SIR Modeling in the Dissemination of Communicable Computer Malware

Daniel Lee, Steven Li, William Li

MATH-2552-QHS

Georgia Institute of Technology
Dr. Yaofeng Su
Spring 2023

1 Introduction

The advent of the digital age has brought with it an environment prone to the propagation of computer malware by malicious individuals. Computer malware is a form of software that contains and runs code that accesses, modifies, and destroys files on a computer (among other actions), all while duplicating itself to be spread to other computers via connections on a network (CISA). Often termed "viruses" due to their replicative and communicative nature, malware can pose significant danger to file integrity and personal privacy.

The two types of malware that will be discussed are worms and trojan horses. Worms are defined as malware that can spread from device to device on their own through local network connections, using code to write themselves onto other computers, all without human intervention. Trojan horses are defined as packages that deceive users into downloading software or opening attachments (commonly through email) that contain hidden malware and in turn spreads the trojan to other computers (CISA).

The danger of computer malware escalates as computing devices become more accessible and are adopted by more people around the world. The most successful of viruses have produced billions of dollars in digital damage. While modern technology has devised protections against the spread of viruses, they are still worth investigating for possibilities of stronger malware in the future.

This report seeks to understand the nature of the spread of worms and trojans through SIR differential equation modeling, which involves tracking the susceptible, infected, and recovered populations and their changes over time. This will be applied to susceptible, infected, and recovered computers. The question of whether these two models produce different patterns or trends in spread will be investigated and compared to real world data to assess the trueness of the model to reality.

2 Analysis

While traditionally used to model the spread of diseases, the SIR modeling method carries over to computer viruses, with modifications to the standard SIR system of differential equations. The standard model is written below:

$$\frac{dS}{dt} = -b\frac{S(t)}{N}I(t)$$

$$\frac{dI}{dt} = b\frac{S(t)}{N}I(t) - kI(t)$$

$$\frac{dR}{dt} = kI(t)$$

This is a system of three nonlinear ordinary differential equations. In this model, $S(t)$, $I(t)$, and $R(t)$ are the number of susceptible, infected, and recovered individuals at time t. b is the number of individuals that a single infected individual can infect per unit time (from all groups), N is the total population, and k is the static proportion of infected individuals that recovers per unit time (Smith and Moore).

2.1 Model and Assumptions

Worms and trojans spread through fundamentally different mechanisms, so different models must be developed for each case.

2.1.1 Worms

Since computer worms can spread on their own without human intervention, if the user does not know about it, there will be no barriers to its spread. Their spread will be similar to the spread predicted by the standard model:

$$\frac{dS}{dt} = -b\frac{S(t)}{N}I(t)$$

$$\frac{dI}{dt} = b\frac{S(t)}{N}I(t) - kI(t)$$

$$\frac{dR}{dt} = kI(t)$$

This model relies on several assumptions. It is assumed that users do nothing to prevent their computers from acquiring the virus, which is reasonable given that worms operate independently and sometimes are disguised. The model will assume that each infected computer has a fixed number of contacts per unit time that is composed homogeneously of susceptible, infected, and recovered computers. It assumes that all computers have no protections against these worms (which would have been true a few decades ago) and that every computer will eventually get infected.

2.1.2 Trojans

Trojan viruses operate slightly differently and require a modification on their model. While there are different kinds of trojans, the one investigated here will be the type that involves an email with an attachment that, upon opening, will send the same email to other individuals in the email's mailing list. In this case, the user has the option to open the attachment or not.

Suppose the user of a computer in the susceptible population hears about the virus and as a result will choose not to open the email attachment. The probability that any individual with a susceptible computer learns about the virus depends on the number of infected and recovered individuals because with more infections/recoveries, there will be greater public knowledge. In the act of learning about the virus, the user will not open the email and the susceptible computer becomes a recovered computer. The term $c(I+R)$, where c is a fractional constant dictating the number of people that will be informed of the virus per infection or recovery per unit time, will be subtracted from the susceptible population and added to the recovered population.

This results in the following model for trojans:

$$\frac{dS}{dt} = -b\frac{S(t)}{N}I(t) - c(I+R)$$

$$\frac{dI}{dt} = b\frac{S(t)}{N}I(t) - kI(t)$$

$$\frac{dR}{dt} = kI(t) + c(I+R)$$

This model operates under another set of assumptions. The model assumes that each infected computer has a homogeneously composed fixed number of contacts per unit time. It assumes that all computers have no protections against trojans (which would have been true a few decades ago). It also assumes that infected and recovered individuals will inform a consistent proportion of susceptible individuals per unit time.

2.1.3 Methods of Analysis

Euler's method will be used to approximate the solutions to the curves, and solution plots will be used to observe the changes in the population composition over time. This will first be performed in a hypothetical population of a million computers for the worm model and the trojan model to compare patterns and trends; then, the models will be employed against real data to assess their accuracy. To use Euler's method, variables within the model and initial conditions must be defined.

For the hypothetical worm model, it will be assumed that a single computer in a population of one million computers begins with the worm and that 40% of infected individuals recover per unit time.

$$S(0) = 999999$$

$$I(0) = 1$$

$$R(0) = 0$$

$$N = 1000000$$

$$b = 10$$

$$k = 0.4$$

The same conditions will be assumed for the trojan model, in addition to the condition that users of infected and recovered computers inform 5 people per day about the trojan.

$$S(0) = 999999$$

$$I(0) = 1$$

$$R(0) = 0$$

$$N = 1000000$$

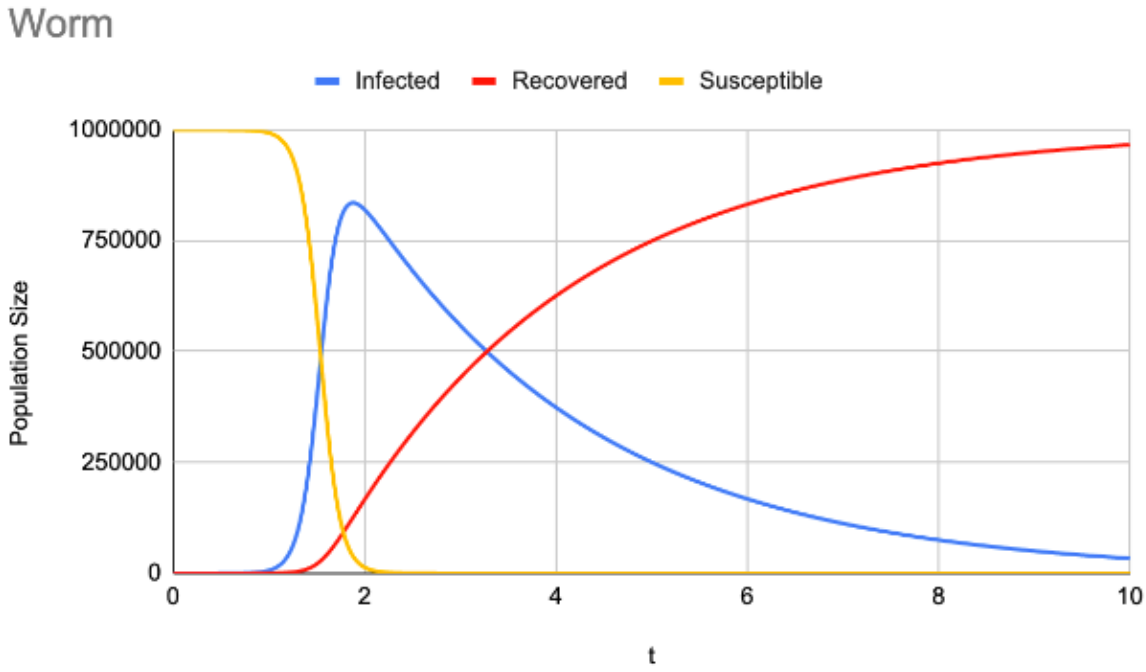
$$b = 10$$

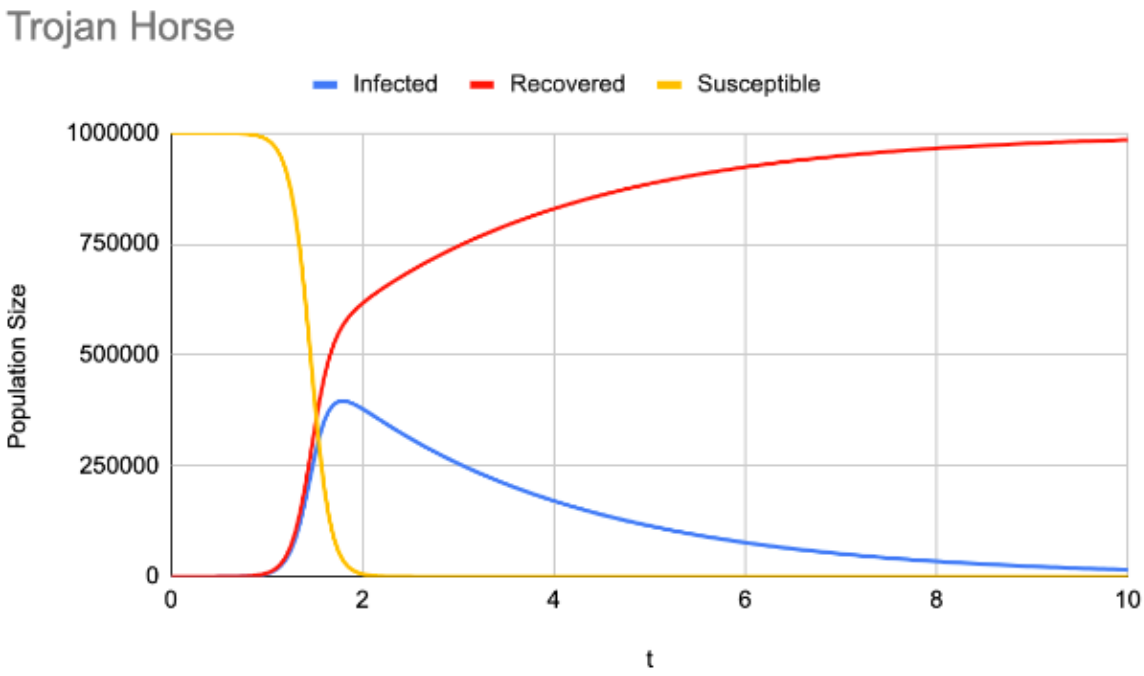
$$k = 0.4$$

$$c = 5$$

2.2 Results

The changes over time in the hypothetical population of computers can be seen in the solution plots. Euler's method was used with the above conditions with step size $t = 0.01$ for both the worm model and the trojan model. The code used to generate the plots in the hypothetical model can be found in the Appendices A and B, respectively.





The difference between the worm trajectory and the trojan trajectory was minor but notable. For every t, the value of the infected curve is higher in the worm group than the trojan group, with the worm group having a higher absolute peak. The rate of increase of the recovered population is also lower in the worm group compared to the trojan group. These two observations make sense, given that the only difference was that some members of the susceptible population in the trojan group were transferred directly to the recovered group. Thus, the quantity of infected computers significantly decreases with greater knowledge and awareness of trojan malware, which could serve as an effective method.

2.2.1 Real World Application

When applying the model to real world data, the accuracy of the model begins to break down. The worm and Trojan model was applied to the CodeRed and the Melissa viruses, respectively.

CodeRed was a buffer overflow worm that spread itself to 359000 computers in 14 hours with no human intervention (Moore and Shannon). Melissa was an email attachment trojan virus that spread to roughly 100000 computers in the span of 3 days in 1999 (United States). These measures will serve as real world trajectories that the model will attempt to predict

There were no available measures of the true values of the initial conditions, so instead, initial conditions were modified to produce those real world trajectories. The value $N = 170000000$ is based on the numbers of computers connected to the network in 2000, which is around when both viruses emerged. The following conditions roughly produced the spread pattern of CodeRed:

$$S(0) = 170000000$$

$$I(0) = 1$$

$$R(0) = 0$$

$$N = 170000000$$

$$b = 1.36$$

$$k = 0.4$$

And the following conditions roughly produced the spread pattern of Melissa:

$$S(0) = 170000000$$

$$I(0) = 1$$

$$R(0) = 0$$

$$N = 170000000$$

$$b = 1.36$$

$$k = 0.4$$

$$c = 0.15$$

Clearly, these conditions are not the only combinations that could produce the results, but it is evident that they are all far from the true initial conditions; the values for each variable are unreasonably small to be valid. To improve it as a predictive model, additional factors should be included and more accurate initial conditions should be obtained.

3 Conclusion

SIR models are a common way of demonstrating the spread of communicable entities through a population. While the model developed had predictive capacity, it was effective as a tool to conduct hypothetical comparisons between different forms of computer malware. It was observed with a hypothetical model that the spread of email-based trojan viruses can be significantly inhibited with knowledge of how the virus operates compared to worms that are harder to detect naturally, decreasing the quantity of computers suffering from infections.

This method produced a multitude of limitations. One major limitation is that values for b, k, and c are difficult to predict. Furthermore, Euler's method produces a degree of error due to finite step sizes, causing an overestimation in the susceptible curve, and underestimation in the recovered curve, and both in the infected curve. Finally, there are countless other factors that influence the rate and reach of computer virus spread that are not encompassed by the model, producing inaccurate predictions.

The development of technology may produce even more dangerous viruses in the future, but differential equation models will be an invaluable aid in combating these malicious entities.

4 Appendix A

```
//WORM EULER'S METHOD
class Main {
  public static void main(String[] args) {
    double susceptible = 999999;
    double infected = 1;
    double recovered = 0;
    double t = 0;
    for (int i = 0; i < 1000; i++){
      double sChange = susceptibleChange(susceptible, 0.01, infected);
      double iChange = infectedChange(infected, susceptible, 0.01,
infected);
      double rChange = recoveredChange(recovered, 0.01, infected);
      susceptible+=sChange;
      infected+=iChange;
      recovered+=rChange;
      t+=0.01;
      // if (i % 5 == 0){
      //   System.out.println();
      // }

    }
  }
  public static double susceptibleChange(double initial, double step,
double inf){
    double s = initial;
    s = step*(-b*(s/1000000)*inf);
    return s;
  }
  public static double infectedChange(double initial, double sus,
double step, double inf){
    double i = initial;
    i = step*(b*(sus/1000000)*inf - k*inf);
    return i;
  }
  public static double recoveredChange(double initial, double step,
double inf){
    double r = initial;
    r = step*(k*inf);
    return r;
  }
}
```

5 Appendix B

```
//TROJAN EULER'S METHOD
class Main {
  public static void main(String[] args) {
    double susceptible = 999999;
    double infected = 1;
    double recovered = 0;
    double t = 0;
    for (int i = 0; i < 1000; i++){
      double sChange = susceptibleChange(susceptible, 0.01, infected,
recovered);
      double iChange = infectedChange(infected, susceptible, 0.01,
infected);
      double rChange = recoveredChange(recovered, 0.01, infected,
susceptible);
      susceptible+=sChange;
      infected+=iChange;
      recovered+=rChange;
      t+=0.01;
      // if (i % 5 == 0){
      //   System.out.println(inf);
      // }
    }
  }
  public static double susceptibleChange(double initial, double step,
double inf, double rec){
    double s = initial;
    s = step*((-b*(s/1000000)*inf - c*(s/1000000)*(inf + rec)));
    return s;
  }
  public static double infectedChange(double initial, double sus,
double step, double inf){
    double i = initial;
    i = step*(b*(sus/1000000)*inf - k*inf);
    return i;
  }
  public static double recoveredChange(double initial, double step,
double inf, double sus){
    double r = initial;
    r = step*((k*inf) + c*(sus/1000000)*(inf + r));
    return r;
  }
}
```

6 Works Cited

CISA. "Virus Basics: CISA." Cybersecurity and Infrastructure Security Agency, 17 Mar. 2023, https://www.cisa.gov/news-events/news/virus-basics.

Moore, David, and Colleen Shannon. "The Spread of the Code-Red Worm (CRV2)." CAIDA, 30 July 2020, https://www.caida.org/archive/code-red/coderedv2_analysis/.

Smith, David, and Lang Moore. "The Sir Model for Spread of Disease - the Differential Equation Model." Mathematical Association of America, Convergence, Dec. 2004, https://www.maa.org/press/periodicals/loci/joma/the-sir-model-for-spread-of-disease-the-differential-equation-model.

United States. General Accounting Office. "Information Security: The Melissa Computer Virus Demonstrates Urgent Need for Stronger Protection Over Systems and Sensitive Data," General Accounting Office, Apr. 1999.