EPIDEMIOLOGICAL MODELS OF COMPUTER VIRUSES

Submitted by
TAN WAI LIP VINCENT

Supervised by
DR. CHANG EE-CHIEN

In partial fulfilment of the requirements for the degree of
Bachelor of Science with Combined Honours in Computational Science and Mathematics

at
Department of Computational Science
National University *of* Singapore
2001/2002

Abstract

Aside from their obvious physical differences, biological and computer viruses possess similar properties and behaviour. This motivated us to apply the techniques used in biological epidemiology to studying the propagation of computer viruses. Most epidemiological models assume a homogeneous characteristic, which is usually not the case with computer viruses. Thus we incorporated topological factors (computer network connectivity) to a standard epidemiological model by using graphs. We conducted further investigations into the study of three other factors: an advance alarm system, a natural response system and a periodic activity system. Our computer simulations show that any factor (such as a larger number of initial infectives), that can affect the relation between an infected computer and a virus-free computer, will play a part in determining whether an epidemic occurs. However, only through topological factors, together with the advance alarm, natural response and periodic activity systems, will the resulting simulation behave in a similar manner as that of a real world computer virus epidemic.

Acknowledgement

I want to thank my supervisor for helping me understand certain concepts and for keeping me focused on the objectives. A word of gratitude to my family for their subtle but heartfelt encouragement.

Contents

List of Figures

Chapter 1

Introduction

Perhaps it is inevitable that computer viruses came into being. When we used computers to help with our repetitive tasks, automating them to do those tasks is the next logical step. At their heart, computer viruses are just computer programs, as are the desktop software used daily nowadays. What separates viruses from normal programs is that viruses encroach upon the user without permission. This general distinction also includes benign programs such as those of maintenance of computers [6]. There is no specific definition for a virus, but the general accepted idea is that a virus is a program that executes procedures without the user's permission (and usually unwanted as well) and that it also attempts to spread (or has the ability to spread) to other computers (or files). For our purposes, we are concerned with the self-replication property. This is actually the usual definition of a worm, but we shall refer all self-reproducing rogue programs as viruses.

From a programming or artificial intelligence point of view, one cannot ignore the fact that a few simple rules can characterise an artificial life form [14, 15] as complex as biological life forms. Cohen [6] commented on the possibility of living with viruses. Judging from the current trends, computer viruses are evolving faster than any life form on earth. Total extinction is a possibility, but until it is widely achievable, we will have to accept the compromise of coexisting with a low number of viruses.

In this chapter, we shall begin introducing some epidemiological concepts to provide a basic understanding of the theories behind mathematical models. Next in chapter two, we discuss past work on epidemiological models, whose origins come from biological research. Armed with this knowledge, we start formulating models based on graphs to include localisation properties in chapter three using random graphs. We introduce the tree graph in

chapter four to improve on the random graph. The advance alarm, natural response and periodic activity systems are studied closely in the chapter as well. Finally, we propose a new graph model as another representation of the Internet in chapter five. Comparisons between the new graph model, random graph and two-dimensional planar graphs (spatial model), as well as the testing of different combinations of the three systems will also be done.

1.1 Basic epidemiological principles

Before we can discuss about the mathematical models applied to viral epidemics, we have to introduce certain concepts in the study of viruses. These concepts had been identified during extensive medical research. They were given special significance in designing realistic mathematical models as they affect the models' theoretical formulation directly. We shall use the term "individual" to denote simple units capable of receiving or transmitting a virus, be they human beings or computers.

1.1.1 Infection

The notion of infection is intuitive, but how does one formulate it in a mathematical equation? Most models simplify the concept either by using a constant rate at which the number of infected individuals increases or decreases, or casting it as a probability to be a measure of how likely it is to catch an infection. This simplification, of course, does not take into account *how* the individuals come into contact with each other. However, it does make the problem more tractable, often rendering the model accessible to rigorous mathematical analysis.

1.1.2 Immunity

Closely related to (or perhaps more correctly, the opposite of) infection is the ability to resist the invasion of the virus. Mathematical models usually classify individuals as being in one of these distinct states: *susceptible* (still vulnerable to the virus), *infected* (caught and now spreading the virus) and *removed* (as in removed from the chance of contracting the virus again).

In the model proposed by Kermack and McKendrick (1927), the population of the immune depends on the infectives. This is true of biological diseases, where one has to survive the virus before immunity is gained, if at all. Certain diseases such as tuberculosis, meningitis and gonorrhoea do not confer immunity to reinfection. The individual under attack by the virus may already possess an innate biological defence system that grants immunity, even though there was no previous exposure [3]. Inoculation (in the case of smallpox, for example) was given to individuals *before* they were infected. Perhaps antiviral software will evolve to the point where it is capable of removing a new virus without human intervention. Although some users updated their antivirus software before their computers get infected, and are thus immune before being infected, the time scale is too short to justify a separate rate describing the change from susceptible to removed. Current protection utilities can already detect variants of a virus that had been studied, but a completely different virus usually requires the attention of a dedicated virus analyst.

1.1.3 Latent period, incubation period and serial interval

A latent period is the duration of time starting from the stage when an individual is first infected but not infectious, and ends at the time when the individual becomes infectious. This does not contradict the definition for carriers; Carriers continue to infect without unleashing its payload, as opposed to simply residing parasitically on its host.

An incubation period starts from the point of infection up to the time the symptoms appear. The stage where the individual becomes infectious comes before the symptoms appear meaning the latent period is part of the incubation period. The concept of carriers and its slight complication to the mathematical model, arises from this lapse in time of discovery.

The time elapsing from the first case of infection until the second case is called the serial interval [2]. The definition given by Bailey [3] is the period between the first and second *observation of symptoms*, where the second case is directly infected from the first. For our purposes, this difference in definition is relatively trivial, and not critical in the formulations of our models because the serial intervals for current computer viruses are usually too short for it to have any significant impact.

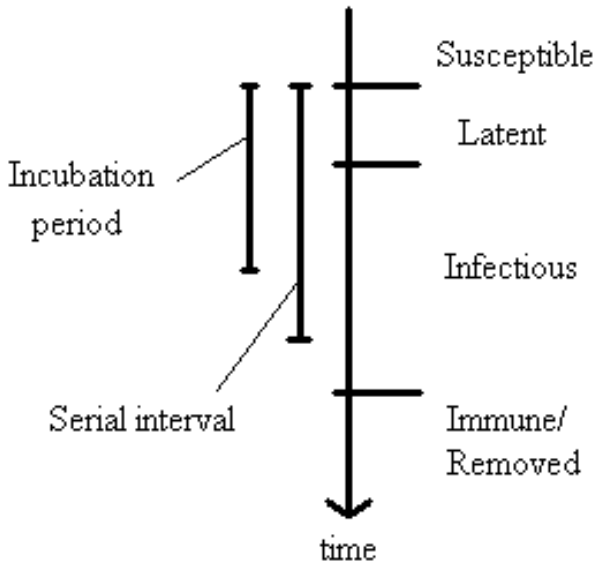

Figure 1.1: Flow of state of an individual. The incubation period is constrained to end in the infectious period, but the serial interval can be shorter or longer (as in this case) than the incubation period.

1.1.4 Carriers

A carrier is an individual capable of spreading the virus but does not seem to have been infected at all. This poses a difficulty in immediate identification of infected individuals. We shall look more closely at the complications involved later.

An interesting point to take note is that while this characteristic of delayed symptom surfacing is probably a genetic side effect of biological viruses, computer viruses possessing this quality are usually deliberately coded with this attribute in mind. It is then a balance of the number of computers the virus can infect and the rate at which the possible damage can be done. The former concentrates on creating a large base of initial infectives, thus increasing the chances of growing into a full-blown epidemic but with a higher risk of premature extinction. The latter devotes attention towards generating enough damage to curtail the progress of its removal.

1.1.5 Vectors

Vectors[1] form a separate population other than the one we are concerned about (humans and computers for the biological and digital contexts respectively) which are susceptible to the same virus. The survival of the virus then depends not only on the primary host population, but also on the secondary vector population. Examples will be given in section six of chapter two to illustrate the idea.

1.1.6 Stratification

The method of spreading by a virus is sometimes affected by certain characteristics of the population. Poliomyelitis affects predominantly children, although adults are just as vulnerable. Thus age is a factor to be considered. The general social behaviour of the population comes into the picture when discussing HIV, since its main mechanism of spreading is through sexual contact. The resulting division of the population into groups (age, sex) or *strata*, creates a more realistic view. Our focus will be on spatial stratification of the Internet (which is not entirely geographical in nature).

1.1.7 Homogeneity

Most mathematical models assume that the population in question mix around frequently and widely. This means that there are no smaller groups of individuals who keep in contact only (relatively speaking) with each other. This is clearly an imprecise view, since the population is automatically segregated into clusters. The smallest non-trivial of the clusters is the family unit, expanding to each family member's social circle, and finally encompassing the entire population.

The Internet originated as a small number of computers connected to each other to ease information processing. These computers are even spatially close to each other. Then new clusters were born and geography began to play a part in processing speed. Science and technology improved and mitigated the physical limitations of hardware connections. One can try tracing the routes taken from one's computer to another computer (or server or website) using `tracert` under the Windows 9X command prompt. For example, `tracert www.yahoo.com` took about 24 intermediate stops to reach

[1] Not be confused with the conventional mathematical usage.

the search engine website on a personal computer in Singapore.

However, the Internet does not seem to possess any readily understandable form of stratification. Geography now does not critically affect the connections ever since technology bridged the physical distances. We shall capture this spatial form by fitting our models on directed graphs[2], which is inherently heterogeneous.

1.1.8 Endemicity

Despite our efforts to eradicate computer viruses, it is disheartening to note that there exist viruses that survive in low numbers in the digital realm. Even though these viruses were discovered and a removal process for them was found, they still persist and continue to inconvenience the unfortunate few who housed them. These hardy viruses are termed as "living in the wild" in the computing world. It is thus helpful to remember that the initial number of infected might have to include those endemic in the population, thus giving a higher number than what we expected.

1.2 Methodology

1.2.1 Deterministic versus stochastic

Epidemiological models can be broadly divided into two classes: deterministic models and stochastic models. Generally speaking, deterministic models provide the population size of the infectives (or whichever population the model was created to characterise) at a particular point of time in an epidemic. These models are deterministic in that their formulations do not incorporate randomness [2]. Examples include differential equations and integral equations with initial conditions.

On the other hand, stochastic models give the probability of the infective population (again, depending on the model) being equal to a certain number at a particular point in time. For example, these models can give the probability that 50% of the total population is infected 1 unit time after the start

[2]The number of edges per vertex or the *degree* of the vertex, is taken to be very much smaller than the total number of vertices in the graph. Otherwise, a (near) complete graph is formed, and the behaviour will be similar to homogeneous models.

of the epidemic.

At first glance, stochastic models seem to be a better choice. However, as the number of computers we consider increase, the fluctuations calculated in a probabilistic treatment start to cancel out each other. The epidemic should thus progress in a relatively more stable manner. Due to the large number of computers we are concerned about, and the mathematical difficulties in handling stochastic models, we shall direct our attention to studying the effects of deterministic models in this project.

1.2.2 Parameter estimation

The early steps into mathematical epidemiology centred on simple concepts and mathematical equations. The formulated models were then fitted with data from real epidemics to test the validity of those models. Now we have a new tool for the formulation and analysis of our models: the computer[3]. Calculations previously done laboriously by hand are now completed in a fraction of the time taken before. However, this does not mean that our models are any easier to analyse.

Mathematical models require the use of variables, such as that representing time. The more complex our conditions, the larger (usually) the number of variables we need. The problem is, we do not know the values of these variables. Without actual data to check, we can only make educated guesses. Even if the required information is obtained, the problem of fitting the data to our models still remain. It is fortunate that we can rely on computers to manage the large amount of numbers. As with any scientific modelling, we must be objective about the conclusions that can be drawn. The model is an approximation to the real situation, data obtained are subject to inaccuracies and there are computational errors involved in the calculations. We must therefore remain sceptical of our results and continue critiques on our models.

1.3 Modelling a computer virus epidemic

The basis of a computer virus epidemic relies on a structure that is not homogeneous. The structure is formed with a combination of links between

[3]Ironically, the very thing that supports computer viral growth is used to analyse computer virus epidemics.

different Internet Service Providers (ISPs), electronic mail (email) between individuals (sometimes of different countries) and a certain amount of program sharing (between people actually living in physical proximity).

The use of directed graphs to capture this unique structure was first undertaken by Kephart and White [5]. This method allows us to explicitly introduce the topological factor into our models. Thus we approach the problem on two fronts. We attempt to predict the course of the epidemic using mathematical formulae and we run computer simulations to test our theories.

The first type of graph will be random graphs, as was studied by Kephart and White. We then turn to tree graphs to imitate the pattern found by Cheswick [18]. Finally, we propose a new graph structure that models frequent program sharing between groups of individuals and a small amount of contact between groups. This model is built with email being the foremost decision factor in mind.

Chapter 2

Epidemic models

Before we can simulate the course of a viral outbreak, the simulation has to be based on certain assumptions or rules. Thus we come to the formulation of epidemic models. There are three main aims to keep in mind [2]: a deeper comprehension of the driving forces behind an epidemic, prediction of an epidemic and control of an epidemic. Firstly, we want to have a better understanding of the different processes involved in the spread of computer viruses. This requires a mathematical structure for analysis work and the fact that ultimately, the computer simulation of the spreading is based on mathematical foundation. For example, Kermack and McKendrick would not have arrived at the simple conclusion that a threshold value had to be exceeded before an epidemic could occur without their model's equations [3].

This brings us to the second aim, which is to predict the future course of the epidemic. This is especially important, since computer viruses operate on a shorter time scale than their biological counterparts, where hours or even minutes are used instead of days and weeks. Antivirus firms sometimes require hours after the surfacing of a new virus to create an antidote. In the meantime, antivirus vendors are swamped with requests for product updates. Networks are clogged with messages from concerned users warning their friends. The media reports on the devastating effect that the new virus is causing, unintentionally creating more chaos. There is a need to know how the virus will affect networks, to be able to assess damages brought about, and be armed with some knowledge about the virus to calm users' minds.

The third aim is to understand how to control the spread of the epidemic, which comes naturally after having foresight of viral spread. Taking medical analogies, the methods most often used are education, immunisation and isolation. Various websites give information on what to do when a virus is

suspected to be present on the user's computer. Even more important is what *not* to do when a virus is detected[1]. Immunisation takes the form of antivirus firms uploading the virus definition (which enables antiviral software to detect the virus in question and disabling it) and their customers updating their own copy of the software. Isolation takes place when the user simply relinquishes contact with the outside world through the computer. Some business companies take this to the extreme and shuts down their entire computer system when a virus is suspected to have infiltrated their network.

Although the numbers we consider are discrete (integers), we shall assume continuous functions of the population. The first reason is that the populations in question are usually large, so the assumption on continuity as well as differentiability is reasonable [1]. The second reason involves the change in population size. If the change in the continuous version is larger than that of the integer version, then we can usually use a continuous function for approximation [7]. The models described here in this chapter had been slimmed down to only the concept they were intended to portray. These models are the SIR model, exponential growth model, logistic model, sequential growth model, carrier model and the vector model. More realistic models would involve a combination of these models.

2.1 SIR model

The acronym takes its initials from (or various forms of) the words *susceptible*, *infected* and *removed* (permanently immune). This type of model was first proposed by Kermack and McKendrick in 1927 [1] to incorporate the successive states of a patient's health into a mathematical form. Simply put, the model is described by

$$\frac{dS}{dt} = -\beta SI \tag{2.1}$$

$$\frac{dI}{dt} = \beta SI - \gamma I \tag{2.2}$$

$$\frac{dR}{dt} = \gamma I \tag{2.3}$$

[1]Basically, users are advised not to email to large numbers of people to warn them, which might start a virus hoax in itself.

where $S = S(t)$, $I = I(t)$ and $R = R(t)$. β and γ are the infection and cure rates respectively.

There is no analytical way to solve the equations. Nevertheless, there are different approaches to take in handling them. Daley [2] divided equation 2.2 by 2.1 to get

$$\frac{dI}{dS} = -1 + \frac{\gamma}{\beta S} \tag{2.4}$$

Integration gives

$$I(t) + S(t) - \frac{\gamma}{\beta} \ln S(t) = S_0 + I_0 - \frac{\gamma}{\beta} \ln S_0$$

where S_0 and I_0 are the initial numbers of susceptibles and infectives respectively. With the assumption that S, I, R are nonnegative, and some observations and results that followed (such as S being strictly decreasing and so on), Daley concluded that S, I and R converges to finite limits S_∞, I_∞ and R_∞. Another conclusion was that $I_\infty = 0$ meaning that the epidemic will eventually die out. Brauer [1] also obtained the equation 2.4, but geared its analysis towards studying the maximum value of $I(t)$ during the epidemic.

Also in the similar class are the SIS and SIRS models. The SIS model assumes that an infective, once cured, returns immediately to the susceptible state with no immunity. An infective in the SIRS model receives temporary immunity after being cured. The protection wears off after some time, leaving the cured infective susceptible again.

2.2 Exponential growth model

In 1978, Thomas R. Malthus proposed a mathematical model of the world's population, which embraced an exponential growth nature. His grim opinion [20] thankfully did not become a reflection of the current world's status. The growth of a lethal computer virus, however, does seem to follow somewhat the exponential growth model at the beginning of its proliferation.

Exponential growth of a virus occurs when the virus can thrive with possession of unlimited resources and no outward competition for those very resources [1]. In a virus's case, its resources are the computer terminals, which act as hosts for its parasitic residence. With the large number of terminals connected in the Internet, we can assume a virus has infinite hosts for

it to infect. If one looks closely at the existing viruses now, very few viruses possess destructive abilities that are powerful enough to fulfil the "end of the world" prophecy. This can be attributed to the fact that most virus writers are not interested in absolute network disruptions [19], their digital creations focussed mainly on propagation of itself. Thus, for the sake of argument, distinct viruses (discounting possibly mutated variants) thrive independently of each other, with no contention of computer host resources (so it is possible for multiple viruses to coexist on a single machine).

We shall assume that the viral spread is a homogeneous system, and that its growth depends on the number of infected computers currently [1]. Expressed mathematically, for a viral population x at time t, a short duration h from time t to time $(t + h)$ results in the approximate number of new infections βhx, where the constant β is the *per capita birth rate* or *infection rate*. Similarly, there exists a constant γ, the *per capita death rate* or *cure rate*, with a corresponding approximate number γhx of "cured" computers. Thus, the total change in infected computers from time t to time $(t + h)$ is given by

$$x(t + h) - x(t) \approx (\beta - \gamma)hx(t) \tag{2.5}$$

Dividing both sides with h, we get

$$\frac{x(t + h) - x(t)}{h} \approx (\beta - \gamma)x(t)$$

Let $h \to 0$, then from calculus, we have

$$\frac{dx}{dt} = (\beta - \gamma)x \tag{2.6}$$

with the assumption that the function $x(t)$ is differentiable.

This differential equation has the general solution $x(t) = ke^{(\beta-\gamma)t}$ where k is a constant. If we impose the *initial condition* that the initial number of infected computers is x_0 at time $t = 0$ (that is, $x(0) = x_0$), then the solution becomes

$$x(t) = x_0 e^{(\beta-\gamma)t} \tag{2.7}$$

2.3 Logistic model

In the exponential growth model, we used a constant growth rate with respect to the total population. This resulted in an unbounded growth process.

To maintain some form of control, we use a simple decreasing function of population $\lambda - ax$ to replace the constant growth rate [1]. This gives

$$\frac{dx}{dt} = x(\lambda - ax)$$

which was first introduced by Verhulst (1838). Its more common form is

$$\frac{dx}{dt} = rx(1 - \frac{x}{K}) \tag{2.8}$$

where $r = \lambda, K = \frac{\lambda}{a}$. It should be noted that equation 2.8 behaves like equation 2.6 when x is small (compared to K). When x is close to K, $\frac{dx}{dt} \approx 0$, meaning that x hardly changes. Restricting x to be in the region $0 < x < K$, the solution to equation 2.8 is

$$x(t) = \frac{Kx_0}{x_0 + (K - x_0)e^{-rt}} \tag{2.9}$$

2.4 Sequential growth

Another way of looking at the growth rate of a virus is through a sequence of numbers. Starting with an initial number of infections x_0, we proceed to the next number using the form $x_{n+1} = f(x_n)$, where the population of the infectives at the next time step x_{n+1} is determined by a function $f(\cdot)$ of the current population x_n. For simplicity, the time interval between each pair of numbers in the sequence is taken to be a constant.

In the trivial case where the number of infections varies as a result of only the constant infection rate β and the cure rate γ, then the difference between the population at the new time step and the current time step is described by

$$x_{n+1} - x_n = \beta x_n - \gamma x_n$$

Rearranging the equation, we get

$$x_{n+1} = (1 + \beta - \gamma)x_n$$

Using $r = 1 + \beta - \gamma$, we obtain a *linear homogeneous difference equation* [1]

$$x_{n+1} = rx_n \tag{2.10}$$

If the initial number x_0 is given, then the difference equation can be easily solved using mathematical induction, with the solution being

$$x_{n+1} = r^{n+1}x_0,\ n = 0, 1, 2, \ldots$$

Now suppose that the function f is

$$f(x_n) = x_n + rx_n(1 - \frac{x_n}{K})$$

where r, K correspond to those in the logistic model. Then

$$x_{n+1} = f(x_n) \Rightarrow x_{n+1} = x_n + rx_n(1 - \frac{x_n}{K})$$

or

$$x_{n+1} - x_n = rx_n(1 - \frac{x_n}{K}) \tag{2.11}$$

which is the logistic difference equation. If $r > 2$, periodic solutions appear [1]. In fact, there exists a solution of period 3 [1] (Saha and Strogatz, 1995), which implies the existence of solutions of all other periods [10, 1] (Li-Yorke Theorem).

Brauer commented that

> the fact that such simple models lead to unpredictable results suggests that experimental results and observations may not be repeatable

It was also suggested that we should focus on values of r giving predictable behaviour. However, we should not close ourselves to the possibility that a system showing irregular behaviour could actually be modelled by simple rules.

2.5 Carrier model

Humans who are afflicted with medical viruses such as typhoid, tuberculosis and poliomyelitis (or polio, its more common name) sometimes do not display symptoms of the sickness until after the incubation period of the virus. These people are called carriers, who appear healthy in all aspects even though they are already capable of transmitting the virus to other humans. This complicates matters as there are now three main groups: the healthy, the carriers and the diagnosed infected. Carriers sometimes remain infectious for a very

long time without succumbing to the virus's effects. Polio, for example, has an incubation period of about one to three weeks. What this means is that the virus has a chance to spread without the interference of remedying actions, a sort of "head start" in the chase by medical practitioners.

In April 1991, the Michelangelo virus was discovered in Sweden and the Netherlands. What it does is that if an infected computer is booted on the birthday of the artist, March 6, the virus will erase important parts of the hard disk. Without specifically checking the computer for viruses, the user may never know of the infection until the virus activates. Even more problematic during the ignorance of the user, is the fact that the virus can infect the boot sectors of any accessed diskettes on the host computer. This enabled the virus to reach out beyond its confinement on its host.

2.5.1 A simple model

The problem of the three interacting groups can be simplified if we assume that the diagnosed infected are cured the moment the infection is discovered [3]. With sufficient vigilance on the part of the user and the expertise of the programmers behind the antiviral software, instantaneous removal of the virus is possible. This leaves the susceptibles and the carriers. Let there be x carriers and y susceptibles at time t. The differential equations describing the process are

$$\frac{dx}{dt} = -\gamma x, \qquad \frac{dy}{dt} = -\beta yx \tag{2.12}$$

With initial conditions $x(0) = x_0$ and $y(0) = y_0$, the equations are easily solved. The solutions are

$$x(t) = x_0 e^{-\gamma t}, \qquad y(t) = y_0 \exp\left(-\left(\frac{\beta x_0}{\gamma}\right)(1 - e^{-\gamma t})\right) \tag{2.13}$$

Obviously, detection of the infection occurs when the first case is reported. However, this usually happens quite some time after the virus is initially introduced into the population, particularly in the carrier model's case. Suppose the introduction of the virus occurs at $t = 0$ as before, and that the viral infection is first discovered at time t_n. At this point of time, the cure rate γ is changed to γ', usually such that $\gamma < \gamma'$, as knowledge of the virus's existence increases vigilance on the part of users, thereby increasing the rate of removal of the virus. After t_n, the equation for susceptibles remains unchanged. The equation for carriers is now

$$\frac{dx}{dt} = \begin{cases} -\gamma x & 0 < t \leq t_n \\ -\gamma' x & t > t_n \end{cases} \tag{2.14}$$

The solution for $0 < t \leq t_n$ are as in equation 2.13, while for $t > t_n$, it is

$$x(t) = x_0 e^{-\gamma t_n - d'(t-t_n)}, \qquad y(t) = y(t_n) \exp\left(\left(-\frac{\beta x(t_n)}{\gamma'}\right)(1 - e^{-\gamma'(t-t_n)})\right) \tag{2.15}$$

2.5.2 An emphasis on the infectives

The previous model places an emphasis on the number of uninfected individuals. From a medical point of view, it is important as it is a priority to maximise the number of people who escapes unscathed. For the duration of the outbreak, the population can be assumed to be constant and is closed. This means that the sum of the susceptibles, the carriers and the diagnosed infected remains relatively constant (due to restricted immigration and emigration). Thus the number of actual cases of infection can be estimated.

However, the number of computers (connected to the Internet) in the world is hardly a stationary number, if one gives some thought to it. New businesses acquire dozens of computers at a time. The digital devices that can be connected to the Internet are not even confined to just personal computers and workstations anymore. Hand-held PDAs (Personal Digital Assistants) and mobile phones are now also open to viral attacks (though not as advanced or deadly as normal computer viruses). Estimating the size of the number of these Internet portals will be difficult, if not impossible. We shall thus focus on the infected and the carriers instead.

Let x be the infective population and y be the carrier population. Suppose that each computer has an average number of θ contacts. This establishes a form of locality and in a certain sense, enforces a closed population. Let $p(x, y)$ be the available number of susceptibles for an infective. Another representation could be $p(x)$, where we assume that carriers can be "infected" to become a known infective. The dependence of p on x and y can be reasoned out if we assume that $p \propto \frac{1}{x+y}$, that is, when x and y are large, p is small and vice versa. The dependence is then easily seen. Obviously, $p \leq \theta$.

In the same manner, we let $q(x, y)$ be the available number of susceptibles for a carrier and that $q \propto \frac{1}{x+y}$. There are two possible choices for the representation of p, but q does not have this option. Infectives cannot

produce new carriers, as any computer suspected of being in contact with the infective is automatically classified as either a new infective or free of infection. Therefore, new carriers must come from only the susceptibles. The differential equations now stand as

$$\frac{dx}{dt} = p\beta_1 x - \gamma_1 x, \qquad \frac{dy}{dt} = q\beta_2 y - \gamma_2 y \tag{2.16}$$

where β_1, γ_1 correspond to infection and cure rates of x and similarly for y.

These equations cannot be solved, but the point is that a single new property or concept rendered the original model difficult to analyse. We shall look at a simplification of p and q, making them functions of time instead. This made the model amenable to some form of analysis and is studied more closely in the subsection on random graphs.

2.6 Vector model

The concept of vectors can be easily visualised using malaria as an example. The agent responsible for spreading the virus is a particular species of mosquito. The female mosquito will suck the blood of an infected individual, thus catching the virus. It will then proceed to suck the blood of another individual, thereby passing the virus to that person. Tracking the progress of malaria thus requires monitoring the states of both human and mosquito populations. With a slight relaxation of definition, HIV can also be considered as a host-vector situation, where the male and female populations are considered as both the host and the vector.

The Brain virus spreads primarily through the use of floppy diskettes. There are probably two reasons for this. The first is that the virus was written such that its infection process does not include a self-replication mechanism. It can infect any diskettes that was used on its host computer, but that was about the extent of its replication ability. The second reason is perhaps a little subtler. The computers at the time the virus was written were manufactured basically as isolated, stand-alone machines. For the virus to spread, it will have to be physically moved from one computer to another. In this case, it was in the form of diskettes.

We shall give a slightly modified version of the vector model given by Bailey[2]. Let x_1, y_1 be the host infective and host susceptible population, with

[2]Bailey used an SIR model, ours will be an SIS model.

β_1, γ_1 as the corresponding infection and cure rates. Let x_2, y_2 be the vector infective and vector susceptible population, with β_2, γ_2 as the corresponding infection and cure rates. We assume that the virus can only be transmitted vector to host and vice versa. The set of equations are

$$\frac{dx_1}{dt} = \beta_1 x_2 y_1 - \gamma_1 x_1, \qquad \frac{dy_1}{dt} = -\beta_1 x_2 y_1 \tag{2.17}$$

$$\frac{dx_2}{dt} = \beta_2 x_1 y_2 - \gamma_2 x_2, \qquad \frac{dy_2}{dt} = -\beta_2 x_1 y_2 \tag{2.18}$$

Detailed discussion of the analysis is found in [3]. The main thing to note is the interdependence between the different populations. Ths concept of vectors added a higher level of complexity than the carrier model. As we incorporate more conditions into our model, the formulation process becomes harder, until we finally turn to computer simulations to test the validity and usefulness of the newly added concepts.

2.7 Other models

Stochastic models are not exactly models *per se*. Most deterministic models possess a stochastic counterpart. It all depends on the description of the virus or the modeler's preference. Considerations include the decision on continuous or discrete time [2] or whether the model is Markovian. Some models which involve only probabilistic treatment are Reed-Frost and Greenwood models, and they are in fact Markov chains as well (Their equations involved binomial terms, hence they are also known as chain-binomials [3, 2]).

When geography plays an important role in viral spread, we consider spatial models. There are different ways to formulate the model. Some models incorporate the spatial densities of the populations instead of the population number. Other considerations include velocity of propagation [3] or diffusion of infection [5].

The models we have considered so far involved only a few states, namely susceptible, infected or removed. There are also multistate models [3], which consider several distinct states. Common uses apply to tubercolosis.

Then there are interference models [3], where two or more (viral) populations vie for (the same) resources. Along the same line of thought are

interaction models, or more specifically, predator-prey models [1]. A common usage is to model the populations of foxes and rabbits (if we assume that foxes rely exclusively on rabbits for food).

Finally, we have the rumour model, used to describe an "infection" of the mind. Our computer virus equivalent would be virus hoaxes[3]. However, virus hoaxes should not be seen as any less harmful. The trouble we had regarding the `sulfnbk.exe` involved the inaccurate diagnosis by antivirus software, resulting in a harmless file being treated as infected. The problem is that this file exists in all computers installed with Windows, thus a lot of computers were "infected". The confusion worsened when popular antivirus software did *not* detect this file as dangerous (because it is not, but it is still susceptible to a real infection and there lies the confusion), and concerned users tried to remove the file themselves, which in turn caused some malfunctions[4].

[3]Visit www.vmyths.com to find out more.

[4]The file is used to restore long file names, but is not essential for running Windows. Details at `http://support.microsoft.com/`, with a search for `sulfnbk.exe`.

Chapter 3

Random Graphs

The Internet has grown into such a massive network that the thorough mapping of the linkages [18] does not form a definite, decipherable pattern. Random graphs are suitable for modelling due to its relative ease of use in simulation work, the simplicity of its structure for analysis and that it possesses the haphazard quality of the structure of the Internet. Hence, we shall first study the spread of viruses on random graphs.

For a set of N vertices, a random graph is constructed by randomly selecting an edge out of the $N(N-1)$ possible edges (a complete graph has $N(N-1)$ edges). For a (bi)directional random graph, the size of the selection pool is doubled (vertex i has one outgoing edge to vertex j, and an incoming one from vertex j), thus yielding an edge set of cardinality $2N(N-1)$. Each vertex in the graph can be viewed as an individual computer. If we disregard the details of infection and discretise the "health" of the vertex into distinct states, then we can attach an infection rate β_{ij} to an edge going from vertex i to vertex j. The cure rate, however, is associated with the vertex i, the value denoted by γ_i.

It is noted that Kephart and White did not mention the bidirectional nature of their model. However, since we are concerned about the outgoing edges, the calculations involved are not affected. Instead of modelling the dynamics of the infected population, their model focused on the *fraction* of the infected individuals with respect to the whole population. The details of infection and removal are simplified by assuming a global infection rate β and cure rate γ.

Assuming the population is large, let $i(t) \equiv \frac{x(t)}{N}$ be the fraction of infected individuals in the population, where $x(t)$ is our usual function for the infec-

tives and N is the population size (or graph size). For an infected vertex, it has an expected number of $p = r(N-1)$ outgoing edges, where r is the probability of an edge being selected. Since the fraction of susceptibles is $1-i$, therefore the expected number of susceptibles available to this infected vertex is $r(N-1)(1-i)$ or $p(1-i)$. The differential equation is then

$$\frac{di}{dt} = \beta p i(1-i) - \gamma i \tag{3.1}$$

with the solution

$$i(t) = \frac{i_0(1-\frac{\gamma}{\beta'})}{i_0 + (1-\frac{\gamma}{\beta'} - i_0)e^{-(\beta p(N-1)-\gamma)t}} \tag{3.2}$$

where i_0 is the initial fraction of infected vertices and $\beta' = \beta p(N-1)$.

3.1 Redefining population closure

However, Kephart's and White's model again assumed a closed population. This concept of closure automatically establishes the balance between the infected and susceptibles (and possibly the removed). This prevented the viral growth from increasing at an even higher rate as well as stabilising the infected population to a homogeneous limit in most classical models.

Previously, we introduced the notion of an average contact number in the section about the carrier model. As a substitute for the closure concept, it fails miserably when the contact number becomes large. The basis of the closure concept is that as the number of infectives grows, the number of susceptibles decrease. We can simulate this effect by making the contact number a function of time, namely $p(t)$ and that $p(t)$ is a generally-decreasing function. By generally-decreasing, we mean that $p(t)$ is not confined to monotonically decreasing but is such that as t increases, p eventually becomes 0 or close to it.

With this new definition of p, the contact number concept will certainly produce a similar effect as the closure concept. However, p is independent of x, the number of infected computers. Can this be justified? Consider this: News of a powerful and recently-birthed virus will soon spread to the masses, by antivirus companies, through the media or via family, friends and colleagues (perfect conditions for the rumor model). All of this reduces the number of available susceptibles. Even people who are suspicious of virus hoaxes will be slightly more wary, though they might not take very elaborate

measures to protect themselves. With this assumption, p can be reasonably taken to be simply dependent on time.

We shall investigate the effects of $p(t)$ by taking p to be either linearly, quadratically or exponentially decreasing. Note that the functional decay refers to $p(t)$, not $x(t)$. The model will be imposed on a simple differential equation

$$\frac{dx}{dt} = \beta px - \gamma x \tag{3.3}$$

with the usual infection and cure rates β and γ respectively, and that $p(0) = p_0$, the initial number of contacts.

3.2 Linear decay of $p(t)$

Suppose $p(t) = p_0 - at$ and $a > 0$. Now $p(t)$ is zero when $p_0 - at = 0$ or $t = \frac{p_0}{a}$. The differential equation is now

$$\frac{dx}{dt} = \beta px - \gamma x = (\beta(p_0 - at) - \gamma)x \tag{3.4}$$

with the solution

$$x(t) = x_0 e^{(\beta p_0 - \gamma)t - \frac{\beta a}{2}t^2} \tag{3.5}$$

Now

$$\frac{dx}{dt} = 0 \Rightarrow x_0 e^{(\beta p_0 - \gamma)t - \frac{\beta a}{2}t^2} \left((\beta p_0 - \gamma) - \beta at\right) = 0$$

which implies

$$(\beta p_0 - \gamma) - \beta at = 0 \quad \text{or} \quad t = \frac{p_0}{a} - \frac{\gamma}{\beta a}$$

This stationary point can be shown to be a maximum by taking a second derivative and it can be shown that $\frac{d^2x}{dx^2} < 0$ at $t = \frac{p_0}{a} - \frac{\gamma}{\beta a}$. There is only one stationary point and this point is the maximum. This maximum can be found by substituting $t = \frac{p_0}{a} - \frac{\gamma}{\beta a}$ back into the equation, giving

$$x_{\max} = x_0 e^{\frac{(\beta p_0 - \gamma)^2}{2\beta a}}$$

Note that this peak occurs just before the time when p becomes zero.

Note that our experiments on the random graph will be done on a 10000-vertex random graph. The graph models in the two subsequent chapter will

follow as closely as possible to a graph size of 10000. We shall follow the model by Kephart and White, using $\beta = 1.0$ and $\gamma = 0.2$. Their simulations yielded a homogeneous limit of 80 infected vertices, corresponding to 80% infection. We shall thus use 8000 as a guide for our homogeneous limit. If we assume that $p_0 = 20$, $x_0 = 1$ and $x_{\max} = 8000$, then solving for a, we obtain $a = 21.811$ and the peak occurs at $t = 0.9078$.

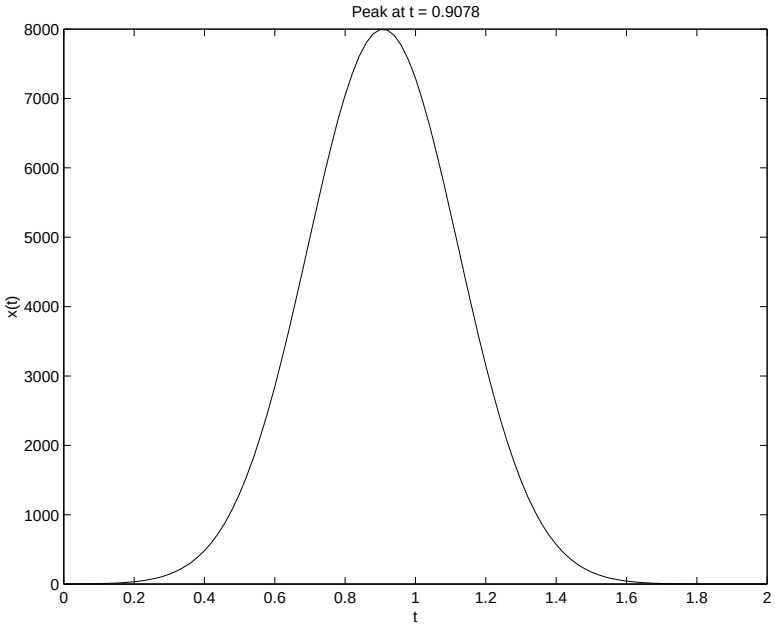

Figure 3.1: A peak of 8000 with linear decay

Since $p(t) = 0$ at $t = \frac{p_0}{a}$, any furtherance of x beyond $t = \frac{p_0}{a}$ is unrealistic. For $t > \frac{p_0}{a}$, the equation for $\frac{dx}{dt}$ reduces to the simple form

$$\frac{dx}{dt} = -\gamma x \tag{3.6}$$

with solution

$$x(t) = x_0 e^{\frac{\beta p_0^2}{2a} - \gamma t} \tag{3.7}$$

3.3 Quadratic decay of $p(t)$

To simulate a quadratic decay of $p(t)$, we let $p(t) = p_0 + bt - at^2$, where $a > 0$ and no limitation of sign on b. Following the format before, we find

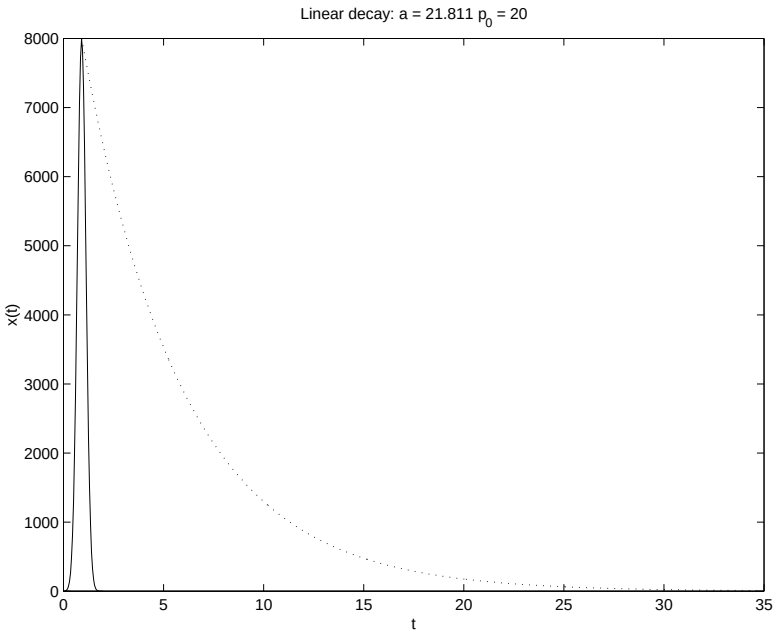

Figure 3.2: The dotted line shows the modified course of x after p reaches zero.

that $p(t) = 0$ at $t = \frac{b+\sqrt{b^2+4ap_0}}{2a}$, ignoring the negative value for t. The corresponding differential equation and its solution are

$$\frac{dx}{dt} = (\beta(p_0 + bt - at^2) - \gamma)x \tag{3.8}$$

$$x(t) = x_0 e^{(\beta p_0 - \gamma)t + \frac{\beta b}{2}t^2 - \frac{\beta a}{3}t^3} \tag{3.9}$$

The computation of the peaking time is now slightly more complicated. We shall simplify matters by assuming $p_0 \geq 1$ and that $\gamma < \beta$. Then solving for t by equating $\frac{dx}{dt}$ to zero, we can again ignore the negative value and get

$$t = \frac{b + \sqrt{b^2 + 4ap_0 - \frac{4a\gamma}{\beta}}}{2a}$$

Again, note that the peak occurs just before p becomes zero, similar to the linear decay version.

We shall let $b = 0$ to further mitigate the complications, thus appointing a as the dominating factor. Similarly, we can solve for a and t when the peak occurs with $x_{\max} = 8000$ to get $a = 42.7134, t = 0.68428$.

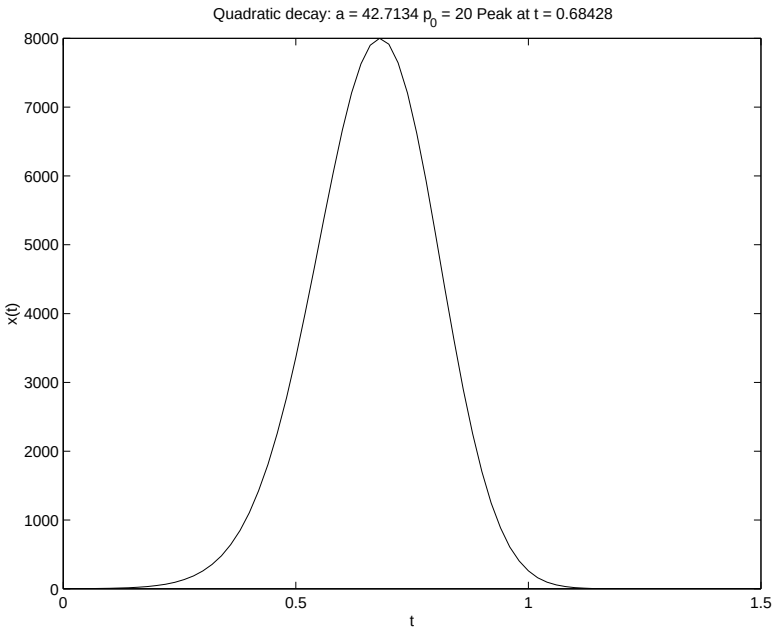

Figure 3.3: Quadratic decay of $p(t)$

The progress of x after the point of time when $p = 0$ is exponential decay, with the calculations of the new x nearly identical to that in the previous section.

3.4 Exponential decay of $p(t)$

The functional form of $p(t)$ will be taken to be $p(t) = p_0 e^{-at}$ in this section. Using the same method in the previous two sections, the infected population is dictated by

$$x(t) = x_0 e^{\frac{\beta p_0}{a}(1-e^{-at})-\gamma t} \tag{3.10}$$

Solving for $\frac{dx}{dt} = 0$, we get $t = \frac{\ln \frac{\gamma}{\beta p_0}}{-a}$. If we assume that $0 < \gamma < \beta$ and $p_0 \geq 1$, then t is positive. Due to its mathematical formula, $p(t)$ is never zero. However, we can analyse the period of time when $p(t) > 0$. Solving for t, we find that this occurs when $t < \frac{\ln p_0}{a}$, which encompasses the point of time when the peak occurs.

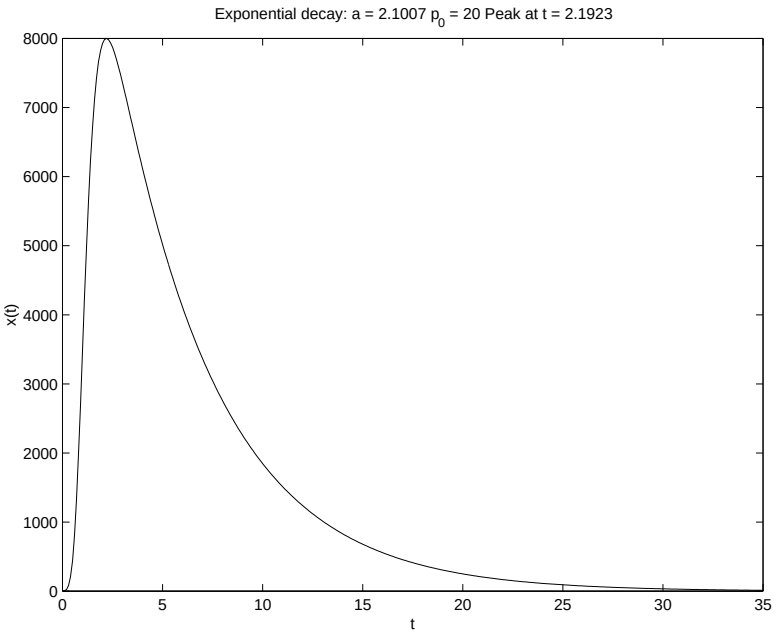

Figure 3.4: Exponential decay of $p(t)$

Chapter 4

Tree graphs

The random graph has one fundamental flaw with regards to modelling the Internet. Since its edges are randomly generated, in general, the graph so obtained is non-hierarchical. However, the Internet has a hierarchical structure, such as an ISP (Internet Service Provider) with its subscribing clients, or the internal email system between employers and employees. This is detrimental to its usage as a topological approximation to the Internet. Furthermore, it is not certain that a random graph is connected as a whole [12, 13]. This phenomenon can be verified using $\beta = 1.0$ and $\gamma = 0.2$, one can verify that as the total number of vertices increase, the stabilising number of infected vertices start to break away from the homogeneous limit of 80%.

To correct this structural defect, we shall explicitly construct a connected hierarchical graph by using trees[1]. We arrange all the vertices on a huge tree structure, imitating the map by Cheswick [18]. To simplify construction, we shall use regularly-structured trees, the simplest of these being the binary tree.

The hierarchical model by Kephart and White [5] can be described as a "virtual" tree. They placed the vertices on the ends of a binary tree (or leaf vertices). The vertices in the body of the tree are used to denote different levels of infection rate. Their model thus automatically creates clusters of vertices with high contact rates within each cluster.

We shall form our model on a graph that is almost a tree. For an average contact number p, every vertex (except the leaf vertices) satisfy the condition of connecting to p other vertices[2]. The leaf vertices are connected to only

[1] A tree is defined as a graph with no cycles in graph theory.

[2] More precisely, it should be $p + 1$, but this fails for the root vertex. For convenience,

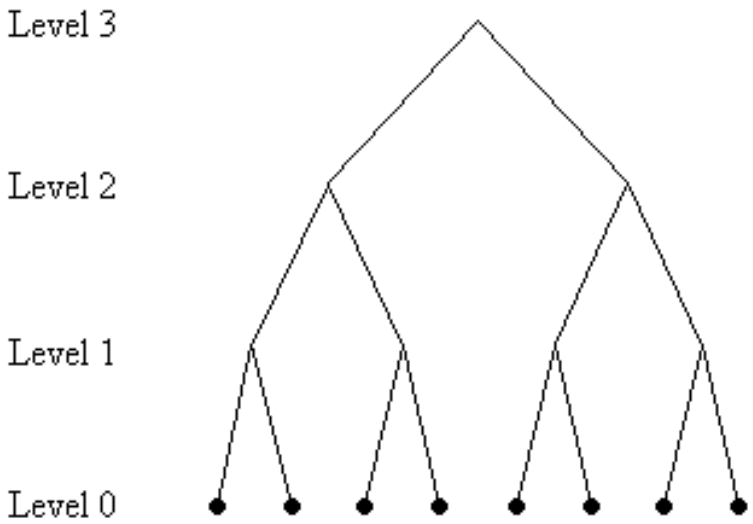

Figure 4.1: "Virtual" tree. The black dots represent the vertices.

their parent vertex. Joining each leaf vertex to every other leaf vertices of the same parent, we form cliques at the last level of the tree, shown in figure 4.2.

4.1 Applying the average contact concept

Previously in the carrier model, we introduced the concept of an average number of contacts per infective. However we did not conduct analysis due to the unknown form of the function $p(t)$. In the section in random graphs, we gave suggested explicit forms of $p(t)$, but the results are extremely sensitive to the parameters of $p(t)$. Small changes to a shifts the peak value from its original homogeneous limit by a large difference.

We shall study three concepts to the furtherance of the generally-decreasing nature of $p(t)$, namely: *advance alarm*, *natural response* and *periodic activity*. First is the "kill signal", suggested by Kephart and White. In simple terms,

we assume that there are p neighbours.

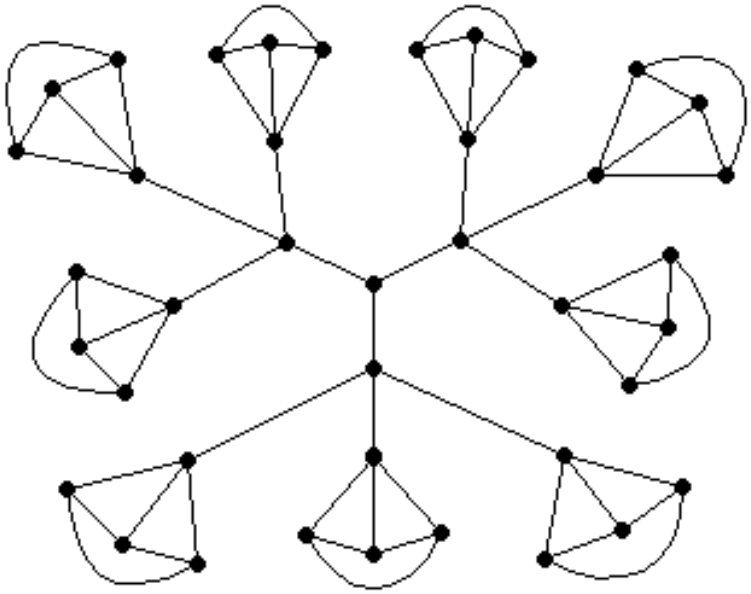

Figure 4.2: Augmented tree graph

it simply means that an individual once infected, informs its neighbours that they too might be infected.

Next, natural response is when an individual becomes more vigilant immediately after infection, but as time passes without any new infection, that individual becomes less careful. This effectively is the SIRS model, but instead of transitioning discretely from immune to susceptible, an individual gradually loses its immunity, becoming more susceptible as time passes.

Finally, after examining the growth of actual viruses (however scarce or inaccurate the data was), we found that typical viral growth has periodic properties. It increases, then decrease, and then increase again, though to a lower number. It continues the increase-decrease pattern, the local peak of each increment being lower than the last one. What is more interesting is the fact that the average period is about 24 hours. This can be explained if we assume that in a local area, most of the computers in that area are switched on and off at roughly the same time. The duration of time between switching on and off of the computers can be regarded as normal working hours. Stretch this to include computer servers (which in essence, are never

switched off, except for maintenance) and personal computers (which could be switched on from say 2 to 10 hours), and we get a global computer uptime of about 24 hours.

4.2 Simulations

We did simulations to evaluate the three concepts separately. Without imposing all three conditions, the tree graph gave similar results as the random graph, quickly attaining an 80% homogeneous limit. When advance alarm was activated, we found that almost all the simulation runs for both types of graphs reached zero infections[3]. The number of infected vertices in the tree graph usually peak at about 600 and then fall again to zero after about 30 units of time. The number of infected vertices in the random graph behaved similar, but with a higher peak of about 2000 and a shorter life span of about 15 units of time. The significantly higher peak by random graphs is due to the relatively higher level of homogeneity of random graphs, as opposed to the limited structured linkages in the tree graph. The shorter existence is a result of the "rebound" effect by the high peaks. More infected vertices means more vertices are aware of the existence of the infection, thus raising the general level of vigilance, resulting in a sharp decline of infected vertices. If this occurs fast enough, the rebound effect will push the viral infection to extinction. This means the advance alarm system is an extremely effective tool in deterring the growth of computer viruses. The fact that epidemics still occur is probably mainly due to difference between the real-world global vigilance state and the ideal total vigilance state.

Activating the natural response system on random graphs gave two general results. The first one is that the virus still dies out, usually at about 10 units of time. The second gave a high peak of 3000 infected vertices, but the rebound effect did not produce extinction. Instead, a stable endemic state with about 100 infected vertices occurred. Tree graphs gave the peak-then-death result as before in the advance alarm system.

When periodic activity was imposed, a pronounced recurring fluctuation appeared after a homogeneous limit was reached. The fluctuations corresponded to the uptime condition. We used 24 units of time to represent a day and the resulting simulation runs reflected this period, shown in figure

[3]Results for random graph simulations without any of the three concept are similar to those in [5]

4.3.

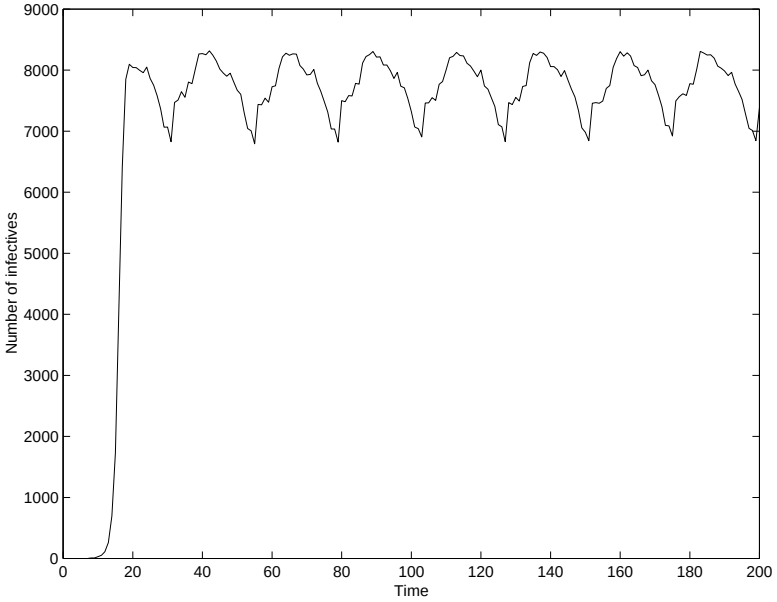

Figure 4.3: Periodic fluctuation in tree graph

Simulations were run on the tree graph to investigate the effects of the three systems. Results from the original tree graph gave zero infections very early and is almost guaranteed. So we modified the original graph by adding an additional random edge to each vertex. Typical results are shown in figures 4.4 and 4.5. When the natural response system was involved, the "valleys" between each peak were flatter, wider and lower. This means that for a relatively long period of time, the number of infectives was very small. For extinction of the virus to be possible, it would be logical to attack during these periods, such as performing virus checks at the end of the periodic activity cycle (which coincides with the period with low numbers).

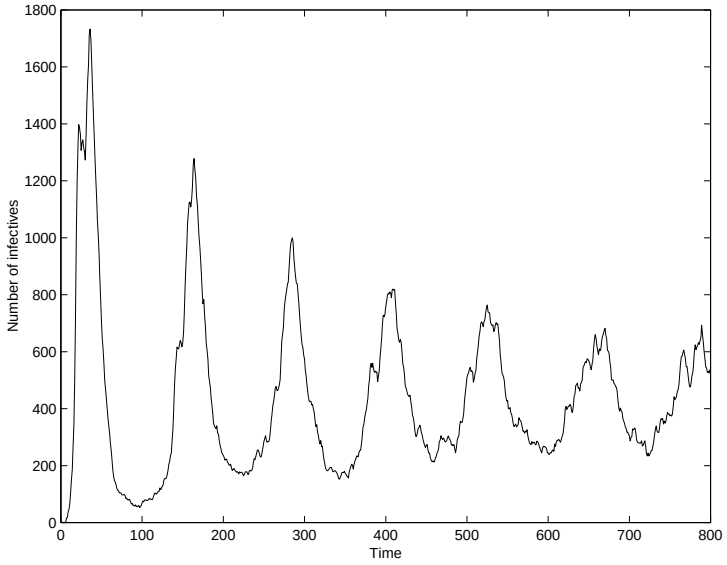

Figure 4.4: Tree graph with advance alarm and periodic activity activated.

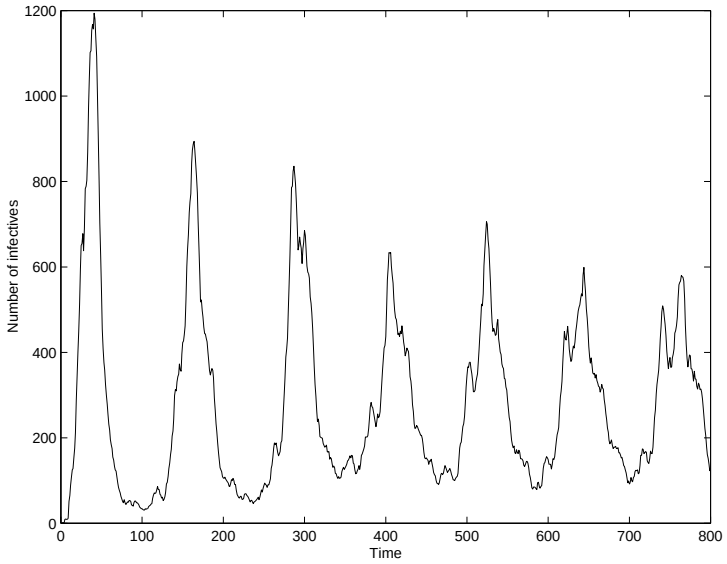

Figure 4.5: Tree graph with all three systems activated.

Chapter 5

Proposed Internet model

Cheswick captured the topology of the Internet by tracing various ISPs, which intuitively gives a tree-like structure. However, the most common mode of virus transmission currently is by email. The Goner worm spreads by pilfering the email addresses in the local address book and then sends a copy of itself to names in its stolen list. With this mode of infection in mind, we need to incorporate locality and some randomness into our graph. First we form cliques to establish localisation. Then we add edges (randomly) to join a clique to another clique.

5.1 Simulations

Using the advance alarm, natural response and periodic activity systems described previously, we run simulations on this new graph structure. The proposed graph gave similar results as the tree and random graph without imposing any of the three systems, reaching a homogeneous limit quickly and then stayed there with small fluctuations. Results with only one of the three systems are similar to those of the tree graph. However, with all three systems activated, the simulations gave a growth pattern similar to those of figure 5.2, disregarding differences in peak values.

5.2 Comparisons

For the convenience of discussion in this section, simulations will be assumed to take the advance alarm and periodic activity systems. Differences will be

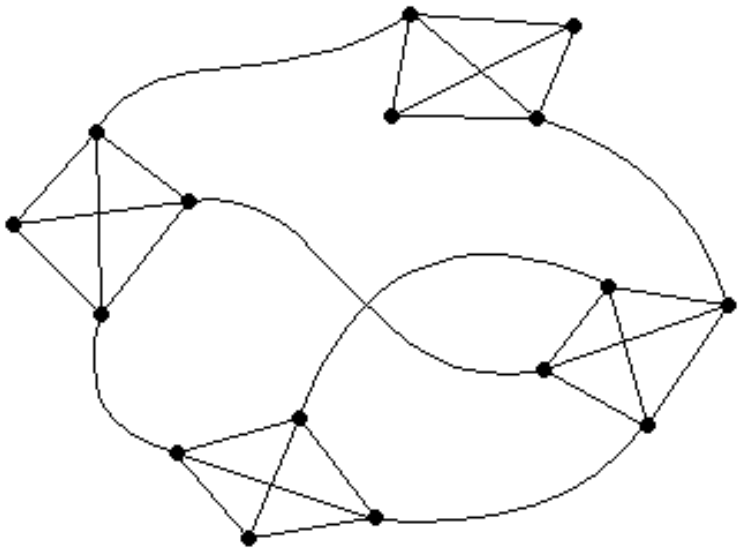

Figure 5.1: Proposed graph model formed with random linkages between clusters

the presence or absence of the natural response. We start by revisiting random graphs. Simulations without natural response gave very high first peaks, then falling rapidly to reach a low homogeneous limit with stable fluctuations.

Random graph simulations with natural response behaves similarly, but with lower peaks. We believe that the unusually high first peaks is due to the high connectivity of the random graph (as mentioned earlier while discussing tree graphs). Although the number of edges per vertex is the same as (or as close to) other graph models, the number of vertices or edges on a path between any two vertices[1] is small, thus a large part of the graph is reached before the advance alarm system can impede the progress.

The spatial graph is constructed by placing the vertices on a two dimensional, 100 by 100 square matrix. The sides of the matrix are then wrapped around to form a torus. Simulations without natural response produced extremely erratic fluctuations.

We also investigated two other factors combined with all three systems on the proposed graph. For the first factor, instead of only one initial infective, we randomly planted twenty. The effect is similar to the original plot. This

[1] In graph theory, this would be the diameter of the graph.

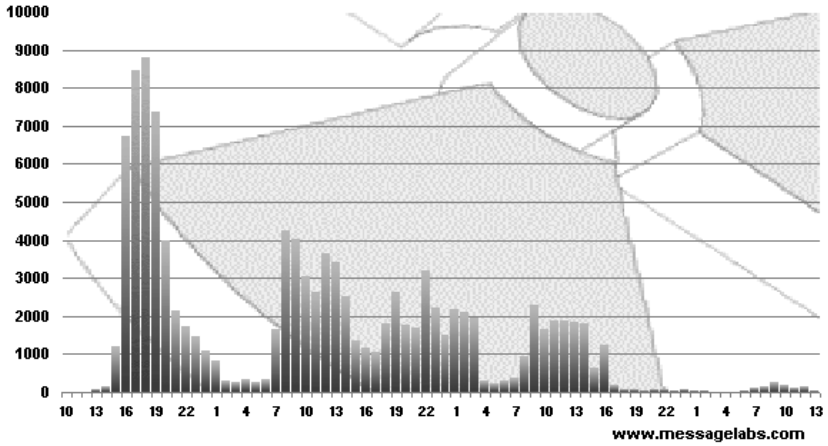

Figure 5.2: A typical real world example of a virus growth process. This depicts the course of the Goner worm growth and will be used as the standard model for our simulation work.

suggests that having a large number of initial infectives does not change the general course of the epidemic, only that survival is more assured. We also tried delaying the imposition of the advance alarm system by 24 units of time. Naturally, the first peak is higher than in the original plot, but as soon as the alarm system was activated, the course of the epidemic reverted to its original shape. This strengthens the theory that the advance alarm system is an extremely powerful deterrent.

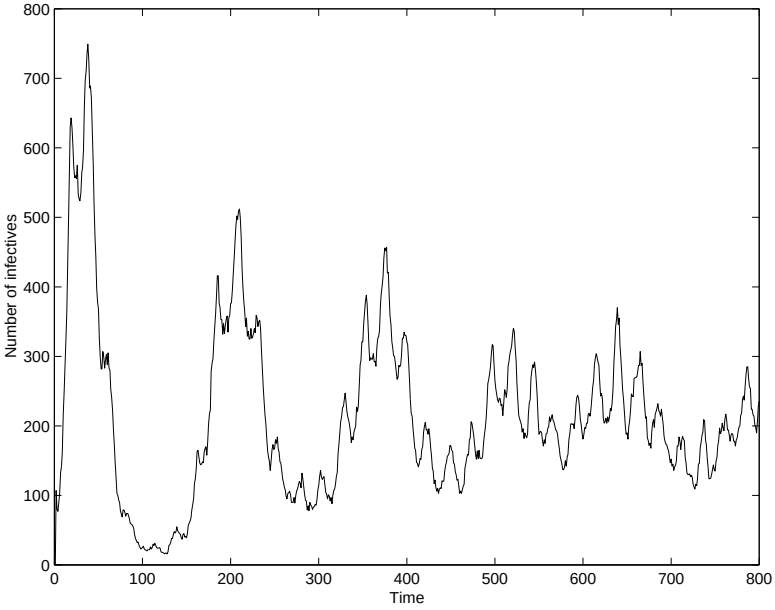

Figure 5.3: Proposed graph with all three systems activated.

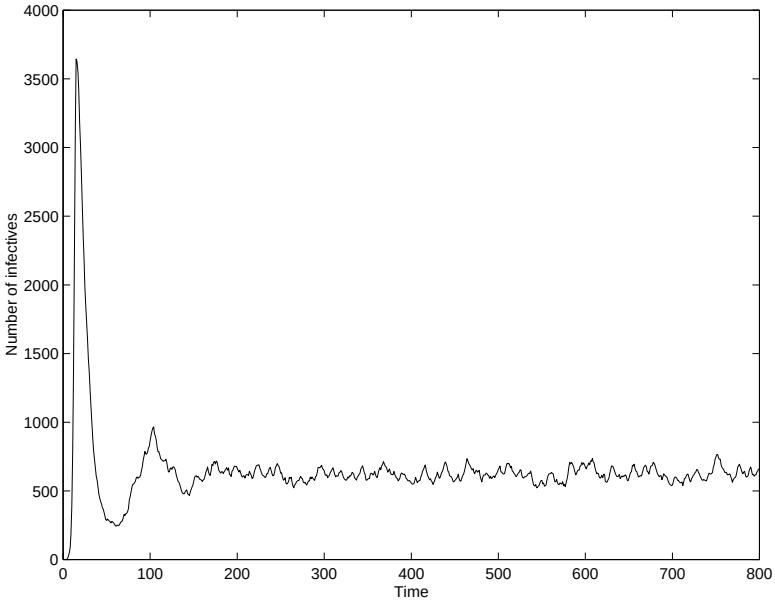

Figure 5.4: Random graph with advance alarm and periodic activity.

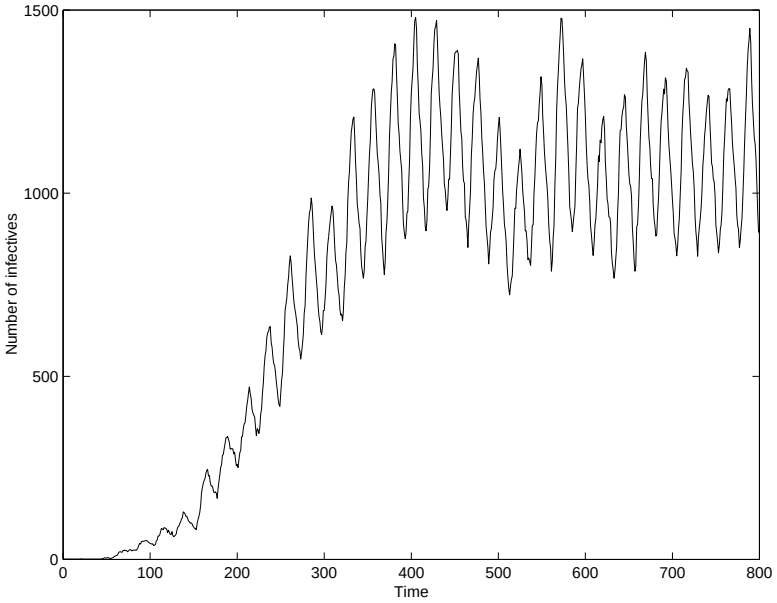

Figure 5.5: Spatial graph with advance alarm and periodic activity.

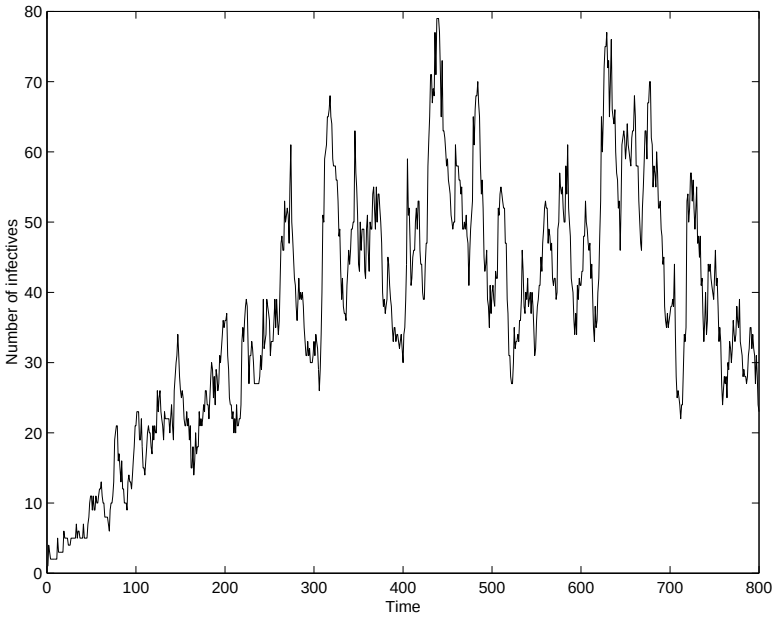

Figure 5.6: Spatial graph with all three systems activated.

Chapter 6

Consolidation

It has been proven that antivirus software cannot detect every virus [6], not to mention disabling all of them. However, it is also a known fact that no virus thus far has been able to bring down the entire global network. An observation on current virus fighting is that, as imperfect as antivirus software are, they still remain an effective impediment to the spread of viruses. Another point to note is the spread of *news* of the virus, which is the advance alarm system discussed in the section on tree graphs. Although it proved to be very effective in bringing down the number of infectives in our simulations, we assumed that the warning to other individuals is passed via the very same links as the virus infection. Contacts between two individuals might be different in the physical world. A concerned user might use the telephone to pass the warning, or simply inform others face to face. Another graph set could be used to model the contact links, but it will require twice the computation work. The model also did not take into account the real world problem of bandwidth consideration.

The models discussed in chapter 2 do not explain why current virus epidemics seem to grow at an exponential rate but eventually fall off to a low number, and are possibly endemic as well. It is believed that the real situation is probably a combination of those models, as well as certain other factors such as the periodic activity system that are unique to computer virus epidemics.

In chapter 3, we looked more closely at the alternative view to the concept of population closure, introduced earlier in the section on carrier models. By using random graphs, we are assuming that individuals have about the same number of contacts each. We proposed that instead of looking at the restriction on the entire population, we examine the effects of controlling

this average contact number. This brings the analysis down from a macroscopic view of global infection to a microscopic view of localised infection. Specific mathematical forms of this concept were given, but the results were extremely dependent and sensitive to its parameters.

Random graphs are known to be disconnected unless they have a relatively large number of edges (discounting multiplicity). To improve on our graph model, we used a tree graph, which retains the property of an average contact number per vertex, the function p, but is topologically different from random graphs. The mathematical formulation is similar to that of random graphs, since the only difference is the change in graph structure but keeping the property of having the same degrees for each vertex. Since the functional form of p is not known, any further educated guesses on its formula will not yield any deeper insights. Thus we abandoned the mathematical approach, and concentrated on computer simulations instead, incorporating those same educated guesses into the computer program. From the simulations, we hope to gain an insight to the workings of p. We might then achieve a better understanding of its properties, which can finally lead to an explicit formulation of p.

The overall results given by our simulations on the tree graph does not seem to tally with the current trends of computer viruses. Most of the simulation runs actually indicated the extinction of the virus. The only explanation we can give involve two properties of the tree graphs. The first property is a high connectivity of the vertices. If there are l levels in the tree, then an infection at the leaf level needs only to pass through a maximum of $2l$ vertices to reach another leaf vertex in another cluster. This allowed the infection to reach a lot of vertices in a short time. The second property is about the increasing clustering effect when one travels from the root vertex (level) to the leaf vertex (level). Each vertex can be thought of as belonging to a cluster at its own level. After the infection established its prevalent existence throughout the graph, the clustering effect of the arrangement of vertices enforced a divide-and-conquer condition. The infected vertices became isolated in their own clusters and soon dies out. This is the rebound effect where the extinction follows closely after a high peak. Activating all the three systems on the tree graph produced subsequent decreasing peaks. However, the peaks are usually preceded by a sharp increase and followed by a sharp decrease. This is slightly different from our defined standard growth course (see figure 5.2).

As encouraging as the findings of simulations on tree graphs are, the

mechanism of viral spread do not operate with such optimistic results. A new structure was proposed in chapter 5. Reiterating the three main factors in the simulations, they are the advance alarm (kill signal), natural response (an increased vigilance after being cured of infection) and periodic activity ("operational uptime" of each vertex). Comparisons were made between four different types of graph models (random, tree, spatial and the proposed model), as well as different combinations of the graph models with the three systems. We found that the proposed model gave the closest approximation to a real epidemic growth.

From all the simulations as well as preliminary formulations of the average contact number p, we believe the properties of p include periodicity and a decreasing magnitude as time passes. However, without more data to compare with, it is difficult to draw further conclusions. Data collection is difficult when dealing with biological viruses, but is even more so with computer viruses. No one likes to admit they were infected, especially business companies since it directly reflects on their credibility and security. The number of known infections is thus not very accurate.

6.1 Future research

Finally, research could be done on fractal topology, where a closer look at a smaller section of the graph (or whatever representation the modeler chooses) reveals a structure similar to a larger section of the graph. This was represented by the tree graph, but perhaps a more mathematically rigorous analysis could be done with fractals.

6.2 Conclusion

The search for simplifying the underlying mechanisms of a computer virus epidemic led us to the use of graph models. Further investigations showed that a good graph approximation is not enough to simulate real viral growth. Three other independent factors were discovered. They were the advance alarm system, natural response system and the periodic activity system. The resulting pattern, generated from these four factors, appear to grow similarly in tandem with real data, with subsequent decreasing, non-sharp peaks. Separate investigations on each factor did not yield similar effects. Hence, we

believe that these three systems, together with the graph structure of our proposed model, form a close estimation of a real computer virus epidemic.

We believe that computer viruses, though difficult to eradicate completely, could (and should) be kept to a manageable level. The investigations conducted here are steered towards that direction. In doing so, we hope that the results are helpful in providing a deeper understanding of the workings of a computer virus epidemic.

Bibliography

[1] Fred Brauer, Carlos Castillo-Chávez, *Mathematical Models in Population Biology and Epidemiology*, Springer-Verlag New York, Inc., 2001

[2] Daryl J. Daley, J. Gani, *Epidemic Modelling: An Introduction*, Cambridge University Press, 1999

[3] Norman T. J. Bailey, *The mathematical theory of infectious diseases and its applications*, second edition, Charles Griffin & Company Limited, 1975

[4] Paul Waltman, *Deterministic Threshold Models in the Theory of Epidemics*, Springer-Verlag, 1974

[5] Jeffrey O. Kephart, Steve R. White, *Directed-Graph Epidemiological Models of Computer Viruses*, Institute of Electrical and Electronics Engineers, 1991

[6] Frederick B. Cohen, *A short course on Computer Viruses. Second Edition*, John Wiley & Sons, Inc., 1994

[7] Clifford Henry Taubes, *Modelling differential equations in biology*, Prentice Hall, Inc., 2001

[8] Edward Beltrami, *Mathematics for dynamic modelling, second edition*, Academic Press, 1998

[9] F. Reif, *Fundamentals of statistical and thermal physics*, McGraw-Hill, Inc., 1965

[10] Denny Gulick, *Encounters with chaos*, McGraw-Hill, Inc., 1992

[11] Sheldon M. Ross, *Stochastic processes, second edition*, John Wiley & Sons, Inc., 1996

[12] Béla Bollobás, *Random graphs*, Academic Press Inc. (London) Limited., 1985

[13] Svante Janson, Tomasz Luczak, Andrzej Rucinski, *Random graphs*, John Wiley & Sons, Inc., 2000

[14] John von Neumann, *Theory of Self-Reproducing Automata*, University of Illinois Press, 1966

[15] John H. Holland, *Adaptation in natural and artificial systems*, Massachusetts Institute of Technology, 1992

[16] Daniel Dorling, David Fairbairn, *Mapping: Ways of Representing the World*, Addison Wesley Longman Limited, 1997

[17] Warren G. Kruse II, Jay G. Heiser, *Computer Forensics: incident response essentials*, Addison-Wesley, 2001

[18] Bill Cheswick, Hal Burch, *Internet Mapping Project*, `http://www.cs.bell-labs.com/who/ches/map`, 1998-2001

[19] Sarah Gordon, *The Generic Virus Writer*, `http://www.research.ibm.com/antivirus/SciPapers/Gordon/GenericVirusWriter.html`, 1994

[20] Thomas R. Malthus, *An Essay on the Principle of Population*, `http://socserv2.socsci.mcmaster.ca/~econ/ugcm/3ll3/malthus/popu.txt`, 1798