SECURITY AND COMMUNICATION NETWORKS
Security Comm. Networks 2015; **8:**2561–2579
Published online 28 January 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/sec.1186

REVIEW ARTICLE

Mathematical modeling of the propagation of malware: a review

Angel Martín del Rey*

Department of Applied Mathematics, Institute of Fundamental Physics and Mathematics, University of Salamanca, Spain

ABSTRACT

In the present work, we offer a critical review of the mathematical models that have been proposed to date to simulate malware propagation in a network of computers or mobile devices. We analyze the models proposed determining the deficits and possible alternatives for improving them. Copyright © 2015 John Wiley & Sons, Ltd.

KEYWORDS
malware; mathematical modeling; computational simulation; differential equations; cellular automata

***Correspondence**
Angel Martín del Rey, Department of Applied Mathematics, University of Salamanca, Spain.
E-mail: delrey@usal.es

1. INTRODUCTION

Malware is currently one of the main threats to information security. Far from decreasing, his threat (and the effects thereof) will expand considerably in coming years, mainly because of improvements in its techniques and goals (APTs, Crimeware, etc.) and the progressive implantation of the Internet of Things. The struggle against malware is led from different areas, ranging from the awareness of the user to adopt security measures to the development of antimalware software by specialized companies, passing through the setting up of adequate security policies in different agencies and companies, and so on. The great missing link in this scenario is the development of software simulating malware propagation.

This type of application, so widely used in other fields such as the propagation of infectious diseases or of forest fires (see, for example, [1,2]), would be of great use to managers because these applications would allow them to simulate the behavior of the propagation of malicious code in a network, test the effectiveness of countermeasures and, in sum, make suitable decisions regarding the containment of the propagation or at least the minimization of its noxious effects. Simulation software is derived from the computational implementation of a given mathematical model. Thus, the development of these models that try to explain the behavior of the propagation of malicious code is crucial. Very few models have been published in the scientific literature for such purposes; most are based on the paradigm inherited from Mathematical Epidemiology (and more specifically, on the model of Kermack and McKendrick, proposed during the first quarter of the 20th century), such that although they have a sound mathematical basis, some problems emerge when they are applied to real-life situations.

The chief aim of the present work is to offer a critical analysis of the mathematical models proposed, determining the strong and weak points and, from there, determine possible (and alternative) future lines of investigation in the study of the malware propagation. More precisely, the great majority of proposed models are based on differential equations. They are analyzed determining their deficits and the possible alternatives to improve them. Specifically, we will focus our attention on discrete models based on cellular automata to overcome the drawbacks exhibited by differential equation models. Some of these models are described, and the main future lines of research are detailed.

The remainder of this paper is organized as follows. In Section 2, we introduce the concept of malware, indicating the main types, their manner of propagation, and their impact. In Section 3, we address the notion of mathematical modeling, detailing the different classes of model that can be designed. The mathematical models that have been designed to attempt to simulate the propagation of malware are explored and analyzed critically in Section 4, and finally, in Section 5, we present some conclusions and future lines of investigation.

2. MALWARE

Malicious code (or in its English form, malware) is the generic term used to designate any informatics program created deliberately to carry out an unauthorized activity that, in many cases, is harmful to the system in which it has been lodged. The unauthorized activity of malware (payload) may range from a simple erasure of files to the retrieval and later use of private and/or confidential information (web sites visited, contact lists, passwords, account numbers, etc). Sometimes, the activity performed by the malware may provide some benefit to its creator or disseminator [3].

Over the past decade, there has been a huge rise in the number of types of malware created and, accordingly, their effects. This trend remains constant, and according to McAfee data, at the end of June 2014, the total number of specimens was close to 200 million [4]. According to a study reported by PandaLabs [5], the mean number of computers infected by malware is currently 31.88%, the countries with the highest infection rates being China (52.26%), Turkey (43.59%), Peru (42.14%), and Bolivia (41.67%). At the opposite pole, the countries least infected are Sweden (21.03%), Norway (21.14%), and Germany (24.18%). The economic losses caused by malware activity in its different scenarios (government agencies, companies and individuals) are huge and have been estimated at thousands of millions of dollars per year.

The propagation of malware occurs through different vectors: the use of removable devices (CD-ROMS, USB memory sticks, external hard drives, etc), the sending of infected files or links to malicious web sites through e-mail messages, the downloading of infected files from malicious web sites, the sending of infected files through Bluetooth, MMS, SMS, and so on. Attending to the main characteristics regarding the form of propagation, functionality and harm caused by malware, this can be classified in different types [3].

(1) *Computational viruses* (file infectors, boot-sector infectors, system infectors, macro virus, script virus, etc.). This is a form of malware which is capable of copying itself and spreading to other computers by attaching themselves to various programs and executing code when the user launches one of those infected programs.
(2) *Computer worms.* This type of malware spreads over computer networks by exploiting operating systems vulnerabilities. Sometimes, computer worms are classified as a particular type of computer viruses, but there are several characteristics that distinguish worms from computer viruses. The most important difference is that computer worms have the ability to self-replicate and spread independently, whereas viruses depend on the human activity to spread. In this sense, the propagation of computer worms is often based on the sending of mass e-mails with infected attachments to user's contacts.
(3) *Trojans* (downloader, banking, backdoor, etc.). They are a type of malware that disguises itself as a normal file or program to trick users into downloading and installing malware. A trojan provides a malicious party remote access to the computer where it is installed in order to steal data, install more malware, use the computer in botnets, and so on.
(4) *Rootkits.* This is a type of malicious software designed to remotely access or control a computer without being detected by users or security programs. In this sense, rootkit prevention, detection, and removal are difficult because of their stealthy operation.
(5) *Spyware.* It is a type of malware that functions by spying on user activity without his/her knowledge. Spyware spreads by exploiting software vulnerabilities, bundling itself with legitimate software, or in trojans.
(7) Others. Addware, logic bombs, keyloggers, ransomware, hoaxes, RATs, and so on.

However, in light of the constant evolution of malware today, it is often very hard to classify a new threat as being of one type or another because the same agent can show varying characteristics or functionalities. The most widely developed malware today is the type that can be classified as a trojan, and most of the infections produced are due to this type of malware; this is followed at a considerable distance by those classified as viruses or worms (Figure 1).

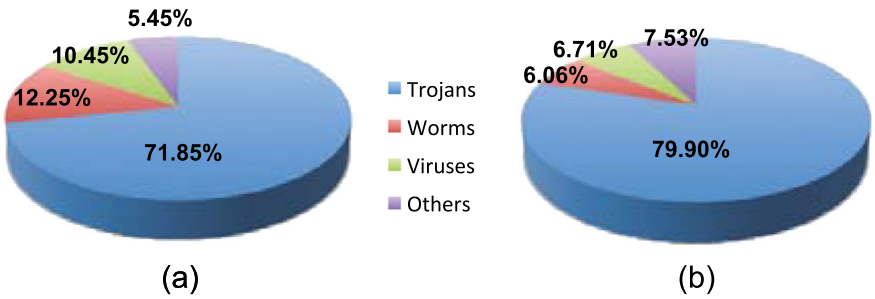

Figure 1. (a) New malware strains in the first quarter of 2014 and (b) malware infections by type in the first quarter of 2014. (Source: *Quarterly Report Pandalabs: January–March 2014*)

3. MATHEMATICAL MODELING

Mathematical modeling is a discipline that aims at transferring problems that arise within a given scientific field or technological field to mathematical languages so that the theoretical and numerical analyses performed on these problems will provide information for better understanding the mechanisms driving the phenomenon in hand. It is a modern research tool that can be considered a complement to theory and experimentation in scientific investigation [6]. In fact, mathematical modeling is the best option when attempting to gain knowledge from experiments that are either very expensive to perform or that would demand an excessive amount of time or would be dangerous. Mathematical modeling can be encapsulated within the so-called mathematical models, which, roughly speaking, are merely representations simplified in mathematical terms of a given idea. Essentially, they are characterized by making assumptions about the variables (the characteristics that change), the parameters (the characteristics that do not change), and the functional relations between the variables and parameters that govern the dynamic of the variables. Thus, mathematical models encompass hypotheses about the systems studied and allow us to compare these hypotheses with the empirical data.

The scheme followed by all mathematical modeling processes is shown in Figure 2. The starting point is marked by the real problem to be solved (usually consisting of the determination of the laws governing the dynamic of a certain phenomenon). Following this, it is necessary to identify and select the factors that will describe the most important aspects of the phenomenon: the principles that drive the evolution of the phenomenon, the physical laws involved, variables, parameters and the relations between them, and so on. Moreover, determining the other characteristics that are not to be taken into account with this simplification process, it is possible to obtain what is called a working model.

It is then necessary to express the working model in mathematical terms. To accomplish this, we must determine the equations whose solutions describe it: we thus obtain the mathematical model, which must then be implemented computationally, giving rise to the computational model. Executing this model allows us to perform simulations from which results and conclusions can be inferred. The interpretation of such results and their comparison with the empirical data obtained from observation of the real phenomenon will allow us to determine the efficiency of the mathematical model developed. If it is found that the predictions match what happens in reality, we can state that the model is suitable, if not, it is necessary to begin the modeling process again to obtain a more refined product. Mathematical models can be classified on the basis of the different characteristics defining them. Thus, among others, the following variants can be found:

(1) Deterministic versus stochastic models. Deterministic models are those whose parameters and variables are not random, that is, they do not follow any probabilistic distribution, unlike the case of stochastic models.
(2) Continuous versus discrete models. Continuous models are those in which the variables may take an infinite and uncountable number of values within a given range; in discrete models, some or all the variables take a finite number of values.
(3) Global versus individual models. When aiming to simulate the behavior of a complex system formed by multiple elements, global models are those that study the dynamic of the system in general,

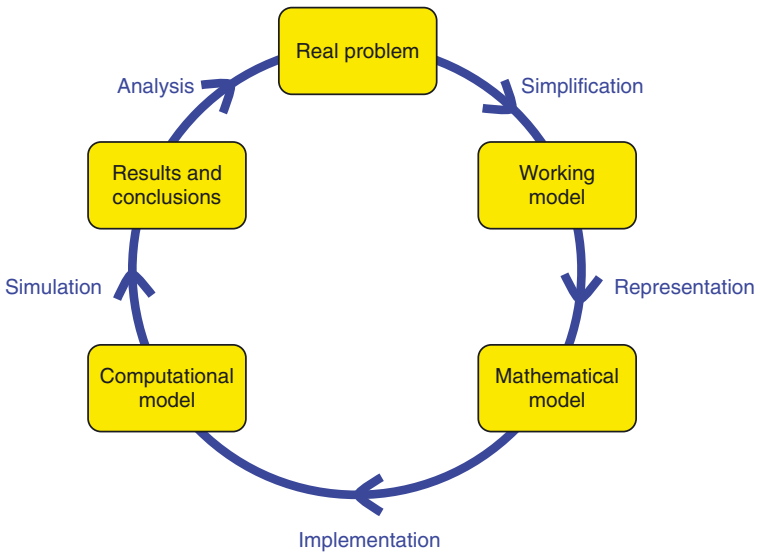

Figure 2. A simple scheme of the mathematical modeling process.

providing the global evolution of the system. In contrast, individual models are those focused on the study of the particular dynamic of the different elements forming the system.

The mathematical paradigm on which the model is based determines the model's classification. The large majority of mathematical models are based on the use of ordinary differential equations or on partial differential equations, which mean that such models are continuous in nature. By contrast, if the mathematical tools used are cellular automata, neuronal networks, finite-state machines, agent-based models, and so on, we would obtain discrete models. Intermediate between these are mixed models (those in which some variables are continuous and others discrete) based on recurrence equations, Boolean delay equations, and so on.

4. MATHEMATICAL MODELS FOR SIMULATING THE MALWARE PROPAGATION

The mathematical models developed to explore the propagation of malware are based on models designed to study the dissemination of infectious diseases. This is due to the similarity between the behavior of biological viruses, fungi, or prions and that of malware (computer viruses, computer worms, etc). Thus, many of the properties of the former are translated and reflected in the latter [7–10]. The analysis of mathematical models made here has a dual aim: on one hand, we perform a quantitative analysis that will allow us to determine the evolution, the impact, and the future perspectives of such models, and on the other hand, we carry out a qualitative analysis, studying the structure of the models from the mathematical tools used, and determining their advantages and disadvantages.

4.1. Quantitative study

The number of mathematical models that have appeared in the scientific literature addressing the study of the propagation of malicious code is very low. On performing a search in three of the main databases: Scopus from Elsevier (which contains nearly 50 million documents), Web of Science from Thomson–Reuters (with about 36 million documents), and IEEE Xplore from IEEE (3.6 million documents) of the chains 'Mathematical model' and 'Malware' (chain 1), 'Mathematical model' and 'Computer virus' (chain 2), and 'Mathematical model' and 'Computer worm' (chain 3), few documents were retrieved, as can be seen in Table I. In particular, in the case of Scopus, the number of references found was 0.015%, 0.045%, and 0.108% of the works indexed in that database addressing malware, viruses, and computational worms, respectively. Moreover, on comparing the total number of works in which mathematical models are proposed and referenced in the Scopus database, very low percentages are obtained: 0.000051%, 0.00017%, and 0.00012%, respectively. Accordingly, it appears that in this case, there is no correlation between the importance of the problem and the number of works published.

The first mathematical studies aimed at predicting the behavior of an epidemic of malware were carried out in the time of Kephart and White [11–13] at the beginning of the 1990s. Since then, and up to 2000, the number of papers appearing in the scientific literature has been fairly scarce, never surpassing 2–3 publications/year (Figure 3). It was then, at the start of the current century, that the number of proposals published began to grow, reaching a peak in 2005–2006. Since then, a decline has occurred in the number of publications such that we are now appearing to regressing to the pre-2000 situation.

As may be seen from Figure 3, from the birth of the term virus to the end of the decade of the 1980s [7], investigators have been working on mathematical models, although it is true that during the early years, these did not have much academic impact (measured according to the number of publications). Only during the first decade of the 21st century was there a noteworthy increase in the number of models proposed. This could have been due to the large number of specimens detected during those years and the media impact they generated. As from 2000, there were massive infections mainly because of worms such as *I Love You*, *Mydoom*, *Netsky*, or *Sasser*. Accordingly, public awareness increased and this had its own repercussions on the scientific community.

Although in the second decade of the 21st century malicious code continues to increase, there has been a reduction in the number of models proposed. This may be because that such models presuppose that propagation occurs in a network of computers, from fixed entities (and readily

Table I. Number of papers published and indexed (July 2014) in Scopus, Web of Science and IEEE databases.

Text chain	Scopus	WoK	IEEE
'Mathematical model' and 'Malware'	55	5	218
'Malware'	3642	1946	6358
'Mathematical model' and 'Computer virus'	185	7	137
'Computer virus'	4078	1011	2193
'Mathematical model' and 'Computer worm'	129	5	80
'Computer worm'	1198	72	1046

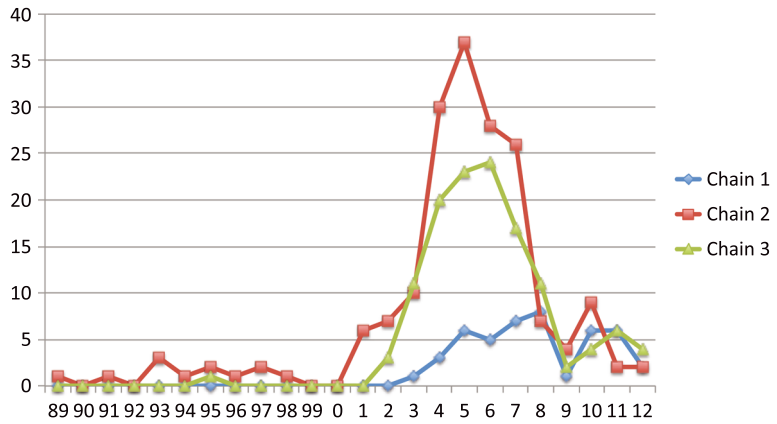

Figure 3. Evolution of the number of papers indexed in Scopus.

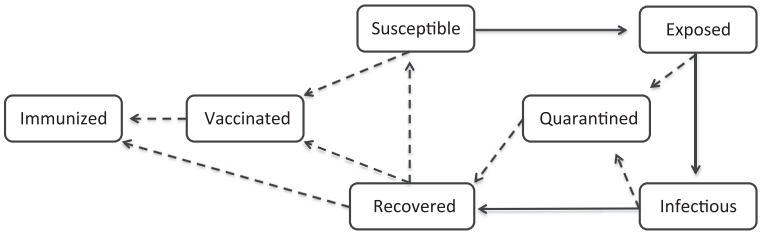

Figure 4. Illustrative diagram of the evolution of the different types of behavior.

modeled), but the social reality has taken a different turn: in these years, the paradigm has changed and we now see increasing use of other types of device that are able to connect users with the Internet (the natural habitat of malware): tablets, smartphones, and so on. In addition, these new ways of connecting favors the (inexorable) development of the Internet of Things. Accordingly, we are creating a scenario in which nearly all objects are susceptible to being connected to the Internet and we can access these at any time and from any place. In this situation, the mathematical models already created are obsolete by virtue of their own nature because, as we shall see in the succeeding text, they are unable to efficiently capture the new paradigm of mobility and universal connection currently being implanted. Considering this new scenario, in the last few years, some works dealing with the design of mathematical models to simulate mobile malware spreading have appeared. They are usually based on alternative mathematical tools of discrete nature.

4.2. A critical analysis of mathematical models

4.2.1. General characteristics.

As mentioned earlier, mathematical models that simulate the propagation of malware are based on those developed to study the behavior of the propagation of infectious diseases. Thus, they have inherited the basic characteristics defining these models, that is, the class into which the population is divided, the nature of the model, and the mechanisms that govern the dynamic of the infection.

The epidemiological models are compartmental, that is, the population (through which the infectious disease is propagated) is divided into different types of behavior bearing in mind the characteristics of the disease: susceptible, exposed (with or without symptoms), infectious, recovered, quarantined, vaccinated, isolated, and so on. The denominations translated to the case of the study of malware propagation correspond thus to the following: susceptible (computers that have not been infected by malware), exposed (computers that have been infected by malware but have not yet been activated), infectious (computers in which the malicious code is in active mode, thus having the capacity to perform its duty and propagate to other computers of the network), quarantined (computers detected as having been infected and are withdrawn from the network to eliminate the virus), recovered (computers in which the malware has been eliminated), and so on.

Thus, we find susceptible-infectious-susceptible (SIS) models, susceptible-infectious-recovered (SIR) models, susceptible-exposed-infectious-recovered (SEIR) models, susceptible-exposed-infectious-quarantined-recovered (SEIQR) models, SEIRS-V models (the vaccinated class is included) or variants SIRS, SEIRS, SEIQRS, SEIQV (susceptible-exposed-infectious-quarantined-vaccinated), and so on. Figure 4 shows the diagram of the most usual types of behavior and dynamic among them.

Accordingly, in models exploring malware, we can find the same compartment. Thus, proposals have been made for SIS (see, for example, [11,14–18]), SIR (see, for example, [19–24]), SEIR (see, for example, [25,26]), SEIRS [27–29], SEIQS [30], SEIQRS [31], SEIRS-V [32], SEIS-V [33], SEIQV [34], SIRS [35–37], SEI [38,39], SIC [40], SIRQ [41], and so on. It may be seen that there is no single compartmental model that has been addressed in a greater number of works but that there is a certain homogeneity in the compartmental models proposed.

These models can also be classified attending to the nature and the mathematical tools on which they are based. In this sense, we find deterministic models (see, for example, [20,22,28,32,42–45]) and stochastic models (see, for example, [12,14,21,46–49]). Deterministic models are usually based on differential equations or difference equations. On the other hand, stochastic models can be classified into two basic types: those based on Markov chains (discrete time Markov chain models or continuous time Markov chain models), where the state variable is discrete, and those based on stochastic differential equations, where both time and state variables are continuous. The deterministic models provide good results when the population is very large: greater than 10^5–10^6 [41], whereas stochastic models are more appropriate when attempting to simulate the propagation of malware in small computer networks, usually between 10^2 and 10^5–10^6.

Most of the models proposed (be they deterministic or stochastic) can be classified as global models because they study the dynamic of the population overall without taking into account the local interactions between individuals beyond what is reflected in the parameters. In contrast, there are very few individual-based models. As far as we are aware, only four have been proposed and all of them are based on cellular automata [30,50–52].

Continuous models predominate because they are based on differential equations [53]; discrete models are usually based on cellular automata or agent-based models (see, for example, [54]). The aim of the large majority of the models proposed is to study the dynamic of the different compartments into which the population is divided, that is, knowledge of the number of susceptible, exposed, infectious, and so on, computers present at each moment of time, and the evolution of their numbers with time. However, there are some models in which the specific aim differs from this. Thus, in [44], a study is reported of the propagation of one or several specimens of malware among the different software components of a single informatics system; in [55], the author proposes a model to study the effect caused by malware on the response of an informatics system challenged by such attacks. In [56], the authors show how resources and costs influence the epidemic dynamics of computer viruses in scale-free networks: it is possible to control its spread if the resources are restricted and the costs significantly increase.

The great majority of models are devoted to the study and simulation of computer worms (see, for example, [41,45,48,49]) and computer virus spreading (see, for example, [30,38,39,42]), although some proposals have been appeared for unspecified malicious software [25,31,36,54,55]. Furthermore, there is a minority of models devoted to other types of malware [24]. In the case of computer worms, the great majority include the exposed compartment (see, for example, [26–29,32–34]), whereas in the case of computer viruses, the majority are SIS or SIR models [14–16,19,20,22,35,40]. Notice that this is a curious fact because, as is mentioned in Section 2, the spreading of computer viruses is possible once the user runs the program where it is attached. As a consequence, there must be latent period (along which the computer belongs to exposed).

As mentioned in Section 3, all mathematical models are characterized by three elements: the variables studied, the parameters used, and the functional relationships governing the dynamic (considering the variables and the parameters). In the case of the simulation of the propagation of malware (that is, study of the evolution of the different compartments with time), the variables employed are the number of devices (usually computers) found in some of the types considered: susceptible computers, infectious computers, exposed computers, and so on.

The usual parameters used in modeling are as follows: the rate of infection or transmission rate (which depends on the contact between individuals and the probability that a contact leads to transmission), the rate of recovery (due to the effect of antivirus applications), the number of computers removed from the network, the probabilities of passing from one compartment to another, the probability of the acquisition of immunity (transient or indefinite), the latency period, the period of immunity, and so on. The use of one or the other depends on the model implemented and the type of malware considered.

The evolution of the different compartments is governed by the functional relationships that take into account the parameters introduced in the model. These relationships can be orchestrated with different mathematical tools, some of which have been mentioned earlier. Thus, those most used are ordinary differential equations. Also used are difference equations, Markov chains and, to a much lower extent, cellular automata.

4.2.2. Models based on differential equations.

As stated earlier, most of the mathematical models designed to study the propagation of malware are based on the use of differential equations. This is mainly because ordinary differential equations and partial differential equations are important pillars in Mathematical Modeling. Because the models studied here were inspired in models proposed for Mathematical Epidemiology, their structure is very similar and in some cases identical. Undoubtedly, the pillars on which models based on differential equations rest are the models of Kermack and McKendrick [57], Hethcote [58] and that of Diekmann and Heesterbeek [59].

The use of differential equations allows one to perform a detailed mathematical analysis of the model in question. The behavior of these models mainly depends on a

DOI: 10.1002/sec

threshold parameter termed the basic reproductive number, R_0, which will determine the stability of the disease-free equilibrium and of the endemic equilibrium [60]. The basic reproductive number is defined as the number of secondary infections caused by a single infected individual in a completely susceptible population. Thus, if $R_0 \leq 1$, the infection will be reduced (the number of infected individuals decreases until they disappear), a stable state of infection-free equilibrium being reached. If, by contrast, we see that $R_0 > 1$, then the infection will propagate (the number of infected individuals will increase), a state of stable endemic equilibrium being attained.

The appearance of the model proposed by Kermack and McKendrick in 1927 is the most important milestone in Mathematical Epidemiology. This is an SIR model in which the size of the population remains constant. Susceptible individuals are those that may contract the disease, infectious individuals are those who have acquired the disease and are able to transmit it, and finally, recovered individuals are those who were infected but, for different reasons, are currently non-infectious (they have been isolated, they have recovered, acquiring immunity, or they have died, etc). Accordingly, three time-dependent variables are considered: the number of susceptible individual $S(t)$, the number of infectious individuals $I(t)$, and the number of recovered individuals, $R(t)$. Because it is assumed that the population remains constant, we have that $S(t) + I(t) + R(t) = N$ for any moment of time t, where N represents the total number of individuals forming the population. In addition, two parameters are taken into account, namely, the transmission rate, a, (the product between the contacts among individuals and the probability of transmission per contact) and the recovery rate, b, (rate at which an infected individual recovers per unit of time). The dynamic of the model is governed by the following set of ordinary differential equations:

$$\begin{cases} S'(t) = -aI(t)S(t) \\ I'(t) = aI(t)S(t) - bI(t) \\ R'(t) = bI(t) \end{cases} \quad (1)$$

The notation $S'(t)$, $I'(t)$, and $R'(t)$ indicates the respective derivatives of $S(t)$, $I(t)$, and $R(t)$ with respect to time t, that is, the variation over time of these magnitudes. The first equation of the system, $S'(t) = -aS(t)I(t)$, establishes that the variation in the number of susceptible individuals with time depends on the number contacts between susceptible individuals and infectious individuals and on the transmission rate (mass action law). The second equation of the system, $I'(t) = aNS(t)I(t) - bI(t)$, indicates that the variation in the number of infectious individuals is the difference between those newly infectious and infectious individuals who have recovered. Finally, the third equation, $R'(t) = bI(t)$, shows that the increase in recovered individuals is proportional to the number of infectious individuals where the recovery rate itself being the proportionality constant.

From the mathematical study of the model, we obtain that $R_0 = a/b$. Consequently, if $a \leq b$ ($R_0 \leq 1$), then the system reaches a state of equilibrium without infection, whereas if $a > b$, then there is an initial increase in the number of infected individuals. An example of the evolution of the different compartments can be seen in Figure 5 in different cases: in part (a), we consider the parameters $a = 0.25, b = 0.1$ such that $R_0 = 2.5 > 1$ and the epidemic will occur; in part (b), the values of these parameters are $a = 0.25, b = 0.25$ and hence, $R_0 = 1$ and the number of infected individuals will not increase (disappearing with time).

In the equations stated in 1, the term $aI(t)$ is called force of infection and it is the main characteristic of density-dependent transmission models. In the opposite, the frequency-dependent transmissions models suppose that the force infection is given by $a\frac{I(t)}{N(t)}$.

As mentioned previously, the Kermack and McKendrick model is the fundament on which all epidemiological models based on differential equations are constructed; this can also be applied to models of malware propagation. Among these, we could cite the following as paradigmatic examples

(1) *The model of Mishra and Saini* [25]. This is a SEIRS model in which the following assumptions are made:

(i) Any computer in the network is susceptible.

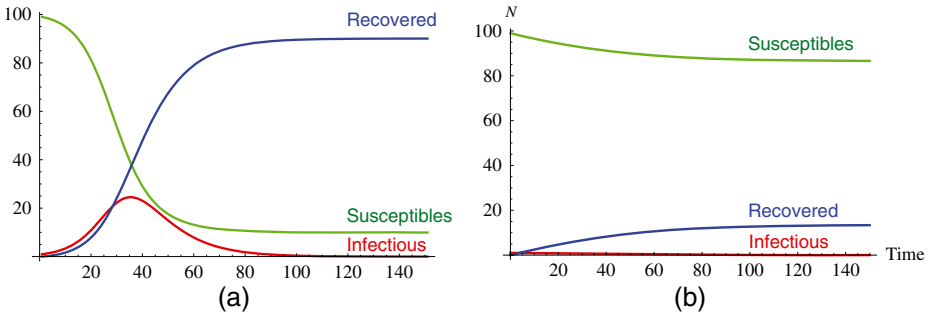

Figure 5. Dynamic of the SIR model due to Kermack and McKendrick with $N = 100$ individuals and $S(0) = 99, I(0) = 1$. (a) $a = 0.25, b = 0.1$ and (b) $a = b = 0.25$

(ii) The rate of removal from the network not caused by the effect of malware is constant and equal for any of the elements of that computer network.
(iii) The rate of removal from the network caused by the malware is constant.
(iv) The latency period, ω, and the immunity period, τ, are constant.
(v) The times of undergoing states of exposure, infection, and quarantine follow an exponential distribution.
(vi) When a computer is no longer infectious, it recovers, acquiring transient immunity with a probability p and is removed from the network owing to the effects of malicious code with probability $1-p$.

The system of differential equations governing this is as follows:

$$\begin{cases} S'(t) = bN(t) - \frac{\gamma}{N(t)}S(t)I(t) - \mu S(t) \\ \qquad +\alpha I(t-\tau)e^{-\mu t} \\ E'(t) = \frac{\gamma}{N(t)}S(t)I(t) \\ \qquad -\frac{\gamma}{N(t-\tau)}S(t-\tau)I(t-\tau)e^{-\mu\omega} \\ \qquad -\mu E(t) \\ I'(t) = \frac{\gamma}{N(t-\tau)}S(t-\tau)I(t-\tau)e^{-\mu\omega} \\ \qquad -(\mu+\epsilon+\alpha)I(t) \\ R'(t) = p\alpha I(t) - \alpha I(t-\tau)e^{-\mu\tau} \\ \qquad -\mu R(t) \end{cases} \tag{2}$$

such that $E(t)$ is the number of exposed computers at time t, b is the rate of appearance of new computers in the network, μ is the rate of removal from the network owing to causes not due to the malware, ϵ is the rate of removal from the network owing to the action of the malware, α is the rate of recovery, and γ is the transmission rate, that is, the product of the mean number of contacts of a computer (per unit time) and the probability that the malware will be transmitted during that time. Note that this is a frequency-dependent transmission model. With these premises, it may be seen that the basic reproductive number of this model is as follows:

$$R_0 = \frac{\gamma e^{-b\omega}}{b+\alpha+\epsilon} \tag{3}$$

This model is an extension of the SIR epidemic models of Diekmann and Heesterbeek to SEIR type with constant exposed and latent periods. Temporary immunity is considered and it is obtained that the longer the exposed period the system has, the less the chances are that it will be endemic.

(2) *The model of Mishra and Jha* [31]. This is an SEIQRS model that can be defined with the following system of differential equations:

$$\begin{cases} S'(t) = A - aS(t)I(t) - dS(t) + \eta R(t) \\ E'(t) = aS(t)I(t) - (d+\mu)E(t) \\ I'(t) = \mu E(t) - (d+\alpha+\gamma+\delta)I(t) \\ Q'(t) = \delta I(t) - (d+\alpha+\epsilon)Q(t) \\ R'(t) = \gamma I(t) + \epsilon Q(t) - (d+\eta)R(t) \end{cases} \tag{4}$$

where $Q(t)$ is the number of quarantined computers at time t, A is the number of new computers added to the network at each moment in time, d is the rate of removal of computers from the network owing to causes not due to the malware, μ is the rate of passage from the exposed state to the infected state, δ is the rate of passage from the infectious state to quarantine, α is the rate of removal of a computer from the network owing to the action of the malware (as long as the computer is in the infectious state or is in quarantine), γ is the rate of recovery due to the action of the antivirus software, ϵ is the rate of passage from the infectious or quarantined state to the recovered state, and finally, η is the rate of the loss of immunity. Figure 6 shows the flow chart among the different compartments of this model. In this model, we see that the basic reproductive number is given by the expression

$$R_0 = \frac{aA}{d(\mu+\alpha+\delta+\gamma+d)} \tag{5}$$

In this model, the number of contacts is influenced by the size of the quarantined class, and consequently, both the effective infectious period and the basic reproductive number decrease as the quarantine rate increases.

(3) *The model of Mishra and Pandey* [27]. This is an SEIRS model for studying the propagation of computer worms, defined by the following set of differential equations:

$$\begin{cases} S'(t) = b - \lambda S(t)I(t) - pbE(t) - qbI(t) \\ \qquad -dS(t) + \zeta R(t) \\ E'(t) = \lambda S(t)I(t) + pbE(t) + qbI(t) \\ \qquad -\epsilon E(t) - dE(t) \\ I'(t) = \epsilon E(t) - \gamma I(t) - dI(t) - \eta I(t) \\ R'(t) = \gamma I(t) - \zeta R(t) - dR(t) \end{cases} \tag{6}$$

where b is the recruitment rate of susceptible nodes, d is the per capita natural mortality rate, ϵ

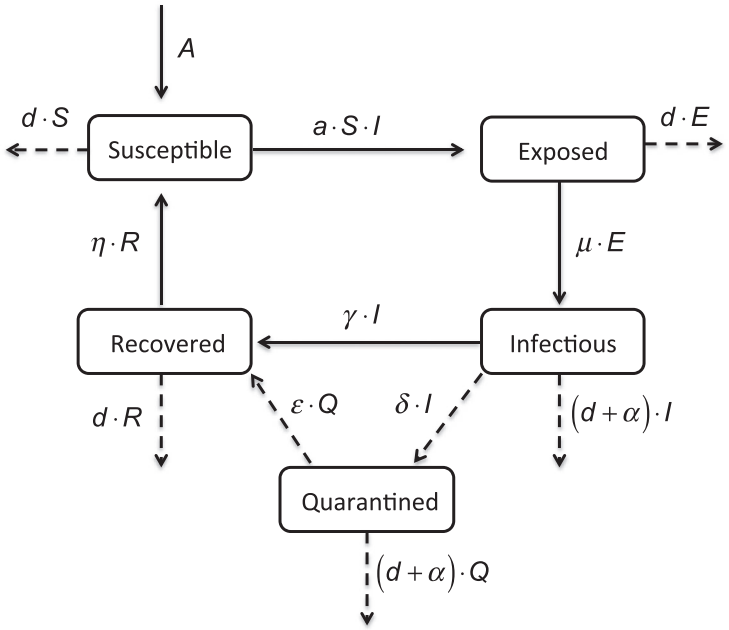

Figure 6. Schematic flow chart representing the dynamic between the compartments of the model of Mishra and Jha.

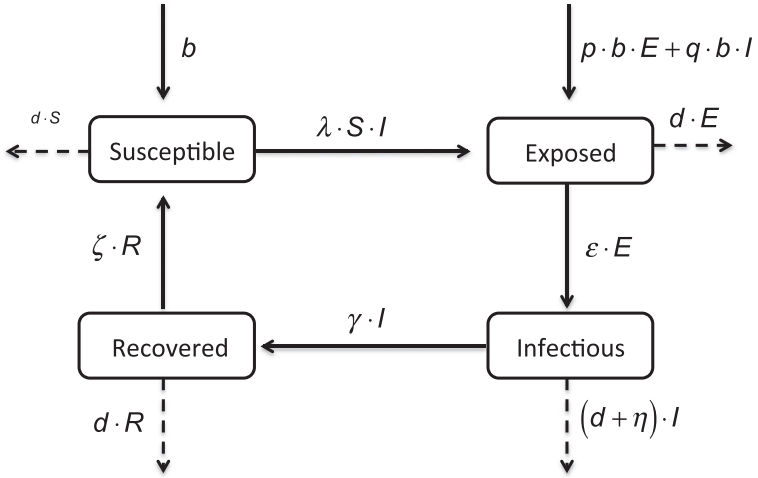

Figure 7. Schematic flow chart representing the dynamic between the compartments of the model due to Mishra and Pandey.

is the rate for nodes leaving the exposed class for infectious class, γ is the rate for nodes leaving the infectious class for recovered class, η is the disease related death rate in the infectious class, and ζ is the rate for nodes becoming susceptible again after recovery.

Moreover, the flow chart between the different compartments is shown in Figure 7. In addition, the basic reproductive number is given by the expression

$$R_0 = \frac{\lambda\epsilon}{(\epsilon + d)(\gamma + d + \eta) - pb(\gamma + d + \eta) - qb\epsilon} \quad (7)$$

The main characteristic of this model is the use of vertical transmission in which there is a constant period of temporary immunity of fixed length following by a temporary recovery period instead of an exponentially distributed period of temporary immunity.

(4) *The model of Zhu, Yang, and Ren* [22]. This is an SIR-type compartmental model in which susceptible individuals and infectious individuals removed from the network are introduced. The set of differential equations governing the model is as follows:

$$\begin{cases} S'(t) &= \lambda_1 - \beta_1 S(t) I(t) - \beta_2 S(t) \frac{R_I(t)}{R_N(t)} \\ & \quad -\mu_1 S(t) \\ I'(t) &= \beta_1 S(t) I(t) + \beta_2 S(t) \frac{R_I(t)}{R_N(t)} \\ & \quad -(\mu_1 + \sigma_1) I(t) \\ R'(t) &= \sigma_1 I(t) - \mu_1 R(t) \\ R'_S(t) &= \lambda_2 - \beta_2 \frac{R_S(t)}{N} + \sigma_2 \frac{R_I(t) R(t)}{N} \\ & \quad -\mu_2 R_S(t) \\ R'_I(t) &= \beta_2 \frac{R_S(t)}{N} - \sigma_2 \frac{R_I(t) R(t)}{N} \\ & \quad -\mu_2 R_I(t) \end{cases} \tag{8}$$

where $R_S(t)$ (*resp.* $R_I(t)$) is the number of susceptible (*resp.* infectious) removable devices at time t, $R_N(t)$ is the total number of removable devices at time t, λ_1 is the recruitment of computers, λ_2 is the recruitment of removable devices, β_1 is the contact infective force between susceptible and infectious computers, β_2 is the contact infective force between computers and removable devices, σ_1 (*resp.* σ_2) is the recovery rate of infectious (*resp.* removable) devices due to the effect of antivirus, μ_1 is the rate at which networked computers are disconnected from network, and μ_2 is the rate at which removable devices break down.

The associated basic reproductive number is as follows:

$$R_0 = \frac{\beta_2^2 + \frac{\mu_2 \beta_1 \lambda_1}{\mu_1}}{\mu_2 (\mu_1 + \sigma_1)} \tag{9}$$

Figure 8 shows the dynamic between the different compartments forming part of the model.

This work provides an excellent start point for understanding the propagation of computer virus through the interactions between computers and external removable devices.

(5) *The model of Toutonji, Yoo, and Park* [28]. This is an SEIRS model whose set of differential equations is as follows:

$$\begin{cases} S'(t) &= -\frac{\beta \alpha}{N} E(t) S(t) - \psi_1 S(t) + \phi R(t) \\ E'(t) &= \frac{\beta \alpha}{N} E(t) S(t) - (\alpha + \psi_2) E(t) \\ I'(t) &= \alpha E(t) - (\gamma + \theta) I(t) \\ R'(t) &= \mu N + \psi_1 S(t) + \psi_2 E(t) + \gamma I(t) \\ & \quad -\phi R(t) \end{cases} \tag{10}$$

The parameters involved in this model are the contact rate β, the state transition rate from exposed to infectious α, the state transition rate from susceptible to recovered ψ_1, the state transition rate from exposed to recovered ψ_2, the state transition rate from infectious to recovered γ, the state transition rate from recovered to susceptible ϕ, the dysfunctional rate θ, and the replacement rate μ. Moreover, the basic reproductive number obtained is the following:

$$R_0 = \frac{\alpha \beta}{(\psi_1 + \phi)(\alpha + \psi_2)} \tag{11}$$

This model takes into consideration accurate positions for dysfunctional hosts and their replacements in state transition. The simulations results showed that all computer worms were able to pervade if $R_0 > 1$, and the malware epidemic died out in the case $R_0 < 1$. Moreover, these results also show the positive impact of increasing security countermeasures in the susceptible state on worm-exposed and infectious propagation waves.

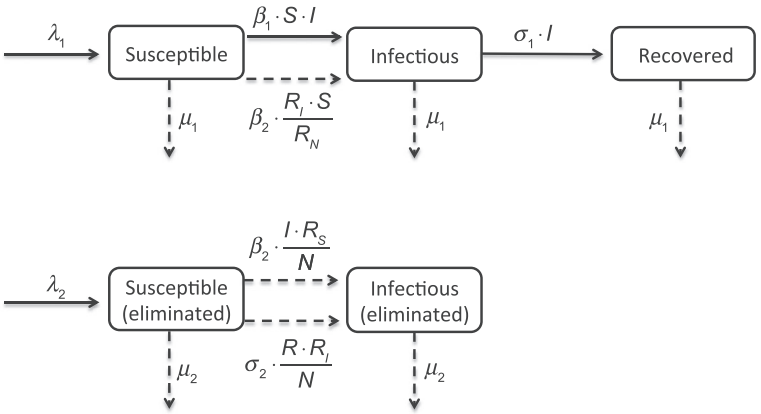

Figure 8. Schematic diagram representing the flow of the compartments of the model of Zhu *et al.*

4.2.3. Discrete models based on cellular automata.

Apart from the models based on differential equations, another series of models based on very different paradigms has been proposed. In the following section, we discuss an individual discrete stochastic model based on the use of cellular automata.

Cellular automata are simple computational models (a particular type of finite-state machines that are able to simulate complex systems efficiently and efficaciously). They comprise a finite number of units called cells that are interconnected via a certain topology, such that at each moment of time, each cell is in one state from among a finite number of possible states. This state changes with the discrete passage of time according to a local transition rule whose variables are states in the previous instant of the cell itself and of its neighbors [61]. There is a wide range of applications of cellular automata in several fields such as Computer science, Biology, Bioinformatics, Cryptography, and Engineering (see, for example, [62,63]). Of special interest are the mathematical models based on cellular automata devoted to the study and simulation of seismicity (see [64] and references therein), the spreading of infectious diseases [65,66], and forest fire propagation [67,68].

The model of Martín and Rodríguez [30] is an SEIQS model in which a study is made of the propagation of a computer virus in a computer network by means of the use of a stochastic cellular automaton. In it, each cell of the automaton is one of the computers of the network and this latter is modeled via a graph (in such a way that two cells/computers are neighbors if there is a connection between them). The cells are found in one of the following states: susceptible, exposed, infected, and in quarantine. The transition functions governing the dynamic of the system are as follows:

(1) Passage from susceptible to exposed. A computer passes from being susceptible to infected when $F = 1$, where F is the following Boolean function:

$$F = X_i^t \vee \left[c_i^t \wedge \left(Y_i^t \vee \bigvee_{v_j \in N_{v_i}} \left(I_{ij}^{t-1} \wedge Z_{ij}^t \right) \right) \right] \quad (12)$$

where $X_i^t = 1$ with probability α_i^t; $X_i^t = 0$ with probability $1 - \alpha_i^t$; $Y_i^t = 1$ with probability $\beta_i^t \cdot \left(1 - pF_i^t\right)$; $Y_i^t = 0$ with probability $1 - \beta_i^t \cdot \left(1 - pF_i^t\right)$; $Z_{ij}^t = 1$ with probability $pV_{ij}^t \cdot \gamma_i^t \cdot \delta_i^t \left(1 - pF_i^t\right)$; $Z_{ij}^t = 0$ with probability $1 - pV_{ij}^t \cdot \gamma_i^t \cdot \delta_i^t \cdot \left(1 - pF_i^t\right)$; and finally, $c_i^t = 1$ if the computer v_i does not have access to Internet at time t and $c_i^t = 0$ otherwise.

Moreover, $\alpha_i(t)$ is the probability that at time t, an external device infected by malicious code will be used; $\beta_i(t)$ is the probability of visiting an infected web site at time t; $pF_i(t)$ is the probability that the anti-virus software will detect and eliminate such malware; $pV_{ij}(t)$ is the probability of receiving an e-mail form the j-th neighbor computer, and $\gamma_i(t)\delta_i(t)$ is the probability of opening an infected file accompanying an e-mail.

(2) The passage from exposed to infected. A computer passes from the exposed to the infected state once the latency time and sleeping time have finished.
(3) The passage from exposed to in quarantine. An exposed computer passes to the quarantine stage with a certain probability that depends on the phase in which it finds itself (sleeping time /latency time).
(4) The passage from infected to quarantine. An infected computer passes from being infected to the quarantine stage with a certain probability in which both the user's behavior and that of the anti-virus software are taken into account.
(5) Passage from quarantine to susceptible. A computer passes from being in quarantine to susceptible once the quarantine period has finished. In this model, the different parameters are addressed individually, that is, the different probabilities of infection (the use of an infected external device, the downloading of infected files from the network, the opening of infected e-mails) are not equal for all the elements of the network but may be different and individual. The same is the case of the action of anti-virus software and with the user's behavior when challenged by an e-mail or extraneous behavior by the computer.

Figure 9 shows the global evolution of the different compartments, and in Figure 10, we see the individual evolution of computers connected to a small local network.

Mathematical study of this model indicates that the stability of the equilibrium without infection is reached when $c_i(t)\beta_i(t)\left(1 - pF_i(t)\right) = 0$.

4.3. The case of mobile malware

Smartphones and other mobile devices play a very important role in our lives. There is an unstoppable rise of the number of such devices and, taking into account some studies, they will grow up to two billion within the next 3 years. The majority of their applications require an Internet access, and consequently, they are exposed to the effects of malware.

Consequently, as in the case of computer networks, the prediction of the behavior of the spreading of mobile malware is very important. Unfortunately, not many mathematical models dealing with this issue have been appeared. In a similar way to which it is done with computer networks, the great majority of these works are based on continuous mathematical tools such as systems of ordinary differential equations [69–75] or recurrence relations [76]. Moreover, also, stochastic models have been proposed [77].

These are well-founded and coherent models from the mathematical point of view, offering a detailed study of the main characteristics of their dynamic: stability,

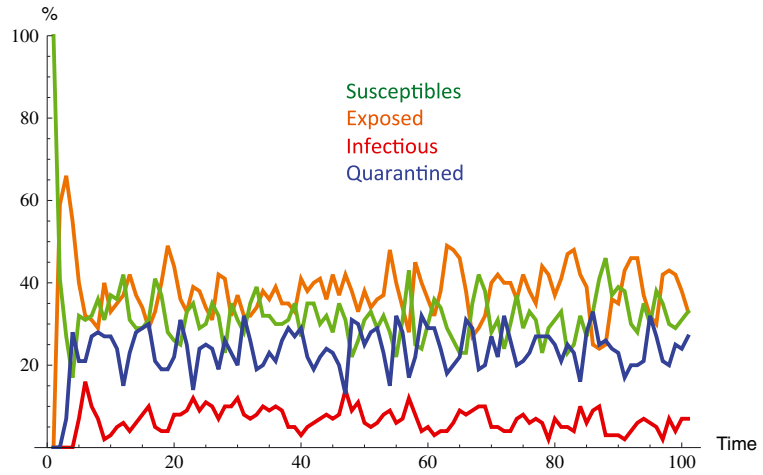

Figure 9. Global evolution of the different classes of computers according to the model of Martín and Rodríguez.

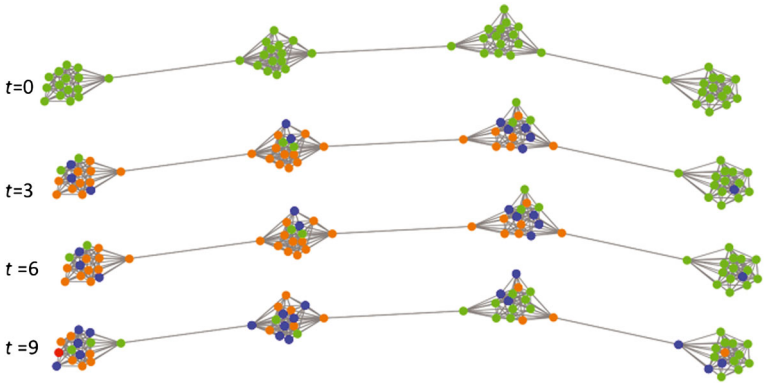

Figure 10. Individual evolution of the computers forming part of a local network.

equilibrium, and so on. Nevertheless, they exhibit the same drawbacks than in the previous case, and these deficiencies could be rectified simply if we use discrete models or individual-based models (as is shown in Section 4.4).

As far as we know, there are only two discrete mathematical models published for mobile malware based on cellular automata. In what follows, both models will be described.

(1) *The model of Peng, Wang, and Yu* [78]. It is a compartmental model where the population is divided into susceptible, exposed, infectious, diagnosed, and recovered smartphones. Every cell stands for a regular geographical area belonging to a square 2D grid which is occupied by at most one smartphone, and these smartphones are fixed in the corresponding cell (that is, movements between cells are not permitted). Moreover, the local transition function mainly depends on two parameters: the infected factor $0 \leq IF_{vu} \leq 1$ and the resisted factor $0 < RF \leq 1$. F_{vu} denotes the infection degree from node v to node u (if $IF_{vu} = 0$, the node has no infection to other nodes, whereas if $IF_{vu} = 1$, the node has stronger infection with other nodes). The resisted factor RF stands for the resistance degree of a node on infection from other nodes (if $RF = 1$, the node has a strong ability to resist infection). Consequently, taking into account these two parameters, the interaction coefficient between a cell C_{ij} and its neighbor C_{kl}, $\Phi_{C_{ij}C_{lk}}$, and the infection index δ can be defined as follows:

$$\Phi_{C_{ij},C_{kl}} = \sum_{m=1}^{N_u} \frac{IF_{vu}}{\sqrt{(i-k)^2 + (j-l)^2}} \tag{13}$$

$$\delta = \frac{\Phi_{C_{ij}C_{lk}}}{RF} \tag{14}$$

where N_u stands for the number of each node's neighbor nodes. Note that $\Phi_{C_{ij}C_{lk}}$ is the strength or likelihood of an infection from C_{lk} to C_{ij}, whereas δ is the ratio of the interaction coefficient between C_{ij} and its neighbors to its resisted factor.

(2) *The model of Martín and Rodríguez* [79]. In [79], a mathematical model based on cellular automata to simulate mobile malware spreading using bluetooth connections is introduced. It could be considered as an improvement of the last mentioned model because (i) more than one smartphone could be placed in a particular cell; (ii) different operative systems are considered; (iii) smartphone movements are permitted; and (iv) more realistic local transition functions are defined. Specifically, in this model, the population is divided into four classes: susceptible, carrier, exposed (or latent), and infectious smartphones. It is based on a paradigm which is far from the paradigm used in [78]. Specifically, two 2D-dimensional cellular automata are used: one of them, $\mathcal{A}_G$, rules the global dynamic of the model, whereas the other one, $\mathcal{A}_L$, governs the local dynamic. The geographical area, where smartphones are placed, is tessellated into several square tiles that stand for the cells of the global cellular automata, such that there can be more than one smartphone in each cell. Moreover, the smartphones placed into a (global) cell stand for the cells of the local cellular automata whose transition rule update synchronously the states of the smartphones. Furthermore, the smartphones can move from one global cell to another at every step of time.

The global cellular automaton, $\mathcal{A}_G$, simulates the global behavior of the system giving at every step of time the number of smartphones which are susceptible, carrier, exposed, and infectious in a certain geographical area. It follows the traditional paradigm for CA, consequently, the whole area where the smartphones 'live' is tessellated into $r \times c$ constant-size square tiles (square grid), and every cell of $\mathcal{A}_G$ stands for one of these square portions of the area. If S_{ij}^t is the number of susceptible smartphones, C_{ij}^t is the number of carrier smartphones, E_{ij}^t is the number of exposed smartphones, and I_{ij}^t is the number of infectious mobile devices in the (i,j)-cell at time t, then the local transition functions are defined as follows:

$$S_{ij}^t = S_{ij}^{t-1} - OS_{ij}^{t-1} + IS_{ij}^{t-1} \tag{15}$$

$$C_{ij}^t = C_{ij}^{t-1} - OC_{ij}^{t-1} + IC_{ij}^{t-1} \tag{16}$$

$$E_{ij}^t = E_{ij}^{t-1} - OE_{ij}^{t-1} + IE_{ij}^{t-1} \tag{17}$$

$$I_{ij}^t = I_{ij}^{t-1} - OI_{ij}^{t-1} + II_{ij}^{t-1} \tag{18}$$

where OS_{ij}^{t-1} stands for the number susceptible smartphones that moved from (i,j) to a neighbor cell at time $t-1$, and IS_{ij}^{t-1} represents the number of susceptible smartphones that arrived at the cell (i,j) at time $t-1$ coming from a neighbor cell and so on.

The local cellular automaton, $\mathcal{A}_L$, simulates the individual behavior of every smartphone of the system, that is, $\mathcal{A}_L$ governs the evolution of the states of each smartphone (susceptible S, carrier C, exposed E, or infectious I). It is a cellular automaton on a graph G, which defines the topology of the cellular space and the neighborhoods. The local transition rules are as follows.

(1) *Transition from susceptible to exposed and carrier*. As is mentioned in the last section, a susceptible smartphone v becomes exposed or carrier when the mobile malware reaches it and this occurs when the user accepts a bluetooth connection from a malicious device. The Boolean function that models the transition from susceptible state to exposed or carrier state is the following:

$$f_L(u) = B_v \cdot \bigvee_{\substack{u \in N(v) \\ s_u^{t-1} = I}} A_{uv} \cdot \alpha_{vu} \tag{19}$$

with

$$B_v = \begin{cases} 1, & \text{with probability } b_v \\ 0, & \text{with probability } 1 - b_v \end{cases} \tag{20}$$

$$A_{uv} = \begin{cases} 1, & \text{with probability } 1 - a_{uv} \\ 0, & \text{with probability } a_{uv} \end{cases} \tag{21}$$

where b_v is the probability to have the bluetooth enabled, and a_{uv} is the probability to accept the bluetooth connection from the smartphone u. Moreover, $\alpha_{vu} = 1$ (*resp.* $\alpha_{vu} = 0$) if the smartphones u and v have the same OS (*resp.* have not the same OS). As a consequence, the state of the node/smartphone v at time t, s_v^t, is

$$s_v^t = \begin{cases} E, & \text{if } s_v^{t-1} = S, f_L(u) = 1 \text{ and } \beta = 1 \\ C, & \text{if } s_v^{t-1} = S, f_L(u) = 1 \text{ and } \beta = 0 \\ S, & \text{if } s_v^{t-1} = S \text{ and } f_L(u) = 0 \end{cases} \tag{22}$$

where the parameter β denotes if the infection comes from an infectious smartphone with the same operative system ($\beta = 1$) or with a different operative system ($\beta = 0$).

(2) *Transition from exposed to infectious*. When the virus reaches the host, it becomes infectious after a period(the latent period t_v^L). Consequently,

$$s_v^t = \begin{cases} I, & \text{if } s_v^{t-1} = E \text{ and } \tilde{t}_v > t_v^L \\ E, & \text{if } s_v^{t-1} = E \text{ and } \tilde{t}_v \leq t_v^L \end{cases} \tag{23}$$

where $\tilde{t}_v$ stands for the discrete steps of time passed from the acquisition of the mobile malware.

(3) *Transition from infectious to susceptible*. If there is a security application installed in the smartphone v, the mobile malware can be detected with a certain probability d_v, then

$$s_v^t = \begin{cases} S, & \text{if } s_v^{t-1} = I \text{ and } D = 1 \\ I, & \text{if } s_v^{t-1} = I \text{ and } D = 0 \end{cases} \quad (24)$$

where

$$D = \begin{cases} 1, & \text{with probability } d_v \\ 0, & \text{with probability } 1 - d_v \end{cases} \quad (25)$$

In Figure 11(a), the global evolution of the different compartments (susceptible—green, carrier—blue, exposed—orange, and infectious—red) is shown when it is supposed that the distribution of the operative systems is as follows: 60% of smartphones is based on Android, 20% is based on iOS, and 5% of mobile devices are based on both Blackberry OS and Windows Mobile. In addition, we state that 20% of Android-based smartphones are infectious at time $t = 0$. As shown, the system evolves to a quasi-endemic equilibrium with periodic outbreaks (this trend appears in all the simulations). As this is an individual-based model, not only the global dynamic of the system can be modeled using cellular automata but also the individual dynamic of each smartphone can be obtained; in this sense, in Figure 11(b), the evolution of the state of every smartphone is shown when all are based on Android OS and 20% of them is initially infectious. In this figure, each row represents the evolution of an individual smartphone where susceptible periods are in green, latent periods are in orange, and infectious periods are in red. Note that, as in the previous case, there is a reinfection for several smartphones, and the disease-free equilibrium is not reached.

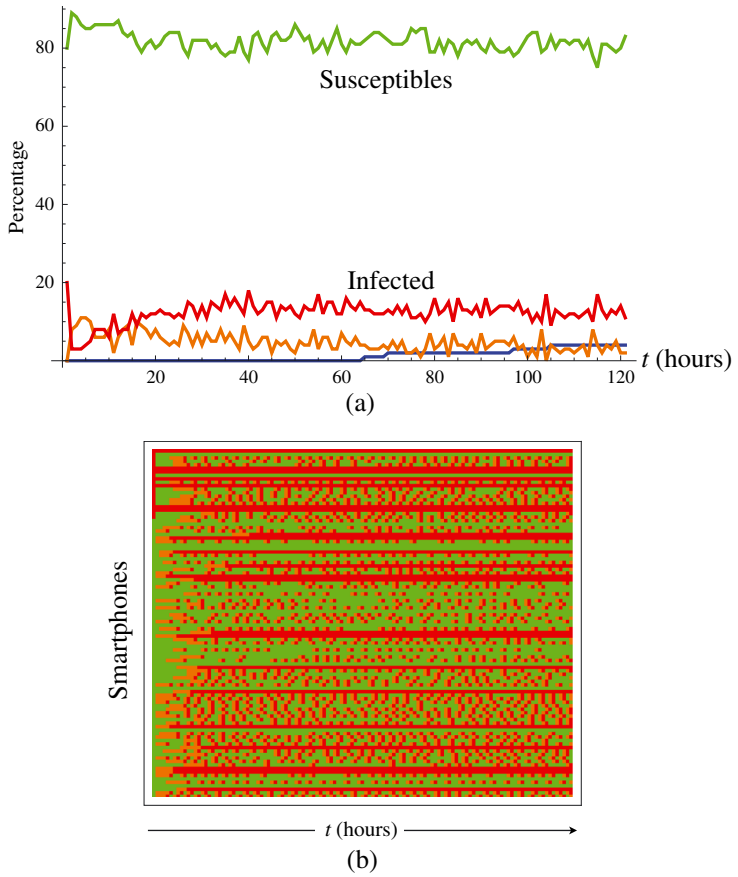

Figure 11. (a) Global evolution when different operative systems are considered. (b) Individual evolution when same operative systems are considered.

4.4. Discussion

As seen previously, most models proposed for the study of the propagation of malicious code, both on computer networks and on mobile environment, are deterministic and are based on systems of differential equations. These are well-founded, coherent models from the mathematical point of view, offering a detailed study of the main characteristics of their dynamic: stability, equilibrium, and so on. Nevertheless, they do have some drawbacks that, owing to their importance, merit attention.

(1) They do not take into account local infections between the computers forming the network. Parameters such as the rate of infection and the rate of recovery are used but they are of a general nature. The value of the parameter is constant for all the elements of the network, or in some cases, it follows a given probability distribution. Accordingly, the use of parameters individualized for each of the elements of the network is not considered. The infection rate could vary from computer to computer as a function of the characteristics of the network administrator or of the actual user of the computer. For example, the infection rate should fall if the user is worried about security, has an anti-virus application running, periodically updates the software, and so on. By contrast, the infection rate should be higher if the user engages in risky practices, such as visiting web sites that could be suspected of harboring malicious code, does not have an antivirus program installed or updated on the computer, and opens files and e-mails indiscriminately. Accordingly, it seems reasonable to search for a mathematical model that will take these aspects into account.
(2) They assume that the elements forming the network (through which the malware is propagated) are distributed homogeneously and that all are connected with one another. Moreover, in the mobile malware case, it is suppose that infected and susceptible devices mix completely with each other and move randomly within an arena of fixed size. When the propagation of malware is analyzed macroscopically (across the whole internet, for example), the results obtained provide a fairly good approximation of what is really happening. However, if we analyze such propagation in local networks, intranets, and so on, the results obtained are manifestly poorer because at microscopic scale, the dynamic is very sensitive to local interconnections.
(3) They are unable to simulate the individual dynamic of each of the elements of the network. It is true that when the size of the network is very large, the overall behavior observed may seem very similar (as regards trends) to what is happening in reality, but the use of essential information is neglected, for example, computers whose operative system is Mac OS should not be affected (in the sense of infected) by malware specifically designed for systems using Windows (although they could be considered as exposed) and so on. Thus, in models based on differential equations, we can obtain good results about the global behavior, but we shall lack information about the individual behavior of each of the computers in the network.

To overcome the first drawback, we have to consider individual-based models that capture the individual characteristics of each element of the system. In addition, a numerical estimation of the model's coefficients that reflect these characteristics must be given; these values could be obtained from the data collected by the network administrators and other responsibilities. For example, the information needed in order to obtain the coefficients of the system is the following: software installed in the devices and software update policy, security software installed in the devices and security measures implemented in the network, and characteristics of the users (users' profile, privacy attitudes, privileges, etc.). The second drawback could be offset if the model is endowed with the appropriate topologies. It is necessary to associate to each transmission vector its topology by means of a graph. For example, in the case of the models simulating the propagation of mobile malware, we can consider as vector transmissions the bluetooth connections (the topology will be defined by an undirected graph where two nodes/smartphones are adjacents if they are in the effective range of the bluetooth), the instant messaging app (the topology will be given by a directed graph representing the app address book), the e-mails (a directed graph modeling the e-mail address book will be used), and so on. Finally, the third deficiency can be overcome also using individual-based models because they provide not only the individual evolution of every element of the system but also its global dynamics.

These three main deficiencies shown by models based on differential equations can be rectified simply if, for example, we use a different type of model based on cellular automata or those based on agents. In these, it would be possible to take into account the individual characteristics of each of the computers or devices that are connected to the network (Figure 12). Moreover, we could consider different network topologies (power law, small-word, and random graphs) and even vary them with time. In this sense, some of the models based on differential equations for modeling malware epidemic spreading in topological networks largely overestimate epidemic spreading speed due to their implicit homogeneous mixing assumptions [80]. In this way, we would have defined a model in which the dynamic varies as a function of the different individual parameters.

The individualized behavior provided by these models of each of the components of the network would be of great use when performing forensic analyses when challenged by security breaches. It would be possible to trace the dynamic of the malicious code and from this draw conclusions about how to improve computer networks.

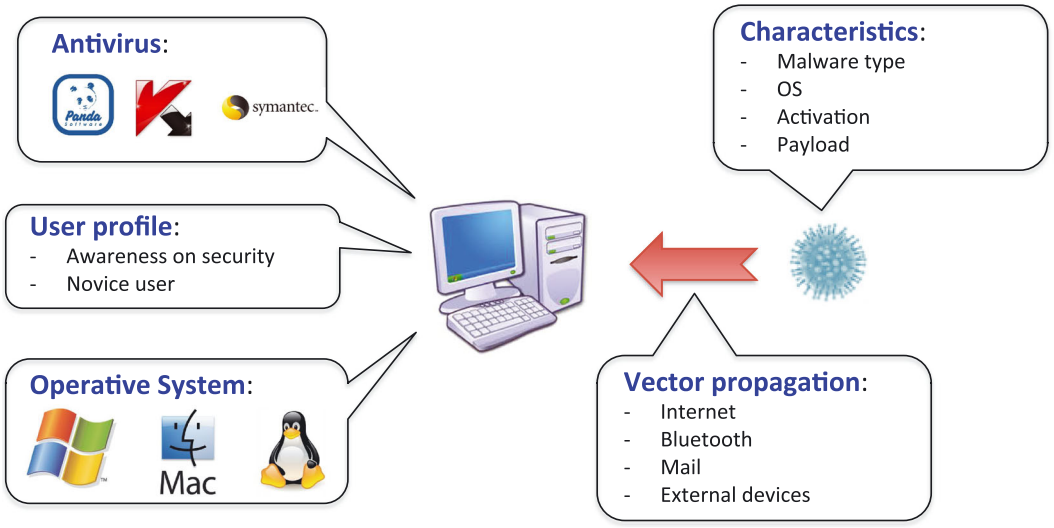

Figure 12. Individual characteristics of each element of the network that could be captured with the use of individual based discrete models.

Nevertheless, this alternative approach exhibits also some problems. The most important are the following: (i) the computational cost is too high when the number of devices or computers is large; and (ii) it is necessary to have several data of all elements (malware and devices characteristics) of the system to accurately determine the simulations.

Finally, some future improvements proposed for models based on cellular automata could be the following:

(1) Not only the bluetooth connections must be considered as transmission vector but also instant messaging apps, downloading from app markets, and so on.
(2) New compartment must be considered: the recovered smartphones (those infected devices where the worm has been successfully detected and removed and are endowed with an immunity period).
(3) Modified Boolean functions ruling the transition from susceptible to exposed or carrier must be defined including a connectivity factor of neighbor cells, a memory infection coefficient (considering the number of reinfections of each smartphone to train it to decrease the probability of a new infection), and modified parameters in order to obtain more realistic simulations.

5. CONCLUSIONS

Here, we address the mathematical models proposed to date to study the behavior of epidemics produced by malicious code in computer networks. Most of them are deterministic and use systems of ordinary differential equations. They are eminently mathematical models that have not been implemented computationally to give rise to a simulator that would help information security managers make decisions.

These models have a series of deficiencies that mean that they are not suitable for making decisions when threatened by malware. Such deficiencies could be overcome by using discrete models (deterministic or stochastic), such as models based on agents or models based on cellular automata.

As far as we are aware, very few models of this type based on cellular automata and the development of models based on agents have been proposed. The results obtained with them have been fairly satisfactory and promising.

Future lines of investigation in this field should address the improvement of models based on cellular automata and the development of agent-based models. Application of these—to the new scenarios the propagation of malware through networks of mobile devices and the Internet of Things—must also be explored in depth.

ACKNOWLEDGEMENTS

This work has been supported by Junta de Castilla y Leon (Spain) and Ministerio de Economia y Competitividad (Spain).

REFERENCES

1. Janssens ML. *Introduction to Mathematical Fire Modeling*, 2nd edn. CRC Press: Lancaster, Pennsylvania, USA, 2000.

2. Vynnycky E, White RG. *An Introduction to Infectious Disease Modelling*. Oxford University Press: New York, USA, 2010.
3. Tipton HF, Krause M. *Information Security Management Handbook*, 6th edn. Auerbach Publications: Boca Raton, Florida, USA, 2010.
4. McAfee labs threats report, 2014. (Available from: http://www.mcafee.com/sg/resources/reports/rp-quarterly-threat-q1-2014.pdf) [Accessed on July 2014].
5. PandaLabs. Quarterly report pandalabs, 2014. (Available from: http://press.pandasecurity.com/wp-content/uploads/2014/05/Quaterly-PandaLabs-Report_Q1.pdf) [Accessed on July 2014].
6. Upadhyay RK, Iyengar SRK. *Introduction to Mathematical Modeling and Chaotic Dynamics*. Chapman and Hall/CRC: Boca Raton, Florida, USA, 2013.
7. Cohen F. Computer virus: theory and experiments. *Computer Security* 1987; **6**: 22–35.
8. Meisel M, Pappas V, Zhang L. A taxonomy of biologically inspired research in computer networking. *Computer Networks* 2010; **54**: 901–916.
9. Murray WH. The application of epidemiology to computer viruses. *Computer Security* 1988; **7**: 139–145.
10. Rice M, Butts J, Miller R, Shenoi S. Applying public health strategies to the protection of cyberspace. *International Journal of Critical Infrastructure Protection* 2010; **3**: 118–127.
11. Kephart JO, White SR. Directed-graph epidemiological models of computer viruses, *Proceedings of IEEE Symposium on Security and Privacy*, Oakland, CA, 1991; 343–359.
12. Kephart JO, White SR. Measuring and modeling computer virus prevalence, *Proceedings of IEEE Symposium on Security and Privacy*, Oakland, CA, 1993; 2–15.
13. Kephart JO, White SR, Chess DM. Computers and epidemiology. *IEEE Spectrum* 1993; **30**: 20–26.
14. Amador J, Artalejo JR. Modeling computer virus with the BSDE approach. *Computer Networks* 2012; **57**: 302–316.
15. Rao NS, Deepshikha J. A deterministic approach for the propagation of computer virus in the framework of linear and sinusoidal time variation of birth rate of virus. *Proceedings of ICISTM 2011, Communications in Computer and Information Science* 2011; **141**: 206–213.
16. Sanders J, Noble B, Van Gorder RA, Riggs C. Mobility matrix evolution for an SIS epidemic patch model. *Physica A* 2012; **391**: 6256–6267.
17. Wang Y, Cao J, Jin Z, Zhang H, Sun GQ. Impact of media coverage on epidemic spreading in complex networks. *Physica A* 2013; **23**: 5824–5835.
18. Tomovski I, Trpevski I, Kocarev L. Topology independent SIS process: an engineering viewpoint. *Communications in Nonlinear Science* 2014; **19**: 627–637.
19. Piqueira JRC, Araujo VO. A modified epidemiological model for computer viruses. *Applied Mathematics and Computation* 2009; **213**: 355–360.
20. Ren J, Yang X, Yang LX, Xu Y, Yang F. A delayed computer virus propagation model and its dynamics. *Chaos Soliton and Fractals* 2012; **45**: 74–79.
21. Wierman JC, Marchette DJ. Modeling computer virus prevalence with a susceptible-infected-susceptible model with reintroduction. *Computational Statistics & Data Analysis* 2004; **45**: 3–23.
22. Zhu Q, Yang X, Ren J. Modeling and analysis of the spread of computer virus. *Communications in Nonlinear Science* 2012; **17**: 5117–5124.
23. Shukla JB, Singh G, Shukla P, Tripathi A. Modeling and analysis of the effects of antivirus software on an infected computer network. *Applied Mathemaics and Computation* 2014; **227**: 11–18.
24. Dagon D, Zou C, Lee W. Modeling botnet propagation using time zones, *Proc. 13th Annual Netw. Dis. Syst. Secur. Symp.*, San Diego, California, USA, 2006; 235–249.
25. Mishra BK, Saini DK. SEIRS epidemic model with delay for transmission of malicious objects in computer network. *Applied Mathematics and Computation* 2007; **188**: 1476–1482.
26. Wang F, Zhang Y, Wang C, Ma J. Stability analysis of an e-SEIAR model with point-to-group worm propagatio. *Communications in Nonlinear Science* 2015; **20**(3): 897–904.
27. Mishra BK, Pandey SK. Dynamic model of worms with vertical transmission in computer network. *Applied Mathematics and Computation* 2011; **217**: 8438–8446.
28. Toutonji OA, Yoo SM, Park M. Stability analysis of VEISV propagation modeling for network worm attack. *Applied Mathematical Modelling* 2012; **36**: 2751–2761.
29. Yang Y. A note on global stability of VEISV propagation modeling for network worm attack. *Applied Mathematical Modelling* 2015; **39**(2): 776–780.
30. Martín del Rey A, Rodríguez Sánchez G. A discrete mathematical model to simulate malware spreading. *Intenational Journal of Modern Physics C* 2012; **23**: 1250064.
31. Mishra BK, Jha N. SEIQS model for the transmission of malicious objects in computer network. *Applied Mathematical Modelling* 2010; **34**: 710–715.
32. Mishra BK, Keshri N. Mathematical model on the transmission of worms in wireless sensor net-

work. *Applied Mathematical Modelling* 2013; **37**: 4103–4111.
33. Mishra BK, Pandey SK. Dynamic model of worm propagation in computer network. *Applied Mathematical Modelling* 2014; **38**: 2173–2179.
34. Wang F, Zhang Y, Wang C, Ma J, Moon SJ. Stability analysis of a SEIQV epidemic model for rapid spreading worms. *Computer Security* 2010; **29**: 410–418.
35. Amador J, Artalejo JR. Stochastic modeling of computer virus spreading with warning signals. *Journal Franklin I* 2013; **350**: 1112–1138.
36. Feng L, Liao X, Han Q, Li H. Dynamical analysis and control strategies on malware propagation model. *Applied Mathematical Modelling* 2013; **16-17**: 8225–8236.
37. Haldar K, Mishra BK. A mathematical model for a distributed attack on targeted resources in a computer network. *Communications in Nonlinear Science* 2014; **19**: 3149–3160.
38. Yang LX, Yang X, Zhu Q, Wen L. A computer virus model with graded cure rates. *Nonlinear Analysis-Real World Applications* 2013; **14**: 414–422.
39. Yang LX, Yang X. A new epidemic model of computer viruses. *Communications in Nonlinear Science* 2014; **19**: 1935–1944.
40. Yang LX, Yang X. The effect of infected external computers on the spread of viruses: a compartment modeling study. *Physica A* 2013; **392**: 6523–6535.
41. Zou C, Gong W, Towsley D. Code red worm propagation modeling and analysis, *Proceedings of 9th ACM Conf. Comput. Commun. Secur.*, Washington DC, USA, 2002; 18–22.
42. Billings L, Spears WM, Schwartz IB. A unified prediction of computer virus spread in connected networks. *Physics Letters A* 2002; **298**: 261–266.
43. Mishra BK, Pandey SK. Effect of anti-virus software on infectious nodes in computer network: a mathematical model. *Physics Letters A* 2012; **376**: 2389–2393.
44. Mishra BK, Saini DK. Mathematical models on computer viruses. *Applied Mathematical and Computation* 2007; **187**: 929–936.
45. Yao Y, Guo L, Guo H, Yu G, Gao FX, Tong XJ. Pulse quarantine strategy of internet worm propagation: modeling and analysis. *Computational Electronics Engineering* 2012; **38**: 1047–1061.
46. Kondakci S. Epidemic state analysis of computers under malware attacks. *Simulation Modelling Practice and Theory* 2008; **16**: 571–584.
47. Kondakci S, Dincer C. Internet epidemiology: healthy, susceptible, infected, quarantined, and recovered. *Security in Communication Networks* 2011; **4**: 216–238.
48. Mishra BK, Goswami RT. Probabilistic e-epidemic model on computer worms, *Proceedings of ICIEV 2012*, Dhaka, Bangladesh, 2012; 1091–1096.
49. Okamura H, Kobayashi H, Dohi T. Markovian modeling and analysis of Internet worm propagation, *Proceedings of 16th IEEE International Symposium Software, Reliability Engineering*, Chicago, IL, 2005; 149–158.
50. Hao J, Yin J, Zhang B. Modeling viral agents and their dynamics with persistent turing machines and cellular automata, *Proceedings of 9th Pacific Rim International Conference Agent Computing and Multi-Agent Systems, Lecture Notes Computer Science*, Guilin, China, 2006, 4088; 690–695.
51. Martín del Rey A. A computer virus spread model based on cellular automata of graphs, *Proceedings of 10th International Conference Artificial Neural Networks, Lecture Notes Computer Science*, Salamanca, Spain, 2009, 5518; 503–506.
52. Song Y, Jiang GP, Gu Y. Modeling malware propagation in complex networks based on cellular automata, *Proceedings of 2008 IEEE Asia Pacific Conference on Circuits and Systems*, Macao, China, 2008; 259–263.
53. Ge ST, Tang GY, Yang X, Xu QL, Yu H, Wang PD. Stability analysis of computer virus model system in networks. *Applied Mechanics and Materials* 2013; **278**: 2033–2038.
54. Pan J, Fung CC. An agent-based model to simulate coordinated response to malware outbreak within and organization. *International Journal of Information Security* 2012; **5**: 115–131.
55. Saini DK. A mathematical model for the effect of malicious object on computer network immune system. *Applied Mathematical Modelling* 2011; **35**: 3777–3787.
56. Huang CY, Lee CL, Wen TH, Sun CT. A computer virus spreading model based on resource limitations and interaction costs. *Journal of Systems and Software* 2013; **86**: 801–808.
57. Kermack WO, McKendrick AG. Contributions to the mathematical theory of epidemics. *Proceedings of the Royay Society A* 1927; **115**: 700–721.
58. Hethcote HW. Qualitative analyses of communicable disease models. *Mathematical Biosciences* 1976; **28**: 335–356.
59. Diekmann O, Heesterbeek JAP. *Mathematical Epidemiology of Infectious Diseases*. John Wiley & Sons: Chichester, West Sussex, England, 2000.
60. Hethcote HW. The mathematics of infectious diseases. *SIAM Review* 2000; **42**: 599–653.
61. Wolfram S. *A New Kind of Science*. Wolfram Media Inc.: Champaign, Illinois, USA, 2002.

62. Sarkar P. A brief history of cellular automata. *ACM Computing Surveys* 2000; **32**: 80–107.
63. Hoekstra AG, Kroc J, Sloot PMA (eds.) *Simulating Complex Systems by Cellular Automata*. Springer: Berlin, Heidelberg, Germany, 2010.
64. Jimenez A. Cellular automata to describe seismicity: a review. *Acta Geophysica* 2013; **61**: 1325–1350.
65. Precharattana N, Triampo W. Modeling dynamics of HIV infected cells using stochastic cellular automata. *Physica A* 2014; **407**: 303–311.
66. López L, Burguerner G, Giovanini L. Addressing population heterogeneity and distribution in epidemics models using a cellular automata approach. *BMC Res. Notes* 2014; **7**: 234–245.
67. Karafyllidis I, Thanailakis A. A model for predicting forest fire spreading using cellular automata. *Ecological Modelling* 1997; **99**: 87–97.
68. Yassemi S, Dragićević S, Schmidt M. Design and implementation of an integrated GIS-based cellular automata model to characterize forest fire behaviour. *Ecological Modelling* 2008; **210**: 71–84.
69. Cheng SM, Ao WC, Chen PY, Chen KC. On modeling malware propagation in generalized social networks. *IEEE Communications Letters* 2011; **15**: 25–27.
70. Jackson JT, Creese S. Virus propagation in heterogeneous bluetooth networks with human behaviors. *IEEE Transactions on Dependable and Secure Computing* 2012; **9**: 930–943.
71. Mickens JW, Noble BD. Modeling epidemic spreading in mobile environments, *Proceedings of 4th ACM Workshop on Wireless Security*, Cologne, Germany, 2005; 77–86.
72. Ramachandran K, Sikdar B. On the stability of the malware free equilibrium in cell phones networks with spatial dynamics, *Proceedings of 2007 IEEE International Conference on Communications*, Glasgow, Scotland, 2007; 6169–6174.
73. Ramachandran K, Sikdar B. Modeling malware propagation in networks of smart cell phones with spatial dynamics, *Proceedings of 26th IEEE International Conference on Computer Communications*, Anchorage, Alaska, USA, 2007; 2516–2520.
74. Rhodes CJ, Nekovee M. The opportunistic transmission of wireless worms between mobile devices. *Physica A* 2008; **387**: 6837–6844.
75. Wei X, Zhao-hui L, Zeng-qiang C, Zhu-zhi Y. Commwarrior worm propagation model for smart phone networks. *Journal of China University Posts Telecommunication* 2008; **15**: 60–66.
76. Merler S, Jurmanm G. A combinatorial model of malware diffusion via Bluetooth connections. *PLos ONE* 2013; **8**, art. no. e59468.
77. Peng S, Wu M, Wang G, Yu S. Propagation model of smartphone worms based on semi-Markov process and social relationship graph. *Computer Security* 2014; **44**: 92–103.
78. Peng S, Wang G, Yu S. Modeling the dynamics of worm propagation using two-dimensional cellular automata in smartphones. *Journal of Computational and System Sciences* 2013; **79**: 586–595.
79. Martín del Rey A, Rodríguez Sánchez G. A CA model for mobile malware spreading based on bluetooth connections. *Advances in Intelligent Systems Computations* 2014; **234**: 619–629.
80. Zou C, Gong W, Towsley D. Modeling and simulation study of the propagation and defense of internet email worm. *IEEE Transactions on Dependable and Secure Computing* 2007; **4**: 105–118.