Applying Epidemiology in Computer Virus Prevention: Prospects and Limitations

Weiguo Jin
Department of Computer Science
University of Auckland
wjin003@ec.auckland.ac.nz

Abstract

The impact of a widespread computer virus infection can be profound. Using a biological analogy, computer virus epidemiology develops a theoretical model to predict the rate and extent of propagation of a computer virus infection. This paper describes the prevalence of computer viruses, and surveys several related research works in this field. Furthermore, we describe the significance of applying epidemiological models in computer virus protection and prevention, and discuss their implication in developing anti-virus technologies and policies. We conclude that, despite its limitations, the epidemiological models bring new hope for solving the computer virus problems.

1. Introduction

Computer virus[1] is a piece of software containing malicious code that can install, propagate and cause damage to computer files and data without the knowledge and/or express permission of the user [11]. Computer virus earns its reputation by the ability to attach itself to other programs or make copies of itself, and the ability to cause permanent loss of data and hardware on the host machine [3, 7, 11].

Computer viruses used to be the spotlight in the public media. Though, the impact of and damages caused by computer viruses have been reduced considerably due to the increasing public conscious and the advance of technologies. Nevertheless, it does not mean that we can assume that the problems of computer viruses have been solved. Firstly, the cost to recover from the damage of computer viruses can be potential huge. According to NewsFactor Network [10], in 2001, Code Red and its variants "had an estimated worldwide economic impact of US$2.62 billion". On the other hand, the

[1] There are several terms related to computer virus, such as worms and Trojan horse. In this paper, the term "virus" is used to refer to them as whole without the intention to tell the differences among them.

computer viruses may take advantage of new technologies, i.e. the fast growing Internet, to reach an even broader range of individual systems at a faster spread rate.

Epidemiology is a biological term. It is "the study of the distribution and determinants of health-related states or events in specified populations and the application of this study to the control of health problems" [2]. The biological analogy between the biological virus and its counterpart in computer science, as indicated by its name, implies that "the mathematical techniques which have been developed for the study of the spread of infectious diseases might be adapted to the study of the spread of computer viruses" [7].

Traditional anti-virus approaches are passive and fails to analyze the propagation of computer viruses at a theoretical level. It is necessary to adapt properly augmented mathematical models, which capture the spreading characteristics of computer viruses, to analyze quantitatively the propagation of computer viruses. The consequent benefits include [7]:

- Aid in the evaluation and development of general policies and heuristics for inhibiting the spread of viruses.
- Aid in predicting the course of a particular epidemic and plan resources to deal with the problem.

The rest of the paper is structured as follows. In the next section, we give an overview of previous works on computer virus study and describe its prevalence. Section 3 introduces several epidemiological models that are used to analyze the spreading of computer viruses, and presents theoretical results. We discuss the limitations and implications of the epidemiological models in section 4. In the last section, we draw the conclusion and discuss the future work in this field.

2. Computer Viruses and Their Spread

"Since the first documented reports of microcomputer viruses in the mid-1980's, they have spread throughout the world" [7]. In this section, we give an overview of computer virus, in particular, how computer virus works, the prevalence and current works in computer virus defenses.

2.1. How Computer Virus Works

They can enter the computer system through two main entry points: the disk drives (floppy drive, CD, etc) and the network adapter cards (network or modem card) [11]. There are several properties that distinguish computer virus from other program code [3, 7, 11, 5].

Firstly, it can replicate by either attaching itself to other executable programs or making copies of itself. The infected programs become viruses and spread to other systems. A typical viral infection is explained in the following pseudo code:

```
loop:
   find a random executable file;
   if contains virus signature
         go back to loop;
   else
         add a copy of itself on top of the uninfected file;
```

Secondly, it can be activated under certain circumstances and make malicious actions to the host computer systems. Many viruses contain a destructive sequence that is called payload. It may be triggered by the arrival of a particular date or an action done by the user. The effect of the payload may be disastrous. It can cause permanent loss of data and hardware.

2.2. The Prevalence of Computer Viruses

In order to develop the efficient and effective anti-virus policies and heuristics, we should not only understand how virus works, but also the factors that affect its prevalence, such as the new technologies and new transmission media. It is also necessary to be aware of the different forms computer virus may take.

Explosive development of the Internet technology enables more frequent and easy access to various resources from every corner of the world. Despite those obvious benefits, it also places a great security challenge on top of the computer systems of different governments and organizations. It is assumed that the spread of computer viruses will follow a similar pattern as of the exponential growth of the Internet. Figure 1 [8] shows that the number of different PC-DOS viruses has been growing at a roughly exponential rate from 1988 up to 1993. However, we have not found latest statistics on the number of new viruses. There are not enough evidences that indicate such direct contributions of Internet to the wide spread of computer viruses.

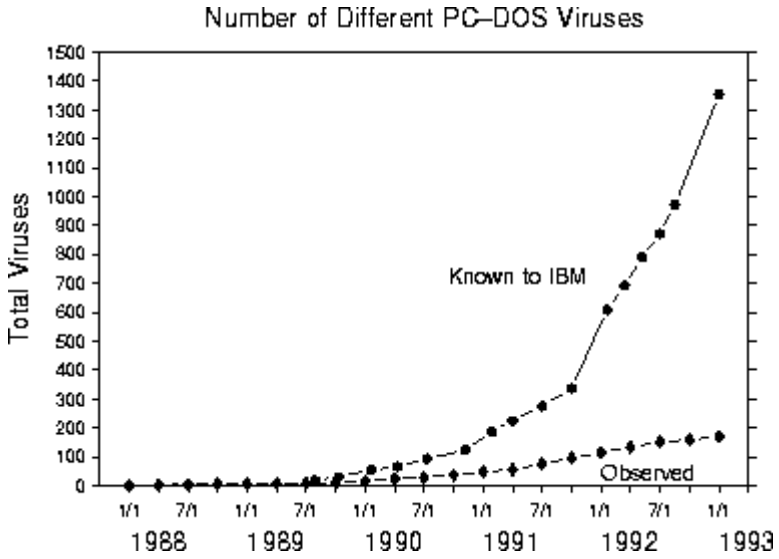

Figure 1. Number of viruses known to IBM and number of viruses observed as a function of time

Next figure, published by LCSA Labs in 6th Annual Computer Virus Prevalence Survey 2000 [1], shows the monthly rate of infection per 1,000 PCs for the first two months of 1996 through 2000. As we can see from Figure 2, there is an increased infection rate of about 10 infections per 1,000 machines per month each year from 1996 to 1998 and from 1999 to 2000. "In 1999, there was a surge in the encounter rate. This result was no doubt the result of the 'mass mail' payload of macro viruses, Internet worms, and the scripting viruses that followed."

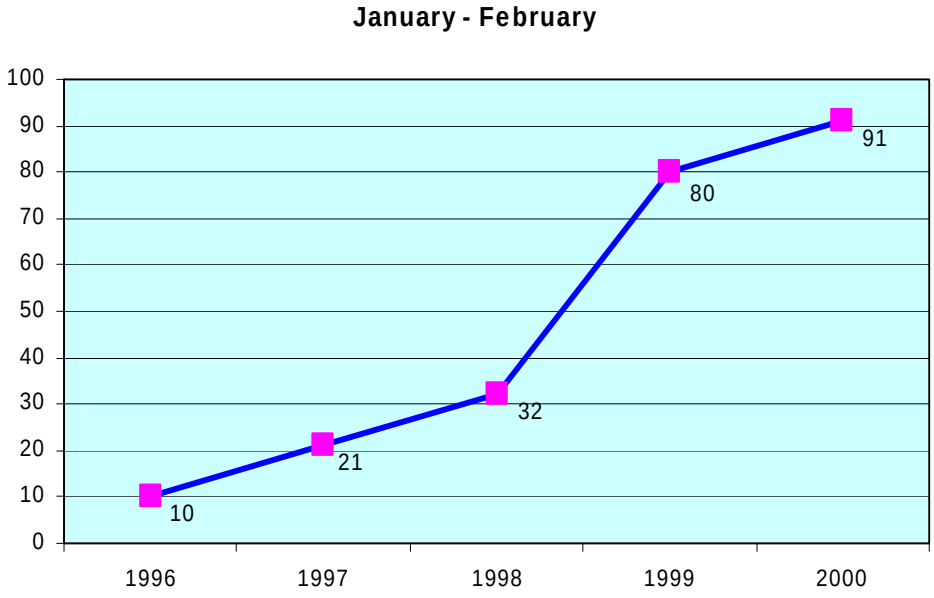

Figure 2. Infection Rates Per 1,000 PCs Per Month – January and February

Computer viruses take various types ranging from original simple DOS file and boot sector viruses to Excel and Word macro viruses [12]. The new-age Internet-enabled scripting viruses, i.e. JavaScript or VBScript viruses, are the latest comers in this field. This is somehow the reflection of the progress of the technologies. The traditional boot sector viruses or file type viruses are effectively disappeared from the scene [1].

2.3. Traditional Computer Virus Defenses

Most of the traditional anti-virus techniques are reactive approaches that rely on finding a particular virus before being able to deal with it well [12]. The most common technique is to use virus scanner, resident or not, to examine files, emails, memory and disk boot sectors for known virus signatures. Some scanners can also detect slight variants of known viruses. Some even incorporate a heuristic function, which allows them to detect some brand-new viruses by guessing at the function of the code. Access control system is another anti-virus technology that works by preventing unauthorized programs from altering other programs. Without access control, a system becomes extremely vulnerable, especially for the networked systems. Some systems use integrity management 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 alert other users when it notices an anomaly due to a virus [9, 4].

In reality, these traditional anti-virus technologies work well for known viruses. But they require frequent updates to deal with new viruses, and may disrupt or prevent legitimate activity to a certain degree.

In his pioneering research work in computer viruses, Dr. Cohen [3] pointed out that there is no perfectly secure against viral attacks. However, the study of epidemiology in computer virus brings a new hope that an imperfect anti-virus technique can be used to prevent the propagation of computer viruses under some conditions. We introduce epidemiology in the next section.

3. Epidemiology in Computer Virus Prevention

The cost caused by the damage of computer viruses can be potentially huge. Various approaches have been proposed to address the computer virus problem theoretically. Adapting and applying mathematical epidemiology to this problem is one such

attempt. The hope is that a science of computer virus epidemiology will benefit from the success of epidemiology in biology [9].

3.1. Basics of Epidemiology

Epidemiology tries to reveal under what conditions will an epidemic happen, and the relationship between the infection rate and the number of infected computer systems along the time scale. There are many factors considered when estimating the probability of virus infections, such as the ability of the virus to replicate, the amount of contact any given machine has with the general population of computers, and the presence of any computers currently infected [6].

In this part, we list definitions of several terminologies that are used to describe a viral epidemic [12, 9, 7]:

- Birth rate – the rate at which a virus attempts to replicate from one machine to another (also known as infection rate). It is denoted as β.
- Death rate – the rate at which a virus is eliminated from infected machines, usually when the user discovers it and cleans it up (also known as cure rate). It is denoted as δ.
- Topology – the patterns of contact along which diseases spread between individuals in a population.
- 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. It is denoted as ρ.

3.2. Modeling of Computer Virus

The models of computer virus epidemiology provide a reasonable qualitative understanding of the conditions under which viruses spread and why some viruses spread better than others [12]. In 1991, Jeffrey Kephardt and Steve White published "Directed-Graph Epidemiological Models of Computer Viruses," [7] the first paper that adapt the mathematical epidemiology to the new problem of computer viruses. In this paper, they proposed an epidemiological model, SIS (susceptible → infected → susceptible) model, on directed graph.

Assumptions. A single computing system is treated as an individual in a population of similar individuals. The details of infection within an individual are ignored. An individual can only have a small number of discrete states, i.e. infected, susceptible or immune. The infection rate is simplified as the probability per unit time that an infected individual will infect an uninfected individual. Similarly, the cure rate is the probability per unit time for an infected individual to be cured [7].

Topologies. The authors further incorporated the topological effect into the mathematical epidemiology. Figure 3 [7] shows a random fully-connected graph with 10 nodes of a homogeneous system. An individual system is represented as a node. The directed edge represents the number of nodes that can be infected. Black filled and unfilled circles represent infected and uninfected nodes, respectively. Each edge is associated with an infection rate, while each node has a cure rate. It is assumed that each node in the graph is equally likely to infect or to be infected by every other node. However, the homogeneous mixing assumption fails when "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" [9]. This observation led to other two topological structures, sparse and localized systems. "It is said to be sparse because each individual has adequate contacts 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 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" [8]. The authors modified the previous random graph model by adding a "weak" link to represent the sparse systems. Two other models, hierarchical and spatial models were used to represent the localized systems. We omitted the details here. Interested readers can refer to [7, 8].

3.3. Theoretical Results

Several different techniques were used to study the behavior of the SIS model. For example, the deterministic approximation, probabilistic analysis and simulation are used to analyze the homogenous model.

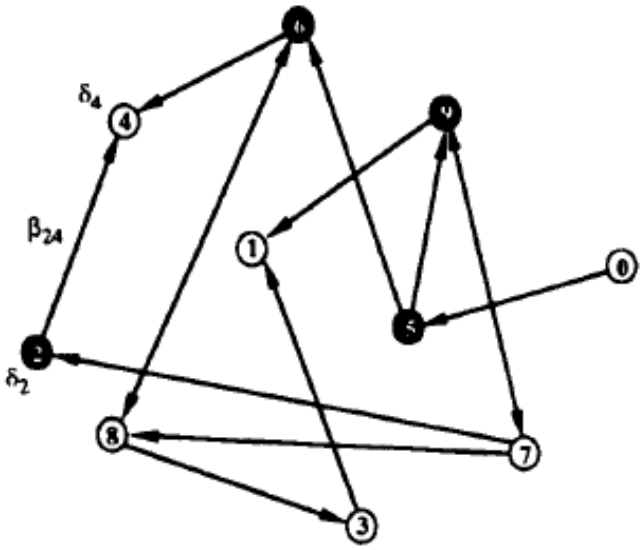

Figure 3. SIS Model on Directed Graph

Figure 4 [7] displays the comparison of the expected number of infected nodes as a function if time in the deterministic and stochastic models. The total population is 100 nodes. The average infection rate $\beta = 1.0$, and the cure rate is $\delta = 0.2$. The black curve is the deterministic average, and the white curve is the stochastic average. Gray area represents one standard deviation about the stochastic average [7].

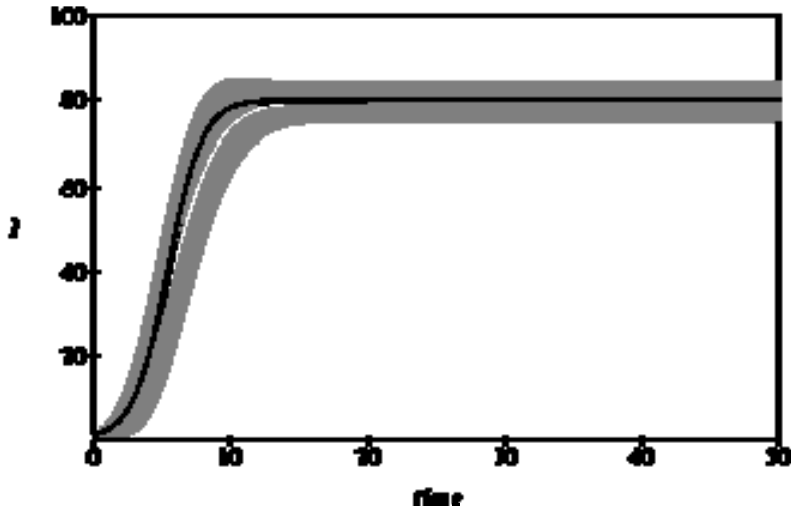

Figure 4. Number of Infected Nodes in the Deterministic and Stochastic Models

The purpose of simulation is to test the accuracy of previous assumptions. A straightforward event-driven implementation is used to simulate the model. Figure 5 [7] compares a typical simulation run on a 100-node graph to the corresponding

deterministic solution with similar parameters used in Figure 4. “The simulation run follows the deterministic solution reasonably well, except that the equilibrium appears to be lower” [7].

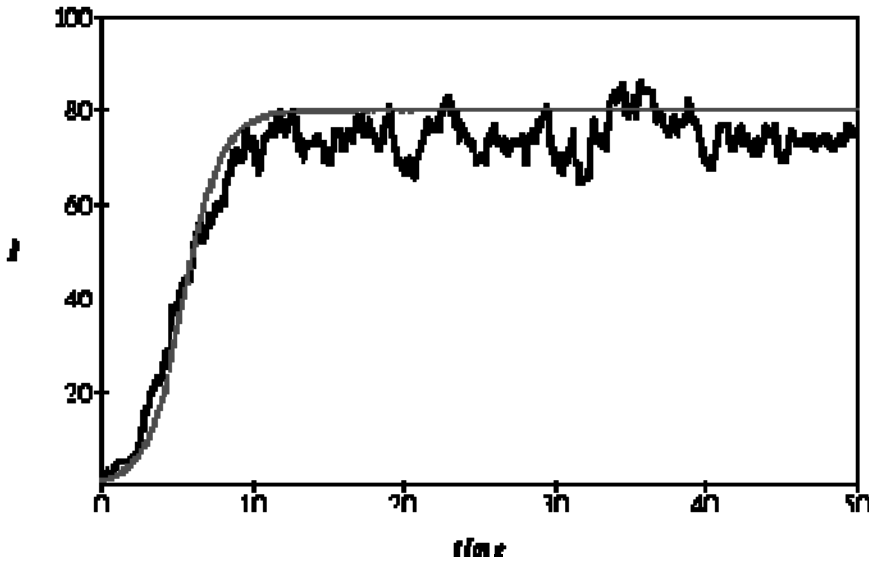

Figure 5. Simulation Govern by Deterministic Theory

By analyzing and simulating viral spread on a variety of topological structures, the following theoretical results have been reached [8]:

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 $\rho_{threshold}$ is diminished to some value less than 1. As the average degree of nodes in the graph diminishes, so does $\rho_{threshold}$, and the probability of an epidemic diminishes (dropping to zero if $\rho_{threshold}$ 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.

4. Implications and Limitations of Epidemiological Models

The concept of epidemic threshold is perhaps the first good news that has been derived from theoretical studies of computer viruses. “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” [8]. We notice that there are thousands of different computer viruses around, but only a few of them have been observed in the viral incidents. “The 10 most frequently observed viruses in 1992 accounted for two-thirds of all incidents. The top two – Stoned and Form – accounted for about one-third of the total” [9]. Most viruses do not spread at all or their spread could not be established firmly, because they are below the epidemic threshold.

4.1. Implications

Epidemiological models show that, by reducing the birth rate and increasing the virus death rate sufficiently, one can push viruses below the epidemic threshold. “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” [9].

Even though we could not have a perfect technology again viral attacks, we still can utilize the available traditional anti-virus technologies (as mentioned in section 2.3) to achieve the better effects. 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, like a resident scanner. Scanners can also act as filters to decrease the viral birth rate [9].

Another extremely effective way to manage the virus problem in organizations is to establish the centralized reporting and response mechanisms. The following policies are recommended to all organizations [8]:

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

4.2. Limitations

Since current epidemiological models are somewhat simplistic. There are still plenty of rooms for improvement. We address a number of limitations in this part [7, 12]:

- In SIS model, an infected node is recovered from viral attack, and becomes susceptible again. This is contradictory to the actuality. We know that individuals will become permanently immune or at least to some degree once they recovered from the infection in biology. It should apply to computer system as well.
- In current models, all of the systems are assumed to have the same birth and death rates. In the real world, this assumption will not hold true. How can we accommodate these variants into the epidemiological model?
- As shown in Figure 4 and 5, when infection rate is greater than cure rate, the number of infections increases exponentially at initial stage, and eventually saturates at equilibrium. However, statistics show that few viruses can stay at such a high infection rate. They will be wiped out of the map finally.

5. Conclusion

It is very hard to measure the impact of computer viruses. It includes not only “hard cost”, e.g. employee hours, but also “soft cost”, such as the denial of services, lost staff productivity and cost to reputation [10]. As pointed out in [1], “the virus risk continues to get worse despite corporate efforts. … Companies are experiencing more and more virus incidents which result in higher and higher virus incident costs each year”.

Current anti-virus software is reactive, not efficient and only works well with the known viruses. However, as one of the efforts to solve the problem of virus, adapting and applying mathematical epidemiology in computer virus enables the qualitative analysis of the conditions under which the viruses spread, and the relationship

between the infection rate, the number of infected computer systems along the time scale. Epidemiological models reveal that there is a well-defined epidemic threshold. "An imperfect defense against computer viruses can still be highly effective in preventing their widespread proliferation, provided that the infection rate does not exceed" the threshold [7].

There are still a lot of works need to done towards a more complex and realistic epidemiological model that hopefully can solve these mysteries around the propagation of computer viruses.

References:

1. L. M. Bridwell and P. Tippett. "ICSA Labs 6th Annual Computer Virus Prevalence Survey 2000". 2000. http://www.trusecure.com/html/tspub/pdf/vps20001.pdf.
2. Central Brain Tumor Registry of the US (CBTRUS). "Glossary: Epidemiology". http://www.cbtrus.org/glossary/epidem.htm.
3. F. Cohen. "Computer Viruses, theory and experiments". In Proc. DOD/NBS 7th Conf on Computer Security, 1984.
4. Computer Help Desk, University of Victoria. "Virus Protection". http://helpdesk.uvic.ca/how-to/support/virus.html.
5. Definition of Computer Virus. http://www.upm.edu.ph/~enrico/notes/virus_definition.html.
6. S. Gordon. "Technologically Enabled Crime: Shifting Paradigms for the Year 2000". Computer and Security, 1994. (Published by Elsevier Press' Computers and Security 1995). Available: http://www.reseach.com/antivirus/SciPapers/Gordon/Crime.html.
7. J. O. Kephart and S. R. White. "Directed-Graph Epidemiological models of Computer Viruses". In Proceedings of the 1991 IEEE Computer Society Symposium on Research in Security and Privacy. Oakland, California. 343-359. May 20-22, 1991.
8. J. O. Kephart and S. R. White. "Measuring and Modeling Computer Virus Prevalence". Proceedings of the 1993 IEEE Computer Society Symposium on Research in Security and Privacy, Oakland, California, May 24-25, 1993; pp.2-15.
9. J. O. Kephart, S. R. White and D. M. Chess. "Computers and Epidemiology". IEEE Spectrum, 20-26. May 1993.
10. J. Lyman. "In Search of the World's Costliest Computer Virus". NewsFactor Network. 2002. http://www.newsfactor.com/perl/printer/16407/.
11. PageWise. "How Computer Viruses Work". http://www.allsands.com/Science/computervirusi_ol_gn.htm.

12. S. R. White. “Open Problems in Computer Virus Research”. Presented at Virus Bulletin Conference. Munich, Germany. October, 1998.
13. S. White, J. Kephart and D. Chess. “Computer Viruses: A Global Perspective”. http://www.research.ibm.com/antivirus/SciPapers/White/VB95/vb95.distrib.html.
14. Virus: a Retrospective. “The Prevalence of Computer Viruses”. http://cse.stanford.edu/classes/cs201/projects-00-01/viruses/social-prevalence.html.