Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 198 (2022) 694–699

Procedia Computer Science

www.elsevier.com/locate/procedia

International workshop on Small and Big Data Approaches in Healthcare (SBDaH)
November 1-4, 2021, Leuven, Belgium

Models for early web-attacks detection and intruders identification based on fuzzy logic

Zhadyra Avkurova[a], Sergiy Gnatyuk[b], Bayan Abduraimova[a], Solomiia Fedushko[c,*], Yuriy Syerov[c], Olha Trach[c]

[a]*L.N. Gumilyov Eurasian National University, Satbayev Str., Nur-Sultan, 010000, Kazakhstan*
[b]*National Aviation University, 1 Liubomyr Huzar Ave, Kyiv, 03058, Ukraine*
[c]*Lviv Polytechnic National University, 12 Stepana Bandery Str., Lviv, 79000, Ukraine*

Abstract

Modern information and communication technologies are implemented in vatious spheres (in particular, sectors of the critical infrastructute), but these technologies are vulnerable to web-attacks and other emerging threats. Early detection of such threats is a relevant and important research task without universal solution for various information and communication systems. In this paper, a method of linguistic terms using statistical data was used for structural and analytical models of the host and network parameters construction. Based on these models, system of the logical rules can be developed to provide the functioning of web-attacks detection system. In the future, given results as well as system of the logical rules can be used to create advanced intrusion detection system based on honeypot technology (or others) for web-attacks detection and intruder identification.

Keywords: Big Data; WEB-attack; Detection; Identification; Honeypot; Fuzzy Logic; Parameter.

1. Introduction

The development of information and communication technology (ICT) creates new types of threats to information security. The intruder in computer systems and networks (for example, WEB-attacks or other negative influences)

* Corresponding author. Tel.: +38 0322 582 595
E-mail address: solomiia.s.fedushko@lpnu.ua

1877-0509
10.1016/j.procs.2021.12.308

occupies a prominent place. To effectively counter this threat, IDS (intruder detection system) is being developed to detect and identify intruders. Early detection is essential and not a simple task for the security side. A typical IDS should perform the following main functions [1]: monitor and analyze the activity of ICS (information and communication system) users; capture system configurations and vulnerabilities; assess the integrity of critical system files and data files; recognize activity patterns that reflect known attacks; perform statistical analysis to detect abnormal behavior; identify violations of a security policy by the system user. The tasks to be solved by IDS can be divided into global and local. Global tasks – recognition of the violator (Vl) and legitimate user – the solution of this problem contains the following stages [2-3]: data collection, filtering, behavior classification - directly recognizing the Vl, report, and response system. As can be seen from the main functions and tasks of IDS, one of the most critical aspects of their functioning is the fixation of the violation of ICS protection and its identification. This is important for critical infrastructure protection, for example, in e-health [4, 5].

2. Development of the structural and analytical models based on host and network parameters

2.1. Basic parameters for intruders identification

In the process of attack, the violator, acting on the system, changes specific parameters, creates or terminates its inherent processes, etc. All these actions are reflected in the state of the system. Evaluating these parameters, you can detect the fact of intrusion into the system. The work of modern IDSs is based on this principle. Thus, the NIDES system performs audits of such processes as logging in, working with files and processes, administration, and fixing errors and failures. Previous works describe the parameters by which the developed system identifies the violator. These parameters (are host settings) include:

- Host Parameters: Username at login, *UID;* Login time, *Tlog;* Frequency of login requests, *Nlog;* Time spent logging in, *TSlog;* Intensity of actions, *I;* Processor time / CPU usage, *CPU;* The amount of RAM load, *Muse;* Number of executable files, *NEF;* The type of files used in the attack, *AtEF;* Number of failures and errors, *NEr;* Process/file execution time, *RTPr/F;* Unusual processes, *UPr;* File transfer to the system, *TrFin;* Files changes, *ModF;*copying / transferring files from the system, *TrFout;* Pressing the keyboard keys, *KS.*
- Network Settings – characteristics of *ARP-, IP-, ICMP-* and *TCP* packages.

Since the process of detection and identification of the violator takes place in conditions of uncertainty, and some of the parameters of the IDS are unclear, the operation of such a system should be based on fuzzy logic. To identify the violator, you can use the logical-linguistic approach and the basic model of parameters, partially described in [4], which will be the basis for constructing the developed IDS. For example, to detect the process of port scanning in section [6] used linguistic variables (LV) "Number of virtual channels" and "Age of virtual channels", and an area [7] LV "Number of simultaneous connections", "Query processing speed", "Delay between requests" and "Number of packets with the same sender and recipient address"- to detect DDOS attacks and spoofing.

The process of detecting and identifying the violator requires determining the necessary parameters and their properties. In this regard, the primary purpose of this work is to build models of standards as are necessary for the operation of IDS in a vaguely defined, poorly formalized environment.

2.2. Method of linguistic terms using statistical data

Consider the method of linguistic terms using statistical data (MLTS) [8]. In contrast, a measure of belonging of the element to the set estimates the frequency of use of the concept, which is given by a fuzzy set to characterize the element. To do this, the values of the linguistic variable (LV) are placed on the universal scale [0; 1] $X=\{x_1, x_2, \ldots, x_n\}$. The method is based on the condition that the same number of experiments falls into each interval of the scale, but this is usually not followed in practice. An empirical table is compiled in real conditions, in which experiments can be unevenly distributed over intervals. Some of them may not be involved, and then the data is processed using a matrix of prompts [9]. May it is necessary to estimate in values of LV deviations of the parameter $\Delta B \in [0, B]$ (B - the maximum possible deviation), which characterizes the current measurements. Next for n = 5 determine the value of LV $\{x_1, x_2, x_3, x_4, x_5\}$. Interval [0, B] and ΔB/B (estimated ratio) divided into k segments (for example, 5), on which the statistics characterizing frequency of use by the expert of the value of drugs for the display of the conclusions gathers. Then the data are entered into the table and processed to reduce the errors made during the

experiment: the table is removed individual elements on the left side and on the right side of which there are zeros in the row. The tooltip matrix is a string whose elements are calculated by the formula:

$$k_j = \sum_{i=1}^{n} b_{ij} = \sum_{i=1}^{5} b_{ij},\ j = \overline{1,5}\ . \tag{1}$$

Next, in the resulting row of the matrix, the maximum element is selected $k_{\max} = \max k_j$, and then all elements of the table are converted by expression

$$c_{ij} = b_{ij} k_{max} / k_j,\ \ i = \overline{1,5};\ j = \overline{1,5}\ , \tag{2}$$

and for columns, where $k_j = 0$ the linear approximation is applied $c_{ij} = (c_{ij-1} + c_{ij+1})/2,\ \ i = \overline{1,5};\ j = \overline{1,5}$. Next, calculate the value of MF (membership function) by the formula

$$\mu_{ij} = c_{ij} / c_{imax}\ ,\ \ c_{imax} = \max_j c_{ij}\ ,\ \ i = \overline{1,5};\ j = \overline{1,5}\ . \tag{3}$$

The described method uses data from statistical studies. Their processing is quite time-consuming because to build a MF of one term, it is necessary to conduct statistical studies of all terms of LV. We construct a model of standards of linguistic variables for fuzzy parameters of violator identification from the set of parameters.

DIO = <UID, Tlog, Nlog, TSlog, I, CPU, MUse, NEF, AtEF, NEr, RTPr/F, UPr, TrFin, ModF, TrFout, KS, ARP, IP, ICMP, TCP>.

2.3. Structural and analytical models

Login time, Tlog. This parameter is based on the fact that the activity of the ICS and users of this system depends on the time of receipt. Usually, the usual greater activity of users to log in is detected on the last day, less - at night. Still, other statistics are possible, determined by the mode of operation of the organization to which the ICS belongs. The nature of these parameters is unclear, due to which it is impossible to conclude the message's illegal activity unambiguously. Thus, in organizations working from 08.00 to 16.00, the probability of who is the user who logs in - the message is lowest at 08.00 and increases over time, reaching a maximum in the years after 16.00. However, it should be changed that in the concepts of honeypot-technology, this parameter loses weight, as any activity on them is considered criminal. Let's evaluate the LV "Level of legitimacy over time". Determine the value of the linguistic variable $\{x_1, x_2, x_3\}$, corresponding {legitimate, suspicious, illegitimate}.

That is $T_{Tlog} = \bigcup_{i=1}^{3} T^i{}_{Tlog} = \{legitimate,\ suspicious,\ illegitimate\}$. We use statistics for B = 24 hours. It is advisable to divide the total interval into 4 intervals [00:00;06:00], [06:00;12:00], [12:00;18:00], [18:00;24:00].

Table 1. Data for LV Tlog

Value of LV	Interval			
	№1	№2	№3	№4
High	0	8	6	1
Middle	2	1	2	3
Low	6	1	1	4

Using expression (1), we define $k_j = \|8\ 10\ 9\ 8\|$, where $k_{max} = 10$, and in accordance with (2) calculate:

$$\|c_{ij}\| = \begin{Vmatrix} 0 & 8 & 6,66 & 1,25 \\ 2,5 & 1 & 2,22 & 3,75 \\ 7,5 & 1 & 1,11 & 5 \end{Vmatrix}.$$

Calculate the MF by formula (3):

$$\left\|\mu_{ij}\right\| = \begin{Vmatrix} 0 & 1 & 0,83 & 0,16 \\ 0,66 & 0,26 & 0,59 & 1 \\ 1 & 0,13 & 0,15 & 0,66 \end{Vmatrix}.$$

For $\bigcup_{i=1}^{3} \mu_{ij}$ accordingly, we find the evaluation relationship $\bigcup_{i=1}^{3} \Delta B_i / B = \{0,25;\ 0,5;\ 0,75;\ 1\}$, and we obtain the following fuzzy numbers:

$$L = \{0/0,25;\ 1/0,5;\ 0,83/0,75;\ 0,16/1\},$$

$$P = \{0,66/0,25;\ 0,26/0,5;\ 0,59/0,75;\ 1/1\},$$

$$N = \{1/0,25;\ 0,13/0,5;\ 0,15/0,75;\ 0,66/1\}.$$

Schedule MF terms LV The login time is shown in Fig. 1.

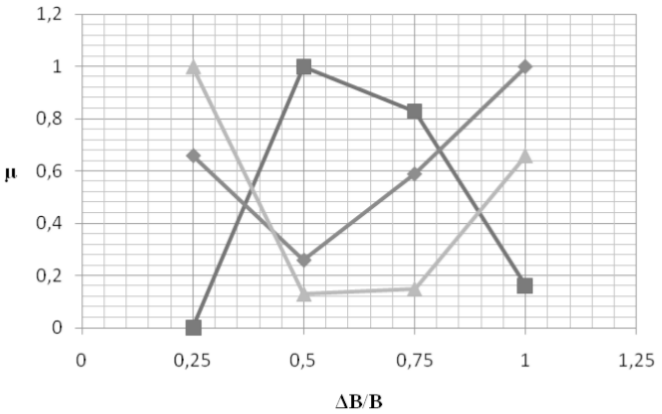

Fig. 1. Linguistic standards of fuzzy numbers for *Tlog*

Frequency of login requests, Nlog. The highest frequency of login requests will be observed during system attacks by bots (including hacker bots, as spammers do not require login). The human offender is also marked by an increased frequency of requests due to attempts to circumvent the protection and the theoretical assumption that he does not have a legitimate login and password, so he will be forced to make at least a few attempts. And the greater the number of attempts, the more likely the violator is trying to enter the ICS. This parameter is also fuzzy.

Using (1) – (3), we can form structural and analytical models for analogically *Nlog* (Fig. 2, Table 2):

$$T_{Nlog} = \bigcup_{i=1}^{5} T^{i}{}_{Nlog} = \{\ low,\ below\ average,\ medium,\ above\ average,\ high\ \}.$$

Table 2. Data for LV Nlog

The value of LV	Interval				
	№1	№2	№3	№4	№5
Low	8	0	0	0	0
Below average	5	2	0	0	0
Average	1	6	4	0	0
Above average	0	2	8	1	0
High	0	0	1	6	6

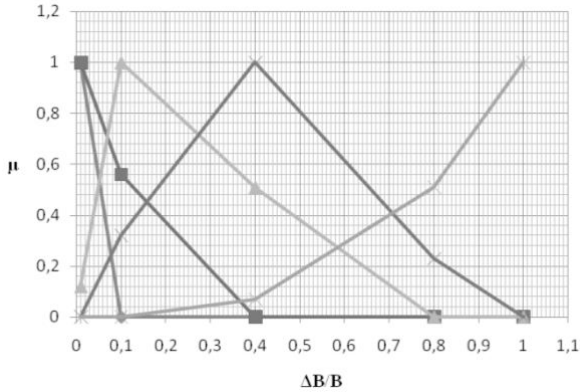

Fig. 2. Linguistic standards of fuzzy numbers for *Nlog*

Time spent logging in, TSlog. A parameter that is closely related to the previous one. The time spent by the infringer is in most cases longer than the time spent by the legitimate user. But it is unclear because it does not allow unambiguous identification.

Using (1) – (3), we can form structural and analytical models for analogically *TSlog* (Fig. 3, Table 3):

$$T_{Slog} = \bigcup_{i=1}^{5} T^{i}{}_{Slog} = \{\ very\ small,\quad small,\ medium,\ large,\ very\ large\}\ \}.$$

Table 3. Data for LV *TSlog*

The value of LV	Interval				
	№1	№2	№3	№4	№5
Very small	9	3	0	0	0
Small	5	10	1	0	0
Medium	1	7	5	0	0
Large	0	1	2	9	2
Very large	0	0	1	6	9

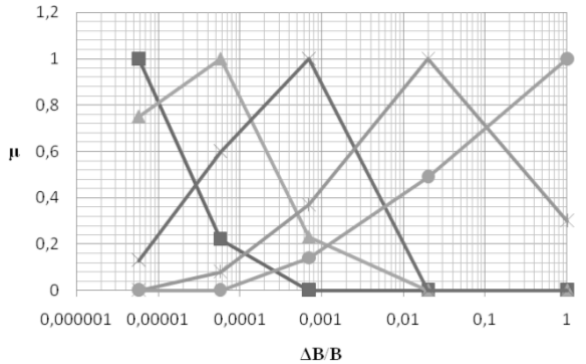

Fig. 3. Linguistic standards of fuzzy numbers for *TSlog*

The intensity of actions, I. Here we mean the number of any user actions, including login/logout, transfer, change, copy files, start/stop processes, etc., per unit time. The intensity may not differ between the human violator and the legitimate user. Still, in bots, it is much higher, so it is essential to identify and delimitate human-bot categories. However, a significant excess of the norm indicates unauthorized automatic systems-violators (bots), a fuzzy parameter because the normal value of the intensity index is complicated to determine.

Using (1) – (3), we can form structural and analytical models for analogically *I* (Fig. 4, Table 4).:

$$T_I = \bigcup_{i=1}^{3} T^{i}{}_{I} = \{low,\ medium,\ high\}.$$

Table 4. Data for LV *TSlog*

The value of LV	Interval			
	№1	№2	№3	№4
Low	7	5	1	0
Middle	0	7	4	0
High	0	1	5	7

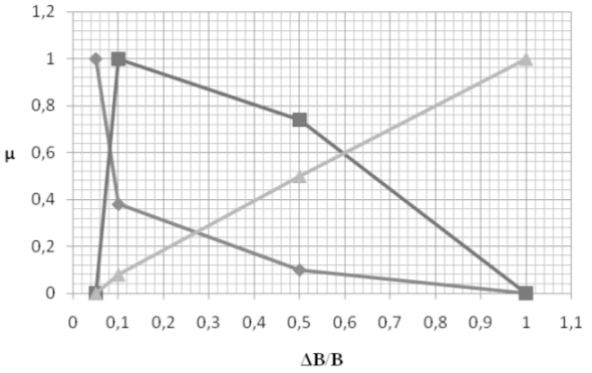

Fig. 4. Linguistic standards of fuzzy numbers for *I*

Analogically, using (1) – (3), structural and analytical models for other parameters (*CPU, Muse, NEF, NEr, RTPr/F*) can be formed and presented. It will be presented in authors extended research study.

3. Conclusions

In this paper, based on MLTS, LV was introduced, and structural and analytical models of parameters were built for *Tlog, Nlog, TSlog, I, CPU, Muse, NEF, NEr, RTPr/F*. Also, for each described LV, MF was calculated, and schedules of their terms were constructed. The formed standards are necessary for forming the logical rules allowing to provide functioning of IDS for WEB-attacks (or other threats) detection and intruder identification. In the future obtained results will be used to build an IDS system based on honeypot technology. The next step rules system development for detecting the violation of ICS and identifying the violator's person.

Acknowledgments

National Research Foundation of Ukraine supported this work within the project "Science for the safety of human and society", "Methods of managing the web community in terms of psychological, social and economic influences on society during the COVID-19 pandemic" (grant number 174/01/0341).

References

[1] M. Khosravi, B. Ladani (2020) "Alerts Correlation and Causal Analysis for APT Based Cyber Attack Detection", *Access,* **8**:162642-162656.

[2] Denning D.E. (1987) "An Intrusion-Detection Model", *IEEE Transactions On Software Engineering,* **SE-13 (2)**: 222-232.

[3] Hu Z., Odarchenko R., Gnatyuk S., Zaliskyi M., Chaplits A., Bondar S., Borovik V. (2020) "Statistical techniques for detecting cyberattacks on computer networks based on an analysis of abnormal traffic behavior", *IJCNIS,* **12 (6)**:1-13.

[4] M. Zaliskyi et al. (2018) "Method of traffic monitoring for DDoS attacks detection in e-health systems and networks", *CEUR,* **2255**: 193-204.

[5] N. Shakhovska et al. (2019) "Development of Mobile System for Medical Recommendations". *Procedia Computer Science*, **155**: 43-50. https://doi.org/10.1016/j.procs.2019.08.010

[6] A. Paradise et al. (2017) "Creation and Management of Social Network Honeypots for Detecting Targeted Cyber Attacks", *IEEE Transactions on Computational Social Systems,* **4(3)**: 65-79.

[7] Svarovskiy S. (1980) "Approximation of membership functions for linguistic variables", *Mathematical issues of data analysis,* 127-131.

[8] M. Zuzčák, P. Bujok. (2019) "Causal analysis of attacks against honeypots based on properties of countries", *IET IS,* **13 (5)**: 435-447.

[9] W. Zhang, B. Zhang, Y. Zhou, H. He and Z. Ding (2020) "An IoT Honeynet Based on Multiport Honeypots for Capturing IoT Attacks", *IEEE Internet of Things Journal,* **7 (5)**: 3991-3999. doi: 10.1109/JIOT.2019.2956173