1-2011

An Exploration of Modeling Techniques for the Study of the Dynamics of E-Mail Viruses

Karin Weule
Montclair State University

ABSTRACT

Title of Thesis: AN EXPLORATION OF MODELING TECHNIQUES FOR THE STUDY OF THE DYNAMICS OF E-MAIL VIRUSES

Karin Weule, Master of Science, 2011

Thesis directed by: Dr. Lora Billings

Department of Mathematical Sciences

We analyze real data sets from two e-mail viruses, the Magistr.b and the Sircam.a to explore how we can use mathematical models to predict the behavior described by the data. Analysis of the data is conducted primarily with computer programming in MatLab. We focus mainly on the use of two continuous models commonly used in the study of biological diseases, the SIS and the SIR models. A discrete modeling approach using agent-based simulations is also explored and revealed to be potentially useful in developing a compartmentalized model that incorporates both SIS and SIR model behavior. The theory behind the continuous models and the programming method for the simulations are described.

The factors that affect the spread of biological infections, such as exposure rates and recovery rates, are factors with similar driving force in cyberspace. The parameters that govern the movement of these computer viruses through the susceptible population of computers on the internet are identified as the contact rate, β, and recovery rate, γ. We use the real data to estimate the values for these parameters and use these values in our models to find the one that best matches the behavior described by the data.

We approximate values for β using a standard method and find that β must be very small to account for an almost linear growth in the infection early on. The recovery rate, γ, is found by taking the reciprocal of the average duration of infection. Unlike biological diseases which take their course in a host for set period of time, these e-mail viruses show durations of infection that vary widely. Using a mean duration of infection calculated from the data, the SIS model reaches a non-trivial endemic state. However, such an endemic state is not supported by the data.

Closer analysis of both data sets reveals that the durations of infection for infected computers actually decreased over time. By applying a time-dependent $\gamma(t)$, we are able to modify the behavior of the SIS model. We are able to approximate the shape of the latter half of the time series. In a similar fashion, we apply a range of linear functions for $\gamma(t)$ to the SIR model. Using very small β, we can approximate the shape of the first half of the time series.

We find that the introduction of a variable γ modifies the behavior of both models in such a way that it remains unclear which model best reflects the behavior of our viruses. Only qualitative fits were achieved with the Magistr.b virus and both the SIS and SIR models. It is possible that a precise match to either of the continuous models could not be achieved because the dynamics of these viruses involve both SIS and SIR behavior. That is, some of the infected computers become

completely disabled by the infection and thereby enter the Removed class of an SIR model, while others are repaired and enter the Susceptible class of an SIS model. Computers with longer durations which have significant lags in time between detections suggest the possibility of re-infection consistent with the SIS model. The development of a compartmentalized model using discrete agent-based simulations may provide us with a better fit to the data and is described as a future direction for the work put forth in this paper.

The results of this project demonstrate that, even without achieving a precise match to a model, we are able to reveal the existence of a time-dependent $\gamma(t)$. We show that by decreasing the recovery time for infected computers, *i.e.* by increasing $\gamma(t)$, we can drastically reduce both the magnitude of an outbreak and the time it takes for the population to reach a disease-free equilibrium.

MONTCLAIR STATE UNIVERSITY

AN EXPLORATION OF MODELING TECHNIQUES FOR THE STUDY OF THE DYNAMICS OF E-MAIL VIRUSES

by

Karin Weule

A Master's Thesis Submitted to the Faculty of

Montclair State University

In Partial Fulfillment of the Requirements

For the Degree of

Master of Science in with a Concentration in Pure and Applied Mathematics

January 2011

School College of Math and Sciences

Department Mathematical Sciences

[redacted]

Dr. Robert Prezant
Dean

12/16/10
December 2010

Thesis Committee:

[redacted]

Dr. Lora Billings
Thesis Sponsor

[redacted]

Dr. John Stevens
Committee Member

[redacted]

Dr. Eric Forgoston
Committee Member

[redacted]

Dr. Helen M. Roberts
Department Chair

AN EXPLORATION OF MODELING TECHNIQUES FOR THE STUDY OF THE DYNAMICS OF E-MAIL VIRUSES

A THESIS

Submitted in partial fulfillment of the requirements

For the degree of Master of Science in Pure and Applied Mathematics

By

KARIN WEULE

Montclair State University

Montclair, NJ

January 2011

Contents

List of Figures

1 Introduction

Our technological world is increasingly dependent on having our computers running as continuously and safely as possible. However, the pervasive nature of computer viruses poses a constant threat to the seamless connectivity of our computer systems. Equally as random and as patterned as human contact, computer connectivity in the virtual world of shopping malls, clubs, games, schools and work renders computers as vulnerable to infections as humans. It makes sense then, in order to gain a better understanding of the dynamics of the spread of computer viruses, that we look to the field of mathematical biology. Over a century of progress has been made in the study of the spread of human disease through mathematical modeling. The factors that affect the spread of biological infections, such as exposure rates and recovery rates are factors with similar driving force in cyberspace.

Among the commonly used models are the continuous models which use ordinary differential equations (ODEs) and the discrete models like Markov chains and agent-based simulations. This paper focuses primarily on the use of continuous models as we analyze real data sets from two e-mail viruses, the Magistr.b and the Sircam.a. We explore how well we can use the continuous models to predict the behavior described by the data. It is then possible to predict how effective certain intervention strategies may be in curbing outbreaks, minimizing the endemic state of a virus, and leading it to extinction.

The models used are the SIS and the SIR models, which differ primarily in how the population is classified. Brief descriptions of the models are given below. We identify the classes in our population (susceptible, exposed, infected, and recovered or removed) and the parameters that govern the movement of a disease through the population, specifically the contact rate and recovery rate. We use the real data to estimate the values for these parameters and use these values in several mathematical models to find the one that best matches the shape of the real time series.

While many illnesses, like the common cold, take their course in a host for set period of time, these viruses show durations of infection that vary widely. In fact, it was found using compartmental analysis that the duration of infection decreased over time. One possible reason for this trend is an increased awareness by users that their computers are infected and need to be repaired. It will be shown that decreasing duration length for infections can drastically reduce the magnitude of an outbreak and bring a population to a disease-free equilibrium more quickly. This would imply that improvements in rapid response strategies on the part of IT professionals, e-mail providers, and anti-virus software providers could play an important role keeping our computers clean and healthy.

2 The Models

In the SIS model, the susceptibles, S, come into contact with infectives, I, at a particular rate, β, which is often referred to as the contact rate. An infective

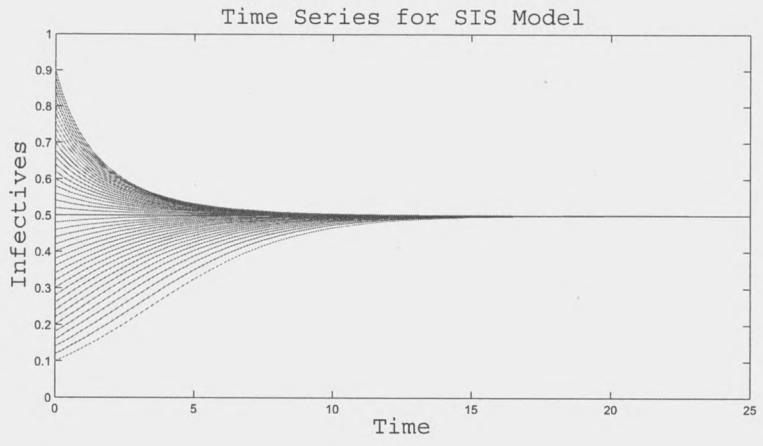

Figure 1: Time series of expected values for the percentage of total population that is infected from the SIS model. Parameters are set to $\beta = 0.8$ day^{-1} and $\gamma = 0.4$ day^{-1}. Initial conditions range from $i = 0.9$ to 0.1.

recovers from the disease at a particular rate of recovery, γ, and thereby becomes susceptible again. In this type of model, the disease can remain endemic in the population with a perpetual cycle of re-infection and recovery. The existence of an endemic state depends on the initial conditions (the number of infectives), as well as the rates of infection and recovery. The schematic looks like this:

$$S \xrightarrow{\beta IS} I \xrightarrow{\gamma I} S.$$

The SIS model is governed by the following ordinary differential equations:

$$S' = -\beta IS + \gamma I \tag{1}$$
$$I' = \beta IS - \gamma I \tag{2}$$

where β is the contact rate and γ is the rate of recovery. Note that $1/\gamma$ is the average length of time that an infective remains in the infected class and the units for both β and γ are day^{-1}. While this system can be studied using integers for the sizes of both groups in this population, we can also consider a normalized system with the variables s and i representing fractions of the population, n, where $n = 1$, $0 \leq s \leq 1$, $0 \leq i \leq 1$, and $s + i = 1$

There are two steady states that correspond to the differential equations (1) and (2). Since there are only two classes of individuals in this model, we can use the equation, $s + i = 1$, to simplify our mathematical analysis. We substitute $1 - i$ for s into Eqn. (2) and solve for the steady state, (s^*, i^*), algebraically:

$$\beta i^*(1 - i^*) - \gamma i^* = 0, \tag{3}$$

gives us the trivial, disease free solution,

$$i^* = 0, \tag{4}$$

and the non-trivial, endemic solution,

$$i^* = 1 - \gamma/\beta. \tag{5}$$

For this simple model the only variable factors are the parameters, β and γ, which relate to one another in what mathematical biologists term the basic reproductive ratio, R_0, [1]. When considering the spread of a disease, a reproductive ratio determines whether the number of infections will grow or die out. We want to know how many new infectives are produced as the result of a single infective's contact with the susceptibles in the population, keeping in mind that the total number of individuals in the infective class is mitigated by the number of infectives that recover and leave that class. When the contact rate exceeds the recovery rate, the reproductive ratio is greater than one and the disease spreads. When the ratio is less than one, the disease eventually dies out.

In a fixed population, we can interpret β to be the rate at which a single infective makes infectious contacts. As mentioned earlier, $1/\gamma$ is the length of time that an infective remains in the infected class and remains a threat to other susceptibles. The product of these two terms, then, would give us the expected number of infections caused by any one infective individual, as shown in Britton [1]. Therefore, in models like the SIS, the basic reproductive ratio is

$$R_0 = \beta/\gamma. \tag{6}$$

Thus, if $R_0 < 1$, the disease dies out. If $R_0 > 1$, the disease remains endemic in the population. Figure 1 shows the stability of the endemic state for $R_0 > 1$ given various initial conditions. The stable endemic state, with 50% of the population infected ($i = 0.5$), is consistent with the solution derived with Eqn. (5).

The SIR model provides a different picture of the progression of a disease. We add a class of individuals, called the recovered or removed class, R, which are no longer susceptible to infection. The scenario is as follows:

$$S \xrightarrow{\beta IS} I \xrightarrow{\gamma I} R$$

The SIR model is governed by the following ordinary differential equations:

$$S' = -\beta IS \tag{7}$$
$$I' = \beta IS - \gamma I \tag{8}$$
$$R' = \gamma I \tag{9}$$

where β is the contact rate and γ is the rate of recovery. For a fixed population, we can reduce this system to Equations (7) and (8), since R can calculated directly from the solutions S and I. The steady state requires that I be zero, but does not restrict the other variables other than $S + R = N$, or, in a normalized system, $s + r = 1$. Therefore, the only steady state is $(S^*, I^*, R^*) = (\hat{S}, 0, \hat{R})$, where $0 < \hat{S} < N$ and $\hat{R} = N - \hat{S}$, determined by the initial condition.

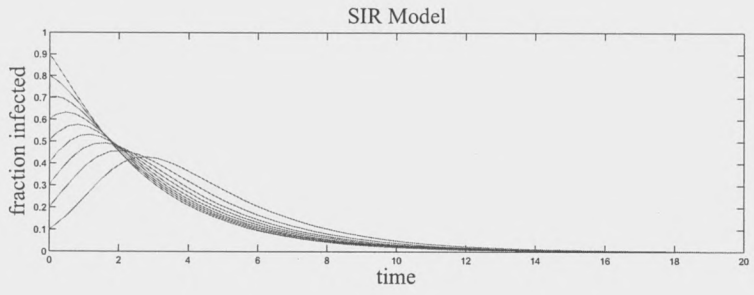

Figure 2: Time series of expected values for the percentage of total population that is infected from the SIR model using parameters $\beta = 1.6$ day^{-1} and $\gamma = 0.4$ day^{-1}. Initial conditions range from $i = 0.9$ to 0.1.

Since the individuals leaving the infected class do not re-join the susceptible group, the dynamics of the SIR model differ from those of the SIS model. All solutions of the SIR model will have the disease die out in time. The basic reproductive ratio, R_0, and the initial conditions determine if the number of infectives increase into an outbreak before the decrease to die out. If $R_0 > 1$ and $s > 1/R_0$, the infection burns through the population at an exponential rate for the first few days to achieve an early maximum outbreak. Then the susceptible pool shrinks, leaving less targets for infection and, as more individual leave the infected class, the number of infected declines, at first, slowly, then exponentially, as shown in Fig. 2.

These two dynamical systems capture the general behavior of a computer virus as well. They will form the basis on which we model the data in the following section, since we explore both reinfection and removal possibilities for recovered computers.

3 The Data

Data sets for two viruses, the Magistr.b and the Sircam.a, were used for this modeling project. The data were obtained from Message Labs, an internet provider which keeps logs of the date and time that any of its users received an e-mail with a known virus. To protect user confidentiality, a hash number was assigned to each IP address from which an infected e-mail was sent. This hash number is nothing more than an identification number that replaces a unique IP address. An example of the data format from the Sircam.a data set:

date	time	hash #
7/17/2001	7:27	1
7/17/2001	10:56	1
7/17/2001	14:17	1
7/17/2001	18:53	1
7/17/2001	19:33	2
7/17/2001	19:35	3
7/17/2001	19:47	2
7/17/2001	19:52	4
7/17/2001	20:00	2
7/17/2001	20:13	2
7/17/2001	20:16	5

To generate a time series for each virus, the date and time are converted to relative time in days, where the first detection is $t = 0$ days and subsequent detections are recorded as the difference in time from that first detection in the unit of days. For example, $t = 1.5$ days would represent a detection that occurs 36 hours after the first detection.

4 Magistr.b Virus

Magistr.b can spread three ways: by e-mail, on a local area network, or through shared disks and mainly affects users with the Microsoft Outlook, Eudora or Netscape email client in the operating systems Windows 95, 98, Millennium and 2000. In the process, it may destroy sectors of the hard drive and erase the cmos/bios [2]. The virus is triggered when a user opens an infected attachment to an e-mail message. Most of the attachments and subject headings are taken from the host PC so that often the recipient trusts the message and opens the attachment. Magistr.b then scans the user's address book, then runs its own internal e-mail program to send messages to everyone in the book.

Since it requires user action to initiate spread to other computers, the Magistr.b spreads more slowly than other viruses. Another factor affecting its spread is a package of executable files, termed a payload, which the virus activates on the host PC approximately one month after infection. The payload attacks the CMOS and BIOS of machines running machines running Windows 95, 98, and ME, which is less secure than Windows NT and 2000 machines and can delete or overwrite sectors of the hard drive. The CMOS is necessary to boot the PC, therefore its destruction renders the computer inoperable until it is repaired. On the one hand, dormancy feature allows the virus to be propagated without the host user's knowledge for one month. Even using the same address book, the messages change, thus increasing the probability that a recipient will unwittingly open one of the infected e-mails sent during that month's time. On the other hand, the destructive nature of this virus renders PCs with less secure operating systems inoperable, which would slow the spread.[2]

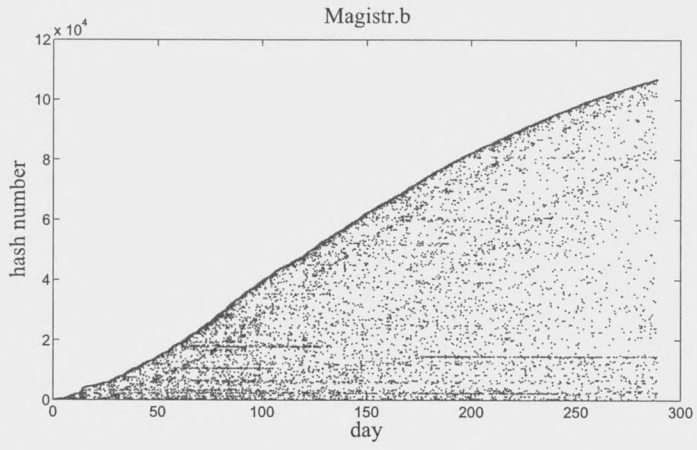

Figure 3: The detections reported in the Magistr.b virus data set. Each point represents the time a detection was recorded for a hash number. It should be noted that this is a finite data set of 288 consecutive days of activity. The new hash numbers are assigned consecutively in time, while repeat detections for previously recorded hash numbers are recorded. The virus continues to spread after the time frame shown.

The first occurrence of a Magistr.b infection was detected in March 2001.The data we used was gathered between September 4, 2001 and June 20, 2002. The most recent detections that could be found among various technical support sites for computer viruses was April 27, 2009 [2]. Current anti-virus software packages still provide protection from Magistr.b, so it continues to pose a threat. Symantec, an anti-virus software developer, currently gives this virus a rating of 2 out of 5 for severity [4]. The decline in infections is mostly due to the availability of detection and removal software. It is also likely that this virus has lost its ability to spread due to a decrease in the number of hosts with the older more vulnerable operating systems.

A plot of time vs. hash numbers for the all of the detections in the data set, shown in Fig. 3, demonstrates a steady increase in the number of new computers infected. It also shows that, once infected, many computers continue to send the virus to others for the duration of the data set. Detections for a sample of the hash numbers with longer durations are plotted to show the data points in more detail in Fig. 4. Notice that for some infected computers there are gaps of more than three weeks for detection times.

The density of occurrences in Fig. 3 might be a reflection of reinfection, which would suggest that the SIS model is a good candidate for modeling this virus. On the other hand, the virus is known to destroy sectors of the hard drive which could put an infected computer into the removed class of an SIR model. It is not clear which is the better model for this type of virus. To see a profile of the outbreak, we derive the number of computers infected each day from the data set. We record the first and last detection for each hash number and consider them the endpoint for the interval during which that computer was infected. Then, we count how many

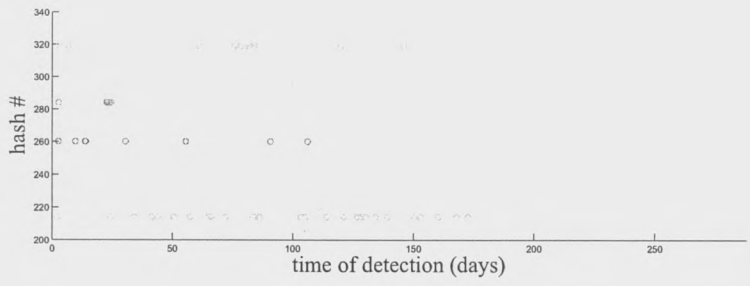

Figure 4: Detection times for a sample of infected computers (hash numbers). It should be noted that virus activity on computers using other network servers are not recorded here and may be taking place during the time gaps. For this reason, we calculate duration of infection as the difference between last and first detection.

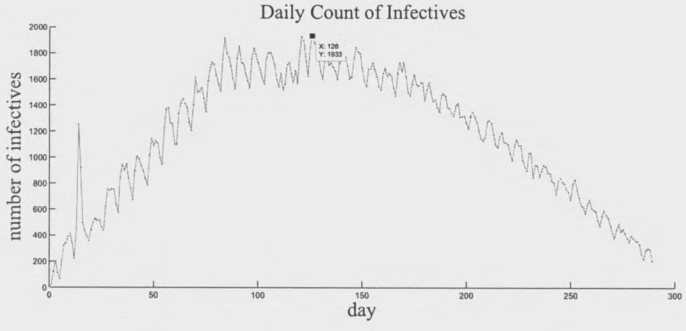

Figure 5: The daily count of number of infected computers derived from the Magistr.b virus data set.

were infected for a given day. The results are plotted in Fig. 5. Since $t = 0$ was defined earlier as the time of the first detection, the graph includes the point $(0, 1)$. To fit each model to the data, we continue by estimating values for the governing parameters, γ and β.

4.1 Finding the contact rate, β

To find an appropriate value for the contact rate, β, we use the number of infectives per day. Both models assume exponential growth in the initial stages of the outbreak. Standard methods suggest that we plot $ln(I)$ vs. time and get a linear fit. But the first 80 days of the data shows fairly linear growth, as shown in Fig. 6. Therefore, the beta that we are considering is quite small.

4.2 Finding the recovery rate, γ

We continue by examining the durations of infection in order to estimate the recovery rate, γ. The duration of infection for each hash number is calculated by

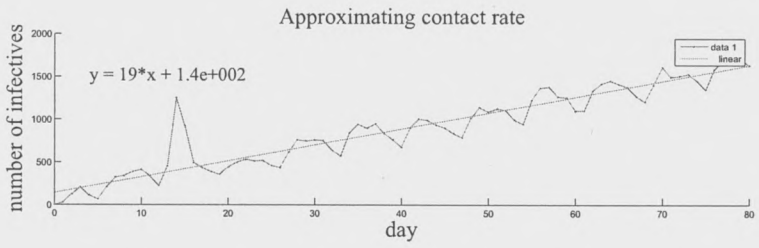

Figure 6: A linear approximation of the number of infections for the first 80 days of Magistr.b virus data set.

taking the difference between the first time and the last time that the computer was detected sending an infected e-mail. Many of the computers infected were detected sending only one e-mail, or several at the same time. For this project, we manipulate the data set in three different ways to see which is the best method for modeling the behavior of the virus. First, we leave the duration values of zero in the data set for all single detections. Second, we assume that duration of infection for single detections is one minute. Third, since a single detection can be considered an even trade of one computer recovering while one is infected, we eliminate those hash numbers from the data set.

Including durations of zero for single detections, the mean duration is 2.1603 days with a standard deviation of 16.3282. In fact, most of the durations are close to zero and the median is zero. This tells us that more than half of the computers sent only one e-mail. The computers with duration close to zero sent only one batch of e-mails almost immediately upon becoming infected. We observe another spike in frequency around 21 days. Using a value of one minute for the duration for single detections yields similar results, showing a mean of 2.1608 days with a standard deviation of 16.3281. However, taking out the zeros yields a mean of 11.8241 days and standard deviation of 36.6751. The median is 19 minutes and the mode is 4 minutes. We will explore how the SIS model behaves using the second and third set of results.

If we calculate the reciprocals of the mean durations for both scenarios (without single detections and with one-minute detections), we find that γ ranges from 0.0846 day^{-1} to 0.4628 day^{-1}, respectively. This gives us a good starting point for choosing a γ for the SIS model. We can use as an example $\beta = 5$ day^{-1}, to generate time series for a range of γ values. See Fig. 7. As predicted by our analysis of the SIS model, $R_0 > 1$ for these values and the solution approaches an endemic state, which is not reflected in the real data for Magistr.b, as shown in Fig. 5. Therefore, we continue with an improvement to the model which captures the time varying nature of the recovery time.

Closer examination of the data shows the duration of infection actually decreases with time, as supported by Fig. 8. While much of the data has only one detection or a very short duration, we see that some hash numbers are detected continuously throughout the data set. It should be noted that some of the drop-off in duration

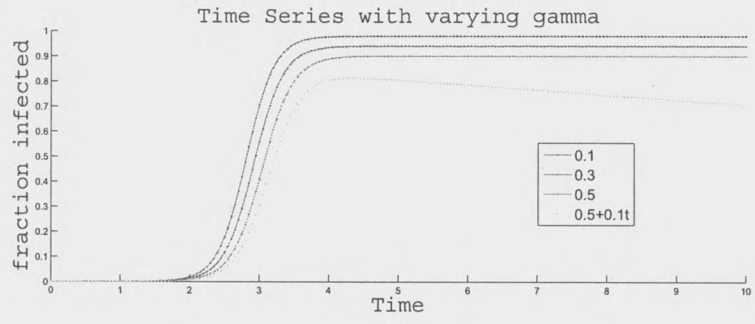

Figure 7: Time series of the number infected predicted by the SIS model with varying γ terms. Fixed values for γ lead to an endemic state, while a time-dependent γ slows the occurrence of new infections and leads to a steady decline.

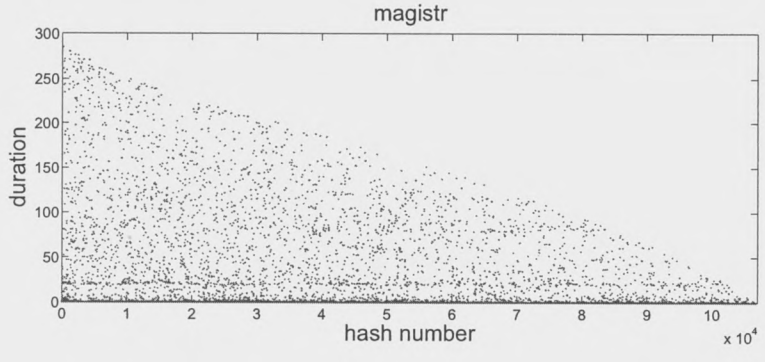

Figure 8: Duration of infection for each observed computer, labeled by hash number. A steady decrease in duration for each successive hash number infected can be observed over a fixed time period. It should be noted that the calculated durations for most computers that are still infected on the 288th day are shorter the actual durations due to the finite nature of the data set.

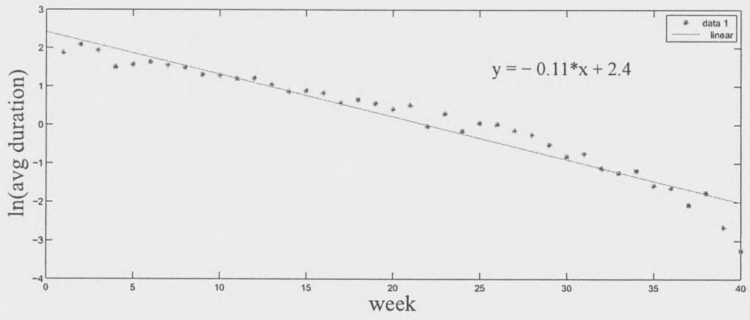

Figure 9: Time series of duration of infection for new infections grouped and counted in seven day intervals. Single detections are assigned a duration of one minute. Notice the monotonically decreasing duration of infection.

for larger (later) hash numbers may be an artifact due to the data set being finite. For example, if the hash number was detected on day 100, there cannot be a duration beyond 188 days. We can argue that, since the data is finite, a shorter duration for a hash number towards the end of set is likely to reflect the absence of a subsequent detection for that computer. But, given an average duration of 11 days for computers that sent more than one e-mail, it is safe to assume that the decrease in duration of infection shown above is real, not artifact.

Figure 8 suggests that the introduction of a time-dependent γ may make the SIS model a viable candidate for this virus. Since γ is the reciprocal of duration, a decreasing duration can be translated into an increasing γ. The fourth curve in Fig. 7 uses $\gamma = 0.5 + 0.1t$ day^{-1} and gives us a hint that this strategy might work.

We begin our analysis of how γ changes over time by separating the data set into subsets of equal time intervals and plotting a time series of the average duration of infection for each interval. To determine how long the intervals should be, we first take a look at the time series of the number of new computers that become infected each day, shown in Fig. 5. By applying the Fast Fourier Transform (FFT) algorithm to this time series, we can analyze the frequencies contained in these plots. We use the routines provided by Matlab. The FFT algorithm reveals a strong peak frequency at 0.143 which can be interpreted as a dominant cycle of 1/0.143, or approximately 7 days. The largest peak is very close to zero, which reflects a quick turn-around time for many computers who were infected and sent out only one batch of e-mails immediately upon infection.

If we use the dominant frequency of 7 days to analyze the average duration of infection, we can expect a fairly smooth plot of duration over time, since each interval will contain a local maximum and minimum. As shown in Fig. 9, the log of the average duration provides us with the best linear fit to this curve. The resulting line is $ln(y) = -0.11x + 2.4$ where y is $1/\gamma$ and x is time in weeks. We adjust the line to reflect daily average duration by dividing the slope by 7. Solving for $\gamma(t)$,

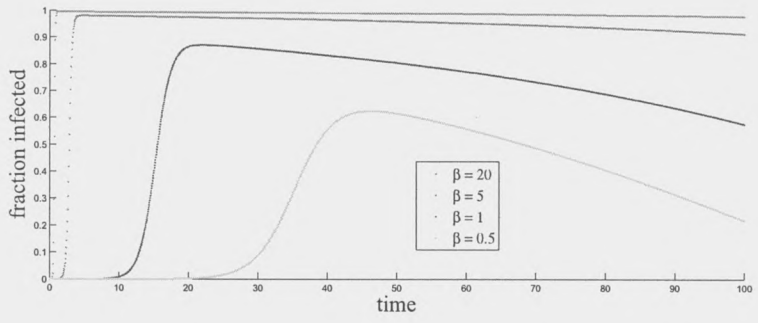

Figure 10: Time series of the number infected predicted by the SIS model for $\gamma(t) = e^{(0.0157t-2.4)}$ day^{-1}.

and using t in days, we find

$$\gamma(t) = e^{(0.0157t-2.4)} \text{ day}^{-1}. \tag{10}$$

We can now introduce the time-dependent parameter into the systems of equations governing the continuous models.

4.3 The SIS and SIR models

First, we introduce the time-dependent $\gamma(t)$ to the SIS model. As shown in Fig. 10, we see that for smaller values of β the curve bears some resemblance to the actual time series. We hypothesize the slow decrease in the outbreak is caused by the basic reproductive number $R_0 = \beta/\gamma(t)$ decreasing below one. Taking a qualitative approach and scaling the data, we fit the SIS model to the Magistr.b daily count data using a very small $\beta = 0.00001$ day^{-1} and several linear functions for $\gamma(t)$ as shown in Fig. 11. Here, rather than normalizing the system, we use a population of 4000 computers with an initial condition of 150 infectives. The dramatic burst in infections that we saw in Fig. 10 is almost linear for a much smaller β. Notice how the severity of the outbreak is reduced as we increase $\gamma(t)$.

In Fig. 12, we show the results for the SIR model using $\gamma(t) = e^{(0.0157t-2.4)}$ day^{-1} as we vary β. Smaller values for β show a curve that resembles the real data. However, the model shows much slower growth in infection in the first 30 days than the real data supports. In Fig. 13, we fit the SIR model to the Magistr.b daily count data using a very small $\beta = 0.00001$ day^{-1} and several linear functions for $\gamma(t)$. We use a population of 4000 computers with an initial condition of 300 infectives, which provides us with a shift in the model that adjusts for the initial slow growth in Fig. 12. While we are able to generate a qualitative fit to the SIR model, the end of the time series, when compared with the Magistr.b daily count, still differs in its essential shape. The model has an exponential decline in infectives, while the data shows a more linear decline.

It should be noted that, were we able to craft the best relationship between the parameters β and $\gamma(t)$, our model would demonstrate what we see in Fig. 13.

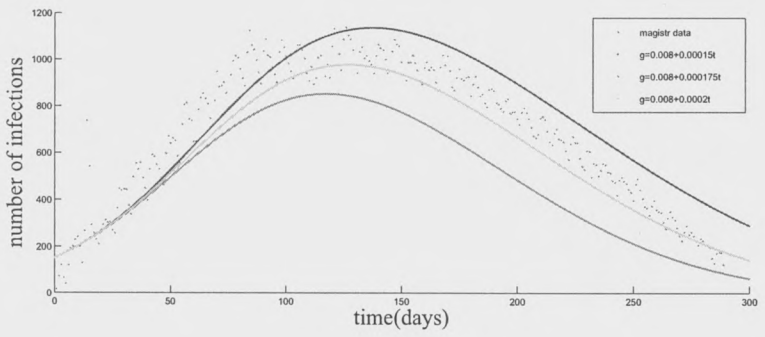

Figure 11: Magistr.b daily count fit by the number infected in the SIS model using time-dependent $\gamma(t)$ day^{-1} and $\beta = 0.00001$ day^{-1}.

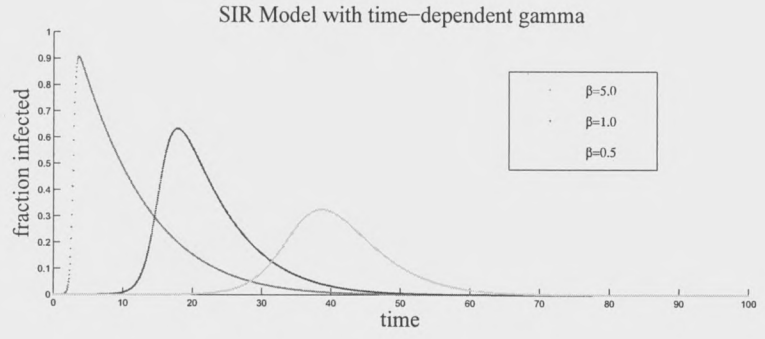

Figure 12: Time series of the number infected in the SIR model using time-dependent $\gamma(t) = e^{(0.0157t-2.4)}$ day^{-1} and varying β day^{-1}.

Without further experimentation with parameter values, it is evident that, for both the SIS and SIR models, the outbreak maximum decreases as we increase $\gamma(t)$. That is, if we can reduce the average duration faster, the outbreak is minimized.

5 Future directions

The approach of using normalized systems of ODEs did not produce models that exactly matched our data. Yet more time can be spent finding better approximations parameter values to see which normalized system, if any, provides a better model for our virus.

We were able to get qualitative fits for the Magistr.b virus with both continuous models. However, the numbers chosen for our trial model-fitting were not based on actual data. One challenge to this project was not knowing our total population size. This problem exists in biology as well, where it may not be sufficient to know the total population of hosts. The number of potential contacts would provide a more accurate assessment of total population size, N. In the case of e-mail viruses,

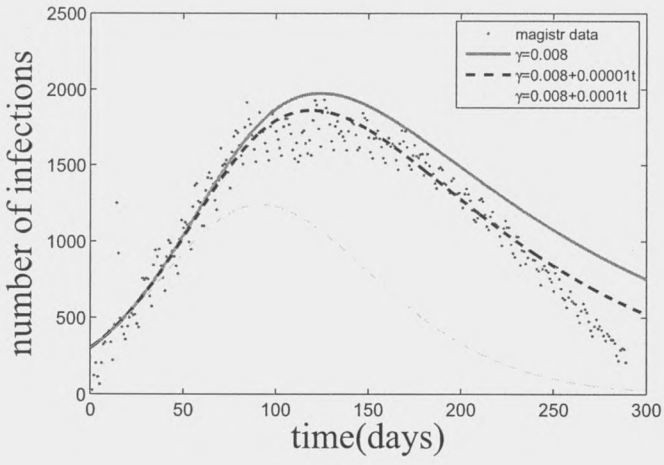

Figure 13: Magistr.b daily count fit by the number infected in the SIR model using time-dependent $\gamma(t)$ day^{-1} and $\beta = 0.00001$ day^{-1}.

the total population of computers on the Internet does not represent the susceptible class. To get reasonable figures for potential hosts for the Magistr.b, for example, it would be necessary to total the sizes of each infected computer's address book.

5.1 Agent-based Simulations

Another challenge was determining which model makes more sense for the Magistr.b virus. It is known that the virus can destroy sectors of the hard drive and erase the cmos/bios, thus rendering the computer, not recovered, but removed, which in the SIR model is the same class. On the other hand the data show significant lags in time for some IP addresses sending infected e-mails, which could imply recovery and reinfection, the scenario set forth by the SIS model. It would be worthwhile to explore another type of modeling approach that could simulate an epidemic that includes both SIS and SIR behavior. The discrete modeling approach of agent-based simulations is a good candidate for this research.

Preliminary work done for this project involved developing the Matlab programs to simulate the spread of disease in discrete time steps. We began by assigning values to an array that represents the initial state of the population. A value of zero represents an individual who is susceptible and a value of one represents an individual who is infected. In the SIS model, these are the only two conditions for any individual in the population. The number of elements in the array, N, is the total population.

We start by evaluating each element of our initial state for its value. For each element, $x(i)$, $i = 1, 2, ...N$, a value of zero prompts us to move to the next element. If $x(i) = 1$, we must now determine whether or not this infective individual will come in contact with any of the susceptibles in the array and infect that susceptible. To do this, we generate a random number, between zero and one, and choose a threshold

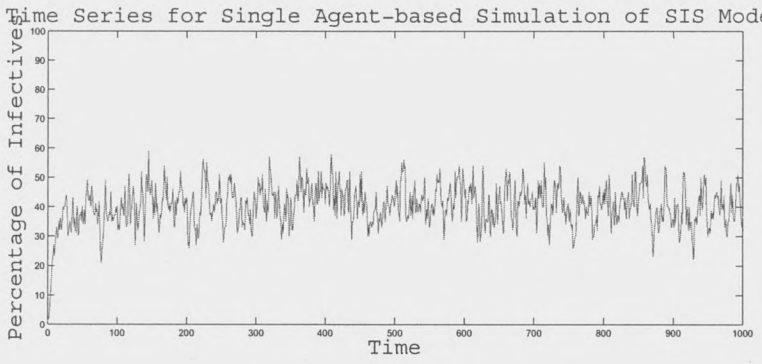

Figure 14: Time series of one simulation, 1000 time steps, of agent based simulation of SIS model. Parameters are set to $N = 100$ computers, $\beta = 0.8$ day^{-1} and $\gamma = 0.4$day^{-1}. The initial state is $I = 1$, or one infected computer.

value that represents the probability that the susceptible comes in contact with the infected individual and that the virus is transferred, β. If the random number is less than β, the susceptible individual becomes infected and we assign it the value of one in the next time step. Otherwise, the element remains zero in the next time step. We then move on to assess the probability of this infective coming in contact with and infecting the next susceptible in the array.

In addition to having the potential to infect susceptibles in the population, an infective in the SIS model also has the potential to become healed, and thereby becoming susceptible. This step involves generating another random number and choosing another threshold value, γ. If the random number is less than γ, the $x(i) = 1$ becomes a zero in the next time step. Otherwise it remains infective. Continuing this process, we find the next infective in the array, and, if one exists, determine its impact on the next state of the system.

Once the next state is determined the process is repeated and a series of time steps results. We can then take the sum of infectives at each time step and plot a time series, Fig. 14, that shows the neighborhood of the endemic state. While the number of infectives bounces around from one time step to the next, we can see that the average number of infectives is approximately 41% of the population, which is within $\sqrt{N}$ of the predicted value for the ODE of 50% ($i = 0.5$). The differential equation model produces a solution curve that is smooth since the system is solved for a population of infinite size, i.e., the solution is based on the limit as N approaches infinity. The agent-based simulation, using a small fixed population, produces an oscillating plot which varies around the smooth curve in Fig. 1.

We can generate a histogram to see how many time-steps in our simulation match the predicted endemic state. Figure 15 shows the results of one run of 1000 time steps, taking the last 90% of the time steps. The mean value of infectives should be fairly close to the predicted value $I = N(1 - \gamma/\beta)$. Running a large number of these simulations, we can take average number of infectives at each time step and plot a time series which can then be compared to the resulting time series

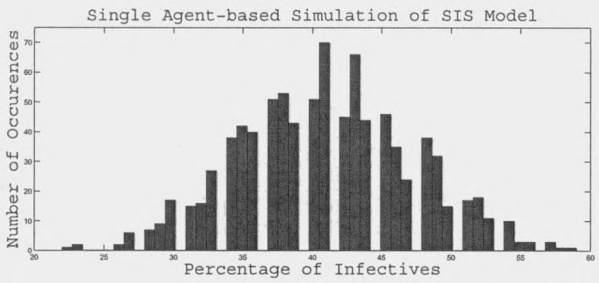

Figure 15: Histogram of agent based simulation of SIS model in Fig. 14. The mean percentage of infective individuals in this run is 41%, which is within $\sqrt{N}$ of the predicted value of 50% ($i = 0.5$).

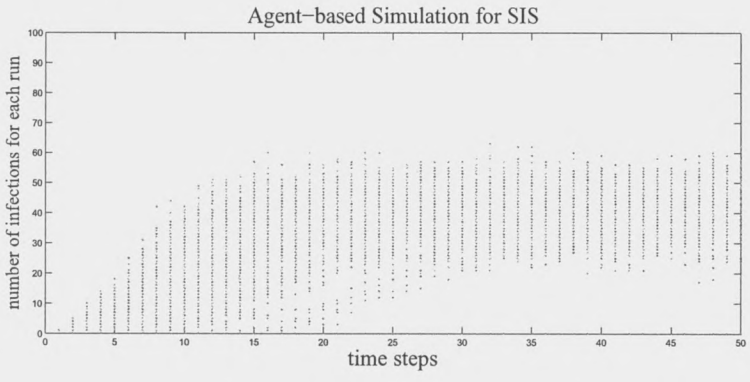

Figure 16: Agent based simulation of SIS model with initial conditions $S = 99$ computers and $I = 1$ computer and parameter values $\beta = 0.8$ day^{-1} and $\gamma = 0.4$ day^{-1}. Note the endemic state.

of the real data. We can repeat the experiment using an agent based SIR model, as shown in Fig. (17).

Future research with these simulations could incorporate a time-dependent $\gamma(t)$ by increasing the threshold for recovery with each time step. In addition, these simulations are ideal tools to test a new hypothesis that both SIS and SIR behaviors make up the dynamics of the Magistr.b virus. We saw in the raw data for Magistr.b that more than half of the hash numbers were detected sending an e-mail only once. These computers would constitute the removed class, R, of the SIR model. The rest of the computers would be classified as the susceptibles, S, of the SIS model. The simulation program could be modified to place a certain percentage of infectives into the recovered, or removed, class while the rest enter the susceptible class. This proposed compartmentalized model may show us behavior that accounts for the linear ascent and decline of the virus, both of which could not be captured entirely by either the SIS or SIR model.

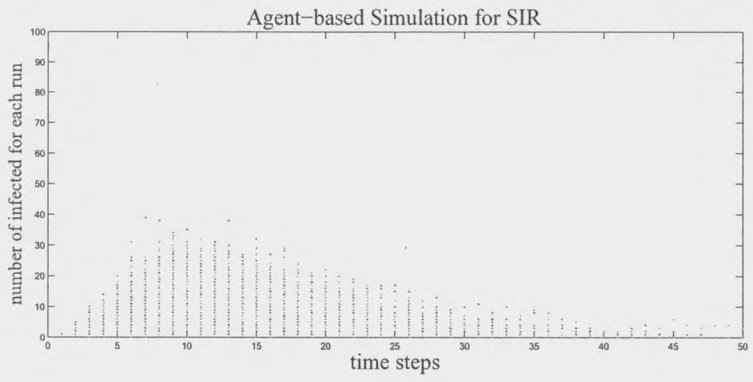

Figure 17: Agent based simulation of SIR model with initial conditions $S = 99$ computers and $I = 1$ computer and parameter values $\beta = 0.8$ day^{-1} and $\gamma = 0.4$ day^{-1}. The simulation follows the trend of the SIR ODE model to a disease free equilibrium.

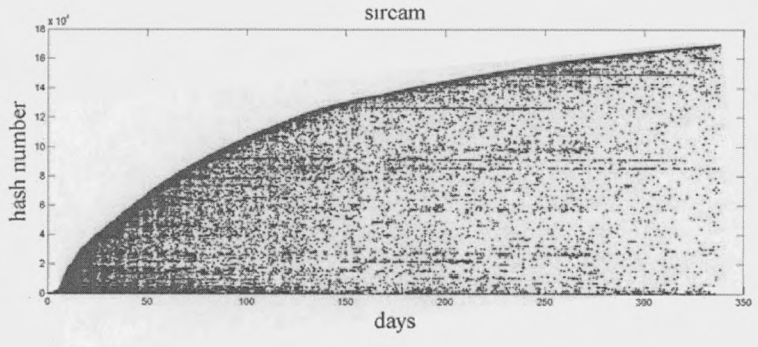

Figure 18: The detections reported in the Sircam.a virus data.

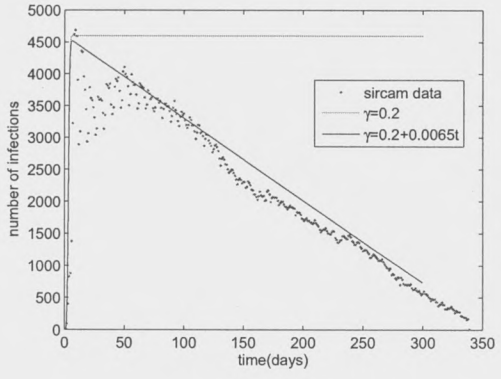

Figure 19: Sircam.a daily count fit by the number infected in the SIS ODE model. The model uses a fixed population of 5000 computers with an initial condition of 2 infectives, $\beta = 0.0005$ day^{-1} and $\gamma = 0.2 + 0.0065t$ day^{-1}.

5.2 Other Data Sets

To gain more insight into the art of modeling this virus, it would be helpful to analyze other data sets for e-mail viruses. Preliminary work on the Sircam.a virus revealed similar challenges to fine-tuning the continuous models. The Sircam.a virus sends itself as well as a clean document with the message "I send you this file in order to have your advice." The virus spreads when the newly infected file is sent to all addresses in the Windows address book and all e-mail addresses found in temporary Internet cached pages. It also spreads to other computers on the local LAN through unprotected network shares. Under certain conditions, the virus can completely fill or even erase the C: drive [3]. A plot of the detections reported can be found in Fig. (18).

Since the spreading mechanism is similar to the Magistr.b virus, the data set for Sircam.a is a good candidate for comparing the modeling approaches used in this project. Using the same qualitative approach that we used for Magistr.b, the SIS model provides a reasonable fit to the Sircam.a data. See Fig. (19) for a graph of the daily count overlaid by the SIS model with time varying $\gamma = 0.2 + 0.0065t$ day^{-1}. This is a work in progress. Applying this technique to the SIR model and using the agent-based simulation approach, we may come closer to understanding the dynamics of these e-mail viruses.

6 Summary & Conclusion

We analyzed sample data sets from two well known viruses, first detected in 2001. This data was limited by the time length of each set which made choosing a model challenging. The data is also a sample of the total population, being limited to detections of virulent e-mails among the users of a particular internet provider. With our attention primarily on the Magistr.b virus, we identified two continuous models, the SIS and the SIR, that could be good candidates for modeling. In the SIS model, recovered individuals return to the susceptible class. An endemic state is possible due to a perpetual cycle of infection, recovery, and re-infection. In the SIR model, the recovered individuals either become immune or die and do not return to the susceptible class. Infectives eventually run out of susceptibles to infect and we have a die-out.

In the case of computers on the internet, those with the targeted anti-virus software and those that have been destroyed by the virus fall into the recovered class of an SIR model. The Magistr.b has the ability to destroy sectors of the hard drive and erase CMOS/BIOS, rendering the computer inoperable. However, the viruses we studied had significantly long gaps between spurts of activity, which may suggest recovery from the virus and re-infection. This could be the result of having a computer cleaned but not fully protected. Or there may be considerable lag between the time of the first detection of a virus and the time it takes for specialists to perfect the anti-virus software and removal tools.

We ran programs in Matlab to elicit more information from the data, particularly

the time series for infectives, and approximations for the parameters that govern the differential equations of the models. The parameter γ is rate of recovery for an infective and is defined as 1/(duration of infection). In biology, γ is derived from an average duration of infection since most diseases take hold in a host for a set period of time. Our data, however, revealed a large standard deviation from the mean duration of infection. Upon closer analysis of the data, we saw that durations of infection decreased for computers infected later in the study. In fact the decrease was found to be steady and exponential. Possible explanations for this decline are an increase in public awareness of how to repair and protect their computers and increase in efficiency and effectiveness on the part of IT professionals to combat the virus as they learn more about it. This decrease in duration translates to an increase in γ over time.

The parameter β is the contact rate. We found Magistr.b to be a slow growing virus having a very small β, on the order of 10^{-5} for the time-dependent γ functions we used. One reason for its slow initial growth is it's mechanism for invasion. An attachment must be opened by the recipient, as opposed to simply opening the message. It is also possible that our rate of contact for this sample data set is skewed as it does not include the infected e-mails sent by each infective to computers outside the network.

The outbreak profile of the SIR made it our first choice for fitting the data. Experimenting with very small β values and both linearly and exponentially increasing $\gamma(t)$, we were able to match the profile of the SIR with our scaled data. We were able to show that introducing a linearly increasing $\gamma(t)$ flattens the typically exponential die-out of the SIR model to the almost linear decline that we see in the time series for Magistr.b. More importantly, by increasing the rate at which $\gamma(t)$ grows (by increasing the rate at which duration decreases), we can minimize the size of the outbreak and achieve die-out of the virus sooner.

The SIS model also proved to be a viable candidate once a time-dependent γ was introduced. Constant γ values produced an endemic state in the SIS model as predicted by the reproductive ratio, β/γ, and our initial conditions. By introducing an increasing $\gamma(t)$, the system reaches a point in time when the ratio passes under the threshold of one and we begin to see a die-out. We were able to show that an increasing $\gamma(t)$ eventually brings the SIS model profile into closer alignment with the data.

The Magistr.b virus has the ability to destroy some computers but not others, depending on the operating system. Therefore, it is possible that the data sets we studied recorded both SIS and SIR behaviors. More complete data would be needed to test this idea. A map of the network of contacts (who infects whom?) in our population could be used to establish parameters for discrete models using agent-based simulations. Preliminary work done here on the development of agent-based simulations of the SIS and SIR models is discussed among the future directions for this project.

In conclusion, the results of this project demonstrate that, even without achieving a precise match to a model, we were able to reveal the existence of a time-dependent γ. On a practical level, we have shown that decreasing recovery time for

infected computers is a critical approach to both minimizing the size of outbreaks and reducing the time it takes to drive an infected population to a disease-free equilibrium. Improvements in rapid response strategies on the part of IT professionals, e-mail providers, and anti-virus software providers will result in a steeper decline in duration of infection, an increasing $\gamma(t)$, and, consequently, minimized interruptions to the smooth flow of information that is so critical to the maintenance of our increasingly digitized world.

References

[1] N. Britton. *Essential Mathematical Biology.* Springer-Verlag, London, 2003.

[2] Tom Mainelli. Magistr.b. `www.pandasecurity.com/...` `/homeusers/security-info/16002/Magistr.B/`, cited March 16, 2001.

[3] Marc Mazuhelli. Sans institute reading room. `www.sans.org/...` `/virus-worm-lessons-learned-sircam-code-red-university-environment-57`, cited August 2001.

[4] Symantec staff. A - z list of all threats and risks. `www.symantec.com/...` `/security_response/threatexplorer/azlisting.jsp?azid=W`, cited December 20, 2001.