Mathematical Models and Software for Modelling the Spread of Malware in Energy Facilities

Leonid Galchynsky[a], Vladyslav Khaidurov[b], Tamara Tsiupii[c] and Tetiana Zhovnovach[d]

[a] *National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 37, Prosp. Peremohy st., Kyiv, 03056, Ukraine*
[b] *Institute of General Energy of NAS of Ukraine, 172, Antonovycha st., Kyiv, 03150, Ukraine*
[c] *National University of Life and Environmental Sciences of Ukraine, 15, Heroyiv Oborony st., Kyiv, 03041, Ukraine*
[d] *Cherkasy Branch of the European University, Smilyanska st., 83, Cherkasy, 18000, Ukraine*

Abstract

The result of the work is an optimization mathematical model and a corresponding description of the implementation of a complex software tool for modeling the spread of malicious software (malware) in modern energy facilities and systems. The developed optimization mathematical model is based on the use of methods of optimization of functions and functionals with constraints in the form of systems of ordinary differential equations (SODEs) with given corresponding initial conditions. To develop process simulation software modules based on the PSIDR mathematical model, stochastic population methods, models and algorithms were used to determine the control parameter at each time step. The use of such optimization methods and algorithms makes it possible to solve more complex tasks that require a procedure for predicting the spread of processes of various origins in general. The developed mathematical model consists in the minimization of costs for the purchase of antiviruses for the protection of relevant systems in energy facilities and systems.

Keywords 1

Malware, prediction, optimization, stochastic model, deterministic model, cellular automaton, energy objects.

1. Introduction

In today's conditions, it is impossible to imagine real life without the use of modern computer technology, smartphones, gadgets and other devices. Every person who lives in today's conditions uses the global Internet. The use of personal computers and mobile devices in combination with the Internet has become an integral part of everyday life. The constant use of the Internet has caused many serious threats to the privacy and security of personal data of one or more users. Every year there is a significant increase in cyber-attacks with an increase in their complexity. Such attacks affect electronic resources, governments, businesses, and individual or legal entities, and cause serious financial and social damage to them. No person is guaranteed 100% protection against damage to their own device by malicious software (malware). For commercial firms and/or banking systems, it is necessary to understand the level of losses over time, which are obtained as a result of the operation of malware in a specific area. In this regard, the issue of predicting the spread of malware in networks and the question of how to respond in the event of malware entering a serious company or firm is quite acute.

Assessing the level of damage requires taking into account many factors, but first of all, it concerns the issue of assessing the dynamics of the spread of the epidemic in computer networks. This issue is quite general, since epidemics in human communities have existed for a long time as a threat factor for

ITTAP'2022: 2nd International Workshop on Information Technologies: Theoretical and Applied Problems, November 22–24, 2022, Ternopil, Ukraine
EMAIL: hleonid@gmail.com (A. 1); allif0111@gmail.com (A. 2); ts.tamara19@gmail.com (A. 3); z.ta@ukr.net (A. 4)
ORCID: 0000-0002-3805-1474 (A. 1); 0000-0002-4805-8880 (A. 2); 0000-0003-2206-2897 (A. 3); 0000-0003-1037-4383 (A. 4)

people and bring huge losses. Therefore, epidemiology has accumulated a significant baggage of models of the dynamics of epidemics and estimates of damages from them. There is a lot of literature on this subject, in particular [1–3]. And therefore, it is worth taking into account this accumulated experience. At the same time, it is necessary to take into account that epidemics in computer networks must differ from epidemics in human communities and the economy [3-5]. The application of epidemiology models for computer networks should be critically reconsidered, in addition, new approaches should be proposed [6; 7].

2. The main part of the article

The problem of forecasting and the problem of optimization are quite often related to each other. To carry out effective forecasting in any field, it is necessary to minimize errors on those data, which are presented in the form of statistics for carrying out the training procedure with or without a teacher [8].

Below, in this section, an analysis of some existing models for predicting the spread of malware is carried out and a description of various methods and algorithms, mathematical foundations, which are applied to statistical data processing, as well as based on gradient methods for finding extrema of multidimensional functions [9] is performed. Effective forecasting, in particular, forecasting the spread of malware in real objects and systems, requires the use of up-to-date machine learning methods, including optimization methods.

2.1. Mathematical models and software complex for modeling the spread of malicious software

Several approaches are used to develop a complex software application for modeling the spread of malicious software. The first approach is to use a modification of classical deterministic models (DMs) to investigate the state of propagation of malware epidemics. The second approach is more universal - modeling the spread of malicious software based on stochastic models (SM), for example, using random processes, Markov chains, cellular automata, neural networks, etc. [5; 6]. Models based on classical systems of ordinary differential equations (SODE) and Cauchy problems for them are described below, as well as general principles of modeling these problems using classical cellular automata using different neighborhoods – Neumann neighborhood and Moore neighborhood.

2.1.1. Deterministic mathematical models of the problem of spreading malicious software

SIR malware propagation model. As a basic mathematical model, we will take the classic SIR model [8]. This model can describe the spread of many different processes, and these processes can be both physical and technical, and biological or economic.

Let us have some network that connects several computers. It can be a local network, or a network that has access to global Internet resources. It is clear that the introduction of a malware onto one personal computer can lead to the spread of this software throughout the network. Of course, the speed of software distribution through this network is governed by various factors [6; 7]. Such factors include, for example:

- activity of users in this network in general;
- the target orientation of the software and the speed of its distribution, which is controlled by the software code from which it is composed;
- a combination of factors;
- other factors.

We will assume that the target nodes of the network can be in one of three given states:

- non-infected nodes of the network, which can be exposed to the entry of malware into them. Let's denote them as S – susceptible;

– infected nodes of the network, i.e., those that have already been infected with a malware infection. Regarding the operation of such nodes, it is necessary to take immediate measures to preserve (including integrity, if possible) data on such nodes. Let's denote them as *I* –infective;
– network nodes that are not conducive to infection, i.e., those on which appropriate protection measures are installed (for example, network screens, anti-virus systems, etc.), and in such nodes, the corresponding of malware has been removed. We will call such nodes as *R* – removed.

The process of the spread of malware depends on only a few parameters:
– the total number of system nodes, which is divided into the number of uninfected nodes, the number of infected nodes and the number of restored (unfavorable) nodes in a specific network;
– the speed of propagation of malware;
– the speed of treatment of network nodes in the event of a specific of malware falling into it (the speed of response and adoption of appropriate protective measures regarding the treatment of network nodes).

It is assumed that the number of network nodes is constant during the entire period of research of this network. This means that nodes are not added to or removed from the network. Let's describe this model mathematically:

Let *N* – the total number of computers (nodes) in the network:

$$S(t) + I(t) + R(t) = N.$$

Changes in the state of the network can be described in the form of SODE, which has the following form:

$$\begin{cases} \dfrac{dS(t)}{dt} = -\dfrac{-\beta I(t)}{N} S(t), \\ \dfrac{dI(t)}{dt} = \dfrac{\beta I(t)}{N} S(t) - \sigma I(t), \\ \dfrac{dR(t)}{dt} = \sigma I(t), \\ \dfrac{dS(t)}{dt} + \dfrac{dI(t)}{dt} + \dfrac{dR(t)}{dt} = 0. \end{cases} \quad (1)$$

In systems of differential equations (1) β – the speed of infection of system nodes, a σ – the speed of treatment of a system node (the speed of taking measures to ensure the protection of an already infected system node).

It is clear that the reproduction of the dynamics of the mathematical model (1) requires setting the initial conditions of the species system:

$$S(0) = S_0, \;\; I(0) = I_0, \;\; R(0) = R_0. \quad (2)$$

Now the mathematical model (1) together with the initial conditions (2) to it sets Cauchy problems, which can be solved by the usual approximate methods, for example, Euler methods or Runge-Kutta methods.

Malware distribution model SAIR. Unlike the *SIR* model, in this model all computer network nodes will be divided into 4 fixed categories:
– the number of uninfected nodes (personal computers) of the network that may be infected is denoted as *S*;
– the number of network nodes (personal computers) that have an installed protection system, for example, an anti-virus system, denoted as *A*;
– the number of infected network nodes (personal computers) is denoted as *I*;
– the number of nodes (personal computers) of the network, which have been cured, denoted as *R*;

The spread of the process of infection of computer network elements can be presented in the form of SODE as follows:

$$\begin{cases} \dfrac{dS(t)}{dt} = -\dfrac{-\beta I(t)}{N} S(t), \\ \dfrac{dI(t)}{dt} = \dfrac{\beta I(t)}{N} S(t) - \sigma I(t), \\ \dfrac{dR(t)}{dt} = \sigma I(t), \\ \dfrac{dS(t)}{dt} + \dfrac{dI(t)}{dt} + \dfrac{dR(t)}{dt} = 0. \end{cases} \quad (3)$$

In model (3), the initial conditions of type (2) should also be taken into account, but for nodes of all 4 states in order to start the process of predicting the spread of malware in the corresponding network system.

It should also be noted that in model (3): N – the total number of new nodes added to this network; μ – the frequency of removal of system nodes that cannot be restored for reasons not related to the results of malware; β_{SI} – frequency of infection of favorable nodes of the network; β_{AI} – the frequency of infection of system nodes on which protection systems such as antivirus systems or network screens are installed; δ – the frequency with which the nodes exposed to malware fail; σ_{IS} – frequency of treatment of network objects (nodes) that were infected; σ_{RS} – the frequency of treatment of infected nodes of the system (network) with the participation of the operator; α – the intensity of the appeal of favorable ones to those who establish a protection system. If we take into account the fact that new nodes are not added to this network, and already existing nodes of the network fail only due to the influence of malware, i.e. $N = \mu = 0$, then the mathematical model of the propagation of malware takes the following form:

$$\begin{cases} \dfrac{dS}{dt} = -\alpha SA - \beta_{SI} SI + \sigma_{IS} I + \sigma_{RS} R, \\ \dfrac{dI}{dt} = \beta_{SI} SI + \beta_{AI} AI - \sigma_{IS} I - \delta I, \\ \dfrac{dR}{dt} = \delta I - \sigma_{RS} R, \\ \dfrac{dA}{dt} = \alpha SA - \beta_{AI} AI. \end{cases} \quad (4)$$

For the system of the form (4), the initial conditions of the form (2) are also set, but for all 4 states of the nodes in the computer network.

Malware distribution model PSIDR. Any mathematical model that describes some process, system or set of systems must take into account many factors.

The mathematical model involves two main stages. The first stage consists in the fact that malware propagates quite quickly and freely, that is, the nodes of the computer network have only two states – *S* and *I*.

The second stage provides for the procedure of responding to the existing malware in the computer network. During a certain period of time, the malware is identified in the network. Next, the procedure for treating nodes (personal computers) of the network is performed. That is, xi computers that were not infected are automatically "vaccinated", that is, an increased protection system is imposed on them. Nodes of the system, which have already been infected, form the so-called "immunity" and are cured from malware entered into them. As before, there are two parameters in the mathematical model, which characterize the speed of propagation of malware – the parameter μ and the treatment rate of the network nodes is the speed parameter δ.

In the mathematical model *PSIDR* again: *I* – the total number of infected elements (nodes) of the computer network. These elements of the network are the spreaders of malware. *S* – the total number of uninfected nodes of the computer network. These network nodes can be exposed to infection during the entire time of existence in this network. *R* – the total number of cured computer network elements. These nodes have a so-called "immunity" to infection. The last type of nodes in the system is this *D* – the total number of infected objects found in the computer network system.

The description given above makes it possible to write down a suitable mathematical model for predicting the spread of malware in this network. The mathematical model looks like this:

$$\begin{cases} \dfrac{dS(t)}{dt} = -\beta S(t)I(t) - \mu S(t), \\ \dfrac{dI(t)}{dt} = \beta S(t)I(t) - \mu I(t), \\ \dfrac{dR(t)}{dt} = \sigma D(t) + \mu S(t), \\ \dfrac{dD(t)}{dt} = \mu I(t) - \sigma D(t), \\ \dfrac{dS(t)}{dt} + \dfrac{dI(t)}{dt} + \dfrac{dR(t)}{dt} + \dfrac{dD(t)}{dt} = 0. \end{cases} \quad (5)$$

There are three parameters in the mathematical model (5). All these three parameters characterize speeds, in particular: β – the speed of infection of network nodes; σ – the speed of treatment of network nodes; μ – the speed of identification of network nodes with malware. The peculiarity of this model is that the antivirus protection must be updated as early as possible, since the delay in updating the antivirus protection in each network node leads to a greater spread of WPS in this network.

For the mathematical model (5), initial conditions of the form are given:

$$I(0) = I_0;\ S(0) = S_0;\ D(0) = R(0) = 0. \quad (6)$$

In this case, we obtain a Cauchy problem of the form (5)–(6), which can also be solved by any explicit approximate method, for example, the Euler method or the Runge-Kutta method.

To find an approximate solution of the Cauchy problem (5)–(6), the classical Euler method will be used. In the third section, the stages of designing modules that reproduce the work of a mathematical model based on the SODE of the form (5)–(6) will be given as a basis.

2.1.2. A stochastic model of the spread of malicious software

Cellular automata. Models of any complex real systems and their complexes, as well as simulation models are of high value due to the tasks of forecasting physical and technical, economic, ecological, social and other processes. Forecasting the spread of malware is quite similar to forecasting the spread of epidemiological processes. Epidemiological events caused by viruses, pollution from stationary sources of pollutants or new varieties of bioterrorism can cause significant costs to people and their resources. For this reason, it is important to study and model the spread of pathogens in the population.

A cellular automaton (CA) is an abstract object that consists of a set of cells, each of which has a certain finite state. The state can change over time, based on specific conditions that describe one or another process of spreading the phenomenon.

A cell is a separate element of cellular space, or the so-called "smallest unit of space".

Cellular space is a lattice space that consists of cells, and each cell is in one of several predefined states.

A CA state rule is a rule that governs the transition between cells and their states. The determination of the most appropriate transition rules for the phenomenon or process under study is of particular importance.

The definition of a cell state machine is called "local" because it uses only the neighborhood as input. Neighboring refers to the cells surrounding a particular cell, and they have the ability to influence the next state of that cell. The choice of neighborhood affects the behavior of the cell space. The choice of the appropriate neighborhood depends on the relationship between the elements.

The most important types of neighborhoods are von Neumann neighborhoods and Moore neighborhoods (Figure 1).

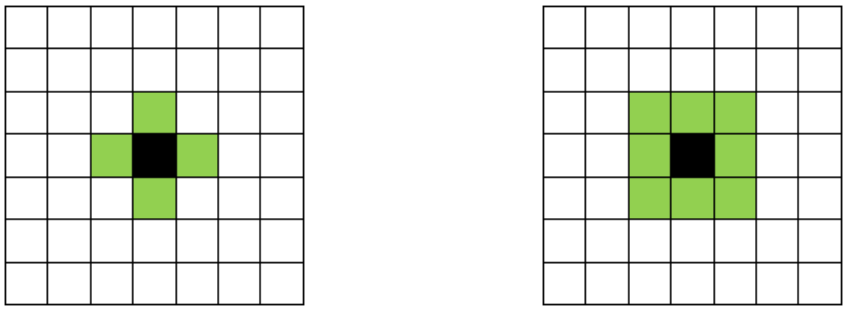

Figure 1: Around a cell in the two-dimensional space of the CA: on the left – the von Neumann neighborhood; on the right - around Moore

In practice, a larger number of neighborhoods and grids are actually used, but Figure 1 shows the ones that will be used in the development of the software for modeling the propagation of malware in a network.

Monte Carlo method. Probabilistic modeling methods are actively used to study the impact of random variability of various quantities on the properties of machines, on chemical or biological processes, on the bearing capacity or reliability of a structure, and in many other cases. A very powerful tool for the study of random processes is the Monte Carlo method.

The Monte Carlo method is a simple computer method that consists of conducting numerous mock experiments with random numbers. Its use is universal and does not require special knowledge of probability theory. The only information required is the ratio between the output and input values, in particular, in this form:

$$y = f(x_1, x_2, \cdots, x_n). \tag{7}$$

The method repeats trials with computer-generated random numbers processed by appropriate mathematical operations. In each "trial" the input variable x_1, x_2, …, x_n are assigned random values, but such that their distributions correspond to the probability distribution of each variable. With these values, the output value y is calculated according to the equation (7).

The generated values can be used to determine the average value or the probability that y will be lower or higher than the selected value y^*, or to define values that will be exceeded (or not reached) with a certain probability (eg time to failure).

The effectiveness of this approach is based on pseudorandom numbers, which are implemented in computing complexes.

Generating random numbers with given distributions. Modern application software often suggests using distributions such as uniform or normal. Random numbers corresponding to other analytically determined distributions can be generated using a uniform distribution.

2.2. A mathematical model of minimizing damages from the spread of malicious software

To build an optimal control model, consider the PSIDR DM, which was described in the first section of this work. In the model of optimal control, we will assume that the parameter μ is a control parameter and it is limited. We will have an optimization problem, which is described in the mathematical formulation below.

Let there be a functional of the form given below:

$$J(\mu(t)) = \int_0^T \left(A \cdot I(\mu(t)) + B \cdot D(\mu(t))\right) dt \to \min,$$

where

A – the fee for personal computers/laptops that were damaged by malware;

B – the fee for updating the anti-virus system for the personal computers and / or laptops under investigation;

T – the final time, during which the study of the behavior of malware in personal computers / laptops was carried out.

From a practical point of view, it is obvious that the possibilities of installing a new antivirus are limited, i.e. $0 \le \mu(t) \le M$.

Then the control problem is described by the following mathematical model. Find the global minimum of the functional:

$$J(\mu(t)) = \int_0^T \left(A \cdot I(\mu(t)) + B \cdot D(\mu(t))\right) dt \to \min, \tag{8}$$

taking into account the limitation that is described in the form of the PSIDR model itself, that is, in the form of the Cauchy problem for SODE:

$$\begin{cases} \dfrac{dS(t)}{dt} = -\beta S(t)I(t) - \mu(t)S(t), \\ \dfrac{dI(t)}{dt} = \beta S(t)I(t) - \mu(t)I(t), \\ \dfrac{dR(t)}{dt} = \sigma(t)D(t) + \mu(t)S(t), \\ \dfrac{dD(t)}{dt} = \mu(t)I(t) - \sigma(t)D(t). \end{cases} \tag{9}$$

The initial conditions of the problem look like this:

$$I(0) = I_0, S(0) = S_0, D(0) = R(0) = 0. \tag{10}$$

The restriction on the desired control parameter looks like this:

$$0 \le \mu(t) \le M. \tag{11}$$

The problem described in the form of mathematical model (8)–(10) can be solved by many methods and algorithms. In most cases, such problems do not have analytical solutions, and if they do, such solutions are not of interest, since real applied problems today can only be solved numerically. The Pontryagin maximum principle can be attributed to the analytical methods of finding the control parameter. The software implementation of the constraint (Cauchy problem) will be based on the use of Euler's classical method for Cauchy problems for ODE and SODE.

To increase the accuracy of solving the direct problem based on the mathematical model (9)–(10) with the known control parameter (11), it is possible to use methods of the Runge-Kutta type 2, 3, 4 and higher orders as desired. Minimization of the functional of the form (8) will be carried out using deterministic methods and population (multi-agent) stochastic methods and algorithms.

2.3. The structure of the developed software and the results of modeling the spread of malicious software

The developed software complex consists of several modules, each of which has a corresponding graphical user interface. The general structure of the software complex is presented in Figure 2 below.

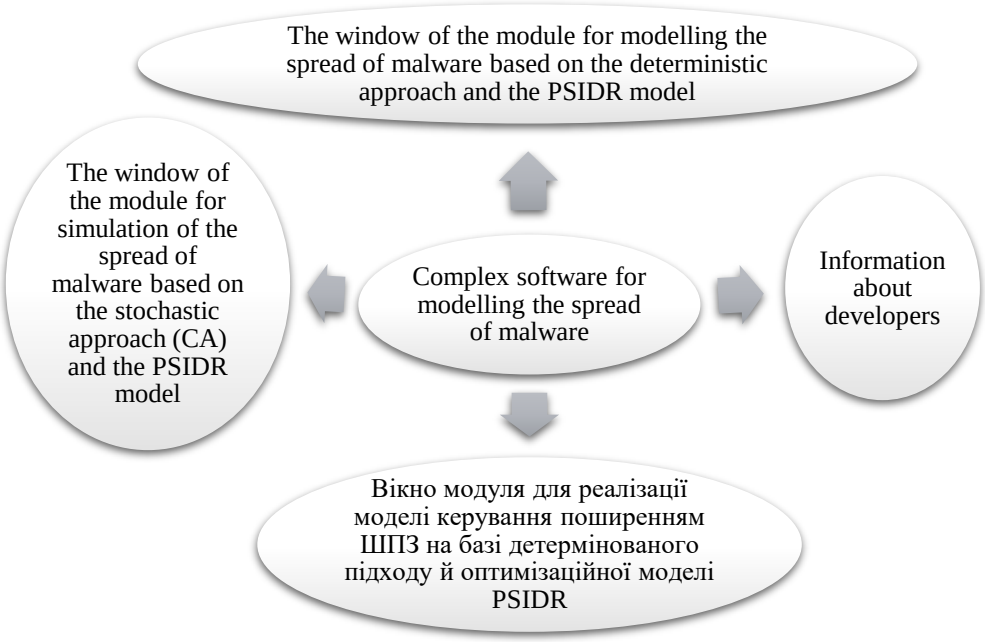

Figure 2: Structure of the software complex for modeling the spread of malicious software

The structure of the developed software complex can be modified by additional modules, since each model works independently of other modules of the software complex.

Figure 3 shows the interface of one of the modules of the complex software application – the implementation of the control model, which involves minimizing the total costs of purchasing and updating antivirus software systems, which is present in modern energy facilities and systems.

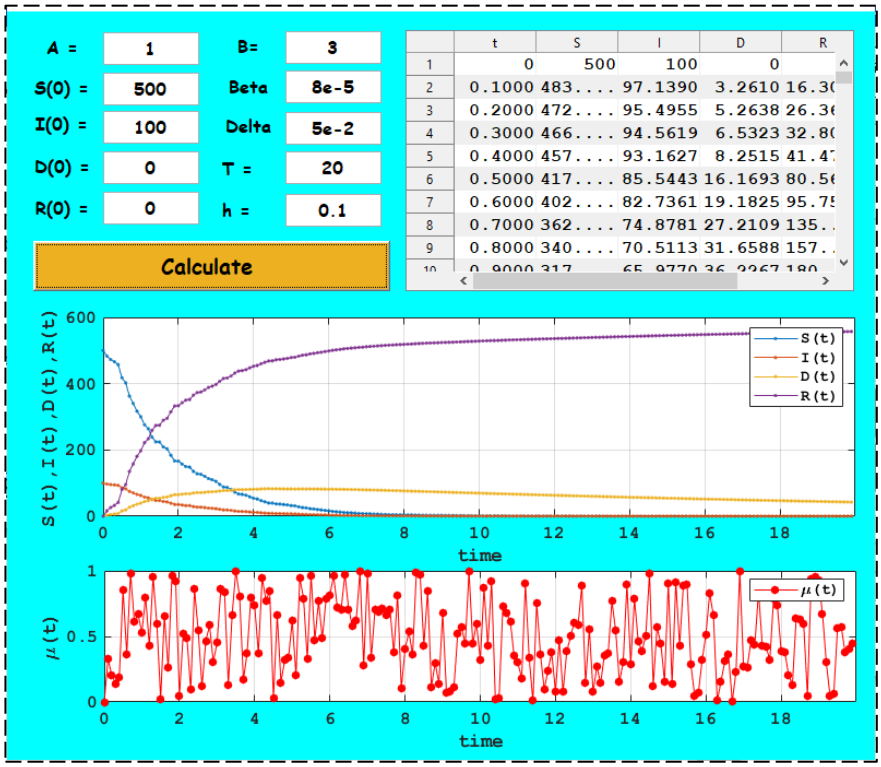

Figure 3: Software interface of the control model module based on the PSIDR model

Figure 4 shows the SM software interface for modeling the spread of malicious software.

The program interface and its logical part are implemented according to the principle of a modular structure in the MATLAB system of applied mathematics.

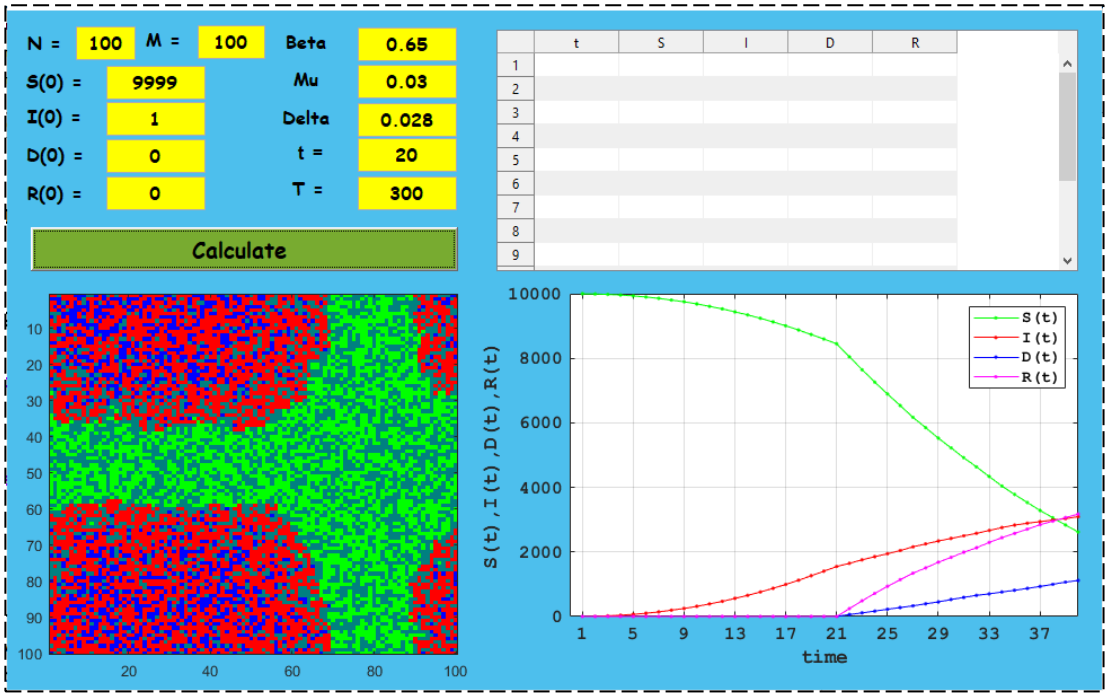

Figure 4: Program interface of the module of the stochastic model of forecasting the spread of malware

The developed software complex can be modified by finalizing the corresponding models of the spread of malicious software for special cases of malware that may occur when working with information systems.

3. Analysis of Practical Results

In the course of simulations using different models, it was determined that SMs are more adapted to modeling the processes of the spread of malware. The effectiveness of SM based on the CA is manifested in the fact that it is possible to simultaneously obtain statistics, which are present when using DM based on differential equations and their systems, and a picture of the spread of the process itself with a time scale. Also, the advantage of the SM based on the CA is the integer number of statistics, which is absent in the simulation of the propagation of malware using the DM based on differential equations and their systems. The disadvantage of SM is the relatively long working time compared to the time allocated for working out DM.

It should also be noted that any implemented models can be modified quite quickly in order to take into account all the factors that affect the processes of the spread of malware in general.

4. General Conclusions

The current rate of development of computer systems requires the maximum possible protection against malware. The entry of malware even into a local information system can lead to serious consequences, which can include the loss of system user data, data extraction from the system, data substitution in this system, even financial losses of a specific company to which the relevant system belongs.

In order to maximally reduce the risks of the formation of various unpleasant and / or catastrophic consequences from the impact of malware on one or another system, various types of modeling are carried out, for example, simulation, which will make it possible to assess the degree of risk and

financial losses that may occur in the company. A comparative analysis of modern methods, tools, and software for modeling the spread of malicious software was conducted. The MALTAB system of applied mathematics was chosen as the development environment for the corresponding software. The system is flexible for software development, the program code is short, logically designed, and the interface is developed quite quickly with the help of a suitable designer. The developed software complex for simulation of the spread of malware consists of 3 main software modules: the module for forecasting the spread of malware based on the PSIDR DM, which is presented in the form of a Cauchy problem for SODE; the module for forecasting the spread of malware on the basis of PSIDR SM, which is programmed on the basis of CA using Neumann and Moore circles; a control model implementation module for the PSIDR DM, which in turn uses population (multi-agent) stochastic methods and algorithms to minimize costs at each time step. All three software modules of the complex have been tested and verified on different input data (initial conditions for simulating the spread of malware in the system). A graphical user interface has been developed for each software module with the possibility of simulating processes for various initial conditions. It should be noted that DMs are less flexible when modeling similar tasks. Their disadvantage is manifested in the fact that the number of studied objects is not always a natural number, and in this case rounding is used according to classical mathematical rules. Also, the DM modification process is complicated compared to the SM modification process. SMs have another significant advantage, which is the transparency of the process. This gives a clearer picture of the process itself and makes modeling much more flexible not only for this kind of tasks.

5. References

[1] Alexeev, A., Henshel, D.S., Cains, M., Sun, Q.: On the malware propagation in heterogeneous networks. In: 2016 IEEE 12th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), pp. 1–5. IEEE (2016)

[2] Brauer, F., Castillo-Chavez, C. (2012). Epidemic Models. In: Mathematical Models in Population Biology and Epidemiology. Texts in Applied Mathematics, vol 40. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-1686-9_9.

[3] Hernández Guillén, J.D., Martín del Rey, Á., Hernández Encinas, L. (2018). New Approaches of Epidemic Models to Simulate Malware Propagation. In: Pérez García, H., Alfonso-Cendón, J., Sánchez González, L., Quintián, H., Corchado, E. (eds) International Joint Conference SOCO'17-CISIS'17-ICEUTE'17 León, Spain, September 6–8, 2017, Proceeding. SOCO ICEUTE CISIS 2017 2017 2017. Advances in Intelligent Systems and Computing, vol 649. Springer, Cham. https://doi.org/10.1007/978-3-319-67180-2_61.

[4] Kuniya, T. Structure of epidemic models: toward further applications in economics. JER 72, 581–607 (2021). https://doi.org/10.1007/s42973-021-00094-8

[5] Karyotis, V., Khouzani, M.: Malware Diffusion Models for Modern Complex Networks: Theory and Applications. Morgan Kaufmann, Amsterdam (2016)

[6] Liu, W., Liu, C., Liu, X., Cui, S., Huang, X.: Modeling the spread of malware with the influence of heterogeneous immunization. Appl. Math. Model. 40(4), 3141–3152 (2016)

[7] Galchynsky L, Pushko A. Modeling the cost estimation of preventive strategies to combat the spread of infectious diseases. A Young Scientist, 2018, vol 4(1). Pp. 120–124.

[8] Khaidurov V., Tsiupii T., Zhovnovach T. Modelling of Ultrasonic Testing and Diagnostics of Materials by Application of Inverse Problems. ITTAP'2021: 1nd International Workshop on Information Technologies: Theoretical and Applied Problems. ITTAP'2021: November 16–18, 2021. Pp. 1–5. CEUR Workshop, 2021, 3039. http://ceur-ws.org/Vol-3039/short25.pdf. ISSN: 1613-0073.

[9] Khaidurov V., Zaporozhets A., Tsiupii T. Creation of High-Speed Methods for Solving Mathematical Models of Inverse Problems of Heat Power Engineering. Springer, Systems Decision and Control in Energy III, vol. 399, 41–74 (2022). https://doi.org/10.1007/978-3-030-87675-3. ISSN: 2198-4182. ISSN: 2198-4182.