Computers and epidemiology

Jeffrey Kephart, David Chess, Steve White

Analogies with biological disease, with topological considerations added, show that the spread of computer viruses can be contained.

Computer viruses have bugged their hosts for half a dozen years or so. Massive outbreaks have been rare. But as society comes to rely ever more heavily on computers, contagious programs are beginning to seem nearly as frightening as biological diseases.

How bad is the problem today? How bad might it become? How might company managers help ensure safe computing environments? For answers to these questions, the behavior of computer viruses must be understood at two levels: microscopic and macroscopic.

The micro level is the focus of hundreds of researchers who dissect and try to kill off the dozens of new viruses written every month. Thanks to Fred Cohen's pioneering theoretical work, done in the early 1980s at the University of California, Los Angeles, computer viruses were understood in minute detail years before they posed even a slight threat.

In contrast, the macro view of computer viruses has lagged. The dearth of information about their prevalence was evident during last year's hullabaloo over the Michelangelo virus [see "The Michelangelo effect," opposite], during which estimates of its prevalence ranged over three orders of magnitude. Similarly, very few 'attempts have been made at modeling the spread of viruses mathematically, and most of these have contained serious flaws.

The situation is being remedied in two ways: by the collection of statistics from actual incidents, and by computer simulation of virus spread. This epidemiological approach -- characterizing viral invasions at the macro level--has led to some insights and tools that may help society to cope better with the threat (and which may aid the study of biological viruses, too).

Today, computer virus epidemiology is an emerging science that reveals that protective measures are definitely within reach of individuals and organizations. Among its findings:

- Computer viruses are far less rife than many have claimed. The rate of PC-DOS virus incidents for medium-sized to large businesses in North America appears to be about one per 1000 PCs per quarter. And fewer machines are caught up in a typical incident if anti-virus measures are in place.
- Few PC-DOS viruses have thrived. Less than 15 percent of the more than 1500 known viruses have ever been observed in a large sample population and most of them only once. The top 10 viruses account for two-thirds of all incidents.
- Because software and diskette sharing tends to be localized, even successful viruses spread at nowhere near the exponential rate that some have claimed. This is good news for the anti-virus industry, which otherwise would have to distribute its software updates even more often.
- Centralized reporting and response within an organization is an extremely effective defense. These policies have more than halved the average incident size within the population monitored by IBM Corp., and can eliminate chronic infections that may afflict even conscientious

organizations.

Biological analogy

Biologists have combined the micro- and macroscopic perspectives on disease to good effect. It turns out that biological diseases and computer viruses spread in closely analogous ways, so that each field can benefit from the insights of the other.

Detailed statistics on disease proliferation date from mid-17th century London. The first major triumph for empirical epidemiology occurred there in 1854 when the city was suffering from a severe outbreak of cholera. Studying the spread of the intestinal disease over time led the physician John Snow to suspect that one of the local water supplies was to blame. A few days after the water source, at his suggestion, was shut down, the cholera epidemic subsided.

A theoretical approach to epidemiology was undertaken in 1760, when Daniel Bernoulli, one of the founders of mathematical physics, decided to model contagion mathematically. A controversial policy for controlling the spread of smallpox advocated inoculating healthy people with an extract derived from the disease's victims. The mortality rate from inoculation was about 1 percent, but those who survived emerged (after a relatively mild case) with lifelong immunity to smallpox. This was considerably better than the 20-30 percent mortality rate from ordinary smallpox, but it was feared that inoculation might ignite too many outbreaks and cause the death of many people who would not have contracted smallpox naturally.

Was the proposed cure worse than the disease? The answer could not be divined by intuition. Bernoulli evaluated the idea quantitatively by developing a mathematical model, using data from mortality tables to estimate its parameters. From a differential equation solution, he concluded that widespread inoculation would increase life expectancy by three years. (A new type of inoculation soon made his analysis moot, however.) Thus the macroscopic approaches of Snow and Bernoulli proved their value even before bacteria and viruses were found to be the cause of disease, late in the 19th century.

Defining terms

Birth rate
: the rate at which a virus attempts to replicate from one machine to another

Computer virus
: a program or piece of a program that, when executed, "infects" another part of a computing system by making a copy of itself. Most PC-DOS viruses infect boot records of disks, or executable programs.

Death rate
: the rate at which a virus is eliminated from infected machines, usually when the user discovers it and cleans it up

Epidemic
: the widespread occurrence of a disease. A disease need not overwhelm a population to be epidemic; it must simply spread through some fraction of it.

Epidemic threshold
: the relationship between the viral birth and death rates at which a disease will take off and become widespread. Above this threshold, the disease becomes a persistent, recurring infection in the population. Below it, the disease dies out.

Epidemiology
: the branch of science that studies the spread of diseases.

Incident rate
: the rate at which virus incidents occur in a given population per unit time, normalized to the number of machines (computers) in the population.

Infected machine
: a computer that contains a virus, and can spread that virus to diskettes or other computers.

Prevalence
: the degree to which a virus is widespread in a population.

Topology
: in epidemiology, the patterns of contact along which diseases spread between individuals in a population.

Virus Incident
: the infection of a number of machines within an organization by a particular virus, due to a single initial infection from outside the organization.

Not until the 1930's, with the advent of electron microscopy and X-ray crystallography, was a start made in elucidating the structure of biological viruses. Their life cycles and biochemistry have been studied intensively since the mid-1940's and have been used in tandem with epidemiology to prevent diseases. The greatest victory of this collaboration was the eradication of smallpox in 1977. Rather than attempt to immunize the entire world's population the World Health Organization collected information about outbreaks and saw to it that only those likely to be in contact with an infected individual were immunized.

Today, the last specimen of the once-mighty virus -- which mothered the invention of inoculation in10th century China and Bernoulli's invention on mathematical epidemiology -- is serving a life sentence in a maximum security facility at the Centers for Disease Control and Prevention in Atlanta, Ga.

For computer viruses, the microscopic view came first, in part because their detailed function and structure is much easier to comprehend than those of biological microorganisms. The world of bits and bytes is, after all, man-made. Computer scientists have no need of sophisticated and expensive tools like electron microscopes, gene sequencers, and graduate students to explore the inner workings of computer viruses. They are happy with a disassembler, a quiet room, and a few minutes or hours of staring at the virus program logic.

The synergy between the micro and macro views in biology has generated many of the 20th century's most important medical advances. The hope is that a science of computer virus epidemiology will benefit from the same synergy.

Constructing a theory

Many, including Cohen and W.H. Murray, a well-known expert in computer security, have suggested applying theories of the spread of disease to computer viruses as well.

One of the simplifications worth borrowing from the biologists is to regard individuals within a population -- in this case, computers and associated hard disks, diskettes, and other storage media -- as being in one of a few discrete states, such as "susceptible" or "infected." Details of the disease within the individual are ignored (for instance, which executable files are infected within the computer).

In epidemiological language, pairs of individuals have "adequate contact" with each other whenever one would have transmitted a disease tot he other if the first had been infected and the second had been susceptible. What constitutes adequate contact can vary quite considerably from one computer virus (or biological disease) to another. The birth rate of a virus is the frequency with which adequate contact occurs.

Offsetting the birth rate is the death rate of the virus -- the frequency with which an individual is cured of the infection. Depending on the disease in question, an individual may become immune to it after infection or, at the other extreme, may become susceptible to it again immediately.

The birth rate of a computer virus is influenced by anything that hinders or promotes its replication, including intrinsic 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 on a diskette or preventive anti-virus software. The virus's death rate is influenced by intrinsic characteristics that might disguise or reveal its presence, by user awareness and vigilance, and by its detection and subsequent removal.

Just as an epidemiological model could be formulated by Bernoulli long before anyone knew the cause of smallpox, so the computer virus model is independent of what determine these rates. The only need is to be able to estimate the rates from empirical data.

A universal feature of the macro models is that, regardless of their specifications, they behave very differently on either side of a sharp threshold (the point at which an epidemic either takes hold or fades away). In the most common models, a virus can spread appreciable among the population only if its birth rate exceeds its death rate.

Such a situation can be shown in a simple simulation of a population of 100 machines (Fig. 1, main graph). Here, an infected machine can infect any other machine directly, and the virus's birth rate is five times its death rate. Exposure creates no immunity; machines may be reinfected immediately after they are cured.

Initially, just one machine is infected. As the simulation begins to run, the number of those with the virus grows exponentially, but levels off when about 80 percent are infected. In this of equilibrium, four out of five adequate contacts product no new infection, because the victim is already infected. So, once this level is reached, the rate at which new infections occur exactly balances the death rate. The equilibrium level depends upon the ratio of the birth and death rates, and can taken on any value between zero and 100 percent.

In some simulation runs, the virus is unlucky and dies out before it spreads very far. This happens when the virus is found and removed before it has reached more than a few machines. The extinction probability is equal to the death rate divided by the birth rate -- meaning 20 percent in this case.

On the other side of the threshold, where the death rate exceeds the birth rate, the virus will be driven to extinction unless some other reservoir of infection periodically reinjects the disease into the population. But this will at worst result in small, short-lived outbreaks with an average size independent of the population size. The inset in Fig. 1 shows such a situation, where all the parameters are as before, except that the virus birth rate is only 90 percent of the death rate.

Real world correction

The powerful concept of an epidemic threshold was discovered by mathematicians early in this century. A few years ago, an analysis of simple

models suggested that the same concept should also apply to the spread of computer viruses. And indeed, the existence of an epidemic threshold is strongly supported by statistics of thousands of virus incidents over the last five years in the large sample population tracked by IBM. This unique database is complied continuously, as infections occur, from hundreds of thousands of DOS personal computers; the sample is international but U.S.-biased, and is typical of virus-conscious Fortune 500 companies.

In cooperation with other virus collectors around the world, IBM's High Integrity Computing Laboratory maintains a collection of PC-DOS viruses, currently including more than 1500 different specimens. Less than 15 percent of them have been observed in the large sample population, and these have rarely appeared more than once.

The 10 most frequently observed viruses in 1992 accounted for two-thirds of all incidents [Fig. 2.] The top two -- Stoned and Form -- accounted for about one-third of the total.

In some cases, computer viruses hardly spread at all, because they are below the epidemic threshold. This concept of epidemic threshold is perhaps the first good news that has been derived from theoretical studies of computer viruses.

An early theoretical result, derived by Cohen, was distinctly depressing. He found that an algorithm capable of distinguishing perfectly between viral and nonviral programs is a logical impossibility. Fortunately, his elegant proof [see "No virus detector is perfect,"] has not halted the development of reasonably good software protection against today's computer viruses.

More good news

For the unattainable goal of perfect detection, the threshold theory substitutes the achievable goal of pushing viruses below the epidemic threshold. It is encouraging that this has already been achieved for many viruses.

Once a virus has fallen below the epidemic threshold, further effort offers diminishing returns (although it does reduce the size of any outbreaks due to reintroduction of the virus from another reservoir, such as infected diskettes in a forgotten desk drawer).

But Cohen's proof cannot be dismissed so easily. Even as workers wipe out one virus, others are being written, some of which are likely to be above the epidemic threshold until anti-virus software is modified to deal with them.

Common defenses

The anti-virus technologies differ in their effects on viral birth and death rates. The virus scanner, the most common, works by examining stored programs for infection with one of a set of known viruses. Scanners often also detect slight variants of know viruses. Some even incorporate a heuristic function, which allows them to detect some brand-new viruses by guessing at the function of the code.

Scanners excel at raising the virus death rate. Someone who installs a scanner and uses it at regular intervals -- say, once a week -- increases the death rate from nearly zero to at least, in this case, once a week. A resident scanner, which is always active in the system examining all programs that are loaded, pushes the death rate even higher, since the virus is detected (an presumably removed) as soon as it is loaded.

Scanners can also act as filters to decrease the viral birth rate. For example, if all new programs arriving at a machine are scanned as they arrive, the effective

birth rate of the machine's neighbors will be lower.

Traditional access control systems are a second kind of anti-virus technology. By preventing unauthorized programs from altering other programs, they can decrease the viral birth rate. Networked systems lacking access control could be swamped by a virus within an hour or two. In his early experiments, however, Cohen showed that even when access controls are in place, viruses can spread quickly and widely without violating those controls.

A third anti-virus technology, sometimes called integrity management, lies somewhere between scanners and access control systems. Its strategy is to detect and prevent virus spread by noticing or preventing the changes viruses make to parts of the computer system. An integrity management system can increase the viral death rate if it notices an anomaly due to a virus and alerts the user. Conversely, the system may note but not warn the user of a change (perhaps because the virus has noticed the protection and has not tried to spread), in which case it is limiting the birth rate.

While scanners work best against known viruses, integrity management systems can guard against larger classes of viruses. Because they look for general methods that viruses use to spread, not for the bit patterns that make up the virus code, they can be more effective on newly devised infectors. Their disadvantage is that they also flag or prevent legitimate activity, and so can disrupt normal work or lead the user to ignore their warnings altogether.

Individual and community

By using both preventative and curative anti-virus technology in tandem, an individual protects his own machine more effectively than by using either technique alone. (Simple mathematical models suggest that the synergy is much stronger than one might guess.) An individual who protects a system out of self-interest also benefits the community by reducing the chance of a virus spreading to other systems.

The importance of lowering a virus's birth rate and raising its death rate is well understood by public health officials. Healthy children are inoculated against measles and tuberculosis patients take medication in part to protect everyone they come in contact with. If the steps taken by individuals put society as a whole below the epidemic threshold, then even the unprotected are unlikely to become infected.

Of all the assumptions woven into the fabric of biological epidemiology, the least applicable to populations of computers is "homogeneous mixing"; the supposition is that every individual in the population is equally likely to infect or be infected by every other individual. But most individuals exchange most of their software with just a few others and never contact the majority of the world's population. Also, their exchanges tend to be localized: if Alice swaps software often with Bob and Carol, chances are that Bob and Carol swap software, too.

What is missing from standard epidemiological theories is this notion of topology -- the pattern of interaction between individuals within a population.

Topological effects are incorporated into epidemiological models by representing individuals as nodes and their contacts as lines connecting the nodes. Each line can be characterized with its own viral birth rate, and each node with its own death rate.

Taking topology into account (and it could be useful in biological epidemiology as well) radically changes the picture of how viruses can spread. An enhanced frame from one of millions of simulations that have been conducted [Fig. 3] shows 250 individual systems out of a total of 10 000 in the population. In this example,

the individuals tend to form hierarchically nested groups, with the members of one group exchanging software frequently among themselves, less often outside their department, and even less often outside their organization. The topology in this example is also sparse; each individual is in contact with only a few others.

First, consider the effect of sparsity. Suppose each individual has potentially infectious contact with 52 randomly chosen neighbors. They interact randomly on average once a year with each neighbor. Now imagine a sparser topology in which each individual has just one neighbor, contacted on average once a week.

In both cases, the overall birth rate is once per week. Analysis and simulation show that in the first scenario the epidemic threshold occurs when the death rate equals the birth rate-once per week. (This is precisely what the homogeneous mixing approximation would predict.) In the sparser topology of the second scenario, analysis and simulation show that an epidemic can occur only if the death rate drops below three per week. In general, bottlenecks hamper viral spread in sparse topologies, to the extent of preventing it entirely or making it less pervasive than in dense topologies.

Localization has a different effect. Viruses spread more slowly in localized than in randomized or homogeneously mixed environments, in which growth is at first exponential [Fig. 4]. In another type of logical topology -- the two-dimensional lattice -- the virus's growth is at first merely quadratic. Strongly sub-exponential spread rates occur in local topologies because the virus is forced to circulate in areas that it already occupies, limiting its access to healthy populations.

In some local topologies, such as those in Fig. 4, the equilibrium is essentially the same as in a homogeneous system. In others, it is much lower, and in yet others, the number of infections fluctuates wildly, never seeming to reach any sort of equilibrium.

Case study

Virus incident statistics collected from the sample population are very revealing about how quickly viruses are spreading in the real world and how prevalent they have become. It so happens that the number of infected PCs in the world is roughly proportional to the number of incidents observed in the sample population.

What is the proportionality constant? It can be estimated for North American business sites with 100 or more PCs. This sector of the world was studied by the 1991 Virus Prevalence Survey conducted by Dataquest Inc., the market research firm in San Jose, Calif.

Two conclusions may be drawn by re-interpreting the raw data, which was supplied by Peter Tippett of Certus, a division of Symantec, namely, that the incident rate was about the same as in the sample population, and that an average of about three or four were infected in the course of an incident. Thus a rough estimate of the rate at which a given virus infects machines in this type of environment can be obtained by multiplying the measured incident rates by a factor of three or four.

Figure 5 shows the observed incident rates as a function of time for some of the most common viruses. During 1990 and 1991, the Stoned, 1813 (also known as Jerusalem), and Joshi viruses proliferated at a far less than exponential rate (perhaps roughly linearly) for a year or two. Then they leveled off at a few incidents per 10 000 PCs per quarter.

The Form virus began slowly, but in the third quarter of 1991, it began to take off as strongly as the Stoned virus had in early 1990. It is rumored that the Form invaded the master diskette used by a software distributor. The diskette duplicator

may have saved the virus the trouble of copying itself, injecting maybe thousands of infected diskettes into the world. This blunder could easily have given Form the impetus to surpass Stoned last summer, when it became the world's most prevalent virus.

None of the viruses in the IBM sample population is multiplying at anything like an exponential rate, supporting the intuitive notion that software exchange is highly localized. This is good news. One widely publicized theory of computer virus replication implicitly assumed homogeneous mixing, predicting exponential growth for all viruses. In 1991, its author estimated that the 1813 (Jerusalem) virus had an exponential doubling time of, at most, 2.6 months, and a population of at least 500 as of October 1989. Extrapolating from this, by April 1993 over 25 million PC -- one fourth of the world's total -- would be infected by the 1813 virus!

Because of the nature of the virus, that level of infection would have crippled the PC-DOS world because a bug in the 1813 virus causes it to reinfect some programs.

Acquiring 1813 additional bytes every tin they become infected, these programs eventually overflow the conventional 640 kilobytes provided by DOS, and can no longer run. Fortunately, the prediction runaway infection was wildly inaccurate. In fact, Fig. 6 suggests this virus is on the decline, and that the war of 1813 is turning in the community's favor.

That even the most common of viruses are not very common is also good news for users. Just before the Michelangelo scare in early 1992, the Stoned incident rate appeared to be leveling off at 0.2 incident per 1000 PCs per quarter. Assuming that each virus incident affects three or four machines, Stoned was infecting under 0.8 of every 1000 PC-DOS machines per quarter in business environments. To obtain the fraction of machines infected with Stone at a given moment in time; this 0.8 should be multiplied by the average duration of the infection. Assume (generously) that this is half a year. Then far fewer than 1.6 in 1000 of these machines would have fallen prey to Stoned at any given moment during late 1991. In certain other environments, such as universities, the average incident size and duration may be much larger; the percentage of infected machines would be correspondingly higher.

As pleasing as it is to users and the anti-virus community, the low level at which viruses plateau is puzzling. The simplest models yield as low an equilibrium only when the birth rate barely exceeds the death rate. Can every single one of the most successful viruses be so delicately balanced on the edge of becoming epidemic? What mysterious force is preventing rampant plagues? Something important is surely missing from the simple model.

Perhaps it's the human element. When a person encounters a computer virus, or hears a colleague has fallen victim to one, he or she becomes more vigilant (most probably through the purchase of antivirus software). Unlike exposure to biological diseases, contact with one computer virus can actually confer immunity on all the computer viruses the anti-virus software can handle. Next, victims sometimes tell their friends, who then rush to check their own machines for infection. New models incorporating this effect show that word of mouth is surprisingly powerful even when not used to its full extent. A similar principle can engender successful anti-virus policies, as shall be seen shortly.

Steps for companies

The new understanding of viral epidemiology can enable effective antiviral policies for companies. From a company's perspective, its machines are under constant attack from the outside world [Fig. 6 again]. Sooner or later, a computer virus will find its way inside. The rate at which this occurs depends on the number

of infections in the world, the number of company machines, the frequency of software exchange between the company and the outside world, and the effectiveness of the company's precautionary measures.

The invasion marks the beginning of a virus incident. The virus may then spread to several more computers within the company before it is discovered. Once all the machines it has infected are found and cleaned up, the incident is over. The total number of machines infected is referred to as the size of the incident.

An organization should have two complementary goals: to reduce the influx of viruses and to stop those that get in from spreading. Epidemiology has much to say about how the second objective should be achieved.

The best approach a company can take today is to encourage users to inform a central agency about their machines' infection, and to have the central agency respond by helping those users clean up their machines and then check neighboring machines for infection. The sample population tracked by IBM has been doing this for several years. (It is a shame that some organizations do exactly the opposite: punishing employees whose machines are found to be infected!)

The effect of this policy can be dramatic. Suppose that the company has no organized anti-virus policy. It is likely that Form, Stoned, and other viruses that are above the epidemic threshold in the rest of the world will also be above the threshold inside the company. Once such a virus gets in, it has the potential to ignite a pervasive, persistent infection similar to the above-threshold case of Fig. 1. Even companies that have distributed anti-virus software widely within their organizations could be at risk.

With effective central reporting and response -- where the entire incident is cleaned up as soon as any machine is found to be infected -- the situation is similar to the below threshold case in Fig. 1. Mathematical analysis shows that this occurs even if the virus's birth rate exceeds its death rate. There are two desirable consequences. First, the average incident size remains quite small, no longer scaling with the size of the organization. Second, the incident is finite in duration, rather than infinite.

Experience confirms the value of central reporting and response, coupled with the dissemination of anti-virus software. While these policies were being instituted within the sample population, the average incident size was 3.4 PCs. More than five PCs were involved in 12 percent of the incidents, but these large incidents were responsible for 60 percent of all machine infections [Fig. 7].

Once the policies were firmly in place, the percentage of large incidents fell. In 1992, the average incident size was less than 1.6 PCs. Only 2.5 percent of the incidents were large, and they accounted for just 27 percent of all machine infections [Fig. 8].

Even within the sample population, infections sometimes persist in an organization. When the number of reports of a particular virus in a particular location is well above the statistical average, the virus may be spreading internally, rather than penetrating repeatedly from the external world.

Resources can then be marshaled to help that organization quell the incident. Accurate statistics thus help a company to focus its anti-virus resources on the areas that need them most, keeping costs down. They also help a company to monitor its own progress in reducing the influx of viruses (through the measured incident rate) and in limiting internal spread (through the measured average incident size).

As a final, powerful argument for central reporting and response, recall the

World Health Organization's sensational victory over smallpox, in which an analogous strategy of well-targeted immunization played a key role.

Prospects

The new science of computer virus epidemiology has already yielded a much better understanding of viral spread, the factors governing it, and how to control it. In order to make theories more quantitative and predictive, ways must be found of characterizing the world's software exchange. From user surveys and automatic monitoring techniques, we hope to learn enough about individual behavior to further influence virus trends occurring within organizations and indeed throughout the world.

Complete eradication of all viruses is impossible as long as there are malicious programmers. Combining microscopic and macroscopic solutions, however, holds out the hope of reducing the problem to the nuisance level.

To probe further

- The microscopic view of biological viruses can be found in Viruses, by Arnold J. Levine (Scientific American Library, W. H. Freeman, New York, 1992). The macroscopic view of diseases is described in Plagues and Peoples, by William H. McNeill (Doubleday, New York, 1977). The classic textbook on mathematical epidemiology is Norman T.J. Bailey's The Mathematical Theory of Infectious Diseases, now in its second edition (Oxford University Press, New York, 1987). An illustrated account of modern-day practical epidemiology is given in the article "The Disease Detectives,' by Peter Jaret, National Geographic, January 1991, pp. 114-140.
- Computers & Security, >Vol. 6 (February 1987) contains the article by Fred Cohen that defined computer viruses as they are known today: "Computer Viruses: Theory and Experiments," pp. 22-35.
- IEEE Spectrum previously examined data security in general, and the microscopic view of viruses in particular, in a multi-part special report in August 1992.
- More detailed versions of the authors macroscopic virus studies can be found in the Proceedings of the 1991 IEEE Computer Society Symposium on Security and Privacy, Oakland, Calif., May 20-22, 1991; in the Proceedings of the Fourth and Fifth Annual Computer Virus and Security Conferences, New York City, March 14-15, 1991, and March 12-13,1992; and in the Proceedings of the Second International Virus Bulletin Conference, Edinburgh, Scotland, Sept. 2-3, 1992.