rho)$, where N is the number of nodes. Below the epidemic threshold ($\beta < \delta$; $\rho > 1$), small outbreaks may occur whenever the disease is introduced into the population, but they can not be sustained for long.

2. In sparse systems, the epidemic threshold still exists, but the critical ratio $\rho_{\text{threshold}}$ is diminished to some value less than 1. As the average degree of nodes in the graph diminishes, so does $\rho_{\text{threshold}}$, and the probability of an epidemic diminishes (dropping to zero if $\rho_{\text{threshold}}$ slips below ρ). Even when an epidemic does occur, the growth rate is slowed, and the equilibrium level of infection depressed below what it would be in the corresponding homogeneous system.

3. In localized systems, the epidemic threshold and the equilibrium level of infection may or may not be affected. What is certain is that the growth in the number of infections with time is slowed qualitatively, becoming strongly sub-exponential.

[2] Or possibly dies of it, in the case of a biological host.

[3] These conclusions hold when individuals become susceptible immediately after they are cured of the disease.

3 Two New Models

In this section, we present two new epidemiological models which may help to explain some observed phenomena, and which suggest some new ideas which can be incorporated into anti-virus technology and policies.

3.1 Kill Signals

In all epidemiological models of which we are aware, an individual's cure takes place independently of that of any other individual. However, consider the following scenario. One day, Alice discovers that one of the programs she uses on her PC is infected with a virus. She eradicates it by any one of a number of procedures. In most models, this would be the end of the story. However, in this case Alice takes it upon herself to inform her friends Bob, Carol, and Dave, with whom she remembers having exchanged software sometime during the last few weeks. Bob already has a virus scanner, so he runs it and finds that, lo and behold, he has the virus, too. Carol and Dave install anti-virus software on their machines, whereupon Carol finds that she is infected, too. Fortunately for Dave, he turns out not to be infected. Bob and Carol, after cleaning up their PCs and diskettes, follow Alice's example and tell *their* friends, and the process continues until finally no one who receives the "kill signal" (the warning about possible viral infection) finds that they have the virus.

Various assumptions can be made about how the kill signal works. If one assumes that the kill signal is delivered and acted upon much more rapidly than the virus can spread, it can be shown that the virus can be pushed below the epidemic threshold even if the infected individual only delivers the kill signal to a fraction of its associates.

Figure 2 summarizes the results of nearly 200,000 simulation runs on random graphs of 100 nodes with average degree 10 (*i.e.* on average, each node had 10 neighbors). The ratio ρ was fixed at 0.2. At the beginning of each simulation run, an entirely new random graph was generated, and one randomly-chosen node was infected initially. At the moment that a node became cured, it tried to send a kill signal to each of its neighbors, which received it with probability p_{kill}. Upon successful receipt of the signal, an infected node would immediately become cured and would instantly try to send a kill signal to each of its neighbors. Any uninfected node would remain uninfected, and would *not* send the signal on to its neighbors. Given the relatively high degree of the graph and the random nature of the connections, the homogeneous approximation gives a good estimate of the epidemic probability when $p_{\text{kill}} = 0$: $1 - \rho = 0.8$. As the kill signal probability p_{kill} is increased, the epidemic probability remains high until a sharp threshold is reached near $p_{\text{kill}} = 0.25$, beyond which the probability for there to be an epidemic drops abruptly to zero. In other words, for these parameters, extinction of the virus is inevitable if more than 2.5 or 3 out of a typical node's 10 neighbors receive and heed the kill signal.

What if the kill signal is *not* transmitted instantaneously? One way to model this is to treat the kill signal as an *epi*-epidemic — a sort of anti-virus epidemic of species K riding on the back of the virus epidemic (species V). We can then assign the kill signal K its own intrinsic "adequate contact" and "death" rates. Specifically, a kill signal is born at a node whenever the virus dies there. Adequate contact among nodes occurs at the rate β_K. In order for adequate contact to result in infection by a K, the intended victim must be infected with V. In order to prevent the kill signal from ricocheting around the system and using up computational resources long after V has been eradicated, we introduce a death rate δ_K. The properties of V remain essentially the same as in standard models; V can only infect nodes that are not infected with K or with V.

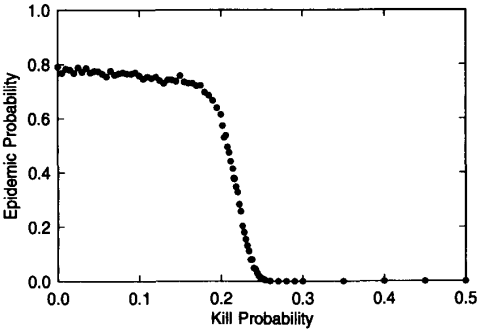

Figure 2: Kill signals can be extremely effective, even when only a fraction of the neighbors of an infected node receive and heed them. Each point summarizes the result of 2500 simulation runs on random graphs of 100 nodes with average degree 10, and indicates the fraction of those runs which resulted in epidemics. The ratio ρ was 0.2.

In a homogeneous system, deterministic analysis (valid for sufficiently large systems) leads to the following coupled pair of nonlinear differential equations:

$$\begin{aligned} \frac{dv}{dt} &= \beta v(1-v-k) - \delta v - \beta_K vk \\ \frac{dk}{dt} &= \beta_K vk - \delta_K k + \delta v \end{aligned} \qquad (1)$$

Equation 1 can be solved numerically to yield $v(t)$ and $k(t)$, the fraction of nodes occupied by V and K, respectively.

Analysis of the solution shows that the epidemic threshold for the virus is unaffected by the kill signal parameters; it remains at $\rho_{\text{threshold}} = 1$. The kill signal has no intrinsic epidemic threshold; it can survive as long as there are viruses upon which to feed, regardless of the relative values of δ_K and β_K. Unlike the first type of kill signal, this second type fails to alter the epidemic threshold. However, it can still be quite effective. By setting the left hand sides of Eq. 1 to zero, one can show that the equilibrium virus population can be made arbitrarily small either by setting δ_K sufficiently low or by setting β_K sufficiently high.

The kill signal parameters δ_K and β_K have several interesting limits. The limit $\delta_K \to 0$ has a simple interpretation: each individual acquires permanent immunity after exposure to and recovery from the virus V. The standard SIR models of mathematical epidemiology (*susceptible* $\to$ *infected* $\to$ *recovered*) are obtained by further taking the limit $\beta_K \to 0$. In this case, K isn't really a signal passed among neighboring nodes; it just appears spontaneously whenever a node is cured of V and remains there for eternity, protecting that node from infection by V. In such models, the equilibrium virus population is always zero; the question of interest is how many nodes *ever* become infected; this is determined by the virus rates β and δ [8]. In the limit $\beta_K \to \infty$, the kill signal becomes instantaneous. If δ_K is finite, this reproduces the $p_{\text{kill}} = 1$ limit in the first kill-signal model.

Figure 3a illustrates the population dynamics of V and K for a particular set of parameters for which V is above the epidemic threshold: $\delta = 1.0$, $\beta = 5.0$, $\delta_K = 0.1$, and $\beta_K = 0.5$. Thus $\rho_K = \rho = 0.2$, but the life cycle of K is ten times slower than that of V. In the absence of the kill signals, the fraction of nodes infected with the virus in equilibrium would have been $1-\rho = 0.8$. The kill signals strongly suppress the equilibrium virus population — by a factor of nearly 16 in this example. The predator-prey oscillations and high initial peak in the fractional virus population could be quelled by making β_K much larger. In a practical situation, one might also wish the kill-signal population to be fairly low; this could be achieved by making δ_K larger (at the expense of increasing the equilibrium virus population).

What is the effect of kill signals in non-homogeneous topologies? In sparse topologies, epidemics *can* be eliminated entirely, as illustrated by the simulation run shown in Fig. 3b. In this case, all parameters are the same as in Fig. 3a, but the topology is a random graph of 10,000 nodes with average degree 2.0. After a short-lived growth spurt, the virus population becomes extinct near $t = 16.4$. Their supply of viruses having run out, the kill signal population decays exponentially due to the death rate δ_K, and becomes extinct near $t = 89.1$. In all 200 simulation runs that were conducted under these conditions, V and K never came close to surviving up to the time limit $t = 1000$. Thus it appears that the sparsity of the graph has plunged the system below the epidemic threshold.

In local topologies, kill signals introduce some interesting new effects. Fig. 3c shows the populations of V and K for a typical simulation run on a 100-by-100 square lattice (wrapped around in both dimensions to form a torus). Each node (vertex) of the lattice is only able to infect its eight nearest neighbors. Although the rates β and δ are identical to those used in the other two topologies presented in Fig. 3, the population dynamics are remarkably different. Initially, the population growth of V and K are quadratic rather than exponential. This is in accord with previous studies which did not include the kill signal [6], in which it was found that the virus population grew as t^D in a D-dimensional lattice (t represents time). However,

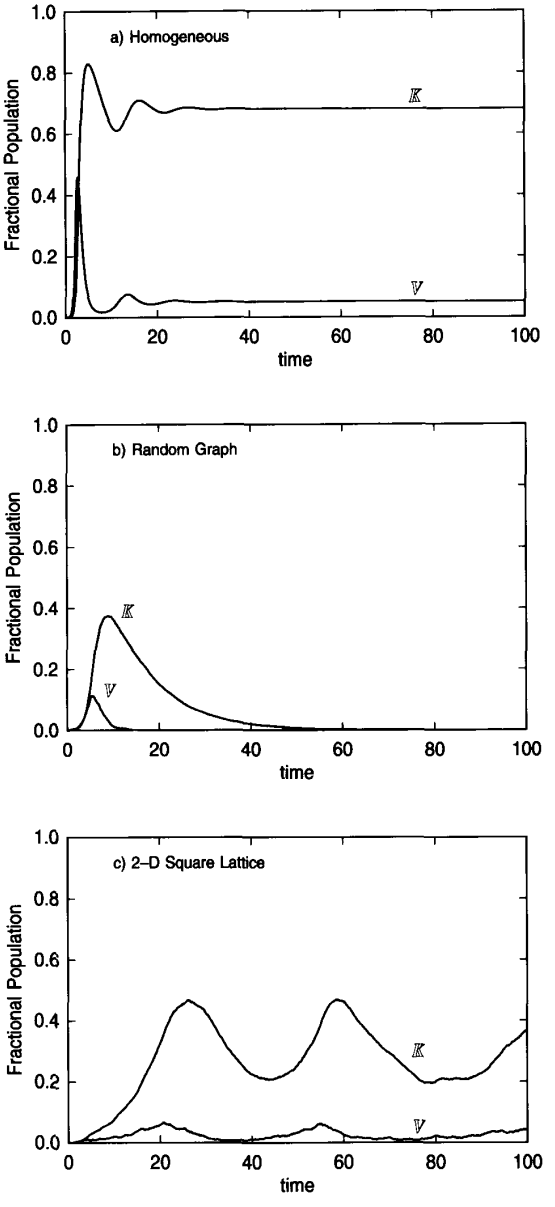

Figure 3: The effect of various topologies on the population dynamics of viruses and kill signals: a) homogeneous mixing model, b) 10000-node random graph with $d = 2.0$, and c) 100-by-100 square lattice wrapped around to form a torus. The rates are $\delta = 1.0$, $\beta = 5.0$, $\delta_K = 0.1$, and $\beta_K = 0.5$ in all cases. The homogeneous mixing curves are numerical solutions of a coupled pair of differential equations in which the initial fractional populations of V and K were 0.0001 and 0.0, respectively. The 2-D square lattice and random graph curves were obtained from typical simulation runs in which initially just one node (out of 10000 total) was occupied with V and none with K. (Recall that a K is born whenever a V dies.)

after a while, large undamped oscillations develop in the populations of V and K, and they are centered at a value which is substantially lower than in the homogeneous topology. Large undamped oscillations are not peculiar to spatial topologies; simulations on other topologies suggest that this phenomenon is generally characteristic of local topologies. Locality allows separate pockets of V and K to periodically develop, interact, separate, and then interact again [7].

These theoretical results on kill signals are exciting because they suggest a very cost-effective technique for thwarting viral spread. A number of different implementations can be considered, including user education (getting people to tell their friends if they discover a computer virus) and organizational policies which encourage users to report virus incidents to a central agency, which can then ensure that machines in the vicinity of the infected machine are scanned for viruses (and cleaned up if necessary). We are currently examining the feasibility of a technological implementation of kill signals for use in networks and other multi-user systems.

3.2 Viral Spread in Organizations

The second new model views viral spread from the perspective of an organization. This establishes a connection between important theoretical parameters and quantities that we can (and have) measured in our studies of virus incidents. In addition, it suggests an important strategy for limiting viral spread within organizations.

From an organization's perspective (Fig. 4), the world is full of computer viruses that are continually trying to penetrate the semi-permeable boundary that segregates the organization from the external world. At a rate depending on the number of computer virus infections in the world, the number of machines in the organization, and the permeability of the boundary, a computer virus will sooner or later make its way into the organization. This marks the beginning of a *virus incident.* After the initial penetration, the virus may spread among several other machines within the organization. Eventually, some user will discover that his machine is infected, and take steps to eliminate it. In the ideal case, that user will also inform either his neighbors or some central agency, which will then look for the virus on neighboring machines. The incident terminates when all machine infections stemming from the initial one are cleaned up.

An organization should have two goals: to limit the influx of viruses and to limit internal spread whenever a virus does manage to penetrate the organizational boundary. Centralized reporting and response can provide much valuable information about these two aspects of the organization's success in dealing with the virus problem. The number of *incidents* reflects the success of the organization in filtering out infectious contacts with the external world. It can also be used to infer the relative trends in virus prevalence in the external world, provided that the organization or collection of organizations being monitored is large enough to yield decent statistics. We shall carry this out in Section 4. The average incident *size* (the num-

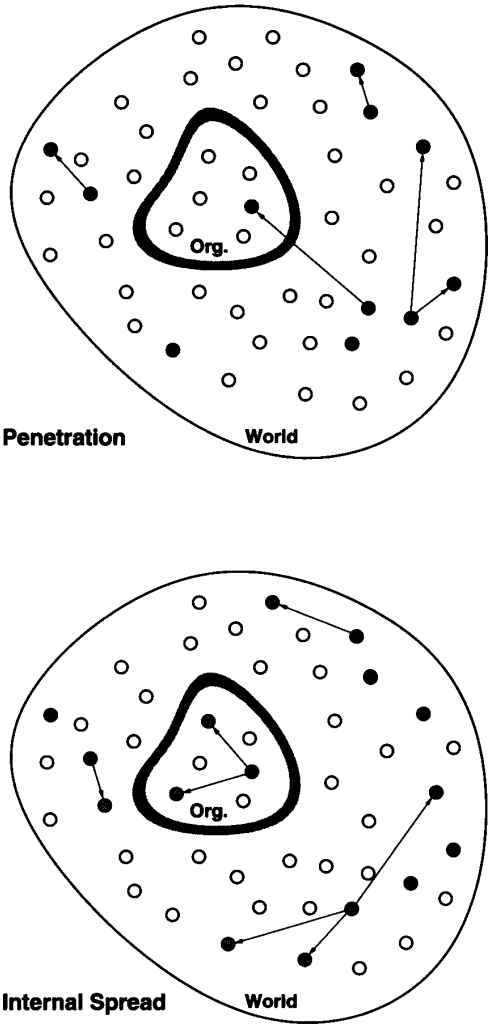

Figure 4: Computer virus spread from an organization's perspective. White circles represent uninfected machines, black circles represent infected machines, and gray circles represent machines in the process of being infected. Throughout the world, computer viruses spread among PCs, many of them being detected and eradicated eventually. Left: Occasionally, a virus penetrates the boundary separating the organization from the rest of the world, initiating a virus incident. The frequency with which this occurs depends upon the fraction of infected machines in the world, the number of machines in the organization, and the success of the organization in filtering out infectious contacts with the outside world. Right: The infection has spread to other PCs within the organization. The number of PCs that will be infected by the time the incident is discovered and cleaned up (the *size* of the incident) depends upon inherent characteristics of the virus and the effectiveness of the organization's anti-virus policies, particularly the extent to which anti-virus software is being used.

ber of infected machines per incident) reflects the organization's success in limiting the spread of viruses once they get into the organization. The next two subsections treat these two characteristics of virus incidents from a theoretical point of view.

3.2.1 Viral Influx

Let us consider the relationship between world-wide virus prevalence and the number of virus incidents observed in a large sample population as a function of time.

If the number of machines in the organization and the permeability of the organizational boundary remain constant, the number of incidents per unit time $I(t)$ is proportional to the number of infected machines in the world (at least the part of the world with which the organization comes in contact). However, all that a central reporting agency can record is $I_{obs}(t)$, the number of incidents *observed* per unit time.

If the δ and β are reasonably constant with time, the observed incident rate $I_{\rm obs}(t)$ is related to the actual incident rate $I(t)$ via

$$I_{\rm obs}(t) = \int_0^\infty d\tau Q(\tau) I(t-\tau) \qquad (2)$$

where $Q(t)$ is the probability density for the incident duration to be t. To a rough approximation, $I_{\rm obs}(t) = I(t-Q)$, where Q is the average incident duration. In other words, $I_{\rm obs}(t)$ is approximately a time-delayed version of $I(t)$.

The assumption of constant δ and β usually holds, but was violated severely during the period of Michelangelo Madness, as we shall show in Section 4. In the future, more sophisticated theories may in fact allow us to estimate Q from incident statistics taken in the months surrounding March 6, 1992, the much-publicized trigger date of the Michelangelo virus.

3.2.2 Internal Spread

We can get some insight into the second issue — that of internal spread — by the following simple model.

Let us assume that central reporting and response are perfectly effective, so that an incident is completely cleaned up as soon as any machine is found to be infected. We wish to know:

1. How many machines are typically infected before the incident is discovered and cleaned up (*i.e.* what is the distribution of incident sizes and its average)?
2. What is the average duration of an incident, both in general and as a function of the incident size?

To make the problem tractable, let us assume that homogeneous mixing applies within the organization. Then, an excellent approximation to the distribution of incident sizes can be derived as follows. Suppose that a virus has infected a machine in an organization, and that after some period of time the number of infected machines stemming from this initial event is n. The next event will be either a birth (resulting in $n+1$ infections) or a death (resulting in 0 infections, assuming that the clean-up is instantaneous and contemporaneous with detection by one of the machines). The rate at which deaths occur is simply $n\delta$, and the rate at which births occur is $n(1-n/N)\beta$, where N is the total number of machines in the organization. Thus the probability of going from n infections to $n+1$ infections is

$$p_{n\to n+1} = \frac{(1-n/N)}{(1-n/N)+\rho} \approx \frac{1}{1+\rho} \qquad (3)$$

with the approximation being valid to the extent that $n \ll N$. Then the probability that an incident will be discovered and cleaned-up after n machines are infected is:

$$P(n) = (1-p_{n\to n+1}) \prod_{i=1}^{i=n-1} p_{i\to i+1} \approx \frac{\rho}{(1+\rho)^n} \qquad (4)$$

Thus the size distribution is very nearly exponential, with mean μ given by:

$$\mu = \sum_n nP(n) \approx 1 + \frac{1}{\rho} \qquad (5)$$

which is valid provided that $\rho N \gg 1$ (or equivalently $\mu \ll N$). Previously, we found that in the absence of centralized response, an epidemic can occur if $\rho < 1$. However, Eq. 5 shows that, given perfect centralized response, the average incident size is $\ll N$ even when $\rho < 1$, provided that ρ is not so small as $O(1/N)$.

Note that, if the average incident size is less than two, the organization is below the epidemic threshold, and viruses would not propagate much even if central response were suddenly eliminated. However, if the average incident size is greater than two, the organization is intrinsically above the epidemic threshold, and elimination of central response would make the organization highly susceptible to widespread propagation of any virus that happened to enter it.

As a first step in deriving the distribution of incident *durations*, we can calculate the probabilities $p(n,t)$ for there to be n infections at time t. Suppose that there are n infected machines at time t. Then the probability per unit time of making a transition to $n+1$ infected machines is $R_{n\to n+1} = \beta n(1-\frac{n}{N})$. The probability per unit time of discovering the virus on one machine (and thus making an instantaneous transition to 0 infected machines) is $R_{n\to 0} = \delta n$. From these considerations we obtain the coupled differential equations:

$$\begin{aligned} \frac{dp(n,t)}{dt} &= -p(n,t)\left[R_{n\to n+1} + R_{n\to 0}\right] + \\ &\quad p(n-1,t)R_{n-1\to n}, \end{aligned} \qquad (6)$$

valid for $n \geq 1$. $p(0,t)$ can be obtained either from the rate equation:

$$\frac{dp(0,t)}{dt} = \sum_{n\geq 1} p(n,t)\delta n \tag{7}$$

or the normalization condition:

$$p(0,t) = 1 - \sum_{n\geq 1} p(n,t). \tag{8}$$

Typically, we are interested in solving Eq. 6 given the initial condition $p(1,t) = 1$; $p(n,t) = 0$, $n \neq 1$.

If we make the approximation $n \ll N$, we can solve Eq. 6 analytically. Consider the equation for $p(1,t)$:

$$\frac{dp(1,t)}{dt} = -p(1,t)\left[\beta + \delta\right]. \tag{9}$$

Given the initial condition $p(1,0) = 1$, we immediately obtain:

$$p(1,t) = e^{-(\beta+\delta)t} \tag{10}$$

The equation for $p(2,t)$ is:

$$\frac{dp(2,t)}{dt} = -p(2,t)\left[2(\beta + \delta)\right] + \beta p(1,t). \tag{11}$$

Using the method of integrating factors and the initial condition $p(2,0) = 0$, we obtain the solution:

$$\begin{aligned} p(2,t) &= \int_0^t e^{2(\beta+\delta)(t_1-t)}\beta p(1,t_1)dt_1 \\ &= \frac{\beta}{\beta+\delta}\left[1 - e^{-(\beta+\delta)t}\right]e^{-(\beta+\delta)t} \end{aligned} \tag{12}$$

In general, the solution for $p(n,t)$ can be expressed as a convolution involving $p(n-1,t)$:

$$\begin{aligned} p(n,t) &= \int_0^t e^{n(\beta+\delta)(t_1-t)}(n-1)\beta p(n-1,t_1)dt_1 \\ &= \left[\frac{\beta}{\beta+\delta}\left[1 - e^{-(\beta+\delta)t}\right]\right]^{n-1} e^{-(\beta+\delta)t}, \end{aligned} \tag{13}$$

as can be shown by induction. To obtain $p(0,t)$, we can insert Eq. 13 into the normalization condition given by Eq. 8. Summing the resulting geometric series, we obtain:

$$p(0,t) = \frac{\delta(1 - e^{-(\beta+\delta)t})}{\delta + \beta e^{-(\beta+\delta)t}} \tag{14}$$

It is straightforward to verify that this solution for $p(0,t)$ also satisfies the rate equation (Eq. 7). As one would expect, $p(0,t)$ increases monotonically from 0 at $t = 0$ towards 1 as $t \to \infty$.

Having obtained analytic formulas for the probabilities $p(n,t)$ of n infections at time t, we can now use them to calculate several quantities of interest. As a simple warmup exercise, we can calculate the distribution of incident sizes, which was derived earlier by another method. The probability for there to be n infections at time t followed by a transition to 0 infections at some time t' in the infinitesimal interval $t < t' < t + \Delta t$ is $p(n,t)\delta n \Delta t$. Integrating over all possible "extinction" times t, we obtain the probability $P(n)$ that the incident size was n:

$$\begin{aligned} P(n) &= \int_0^\infty dt p(n,t) n \\ &= \frac{\rho}{(1+\rho)^n} n \int_0^1 dx x^{n-1} \\ &= \frac{\rho}{(1+\rho)^n}, \end{aligned} \tag{15}$$

in agreement with the result given by Eq. 4. (The substitution $\left[1 - e^{-(\beta+\delta)t}\right] \to x$ was made in going from the first line to the second in the above derivation.)

The duration distribution $Q(n,t)$ for an incident of size n is simply the extinction time distribution normalized such that $\int_0^\infty dt Q(n,t) = 1$ for all n:

$$\begin{aligned} Q(n,t) &= \beta n (1+\rho)^n p(n,t) \\ &= (\beta+\delta) n \left[1 - e^{-(\beta+\delta)t}\right]^{n-1} e^{-(\beta+\delta)t}. \end{aligned} \tag{16}$$

To obtain the average duration $Q(n)$ of an incident of size n, we need to solve the following integral:

$$\begin{aligned} Q(n) &= \int_0^\infty dt\, t Q(n,t) \\ &= -\frac{n}{\beta+\delta}\int_0^1 dx\, x^{n-1}\ln(1-x) \\ &= \frac{1}{\beta+\delta}\sum_{j=1}^{n}\frac{1}{j}. \end{aligned} \tag{17}$$

For sufficiently large n, Eq. 17 is approximately

$$Q(n) \approx \frac{1}{\beta+\delta}\left[\ln(n) + \gamma + \frac{1}{2n}\right], \tag{18}$$

where $\gamma = 0.57721\ldots$ is Euler's constant. Thus the expected duration of an incident scales logarithmically with its size. This can be attributed to the exponential growth in the number of infections with time, a hallmark of the homogeneous approximation.

To obtain the overall duration distribution $Q(t)$, we can average the distribution $Q(n,t)$ over all incident sizes n (using the weighting factor given by Eq. 4). Alternatively (and more simply), we can note that

$$\begin{aligned} Q(t) &= \frac{dp(0,t)}{dt} \\ &= \left[\frac{\beta+\delta}{\delta+\beta e^{-(\beta+\delta)t}}\right]^2 \delta e^{-(\beta+\delta)t} \end{aligned} \tag{19}$$

Finally, the overall average duration Q is given by:

$$\begin{aligned} Q &= \sum_{n=1}^{\infty} \frac{\rho}{(1+\rho)^n} Q(n) \\ &= \frac{1}{\beta+\delta} \sum_{j=1}^{\infty} \frac{1}{j} \sum_{n=j}^{\infty} \frac{\rho}{(1+\rho)^n} \\ &= \frac{1}{\beta+\delta} \sum_{j=1}^{\infty} \frac{1}{j} \left[\frac{1}{1+\rho}\right]^{j-1} \\ &= \frac{1}{\beta} \ln(1+\rho^{-1}) = \frac{\ln(\mu)}{\beta}. \end{aligned} \tag{20}$$

In the above derivation, the order of summation was switched in going from the first line to the second, and the fourth line was obtained from the third by identifying the Taylor series expansion for $\ln(x)$. Of course, the same result could have been obtained by performing the integral $Q = \int_0^\infty dt\, tQ(t)$.

The rates β and δ figure prominently in the various expressions for probability distributions and averages of the incident size and incident duration. By measuring the average incident size μ in a particular organization with good central reporting and response, we might hope to use Eq. 5 to estimate $\rho = \frac{\delta}{\beta}$ in that organization. In order to estimate β and δ separately, we could combine this estimate of ρ with a measurement of the average incident duration and use Eq. 20.

For several reasons, such an exercise might be difficult. Although data on the incident size distribution can be collected (see Section 4), data on incident durations are very difficult to obtain because it is hard to tell when an incident began. In addition, there are several idealizations in this particular model that may not reflect the real world. In principle (if they can be measured), the various probability distributions derived in this section can be used as independent checks of the validity of the approximations made. For example, in a population of individuals in which β and δ vary somewhat from one individual to another, we might expect the distribution of incident sizes to deviate from the exponential distribution predicted by Eq. 4. Indeed, as will be seen in the next section, the incident size distributions of our sample population exhibit a non-exponential tail. Another potential difficulty is the use of the homogeneous-mixing approximation in deriving these results. In the future, simulations will be used to assess the degree to which topology alters the theoretical results of this section. We expect the results for incident duration to be affected significantly because they appear to contain quantities associated with exponential growth. The results for incident size may be somewhat less affected, because they do not depend on the time scales involved.

Thus, a model based on the organizational perspective has the potential to help us measure important theoretical parameters, but attempts to do so now are probably premature. In the future, by incorporating topological and other effects into the theory and by finding ways of measuring either the average incident duration, β, or δ, we should be able to tie many attributes of virus incidents together and to estimate parameters that will help us predict virus spread on a global scale.

An additional point should be rescued from the morass of equations and emphasized very clearly here. Central reporting and response appears to be a powerfully effective policy. Even if an organization is intrinsically above the epidemic threshold, central reporting and response prevent the incident size from scaling with the number of machines in the organization. Not only do incidents remain small; their duration is finite (rather than infinite). As will be seen in the next section, our virus prevalence statistics also suggest that organizations should adopt this policy.

4 Virus Prevalence Statistics

For several years, we have collected statistics on virus incidents in a well-monitored population of several hundred thousand PCs as they occurred. For each incident, we recorded where and when it was reported, the number of infected PCs and diskettes, and the virus involved. As could be inferred from Section 3.2, this method requires that the population under observation possess three important characteristics:

1. Anti-virus software in regular use by users. Users must have the means to determine if they are infected. If they are, they must have a reliable way of determining the identity of the virus.

2. Educated users. Users must know what viruses are, how to use anti-virus software, and to whom they should report an infection if they discover one.

3. Central reporting. There must be a central reporting facility that collects information about virus incidents.

The particular sample population that we have chosen to study is international, but biased towards the U.S. It is stable, both in makeup and in size. We believe it to be typical of Fortune 500 companies possessing the three important characteristics cited above, plus active central response to incidents.

Of course, these characteristics are not typical of many other environments, so some of our results may not be representative of universities, home users, and other businesses which lack these characteristics. Nonetheless, from our observations of this population we are able to infer much about worldwide computer virus prevalence, as was explained in Section 3.2.1. We believe that our statistics provide the most accurate picture of virus prevalence that has yet been obtained [4].

[4] For a detailed critique of other attempts to gather statistics on virus prevalence, see references [4] and [5].

4.1 Incident Size Distribution

First, we present some interesting results on the distribution of incident sizes in our population which support our theoretical conclusion that central reporting and response can be quite effective. Fig. 5a shows the distribution of incident sizes during a six-month period when the above-mentioned anti-virus strategies were first being deployed in the various components of our sample population.

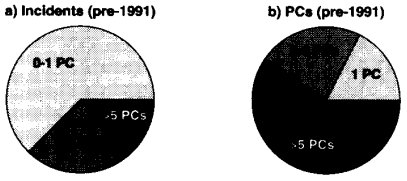

Figure 5: a) Fraction of incidents of given size during six-month periods when strategies were first being deployed. b) Fraction of infected PCs involved in incidents of given size during the same time period.

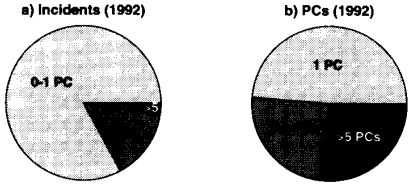

Figure 6: a) Fraction of incidents of given size during 1992. b) Fraction of infected PCs involved in incidents of given size during 1992.

During this period, the average incident size was 3.4 PCs. Most (63%) of the incidents involved just zero or one PCs. (The incident size is defined to be zero if a foreign diskette is caught before it can infect any of an organization's PCs.) Only 12% of the incidents involved more than 5 PCs. However, Fig. 5b presents a different view of the same data. Even though incidents larger than 5 PCs were fairly rare, they accounted for 60% of the total number of infected PCs. Thus the larger incidents actually accounted for most of the problem! Fig. 6 shows the corresponding distributions for 1992, after the anti-virus strategies had been in place for some time. The average incident size was cut by more than a factor of two to just 1.6 PCs. In the vast majority of cases (83%), the infection was caught before it could infect more than one PC. Only 2.5% of the incidents involved more than 5 PCs, and these large incidents accounted for only 27% of the total number of infected PCs.

It should be noted that these incident size distributions do not have the exponential form predicted by Eq. 4. For example, for the 1992 data, the average incident size of 1.6 leads to an estimated $\rho = 1.67$ (using Eq. 5.) (To the extent that the approximations of Section 3.2.2 are valid, the fact that the average incident size is less than two indicates that the population as a whole is intrinsically below the epidemic threshold.) For an exponential distribution with this average, the percentage of incidents involving no more than 1 PC should be 62.5%, the percentage involving 2 to 5 PCs should be 36.8%, and the percentage involving more than 5 PCs should be just 0.7%. The percentages that were actually observed were 83.0%, 14.5%, and 2.5%, respectively. Thus the tail of the distribution is noticeably longer than exponential. This may be due to a certain amount of variation in the β and δ rates among the various members of the sample population.

In any case, the net effect of the anti-virus policies introduced a few years ago was to create a more hostile environment for computer viruses, reducing the average incident size by a factor of two in this instance. In organizations which have not yet implemented active response policies, we can expect the average incident size to be larger than the 1.6 PCs that we have attained. As a check on this, we have been able to compare our results with those obtained by Dataquest [3]. Unfortunately, the question they asked their survey participants confused the distinction between incidents and infected machines. However, by making some assumptions about how the survey participants interpreted the question [4, 5], we find that, in the third quarter of 1991, the average incident size among the organizations surveyed by Dataquest was roughly between 2.4 and 3.2. This is reasonably close to the figure of 3.4 PCs that we observed in our population when anti-virus policies were just being put into place.

We are aware of some conscientious organizations not included in our sample population which, despite having purchased a site license for anti-virus software, suffer from persistent, chronic infections. These organizations appear to be above the epidemic threshold. The theoretical results of Section 3.2.2 indicate that, by implementing central reporting and response, these organizations could bring virus incidents to a swift termination without doing anything to change β and δ.

4.2 Worldwide Virus Prevalence

In the remainder of this section, we shall look at statistics which typify not just our sample population, but which also reveal much about virus prevalence in the world as a whole.

We have maintained a current collection of known viruses by working cooperatively with other virus collectors. At any given moment in time, the number of viruses for which we have signatures in our virus scanner is a conservative estimate of the number of different viral strains in the world. The number of viruses which are actually spreading is taken to be the number of viruses that we have seen in at least one actual incident.

Figure 7 shows that the number of different viruses

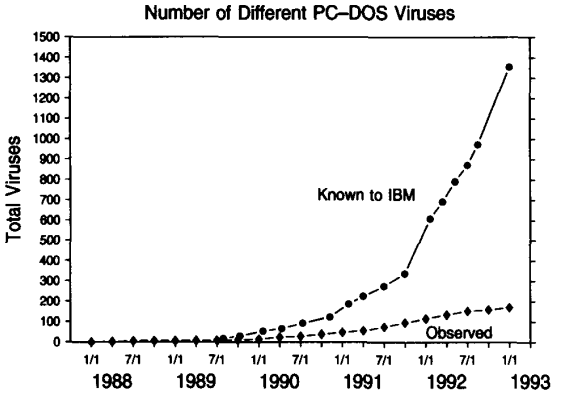

Figure 7: Number of viruses known to us (those we have collected and analyzed) and number of viruses "in the wild" (observed by us in actual incidents) as a function of time.

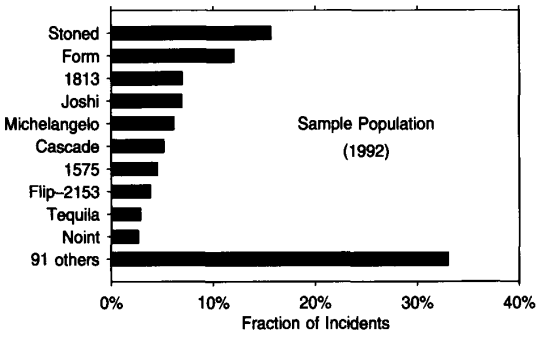

Figure 8: Relative frequency of incidents involving the most common viruses during 1992.

that have been written has grown dramatically during the last four years. So far, it has been growing at a roughly exponential rate, doubling approximately every 7 months. During the last two years, the number of viruses that we have seen in real incidents has consistently been approximately 15% to 20% of the total number in our collection, and a majority of these have only been seen once or twice. We suspect that a significant portion of the viruses which have been seen rarely or never are below the epidemic threshold. Others have just been unfortunate so far, and might conceivably get a lucky break someday that will enable them to spread appreciably [5].

Figure 8 emphasizes the point that a few viruses account for many, but certainly not all, of the observed incidents. The ten most common viruses accounted for 67% of the incidents, with the remaining 33% being distributed among 91 different viruses, over half of which were seen only once. A number of other viruses that we have seen in previous years were not observed at all during 1992. This leaves well over 1000 viruses in our collection that we have never observed at any time. It is interesting to note that the relative market share of the top two viruses has been declining steadily. In 1990, the Stoned and 1813 (Jerusalem) together accounted for 51% of all incidents. In 1991, this dropped to 34%. In 1992, Form supplanted 1813 as the second-most common virus; the Stoned and From viruses together accounted for just 28% of the observed incidents. This decrease is not due to decreased prevalence; it is due to the fact that new viruses are continually entering the field. Some of these newcomers, notably Joshi and Form, are proving to be rather successful. Other new viruses are seen only rarely, but there are so many of them that the fraction of incidents in the "Other" category is growing rapidly.

[5]Recall from Section 2 that viruses which are above the threshold can still die out if they fail to get a good start in life.

Figure 9 shows the observed incident rate of the five most common viruses of 1992 as a function of time. Except for the first half of 1992, it can also be taken to be the relative frequency of these viruses in the world as a whole. (During the first half of 1992, the Michelangelo scare caused an anomaly which disturbed the proportionality between observed and actual incidents; this will be discussed in Section 4.3.)

If we restrict our attention to the fourth quarter of 1991 and earlier (in order to avoid the confusion of the Michelangelo effect), a common pattern is evident in Fig. 9. Viruses appear to increase in prevalence at an approximately linear rate for a period of six months to two years, and then plateau at a very low level. Some, such as 1813, appear to decline after a relatively stable period. Bouncing Ball (not shown) had been in apparent equilibrium for several years, perennially appearing in the list of the 5 most common viruses; during 1992, its prevalence declined precipitously, to about one fourth of its former equilibrium level. This may indicate that viruses like 1813 and Bouncing Ball have fallen below the epidemic threshold, possibly because anti-virus software is being used more widely than it had been. The Brain is a prime example of a virus which is nearly extinct.

The viruses which are increasing in prevalence are clearly above the epidemic threshold, but their strongly sub-exponential spread rate points to highly localized software sharing. The ones which are approximately stable in prevalence are apparently in equilibrium. It is somewhat surprising that the equilibrium is at such a low level: approximately 0.2 incidents per 1000 PCs per quarter for Stoned, the most prevalent virus, at the end of 1991. To estimate the number of infected machines that this represents, we must multiply this figure by the average incident size for the world. Not knowing the extent to which most organizations are protected against viruses, this is difficult to estimate, but in any case it is clear that the fraction of the world's machines which are infected with any particular PC-DOS virus is exceedingly small. If we accept the simple theory of Section 2, this can only be explained if the birth rate is infinitesimally larger than

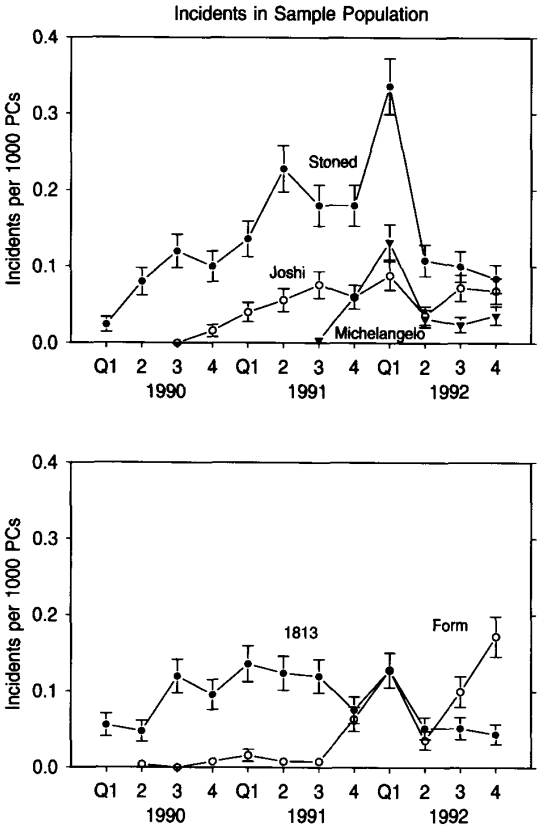

Figure 9: Number of incidents involving five of the most common viruses as a function of time. The units (incidents per 1000 PCs per quarter) pertain to our sample population only, but the curves should also be reasonable estimates of the *relative* worldwide prevalence of each virus. The data points are bracketed by bars indicating the statistical sampling error that one would expect given the number of observed incidents. (The bars do *not* represent errors in the measured data.)

the death rate. This seems very unlikely, especially since several viruses are apparently above threshold, but none have become very prevalent.

We suspect that a combination of two factors is decreasing the equilibrium. First, kill signals are probably operating informally, *i.e.* some people tell their friends when they discover that they are infected. Depending on the assumptions that one puts into the theoretical kill signal models of Section 3.1, this can decrease the equilibrium infection rate by a very substantial factor. Second, it is conceivable that, when someone experiences a computer virus, or hears that someone they know became infected, they become more vigilant. Unlike biological diseases, exposure to one computer virus can actually confer immunity against nearly all computer viruses. As was mentioned in Section 3.1, such an immunization effect can also be accommodated within the kill signal model by setting $\delta_K = 0$. This would also help to explain an extremely low equilibrium level of computer virus infections.

It should be noted that our observations completely contradict the predictions made by Tippett [2]. In March, 1990, he predicted that by March, 1992 there would be approximately 8 million infected PCs in the world — an 8% infection rate. He claimed that the 1813 (Jerusalem) virus would continue to double in prevalence every 1.5 to 2.6 months. In fact, according to Fig. 9, its prevalence remained remarkably stable over a long period of time following that prediction, and today it appears to be declining. He also predicted that, for any virus, exponential growth would continue until approximately 20% of the computer population was infected, after which its prevalence would continue to increase at a slower rate. Note that, even for those viruses in Fig. 9 which have increased in prevalence, it would be difficult to claim that the growth has been *exponential*! We attribute this to highly localized software sharing, as was described in Section 2.

Figure 10 shows the incident rate from all viruses as a function of time in our sample population. It can also be interpreted as the relative frequency of all viruses in the world as a whole. (Again, this is not the case during the first half of 1992 for reasons that will be explained in Section 4.3.) During the last quarter of 1991, about 0.1% of the PCs in our sample population became infected by some external source.

This small but rising increase is due to two separate factors. First, some individual viruses are becoming more prevalent. Second, there has been an increase in the number of different varieties of successfully-spreading viruses (*e.g.* the Joshi, which as shown in Fig. 9a first appeared in our sample population in late 1990). It should be recognized that the statistic shown here is distinct from that presented in Figs. 7 and 9, and can be thought of as a somewhat complicated combination of the two of them.

It is interesting to calibrate our measurements of the total virus incident rate against those of Dataquest [3], which sampled a much greater diversity of organizations. Unfortunately, a direct comparison with their results is not possible because they reported the percentage of *organizations* which experienced at least one incident during given time inter-

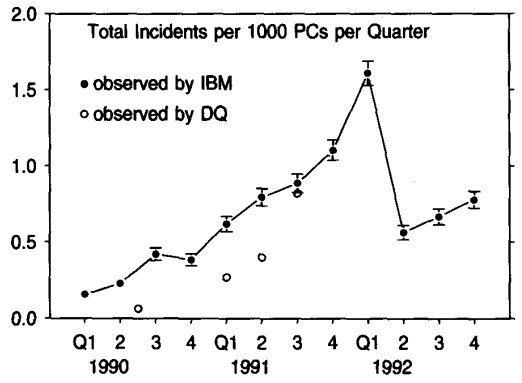

Figure 10: Total number of virus incidents in sample population as a function of time. The units (incidents per 1000 PCs) pertain to our sample population only, but the curve should also be proportional to the worldwide prevalence of all computer viruses as a function of time. The data points are bracketed by bars indicating the statistical sampling error that one would expect given the number of observed incidents. Estimates derived from the raw survey data collected by Dataquest are displayed as well.

vals. The organizations ranged over more than two orders of magnitude in size. However, by re-examining the Dataquest raw data, and taking into account the distribution of organization sizes, we have been able to de-convolve their results so as to provide an estimate of the number of incidents per 1000 PCs [5]. The Dataquest rate of approximately 0.81 incidents per 1000 PCs for the third quarter of 1991 is in the same range as our own observation of 0.90 incidents per 1000 PCs for that quarter. This suggests that the incident rate within our sample population is roughly the same as that in the rest of the world (at least the portion of the world sampled by Dataquest — North American businesses and educational and governmental institutions.) This is consistent with the fact that our sample population is not taking any unusual precautions to prevent viruses from penetrating the organizational boundary; the special anti-virus policies that were instituted a few years ago are designed to limit the *size* of incidents, not their frequency. The fact that the Dataquest data for the year 1990 is considerably lower than ours may indicate fading memory on the part of the survey respondents (who took the survey in October 1991), or a lesser awareness of the virus problem during 1990 than during 1991.

4.3 The Michelangelo Effect

The anomalous peaks in Figs. 9 and 10 in early 1992 require an explanation. Our data collection was strongly perturbed during that time by a very peculiar event: Michelangelo Madness [6]. Although this perturbation has greatly complicated the interpretation of our data, our struggle to cope with it has been instructive.

Figure 11 shows the number of observed incidents during two-week periods in 1992 for the Michelangelo and Stoned viruses, and for the total over all viruses except for these two. A quick glance shows that all three trends have approximately the same shape: a sharp rise to a peak on the two-week period ending on March 6^{th}, followed by a dip that bottoms out during April and a gradual recovery towards the rate that prevailed at the beginning of the year.

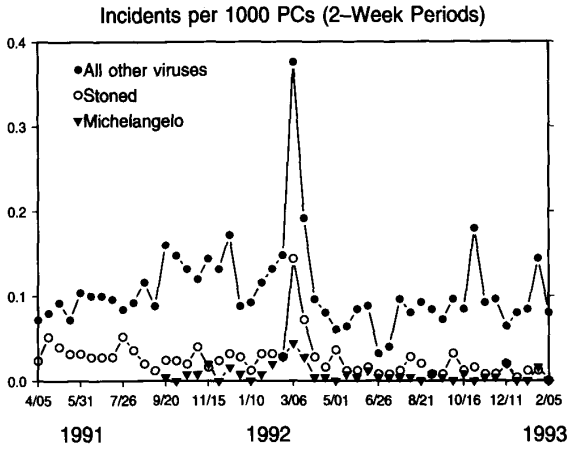

Figure 11: Number of virus incidents reported per 1000 PCs for Stoned, Michelangelo, and all other viruses during two week periods ending with the indicated date. The dashed line indicates the incident rate for all viruses other than Stoned and Michelangelo during the fourth quarter of 1991 (normalized to a two-week period by multiplying by $\frac{2}{13}$).

This behavior is completely at odds with the behavior in Figs. 9 and 10, in which dynamical trends occurred on much slower time scales. Furthermore, it is difficult to believe that the actual incident rate of Michelangelo, Stoned, and the sum total of all other viruses just happened to undergo a huge fluctuation in unison.

There is a simple explanation. In fact, the *actual* incident rate was not fluctuating as wildly as the *observed* incident rate shown in Fig. 11. We can surmise from this data that, about a week or two before March 6^{th} (the date on which Michelangelo was slated to damage the file system on hard disks), many users both inside and outside our sample population decided to scan their disks because they were concerned about being infected with Michelangelo. Since incidents were discovered sooner than they ordinarily would have been, there was a short-lived but dramatic peak in the observed incident rate. This depleted the reservoir of infection in our sample population, resulting in a noticeable dip in the observed infection rate during April, by which time the scanning rate had most likely returned to normal. This also explains the

[6] A particularly virulent form of March Madness, transmitted via casual exposure to newspapers or the evening news. Typical symptoms include hysteria, acute panic, and last-minute purchases of anti-virus software.

similarity in shape of the three curves in Fig. 11. In the course of scanning, a user would naturally have found *any* virus that happened to be on his or her system. One further anomaly should not be surprising in the light of our explanation: during the four weeks prior to March 6^{th}, eleven new viruses were seen for the first time in our sample population — a record high.

In slightly different terms, our usual assumption that the observed incident rate lags the actual incident rate by some *constant* amount (the average incident duration) broke down for the first time due to a sudden, pervasive, but temporary alteration in user behavior during late February and early March of 1992. We are currently trying to refine the theoretical analysis presented in Section 3.2.1 in order to help disentangle two effects:

- An actual reduction in the world's computer virus population (and hence the *actual* incident rate) due to scanning by users all around the world.
- The boom and bust in the observed incident rate due to scanning by users in our sample population.

If we succeed, we will be able to determine the extent to which the world's computer virus population was set back by Michelangelo. In addition, we may be able to estimate the average lag time Q, and hence β and δ for several of the most prevalent viruses.

In the aftermath of Michelangelo, the equilibrium level of infection for common viruses such as Stoned and 1813 appears to have dropped significantly (Fig. 9), as has the total virus incident rate (Fig. 10). Michelangelo Madness had a salutary effect on the world's virus population, perhaps reducing it by a factor of two overall. However, one would wish that this reduction had been achieved by more orderly, less costly means. We believe that our two-pronged (statistical and mathematical) epidemiological approach can help us devise more sensible ways to achieve even more dramatic reductions in worldwide virus prevalence.

5 Conclusion

A mutually-supportive combination of theory and observation has enabled us to infer much about computer virus prevalence and the factors which influence it.

Computer viruses are considerably less prevalent than many have claimed. The rate of PC-DOS virus incidents in medium to large North American businesses appears to be approximately 1 per 1000 PCs per quarter; the number of infected machines is perhaps 3 or 4 times this figure if we assume that most such businesses are at least weakly protected against viruses. Businesses with virtually no anti-virus protection can probably expect a higher rate than this, but we have no data on which to base an estimate.

Gradually, computer viruses are becoming more prevalent. This is not because any one viral strain is getting out of hand; it is because the number of different viruses is growing with time. Most viruses that are written appear to be below the epidemic threshold. Of the ones that we have seen, just a small minority account for a substantial majority of the incidents. The ones that are most successful seem to increase in prevalence for a year or two at a strongly sub-exponential rate (approximately linear!) and then level off at a very low level of incidence. This qualitatively slow spread rate indicates that software exchange is highly localized. It is good news for known-virus technology; it means that updates can be sent out less frequently than would be required if the growth rate were exponential. Even more so, it is very good news for all PC users, who should be thankful that previous predictions of exponential growth were so far off the mark.

Furthermore, previous claims about the ineffectiveness of virus scanning are discredited. Simple epidemiological models show that, by increasing the virus death rate sufficiently, one can push viruses below the epidemic threshold. Virus scanners are an effective way to increase the death rate, particularly if they are designed such that they scan periodically without any prompting from the user.

Finally, our observations and our theoretical analysis of the effect of centralized reporting and response suggest that this is an extremely effective way to manage the virus problem in organizations. We strongly recommend the following policies to all organizations:

1. Make sure that users use anti-virus software.
2. Make sure they know what viruses are and who to contact if they find one.
3. Make sure that the people they contact remove the reported infection (and others connected with it) quickly.

These policies have helped to cut the average incident size by more than a factor of two within our sample population. Furthermore, the information collected by the central agency can be used to assess the organization's progress in dealing with the computer virus problem.

Theoretical results on kill signals suggest that they are highly effective in reducing the virus threat. In the not-too-distant future, we plan to implement them in networks of PCs.

As time passes, our knowledge and understanding of the computer virus problem is bound to increase. With more data, trends in computer virus prevalence will become clearer. In addition, the theory will continue to advance in a number of directions. Currently, we can only say that the topology of software exchange among the world's computers has a very important effect, and that the global trends appear to indicate that it is highly localized. In order to make our theories more quantitative and predictive, we must find ways of characterizing the world's topology. From user surveys and automatic monitoring techniques, we hope to obtain enough information about individual behavior to be able to predict and to influence the future course of computer virus trends within organizations and throughout the world.

Acknowledgments

We are grateful to Alan Fedeli for establishing and overseeing IBM's central virus reporting agency, to Ralph Langham, Julian Banks, and Yann Stanczewski for collecting the data presented here, and to Dave Chess for his help in analyzing it. The kill signal idea owes its existence and its name to Bill Arnold. We also thank Peter Tippett of Certus (a division of Symantec) for supplying us with the raw data from the 1991 Dataquest survey.

References

[1] J. McAfee, quoting expert sources on The MacNeil/Lehrer News Hour, March 5, 1992.

[2] P.S. Tippett, "The Kinetics of Computer Virus Replication: A Theory and Preliminary Survey," *Safe Computing: Proceedings of the Fourth Annual Computer Virus and Security Conference,* New York, New York, March 14-15, 1991, pp. 66-87.

[3] Dataquest, "Computer Virus Market Survey for National Computer Security Association", 1991.

[4] J.O. Kephart and S.R. White, "How Prevalent Are Computer Viruses?," *Proceedings of the Fifth International Computer Virus and Security Conference,* March 12 – 13, 1992, New York, pp. 267-284.

[5] J.O. Kephart and S.R. White, "Measuring Computer Virus Prevalence," *Proceedings of the Second International Virus Bulletin Conference,* Edinburgh, Scotland, September 2–3, 1992, pp. 9-28.

[6] J.O. Kephart and S.R. White, "Directed-Graph Epidemiological Models of Computer Viruses," *Proceedings of the 1991 IEEE Computer Society Symposium on Research in Security and Privacy,* Oakland, California, May 20-22, 1991, pp. 343-359.

[7] J.O. Kephart, "How Topology Affects Population Dynamics," Proceedings of Artificial Life 3, Santa Fe, New Mexico, June 15-19, 1992.

[8] Norman T. J. Bailey, *The mathematical theory of infectious diseases and its applications,* second edition, Oxford University Press, New York, 1975.

[9] J.O. Kephart and S.R. White, "Commentary on Tippett's 'Kinematics of Computer Virus Replication'," *Safe Computing: Proceedings of the Fourth Annual Computer Virus and Security Conference,* New York, New York, March 14-15, 1991, pp. 88-93.