ECONTECHMOD. AN INTERNATIONAL QUARTERLY JOURNAL – 2016, Vol.5, No.4, 61-67

INVESTIGATION OF AGENT-BASED SIMULATION OF MALICIOUS SOFTWARE

D. Chumachenko, S. Yakovlev

National Aerospace University "Kharkiv Aviation Institute"; dichumachenko@gmail.com

Received November 08.2016: accepted December 09.2016

Abstract. Epidemics of malicious software are actual problem and network worms are one of the most important issues. Identifying trends in network worm distribution, finding the factors that influence the spread of the Internet worm will help to identify the effective preventive and precautionary measures to prevent epidemics of malicious software. To solve the problem of the development of advanced security mechanisms against network worms, different approaches to modeling the spreading of worms have been studied. Deterministic models of propagation of computer viruses in a heterogeneous network, taking into account its topological and architectural features have been analyzed and improved. Agent-based model of network worm propagation have been developed. Simulated model is based on epidemic approach to modeling. SAIDR structure of agent-based model has been used for simulation of malicious software of "network worm" type. A comparative study of developed mathematical models has been conducted. Comparative graphs of the dependence of the infected nodes number on the time of the computer system functioning in the propagation of the epidemic have been built. Research carried out by the example of the Code Red worm propagation.

Key words: agent-based simulation, imitation model, malicious software, network worm, Code Red.

INTRODUCTION

The widespread use and easy access to the Internet makes it a prime target for malicious activity. In particular, the Internet has become a powerful mechanism for spreading of malicious software. Network worms, self-contained programs that propagate via computer networks using automatic search, attacks and infection of remote computers, has being developed almost 30 years since the first Morris worm. Under modern conditions, the computer infrastructure is more vulnerable than ever before, because the speed of technological development is much higher than the development of protective measures speed [1]. Incidents of Code Red and Nimda worms in 2001 showed how vulnerable computer networks and how fast can worm spread. Moreover, Weaver introduced certain principles of worm development, using which they could spread even more rapidly [2]. In order to protect from possible attacks of worms in the future, you must understand the different properties: patterns of spread of worms throughout their life cycle, development of patches, awareness and other human countermeasures, network topology, etc.

Development of an accurate model of Internet worm will give an idea about its behavior. This will reveal weaknesses in the dynamics of a network worm, as well as forecast of its distribution to assess the damage from the activities of the worm. In epidemiological studies, there are a number of deterministic and stochastic models for spreading of viral diseases [3], as well, some models exist to simulate the spread of Internet worms.

THE ANALYSIS OF RECENT RESEARCHES AND PUBLICATIONS

Kephart, White and Chess from IBM have conducted a series of network worm simulation experiments from 1991 to 1993, based on epidemiological models of virus infections [4]. All traditional epidemiological models have homogeneous nature, i.e., any infected host can equiprobable infect any susceptible host [3]. Taking into account the local interactions of viruses at that time [5], epidemiological models have been applied to some non-homogeneous networks: random graphs, two-dimensional lattice and a hierarchical graph of type "tree". Despite the fact that at that time the hypothesis of local interaction was correct because of the sharing of information storages, in the modern world it is not suitable for simulating the behavior of network worms, since the vast majority of worms are distributed via the Internet and can hit the target directly. In addition for modeling, model Susceptible-Infected-Susceptible (SIS), which suggests that healed computer can be infected again immediately, has been studied.

Wang and his colleagues presented the results of a simulation of a simple network worm in clustering and hierarchical tree-like networks [6]. They have shown that selective immunization in some network topologies can significantly slow down the spread of the virus. On the other hand, their conclusion was based on a hierarchical tree topology, which is not suitable for the Internet network.

The epidemic of the Code Red worm in July 2001 has stimulated active research activities of simulation and analysis of the worm on the Internet. Staniford and colleagues used classical epidemic simple equations for simulating spreading of the Code Red worm immediately after the incident, July 19th [7]. Their model fairly accurately matched the actual data with the limited observations. Moore has provided the observed data, and the analysis of the behavior of the worm Code Red [8]. Weaver suggested some of the approaches in the design of worms, which can be used to spread malicious code faster than Code Red and Nimda worms [2].

Previous studies on modeling of spreading of network worms neglected the dynamic effects of human countermeasures to the epidemic. Wang and his colleagues looked at protection against the epidemic by means of immunization. However, they only take into account the static immunization, which means that part of the hosts is vaccinated prior to the spreading of network worm. But in real conditions, human countermeasures are dynamic actions which play a major role in reducing the rate of network worm propagation and prevent outbreaks. Many new viruses and worms come out almost every day. Most of them, however, go out without infecting a large number of computers due to the timely implementation of human countermeasures.

The simulation of epidemic process assumes that the level of viral infection is constant. The epidemic modeling process assumed that the level of viral infection is constant. The above-described models of Internet viruses and network worms estimate the time required to search for target of infection of host, regardless of whether it is already infected or not, as a constant. Weaver considered the level of infection, as a random variable, given the unsuccessful attempts to IP-scan of the network worm. Weaver considered the level of infection, as a random variable, given the unsuccessful attempts to scan the IP-worm. However, the average value of rate of infection propagation is still assumed as constant, which is true for simulating epidemics of disease, but it may not be true for Internet viruses and worms.

OBJECTIVES

Modeling of propagation of network worm will identify trends in its distribution, find the factors that influence the spread of the Internet worm. Also model helps to identify the effective preventive and precautionary measures to prevent epidemics of malicious software. To solve the problem of the development of advanced security mechanisms against network worms, different approaches to modeling the spreading of worms were studied. The aim of given research is to develop the agent-based model of malicious software of network worm type by example of Code Red worm.

THE MAIN RESULTS OF THE RESEARCH

SEIQR model. According to this model structure, objects are divided into five groups: Susceptible (S), Infected (I), Exposed (E), Removed (R), Quarantine (Q) (Fig. 1). Here, transitions of states can happen in the following ways:

– from Exposed to Removed,
– from Exposed to Infected,
– from Infected to Quarantine and subsequently to Removed,
– from Infected to Removed.

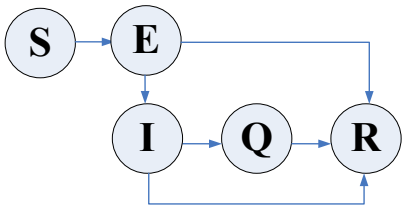

Fig. 1. State changes in SEIQR model

Dynamics of system is described by following differential equations:

$$\begin{cases} \dfrac{dS(t)}{dt} = -\dfrac{\beta I(t)}{N} S(t), \\ \dfrac{dE(t)}{dt} = \dfrac{\beta I(t)}{N} S(t) - (\alpha + k)E(t), \\ \dfrac{dI(t)}{dt} = \alpha E(t) - (\gamma + \delta)I(t), \\ \dfrac{dQ(t)}{dt} = \delta I(t) - \gamma Q(t), \\ \dfrac{dR(t)}{dt} = kE(t) + \gamma(Q(t) + I(t)). \end{cases} \quad (1)$$

where: β is the infection rate, δ is quarantine rate, k, γ are treatment rates, α is transition rate from a latent state E to the infected I (i.e., the average time spent in a latent state).

Results of the Code Red Worm propagation are shown in Fig. 2. Results show that the implementation of additional types of object management and the possibility of implementation of network nodes in quarantine improves the accuracy of the final result on the condition that the anti-virus software is updated.

Progressive SIDR model. Progressive SIDR model or PSIDR model takes into account two factors (Fig. 3):

1. The classical SI model works at early stage of epidemic distribution. Its duration is π.

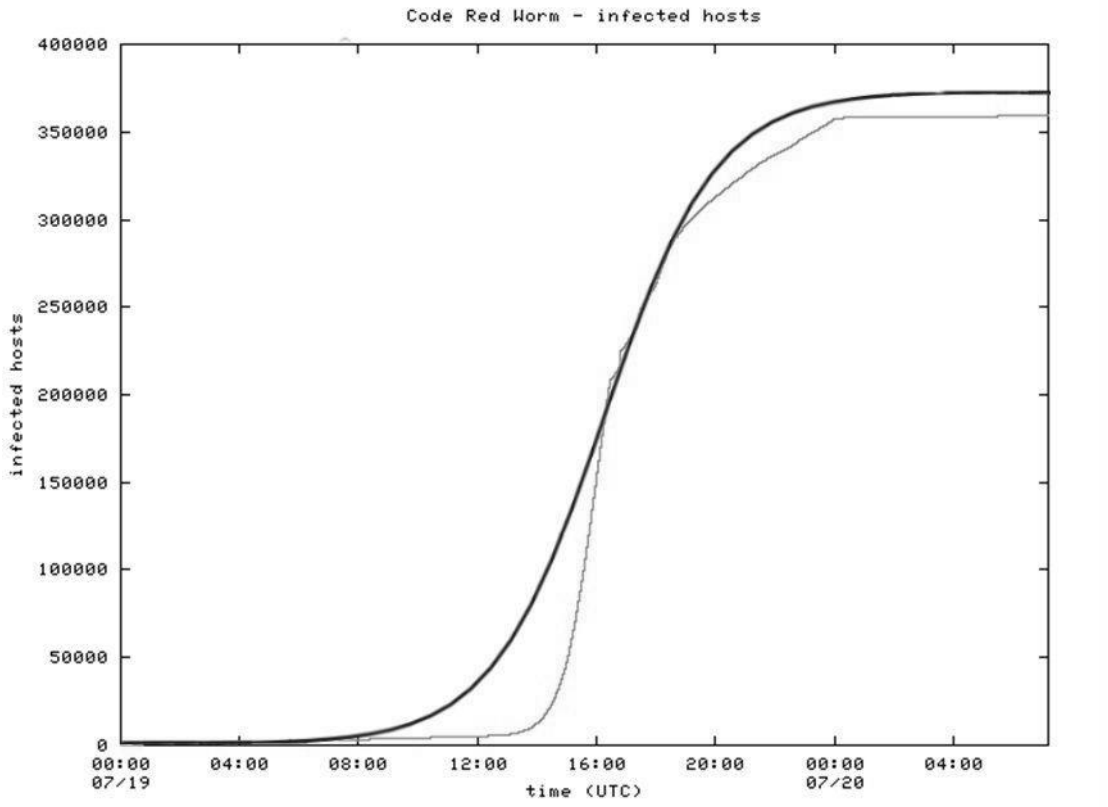

Fig. 2. SEIQR model of Code Red propagation (black) and actual Code Red propagation (grey)

2. Further development of the epidemics is described by introducing an additional state *D* (Detected), in addition to the states *S*, *I* and *R*. The node is said to be in the state D, when the presence of the worm was already discovered, but active counteraction has not yet begun.

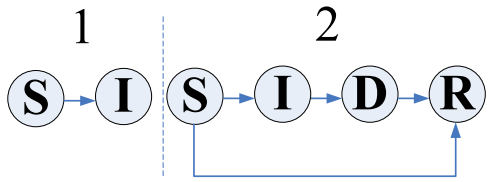

Fig. 3. State changes in PSIDR model

Infection and recovery of network nodes are performed as before, with the average speed β and γ, a transition from state *I* to an "intermediate" state *D* with the speed of μ nodes per time unit.

At stage 1, when $t < \pi$ the following condition applies:

$$S(t) + I(t) = N, \tag{2}$$

and the system is described by following equations:

$$\begin{cases} \dfrac{dS}{dt} = -\beta SI, \\ \dfrac{dI}{dt} = \beta SI. \end{cases} \tag{3}$$

At stage 2, when $t \geq \pi$ the following condition applies:

$$S(t) + I(t) + D(t) + R(t) = N, \tag{4}$$

and the system is described by following equations:

$$\begin{cases} \dfrac{dS}{dt} = -\beta SI - \mu S, \\ \dfrac{dI}{dt} = \beta SI - \mu I, \\ \dfrac{dD}{dt} = \mu I - \delta D, \\ \dfrac{dR}{dt} = \delta D + \mu S. \end{cases} \tag{5}$$

Results of the Code Red Worm propagation are shown in Fig. 4.

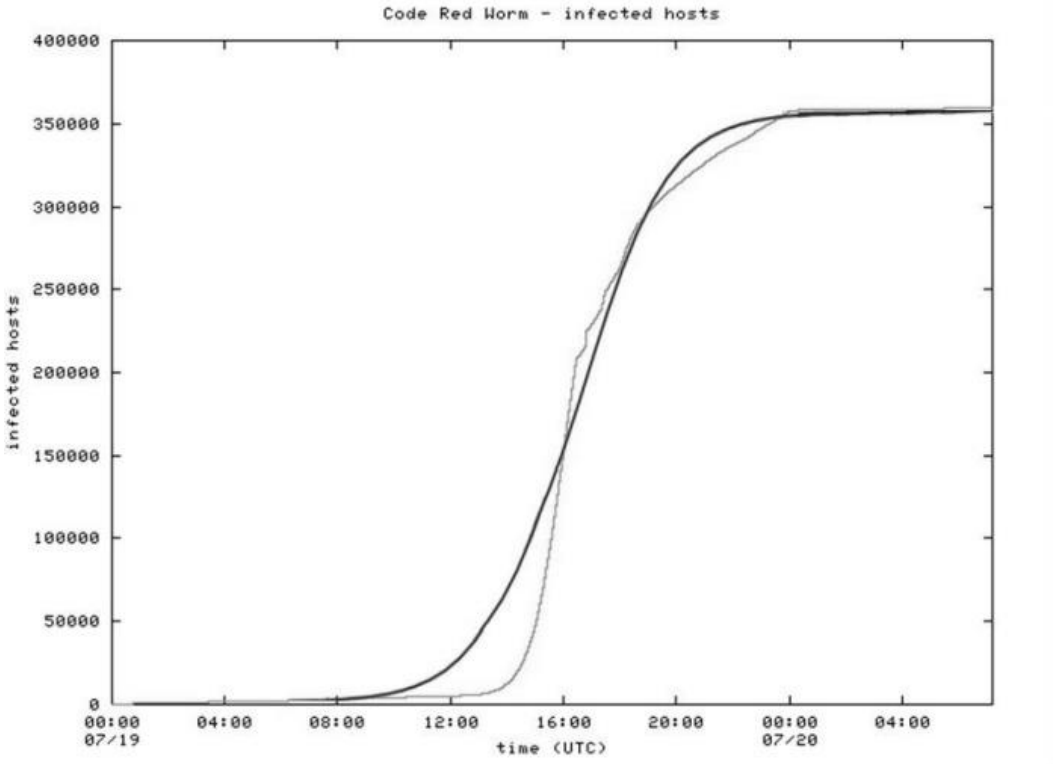

Fig. 4. PSIDR model of Code Red propagation (black) and actual Code Red propagation (grey)

Agent-based simulation. The first research on the epidemic process spreading in medicine, which allowed to make a step forward compared to the classical models, was presented in [9] (Baroyan-Rvachev Model). Actually, this model is the same as the classic models as it is based on the system of differential equations. However, it uses the concept of the model time and takes into account the length of a particular state. However, the abovementioned disadvantages of the model were not fully resolved.

To eliminate these drawbacks, the agent-based simulation was developed. It allows taking into account a large number of factors influencing the process of malware distribution [10-14].

Adequacy of the simulation model is highly dependent on the number of agents in the system. The use of large populations and detalization of properties of agents leads to the need of using the most modern information tools and technologies, particularly the algorithms that are optimal in the number of operations performed by the machine. In agent-approach the simulation process is based on the construction and processing of the event queue, which can be divided into two types:

1. Changing the agent state from the point of view of the external environment (the physical position of agent);
2. Changing the internal state of the agent. Events of this type arise from the interaction of the agent with other agents, as well as with the environment.

The objective is to find and use such a set of properties and methods of agents, which would allow the use of agent-based approach in the greatest way.

In this paper, we propose a formal description of the constructed agent-based model of the Code Red worm propagation. Agent can be viewed as a set of attributes:

$$a = \langle s, s_t, c, l \rangle, \; a \in A, s \in S, c \in C, \qquad (6)$$

where: s_t is time that agent is in state s;

A – amount of agents;

S – set of possible states;

l – life duration;

C – set of cells of working area.

The set of states of the agent is determined beforehand and is permanent. In our model we define set of states as:

$$S = \{Susceptible, Antidotal, Infected, Detected, Recovered\}.$$

This set of states is based on the idea of dividing the entire population into subsets based on their states according to epidemic status. The proposed set describes the model as an analog of expanded SAIDR model.

Fig. 5 shows the transitions between states:

- Susceptible – agent is not infected. This state is only applicable to agents that can be infected with a particular worm.
- Antidotal – agent is not infected. Agents in that state have the installed antivirus software, so it cannot be infected.
- Infected – agent is infected and can propagate infection to the other hosts.
- Detected – agent's infection is detected by antivirus software and isolated from the network.
- Recovered – agent is cured and is not vulnerable to this particular worm anymore.

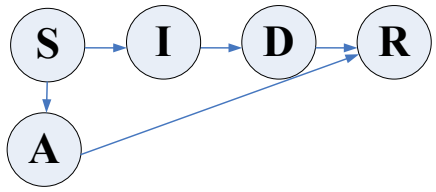

Fig. 5. State changes in expanded SAIDR model for agent-based system

Network is described as a set of cells. Decomposition of workspace leads to a set of cells as the contingent of abstract objects, that include a set of agents and one infected agent.

Thus, a cell can be defined as:

$$c = \left\langle z, \widetilde{A}, i \right\rangle, z \in Z, \widetilde{A} \subset A, i \in I, \qquad (7)$$

where: I is a set of all infected agents; Z is a set of all zones.

In our model, infection is possible only directly from one agent to another. If we treat propagation as a direct communication from one host to another, computation will be inefficient. Processing of such communication is very hard. To simplify the computations, we suggest treating the infection of agent as a fact of agents belonging to the same cell. It will allow taking into account the interaction between agents and significantly reducing the performance loss.

The propagation is implemented in the following way. At first, the possibility of communication between two hosts is checked. The communication and subsequent infection is considered to be possible if a cell of the currently processed agent includes other agents. Each pair of agents of the cell is processed and if at least one of them is infected, it is considered that the communication happens with a certain probability.

Method of agents operating in the crucial network consists of three sub-functions: perception, decision and transformation.

Perception sub-function:

$$Per : CE \rightarrow F_{in} \qquad (8)$$

provides a selection of information from the environment and the assignment of values of input attributes.

Decision sub-function

$$Dec : F_{in} \rightarrow F_{out} \qquad (9)$$

determines the values of the output variables from the values of the input.

Transformation sub-function

$$Tran : F_{out} \rightarrow CE' \qquad (10)$$

changes the state of the environment by performing the translation operation of sets of elements from one to another in accordance with the rules, as well as removes the elements of the sets.

The general structure of the constructed agent-based system can be expressed by tuple:

$$MAS=\{Ind, Prp, Atr, Inp, Out, Str\}, \qquad (11)$$

where: *Ind* is the name of the system, *Prp* is the aims of the systems, *Atr* is general system descriptions, *Inp* is the entrance of system, *Out* is the output of the system, *Str* is a structure of the system *Str={CE, R}*, and *CE* are component elements of the system, *R* are communications of components.

The prototype of developed agent-based model of the Code Red worm spreading was realized using NetLogo agent-based software (Fig. 6).

The advantage of using a given model is its simple interface. As you can see from screenshot (Fig, 6), user of developed model can change input parameters (number of hubs in the given network, infection rate, test frequency etc.) and observe plots of worm dynamics and system behavior in real time. However, the implementation of the proposed model with a large number of agents in NetLogo software requires very high processing power. This does not allow simulating the real processes occurring in a network with a large number of hosts. To solve this problem, the model was developed in Python programming language. The simulation results show the behavior that is similar to real statistical data (Fig. 7).

This research allows us to suggest a hypothesis that time series corresponding to the number of infected hosts must not exceed a certain threshold value for the non-arising of the epidemic. Thus, the dynamical system describing the agent-based model should be the reflection of a finite set I in itself, *MAS:* $I \rightarrow I$, that is a sign of chaotic dynamics for nonlinear mappings. Thus, it seems to be relevant to create conditions in relations (8)-(10), by which it is possible to judge about the stability of the dynamics of agent-based system.

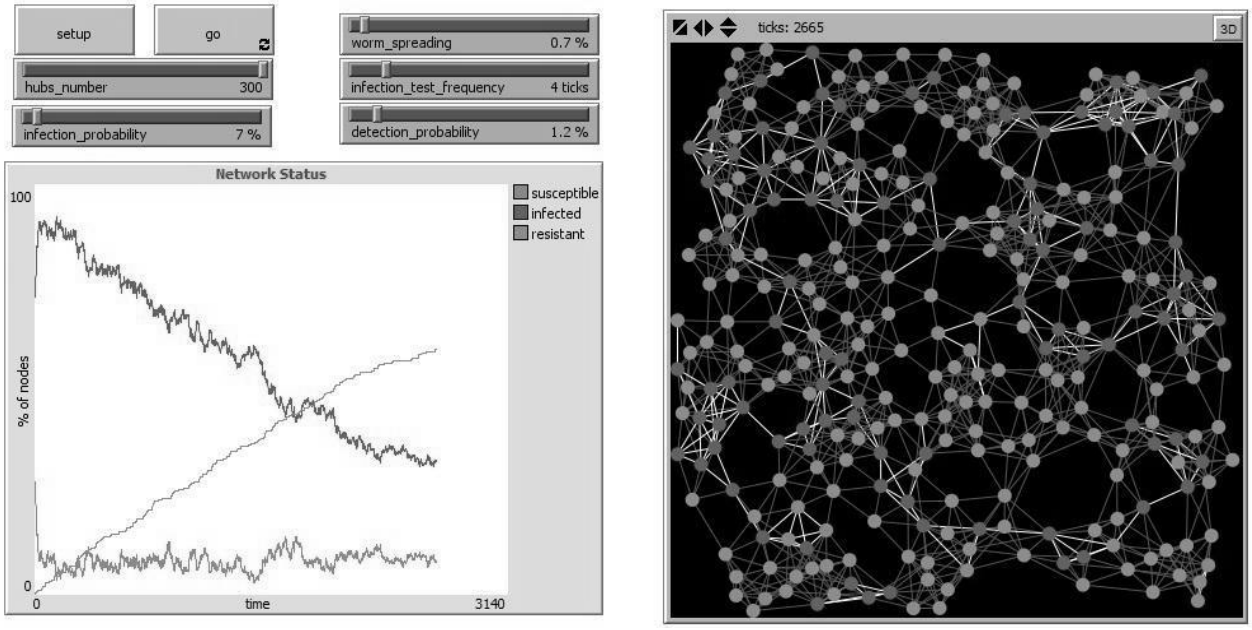

Fig. 6. Model in NetLogo

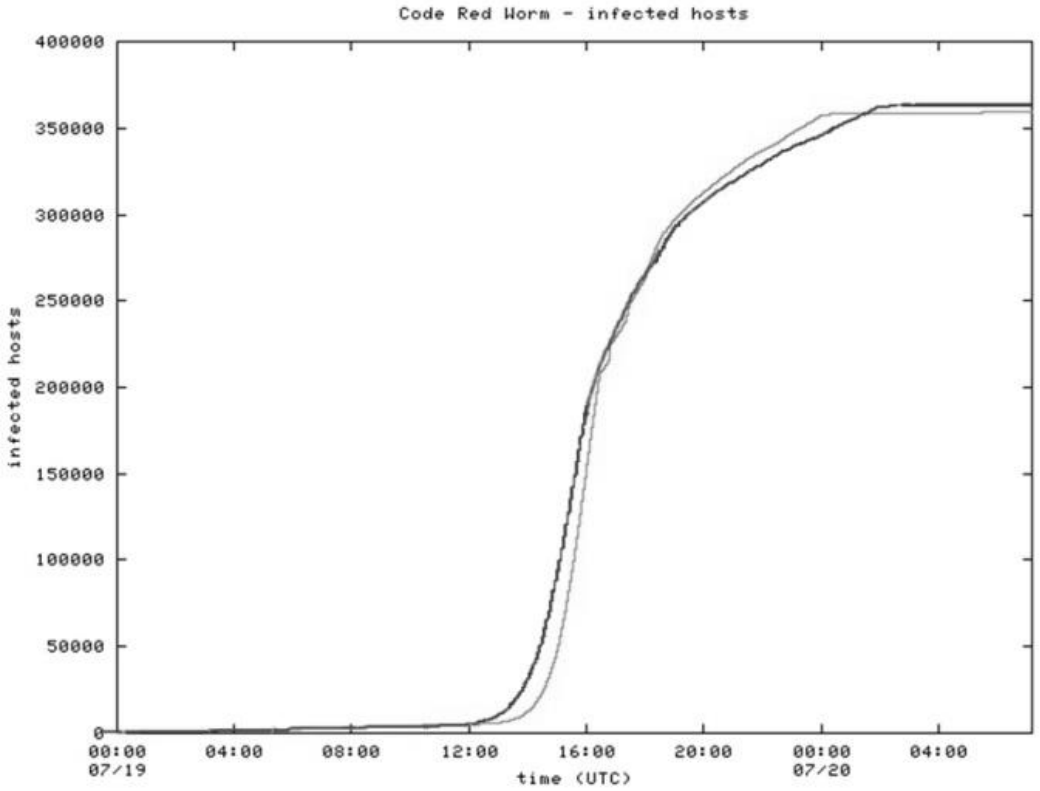

Fig. 7. Agent-based model of Code Red propagation (black) and actual Code Red propagation (grey)

CONCLUSIONS

1. Agent-based model of the malicious software have been developed on the basis of existing models and modeling approaches by the example of the Code Red Worm. Agent-based simulation showed the most close to the real data results.

2. Developed agent-based model is universal for analyzing the dynamics of the behavior of any kind of malicious software of type network worm. In given study the dynamics is shown by the example of the Code Red worm. The adequacy of the model has been checked with real statistics of the Code Red worm incidence.

3. Analysis and comparative research of developed mathematical models led to the conclusion about improving of their accuracy in comparison with the known mathematical models. That led to the conclusion about the appropriateness of using developed model in computer networks. At the same time agent-based simulation results allow practically repeat a real worm behavior.

REFERENCES

1. **Nachenberg C., 2000.** The Evolving Virus Threat. 23rd NISSC Proceedings.

2. **Weaver N., 2001.** Warhol Worms: The Potential for Very Fast Internet Plagues.

3. **Chernyshev Y., Chumachenko T., Chumachenko D., Tovstik A., 2012.** System of Simulation of Epidemic Diseases Spreading. Proceedings of East West Fuzzy Colloquium 2012 (19th Zittau Fuzzy Colloquium, September 5 – 7, 2012), 154 – 161.

4. **Kephart J., Chess D., White S., 1993.** Computers and Epidemiology. IEEE Spectrum 30 (5). 20 – 26.

5. **Kephart J., White S., 1993.** Measuring and Modeling Computer Virus Prevalence. Proceedings of the IEEE Symposium on Security and Privacy. 2 – 15.

6. **Wang C., Knight J., ElderM., 2000.** On Viral Propagation and the Effect of Immunization. Proceedings of 16th ACM Annual Computer Applications Conference. 246 – 256.

7. **Staniford S., Paxson V., Weaver N., 2002.** How to Own the Internet in Your Spare Time. 11-th Usenix Security Symposium. 149 – 167.

8. **Moore D., Shannon C., 2003.** The Spread of the Code-Red Worm. Center for Applied Internet Data Analysis.

9. **Baroyan O., Rvachev L., 1978.** Forecasting of influenza epidemics in USSR. Medicine #2.131-137. (in Russian).

10. **Wooldridge M., 2009.** An Introduction To Multiagent Systems. 468.

11. **Chumachenko T., Chumachenko D., Sokolov O., 2013.** Multiagent simulation of the hepatitis B epidemic process. Online journal of public health informatics, 5 (1).

12. **Boyko N.**, **Kutyuk O., 2016.** Basic concepts of evolution in agents calculating and agents system. ECONTECHMOD. An International Quarterly Journal, Vol. 05, No. 2. 69-76.

13. **Chernyshev Y., Chumachenko D., Tovstik A., 2013.** Development of intelligent agents for simulation of hepatitis B epidemic process. Proceedings of East West Fuzzy Colloquium 2013 (20th Zittau Fuzzy Colloquium, September 25 – 27, 2013). 161 – 168.

14. **Bobalo Y., Politanskyi R., Klymash M., 2015.** Traffic simulation in a telecommunication system based on queuing systems with different input flows. ECONTECHMOD. An International Quarterly Journal, Vol. 04, No. 1. 11-16.