Egyptian Informatics Journal 24 (2023) 61–69

Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

A computationally efficient method for assessing the impact of an active viral cyber threat on a high-availability cluster

Ahmed Altameem [a], Mohammed Al-Ma'aitah [a], Viacheslav Kovtun [b,*], Torki Altameem [a]

[a] Computer Science Department, CC, King Saud University, 11451, 95, Riyadh 11362, Saudi Arabia
[b] Computer Control Systems Department, IITA, Vinnytsia National Technical University, Khmelnitske Shose, 95, Vinnytsia 21000, Ukraine

ARTICLE INFO

Article history:
Received 20 July 2022
Revised 14 October 2022
Accepted 30 November 2022
Available online 10 December 2022

Keywords:
Information security
Viral cyber threat
High availability cluster
Markov model
Quantitative metric
Computational efficiency

ABSTRACT

The field of computer science, like its sub-field of cyber threat modelling, is rapidly evolving. The prerequisites for key changes can be summarized as follows: cyber threats are evolving; there are leaks of special services tools; agile development methodology is being introduced everywhere; the boundaries of the object of protection are blurred; the scope of application of artificial intelligence is expanding; potentially vulnerable API integrations are increasingly being used. These factors lead to the fact that the processes of analysis of cyber threats, analysis of protective measures, generalization of data, and development of protective tools should now be considered continuous, not discrete. At the same time, the cost of cybersecurity increases like an avalanche in an attempt to avoid reputational and information losses. The only way to avoid this tendency is to apply a rational, scientific, accurate method of cognition to these processes. Thus, the creation of mathematical models of processes in the field of cybersecurity is now more relevant than ever. The article is devoted to the investigation of the process of the influence of an active viral cyber threat on a high-availability cluster in the paradigm of the provisions of the theory of Markov processes, graph theory and the theory of mathematical analysis. The main contribution of the research is a formalized computationally efficient method of approximate estimation of the average number of affected elements of the target high-availability cluster under the influence of an active viral cyber threat. Also, a criterion that allows estimating the trend of the quantitative parameter of the metric of the model of the studied process at medium and long time intervals is proposed. To obtain the declared scientific result, the authors: - formulated a Markov model of the process of the influence of an active viral cyber threat on a high availability cluster; - substantiated a compact metric for accurate assessment of the average number of cluster elements affected by an active viral cyber threat at any time; - formulated a computationally efficient method of approximate estimation of the parameter of the mentioned metric for the model of the target studied process; - proposed a criterion that allows researchers to evaluate the trend of the parameter of the mentioned metric for the model of the target researched process at medium and long intervals of time. The adequacy of the formulated method has been proven empirically.

1. Introduction

Modern machine learning methods already make it possible to create evolving viral cyber threats that are invulnerable to typical protective mechanisms of information systems. Undoubtedly, soon the toolkit of attackers will allow the automatic creation of new viral cyber threats, hardened by pseudo-evolutionary selection to the specified properties of the target information environment. Such viral cyber threats will have developed heuristic properties and swarm organization. Creating models of the spread of such cyber threats to investigate their behavior and develop countermeasures is an urgent scientific and applied task.

The classical and still relevant conceptual basis for describing the dynamic process of the spread of viral cyber threats is the *SI* and *SIR*

* Corresponding author.
E-mail addresses: ahaltameem@ksu.edu.sa (A. Altameem), malmaitah@ksu.edu.sa (M. Al-Ma'aitah), Kovtun_v_v@vntu.edu.ua (V. Kovtun), altameem@ksu.edu.sa (T. Altameem).

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.eij.2022.11.002

models [1–5]. When creating the *SI* model, it is assumed that an arbitrary computer in the target finite computer network can be in one of two states – vulnerable (*S*) and infected (*I*). A virus cyberthreat propagates through a network from infected to randomly selected vulnerable computers at a constant average rate. In the *SIR* model, an invulnerable state (*R*) is added to the two states already mentioned in the description of the *SI* model, in which an arbitrary computer of the network can be, to which the target computer can go only from the state I (overcoming a cyber infection). Accordingly, the SIR model additionally takes into account the rate of immunization of computers (nodes) in the target network. In the SIR model (to the already mentioned two states in the description of the *SI* model, in which an arbitrary computer of the network can be), an invulnerable state (*R*) is added, to which the target computer can transit only from the state *I* (overcoming a cyber infection). Accordingly, the *SIR* model additionally takes into account the rate of immunization of computers (nodes) in the target network.

Classically, *SI* and *SIR* models are formalized based on kinetic differential equations, but there are also original approaches, for example, in works [6–9], the differential model of the hydrodynamic process became the basis. The usual "differential" approach is based on the assumption that the number of infected nodes in a computer network is a continuous function of time. A typical *SI*- or *SIR*-like model, built based on differential equations, is a system of equations of this type (mostly nonlinear differential equations of the first order), where each equation characterizes a certain class of node of the studied computer network and describes the permissible transitions between states and material balance (a set of controlled variables and free members). The coefficients for controlled variables in such equations characterize the settings of protective mechanisms. The solution of the corresponding system of differential equations is usually interpreted in the context of determining recommendations for the intensity of renewal of protective mechanisms.

Without changing conceptually, *SI* and *SIR* models are developing methodologically [10–12]. For example, in the work [10], the concept of a computer network node is revealed as a set of two elements ("server" and "client"), and the speed of the spread of cyber infection is considered to be different. In this model, each computer is also characterized by the probability of re-infection, which, according to the authors, allows for taking into account the polymorphic nature of modern viral cyber threats. This direction of development of *SI* and *SIR* models continues. For example, in works [11,12], for computer network nodes of the "server" class, the set of states includes vulnerable (*S*), infected (*I*), and immune (*A*) states, and for computer network nodes of the "client" class, the set of states includes susceptible (*S*), infected (*I*), non-susceptible (*R*) and immune (*A*) states. The mentioned works differ in their interpretation of the concept of "immune state" as, for example, "temporarily invulnerable", "highly invulnerable", etc.

The expansion of the nomenclature of classes of nodes and states is accompanied by a symmetrical expansion of the set of characteristic parameters that allow describing the corresponding models. For example, in well-known models [13–16] parameters such as the epidemiological threshold, waiting time for infection, replication coefficient, probabilities of infection and immunization, node invulnerability time, etc. are entered. However, the typical statement of the research problem in such models is limited to the search for equilibrium points of the system of differential equations and the analysis of the asymptotic behavior of the solutions found, which are associated with the corresponding modes of cyber infection progress.

1.1. A critical review of current models of the development of cyber infections

There are known attempts to take into account the heterogeneous architecture of modern computer networks within the conceptual limits of *SI* and *SIR* models (for example, the *NSIDR* model [17–19]). These architectural properties are taken into account by multiplying the terms of certain differential equations in the model (system of equations) by empirically determined constant coefficients. This trend also includes models aimed at describing the mechanism of controlling the number of inter-computer connections in the target network in conditions of an active viral cyber threat [20–22]. However, the mathematical apparatus of differential equations do not provide researchers with sufficient freedom for an adequate description of such nuances. It was this circumstance that led to the appearance of a wide range of *SI* and *SIR* models implemented in the theoretical basis of graph theory [12,20,23]. Undoubtedly, such models perfectly reproduce the topology of the target computer networks but lose to "differential" models in the nomenclature of effects that can be simulated. However, the situation changes when combining the capabilities of graph theory with machine learning methods [24–26]. For example, in works [25,26], a computer network is modelled by a probability graph, the vertices of which are described by variables that regulate the probabilities of the states of network nodes, and the edges determine the interaction between the variables of the model. The influence of a viral cyber threat in such a model is described as a cellular automaton, in the form of a finite set of rules.

There are also known works where the progress of the influence of a viral cyber threat is described by the method of comparisons [12,27,28]. For example, the analysis of such influence is carried out simultaneously by the methods of autoregression and Fourier analysis. These methods make it possible to predict the trend of the progress of the cyber infection, and the use of different bases for creating a regression model is aimed at increasing the accuracy of the forecast (due to the increase in the computational complexity of the analysis).

If a computer network focused on a critical use is investigated, then models that take into account the mechanism of the topology of blocking nodes of a conglomerate of computer networks in the event of the registration of a viral cyberthreat become relevant. A representative of the models that describe the process of functioning of such a mechanism is a Cayley tree structure with a random number of connections [14,15]. This mathematical apparatus allows researchers to calculate the probability of infection of specific nodes depending on their distance from the source of cyber infection in the topology of a heterogeneous network, taking into account its scale. Such opportunities are provided by the authors of the works [20,23,29] using the large-scale graph technology, which allows for taking into account the processes of hierarchical growth of the network structure. A specific parameter that should be determined when creating such models is the percolation threshold, which is understood as the proportion of blocked nodes at which the target computer network as a whole is unable to implement its functional purpose. When defining this parameter (functional purpose), the classes to which blocked nodes or network sectors belong should be taken into account.

Combining the possibility of taking into account the topology and characteristics of the investigated computer network with a mathematically justified and relatively uncomplicated background allows the mathematical apparatus of Markov chains [11,12,20,30,31]. There are well-known models for describing the impact of viral cyber threats based on stochastic models of interactive Markov chains, in which the state of computer network nodes in each subsequent period does not depend (depends) on the state of a specific node and nodes connected to it in the current period. Naturally, we are talking about Markov and semi-Markov chains, which can be studied both in discrete (most often) and continuous time. It is no secret that the assumption of the continuity of the process of the spread of a viral cyberthreat over time, which is the cornerstone of the models of the development of cyber infection formalized on a differential basis, does not correspond to real-

ity (especially in the short periods). This fact is taken into account in the paradigm of discrete Markov chains "by default". However, Markov models of the process of the spread of a viral cyberthreat have a characteristic drawback, which consists of the rapid nonlinear growth of the computational complexity of the process of calculating the required parameters with the growth of the scale of the investigated computer network. This circumstance determines the search for a method of simplifying this computational process while maintaining the adequacy of the output result.

Taking into account the strengths and weaknesses of the mentioned analogues, we will formulate the necessary attributes of scientific research.

The ***object*** of the research is the process of the influence of an active viral cyber threat on a high-availability cluster.

The ***subject*** of the research is the provisions of the theory of Markov processes, the theory of graphs, the theory of mathematical analysis, and the theory of probability and mathematical statistics.

The ***aim*** of the research is to formalize a computationally efficient method for estimating the state of the model of the studied process at medium and long intervals of time.

The ***objectives*** of the research are:

- strict mathematical formalization of the research object model;
- determination of quantitative metrics for evaluating an arbitrary instance of the research object in the parametric space of the created mathematical model;
- formalization of a computationally efficient method of calculating the selected metric for the model of an arbitrary instance of the research object;
- formalization of the criterion for approximate assessment of the state of the model of the research object at medium and long intervals of time.

The ***main contribution*** of the research is a formalized computationally efficient method of approximate estimation of the average number of affected elements of the target high-availability cluster under the influence of an active viral cyber threat. Also, a criterion that allows estimating the trend of the quantitative parameter of the metric of the model of the studied process at medium and long time intervals is proposed.

The ***highlights*** of the research are:

- a Markov model of the process of the influence of an active viral cyber threat on a high-availability cluster (expressions (1), (2));
- a metric for accurate estimation of the average number of cluster elements affected by an active viral cyberthreat at an arbitrary moment in time (expression (4));
- a computationally efficient method of approximate estimation of the parameter of the mentioned metric for the model of the target object of research (expressions (9), (10));
- a criterion that allows us to evaluate the trend of the parameter of the mentioned metric for the model of the target researched process at medium and long intervals of time (expression (13)).

2. Models and methods

2.1. Setting up the research

Consider a connected directed graph with n vertices as an abstract model of a high-availability cluster architecture. The edges of the graph represent information communication channels between elements (computer servers) of the cluster. The vertices of the graph represent cluster elements that are potential targets of a viral cyber threat. At an arbitrary discrete moment $t \geqslant 0$, each element of the cluster can be either in a potentially vulnerable state to a viral cyberthreat v, or in a state affected by a viral cyberthreat a. The state of all cluster elements at an arbitrary moment in time is defined as $v(t) + a(t) = n$.

Let's assume that during a time cycle, an arbitrary element of the cluster can freely transition into one of two permissible states: $v \to a, a \to v$. The quantitative phenomenon of such a transition can be characterized by such parameters as:

- the probability of transmission of a viral cyberthreat from an affected element to a potentially vulnerable one: p_{a+};
- the probability of transition of the affected element to a potentially vulnerable state $(a \to v)$: p_{av};
- the probability of a potentially vulnerable element transitioning into an affected state $(v \to a)$: p_{va};
- the connectivity of the cluster graph. The parameter c relates the value of the number of vertices n in the studied graph to its average degree $\bar{k}$ by the ratio $c = \bar{k}/n$ (as $c = 0$ there is no informational communication between the cluster elements, and $c = 1$ the cluster graph is fully connected).

Note that the parameter p_{va} depends on the values of the parameters a, p_{av}, c: $p_{va} = f(a, p_{av}, c)$. It is important to investigate the nature of this functional dependence.

For the sake of clarity, let's generalize the parametric description of the state in which an arbitrary element of the cluster is with the UML state diagram presented in Fig. 1.

We will analytically describe the process of the influence of a viral cyberthreat on the elements of a high-availability cluster by a discrete Markov chain ξ, the states of which are determined by the number of affected elements of the cluster a, that is, the total number of states is equal to $n + 1$: $a = \overline{\{0, n\}}$.

The dynamics of the chain $\xi(t)$ is determined by the transition probabilities $p_{a,a'} \in P$. Based on the provisions of the theory of Markov processes [11], we formulate expressions for calculating parameters $p_{a,a'}$ in the form of expression

$$p_{a,a'} = \sum_{i=\max\{0,a-a'\}}^{\min\{a,n-a'\}} K_a^i p_{av}^i (1-p_{av})^{a-k} \times \times K_{n-a}^{a'-a+i} p_{va}^{a'-a+i} (1-p_{va})^{n-i-a'}, \quad (1)$$

where $K_a^i \equiv \frac{a!}{i!(a-i)!}$ is a binomial coefficient whose value is equal to the number of combinations from a to i.

Based on expression (1), we formulate the expression for calculating the probability p_{va}:

$$p_{va} = \sum_{i=0}^{a} K_a^i \left(1 - (1-p_{a+})^i\right) c^i (1-c)^{a-i} \quad (2)$$

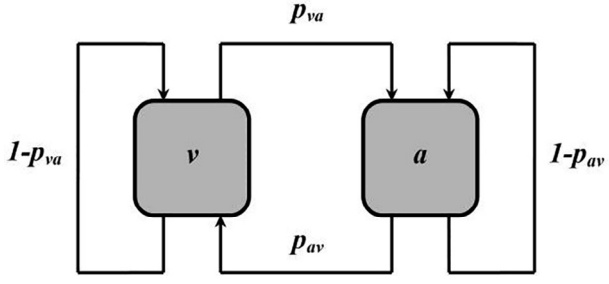

Fig. 1. UML state diagram of an arbitrary element of a high-availability cluster under the influence of a viral cyber threat.

If we expand the expression (2) into a Taylor series and discard all terms higher than the first order, we get the expression $p_{va} = p_{a+}ca$. This linear expression is an analogue of the classic model of the progress of epidemics of viral cyber threats, proposed by Kefard and White [32]. This circumstance indirectly confirms the adequacy of the theoretical postulates made.

A prerequisite for the quantitative characterization of the development of the researched process is the determination of all elements of the matrix of transition probabilities $P = (p_{a,a'}), a, a' = \overline{1,n}$, which are calculated according to expression (1). In particular, considering the defined matrix P, the probability of a state a at a time t can be calculated using the expression

$$p_a(t) = \sum_{a'=0}^{n} p_{a'}(0) p_{a,a'}^{(t)} \tag{3}$$

where $p_a(0)$ is the probability of states of chain ξ at the initial moment $t = 0$:$\xi(0)$; $p_{a,a'}^{(t)}$ is an element of the matrix P^t, which is the matrix P raised to the t th degree.

With a known number of affected elements of the cluster at the moment $t = 0$ (parameter $a_0 = a(t)|_{t=0} = a(0)$), the probability $p_a(0)$ is defined as

$$p_a(0) = \begin{cases} 1 \forall a = a_0, \\ 0 \forall a \neq a_0. \end{cases}$$

In turn, with known probabilities (3), we can estimate the average number of affected cluster elements at a time t using the expression

$$A(t) = \sum_{a=0}^{n} a p_a(t) \tag{4}$$

Parameter (4) is a basic metric that will characterize the development of the research process. The following theoretical material will be formulated around the parameter $A(t)$.

2.2. Approximate estimation of the average number of affected elements of a high-availability cluster under the influence of a viral cyberthreat

Metric (4) is the most compact representative characteristic of the development of the studied process. However, to directly calculate the parameter $A(t)$ using expression (4), it is necessary to raise the transition probability matrix P to a degree t. The computational complexity of this operation rapidly non-linearly increases with an increase in the value of t (expression (3)) and an increase in the number of elements in the cluster n (the dimension of the matrix P is n × n, and the calculation of each element of this matrix according to expression (1) is accompanied by the calculation of the coefficient K_a^i, where $a(n)$, $i(n)$). All these circumstances encourage the search for a computationally efficient concept for calculating the approximate value of the parameter $A(t)$.

Let's start by writing out the multiplier $\left(1 - (1 - p_{a+})^i\right)$ in expression (2):

$$p_{va} = \sum_{i=0}^{a} K_a^i c^i (1-c)^{a-i} - \sum_{i=0}^{a} K_a^i c^i (1-c)^{a-i} (1 - p_{a+})^i \tag{5}$$

The minuend in expression (5) is the binomial expansion of the one. Let's expand the subtrahend from expression (5) using Newton's binomial formula:

$$\sum_{i=0}^{a} K_a^i c^i (1-c)^{a-i} (1-p_{a+})^i = \\ = \left(c(1-p_{a+}) + 1 - c\right)^a = (1 - cp_{a+})^a.$$

Let's rewrite expression (5) taking into account what was just obtained:

$$p_{va} = 1 - (1 - cp_{a+})^a \tag{6}$$

It follows from expression (6) that the parameter p_{va} functionally depends on the combination of parameters c and p_{a+}:$p_{va} = f(cp_{a+})$. Since the parameters c and p_{a+} enter the expression (1) only transitively, through the parameter p_{va}, we can narrow the set of input parameters of the model to three elements: $\{\eta = cp_{a+}, p_{av}, a_0\}$.

It should also be noted that $\lim_{a\to\infty} p_{va} = \lim_{a\to\infty} \left(1 - (1-\eta)^a\right) = 1$ that is, the value of the parameter p_{va} increases monotonically with the increase in the value of the parameter a (the number of affected cluster elements). However, $a \leqslant n$ therefore $\max(p_{va}) = 1 - (1-\eta)^a$.

Finally, if $0 < \eta < 1$ we expand expression (6) into a Taylor series, we get:

$$p_{va} = a\eta - \frac{a(a-1)}{2!}\eta^2 + \frac{a(a-1)(a-2)}{3!}\eta^3 + \ldots \tag{7}$$

The linear approximation of this polynomial will be the expression $p_{va} = a\eta = acp_{a+}$, which is identical to the one already mentioned in section 2.1, proven adequate in the Kefard and White model. This circumstance analytically confirms that the mathematical transformations of expression (2) we made in section 2.2 did not lead to the original model losing its adequacy.

Let's move on to the formalization of the recurrent expression for the approximate estimation of the parameter (4). Let the average number of affected elements in the cluster be $\tilde{A}(t)$ at an arbitrary moment in time $t \geqslant 0$. Then the probability that at the moment $t+1$ a potentially vulnerable element will transit to the affected state will be determined by the expression

$$p_{va} = 1 - (1-\eta)^{\tilde{A}(t)} \tag{8}$$

Based on expression (8), we can argue that at the time $t+1$, on average,

$$p_{va}(1 - \tilde{A}(t)) = (n - \tilde{A}(t))\left(1 - (1-\eta)^{\tilde{A}(t)}\right)$$

elements of the cluster that are potentially vulnerable at a time t will transit to the affected state. At the same time, at moment $t+1$, $p_{av}\tilde{A}(t)$ elements of the cluster, which were in the affected state at the moment t, will transit to the potentially vulnerable state. Summarizing what has been said, we formulate an expression for estimating the average number of affected elements of the cluster at the moment $t+1$:

$$\tilde{A}(t+1) = (1 - p_{av})\tilde{A}(t) + \\ + (n - \tilde{A}(t))\left(1 - (1-\eta)^{\tilde{A}(t)}\right). \tag{9}$$

Naturally, when applying expression (9), it should be taken into account that at the time $t = 0$ there are a_0 elements of the cluster in the affected state:

$$\tilde{A}(0) = a_0 \tag{10}$$

It can be expected that the function $\tilde{A}(t)$ will qualitatively approximate the etalon function $A(t)$ for a fairly wide range of values of the controlled parameters $\{\eta, p_{av}, a_0\}$. Moreover, the approximation error will decrease as the values of the parameters a and t increase. It is rather difficult to strictly analytically estimate the error of approximation of the function $A(t)$ by the recurrent sequence (10)-(9). However, the analytical form of the function (9) allows us to predict that as the number of elements of the cluster n increases, the deviation of the values of the function $\tilde{A}(t)$ from the corresponding values of the function $\tilde{A}(t)$ will decrease.

At the same time, with the fixed values of the parameters a_0 and n, for certain values of the parameters η and p_{av}, the limit expression $\lim_{t\to\infty} A(t) = \underset{\infty}{A} = 0$ will be fulfilled. Judging by the analytical form of function (9), for the same range of values of the tuple $\{a_0, n, \eta, p_{av}\}$, the limit value of the function $A^{\sim}(t)$ will be different from zero: $\lim_{t\to\infty} A^{\sim}(t) = \underset{\infty}{\tilde{A}} \neq 0$. The contribution of this collision to the resulting error of approximation of the function $A(t)$ by the function $A^{\sim}(t)$ should be investigated by numerical methods. Finally, based on analytical expressions (4) and (9), it can be asserted that the functions $A(t)$ and $A^{\sim}(t)$ react differently to the initiating value (parameter a_0). It can be expected that the function $A(t)$ is more sensitive to the value of the parameter a_0 than the function $A^{\sim}(t)$. It can also be predicted that this discrepancy will quickly level off with an increase in the value of the parameter a. However, this statement also needs experimental verification.

Before proceeding to the experimental verification of the hypotheses made above, let us pay attention to the analytical form of the limit expression for the function $A^{\sim}(t)$, that is $\lim_{t\to\infty} A^{\sim}(t) = \underset{\infty}{\tilde{A}}$. Let's rewrite expression (9) in the context of applying the notation $\underset{\infty}{\tilde{A}}$:

$$\underset{\infty}{\tilde{A}} = (1 - p_{av})\underset{\infty}{\tilde{A}} + \left(n - \underset{\infty}{\tilde{A}}\right)\left(1 - (1-\eta)^{\underset{\infty}{\tilde{A}}}\right) \quad (11)$$

Equation (11) does not depend on the initial value (10). This fact allows us to state that the difference between the limit values of $\underset{\infty}{A}$ and $\underset{\infty}{A^{\sim}}$ is not affected by the parameter a_0. For the convenience of further analytical operations, let's simplify expression (11) to the form

$$xp_{av} = (1-x)(1-\beta^x),\ p_{av} > 0 \quad (12)$$

where $x = \underset{\infty}{A^{\sim}}/n$ is the relative number of affected cluster elements at the limiting moment $t \to \infty$ and $\beta = (1-\eta)^n$ is an auxiliary designation.

Expression (12) is a transcendental equation concerning the variable $x \in [0,1]$, the analytical solution of which does not exist. However, the approximate solution of equation (12) can be obtained by numerical methods, including zero-order ones. To a large extent, this solution can be characterized in advance. In particular, the number x = 0 will always be among the roots of equation (12). Non-zero values $\underset{\infty}{A^{\sim}}$ will correspond only to roots belonging to the segment (0,1]. The absence of such roots will indicate that $\underset{\infty}{\tilde{A}} = 0 \forall a_0$.

Let us introduce an infinitely differentiable function $f(x) = \frac{(1-x)(1-\beta^x)}{p_{a+}}$ with a negative second derivative $f''(x) = \frac{\beta^x}{p_{av}} \ln\beta(2 - (1-x)\ln\beta)$ on the unit segment. The first derivative of the function $f(x)$ decreases monotonically on the unit segment. For the limit values, we write: $f'(0) = -\ln(\beta)/p_{av} > 0, f'(1) = (\beta - 1)/p_{av} < 0$. This means that there is a single point $\varphi \in (0,1)$ at which $f'(\varphi) = 0$, that is, the function $f(x)$ monotonically increases during the interval $[0, \varphi)$, and monotonically decreases during the interval $(\varphi, 1]$.

The presented results of the analysis of the function $f(x)$ make it possible to formulate a condition under which the solution of equation (12) (equivalent to the equation $x = f(x)$) will not contain roots on the segment $(0,1]$. This condition: if the tangent $y = -x\ln\beta/p_{av}$ to the function f(x) at the point x = 0 lies under the line y = x (only under such circumstances the graphs of the functions $y = f(x)$ and y = x will not intersect on the unit interval). On this basis, the criterion for the absence of nonzero roots of equation (12) on the unit interval looks like this:

$$-\ln\beta/p_{av} \leqslant 1 \equiv (1 - cp_{a+})^n \geqslant \exp(-p_{av}) \quad (13)$$

If criterion (13) is fulfilled for the investigated process of the influence of an active viral cyber threat on the target cluster of high availability, then over time ($t \to \infty$) the protective mechanisms of the information system will overcome the negative impact on their own. If criterion (13) is not fulfilled (threshold value $T(p_{av}) = -n\ln(1 - cp_{a+})$), then it will not be possible to overcome the negative influence without external intervention. It should be noted that the expression for calculating the threshold value $T(p_{av})$ was derived based on the investigation of the limit approximate value of the parameter A(t), i.e. $\underset{\infty}{A^{\sim}}$. This means that the value of the parameter $T(p_{av})$ is an approximate optimistic estimate of the exact threshold value. However, this circumstance does not negate the fact that as the value t increases, criterion (13) turns from a law rule.

Therefore, based on the Markov model of the process of the influence of a viral cyberthreat on a target high-availability cluster described in Section 2.1, a computationally efficient method of approximate estimation of the average number of affected cluster elements under the influence of a viral cyber threat is formulated in Section 2.2, and a criterion is defined that allows investigating the trend of the progress of the researched process on medium and long periods.

3. Experiments

To present the functionality of the proposed method of approximate estimation of the average number of affected elements of the target cluster of high availability under the influence of an active viral cyber threat, we will apply the capabilities of simulation modelling.

The specialized MathWorks MATLAB software package was chosen to implement the simulation model. The justification for such a choice is that the functions of the toolboxes of this software platform have been tested worldwide and their adequacy has been empirically proven. Using mostly the Hidden Markov Model Toolbox, we created the following custom functions:

- MyPoission(a,λ) is a function for modelling a stochastic value with a Poisson distribution, where a is the number of events, λ is their intensity;
- *MyState* (X, Λ, M, R, a) is a function for implementing the transition of a discrete Markov chain to the next state from the current X, which is characterized by a tuple of sets $\langle \Lambda, M, R \rangle$, where $\Lambda = \left\{\lambda_1, \lambda_n\right\}$ is a set of intensities of viral cyberthreat flows affecting the studied cluster, $\lambda_i \geq 0$; $M = \left\{\mu_1, \mu_n\right\}$ is the set of intensities of flows of cyber-immune reactions of cluster elements, $\mu_i \geq 0$; $R = \left\{r_1, r_n\right\}$ is the set of probabilities of neutralizing the effects of viral cyberthreat by protective mechanisms (returning cluster elements from an affected to a potentially vulnerable state), $0 \leq r_j \leq 1$;

MyDMC (Λ, M, R, a_0, t = 0) is a function for implementing a parameterized instance of a discrete Markov chain with an initiating effect a_0 and calculating for the current moment $t \geqslant 1$ the exact value of the quantitative metric $A(t)$ according to expression (4) and the approximate value of the quantitative metric $A^{\sim}(t)$ according to expressions (9), (10).

Let us focus on the description of the function *MyState*. For a discrete Markov chain ξ, the state X is determined by a pair of parameters (a, t), where $a \in \left\{0, n\right\}$ is the identifier of state X and t is the moment when the system transitions to state X. The func-

tion *MyState* implements the transition of the system from the current state X=(a,t) to the next state $X' = (a', t')$:$X \to X'$. The values of parameters a', t' are determined depending on the value of the parameter a. If:

- a = 0 (all elements of the cluster are in a potentially vulnerable state / a viral cyber threat to the cluster is not implemented), then the function *MyState* generates a set of stochastic values $T = \{\tau_i\}, i = \overline{1,n}$, using the function *MyPoisson*. Let $\tau_j = \min\left\{\tau_1, \overline{\tau_n}\right\}$ for $j \in \left\{\overline{1,n}\right\}$, then we take $a' = a + j$ and $t' = t + \tau_j$, that is,$X' = (j, t + \tau_j)$;
- - $a \in \left\{\overline{1,n}\right\}$ (the elements of the cluster are negatively affected in the form of a viral cyberthreat), then the function *Mystate* generates, using the function *MyPoisson*, a stochastic value τ distributed according to the Poisson distribution law with parameter λ_a. We accept $t' = t + \tau$. Using the standard MATLAB function *Rand*, we obtain a stochastic number $\times$ uniformly distributed over the interval $[0, 1]$. If the inequality $r_a > x$ holds, then we take a',=0 otherwise, we take a'=a + r;
- a = n (all elements of the cluster are in an affected state, that is, the protective mechanisms did not cope with the viral cyberthreat), then the function *MyState* accepts a'=n and t'=t. Accordingly,$X' = X = (a = n, t)$.

Functioning according to the algorithm described above, the function *Mystate* allows for the discrete Markov chain ξ to calculate the sequence of states of the form $X_0 = (a_0 = 0, t_0 = 0) \to X_1 = (a_1, t_1) \to \ldots \to X_n = (a = n, t_n)$. The initialization of the model (state X_0) is implemented by the function *MyDMC* (Λ, *M*, *R*, a_0, t = 0). Each subsequent iteration is implemented by calling the function and ends by calling the function, returning the output values of and. Transitions occur a predetermined number of times or until the model enters the state.

To conduct experiments, it was necessary to obtain data for a tuple of sets. Such information was kindly provided by the staff of the Situation Center of the Information Technology Department of Vinnytsia City Council (Vinnytsia, Ukraine) (hereinafter referred to as SC). The staff of the SC department supports the functioning of the distributed information and communication system (high availability cluster), which manages video surveillance and traffic lights on the roads of Vinnytsia. There are more than 1000 elements (servers, workstations, client computers) involved in information exchange and are potentially vulnerable to viral cyber threats in the architecture of the information and communication system of the SC.

The analysis of the SC system operation logs for the period from 09/01/2020 to 09/01/2021 (365 full days) in the context of detecting cases of implementation of viral cyber threats, summarized in the set, made it possible to determine the following input data for modelling:

$N_a = 3,\ \Lambda = (4.27, 3.96, 1.12)$

$M(0.91, 0.41, 0.94)$

$R = (0.09, 0.39, 0.0.37)$

where N_a is the number of recorded cases of implementation of viral cyber threats against the SC cluster for the censored period.

The obtained data made it possible to calculate the shown in Figs. 2 and 3 dependences of $A(a_0, n, \eta, p_{av})$ and $A^{\sim}(a_0, n, \eta, p_{av})$, respectively. At the same time, the author's software was launched with such initiating data as $a_0 = \{1, 5, 10, 25, 50, 60, 70\}$. In addition, such values of the established parameters of the model (1)-(2) were specified $n = 100, \eta = 0.001, p_{av} = 0.05$.

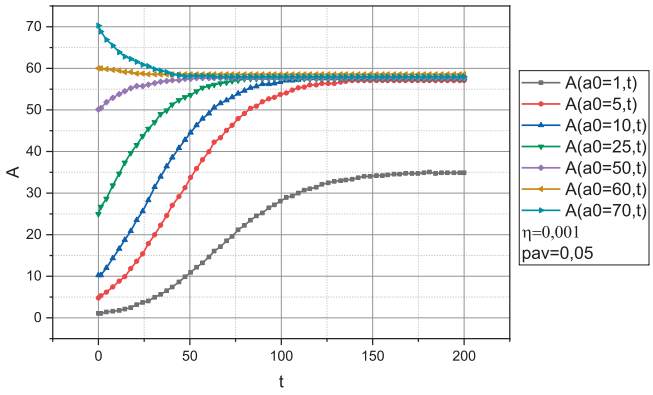

Fig. 2. Graphs of function $A(a_0, n, \eta, p_{av})$ at $a_0 = \{1, 5, 10, 25, 50, 60, 70\}, n, \eta, p_{av} =$ const.

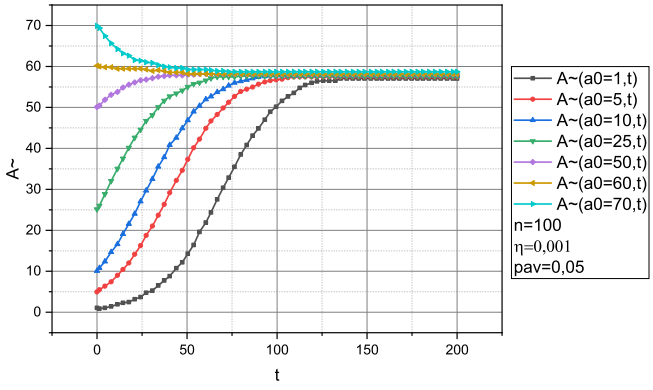

Fig. 3. Graphs of function $A^{\sim}(a_0, n, \eta, p_{av})$ at $a_0 = \{1, 5, 10, 25, 50, 60, 70\}, n, \eta, p_{av} =$ const.

We did not neglect the part of the material presented in section 2.2, devoted to the formalization of the trend evaluation of the quantitative parameter of the metric of the model of the studied process (that is, the process of the influence of an active viral cyberthreat on a high-availability cluster) over medium and long periods.

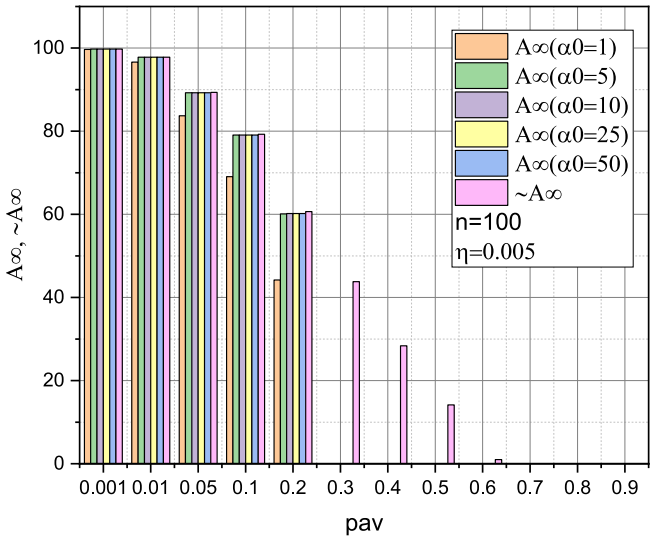

Fig. 4. Diagrams for functions $A(a_0, n, \eta, p_{av})$, $\underset{\infty}{\widetilde{A}}(a_0, n, \eta, p_{av})$ at $a_0 = \{1, 5, 10, 25, 50\}$, $p_{av} = \{0.01, 0.1, 0.2, \ldots, 0.9\}$, $n, \eta =$ const.

Table 1
Detailed presentation of information from Fig. 4.

p_{av}	$\underset{\infty}{A}$ $(a_0 = 1)$	$\underset{\infty}{A}$ $(a_0 = 5)$	$\underset{\infty}{A}$ $(a_0 = 10)$	$\underset{\infty}{A}$ $(a_0 = 25)$	$\underset{\infty}{A}$ $(a_0 = 50)$	$\underset{\infty}{\tilde{A}}$
0,001	99,658	99,779	99,779	99,779	99,779	99,7806
0,01	96,6112	97,8012	97,8012	97,8012	97,8012	97,8176
0,05	83,7142	89,2378	89,2379	89,2379	89,2379	89,3396
0,1	69,0677	79,0778	79,0808	79,0808	79,0808	79,268
0,2	44,2002	60,1127	60,2082	60,2084	60,2084	60,659
0,3	0	0	0	0	0	43,7863
0,4	0	0	0	0	0	28,3589
0,5	0	0	0	0	0	14,153
0,6	0	0	0	0	0	0,99239

For the corresponding experiment, we used already developed software, which was launched at such values of fixed parameters as $n = 100$,$\eta = 0.005$. With such a configuration of input parameter values, the threshold value $T(p_{av})$ derived from criterion (13) is equal to 0.5023. For the parameter p_{av} (probability of the transition of the affected element to a potentially vulnerable state) changing in the range $(0.001, 0.1, 0.2, \ldots, 0.9)$, we calculated the value of the parameter $\underset{\infty}{\tilde{A}}(p_{av})$ using equation (12). Further, with the help of algorithms standard for the theory of Markov processes, we calculated the valid limit values of the parameter $\underset{\infty}{A}(p_{av})$ for the same set of input data ($p_{av} = (0.001, 0.1, 0.2, \ldots, 0.9)$,$n = 100$, $\eta = 0.005$), additionally varying the value of the parameter $a_0 = \{1, 5, 10, 25, 50\}$ (the number of affected cluster elements at the time $t = 0$).

The calculated values of the functions $\underset{\infty}{A}(a_0, n, \eta, p_{av})$ and $\underset{\infty}{\tilde{A}}(a_0, n, \eta, p_{av})$ are visualized in the form of the corresponding diagrams in Fig. 4 and Table 1.

4. Discussion

Let's start the discussion of the scientific and experimental results presented in the article, focusing on Figs. 2 and 3. The graphs shown in Fig. 2 characterize the Markov model of type (1)-(2) of the information and communication system of SC in the metric (4), calculated by classical methods of the theory of Markov processes. In practice, this is implemented in the form of software based on globally proven Hidden Markov Model Toolbox algorithms of the specialized MathWorks MATLAB software package. These prerequisites allow us to consider those shown in Fig. 2 reference results. Accordingly, as visualized in Fig. 3 results characterize the new mathematical apparatus proposed by the authors in section 2.2, generalized by the recurrent sequence (10)-(9), for the approximate calculation of the metric (4) for the model (1)-(2) of the target researched process.

It can be visually verified that the function $A^{\sim}(t)$ qualitatively approximates the etalon function $A(t)$ for a fairly wide range of values of the controlled parameters $\{\eta, p_{av}, a_0\}$. Moreover, the deviation of the values of the function $A^{\sim}(t)$ from the values of the function $A^{\sim}(t)$ decreases with the increase in the value of the parameters a and t. It is noticeable that the functions $A(t)$ and $A^{\sim}(t)$ react differently to the initiating value (parameter a_0). The function $A(t)$ is more sensitive to the value of the parameter a_0 than the function $A^{\sim}(t)$. However, this discrepancy is quickly levelled off with an increase in the value of the parameter a.

In general, comparing presented in Fig. 3 (author's) and Fig. 2 (etalon) results, the mathematical apparatus summarized by expressions (9) and (10) can be recognized as adequate. The author's mathematical apparatus proved to be particularly effective for the calculation of the quantitative metric (9) (an approximate analogue of the metric (4)) of the studied process on medium and long-term time intervals.

Now let's focus on those presented in Fig. 4 results of the investigation of the limit values of metrics (4) and (9). It can be seen that for values of $p_{av} >> T(p_{av})$, the approximate value of $\underset{\infty}{\tilde{A}}(p_{av})$ qualitatively approximates the exact value of $\underset{\infty}{A}(p_{av})$. What is more, this tendency becomes more apparent with the increasing value of the initiating parameter a_0. This circumstance does not contradict the conclusions we made above regarding the comparison of the results shown in Figs. 2 and 3. A significant deviation of the values $\underset{\infty}{\tilde{A}}(p_{av})$ from the etalon values $\underset{\infty}{A}(p_{av})$ is manifested when the values of the parameter p_{av} approach the threshold $T(p_{av}) = 0.5023$. The value of $\underset{\infty}{A}(p_{av})$ is exhausted already at $p_{av} = 0.3$, while the author's mathematical apparatus shows a similar result only at $p_{av} \approx 0.55$. However, at values of the parameter p_{av} outside of the vicinity of the threshold value $T(p_{av})$, the author's and reference mathematical apparatus for estimating the limit value of the average number of affected cluster elements under the influence of a viral cyber-threat show the same results.

As expected, the results of the experiments showed that the approach to determining the threshold value $T(p_{av})$ based on criterion (13) is quite rough, because it is the result of "simplification raised to the absolute." Determination of the confidence interval for the parameter $T(p_{av})$ is a promising direction for further research.

In summarizing, we recall that the linear approximation of the polynomial (7) is the expression $p_{va} = a\eta = acp_{a+}$, which is identical to the adequate, tested model of Kefard and White. This circumstance analytically confirms that the mathematical transformations of expression (2) made by us in section 2.2 did not lead to the loss of adequacy of the original model presented in section 2.1.

5. Conclusions

The field of computer science, like its sub-field of cyber threat modelling, is rapidly evolving. The prerequisites for key changes can be summarized as follows: cyber threats are evolving; there are leaks of special services tools; agile development methodology is being introduced everywhere; the boundaries of the object of protection are blurred; the scope of application of artificial intelligence is expanding; potentially vulnerable API integrations are increasingly being used. These factors lead to the fact that the processes of analysis of cyber threats, analysis of protective measures, generalization of data, and development of protective tools should now be considered continuous, not discrete. At the same time, the cost of cybersecurity increases like an avalanche in an attempt to avoid reputational and information losses. The only way to avoid this tendency is to apply a rational, scientific, accurate method of

cognition to these processes. Thus, the creation of mathematical models of processes in the field of cybersecurity is now more relevant than ever.

The article is devoted to the investigation of the process of the influence of an active viral cyber threat on a high-availability cluster in the paradigm of the provisions of the theory of Markov processes, graph theory and the theory of mathematical analysis. The main contribution of the research is a formalized computationally efficient method of approximate estimation of the average number of affected elements of the target high-availability cluster under the influence of an active viral cyber threat. Also, a criterion that allows estimating the trend of the quantitative parameter of the metric of the model of the studied process at medium and long time intervals is proposed. To obtain the declared scientific result, the authors: - formulated a Markov model of the process of the influence of an active viral cyber threat on a high availability cluster; - substantiated a compact metric for accurate assessment of the average number of cluster elements affected by an active viral cyber threat at any time; - formulated a computationally efficient method of approximate estimation of the parameter of the mentioned metric for the model of the target studied process; - proposed a criterion that allows researchers to evaluate the trend of the parameter of the mentioned metric for the model of the target researched process at medium and long intervals of time.

The adequacy of the formulated method has been proven empirically. visualized in Fig. 3 results characterize the new mathematical apparatus proposed by the authors in section 2.2, generalized by the recurrent sequence (10)-(9), for the approximate calculation of the metric (4) for the model (1)-(2) of the target researched process. It can be visually verified that the function qualitatively approximates the etalon function for a fairly wide range of values of the controlled parameters. Moreover, the deviation of the values of the function from the values of the function decreases with the increase in the value of the parameters and. It is noticeable that the functions and react differently to the initiating value (parameter). It can be seen in Fig. 4 that for values of, the approximate value of qualitatively approximates the exact value of. What is more, this tendency becomes more apparent with the increasing value of the initiating parameter. This circumstance does not contradict the conclusions we made above regarding the comparison of the results shown in Figs. 2 and 3. A.

As it turned out, the simplification of the classical approach proposed by the authors to the calculation of the average number of cluster elements affected by an active viral cyber threat is accompanied by a noticeable error when determining this characteristic parameter at short time intervals. Also, the results of the experiments showed that the author's approach to determining the threshold value based on criterion (13) is quite rough. ***Further research*** is planned to address these limitations.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is funded by Researchers Supporting Project number (RSP-2022R503), King Saud University, Riyadh, Saudi Arabia.

Funding

RSP-2022R503.

Institutional review board statement

Not applicable.

Informed consent statement

Informed consent was obtained from all subjects involved in the study.

Data availability statement

Most data is contained within the article. All the data are available on request due to restrictions e.g. privacy or ethics.

References

[1] Maloth S, Vanmathi C. Attacks on Cyber Physical System: Comprehensive Review and Challenges. Int J Wireless Microwave Technol (IJWMT) 2022;12(5):53–73. doi: https://doi.org/10.5815/ijwmt.2022.05.06.
[2] Aliyev AG. Technologies Ensuring the Sustainability of Information Security of the Formation of the Digital Economy and their Perspective Development Directions. Int J Inf Eng Electron Business (IJIEEB) 2022;14(5):1–14. doi: https://doi.org/10.5815/ijieeb.2022.05.01.
[3] V. Kovtun, I. Izonin, and M. Gregus, "Reliability model of the security subsystem countering to the impact of typed cyber-physical attacks," Scientific Reports, vol. 12, no. 1. Springer Science and Business Media LLC, Jul. 27, 2022. doi: 10.1038/s41598-022-17254-4..
[4] Şimşek A. "Lexical sorting centrality to distinguish spreading abilities of nodes in complex networks under the Susceptible-Infectious-Recovered (SIR) model", Journal of King Saud University - Computer and Information Sciences. Elsevier BV 2022;34(8):4810–20.
[5] Abhishek V, Srivastava V. SIR Epidemic Model under Mobility on Multi-layer Networks. IFAC-PapersOnLine 2020;53(5):803–6.
[6] O. Bisikalo, O. Danylchuk, V. Kovtun, O. Kovtun, O. Nikitenko, and V. Vysotska, "Modeling of Operation of Information System for Critical Use in the Conditions of Influence of a Complex Certain Negative Factor," International Journal of Control, Automation and Systems, vol. 20, no. 6. Springer Science and Business Media LLC, pp. 1904–1913, Apr. 29, 2022. doi: 10.1007/s12555-021-0368-6..
[7] Blavatska V, Holovatch Yu. Spreading processes in "post-epidemic" environments. II. Safety patterns on scale-free networks. Phys A: Stat Mech Appl 2022;591:126799.
[8] Huo H-F, Yang P, Xiang H. Dynamics for an SIRS epidemic model with infection age and relapse on a scale-free network. J Franklin Inst 2019;356(13):7411–43.
[9] Kovtun V, Izonin I, Gregus M. Model of Information System Communication in Aggressive Cyberspace: Reliability, Functional Safety, Economics. IEEE Access 2022;10:31494–502.
[10] Long Y, Zhang Yu, Wu M, Peng S, Kwoh CK, Luo J, et al. Heterogeneous graph attention networks for drug virus association prediction. Methods 2022;198:11–8.
[11] Hu Z, Odarchenko R, Gnatyuk S, Zaliskyi M, Chaplits A, Bondar S, et al. Statistical Techniques for Detecting Cyberattacks on Computer Networks Based on an Analysis of Abnormal Traffic Behavior. IJCNIS 2021;12(6):1–13.
[12] Shahid N, Aziz-ur Rehman M, Khalid A, Fatima U, Sumbal Shaikh T, Ahmed N, et al. Mathematical analysis and numerical investigation of advection-reaction-diffusion computer virus model. Results Phys 2021;26:104294.
[13] Hu Z, Khokhlachova Y, Sydorenk V, Opirskyy I. Method for Optimization of Information Security Systems Behavior under Conditions of Influences. IJISA 2017;9(12):46–58.
[14] Kekül H, Ergen B, Arslan H. Comparison and Analysis of Software Vulnerability Databases. Int J Eng Manuf (IJEM) 2022;12(4):1–14. doi: https://doi.org/10.5815/ijem.2022.04.01.
[15] Kovalskyi B, Dubnevych M, Holubnyk T, Pysanchyn N, Havrysh B. Development of a technology for eliminating color rendering imperfections in digital photographic images. EEJET 2019;1(2):40–7.
[16] Zimba A. A Bayesian Attack-Network Modeling Approach to Mitigating Malware-Based Banking Cyberattacks. IJCNIS 2021;14(1):25–39.
[17] Long L, Zhong K, Wang W. Malicious viruses spreading on complex networks with heterogeneous recovery rate. Phys A: Stat Mech Appl 2018;509:746–53.
[18] A. Doroshenko, K. Obelovska and O. Bilyk, "Risk Analysis of Personal Data Loss in Wireless Sensor Networks," CITRisk'2021: 2nd International Workshop on Computational & Information Technologies for Risk-Informed Systems, Kherson, Ukraine, September 16-17, 2021. http://ceur-ws.org/Vol-3101/Paper19.pdf.
[19] Machado MR, Pantano S. Fighting viruses with computers, right now. Curr Opin Virol 2021;48:91–9.
[20] Pan W, Jin Z. Edge-based modeling of computer virus contagion on a tripartite graph. Appl Math Comput 2018;320:282–91.
[21] Liang X, Pei Y, Lv Y. Modeling the state dependent impulse control for computer virus propagation under media coverage. Phys A: Stat Mech Appl 2018;491:516–27.

[22] Ren J, Xu Y. A compartmental model to explore the interplay between virus epidemics and honeynet potency. App Math Model 2018;59:86–99.
[23] Gao W, Baskonus HM. Deeper investigation of modified epidemiological computer virus model containing the Caputo operator. Chaos Solitons Fractals 2022;158:112050.
[24] Aliyev AG, Shahverdiyeva RO. Scientific and Methodological bases of Complex Assessment of Threats and Damage to Information Systems of the Digital Economy. Int J Inf Eng Electron Business (IJIEEB) 2022;14(2):23–38. doi: https://doi.org/10.5815/ijieeb.2022.02.02.
[25] R. Tkachenko, "An Integral Software Solution of the SGTM Neural-Like Structures Implementation for Solving Different Data Mining Tasks," Lecture Notes in Computational Intelligence and Decision Making. Springer International Publishing, pp. 696–713, Jul. 23, 2021. doi: 10.1007/978-3-030-82014-5_48.
[26] R. Tkachenko and I. Izonin, "Model and Principles for the Implementation of Neural-Like Structures Based on Geometric Data Transformations," Advances in Intelligent Systems and Computing. Springer International Publishing, pp. 578–587, May 12, 2018. doi: 10.1007/978-3-319-91008-6_58.
[27] Dronyuk I, Fedevych O, Kryvinska N. Constructing of Digital Watermark Based on Generalized Fourier Transform. Electronics 2020;9(7):1108.
[28] Thu Thu Khine P, Pa Pa Win H, Ni Tun KN. New Intrusion Detection Framework Using Cost Sensitive Classifier and Features. IJWMT 2022;12(1):22–9.
[29] Gan C, Qian Yi, Liu A, Zhu Q. Search-driven virus spreading on Social Internet of Things: A dynamical perspective. Commun Nonlinear Sci Numer Simul 2022;114:106624.
[30] Surajudeen Adebayo O, Bulus Micah J, Shefiu Olusegun G, O. Alabi I, Abdulazeez L. Development of Secure Electronic Cybercrime Cases Database System for the Judiciary. IJIEEB 2022;14(1):1–16.
[31] Farman M, Akgül A, Ahmad A, Saleem MU, Ahmad MO. Modeling and analysis of computer virus fractional order model. In: Methods of Mathematical Modelling. Elsevier; 2022. p. 137–57. doi: https://doi.org/10.1016/b978-0-323-99888-8.00010-3.
[32] B. Dissanayake M. Feature Engineering for Cyber-attack detection in Internet of Things. IJWMT 2021;11(6):46–54.