# Chapter 9
# Robustness to Malware Reinfections

In Chap. 8, we analyzed a deterministic epidemiological model where an infected node immediately contaminated all its neighbors of the same type. The spreading process was completely determined by the spreading network's topology, the configuration of node types, and the selection of initially infected nodes or seeds. Because no node recovered from an infection, there were no long-term dynamics. The spreading simply stopped when all reachable nodes were infected. Here, we study a stochastic epidemiological model of multimalware outbreaks where arbitrary but fixed probabilities determine whether nodes are infected. Furthermore, nodes recover from infections with given probabilities, only to be reinfected later. An incident from 2007, where the same worm repeatedly infected the internal networks of a Norwegian bank, illustrates how reinfections can occur in real networks.

The stochastic epidemiological model is first used to determine how to immunize unknown hubs on diverse inhomogeneous spreading networks. A simple solution is obtained by generalizing the acquaintance immunization strategy for monoculture networks [25]. Next, the model is analyzed to determine the software diversity required to halt multimalware spreading in homogeneous spreading networks where nodes can be infected multiple times by the same malware. The analysis produces a lower bound on the diversity needed to ensure that, with a high probability, the malware outbreaks do not spread far but, instead, die out quickly. The lower bound is obtained by modifying a "classical" result from network science [87]. A reader unfamiliar with differential equations can skip the development of the bound, since it is not needed to understand the remaining chapters.

## 9.1 Malware Attack on a Norwegian Bank

In March 2007, Viking.gt, a worm well known to anti-malware software vendors at the time, attacked office computers belonging to a large bank in Norway. The worm, most likely residing on an external game site, entered the bank's internal networks via a computer at a branch office and cascaded across roughly 1,000 servers and



K.J. Hole, *Anti-fragile ICT Systems*, Simula SpringerBriefs on Computing 1,
DOI 10.1007/978-3-319-30070-2_9

11,000 office PCs in 190 branch offices. This cascade failure occurred because the anti-malware software running on the bank's computers did not stop infections, even though the anti-malware software was updated with an appropriate signature for the worm. During the attack, branch offices could not assist their customers with many tasks and some offices closed.

About 200 individuals worked two shifts to remove the worm. Because the worm disabled the machines' automatic software update mechanism, the worm had to be manually removed from each machine. The vendor's initial upgrade to the anti-malware software was flawed, allowing infected machines to reinfect cleaned machines over internal networks. The bank decided to close all connections to the Internet to protect their internal systems from further external infection. It then took days to remove the malware from the internal networks.

The next section presents a malware spreading model that allows malware to reinfect nodes. This stochastic model is a generalization of the deterministic explanatory model in Chap. 8.

## 9.2 Stochastic Epidemiological Model

Let a networked computer system be infected by different types of malware. The malware's spreading network is modeled as an undirected graph with $M$ edges and $N$ nodes of different types. The node types represent machines with distinct software at the operating system or application layer and the edges represent virtual communication lines. There is at most one edge between two nodes and no edge connects a node to itself. If there is an edge between two nodes, then these nodes are neighbors. The degree $k$ of a node is the number of neighbors. The nodes' average degree is $\langle k \rangle = (2M)/N$.

As noted in Chap. 8, the topology of the spreading network depends on the software layer and the vulnerabilities exploited to spread the malware. We discriminate between an *inhomogeneous* network containing a few nodes, the hubs, with a very large degree $k \gg \langle k \rangle$ and a *homogeneous* network where all nodes have a degree $k \approx \langle k \rangle$. Any spreading network has $L$ different node types $l = 1, 2, \ldots, L$ for $1 \leq L \ll N$. Each node type occurs $N_l$ times. A node chosen uniformly at random is of type $l$ with probability $N_l/N$ for $N = \sum_l N_l$. When $N_l = N/L$, the diversity is equal to the number of node types $L$ with the convention that a monoculture network with only one type has no diversity [50].

A node of arbitrary type $l$ is either susceptible to an infection or it is infected. If an infection is removed from the node, then it immediately becomes susceptible to a new infection. There are $L$ types of malware, where each type of malware infects a particular software platform, that is, node type. Because there are $L$ nodes types with $L$ corresponding malware types, the complete spreading network can be viewed as $L$ disjoint subnet monocultures, each containing a single node type.

Multiple simultaneous malware epidemics are modeled by $L$ susceptible–infected–susceptible (SIS) models [23, 87] operating on the same network topology but affecting $L$ disjoint subnet monocultures. Initially, all the nodes are susceptible. At time step $t = 0$, the model selects uniformly at random $S$ ($\geq 1$) nodes of each type $l$ and infects them. These $L \cdot S$ initially infected nodes are the seeds. For each time step $t = 1, 2, 3, \ldots$, any infected node of type $l$ infects any susceptible neighbor of type $l$ with *infection probability* $p_l, 0 < p_l \leq 1$. At the same time, any infected node of type $l$ recovers with *recovery probability* $q_l, 0 \leq q_l \leq 1$.

If $q_l > 0$ for some $l$, then a node can repeat the SIS life cycle many times. The result is a stochastic model with long-term dynamics, where it is assumed that the infections and recoveries are updated in a random asynchronous order. When $p_l = 1$ and $q_l = 0$ for all $l$, the SIS models become $L$ susceptible–infected (SI) models. The overall spreading model is deterministic in this case, since malware infects all reachable nodes with 100 % probability. Consequently, the spreading process is completely determined by the network's topology, the configuration of node types, and the selection of seeds. Because no node recovers from an infection, there are no long-term dynamics. The spreading simply stops when all reachable nodes are infected. This special case of $L$ deterministic SI models was first presented in Chap. 8 for $N_l = N/L$.

## 9.3 How to Immunize Unknown Hubs

While we may not know the degrees of many nodes in inhomogeneous spreading networks, it is still possible to immunize hubs in advance of malware outbreaks. The acquaintance immunization strategy [25] provides an elegant solution to the problem of immunizing unknown hubs in a monoculture ($L = 1$): Choose a set of nodes uniformly at random and immunize one arbitrary neighbor per node. While the original set of nodes is unlikely to contain the relatively few hubs in an inhomogeneous network, the randomly selected neighbors are much more likely to be hubs, since many edges are adjacent to high-degree nodes.

We can generalize acquaintance immunization to diverse networks. Assume that it is possible to estimate the number of nodes $N_l$ of each type $l$ in a diverse network, perhaps by estimating the total size of the network and then determining the percentages of different node types in a small part of the network. For some fraction $0 < f < 1$, choose a set of $f \cdot N_l$ nodes of type $l$ uniformly at random such that each node has at least one neighbor of the same type, $l = 1, 2, \ldots, L$. Immunize one randomly selected neighbor of type $l$ per node in the set. When the number of immunized neighbors $f \cdot N_l$ of each type $l$ is large enough, most hubs are immunized [25].

To illustrate acquaintance immunization on diverse spreading networks, we consider an inhomogeneous network with dominant hubs. The network has 10,670 nodes and 22,002 edges. The largest hub has degree 2,312, which is nearly 11 % of the total number of edges. The $L = 7$ node types have different colors and the size of a node is proportional to its degree; that is, hubs are larger than low-degree nodes.

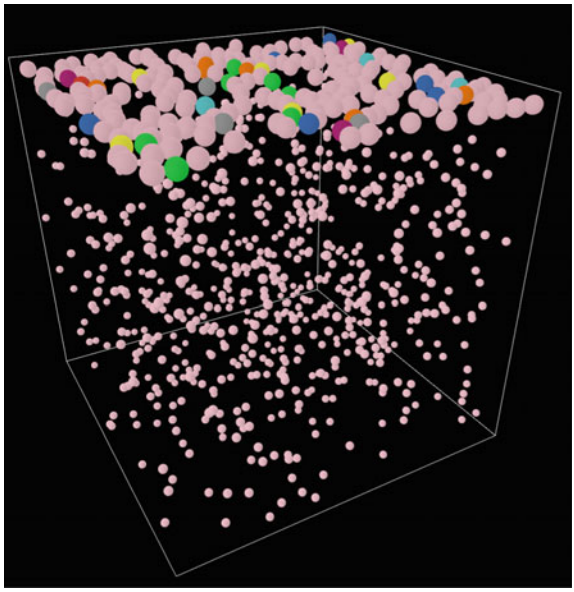

**Fig. 9.1** Acquaintance immunization of a network with enlarged hubs at the *top*. Only immunized nodes and susceptible hubs are shown. Note that most hubs are immunized

A node turns pink when it is immunized. Let the fraction of immunized neighbors be $f = 0.04$ (4 %). Figure 9.1 shows only the immunized pink nodes and the remaining susceptible multicolored hubs after acquaintance immunization. No edges or susceptible low-degree nodes are shown. Note that most of the 216 enlarged hubs are immunized. Assume $S = 20$ seeds per node type for a total of $7 \cdot 20 = 140$ seeds. Let $p_l = 0.06$ and $q_l = 0.04$. When acquaintance immunization is performed in advance, the fraction of infected nodes goes to zero after only a few hundred time steps.

## 9.4 Lower Bound on Required Diversity

In this section we determine a lower bound on the node diversity $L$ needed to make it very likely that new malware outbreaks will die out before they spread to many nodes. We assume that all hubs are immunized, even though acquaintance immunization may miss a few. Because the hubs and their adjacent edges do not partake in the transmission of malware, we model the spreading network after hub immunization as a random homogeneous network with $N$ nodes, average degree $z = \langle k \rangle$, and $N_l = N/L$ nodes of each type, $l = 1, 2, \ldots, L$. The spreading network has a fixed but arbitrary (thin-tailed) degree distribution. Note that modeling the remaining spreading network after hub immunization as a random network is an approximation chosen because random networks are malleable to analysis [23].

The average fraction of infected nodes of type $l$, denoted $h_l$, is estimated by considering the subset of $N/L$ nodes of type $l$. To estimate $h_l$, we extend an analytical

technique for random networks viewed as homogeneous monocultures ($L = 1$) [23, 87]. Each malware outbreak in a network with $L > 1$ node types operates on a subgraph with $N/L$ nodes of the same type. On average, a node has $z/L$ neighbors in the subgraph because the probability that a node is of type $l$ is $N_l/N = 1/L$. Let the spreading rate be $\rho_l = (p_l z)/(q_l L)$ for $q_l > 0$, $l = 1, 2, \ldots, L$, and view $h_l = h_l(t)$ as a continuous-time variable. Representing the expected change in the fraction of infected nodes as the differential equation

$$\frac{dh_l}{dt} = p_l \frac{z}{L} h_l(1 - h_l) - q_l h_l$$

and imposing the stationary condition $dh_l/dt = 0$, we find that the average fraction of infected nodes saturates at $h_l = 1 - 1/\rho_l$ for $\rho_l > 1$. The fraction $h_l$ goes to zero in finite time when $\rho_l < 1$. For a fixed infection probability $p_l$, recovery probability $q_l$, and average degree $z$, the spreading rate $\rho_l = (p_l z)/(q_l L) < 1$ when the number of node types $L > (p_l z)/q_l$. Consequently, $h_l$ goes to zero.

Since we need $h_l$ to go to zero for all $l$, the needed node diversity is lower bounded by

$$L > z \cdot \max_l \left\{ \frac{p_l}{q_l} \right\}, \tag{9.1}$$

where $z = \langle k \rangle$ is the average node degree of the remaining spreading network after hub immunization. The largest spreading rate essentially determines the required diversity $L$.

It is possible to estimate the lower bound in inequality (9.1) for real malware types by estimating the infection probabilities $p_l$ and recovery probabilities $q_l$. However, the inequality is first and foremost important because it shows that multiple simultaneous malware outbreaks with the ability to reinfect nodes will die out before they can spread far when the software diversity is large enough, given that hubs are immunized.

## 9.5 Discussion and Summary

A combination of acquaintance immunization and node diversity prevents malware with the ability to reinfect nodes from creating long-lasting epidemics. Through immunization of most of the hubs and a sufficient increase in node diversity, malware outbreaks are likely to die out quickly. Hence, acquaintance immunization and node diversity together provide robustness to malware reinfection.

As first stated in Sect. 8.7, graph coloring algorithms can be used to ensure that no (or very few) pairs of neighboring nodes have the same color or node type. Coloring algorithms exploit the topology of static spreading networks to reduce the number of node types needed to prevent malware propagation, compared to our simple approach of just randomly assigning node types. Why, then, are we using this simple approach

in both the previous and current chapters when it does not minimize the number of different colors needed to prevent malware spreading? There are two main reasons.

First, while coloring algorithms need central processing to assign node types, our simple scheme requires no central control. We cannot use algorithms requiring central control to assign node types because of their limited scalability. Our goal is a malware-halting technique that scales to millions of nodes. Second, the topologies of the malware spreading networks are not known and, even if they were, networks will vary over time, making it necessary to constantly rerun the coloring algorithms to reassign node types. Hence, we do not let the perfect be the enemy of the good. Instead of trying to come up with sophisticated solutions to make highly complex networks more or less immune to malware spreading, we fight complexity with simplicity [88]. The next chapter suggests and analyzes a simple scalable technique providing anti-fragility to malware with unknown and changing spreading patterns.