On the Malware Front

R.E. Kooij *robert.kooij@tno.nl*
Delft University of Technology &
TNO (Netherlands Organization for Applied Scientific Research)
Brassersplein 2, 2612 CT Delft, the Netherlands

H.J. van der Molen *henk-jan.van.der.molen@hswageningen.nl*
Wageningen University
P.O. box 2, 6700 AA Wageningen, the Netherlands

Abstract

The purpose of this article is to extend related research on the spread of malware in networks and to assess the security impact of certain measures against the spread of malware. We examine the influence of heterogeneous infection and curing rates for a Susceptible-Infected-Susceptible (SIS) model, that is used to describe the spread of malware on the Internet. The topology structure considered is the regular graph, which represents homogeneous network structures. We present a new method to calculate the steady state of heterogeneous populations, for the general case with m subpopulations. Using this method, we give the explicit conditions under which the malware persists in the network. Under the condition that all infection rates in the heterogeneous model are equal, we provide a logistic solution for the model.

Next we give calculation examples which are based on the assumption of two subpopulations and explore this method in more detail, proving that the method produces valid outcomes and that the basic reproduction numbers R for each subpopulation are the only factors determining the steady state situation. The value of R depends on the effectiveness of attacking malware and the defending countermeasures.

We then consider some special cases for subpopulations using this method. In the first case the protection against malware is assumed to be absent within one subpopulation. The calculations show that it pays off for the subpopulations with the best protection to help other, less protected subpopulations.
The second case describes the effect of diversification against malware, when one subpopulation does not share the vulnerabilities with the rest of the population to become infected with malware and propagate that malware. The results show that diversification is an effective countermeasure against the propagation of malware. Based on the market share of the software, we demonstrate how to calculate the 'resistance' of different compartments against malware.

Using statistical data, we finally show that dividing a population in two subpopulations increases the accuracy of the model. Based on this data, we also show that the use of security software does not correlate very well with the number of reported infections.

Keywords: Virus spread, epidemic threshold, heterogeneous networks, diversification.

1. INTRODUCTION

In our current society the Internet represents an enormous societal and economic value. Unfortunately where there is value, crime is soon to follow and on the Internet now many cybercriminals are active and malware is ubiquitous. The term “malware” is defined as a piece of software with a harmful payload, which needs (vulnerabilities in) a specific software package to propagate from an infected system to other systems. In 2010, the average rate of malware in email traffic was 1 in every 284 emails and the average number of malicious web sites blocked each day rose to 3,188. Almost 90% of these blocked sites are legitimate sites, which were compromised [1]. When infected computers spread the infection to other computers, the number of malware sources explodes in a short time. When confronted with such malware avalanches, relying on prevention alone is not realistic any more. It is necessary to identify the factors that control the spread of malware on the Internet, predict how many computers will be infected and how effective countermeasures are.

The purpose of this article is to extend related research on the spread of malware in networks and to assess the security impact of certain measures against the spread of malware. Our malware spreading model is based upon the Susceptible-Infected-Susceptible (SIS) infection model, which arose in mathematical biology, which is often used to model the spread of computer viruses [2], [3], [4]. The SIS model assumes that a node in the network is in one of two states: infected and therefore infectious, or healthy and therefore susceptible to infection. The SIS model usually assumes instantaneous state transitions. Thus, as soon as a node becomes infected, it becomes infectious and likewise, as soon as a node is cured it is susceptible to re-infection.

In epidemiological theory, a crucial notion is the epidemic threshold τ_c, see for instance [2], [3], [5], [6], [7], [8]. If it is assumed that the infection rate along each link is β while the curing rate for each node is δ then the effective spreading rate of the virus can be defined as $\tau=\beta/\delta$. The epidemic threshold can be defined as follows: for effective spreading rates below τ_c the virus contamination in the network dies out, while for effective spreading rates above τ_c the virus is prevalent. In the case of persistence we will refer to the prevailing state as a steady state. The epidemic threshold is related to the so-called basic reproduction number R, see [6]. In fact, the epidemic threshold τ_c corresponds to the case R=1, with virus extinction for R≤1 and virus prevalence for R>1.

Between 1999 and 2009 many articles considered more modeling aspects for homogeneous populations like incubation periods, variable infection rate, a curing process that takes a certain amount of time, adaptive networks and so on, see [6], [7], [9], [10].
In 2009, ref. [11] derived analytical results for the epidemic threshold in the case of heterogeneous curing rates for a specific class of graphs. It is assumed in [11] that the infection rate at every link is the same, namely β. The aim of this paper is to generalize the results from [11] by also considering heterogeneous infection rates.

The rest of the paper is organized as follows. In Section 2 we derive and analyze the spread of viruses in regular graphs in case of m subpopulations, with curing rate δ_i and infection rate β_i, for i=1..m. In Section 3, we discuss the specific case of regular graphs with 2 subpopulations. In the subsequent sections we consider some special cases; in Section 4 we look at the effect of a population without defense and in Section 5 we look at the impact of diversification. In Section 6 we discuss some statistics obtained through Eurostat in order to determine the relation between security software deployed and the percentage of infected computers. We summarize our results in Section 7.

2. VIRUS SPREAD ON REGULAR GRAPHS WITH M SUBPOPULATIONS

In this section, we derive the threshold for the spread of viruses and the steady state of m subpopulations on regular graphs, each with their own curing rate and infection rate. We assume that each node in the connected regular graphs has exactly k neighbors. Denote n_i as the fraction of nodes in subpopulation i, with i = 1..m. Obviously, it holds that $\sum_{i=1}^{m} n_i = 1$. For every node in subpopulation i we denote the curing rate as δ_i, and the infection rate of all incoming links as β_i, with i = 1..m. Our assumptions imply that we are considering bi-directional links where the infection rate in the two directions in general is not equal. The latter condition also reflects the assumption that the rate of infection is determined by the node itself, for instance by the type of software it is running. We will come back to this assumption later on in the paper.

It is important to note that our assumptions imply complete symmetry, each node sees the same fraction of nodes from every subpopulation. So every node has a fraction n_1 of neighbors from subpopulation #1, a fraction n_2 of neighbors from subpopulation #2 and so on. Therefore, the number of subpopulations should not exceed the number of direct neighbors, or m ≤ k.

For subpopulation i at time t, we denote the number of infected nodes as $X_i(t)$ and the fraction of infected nodes as $v_i(t)$. Then, the probability that a randomly chosen node within subpopulation i is infected in the total population with N nodes is $v_i(t) \equiv \frac{X_i(t)}{Nn_i}$.

The rate at which the probability of infection for nodes in subpopulation i changes is due to two processes: susceptible nodes becoming infected and infected nodes being cured. The curing rate for an infection probability v_i is $\delta_i v_i$. The rate at which the probability v_i grows is proportional to the probability of a node in subpopulation i being susceptible, i.e. $1-v_i$. For every susceptible node the rate of infection is the product of the infection rate per node in that subpopulation (β_i) and the probability that on a given link the susceptible node connects to an infected node is$k \sum_{j=1}^{m} n_j v_j$. Therefore, the following system of differential equations describes the time evolution of $v_i(t)$ with i, j =1..m: $\frac{dv_i}{dt} = \beta_i k \left(\sum_{j=1}^{m} n_j v_j \right) (1 - v_i) - \delta_i v_i$ (1)

Note that for $\delta_1 =..= \delta_m$, and for $\beta_1 =..= \beta_m$, the system of equations (1) reduces to a single differential equation, describing the general solution for a homogeneous population with
$v = \sum_{j=1}^{m} n_j v_j$. (2)

For the general case with different curing and infection rates, it is impossible to obtain an explicit solution for the system of equations (1).

Theorem 1. *If the effective spreading rate* $\tau = \sum_{i=1}^{m} \frac{\beta_i n_i}{\delta_i}$ *for a system of m differential equations in Eq. (1), then the epidemic threshold satisfies* $\tau_c = \frac{1}{k}$.

Proof. We will use a Lyapunov function [12] to show that, under the condition$\sum_{i=1}^{m} \frac{\beta_i n_i}{\delta_i} \leqslant \frac{1}{k}$, the origin is a global attractor for {$v_1 \geq 0$, $v_2 \geq 0$, .., $v_m \geq 0$}, hence, that the virus dies out.

Let $V = \sum_{i=1}^{m} \frac{n_i v_i}{\delta_i}$, then we have $\frac{dV}{dt} = \left(k \sum_{i=1}^{m} \frac{\beta_i n_i}{\delta_i} (1 - v_i) - 1 \right) \sum_{i=1}^{m} n_i v_i$. (3)
Because $v_i \geq 0$, it follows that $1-v_i \leq 1$. Therefore Eq. (3) implies that $\frac{dV}{dt} \leq k \sum_{i=1}^{m} \frac{\beta_i n_i}{\delta_i} - 1 \sum_{i=1}^{m} n_i v_i$. Hence under the condition$\sum_{i=1}^{m} \frac{\beta_i n_i}{\delta_i} \leq \frac{1}{k}$ it holds that $\frac{dV}{dt} \leq 0$. The claim follows directly by applying Lyapunov's stability theorem.

Next, we consider the case $\sum_{i=1}^{m} \frac{\beta_i n_i}{\delta_i} > \frac{1}{k}$. We first note that any trajectory of the system (1) can never leave the box B={(v_1, .., v_m) | $0 \leq v_1 \leq 1$, .., $0 \leq v_m \leq 1$}. This follows from $\frac{dv_1}{dt} |_{v_1=0} = \beta_1 k \sum_{i=1}^{m} n_i v_i \geq 0$ and similar inequalities at the borders of the box B. From the construction of the above Lyapunov function V, we can see that for $\sum_{i=1}^{m} \frac{\beta_i n_i}{\delta_i} > \frac{1}{k}$ and for (v_1, .., v_m) in B and sufficiently close to the origin, $\frac{dV}{dt} > 0$. This implies that the origin has an unstable manifold in B. Therefore, since any trajectory of system (1) can never leave the box B, system (1) has an attractor as the ω-limit set and the virus survives. This finishes the proof of the theorem.

If the graphs considered are limited to connected regular graphs where each node has exactly k neighbors, then calculations can be simplified by introducing the basic reproduction numbers $R_i = \frac{\beta_i k}{\delta_i}$. Under the condition $\forall i = 1..\mathrm{m} \rightarrow \delta_i > 0$, Eqs. (1) and (2) lead to:
$\frac{dv_i}{dt} = \delta_i R_i v (1 - v_i) - v_i$ (4)

Theorem 2. *For a system of m differential equations in Eq. (4), the steady state of (v) can be calculated by solving a polynomial equation of order m.*

Proof. Solving Eq. (4) leads to:$\frac{dv_i}{dt} = 0 \rightarrow v_i = \frac{R_i v}{R_i v + 1}$ (5)

If v≠0 then Eq. (2) and (5) lead to: $\sum_{i=1}^{m} \frac{n_i R_i}{R_i v + 1} = 1$ (6)

Thus, when $\sum_{i=1}^{m}(n_i R_i) \leqslant 1 \rightarrow v \leqslant 0; \sum_{i=1}^{m}(n_i R_i) > 1 \rightarrow v > 0.$ (7)

Eq. (7) confirms the epidemic threshold found in Theorem 1. Next, we use Eq. (6) to calculate the steady state: $h(v) = \prod_{i=1}^{m}(R_i v + 1) - \sum_{i=1}^{m} n_i R_i \prod_{j=1, j \neq i}^{m}(R_j v + 1) = 0.$ (8)

So h(v) is a polynomial equation of order m with these preconditions: $\{n_1, .., n_m \in \langle 0..1 \rangle \wedge R_1, .., R_m > 0\}$. Solving h(v) provides the steady state of v and by filling in v in Eq. (5) the steady state of every v_j. This finishes the proof of the theorem.

Theorem 3. *If β_1 = .. = β_m, then Eq. (1) has m – 1 solutions in the form of hyper-planes passing through the origin. The intersection of the hyper-planes is also a solution of Eq. (1) and its dynamics are described by a logistic equation.*

Proof. Assuming $v_j = \lambda_j v_1$ (i=2..m), where the λ_j ‘s are constants, it follows that

$\frac{dv_j}{dt} - \lambda_j \frac{dv_1}{dt}|_{v_j = \lambda_j v_1} \equiv -v_1(c_2 \lambda_j^2 + c_1 \lambda_j + c_0) = -v_1 f(\lambda_j)$ with

$c_2 = n_j k(v_1(\beta_j - \beta_1) + \beta_1); c_1 = n_1 v_1 k(\beta_j - \beta_1) + n_1 \beta_1 k - n_j \beta_j k + \delta_j - \delta_1; c_0 = -n_1 \beta_j k.$ (9)

Hence, if β_1 = .. = β_m, then c_0, c_1 and c_2 are constants. Then, because $c_2 > 0$ and $c_0 < 0$, it follows that $f(\lambda_j)$ has exactly one positive root λ^*_j, for i=2..m. Therefore, the hyper-planes $v_j = \lambda^*_j v_1$ (i=2..m) are solutions of Eq. (1), when β_1=..=β_m. Using the first equation in Eq. (1), we can show that on the intersection of the m – 1 hyper-planes, the dynamics are described by a logistic equation:

$\frac{dv_1}{dt} = \beta_1 k v_1 \left(\sum_{j=1}^{m} n_j \lambda^*{}_j\right)(1 - v_1) - \delta_1 v_1$, (10)

where $\lambda^*_1 = 1$.

This concludes the proof of the theorem.

3. CALCULATION METHOD IN DETAIL FOR TWO SUBPOPULATIONS

In this section, the least complex heterogeneous situation is explored in more detail. By filling in m=2 in Eq. (8) it follows that:

$h(v)=R_1R_2v^2+(R_1+R_2 - R_1R_2)v+1 - n_1R_1 - n_2R_2=0.$ (11)

The method always yields a solution for v, since in Eq. (11) the discriminant d > 0. For two subpopulations with parameters $\{n_1, n_2, R_1, R_2\}$ d is calculated as:

$d=R_1^2R_2^2+(R_1 - R_2)^2+2R_1R_2(R_1 - R_2)(n_1 - n_2).$ (12)

We know from Theorem 1 that for $n_1R_1+n_2R_2 \leq 1$ system (4) with m=2 the virus dies out, i.e. v=0 is the global attractor. Hence we only consider the case $n_1R_1+n_2R_2>1$. Then, because h(0)<0 and h(1)>0, it follows that there is a unique solution 0<v<1 for Eq. (11). A simple calculation shows that the only solution satisfies $v = \frac{R_1R_2 - R_1 - R_2 + \sqrt{d}}{2R_1R_2}$.

Next, we will show that for Eq. (4) with m=2, under the condition $n_1R_1+n_2R_2>1$, the steady state of v corresponds to a stable equilibrium point of the system in Eq. (4).

It follows from the analysis above that for $n_1R_1+n_2R_2>1$ system (4) with m=2 has an equilibrium point located in the region $A=\{(v_1, v_2) \mid 0<v_1<1, 0<v_2<1\}$.

We know from the proof of Theorem 1 that for $n_1R_1+n_2R_2>1$ the origin has an unstable manifold entering A, while trajectories of the system can never leave the region A. Therefore, by application of the Poincaré-Bendixson theorem [12] on A, the ω-limit set for system (4) for m=2, can be either an equilibrium point or an isolated periodic orbit. To rule out the existence of periodic orbits for system (4) with m=2, we can use the Bendixson-Dulac criterion, see [12]. In fact, because periodic orbits cannot intersect $v_1=0$ or $v_2=0$, we can use the Dulac function

$D(v_1, v_2) = \frac{1}{v_1 v_2}$ which leads to

$$\frac{\partial(D\frac{dv_1}{dt})}{\partial v_1} + \frac{\partial(D\frac{dv_2}{dt})}{\partial v_2} = -(\delta_1 n_1 R_1 v_1^2 v_2 + \delta_1 n_2 R_1 v_2^2 + \delta_2 n_1 R_2 v_1^2 + \delta_2 n_2 R_2 v_1^2 v_2)D < 0. \quad (13)$$

Therefore, the system in Eq. (4) with m=2 has no periodic orbits and hence the equilibrium point, corresponding with the steady state v, is globally stable.

Next, we will assess the impact of the fraction nodes of type 1, i.e. n_1, on the value of v, i.e. the fraction of infected nodes. From Eq. (9) it is easy to verify that

$$\frac{\partial v}{\partial n_1} = \frac{R_1 - R_2}{\sqrt{d}}. \quad (14)$$

Hence v is always a monotonic function of n_1, unless R_1=R_2, which corresponds to the homogeneous case. Figure 1 depicts several possibilities for the case R_2=2.

For n_1=0, the homogeneous case were all nodes belong to type 2, the fraction of infected nodes v equals 1–1/R_2=0.5. For the case R_1>R_2, according to Eq. (14), v increases monotonically with n_1 hence for this case the homogeneous case n_1=0 gives the least number of infected nodes.

Clearly, R_1=R_2 is the homogeneous case with v=0.5.

If 1<R_1<R_2 then v decreases monotonically with n_1, while for n_1=1 it holds that v>0. Hence for this case the homogeneous case n_1=1 gives the least number of infected nodes. Finally, if R_1<1 then v decreases monotonically with n_1 while for n_1=1 it holds that v≤0. Hence for this case the optimal situation, where the virus dies out, occurs from $n_1 = \frac{R_2 - 1}{R_2 - R_1}$ onwards.

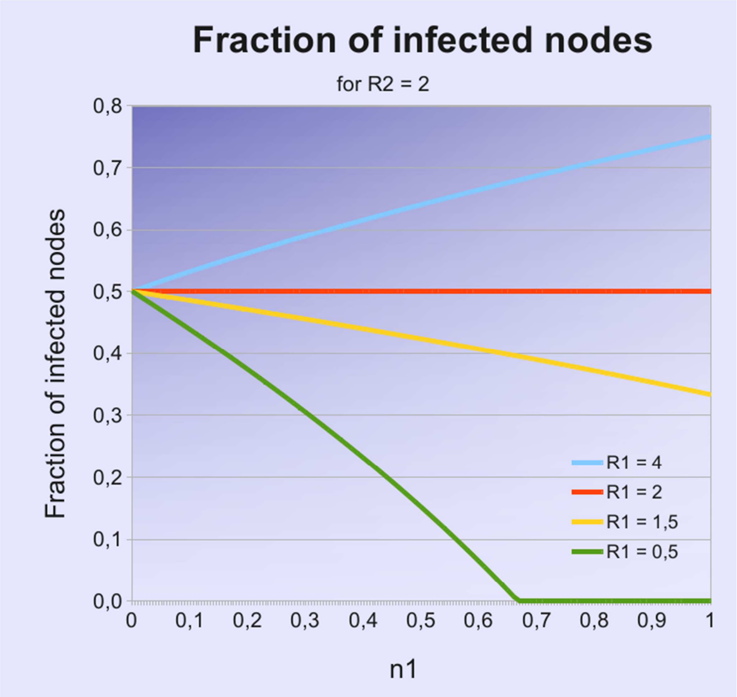

FIGURE 1: The Steady State of v as a function of n_1

We would like to stress once more, that our model is more general than previous models, see e.g. [2], [4], [7, [8], [11], because none of these models can deal with heterogeneous spreading rates. As a simple example, consider the case with two malware populations, with n_1 = 0.7, β_1 = 0.4, δ_1 = 1 and n_2 = 0.3, β_2 = 0.05, δ_2 = 1, where every node has four neighbors, i.e. k = 4. If we would want to apply the results of [2], [4], [7, [8], [11], and we would use for the spreading rate β the mean of β_1 and β_2, then the effective spreading rate would become 0.225, which is below the epidemic threshold 1/k = 0.25. However, applying our, more accurate model, we arrive at the conclusion that the malware persists, because $n_1 \beta_1 + n_2 \beta_2 = 0.295 > 0.25 = 1/k$.

4. SPECIAL CASE #1: MALWARE RESERVE

In this section we consider the case that subpopulation #2 has no defense against malware, i.e. δ_2=0, or equivalently, R_2=∞. Then it follows from Eq. (1) that in steady state, the whole subpopulation #2 is infected, i.e. in steady state v_2=1 holds.

Using Eqs. (2), (5) and taking lim $R_2 \to \infty$, it is easy to show that the steady state fraction of infected nodes v satisfies $v = \frac{R_1 - 1 + \sqrt{R_1^2 + 2(n_1 - n_2)R_1 + 1}}{2R_1}$.

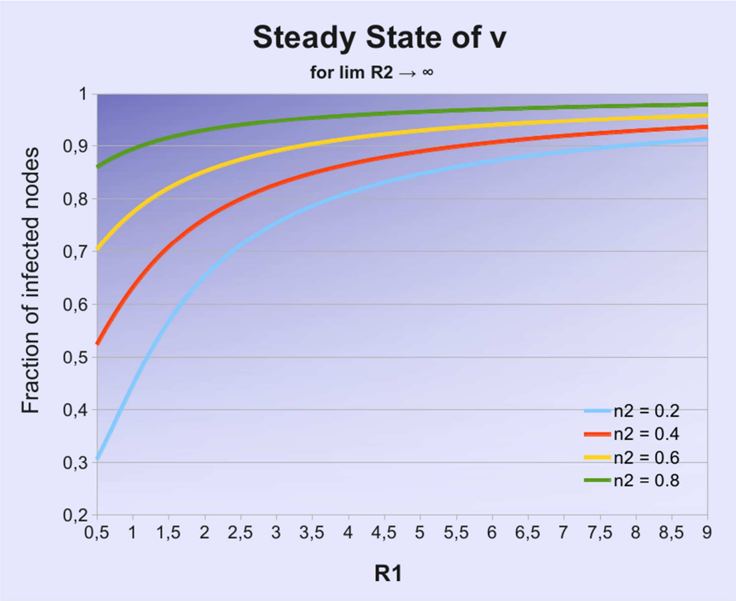

FIGURE 2: Steady state v for the case δ_2=0

Figure 2 shows the steady state v as a function of R_1, for different values of n_2, the fraction of nodes that belong to the malware reserve. We observe that the whole population benefits if the size of the malware reserve is decreased. We also see that even if subpopulation #1 has adequate protection against malware (i.e. R_1<1), they still become infected because of the lack of security for subpopulation #2. For instance, if $R_{1=}$0.5 and n_2=0.2, then in steady state 13% of subpopulation #1 is infected. This is easily verified from Figure 2 and the equality $v = n_1 v_1 + n_2$, which holds for δ_2=0.

5. SPECIAL CASE #3: DIVERSIFICATION AGAINST MALWARE

In this section we consider the case that subpopulation #2 is immune for malware infections, i.e. β_2=0, or equivalently, R_2=0. Then it follows from Eq. (1) that in steady state, the whole subpopulation #2 is uninfected, i.e. in steady state v_2=0 holds.

Using Eq. (12) with R_2=0 it follows that the steady state fraction of infected nodes v satisfies

$$v_1 = 1 - \frac{1}{n_1 R_1}; v_2 = 0 \to v = n_1 - \frac{1}{R_1}. \qquad (15)$$

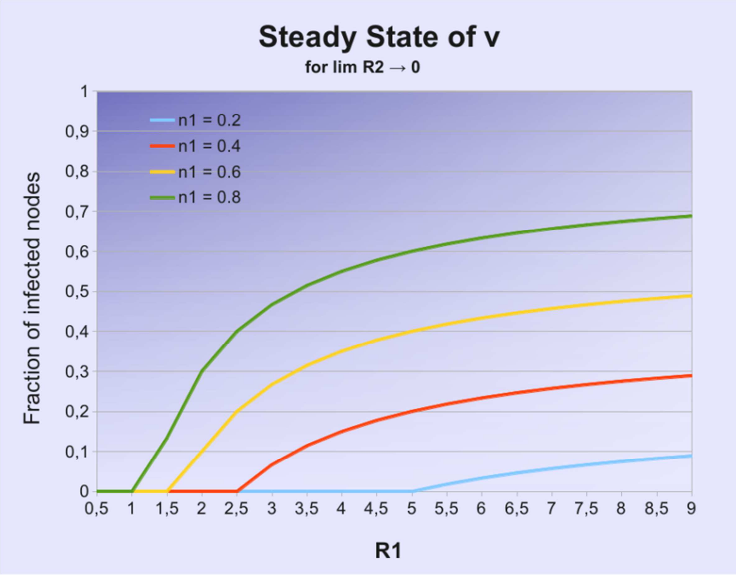

FIGURE 3: Steady state v for the case β_2=0

Figure 3 shows the steady state v as a function of R_1, for different values of n_1, the fraction of nodes not immune to the malware. We observe that the whole population benefits if the size of the immune population increases. We also see that in case of an immune subpopulation, there exists a threshold value for the basis reproduction number R_1, below which the virus dies out. It is clear from Eq. (15) that this threshold value satisfies $R_1 = \frac{1}{n_1}$.

This threshold value for R is the minimum value necessary to “sustain” an infection level above zero in that compartment. The lower the threshold value, the more malware is able to match it. In Table 1 the threshold value (R) was calculated for popular software using market shares [13]. In this example, the software for which the most malware is expected is MS Windows, MS Office, MS Internet Explorer and Mozilla Firefox – in that order. However, market share statistics vary to much to draw explicit conclusions based on the numbers presented.

6. CORRELATING SECURITY MEASURES WITH MODEL PARAMETERS

Like all models, the SIS model is an approximation of reality. It should be applied with care and respect for its limitations and premises. One of its limitations is that it is necessary to assume that the population is completely symmetrical, i.e. the different nodes are distributed evenly in the network.

	Software	Market	Threshold	
Platform	**Platform**	**Share 3Q10**	**R**	**Remark**
Webclient	Windows	0,8821	1,13	
Webclient	MacOS	0,0682	14,66	
Webclient	Linux	0,0108	92,59	
Webclient	Symbian	0,0021	476,19	
Webclient	Blackberry	0,0045	222,22	
Webclient	Other	0,0323		
Webbrowser	IE	0,4622	2,16	
Webbrowser	Firefox	0,2992	3,34	
Webbrowser	Chrome	0,1240	8,06	
Webbrowser	Safari	0,0555	18,02	
Webbrowser	Opera	0,0193	51,81	
Webbrowser	Other	0,0398		
Office Suite	MS Office	0,8800	1,14	(Dutch market share only)
Office Suite	OpenOffice	0,0800	12,5	(Dutch market share only)
Office Suite	Wordperfect	0,0090	111,11	(Dutch market share only)
Office Suite	Other	0,0310		

TABLE 1: Calculation Example of Threshold Values for popular Software

In its simplest, homogeneous form, the value of the single set of parameters of the SIS-model (β, δ, R) are determined by the effectiveness of security measures taken by the defenders and the attempts of the attackers to outsmart them (see Table 2).

	Cyber Defense (Reduce R)	**Cyber Attack (Increase R)**
β	***Prevent malware infection:*** • Intrusion prevention system, firewall, heuristic AV software for on-access scanning • “white-list” software, USB media, e-mail content, web content to download • Configuration: restrict user rights, harden systems, sandboxing, good passwords frequently changed,... • Separate compartments: network, software, user rights, encrypted files • Good procedures for changes / updates • Invest in knowledge and awareness • Better, less vulnerable software • Preventive security audits • Legal software for employees @home	***Increase risk of infection by malware:*** • IP / MAC address spoofing • Multiple attack patterns in malware • Web site offers customized malware • Domain Generation Algorithms (DGA) • Social engineering, imitation of legitimate software, e.g. AV software • Malware on trusted sites • Sharing / stealing knowledge, source code • Commercial and freeware Malware kits • Fuzz testing of software for vulnerabilities • Testing malware with security software • Stealth malware, encryption, code obfuscation • Massive and rapid spread of malware (reverse engineering of patches?) • Targeted malware, APT (“precision ammo”)
δ	***Improve disinfection (detection+correction):*** • Multiple AV packages for scheduled scans • Intrusion Detection System, logging • Management procedures for incidents and changes, including an Incident Response Plan • Invest in knowledge and awareness • Postmortem security audits • Periodically re-install clean software image on all PC’s	***Reduce loss of infected computers:*** • Root kits, anti-virtualization techniques, disable security software and update mechanisms • Encryption, remove trace data, multiple layered code obfuscation • Malware self-activation / self removal under certain conditions • Malware updates faster than AV • Patching of infected computers (!) • Continuity plan for botnet, e.g. rotating web servers, integrate infected computer in >1 botnets, bullet proof hosting of C&C servers • Imitation behavior of legitimate software

TABLE 2: The Battle Between Cyber Attack and Cyber Defense

The infectivity of different occurrences of malware may vary widely, depending on the knowledge of the attacker and the purpose of the malware. For this moment we neglect the differences between malware samples, because we like to focus on the protective measures of the defenders. The infectivity of all malware is then considered to be equal.

When individuals or organizations use different security measures, the parameter (R) is likely to differ. For instance, if one organization prohibits the user to install software, this will reduce the risk of infection considerably.

Correctly estimating the corresponding value of the parameter (R) from the operational security measures is difficult. The Eurostat Newsrelase of Feb 8, 2011 presents a list of statistics on Internet security of the EU countries in 2010.

One of the statistics was the percentage of individuals who reported that they caught a computer infection resulting in loss of information or time using the Internet in the 12 months prior to the survey. Another statistic was the percentage of individuals who used the Internet in the last 12 months and stated that they used IT security software to protect their private computer and data.

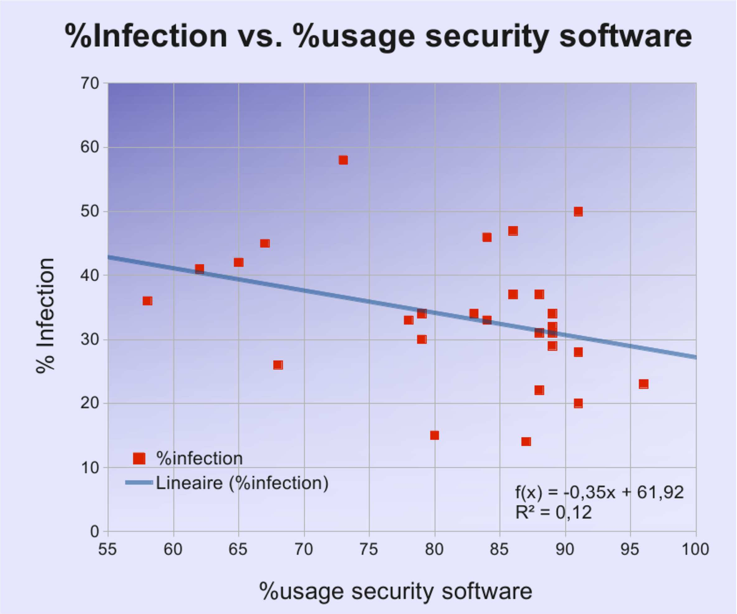

FIGURE 4: Correlating the usage of security software and percentage of infections

After deleting the incomplete data entry of Romania, both statistics were plotting against each other using an XY diagram (see Figure 4). When we treat this population as homogeneous, the lowest value of 14 percent infections and highest value of 58 are far off the mean value of 33.45. The usage of security software vs. the (resulting) percentage of infections seems a clearcut case, but the data shows little correlation - the trend line shows an R^2 value of 0.12. With a total variance of 100.54 for 29 countries, this model's accuracy is low.

We can decrease the total variance by introducing subpopulations with their own mean values. Sorting the list of countries using the percentage of infection, we can divide the population in a group of leaders and a group of laggards. We have determined that the minimal variance of the total population is 41.29 when the group of leaders consists of 22 countries (mean value is 29.14) and the group of laggards consists of 7 countries (mean value is 47.0).

Interestingly, the higher accuracy has little influence on the correlation per subpopulation between the usage of security software vs. the resulting percentage of infections; in fact, for the group of laggards, the correlation line shows that the higher the usage of security software is, the higher the resulting percentage of infections becomes. A possible explanation is that security software is necessary to detect malware infections and that the laggards are more often tricked in using bogus anti-virus software, which in fact is malware. However, even the correlation in the leaders group has dropped to 0.06, so the usage of security software seems to be a poor predictor of the resulting percentage of infections.

7. CONCLUSIONS

We have introduced a new method to calculate the steady state for heterogeneous populations. Based on analysis of this method, we think that a heterogeneous model can be accurately matched with a logistic function. We can also predict that either a minimal or maximum value of the infected fraction of the population occurs when the heterogeneous population becomes homogeneous. Thus, when the security level of the least secure group increases or the fraction of this group decreases, the whole population benefits from this.

The analysis also reveals that a minimum occurs when the population ceases to be a mono-culture, i.e. not every node shares the same vulnerabilities for malware. More diversity is an effective measure against the propagation of malware. Although all separate compartments can attract malware, the existing measures become more effective and the total level of infections is less that in a similar population which all use the same hard- and software. The opposite is also true: the bigger a mono-culture is, the less infectious malware has to be to persist.

Finally, we like to suggest some directions for future research. First of all, since it has been proven difficult to correlate model parameters of the (heterogeneous) SIS model to the use of security software, future research can be directed to derive realistic parameter values for other (clusters of) security measures. Secondly, our presented model assumes a static topology, whereas in real-life, computer networks are changing in time. In line with [10], where homogeneous virus spread for dynamic networks is considered, we suggest to generalize our results for heterogeneous malware populations to dynamic networks. Finally, although our heterogeneous model incorporates different infection and curing rates per subpopulation (β_i, δ_i, R_i) , our results depend on the assumption of a form of complete symmetry in the network, i.e. every node is connected to the same fractions of nodes from each subpopulation. More research is needed to reveal how the results are influenced when this symmetry assumption is dropped.

8. REFERENCES

[1] MessageLabs Intelligence. “2010 Annual Security Report”, December 7, 2010 http://www.inteco.es/file/27gHxrzWsYyeyRTFYq8MuQ [2012-10-05]

[2] J.O. Kephart and S.R. White. “Direct-graph epidemiological models of computer viruses”, Proc. IEEE Computer Society Symposium on Research in Security and Privacy, pp. 343-359, 1991.

[3] R. Pastor-Satorras and A. Vespignani. “Epidemic Spreading in Scale-Free Networks”, Physical Review Letters, Vol. 86, No. 14, April, 3200-3203, 2001.

[4] A. Ganesh, L. Massoulié and D. Towsley. “The Effect of Network Topology on the Spread of Epidemics”, Proc. IEEE INFOCOM.05, Miami, 2005.

[5] N.T.J. Bailey. “The Mathematical Theory of Infectious Diseases and its Applications”, London: Charlin Griffin & Company, 2nd ed., 1975.

[6] D.K. Daley and J. Gani. “Epidemic modelling: An Introduction”, Cambridge University Press, 1999.

[7] Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos. “Epidemic spreading in real networks: An eigenvalue viewpoint”, IEEE Computer Society, 22nd International Symposium on Reliable Distributed Systems (SRDS’03), pages 25—34, Los Alamitos, CA, USA, 2003.

[8] P. Van Mieghem, J. Omic, and R.E. Kooij. “Virus spread in networks”. IEEE/ACM Transactions on Networking, 17(1), 1-14, 2009.

[9] Y. Wang and C. Wang. “Modeling the Effects of Timing Parameters on Virus Propagation”. ACM Workshop on Rapid Malcode, Washington, DC, Oct. 27, 2003.

[10] T. Gross, C. Dommar D’Lima and B. Blasius. “Epidemic dynamics on an adaptive network”, Physical Review Letters 96, 208701–4, 2006.

[11] J. Omic, R.E. Kooij and P. Van Mieghem. “Heterogenous protection in regular and complete bi-partite networks”, Proc. of Networking 2009, Aachen Germany, 11-15 May, 2009.

[12] J. Guckenheimer and P. Holmes. “Nonlinear oscillations, dynamical systems, and bifurcations of vector fields”, New York: Springer, 1983

[13] See for market share used (OS, Browser and Office software) [2012-05-20]: http://marketshare.hitslink.com/operating-system-market-share.aspx?qprid=8; http://marketshare.hitslink.com/browser-market-share.aspx?qprid=0&qpcustomd=0&qptimeframe=M&qpsp=155; www.webmasterpro.de/portal/news/2010/02/05/international-openoffice-market-shares.html