Analysing Worm-Virus Dynamics with Detection and Prevention in Peer-to-Peer Networks: The SIDR Model

Matthew A. Ford

Department of Informatics, University of Sussex, BN1 9QH, UK

mf472@sussex.ac.uk

Abstract – This report analyses the spread of worm viruses with varied attack strategies though peer-to-peer networks, outlining the importance of high-quality detection methods and good user proactiveness for the minimisation of the spread of worms through a network.

1. Introduction

Computer viruses, much like biological pathogens, self-replicate by taking advantage of system weaknesses and attaching themselves to a host program to cause the virus instructions to be executed. *(Chen, T. M., et al., 2004)*

Worm viruses travel from system to system through the networks used to send communication between systems. These were built with the intent of distributing workloads, but instead, they soon became considered to be destructive. *(Spafford, E. H., 1989).*

In 2017, a worm virus known as "WannaCry" infected NHS organizations. This attack encrypted critical information and demanded a ransom. *(NHS England, 2023).* This became one of the most significant virus attacks seen, affecting over a third of systems within the NHS – highlighting the importance of detection and prevention tactics to avoid a full network infection.

The current guidance for infected systems, given by the NHS, is to disconnect from the network straight away. This however is dependent on the responsiveness of users, and the effectiveness of virus detection to halt the spread.

This report explores how the speed of both the detection and preventative action affect the spread of viruses. For this, a real-world dataset of peer-to-peer interactions from the Stanford large network dataset collection *(Leskovec, J., et al., 2004)* is employed to run these experiments. This dataset is ideal for modelling the propagation of worm-like dynamics in the network where each node is representative of a device capable of spreading infections.

While many epidemiological models rely on the commonly known SIR (Susceptible-Infected-Recovered) framework *(Kermack, W. O., et al., 1927)*, these models often fail to account for technological interventions such as detection and isolation. To address this, the SIDR model (Susceptible-Infected-Detected-Recovered) is proposed. This incorporates a "detected" state to simulate the moment a virus is detected by a system's security software. This adds a delay between infection and recovery (user intervention) which is important for accuracy for modelling cyber-attacks.

The strategies of attack will differ throughout the experiment. The three distinct strategies include: a **random** selection of infected nodes, chosen with no regard for the network structure. A strategic attack, where central / high-degree nodes are chosen for initial infection. And a peripheral attack, where the infection is started at the leaf nodes of the network. As a safety feature, only leaf nodes with at least one out-connection can be chosen, as otherwise, the virus would not spread.

The main questions that guide these experiments are:

- How does the responsiveness of detection reduce outbreak size?

- How does proactive recovery reduce the infection?

- What virus-spreading strategy creates the largest outbreak?

The findings from this study aim to inform practical cybersecurity policies, particularly in infrastructural system (e.g. healthcare, and finance) as well as laying the foundations to further models and research in this area. By simulating the infection dynamics under realistic conditions, the importance of detection and user intervention are emphasized in the effort to minimise the risk of any modern digital epidemics.

2. Methods

The exact dataset used in this report was the "p2p-Gnutella08". This data outlines a directed graph in which nodes represent users, and edges represent network connections.

To load the graph, into a processable format "NetworkX" *(Hagberg, A. A., et al., 2008)* was used. This library provides all the necessary tools to visualise and analyse networks.

2.1 | Network Structure

The dataset discussed provided a directed graph with 6,301 nodes and 20,777 edges. As this experiment is based on the effects of an implemented system rather than the structure of the network, it was seen as unnecessary to choose a dataset with double or triple these values.

As seen in *Figure 1*, both the in-degree and out-degree distributions are heavily tailed. This suggests most nodes have a low connectivity. Although the number of in-degrees would be equal to that of the out-degrees, these distributions show that there are many nodes with a low number of each, and some with a high number of connections. This therefore suggests that there are hubs.

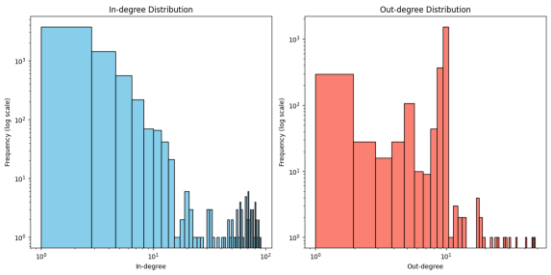

Figure 1: The in-degree and out-degree distributions of the p2p-Gnutella08 network.

The presence of hubs is very common to complex networks such as the internet. This structure ha important values such as reducing the steps to each node but also increasing the vulnerability to attacks.

Through sampling the network, the average connectivity was estimated to be approximately 3.28.

This suggests that each node, on average, is connected to a small number of other nodes.

Although this connectivity is not uniform throughout the network.

Figure 2 shows the core structure of the network. The core nodes (in yellow) have higher connectivity averages and therefore are much more critical to the network structure and stability.

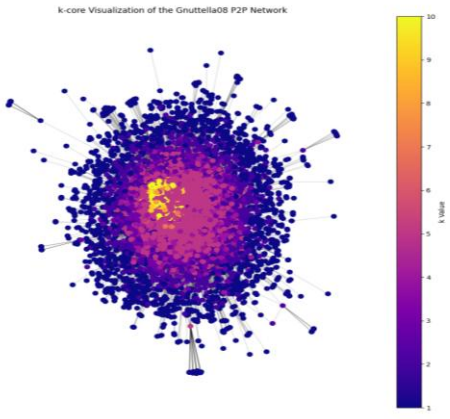

Figure 2: The k-core structure of the network.

The largest connected component of the network has a size of 2068. This component is the densest subgraph of the network.

This section has an important role in supporting peer-to-peer activity. As there are 6,301 nodes, it can be shown that only ~33% of nodes belong to this component. This indicates a fragmented topology where the central cluster is surrounded by smaller, weakly connected components, and supports the layered structure seen in *Figure 2*.

The insight into the structure of the network is important for understanding the dynamics of virus propagation within. Hubs and core nodes act as a form of 'super-spreaders', transmitting the virus rapidly across the network as a result of their high connectivity value.

The average connectivity of 3.28 suggests a limited transmission capability as the centrality is lower hut having a highly connected core significantly reduces the average path length for faster virus spread.

2.2 | SIDR Model

As an extension to the Susceptible-Infected-Recovered (SIR) model, a custom model was created. This being the Susceptible-Infected-Detected-Recovered (SIDR) model.

This model works similarly to the SIR model, by having uninfected nodes being susceptible to infection as infected nodes send messages across the network. As this experiment is to determine the importance of detection and response to a virus, the extra "detected" step shows when a system recognises that it has been infected. This represents the antivirus in the system and metaphorically "tells" the user that the virus has been detected.

From here, the transition between detected and recovered represents the user's responsiveness to the virus detection. As discussed, the advice for any infected systems is to disconnect from the network. Therefore, upon doing so, a node is recovered. From here, a node either stays disconnected, or the virus can be removed after a given time.

Removal of the virus then reconnects the node to the network, making the system susceptible again.

Therefore, this model can also be considered as SIDRS, similar to the commonly known, looping SIS (Susceptible-Infected-Susceptible) model. *Figure 3* visualises the loop for the discussed SIDR model.

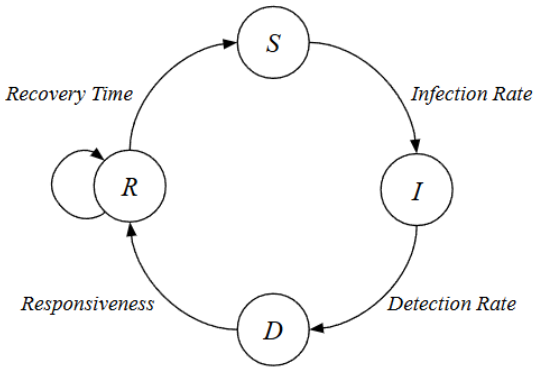

Figure 3: The state loop of the SIDR(S) model.

It should be noted that the recovered state is looped back, representing nodes that choose not to 'remove' the virus and reconnect.

2.3 | Model Dynamics

The dynamics of the model were important to create a realistic simulation, with the attempt to implement awareness and attack strategies.

2.3.1 | Initial Infected Selection

As viruses are commonly started through targeted attacks, it was important to compare the spread, dependent on the initially infected systems. *Figure 4* shows the three methods used for selecting the 60 initial infected.

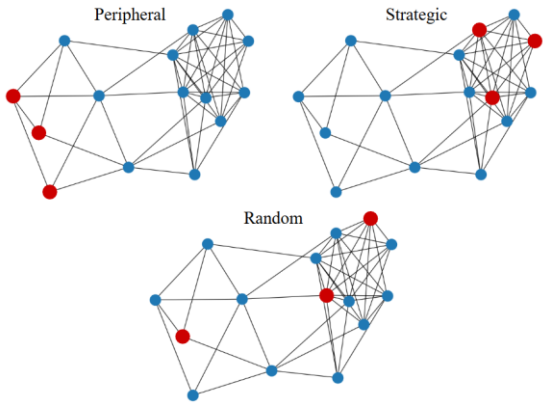

Figure 4: Visualisations of different initial infected selection tactics.

Here there are three methods: peripheral, strategic, and random. The random method, as suggested, selects completely random nodes, without considering its position in the network.

Strategic attacks target the nodes with the highest degree centrality value. This naturally creates a faster spread as hubs are a primary target.

However, peripheral attacks target those with the lowest degree centrality. These are usually the leaf nodes, unless the graph is fully connected. As the network also consists of disconnected nodes or nodes with only outgoing edges, the peripheral nodes were selected from the set of nodes with at least one successor.

2.3.2 | Awareness

As viruses spread through networks, it is common for there to be some social interaction between neighbouring users. For this experiment, this information diffusion is represented as an awareness level which affects the proactiveness of the agents that controls the state transition from D to R.

In the model, the proactiveness of any node is increased if more than 20% of it's neighbours are infected. This formula for P_i^{t+1} is as follows:

$$P_i^{t+1} = \begin{cases} \min\left(1.0, P_i^t + \frac{0.5}{1 + e^{-N_i(I)+2}}\right) if\ N_i(I) > 20\% \\ \max(0.1, 0.99 * P_i^t)\ if\ N_i(I)\ = 0 \end{cases}$$

Where:

$$N_i(I)\ :\ \ Number\ of\ infected\ neighbours$$

The threshold of 20% avoids the model becoming overly sensitive and overreacting to threats. The sigmoid function is used for a smooth buildup, rather than a static linear increase. If a node has no infected neighbours, they become less proactive.

To add further dynamics to the system, the proactiveness values themselves diffuse to neighbours.

Therefore, the proactiveness of any node averages towards that of its neighbours.

$$P_i^{t+1} = 0.9 \cdot P_i^t + 0.1 \cdot \left(\frac{1}{|N_i|}\sum_{j \in N_i} P_j^t\right)$$

Where:

$$N_i\ :\ Set\ of\ successors\ to\ node\ i$$

These dynamics and feedback loops for changing the proactiveness (after it has been assigned a value) adds a further level of realism to the model, making the research much more relevant than results drawn from static graphs.

2.4 | Simulation Parameters

The parameters that can be used to initialise the system are the average proactiveness, the average detection (or antivirus), and the chosen attack strategy (discussed in *Section 2.3.1*).

To add slight randomness to the system, when given the means of the user proactiveness and detection, these were then selected from a random distribution with the given mean as the centre, and a 0.15 standard deviation. This is illustrated in *Figure 5*.

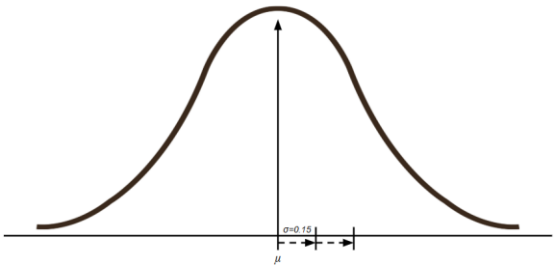

Figure 5: A visualisation of the distribution of initial values, given the mean (μ).

This adds an element of randomness to the initial values, ensuring all node values avoid convergence at the same pace.

2.5 | Evaluation Metrics

The metrics used to evaluate the changes in virus spread dynamic in the network are as follows:

- Susceptible nodes over time
- Infected nodes over time
- Detected nodes over time
- Recovered nodes over time
- R0 value over time

To ensure the results are not just a result of randomness, each parameter combination was run 5 times, and the shown metrics are the average of these runs.

3. Results & Analysis

3.1 | Attack Choices

The results obtained from the parameter sweep provide insight into how different attack strategies influence the spread of a virus across the network. As illustrated in *Figure 6*, which displays the average SIDR curves for each strategy, there are subtle but important differences between the outcomes. Although, some of these differences are difficult to see.

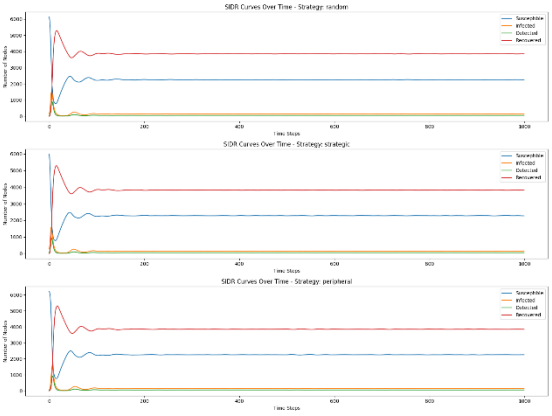

Figure 6: The average SIDR curves for all parameter sweeps, for each strategy.

As expected, strategic attacks on, high-degree nodes, lead to a more aggressive spread of the virus. These hubs help with transmission to many directly connected neighbours which accelerates the outbreak in the early stages. In contrast, peripheral attacks on lower-degree nodes on the outer edges of the network, result in a slower infection rate. This is consistent with the reduced potential for immediate spread in sparsely connected regions.

The recovery dynamics differ between strategies. The peripheral attack shows a more effective and faster decline in infections. This is likely due to the limited connectivity of infected nodes, which as a result limits the spread. However, infections initiated in hubs tend to persist longer, with the virus remaining active over an extended period in the case of strategic attacks. This prolonged activity is due to the higher centrality of the nodes, allowing the virus to propagate further through the network before containment mechanisms have an effect.

Interestingly, the random attack strategy yields similar dynamics to the strategic case. This resemblance may be due to the topology of the network, particularly because of the existence of a giant connected component that contains approximately 33% of the network's nodes. Given that 60 nodes are initially infected in each simulation, approximately 20 of these are likely to fall within this component by chance, including nodes with high centrality.

As a result, even random infections may quickly reach the hubs, producing a cascade effect like that observed in targeted strategic attacks. However, this outcome reflects the specific structure of the network rather than the effectiveness of the random strategy. In larger or more sparse networks, these cascading effects from random attacks would be less common.

The basic reproduction number (R_0) gives the infection rate across the network. As shown in *Figure 7*, the initial R_0 values vary significantly across strategies. Strategic attacks begin with a high R_0, showing the discussed rapid early

spread, while peripheral attacks begin with much lower values.

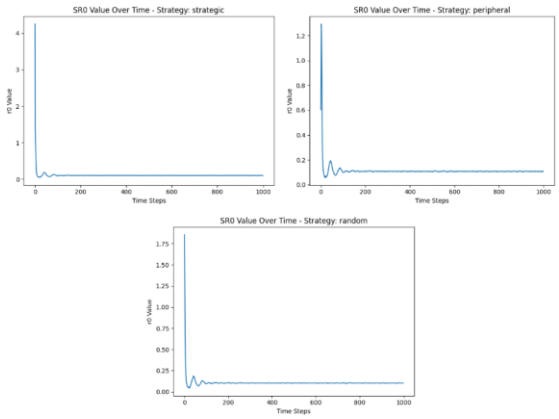

Figure 7: The average R_0 value for every parameter combination, for each strategy.

However, over some time, all strategies converge to similar final R_0 values slightly above zero. This convergence suggests that there is a stable equilibrium state in the system's dynamics, acting as a natural attractor determined by the model's parameters and the underlying network structure, rather than the initial infection strategy.

In summary, the choice of initial infection strategy affects short-term outbreak dynamics. This includes the speed and size of early virus spread. However, in the long run, all strategies appear to lead the system toward a steady state which has a low-level, persistent virus. This behaviour may be a result of the adaptive elements in the model, or inherent stabilizing properties of the network topology. Therefore, on average, while initial strategies can influence the early stages of an outbreak, they do not appear to alter the long-term scale or presence.

3.2 | Antivirus & Proactiveness

To assess the effectiveness of proactiveness and antivirus on the system, the heatmaps in *Figure 8* were created. These show the different parameter values, plotted with the number of steps taken for the virus to go extinct (0 infected). This was seen as the only reasonable metric to use for the comparison, as basing the heatmaps off values such as the final detected, or final recovered nodes would result in a complete favourability towards one of the parameters due to the loop structure seen in *Figure 3*.

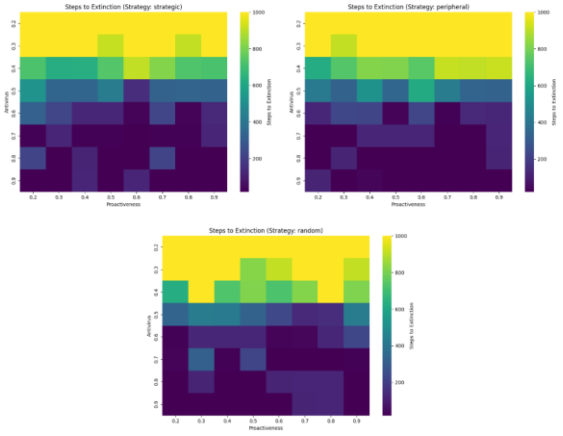

Figure 8: Heatmaps for the antivirus and proactiveness parameter values, showing the steps until a virus goes extinct.

In these heatmaps, it is shown that the higher levels of proactiveness and antivirus detection result in a faster extinction. This is natural to the system as it is based off probabilities, and therefore having higher values means more nodes become recovered more efficiently, leading the viruses' eventual extinction.

For the strategies in relation to these parameters, it was shown that the strategic extinctions take the longest, followed by random and then peripheral. Reinforcing the evidence in the graphs seen in *Figure 6*.

Again, these results are not completely different due to the network topology, which seems to have elements of small-world graphs.

4. Discussion

This experiment demonstrated that both detection mechanisms, and proactive user responses are essential to supressing the spread of worm viruses in peer-to-peer networks, using the SIDR framework. While the initial strategy for selecting nodes to infects (whether targeting strategic, peripheral, or random nodes) influences the early dynamics of the outbreak, these differences tend to converge over time. As a result, the system converges toward a natural attractor that is a result of the feedback in the adaptive dynamics, and the network topology. This happens regardless of the initial conditions.

However, this experiment had some limitations that may have been key influencers to the results seen in *Section 3*.

The first limitation is the static network structure. This is because no nodes were added or removed, and no new connections were formed. As a result, this could have been a limitation on the realism of a system demonstrating traits similar to social networks. Although, it was still a valid assumption to make for closed systems, such as enterprises, or healthcare networks, mirroring the "WannaCry" attack on the NHS *(NHS England, 2023)*.

The second limitation regards the behavioural response of nodes becoming uniform across the network. As a level of realism, discussed in *Section 2.3,* a system of information diffusion was added as a way of simulating realistic communication between nodes to make each other more aware of the virus. However, due to the giant connected component at the core of the network, awareness was spread and increased rapidly. This caused most nodes to gain the same level of proactiveness and therefore homogenized the response to infection. As a result, this reduced the diversity of behaviours, and pushed the system towards a convergence, which was not always desired.

The final limitation was the overall network topology. Again, because of the centralised core, the paths viruses needed to take in order to infect a hub were relatively short. This was the case even when the peripheral nodes were initially infected. This reduced any impact of the different strategies and led to the numerically similar behaviours seen in many of the results presented.

To summarise, while the short-term effects of strategy selection are seen, any long-term behaviour is overshadowed by the effects of network structure, and information diffusion.

For future work in this area, it would be interesting to explore how dynamic networks, or localised topologies react to the spread of a worm virus, which can then better investigate the resilience of real-world systems to total infection.

5. References

Chen, T. M., Robert, JM., (2004). *Statistical Methods in Computer Security ch-19: The Evolution of Viruses and Worms*

https://doi.org/10.1201/9781420030884

Hagberg, A. A., Schult, D. A., Swart, P. J., (2008). Exploring network structure, dynamics, and function using NetworkX

https://networkx.org/documentation/networkx-2.1/citing.html

Kermack, W. O., McKendrick, A. G., (1927). *A Contribution to the Mathematical Theory of Epidemics*

https://doi.org/10.1098/rspa.1927.0118

Leskovec, J., Krevl, A. (2014). *SNAP Datasets: Stanford Large Network Dataset Collection*

https://snap.stanford.edu/index.html

NHS. Malware.

https://cfa.nhs.uk/fraud-prevention/reference-guide/cyber-enabled-fraud/cyber-threats/malware#closedResources

NHS England, (2023). *NHS England business continuity management toolkit case study: WannaCry attack*

https://www.england.nhs.uk/long-read/case-study-wannacry-attack/

Spafford, E. H., (1989). The Internet Worm Program: An Analysis

https://doi.org/10.1145/66093.66095