Modeling Virus Diffusion on Social Media Networks with the SMIRQ Model

Justin Browning
*University of North Texas*, justinbrowning@my.unt.edu

Arnav Mazumder
*University of North Texas (TAMS)*, arnavmazumder@my.unt.edu

Gowri Nanda
*University of North Texas (TAMS)*, gowrinanda@my.unt.edu

Follow this and additional works at: [https://scholar.rose-hulman.edu/rhumj](https://scholar.rose-hulman.edu/rhumj)

Part of the Discrete Mathematics and Combinatorics Commons, and the Ordinary Differential Equations and Applied Dynamics Commons

**Recommended Citation**
Browning, Justin; Mazumder, Arnav; and Nanda, Gowri (2024) "Modeling Virus Diffusion on Social Media Networks with the SMIRQ Model," *Rose-Hulman Undergraduate Mathematics Journal*: Vol. 25: Iss. 1, Article 8.
Available at: [https://scholar.rose-hulman.edu/rhumj/vol25/iss1/8](https://scholar.rose-hulman.edu/rhumj/vol25/iss1/8)
Modeling Virus Diffusion on Social Media Networks with the SMIRQ Model

Cover Page Footnote
We would like to recognize our sponsor, Noelle West. Without her mentorship, we would not have been able to complete our work.
Modeling Virus Diffusion on Social Media Networks with the SMIRQ Model

By Justin Browning, Arnav Mazumder, and Gowri Nanda

Abstract. As social networking services become more complex and widespread, users become increasingly susceptible to becoming infected with malware and risk their data being compromised. In the United States, it costs the government billions of dollars annually to handle malware attacks. Additionally, computer viruses can be spread through schools, businesses, and individuals’ personal devices and accounts. Malware affecting larger groups of people causes problems with privacy, personal files, and financial security. Thus, we developed the probabilistic SMIRQ (pSMIRQ) model that shows how a virus spreads through a generated network as a way to track and prevent future viruses. Our model is an extension of the standard SEIR model used in epidemiology. Notably, our model improves upon SEIR models for this class of problem by accounting for the connections between individuals in social media networks. We modeled this by generating a random-scale free node network via the Barabási-Albert (BA) Algorithm, while maintaining the analytical benefits of SEIR models.

1 Introduction

Computer viruses and malware are a threat to the safety and productivity of human work as they can severely damage computer hardware and software, leak private data, delete files, and send malware to other people’s computers. There are several types of computer malware and viruses which can be activated by interacting with infected digital media. For example, opening an email message with infected attachments, clicking on an unknown URL, or downloading files with unknown content are all common ways that can infect computers with malware [3]. In the growing digital world, people use the internet daily to complete all forms of work, such as checking and sending emails, using social media, and downloading files. With the ever increasing role that digital communication plays on our lives, it is easier to interact and share things with the entire world. However, social networking services (SNS), such as Twitter, Instagram, Hotmail, Gmail, or Facebook, are also a vessel for viruses to be received and sent to others.

Cybercrime, a form of illegal activities involving the internet, puts internet users at risk for becoming infected and compromising their online safety. The United States has some of the largest costs associated with cybercrime reaching around $100 million...
dollars in 2014 [11]. Additionally, the complications of malware-related cybercrime extends beyond just affecting the economy and government, as it can affect civilians and companies [10,11,13]. Some forms of cybercrime involve fraud, identity theft, and scams, but malware attacks are prevalent. The loss of resources and money can go unnoticed by people affected by computer malware. Platforms for banking, sending money, or transferring data are targets for attacks, which risks the safety of important activities [9]. Fortunately, it is possible to protect from malware through the use of various antivirus software that prevents dangerous content from being downloaded in the first place [6].

In the field of biology and epidemiology, a Susceptible, Exposed, Infected, and Recovered (SEIR) model is a mathematical model that is used to show the spread of viruses and diseases in a population. SEIR models account for factors pertaining to modeling the spread of a virus in epidemiology, such as vaccination, sanitation, location, and demographic. These kinds of models have been applied to larger global pandemics, such as COVID-19, Ebola, and West Nile Virus [2, 8]. SEIR models use differential equations to show how each category of the spread of a virus changes over time given a set of parameters. Similarly, with the spread of computer viruses, it is important to work on a model that takes into account the workings of computers and their differences with biological diseases. For example, an account for entering an online application may be restricted or moderated to protect from a virus. A model for computer viruses needs to consider aspects that are specific to computers and the internet rather than modeling how a virus spreads between people and animals. SNS, a common vessel for computer malware to be sent to others can be represented by a simulation of an online network [7].

Modeling the spread of malware can allow for the tracing of where the virus first appeared, which devices were infected and exposed, which devices are sending malware, and how long it takes to fix a device so that it works normally. Mathematical models that intend to show the spread of a computer virus need to consider transmission methods as the way a computer virus spreads is different to how a biological virus spreads. However, previous models that are used to track the spread of a potential computer virus did not consider for the connections between people, such as contacts, follower lists, and online friends.

A random node network can be generated using a range of distributions that involve exponential, log-normal, and power-law probability distributions. All three distributions describe a probability density function for the vertex degree, or number of vertices incident to the vertex under inspection, which scales in size according to the type of distribution. Using the probability distributions, the nodes are added one at a time, and the vertices are connected to each other according to the chosen distribution [1]. Additionally, the generation of these node networks are conducted algorithmically where one node is added at a time rather than clusters of several nodes with similar properties being generated at a time.

A node network can be used to show how different factors affect the spread of a
virus. In our network, every node represents a person within a social network. If there are nodes with more connections, then such nodes can infect other nodes. Results have shown that the node degree of users' on social media websites such as Facebook and MySpace obey scale invariant or power law degree distributions [4]. As a result a randomly generated network which hopes to model virus spreading dynamics over a generalized social media network would best select an algorithm which creates networks with a similar vertex degree distribution.

With our model being an implementation of an epidemiological model over a scale invariant network, the spread of a computer virus is probabilistically independent, as it is based on whether or not the person representing the node activates the virus or not. A model of this type more accurately captures the influence of the "... well connected nodes..." which are "... crucial in epidemic spreading," and additionally indicates that the type of network impacts virus spread [4]. As the node network increases in size, these crucial vertices are more prone to exhibiting cascading effects throughout the network, as modeled in threshold models [12]. This is because if one vertex is surrounded by infected vertices, the surrounding vertices must succeed in enduring far more attempts to infect them. As the degree of such a vertex gets larger the probability of succeeding many infection events in the binomial probability distribution for a large approaches zero. Thus, as the degree connectivity of these vertices increases, clusters of the network would experience larger effects of the spread of a virus. Such an outcome might simulate the effect of virus propagation between closer groups within a network of people.

Our model is novel as we are able to control how the connections between people work through configuring a base set of parameters that affect our data while reworking the SEIR model that is used in epidemiology to study how computer viruses spread by way of a graph theoretical approach. Due to the complex nature of a computer virus, our model can account for several different factors such as the rate of infection and the likelihood of online accounts being deleted or quarantined. Additionally we include categories to account for being messaged or quarantined inside of the network which contributes to the realism of simulating a potential virus. Using the probability of infection and changing certain parameters based on a hypothetical situation, our model can show how connections between people online affect virus transmission.

2 Model & Methodology

3 Results

To reflect the impact various parameters have on our model, we selected a case study of different network topologies, and different parameter values for our simulated virus infection probability. In general, all probabilities used in our simulations are based on simulated values, which are themselves based on values similar to those seen in various
First we will explore our results for a network which is generated with a total of one connection per node as the nodes are added into the network with the BA-Algorithm. This network is named Network A. For Network A, we analyze what occurs on a baseline version of our model and compare that with what occurs on a version of the model which increases the infection rate parameter. Next we will explore the results for a network which is generated with a total of two connections per node as the nodes are added into the network with the BA-Algorithm. This network is named Network B. For Network B, we also analyze what occurs on a baseline version of our model and compare that with what occurs on a version of the model which increases the infection rate parameter. Lastly, we generate descriptive statistics for our results, and use these statistics to analyze the differences between Network A and Network B, with the aim of discussing what impacts adding more than one connection to a node makes on virus propagation.

3.1 Network A Results

Network A, depicted above, is a 250 node network generated with one connection per node as the BA Algorithm generates new vertices and preferentially attaches them [1, 5]. With this Network we simulated the introduction of a virus on this network, using the baseline SMIRQ model which we developed, and simulated what would occur. Due to the stochastic nature of virus propagation, we chose to average the results for 120 simulations, at each day, for 720 days, for each population category. This resulted in the following results, represented graphically:
Figure 2: Simulation Results for the baseline parameter model, with parameters: $\lambda = 0.2500, \eta = 0.0150, \epsilon = 0.0450, \gamma = 0.0010, \phi = 0.0200, \sigma = 0.0100, \zeta = 0.0100$ $D_R = 0.0005$ on Network A.

In the graphed results we identify 6 categories for the population that each node can be; Susceptible, Messaged, Infected, Quarantined, Recovered, and Removed. For our baseline parameters, we can see at time $t = 0$ there are 25 initially infected nodes, or 10% of the initial population. From there we see the susceptible nodes begin populating the messaged category as they start receiving messages from infected nodes. As predicted, as time progresses, the quarantined category manages to outpace the rate of infection, and by day 720, $S(720) = N_T - D_T$ where $N_T$ is the total population of nodes (250), and where $D_T$ is the total number of nodes removed throughout the simulation. This indicates that the virus was completely quarantined and removed from the population.

Next we simulated a version of the model which increased the infection rate probability $\eta$ from $\eta = 0.0150$ to $\eta = 0.0350$ on Network A. Our goal here was to explore what would occur if we simulated a version of the virus which was more infectious, and to show that the model behaves as we would expect for this increase in infection probability. Once again we chose to average the results for 120 simulations, at each day, for 720 days, for each population category.
In the graphical representation of our results, we can see that as we would expect, the increase in the infection rate probability of the virus affects the minimum population of the susceptible population but also the infected population. Additionally we can see the maximum population of the quarantine category increasing, responding to the increased amount of infected nodes in our network. Thus an increase in the number of recovered nodes, resulting from the increase of infection and quarantined nodes.

However, we do not notice a considerable difference in the number of messaged nodes at any point in time. This is due to the power-law distribution of the number of connected edges to a node. In order for the virus to transmit more messages, it would need to be connected to more people. Thus, it needs to reach a highly connected node in the network in order to have many more messages to send, which occurs less frequently with more simulations.

In an effort to show the differences between the two models, we used the data from both simulations to generate the following comparison graphics. These do not represent any additional simulations, but rather, just compare a single category from each of the two models on Network A.

Figure 3: Simulation results for the adjusted infection rate probability model with parameters: $\lambda = 0.2500$, $\eta = 0.0350$, $\epsilon = 0.0450$, $\gamma = 0.0010$, $\phi = 0.0200$, $\sigma = 0.0100$, $\zeta = 0.0100 \frac{D_R}{R} = 0.0005$ on Network A.
Figure 4: Comparison of results between the baseline parameter model and a model with an increased infection rate probability on Network A.

Effectively, these graphical comparisons serve as a summary of our results for Network A between the two models. The most important aspect of our result here is that we can see that increasing a single parameter ($\eta$) resulted in a significant difference in the magnitude of infection, by way of an increase in the population of infected nodes. This one change had a ripple effect among the entire system which we can see when we compare the results between the two models. Nonetheless, by the end of the 730 days, our virus is eventually eliminated from both networks by the action of the Quarantined and Recovered categories. This then allows us to conclude that the behavior of our model is dependent on the configuration of the parameters and so our solutions between the two models are unique.
3.2 Network B Results

Network B, depicted above, is a far messier network generated with 2 connections per node as each node is added into the network. This means that the total pathways for virus traversal is increased by a factor of 2N_T, where N_T is the total population of the network. We simulated Network B using our baseline parameter model with the following results presented graphically:

Figure 5: Network B, generated with the Barabási-Albert Algorithm for 250 nodes.

Figure 6: Simulation Results for the baseline parameter model, with parameters: \( \lambda = 0.2500 \), \( \eta = 0.0150 \), \( \epsilon = 0.0450 \), \( \gamma = 0.0010 \), \( \phi = 0.0200 \), \( \sigma = 0.0100 \), \( \zeta = 0.0100 \) D_R = 0.0005 on Network B.

In our baseline model simulation on Network B, we can see that our Susceptible pop-
ulation does not recover in this instance, as it did in the same simulation on Network A. The overall differences can be attributed to the additional pathway of virus transmission, as this is the only difference between the two networks in aggregate.

From Network B we generated a model which increased the infection rate probability parameter ($\eta$) in an effort to see how this change would impact the behavior of our virus.

![Figure 7: Simulation Results for the increased infection rate probability model, with parameters: $\lambda = 0.2500$, $\eta = 0.0350$, $\epsilon = 0.0450$, $\gamma = 0.0010$, $\phi = 0.0200$, $\sigma = 0.0100$, $\zeta = 0.0100$ $D_R = 0.0005$ on Network B.](image)

Here we can see that in our simulation results there are significant quantitative differences between the increased infection rate parameter simulation and the baseline parameter simulation. This shows that in the instance where a network has an additional connection between each node, the change in a parameter can alter the behavior of the virus diffusion through the network, similar to the same simulation on Network A.

In an effort to show the differences between the two models on Network B, we once again used the data from both simulations to generate the following comparison graphics:
Effectively, these graphical comparisons serve as a summary of our results for Network B between the two models. One of the most important aspect of our results here in Model B is that we can see that increasing a single parameter ($\eta$) resulted in a significant changes to the system, by way of the increasing the population of nodes which are infected. This then had cascading effects on the various populations over time in each category. We can conclude that the virus’ modelled propagation behavior on networks with more connections per node is dependant on the configuration of all the parameters in the model.
3.3 Descriptive Statistics for Networks A & B

In this subsection we present four tables summarizing the results for each of the simulations. These values were generated using standard functions in Python's NumPy library, and represent the average result for each category in the simulation. Due to the stochastic nature of the simulations these values are provided for verification of general results outlined in this section. In doing so we use these values to verify that the descriptive claims made about the different simulation results are reasonably accurate, particularly when comparing results between simulations.

<table>
<thead>
<tr>
<th>Baseline Probability Model</th>
<th>Max</th>
<th>Min</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Network A)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Susceptible</td>
<td>239.876</td>
<td>168.05</td>
<td>205.699</td>
<td>2.29862</td>
</tr>
<tr>
<td>Messaged</td>
<td>34.3417</td>
<td>0</td>
<td>7.56982</td>
<td>0.885841</td>
</tr>
<tr>
<td>Infected</td>
<td>25</td>
<td>0.341667</td>
<td>6.9007</td>
<td>0.700425</td>
</tr>
<tr>
<td>Quarantined</td>
<td>30.2833</td>
<td>0</td>
<td>13.7138</td>
<td>0.894608</td>
</tr>
<tr>
<td>Recovered</td>
<td>24.3667</td>
<td>0</td>
<td>13.4202</td>
<td>0.679671</td>
</tr>
</tbody>
</table>

Table 1: Descriptive Statistics for results in the baseline parameter simulation of Network A.

<table>
<thead>
<tr>
<th>Increased $\eta$ Probability Model</th>
<th>Max</th>
<th>Min</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Network A)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Susceptible</td>
<td>225</td>
<td>107.508</td>
<td>168.673</td>
<td>3.48481</td>
</tr>
<tr>
<td>Messaged</td>
<td>39.95</td>
<td>0</td>
<td>7.64011</td>
<td>0.916606</td>
</tr>
<tr>
<td>Infected</td>
<td>49.8083</td>
<td>1.575</td>
<td>14.2003</td>
<td>1.32864</td>
</tr>
<tr>
<td>Quarantined</td>
<td>62.6917</td>
<td>0</td>
<td>27.6376</td>
<td>1.68746</td>
</tr>
<tr>
<td>Recovered</td>
<td>48.6</td>
<td>0</td>
<td>26.5556</td>
<td>1.2885</td>
</tr>
</tbody>
</table>

Table 2: Descriptive Statistics for the increased infection probability rate simulation of Network A.
Modeling Virus Diffusion on Social Media Networks with the SMIRQ Model

<table>
<thead>
<tr>
<th>Baseline Probability Model (Network B)</th>
<th>Max</th>
<th>Min</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Susceptible</td>
<td>225</td>
<td>74.7667</td>
<td>104.62</td>
<td>1.67566</td>
</tr>
<tr>
<td>Messaged</td>
<td>76.3667</td>
<td>0</td>
<td>33.187</td>
<td>0.916606</td>
</tr>
<tr>
<td>Infected</td>
<td>46.6833</td>
<td>17.6167</td>
<td>24.369</td>
<td>0.793212</td>
</tr>
<tr>
<td>Quarantined</td>
<td>64.2</td>
<td>0</td>
<td>42.7316</td>
<td>1.14951</td>
</tr>
<tr>
<td>Recovered</td>
<td>53.8083</td>
<td>0</td>
<td>38.622</td>
<td>1.28988</td>
</tr>
</tbody>
</table>

Table 3: Descriptive Statistics for results in the baseline parameter simulation of Network B.

<table>
<thead>
<tr>
<th>Increased η Probability Model (Network B)</th>
<th>Max</th>
<th>Min</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Susceptible</td>
<td>225</td>
<td>26.8333</td>
<td>67.8198</td>
<td>2.09921</td>
</tr>
<tr>
<td>Messaged</td>
<td>81.3333</td>
<td>0</td>
<td>20.9953</td>
<td>1.39039</td>
</tr>
<tr>
<td>Infected</td>
<td>87.0583</td>
<td>22.075</td>
<td>34.7814</td>
<td>1.57413</td>
</tr>
<tr>
<td>Quarantined</td>
<td>97.4583</td>
<td>0</td>
<td>62.0606</td>
<td>1.6591</td>
</tr>
<tr>
<td>Recovered</td>
<td>75.35</td>
<td>0</td>
<td>55.3492</td>
<td>1.68517</td>
</tr>
</tbody>
</table>

Table 4: Descriptive Statistics for the increased infection probability rate simulation of Network B.

4 Discussion & Conclusion

The proliferation of digital viruses on social media networks is an issue that affects all of us online, and potentially compromises our personal privacy [10, 11, 13]. Our work extends previous research by adding an additional quarantine category as well as implementing the feature of probabilistic messaging from all infected nodes to each directly connected susceptible node, individually [4,7]. This represents an improvement on existing modeling of social media networks, which do not implement both features and, instead, rely on pSEIR models that do not use a graph theoretical approach to virus transmission. Furthermore we extended the results to include the different kinds of connections that a social media network may exhibit.

In analyzing the differences between the network topology of Network A and Network B, we are limited to hypothesizing that the initial degree generation scheme, given by the BA Algorithm, affects the spread of virus through the network. By easy inspection we can motivate a reason for this, that being that there exists no isometries between spanning subgraphs of the two networks. This fact indicates to us that the results between the two different graphs may not be compared when discussing the dynamics of virus spread. But we may discuss how the inclusion of an additional pathway for virus transmission...
between the two networks with similar parameters alters the topology of the network and what our results indicate about that alteration. To that end, our discussion will center itself on discussing the role that altering the network topology had between results of similar parameter setups and then discuss the dynamics of virus propagation within those networks indicated by altering the $\eta$ value.

Our investigation indicates that between results given in the baseline parameter simulation in Network A and Network B the presence of additional pathways is significant in the persistence and propagation of the virus. This comparison can be done by noting that between Table 2 and Table 4, Network B exhibits a larger mean, max and min value for the infected category. Notably it also has a larger standard deviation implying that there is more variance among these results. We can conjecture that this is due to the cascading effects of larger degree values for those particularly connected vertices in this network topology. Further work here should be done to examine if there is a threshold for which these connections are made "too heterogeneous" such that the system becomes less vulnerable to cascading effects from highly connected nodes such as described in [12].

In comparison graphs given for Network A and Network B, we can see that the presence of the quarantined category affects the dynamics of virus spreading within their respective networks. This is indicated by the fact that as the $\eta$ value is increased, there is a correspondingly large increase in the number of quarantined nodes in the network, and later we see the recovered category at time-steps later in the simulation with population differences that reflect this increase of $\eta$. Thus the quarantined category "collects" up latent infection that cannot be immediately recovered from.

One other notable difference between the two networks is that it appears as though the availability of additional pathways Network B allows for an 'equilibrium' value to occur for the S, E, I, R, & Q states to persist seemingly indefinitely, whereas in Network A the S state tended to 250 while the others tended to 0. In this investigation we were unable to examine longer periods of time or larger networks due to the computational expense of running these simulations. Therefore further investigation with appropriately powerful computers might examine if this remains stable over larger periods of time, on larger networks, and on more heterogeneously connected networks.

While modeling these behaviors on random graphs gives us a good first approximation, the model may only be of marginal use to set bounds on the expected behavior of virus transmission on a network [12]. This bounding should be especially useful if the network exhibits a similar behavior to any of the available random network generating algorithms. For our purposes then the simulation should be considered a good approximate of the dynamics of scale-free networks such as those exhibited on social media websites like Facebook [4]. In particular the model will serve as a better approximation when compared against a standard pSEIR model for the use-case of studying social media malware spread since they do not consider the vulnerability and cascading effects.
of highly connected nodes on non-heterogeneously connected scale-free social media networks. Usually standard $p$SEIR models assume a completely homogeneous mixing of the population in question, and as we have shown the level of connectedness is largely influential to the dynamics of virus spreading. This then indicates the unsuitability of standard $p$SEIR model applied over a homogeneous population for the use case of making good approximations for the behavior of viruses on social media networks.

In order to accurately predict how a virus will propagate on a network, data must be collected about how prone to error the general population is. This includes determining the likelihood of users clicking on unsafe links or downloading malicious code. As well, data must be collected about how likely it is to scan the network to actively quarantine for a virus. Additionally, we must know how likely it is that one can self-recover from this virus, if at all, and then lastly we must know how likely it is for a user to be prone to becoming susceptible for the virus again.

With this method, one can simulate what approximately will occur on their social media network, and prepare to develop actions to mitigate against the impacts that these types of attacks have on it.
References


