Honeypots Network Traffic Parameters Modelling

Roman V. Maximov [1], Sergey P. Sokolovsky [1] and Alexander P. Telenga [1]

[1] *Krasnodar Higher Military School named after the general of the Army S.M.Shtemenko, 4 Krasina ul., Krasnodar, 350963, Russia*

Abstract

One of the relevant and practically significant applications of fractal traffic analysis is modeling of network interaction processes of distributed information systems. The obtained model allows to solve such important information security problem as reproduction of false and (or) hiding of true individual information technologies and (or) structural and functional characteristics of information system or its segments. Three situations of information system functioning are considered, the parameters of network traffic that necessary for the development of the model are highlighted. The Hurst index for the time series of network traffic parameters is calculated, and the coefficients of the van der Pol nonlinear oscillator equation are selected on its basis.

Keywords

Information protection, false network information objects, Hurst index, de-masking features, Van der Pol nonlinear oscillator

1. Introduction

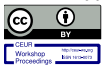
The presence of fractal properties of network traffic was discovered several decades ago, when it was found that it has the property of self-similarity, that is, it looks qualitatively the same at sufficiently large scales of the time axis and exhibits a long-term dependence [1, 2, 3, 4, 5, 6, 7, 8]. In particular, unlike processes that do not have fractal properties, there is no rapid "smoothing" of the process when averaged over the time scale - it retains a tendency to spikes.

Prior to that, the dominant traffic models based on Markov processes had a short-term dependence. They were borrowed from telephone networks and, as applied to computer networks, led to an underestimation of the load. The discovery of the self-similarity of traffic had a significant impact on the subsequent development of client-server information systems and allowed us to rethink the probabilistic and temporal characteristics of such systems.

One of the relevant and practically significant applications of fractal traffic analysis is modeling of network interaction processes of distributed information systems. The obtained model allows not only to predict the system behavior in various critical situations, but also to solve such important information security problem as reproduction of false and (or) hiding of true individual information technologies and (or) structural and functional characteristics of information system or its segments, providing imposition of a false idea about true information technologies and (or) structural and functional characteristics of information system to an adversary.

This problem is most commonly solved by using false network information objects (honeypots) [9], which are designed to distract the adversary from the protected system and collect information about the techniques and tactics used for the attack [10].

The homogeneous structure of modern IP networks gives an attacker an advantage in using the computing resource to carry out computer attacks. This is due to the fact that the homogeneity of network configurations, their hardware and software, allows adversaries to easily and with little computational cost to conduct a large-scale attack on dozens, hundreds or thousands of nodes after successfully conducting a small-scale attack on just one of the nodes of this homogeneous IP-network.

BIT-2021: XI International Scientific and Technical Conference on Secure Information Technologies, April 6-7, 2021, Moscow, Russia
EMAIL: rvmaxim@yandex.ru (A. 1); ssp.vrn@mail.ru (A. 2); telenga@gmail.com (A. 3)
ORCID: 0000-0002-1882-3465 (A. 1); 0000-0002-1396-0284 (A. 2); 0000-0001-6193-0656 (A. 3)


In turn, security services must timely detect and eliminate all potential vulnerabilities, as well as neutralize or reduce the effectiveness of network reconnaissance and implementation of computer attacks. Continuity in time when using traditional methods and means of protection of IP-networks allows the attacker with sufficient ease to calculate the next step of the security system or technological cycle implemented in the IP-network. This technique is called cyber maneuvering [11, 12, 13, 14].

The problem of applying cyber maneuvering techniques is the presence of so-called critical connections in the traffic of the information system, which, for the purpose of data transmission continuity, cannot change their identifiers at one time.

Thus, to achieve the goal of cyber maneuvering and uncompromising operation of protection technologies, false network information objects must generate false (masking) traffic based on the characteristics of real traffic of the information system, which will make noise in the useful network activity and focus the attention of the adversary.

2. Analysis of input data for the model

There are several approaches to describing information system network traffic: as flows and as a sequence of packets ("raw" traffic).

Flows contain header information about network connections between two end devices, such as servers or workstations. Each flow is a collection of transmitted network packets that share some properties. Generally, all transmitted network packets with the same source IP address, source port, destination IP address, destination port and transport protocol within a time window are combined into a single flow [15, 16, 17].

"Raw" traffic typically is a sequence of packets, each containing packet sending time, source IP address, source port, destination IP address, destination port, protocol, packet size, set flags, and a data field in which the payload is written [18, 19, 20].

Based on the above, we can conclude that the parameters that determine the network interaction between the two nodes of the data network of a distributed information system are source IP address, source port, destination IP address, destination port, protocol, packet size, duration of the connection.

Let's take as a reference the network traffic dump recorded from 00:00:00 on 15 March 2021 to 23:59:59 on 21 March 2021. The traffic parameters in the sample under consideration are: date of first session detection, connection duration, protocol, source IP address, source port, destination IP address, destination port, number of packets in session, number of bytes transmitted per session. An example of the entries is shown in Table 1 (some of the columns are hidden), 324533 records in total.

Table 1

Sample of a traffic dump from an local area network

No.	Date first seen	Duration	Proto	Src IP Addr	Src Pt	Dst IP Addr	Dst Pt
0	2021-03-15 00:01:16.632	0.000	TCP	192.168.100.5	445	192.168.220.16	58844
1	2021-03-15 00:01:16.552	0.000	TCP	192.168.100.5	445	192.168.220.15	48888
2	2021-03-15 00:01:16.551	0.004	TCP	192.168.220.15	48888	192.168.100.5	445
3	2021-03-15 00:01:16.631	0.004	TCP	192.168.220.16	58844	192.168.100.5	445
4	2021-03-15 00:01:16.552	0.000	TCP	192.168.100.5	445	192.168.220.15	48888
...	...	...	...	...	...	...	...
324528	2021-03-21 23:59:49.487	0.005	TCP	192.168.220.9	50336	192.168.100.5	445
324529	2021-03-21 23:59:53.710	0.000	TCP	192.168.100.5	445	192.168.220.16	50468

324530	2021-03-21 23:59:53.709	0.002	TCP	192.168.220.16	50468	192.168.100.5	445
324531	2021-03-21 23:59:58.299	0.000	TCP	192.168.100.5	445	192.168.220.6	56281
324532	2021-03-21 23:59:58.298	0.002	TCP	192.168.220.6	56281	192.168.100.5	445

Let us examine which ports were used by the different IP addresses.
The graph for source addresses is shown in Figure 1, and for destination addresses in Figure 2.

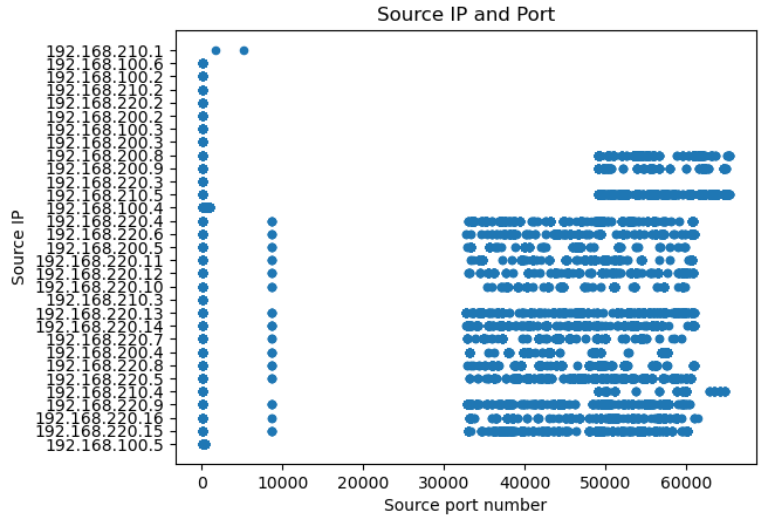


Figure 1: Source IP addresses and used source ports

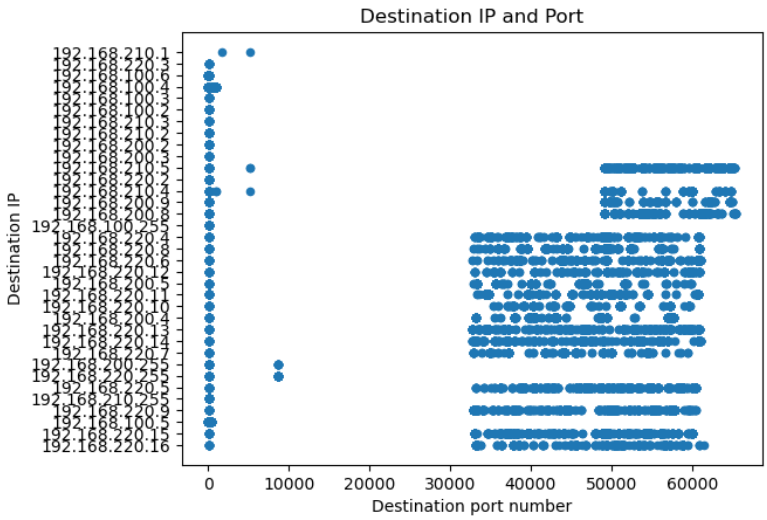


Figure 2: Destination IP addresses and used source ports

In network communication over the TCP transport protocol, clients initiate a connection to servers using the triple handshake algorithm [21], which uses the notion of a socket - an address that includes a port identifier, i.e. the concatenation of an IP address and a TCP port:

1. Sender opens a port to initiate communication. Since administrator privileges are required to access ports below 1000, a number between 10000 and 65535 is usually chosen at random.
2. Sender sends a packet with the SYN flag set to 1 to the recipient. In addition to the SYN flag, the packet header specifies the sender IP, source port, destination IP and destination port. The sender waits for a response from the receiver on the port opened in step 1.

3. If the recipient has a service that uses the port number specified as the destination port, it sends a packet with SYN, ACK flags set to 1 to the source port. Otherwise, it sends a packet with the RST flag.

In Figures 1 and 2 it's easy to see that the IP addresses that are not accessing ports numbered above 10000 are servers that are serving user requests (the large number of reply sessions from these IP addresses on ports above 10000 shows this). This is a de-masking feature that can narrow down the attack vector of an adversary. The goal of a honeypot is to supplement network activity and mask critical network nodes.

Consider now the duration of the sessions, the number of packets and the amount of data transmitted within each session (Figures 3, 4 and 5).

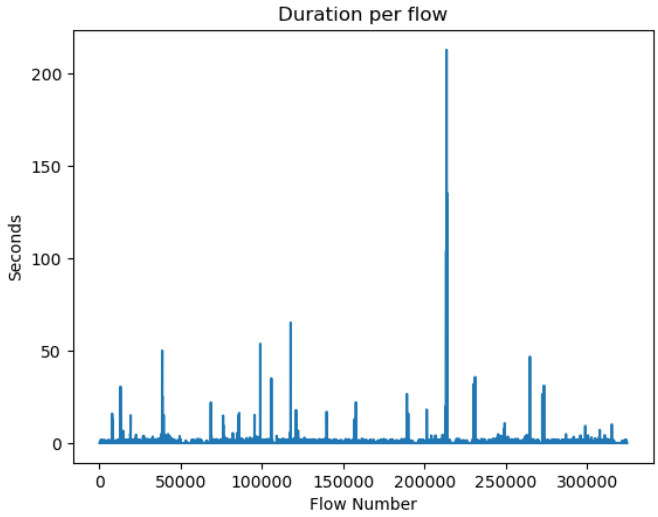


Figure 3: Session duration per flow

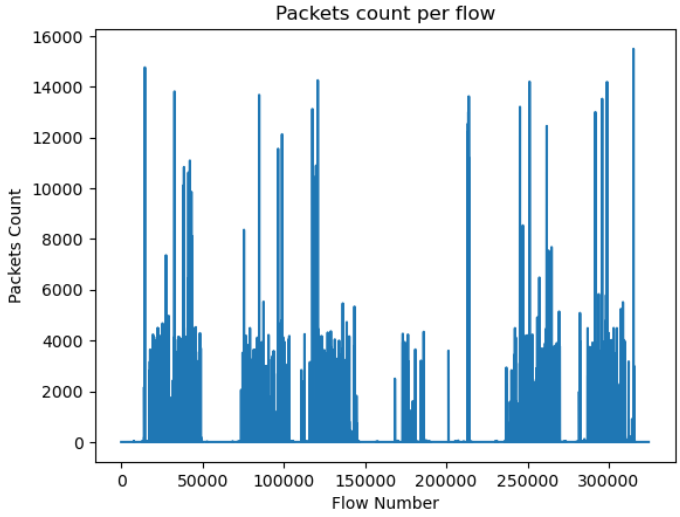


Figure 4: Packets count per flow

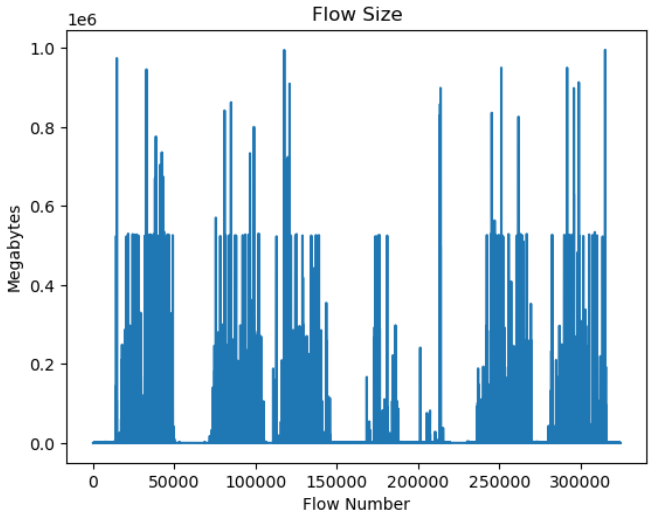


Figure 5: Flow size

3. Model of information system functioning

Let us consider three strategies of information system functioning:

1. Normal mode – the system operates in normal mode, the intensity of traffic does not change.
2. Day/night mode – traffic intensity changes depending on time of day.
3. Critical connections – part of the system degrades due to cyber-attack, traffic of critical connections prevails.

For each of the presented strategies, honeypot must generate its own variant of network traffic to ensure uncompromised functioning of the defenses.

Let us generate an information system model for the first situation.

First we calculate the Hurst index [22] for the time series of source ports for all traffic dump flows using Rescaled range (R/S) analysis [23] (Figure 6).

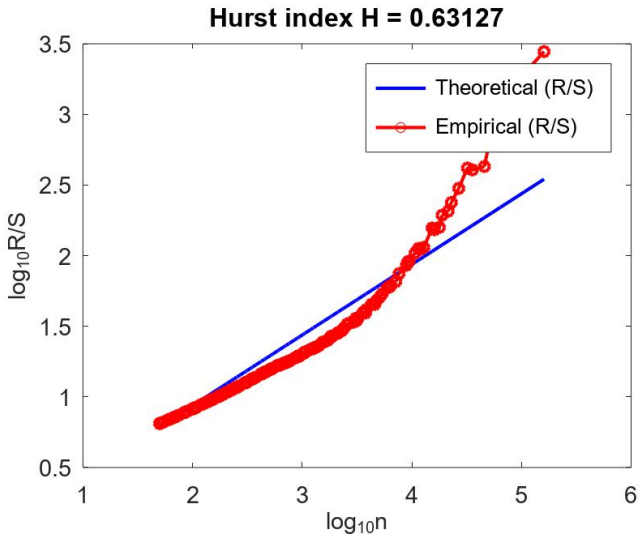


Figure 6: Time dependence R/S in double logarithmic scale and its linear approximation for source ports

After that, we use calculated Hurst index and approximate given time series with the van der Pol nonlinear oscillator equation [24], which has the form $\frac{d^2x}{d^2t} - a(1 - b\frac{dx}{dt}) + x = 0$.

Choosing the coefficients for the practically important case ($a > 0, b > 0$) and solving differential equations with numerical methods, for example, Runge-Kutta of order 4 and 5, we obtain that the closest calculated value of the Hurst index $H_S = 0,63184$ to the value of the Hurst index $H = 0,63127$ of the studied time series of source ports is obtained at $a = 19; b = 20$.

Thus, the processing of the Van der Pol generator model series with the presented coefficients resulted in a dependence that can be considered as a fairly accurate approximation of the empirical series of R/S dependence for a sequence of flows with different source ports, i.e. its mathematical model:

$$\frac{d^2x}{d^2t} - 19(1 - 20\frac{dx}{dt}) + x = 0.$$

Now we move on and calculate Hurst index destination ports time series (Figure 7).

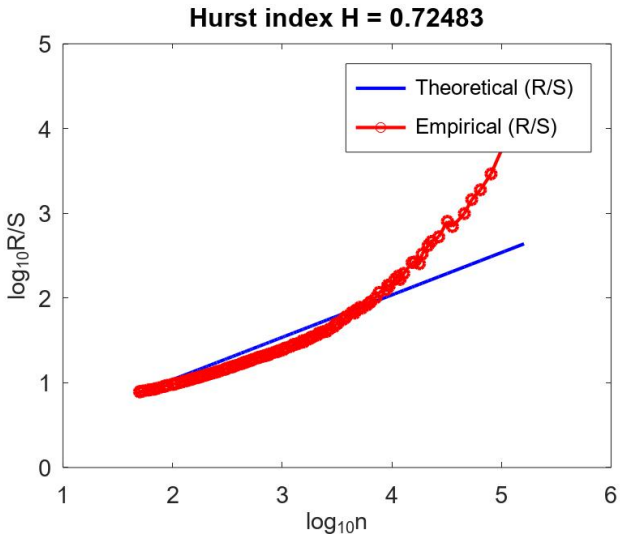


Figure 7: Time dependence R/S in double logarithmic scale and its linear approximation for destination ports

Van der Pol generator model with the closest value of the Hurst index $H_S = 0,72483$ to the value of the Hurst index $H = 0,72483$ has a form

$$\frac{d^2x}{d^2t} - 14.7(1 - 10.1\frac{dx}{dt}) + x = 0.$$

The next one parameter of network traffic is duration of flows. Range scale analysis and Hurst index for given dump is presented on Figure 8.

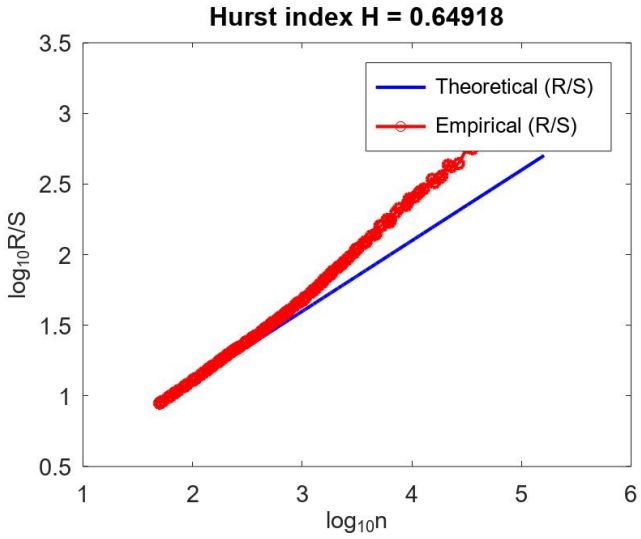


Figure 8: Time dependence R/S in double logarithmic scale and its linear approximation for flows duration

Van der Pol generator model with the closest value of the Hurst index $H_S = 0{,}64897$ to the value of the Hurst index $H = 0{,}64918$ has a form

$$\frac{d^2x}{d^2t} - 10(1 - 10\frac{dx}{dt}) + x = 0 .$$

Now we use Range scale analysis for time series of flows size (Figure 9).

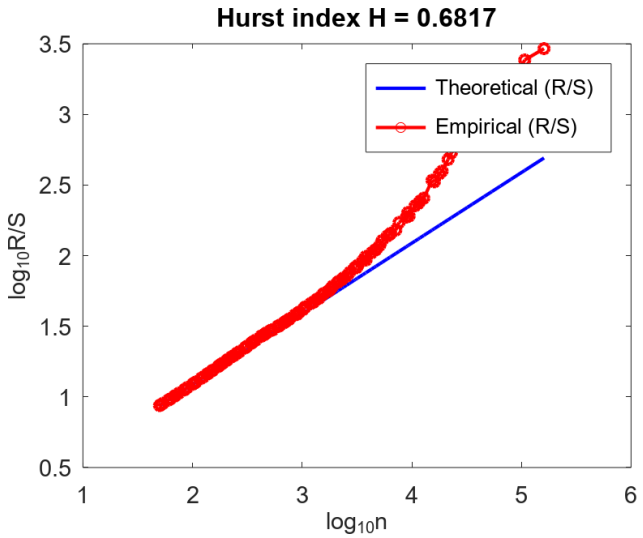


Figure 9: Time dependence R/S in double logarithmic scale and its linear approximation for flows size

Van der Pol generator model with the closest value of the Hurst index $H_S = 0{,}68171$ to the value of the Hurst index $H = 0{,}6817$ has a form

$$\frac{d^2x}{d^2t} - 11(1 - 12.315\frac{dx}{dt}) + x = 0 .$$

Last but not least Range scale analysis for time series of flows packets count (Figure 10).

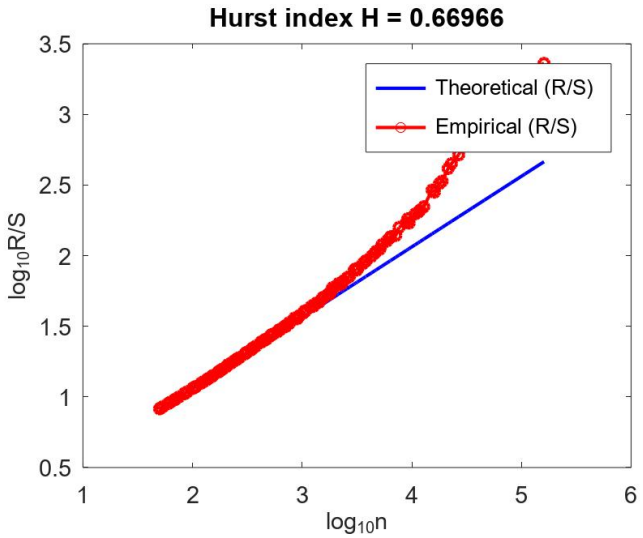


Figure 10: Time dependence R/S in double logarithmic scale and its linear approximation for flows packets count

Van der Pol generator model with the closest value of the Hurst index $H_S = 0{,}66971$ to the value of the Hurst index $H = 0{,}66966$ has a form

$$\frac{d^2x}{d^2t} - 11.3(1 - 11.61\frac{dx}{dt}) + x = 0.$$

So now we have all necessary information to use for honeypot: set of source and destination IP-addresses that could be taken from dump and mathematical models for source and destination ports, flow durations, packets counts and sizes.

Next we examine day/night situation. A sample of network traffic dump for an activity between 08:00:00 and 17:59:59 from 15 March 2021 to 21 March 2021 is shown in Table 2, 171884 records in total.

Table 2

Sample of a traffic dump from a daytime network activity

Date	First seen	Duration	Proto	Src IP Addr	Src Pt	Dst IP Addr	Dst Pt
15.03.2021	08:00:03.047	0.0	TCP	192.168.100.5	445	192.168.220.10	40239
15.03.2021	08:00:03.046	0.002	TCP	192.168.220.10	40239	192.168.100.5	445
15.03.2021	08:00:06.994	0.0	UDP	192.168.210.4	138	192.168.210.255	138
15.03.2021	08:00:08.438	1.501	UDP	192.168.210.4	137	192.168.210.255	137
15.03.2021	08:00:10.766	0.0	UDP	192.168.220.13	137	192.168.220.255	137
15.03.2021	08:00:10.766	0.0	UDP	192.168.220.9	138	192.168.220.255	138
15.03.2021	08:00:10.765	0.0	UDP	192.168.220.10	137	192.168.220.255	137
15.03.2021	08:00:13.569	0.0	TCP	192.168.220.5	51146	192.168.100.5	445
15.03.2021	08:00:13.569	0.0	TCP	192.168.100.5	445	192.168.220.5	51146
15.03.2021	08:00:15.237	0.003	TCP	192.168.100.5	445	192.168.210.4	49159
15.03.2021	08:00:15.235	0.005	TCP	192.168.210.4	49159	192.168.100.5	445

Using a similar methodology described above, we obtain mathematical models for the parameters of the represented traffic.

Source ports:

$$\frac{d^2x}{d^2t} - 9(1 - 10.35\frac{dx}{dt}) + x = 0.$$

Destination ports:

$$\frac{d^2x}{d^2t} - 9(1 - 10.625\frac{dx}{dt}) + x = 0.$$

Connection duration:

$$\frac{d^2x}{d^2t} - 10(1 - 9.389\frac{dx}{dt}) + x = 0.$$

Session size:

$$\frac{d^2x}{d^2t} - 10(1 - 10.2\frac{dx}{dt}) + x = 0.$$

Number of packets per session:

$$\frac{d^2x}{d^2t} - 9.125(1 - 10.585\frac{dx}{dt}) + x = 0.$$

A sample network traffic dump for an activity between 00:00:00 and 07:59:59 from 15 March 2021 to 21 March 2021 is shown in Table 3, 152649 records in total.

Table 3

Sample of a traffic dump from a nighttime network activity

Date	First seen	Duration	Proto	Src IP Addr	Src Pt	Dst IP Addr	Dst Pt
15.03.2021	00:01:16.632	0.0	TCP	192.168.100.5	445	192.168.220.16	58844

15.03.2021	00:01:16.552	0.0	TCP	192.168.100.5	445	192.168.220.15	48888
15.03.2021	00:01:16.551	0.004	TCP	192.168.220.15	48888	192.168.100.5	445
15.03.2021	00:01:16.631	0.004	TCP	192.168.220.16	58844	192.168.100.5	445
15.03.2021	00:01:16.552	0.0	TCP	192.168.100.5	445	192.168.220.15	48888
15.03.2021	00:01:16.631	0.004	TCP	192.168.220.16	58844	192.168.100.5	445
15.03.2021	00:01:17.432	0.0	TCP	192.168.220.9	37884	192.168.100.5	445
15.03.2021	00:01:17.431	0.0	TCP	192.168.100.5	445	192.168.220.9	37884
15.03.2021	00:01:17.432	0.0	TCP	192.168.220.9	37884	192.168.100.5	445
15.03.2021	00:01:17.827	0.0	UDP	192.168.210.4	138	192.168.210.255	138
15.03.2021	00:01:21.703	0.0	TCP	192.168.100.5	445	192.168.220.5	51146

Model for the nighttime network activity as follows.

Source ports:

$$\frac{d^2x}{d^2t} - 8(1 - 7.95\frac{dx}{dt}) + x = 0 .$$

Destination ports:

$$\frac{d^2x}{d^2t} - 7.6(1 - 9.55\frac{dx}{dt}) + x = 0 .$$

Connection duration:

$$\frac{d^2x}{d^2t} - 9.7(1 - 10.755\frac{dx}{dt}) + x = 0 .$$

Session size:

$$\frac{d^2x}{d^2t} - 8(1 - 10.523\frac{dx}{dt}) + x = 0 .$$

Number of packets per session:

$$\frac{d^2x}{d^2t} - 8(1 - 11.24\frac{dx}{dt}) + x = 0 .$$

And now we make a model for the last situation, when we need to degrade some services or make a cyber-maneuver due a cyber-attack and critical connections prevails in a network traffic as a result, 274204 records in total (Table 4).

Table 4

Sample of a traffic dump from a critical connections network activity

Date	First seen	Duration	Proto	Src IP Addr	Src Pt	Dst IP Addr	Dst Pt
15.03.2021	00:01:16.632	0.0	TCP	192.168.100.5	445	192.168.220.16	58844
15.03.2021	00:01:16.552	0.0	TCP	192.168.100.5	445	192.168.220.15	48888
15.03.2021	00:01:16.551	0.004	TCP	192.168.220.15	48888	192.168.100.5	445
15.03.2021	00:01:16.631	0.004	TCP	192.168.220.16	58844	192.168.100.5	445
15.03.2021	00:01:16.552	0.0	TCP	192.168.100.5	445	192.168.220.15	48888
15.03.2021	00:01:16.631	0.004	TCP	192.168.220.16	58844	192.168.100.5	445
15.03.2021	00:01:17.432	0.0	TCP	192.168.220.9	37884	192.168.100.5	445
15.03.2021	00:01:17.431	0.0	TCP	192.168.100.5	445	192.168.220.9	37884
15.03.2021	00:01:17.432	0.0	TCP	192.168.220.9	37884	192.168.100.5	445
15.03.2021	00:01:21.703	0.0	TCP	192.168.100.5	445	192.168.220.5	51146
15.03.2021	00:01:21.701	0.003	TCP	192.168.220.5	51146	192.168.100.5	445

A model of information system functioning with the prevalence of critical connections is presented below.

Source ports:

$$\frac{d^2x}{d^2t} - 8(1 - 11\frac{dx}{dt}) + x = 0 .$$

Destination ports:

$$\frac{d^2x}{d^2t} - 8(1 - 11.355\frac{dx}{dt}) + x = 0 .$$

Connection duration:

$$\frac{d^2x}{d^2t} - 11(1 - 12\frac{dx}{dt}) + x = 0 .$$

Session size:

$$\frac{d^2x}{d^2t} - 14(1 - 11.405\frac{dx}{dt}) + x = 0 .$$

Number of packets per session:

$$\frac{d^2x}{d^2t} - 14(1 - 10.005\frac{dx}{dt}) + x = 0 .$$

4. Conclusions

The developed models allows to determine the probabilistic and temporal characteristics describing the states of the client-server information system functioning process at various strategies of establishment and maintenance of connection parameters by the interacting parties, which allows to estimate the state of the client-server information system and to adjust the parameters of protection systems against network reconnaissance.

The novelty of the developed model consists in the application of modified algorithms of fractal analysis to assess the characteristics of network traffic to improve the reliability, accuracy and validity of honeypots.

5. References

[1] P. Dymora, M. Mazurek, Influence of Model and Traffic Pattern on Determining the Self-Similarity in IP Networks. Applied Sciences, 11, (2021), 190. doi:10.3390/app11010190.

[2] A. Guerrero-Ibanez, J. Contreras-Castillo, R. Buenrostro, A. B. Marti, and A. R. Munoz, A policy-based multi-agent management approach for intelligent traffic-light control, IEEE Intelligent Vehicles Symposium, University of California, San Diego, USA, June 2010. doi:10.1109/IVS.2010.5548133.

[3] A. Bhattacharjee, S. Nandi, Statistical analysis of network traffic inter-arrival, 2010 The 12th International Conference on Advanced Communication Technology (ICACT), 2010, pp. 1052-1057.

[4] Z. Fang, J. Wang, B. Liu, and W. Gong. Double pareto lognormal distributions in complex networks, Handbook of Optimization in Complex Networks, 2011, pp. 55–80, doi:10.1007/978-1-4614-0754-6.

[5] A. Ghosh, R. Jana, V. Ramaswami, J. Rowland, and N. K. Shankaranarayanan. Modeling and characterization of large-scale wi-fi traffic in public hot-spots. In INFOCOM, 2011. doi:10.1109/INFCOM.2011.5935132.

[6] O. I. Sheluhin, S. M. Smolskiy and A. V. Osin, Self-Similar Processes in Telecommunications, Wiley, London, 2007.

[7] Voronchikhin I., Ivanov ., Maximov R., Sokolovsky S. Masking of distributed information systems structure in cyberspace. Voprosy kiberbezopasnosti, 2019, No 6, pp. 92-101. DOI: 10.21681/2311-3456-2019-6-92-101. (In Russ.)

[8] Kuchurov V., Maximov R., Sherstobitov R. Model and technique for abonent address masking in cyberspace. Voprosy kiberbezopasnosti, 2020, No 6 (40), pp. 2-13. DOI: 10.21681/2311-3456-2020-06-2-13. (In Russ.)

[9] N. Bhagat and B. Arora, "Intrusion Detection Using Honeypots," 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), 2018, pp. 412-417, doi: 10.1109/PDGC.2018.8745761.
[10] R. Kwon, T. Ashley, J. Castleberry, P. Mckenzie and S. N. Gupta Gourisetti, Cyber Threat Dictionary Using MITRE ATT&CK Matrix and NIST Cybersecurity Framework Mapping, 2020 Resilience Week (RWS), 2020, pp. 106-112, doi:10.1109/RWS50334.2020.9241271.
[11] S. Applegate, "The principle of maneuver in cyber operations," in In 2012 4th International Conference on Cyber Conflict (CYCON 2012), 2012
[12] Allen, Patrick D. "Cyber Maneuver and Schemes of Maneuver: Preliminary Concepts, Definitions, and Examples." The Cyber Defense Review, vol. 5, no. 3, 2020, pp. 79–98. JSTOR, www.jstor.org/stable/26954874.
[13] P. Beraud, A. Cruz, S. Hassell and S. Meadows, "Using cyber maneuver to improve network resiliency," 2011 - MILCOM 2011 Military Communications Conference, 2011, pp. 1121-1126, doi: 10.1109/MILCOM.2011.6127449.
[14] P. Beraud, A. Cruz, S. Hassell, J. Sandoval and J. J. Wiley, "Cyber defense Network Maneuver Commander," 44th Annual 2010 IEEE International Carnahan Conference on Security Technology, 2010, pp. 112-120, doi: 10.1109/CCST.2010.5678724.
[15] S. Choudhary, "Usage of Netflow in Security and Monitoring of Computer Networks," Interntional Journal of Computer Applications, vol. 68, no. 24, pp. 17–24, 2013, doi:10.5120/11727-7362.
[16] R. Hofstede, P. Celeda, B. Trammell, I. Drago, R. Sadre, A. Sperotto, and A. Pras, "Flow Monitoring Explained : From Packet Capture to Data Analysis with NetFlow and IPFIX," IEEE Communications Surveys & Tutorials, vol. 16, no. 4, 2014, doi: 10.1109/COMST.2014.2321898.
[17] M. Marchetti, F. Pierazzi, M. Colajanni, and A. Guido, "Analysis of high volumes of network traffic for Advanced Persistent Threat detection," Computer Networks, vol. 0, pp. 1–15, 2016, doi:10.1016/j.comnet.2016.05.018
[18] B. Alothman, "Raw Network Traffic Data Preprocessing and Preparation for Automatic Analysis," 2019 International Conference on Cyber Security and Protection of Digital Services (Cyber Security), 2019, pp. 1-5, doi: 10.1109/CyberSecPODS.2019.8885333.
[19] J. J. Davis and A. J. Clark, "Data preprocessing for anomaly based network intrusion detection: A review," Comput. Secur., vol. 30, no. 6-7, pp. 353–375, 2011, doi:10.1016/j.cose.2011.05.008
[20] Michael J. De Lucia, Paul E. Maxwell, Nathaniel D. Bastian, Ananthram Swami, Brian Jalaian, and Nandi Leslie "Machine learning raw network traffic detection", Proc. SPIE 11746, Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications III, 117460V, 2021, doi:10.1117/12.2586114
[21] J. Postel, "Transmission Control Protocol," IETF RFC 793, September 1981.
[22] H. Hurst, R. Black, Y. Simaika, Long-Term Storage: An Experimental Study, Constable, London, 1965
[23] Raimundo, M.S.; Okamoto, J., Jr. Application of Hurst Exponent (H) and the R/S Analysis in the Classification of FOREX Securities. Int. J. Model. Optim. 2018, 8, 116–124.
[24] J. He, J. Cai, Design of a New Chaotic System Based on Van Der Pol Oscillator and Its Encryption Application, Mathematics, 2019, 7, 743. doi:10.3390/math7080743.