SIR modeling of computer malware endemics

A stochastic model of how malware spreads in a homogenous population.

Nils Hallerfelt, Noah Åkesson, Isak Källsmyr
Supervised by Professor Mario Natiello

1 Abstract

This report models the spread of computer malware in a limited and closed network of computers with the help of various SIR models, specifically the SIR, SDIR and SDIR-V versions, the latter of which introduces a form of vaccination against the virus. To do this, Markov Jump processes were used in form of the Kendall algorithm of infection propagation. It is found that the SIR-type models work quite well for this purpose, with results looking as one might expect over larger periods of time. It is also clear that the vaccination element in the SDIR-V model drastically reduces the number of infected computers. Additionally, an economic analysis is made evaluating the cost of damages caused by the malware against the cost of vaccination. This has found that unless the damage of the malware is very limited, the most cost-effective action is to use vaccination.

Lunds universitet
Lund, Sweden
February 2022

Contents

2 Introduction

Malwares, or computer viruses, are harmful programs that are purposefully made to cause harm, by either stealing personal information, passwords, bank information or disrupt society. The first ever definition of computer viruses was coined by Fred Cohen in 1986 as 'A program that can infect other programs by modifying them to include a, possibly evolved, version of itself' (John Love 2018). In monetary terms malwares are projected to globally have caused 6 Trillion dollars of damages by 2022 (Branka Valeka 2021), and this is expected to significantly increase as we become even more dependent on technology. Not only do they cause monetary damages, but also personal, in 2013 alone 3 billion email accounts were compromised (ibid.). Furthermore, and much more terrifying, in 2005 Stuxnet was developed in a collaboration between the USA and Israel, with the goal of disrupting Irans nuclear program. Beginning to spread in 2010 it has since then been responsible for the destruction of one-fifth of Irans nuclear centrifuges, by targeting PLCs (programable logic controllers) (Kelley 2013). The potential harm that could be caused if hostile nations or organizations had access to such destructive tools is quite worrisome, being able to paralyze infrastructure or in the worst case cause nuclear meltdowns. As such it is paramount that people are aware how these spread and work.

Computers are prone to infections much like humans, although they might at first glance seem different from how diseases work in humans, but in reality they are much alike in how they spread. Typical human infections start by susceptible humans being exposed to an infection, this put them in a delitescent state, which might result in them recovering or getting infected. Infected individuals might then infect other susceptible humans. There are then two outcomes, either the human dies or they recover, which temporarily makes them immune to the new infections. This is the same way a computer virus spreads, being able to get infected, then continue spreading it to other computers, then recover by installing anti virus software and then the virus evolving causing the computer to once again be susceptible. However, one key difference is that a computer can never truly die so to speak, but can always be restored to its original state as long as the hardware is intact. Knowing how this process works and how to prevent it is crucial in preventing devastating malwares from spreading.

In this project we aim to model the spread of malicious malwares using Kendall simulation algorithms using Python, under various different circumstances. Kendall simulations are stochastic and hence differ from the more commonly used differential equation models.

2.1 Purpose

The purpose of this report is to examine whether various versions of the SIR model of infection can be used to study the spread of malicious malware in a limited network of computers. Three different models will be studied, the simple SIR model, and the two modifications of it known as SDIR and SDIR-V. These models will be simulated, and the results are presented below. Additionally, a brief economic analysis is made regarding how the cost of damages from the virus compares to the cost of so called vaccination.

2.2 Problem formulation

- Can a SIR model be used to model the spread of computer malware within a contained network?
- How is the SIR model affected if an element of vaccination is introduced to protect the individuals in the model from the malware?
- How does the cost of such a vaccination compare to the difference in cost between a vaccinated population and an unprotected population?

2.3 Literature

When writing this rapport on how malware spreads in closed homogeneous systems some articles and lecture notes were used for reference. Lecture notes by Mario A Natiello, 'Kendall simulation algorithm' (2014) were used to grasp the main concepts of stochastic population dynamic simulations. However, the primary article used was the 'Stability analysis of VEISV propagation modeling for network worm attack' by Ossama A.Toutonji Seong-MooYoo and Moongyu Park (2012) (Toutonji, Yoo, and Park 2012), where we got our parameter values from and some ideas on how malwares spread in computers. Further more, an article from Mario A. Natiello and Hernán G. Solari about 'Modelling population dynamics based on experimental trials with genetically modified (RIDL) mosquitoes' (2020) were also used in our research into how to implement the Feller-Kendall algorithm. (Natiello and Solari 2020).

3 Method

To implement the SIR, SDIR, SDIR-V models discrete-time-simulation has been used and implemented with the help of the programming language Python. The implementation is based on the theory of Kolomogrov Forward Equations (also known as Markov Jump Processes). Starting out simple with implementing the SIR-model and the different parameters and events for this, the implementation has been developed into SDIR and SDIR-V models by adding parameters and events. Thus the model for the SDIR and SDIR-V are more complex, with more parameters and events.

When the implementation was complete the work shifted focus to finding sufficiently good values for the different parameters for respective model. To a large part this work consisted of literature study. After sufficiently good parameter values had been found, the simulation process begun. Since the implemented process is stochastically random, 100 simulations were done for each model. The data obtained from this was then used to find the "average simulation" for the SIR, SDIR, SDIR-V models, respectively, by taking the average population value of infected. susceptible, infected etc. for each discrete time-step.

3.1 The Algorithm

To model the infection spread the simulation algorithm created by Kendall was used. It consists of iterating through the following steps:

- Place yourself immediately after an event has occurred and compute the new populations.
- Compute $R = \Sigma_\alpha W_\alpha(X)$. Generate an exponentially distributed time T for the next event, with parameter R. Where W_α is the probability of an event 'a' happening per time unit.
- Generate a uniformly distributed number s in $[0, R]$ and pick which event has occurred as follows: if $\Sigma_{\alpha=1}^{\beta-1} W_\alpha(X) \leq s \leq \Sigma_{\alpha=1}^{\beta} W_\alpha(X)$ we say that event β has occurred.
- Update time, events and populations, and repeat the process.
- If final time has been reached, stop.

3.2 Populations

To model the spread of the virus in question we will have to divide the "population" of computers into a few groups, depending on how they've been affected by the endemic.

- Susceptible computers (S)
- Delitescent computers (D)
- Infected computers (I)

- Recovered computers (Re)
- Vaccinated computers (V)
- Total number of computers (N)

The variable S represents all computers who have not been exposed to the virus and have no protection against it. Initially, most computers are in this group. The variable D represents the computers who have recently been exposed to the virus, but who have not yet been infected. It would require user action in the form of clicking on some malicious link to be infected. The computers who have been infected by the virus and whose functionality has been negatively affected by it, and are capable of spreading it to other computers are represented by I. Initially, there are only a few computer in this group, and these are patient zero. All the computers who have "recovered" from the virus are represented by Re, meaning that they have been restored to their normal state and can no longer spread the virus. The variable V represents the computers that have acquired some sort of protection against the virus, for instance an antivirus program. Finally, N is the total amount of computers, or the sum of all the above groups.

3.3 Economic model

The report will also evaulate how large of an impact will be created by introducing an element of vaccination, as per the SDIR-V model, in contrast to the population having no protection against malware. This will be done by comparing the cost of damages generated by the malware against the cost of vaccination of a large section of the population. More specifically, a threshold of damage generation will be found at which it is beneficial for an individual to pay for protection against the malware rather than taking the risk of being exposed to it. This is due to the vast variation of severity of damage that different malware may cause, which leads to an estimation of the economic cost of these damages being rather substantial in work load and thus outside the scope of this report.

3.4 Events

In the duration of the process some events will occur that radically change the dynamic of the processes. At the beginning the process will be in a stable state with almost all computers being susceptible. This situation is modified when some external factor changes, and the malware is introduced into the process, directly infecting some computers. This could be done using direct email, infected USB sticks or a infected link posted on the likes of Facebook or Twitter. These events are shared in all three different models, however add

This is the same for all our three different models, however, after that they have different events that can happen.

3.4.1 SIR events

After the virus is introduced, there are just a few events that can happen. Infected (I) computers can spread the malware to susceptible (S), resulting in (S) going to (I). This event is affected by α, which is the chance of infection for susceptible computers. The other event that can happen is that infected (I) computers recover (Re), (I) going to (Re). This in turn is affected by the recovery rate, γ. In the simplest SIR model, there are no other events.

- Computer is infected, moving from S to I, determined by $(\frac{\alpha S(t)I(t)}{N})$
- Computer recover from infection, moving I to Re $(\gamma I(t))$

The populations and events of the SIR model can be seen in figure 1. The parameter values used for the SIR simulation are shown in table 1, the values were taken from (Toutonji, Yoo and Park, 2012).

Parameter	Value	Notes
S(t)	$100,000$	Number of susceptible computers
I(t)	100	Number of infected computers
Re(t)	0	Number of recovered computers
α	0.8	Chance of infection when susceptible
γ	0.5	Recovery rate from infection

Table 1: Parameters for the SIR model

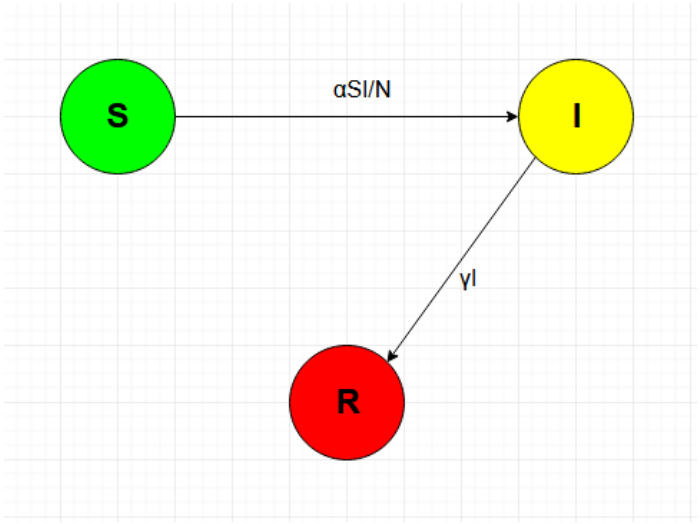

Figure 1: A visualisation of the SIR model

3.4.2 SDIR events

In the more complex SDIR model new events and one new population is introduced, the delitescent population, something in-between infected and susceptible, where individuals do not yet know whether they are infected or not. An example of this is when a computer has received an infected email but has not yet pressed the infected link, hence the outcome is determined by the action of the user. Such an event is affected by β, the contact rate between suceptible and infected computers. In the next stage there are two possible outcomes, either they become infected or they recover, the probabilities of which are denoted by α and ϵ respectively. If they become infected they start trying to infect other computers in the system, as they forward the malware to others. During the infected stage the user will notice symptoms, just like for a normal disease in humans, and take actions to recover. Once the infected computer is recovered it moves to Recovered computers (Re) and is immune to new infections. The events are summarized as the following:

- Computer is exposed to the virus, moving from S to D ($\frac{\beta S(t)I(t)}{N}$)
- User action causes virus to infect computer, moving from D to I ($\alpha D(t)$)
- User action prevent infection, computer moves from D to Re ($\epsilon D(t)$)
- User takes action to protect computer, moving from I to Re ($\gamma I(t)$)

The model is summarized by figure 2, and the parameter values used for the SDIR simulation are shown in table 2 and parametervalues from (Toutonji, Yoo and Park, 2012).

Parameter	Value	Notes
S(t)	100,000	Number of susceptible computers
D(t)	0	Number of delitescent computers
I(t)	100	Number of infected computers
Re(t)	0	Number of recovered computers
V(t)	0	Number of vaccinated computers
β	80	Contact rate between infected and susceptible computers
α	0.8	Chance of infection when delitescent
ϵ	0.2	Chance of not getting infected when delitescent
γ	0.5	Rate of recovery after infection

Table 2: Parameters for the SDIR model

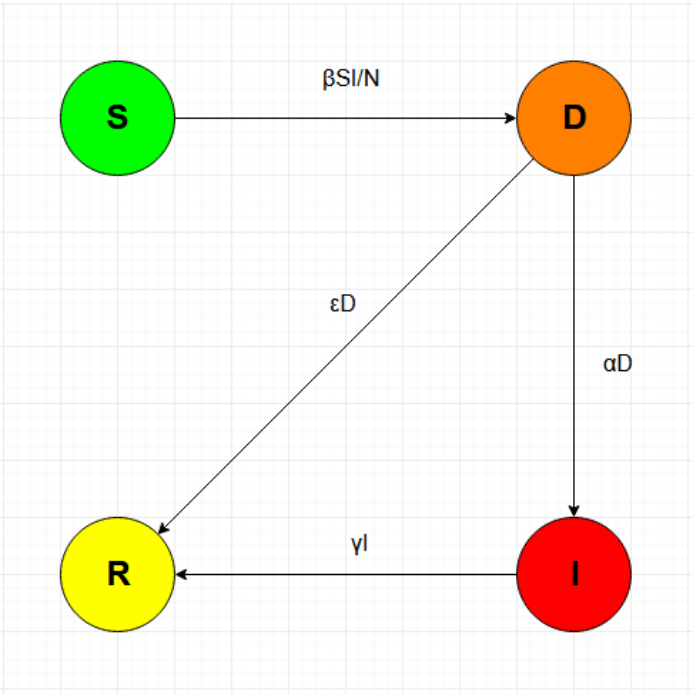

Figure 2: A visualisation of the SDIR model

3.4.3 SDIR-V events

Finally there is the most complex model where vaccination and reinfection is introduced to get another dynamic. Here a new stage is introduced where recovered computers can become susceptible again, the rate of which is denoted by ω. This is the case for computers as well, just like normal viruses that affect humans computer viruses adapt as well (Castro, Schmitt, and Dreo 2019). Furthermore, infected and susceptible computers can get vaccinated, and contrary to the protection from recovery, this vaccination is permanent. The rate of vaccination for both susceptible and infected computers is μ The events in this model are the following:

- Computer is exposed to the virus, moving from S to D ($\frac{\beta S(t)I(t)}{N}$)
- User action causes virus to infect computer, moving from D to I ($\alpha D(t)$)
- User action prevents infection, computer moves from D to Re ($\epsilon D(t)$)
- User takes action to restore computer, moving from I to Re ($\gamma I(t)$)

- The virus mutates and as such computers can be infected again, moving from Re to S ($\omega Re(t)$)
- Infected computer buys antivirus software, moving I to V, this only happens after a certain threshold of computers are infected ($\mu I(t)$)
- Susceptible computer buys antivirus software, moving S to V, this only happens after a certain threshold of computers are infected ($\mu S(t)$)

The parameter values used for the SDIR-V simulation are shown in table 3, and the SDIR-V model is visualised in figure 3. The values for the parameters were taken from (Toutonji, Yoo and Park, 2012).

Parameter	Value	Notes
S(t)	100,000	Number of susceptible computers
D(t)	0	Number of delitescent computers
I(t)	100	Number of infected computers
Re(t)	0	Number of recovered computers
V(t)	0	Number of vaccinated computers
β	80	Contact rate between infected and susceptible computers
α	0.8	Chance of infection when delitescent
ϵ	0.2	Chance of not getting infected when delitescent
γ	0.5	Rate of recovery after infection
ω	0.0005	Rate of loss of protection
μ	0.1	Vaccination rate for infected and susceptible computers

Table 3: Parameters for the SDIR-V model

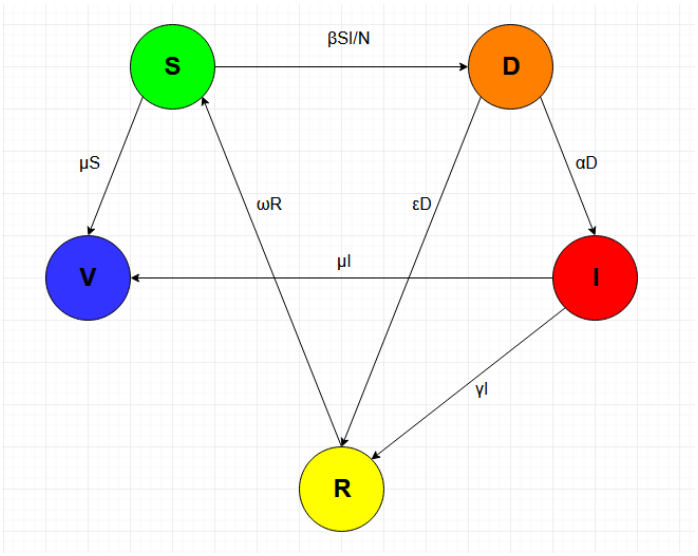

Figure 3: A visualisation of the SDIR-V model

4 Results

4.1 SIR

Figure 4 below and figure 7 in the appendix show the results of simulating the basic SIR model with the parameter values in table 1. We see that the infected and recovered poulations intitially grow together as the susceptible population shrinks. Eventually, the susceptible population becomes so small that there are not many new computers to infect, thus the infected population starts declining, and dies out rather rapidly as most infected computers move to the recovered population.

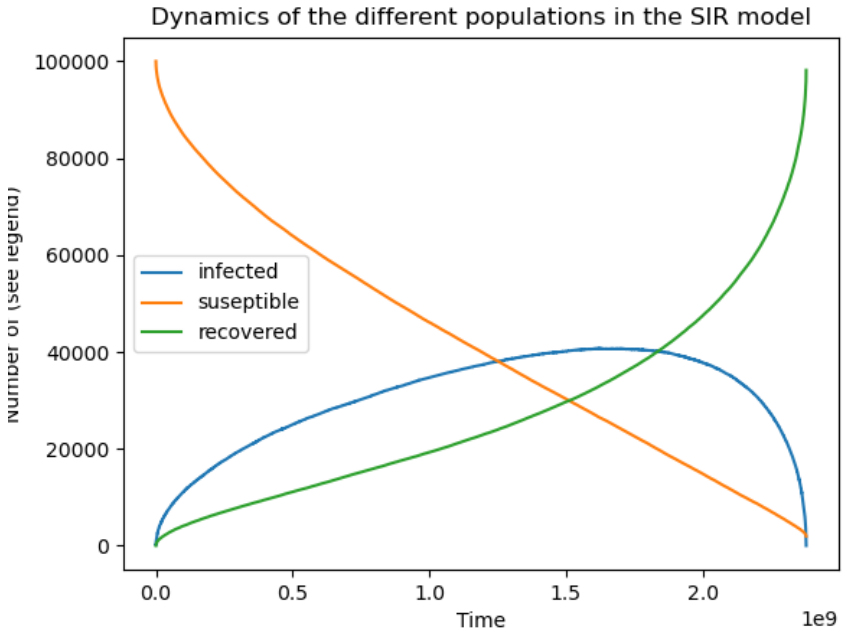

Figure 4: SIR simulation results as an average of 100 simulations

4.2 SDIR

Figure 5 and figure 8 in the appendix show the results of simulating the SDIR model with the parameters presented in table 2. Here we see that the delitescent group initially grows rapidly, as the susceptible group shrinks in reverse proportionality to it. However, the infected and recovered populations remain low until the later part of the simulation, where the susceptible group reaches zero. Then, the delitescent group starts rapidly declining and the infected and recovered populations start increasing. At this point, the entire population quickly reaches the recovered group as the infected group has a short spike before rapidly declining.

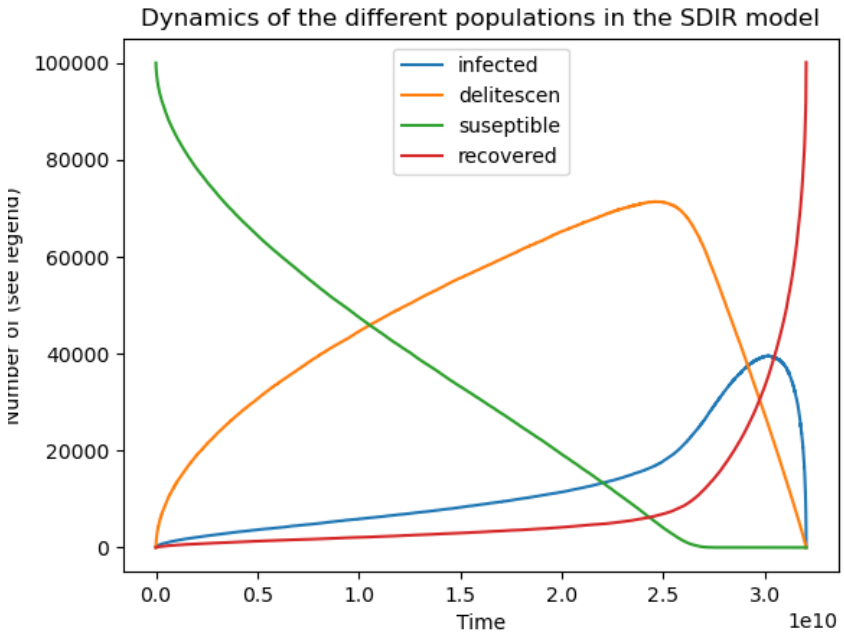

Figure 5: SDIR simulation results as an average of 100 simulations

4.3 SDIR-V

Figure 6 below and 9 in the appendix show the results of the simulation of the SDIR-V model, with the parameters for table 3. We see that this one initially behaves similarly to the SDIR model, but at a certain point in time, a vaccine is introduced, and a certain number of computers from each group receives this vaccine and moves towards the vaccinated population. Here we see that while the delitescent and recovered populations keep increasing as the susceptible population decreases, the infected population never reaches the peak it had in the previous models.

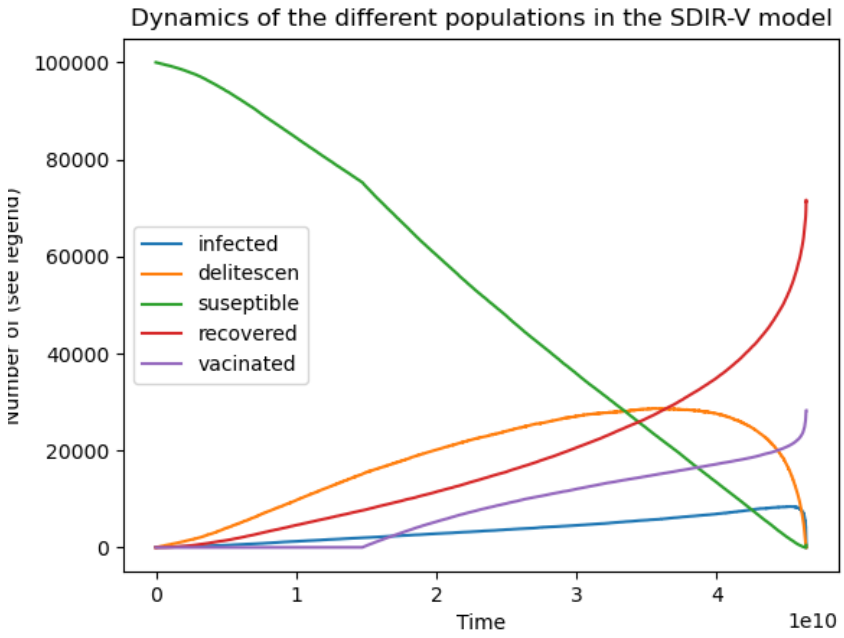

Figure 6: SDIR-V simulation results as an average of 100 simulations

4.4 Cost of malware and protection

First we want to find the average daily cost of an antivirus program, let's call this cost cv A comparison of 10 of the most popular antivirus programs show that the price ranges greatly, from 7$ per license to over 65$, but the average price was found to be $15,75$$ per license.(Priceithere.com 2022) These are monthly subscription plans, so a vaccinated computer is assumed to have an antivirus with the above average, but with the daily cost $cv = \frac{15,75}{30} = 0,525$$

To find out the total amount of damage we must know how many computers were affected by the virus, and for how long. This is done by measuring the total amount of infected and vaccinated computers in the simulation for every time unit, and adding this number for all time unit in the simulation. This was done in both the SDIR and SDIR-V models, to compare the amount of damage done. An average of three runs in each of the model was found. For the SDIR model, this average was found to be $TIa = 5400522078$. For the the SDIR-V model, the infection average was $TIb = 986580064$ and the vaccination average was $TV = 2764124203$. The a and b in this case simply denote whether the infected group was in the SDIR or SDIR-V model. Now let's assume the entire simulation covers a period of exactly one year, then we find how many computers are infected and vaccinated per day. This figure is $dIa = 14795951$ infected for the SDIR model, $dIb = 2702959$ infected and $dV = 7572943$ vaccinated for the SDIR-V model. With the above figure we can calculate the total cost of vaccination per day to be $dV * cv = 7572943 * 0.525 = 3975795$$.

There is a wide variety of damage that malware can do, and much of it is unknown and even actively hidden by its creators, thus it is difficult to create an economic evaluation of this damage for an arbitrary virus. Instead, we present a level of economic damage where using protection against malware by comparing the total cost of the damage in the SDIR model, and the sum of damage cost and cost of vaccination in the SDIR-V model. This is done with the following formula, and with the values found above we get the following result

$$dmg = \frac{dV * cv}{dIa - dIb} = \frac{3975795}{14795951 - 2702959} = 0.32877\$ \tag{1}$$

If the damage being done to the computer in question is greater than this figure, it would be economically beneficial to take the necessary steps to protect it.

5 Discussion

5.1 Using stochastic SIR-models to simulate malware propagation

Although it is difficult to motivate if SIR-models are suitable or not for simulating malware propagation without having real-world data to compare with, some conclusions can be made. To begin with, we found that the spread of malware is not much unlike the spread of infections in biological populations. Most, if not all, of our parameters, population groups, and events have an heritage in how infections spread in biological populations. And with this approach we received reasonable results. If this approach is "the best", we cannot answer. On one hand it feels intuitive, but on the other hand it may have resulted in misconceptions. That is misconceptions about how events in the infected network occurs, and how the population reacts in network environments, thus making the model insufficient.

5.2 Vaccination in SIR models of networks

As mentioned in the section above, we have used a vaccination approach much similar to how it might have worked in biological populations. The result generated showed, not unsurprisingly, that vaccinating the population reduced the number of infected individuals. This is linked to a lesser direct cost, as fewer individuals are lost to the malware. However the cost of producing vaccines must be accounted for. With our much simplified economic model it showed to be cost-effective to produce vaccines in most cases, however if the damage caused by the malware is very limited, it might be beneficial to instead take the risk of exposing yourself to such malware rather than pay the cost of an antivirus program. To find an "optimal strategy" for when to vaccinate and how much, where producing tempo of vaccines increases cost, is a substantially more difficult question, and outside of the scope for this report.

There is one major flaw in the economic analysis, and that is the assumption that the duration of the simulation is that of one year. There is no real way of knowing the timespan of the simulation, as the units of time are completely arbitrary and non-uniformally distributed. The task of finding how long such an endemic might last proved more difficult than anticipated and thus, due to time constraints, was never finished.

5.3 General thoughts about the modeling process

During the report different types of models were presented, and some general conclusions about using SIR-models to simulate the propagation of malware in homogeneous networks can be drawn. To begin with, it has become obvious that the propagation of the introduced malware differs largely based on what model is used to simulate the process. The group state "infected individuals" is included in all of the examined models, however the group behaves differently. The total number of infected individuals (the "integral" of the infected curve) varies largely between models, in the original SIR model it is much grater than in both the SDIR and SDIR-V models. Furthermore the SDIR-V model generates the least amount of infected individuals, maybe not so surprisingly. Thus it is of utter importance to choose a model that is a "good enough" representation of the real world. Including many parameters and events may ideally be a better representation of how the world behaves, but it must be taken into consideration that finding suitable values for parameters is hard, and can almost never be done without any error. This brings an element of insecurity into the models that one must be aware of.

References

Branka Valeka (2021). *44 Worrying Malware Statistics to Take Seriously in 2022*. URL: https://legaljobs.io/blog/malware-statistics/ (visited on 02/18/2022).

Castro, Raphael Labaca, Corinna Schmitt, and Gabi Dreo (2019). *AIMED: Evolving Malware with Genetic Programming to Evade Detection*. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8887384 (visited on 02/18/2022).

John Love (2018). *A brief history of Malware - Its evolution and impact*. URL: https://www.lastline.com/blog/history-of-malware-its-evolution-and-impact/ (visited on 02/18/2022).

Kelley, Michael (2013). *The Stuxnet Attack On Iran's Nuclear Plant Was 'Far More Dangerous' Than Previously Thought*. URL: https://www.businessinsider.com/stuxnet-was-far-more-dangerous-than-previous-thought-2013-11?r=US&IR=T (visited on 02/18/2022).

Natiello, Mario A. and Hernán G. Solari (2020). *Modelling population dynamics based on experimental trials with genetically modified (RIDL) mosquitoes*. URL: https://www.sciencedirect.com/science/article/pii/S0304380020300582 (visited on 02/18/2022).

Priceithere.com (2022). *How much does antivirus software cost in 2022?* URL: https://priceithere.com/antivirus-software-cost/ (visited on 02/18/2022).

Toutonji, Ossama A., Seong-Moo Yoo, and Moongyu Park (2012). *Stability analysis of VEISV propagation modeling for network worm attack*. URL: https://www.sciencedirect.com/science/article/pii/S0307904X11006172 (visited on 02/18/2022).

6 Appendix

6.1 Figures

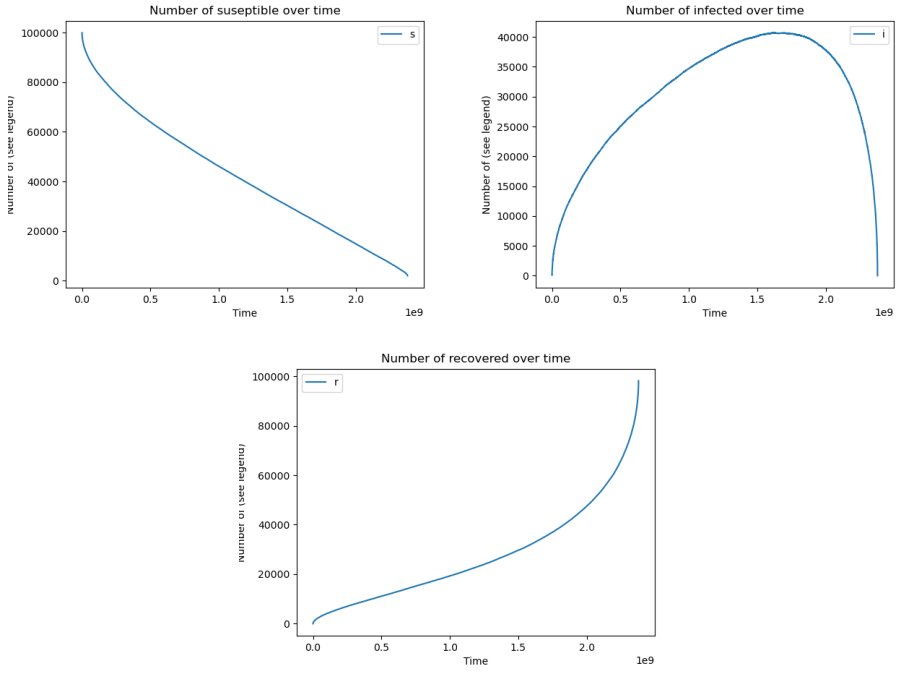

Figure 7: SIR simulation per population

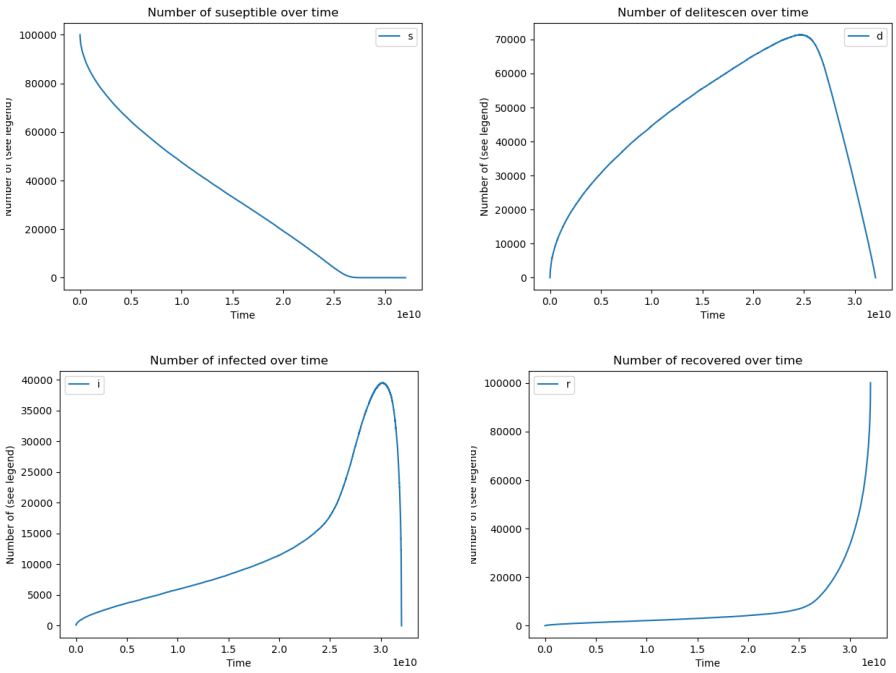

Figure 8: SDIR simulation per population

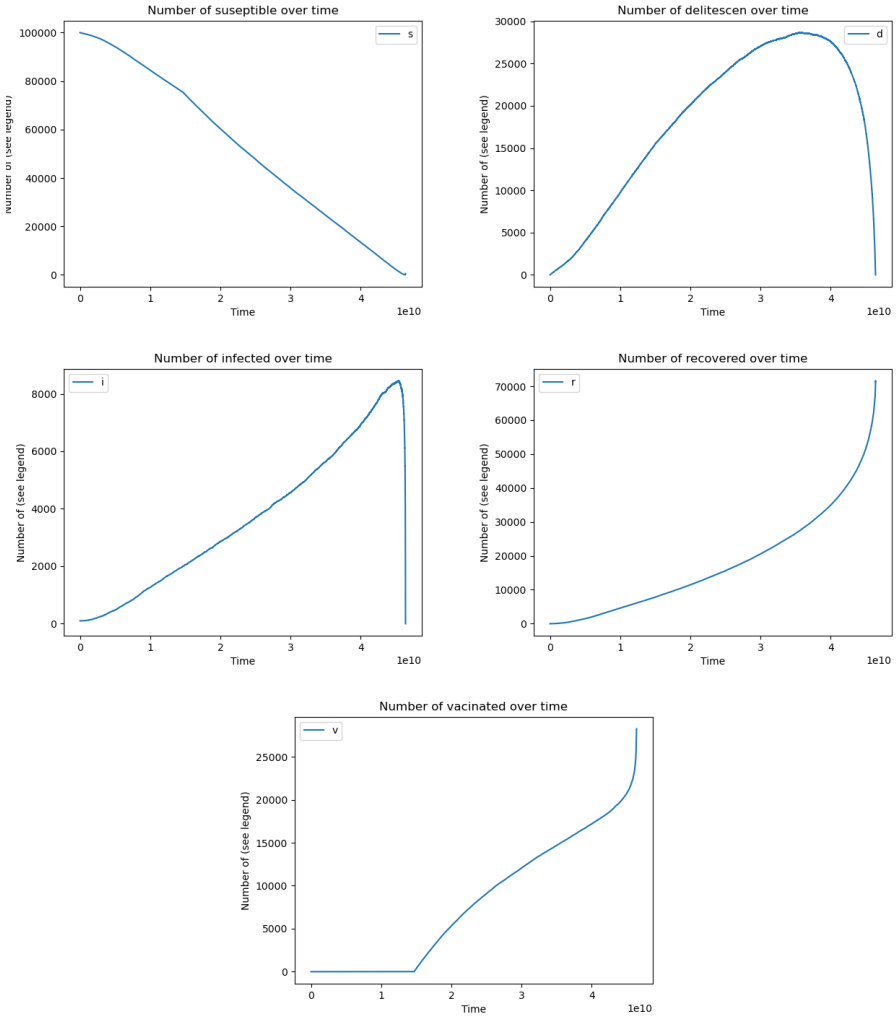

Figure 9: SDIR-V simulation per population