Methodology for Substantiating the Characteristics of False Network Traffic to Simulate Information Systems

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

Simulation of false information traffic to protect the structural and functional characteristics of the information system is not an easy task in view of self-similarity of its statistical properties in IP-networks not only in the current moment, but also retrospectively. We have developed a methodology to substantiate the characteristics of false network traffic to simulate information systems, allowing to solve the problem of maximum likelihood of false network traffic by pseudophase reconstruction of the dynamic system attractor, which approximates the time series of the protected object information traffic.

Keywords

Information protection, false network information objects, Hurst index, de-masking features, pseudo-phase reconstruction

1. Introduction

Information protection measures in state information systems currently include:

- hiding the architecture and configuration of the information system;
- creation (emulation) of false information systems or their components designed to detect, register and analyze the actions of intruders in the process of implementing threats to information security;
- reproduction of false and (or) concealment of true individual information technologies and (or) structural and functional characteristics of the information system or its segments, ensuring the imposition of a false idea on the offender about the true information technologies and (or) structural and functional characteristics of the information system.

This is due to the fact that a sufficiently large number of computer attacks are reconnaissance in nature in order to obtain information about the composition, structure and algorithms of the functioning, location and ownership of information systems, as well as data stored, processed and transmitted in such systems. Along with the threats to information security, related to the dialog interaction of the intruder and information system (in particular – automated network scanning tools), the reconstruction of structural and functional characteristics of information system is aimed at the threat of determining its topology, uncompromisingly implemented by the analysis of network traffic. The result is revealing the topology of the cyberspace distributed information system, determining the importance of its nodes witch could be used by an intruder to implement planned APT-attacks (advanced persistent threat, targeted cyberattack) [1].

The task of implementing the above information protection measures is solved by false (masking) information traffic, which is understood as a set of false (masking) message packets formed by network information objects in order to manage the demasking signs of information systems functioning algorithms: the intensity of traffic between topologically localized network information objects of a distributed information system, network interaction protocols and hierarchical levels (ranks) of its elements.

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)

For this purpose, false network information objects (also called "deceptive" systems) are used, which implement dialog interaction with the intruder, which leads to "depletion" of its computing resources and compromises the processes of tampering.

2. Methodology description

It is known the way to protect computer networks [2, 3, 4], which implements the technique of cyber maneuver - periodic (time-synchronized) or uncontrolled (random) change of network settings of the protected system (used address space and port numbers of subscribers) -, where upon detection of an intruder the DHCP server forcibly stops leasing current IP addresses by legitimate subscribers of the information system and sends them new network settings containing IP addresses from a different subnet not known to the intruder in advance. This task is complicated by the presence in the information system of critical applications and network traffic between them (so called "critical connections"), interruption of which is undesirable or impossible. As a result of network intelligence, they will be vulnerable to computer attacks. In order to eliminate this threat, it is not just necessary to transfer subscribers (network information objects) in another subnetwork, but also to "load" the network information objects of the compromised configuration with the functions of false objects. We need to maintain false information traffic between them, which has statistical characteristics of the compromised information system, so that the cyber maneuver will not be detected by an intruder.

The following variants of solving the problem of false (masking) information traffic are possible.

Pre-recording of information system network traffic and subsequent sending of saved packets to the network. In this case, the time interval between packets is taken from the network traffic record. Disadvantages of this approach, associated with the need to store large amounts of data, are obvious. In addition, with a relatively small number of subscribers in the information system it is possible to detect the fact of re-use (cloning) of traffic.

Another, and more preferable, method is the generation of false (masking) information traffic based on the characteristics of real traffic, which is performed by a false network information object. In order to ensure maximum plausibility of false network traffic, its statistical properties must match the statistical properties of the information traffic of the protection object. Otherwise the application of such protection measures will be compromised, and the goals of simulation will not be achieved.

Thus, there is a contradiction between the need to implement an effective masking exchange and the lack of a unified way to assess the characteristics of network traffic for the false information exchange generation in the system.

Simulation of false information traffic for protection of information system structural and functional characteristics is a complex task in view of self-similarity of its statistical properties in IP-networks [5, 6, 7, 8] not only in the current moment, but also retrospectively. This means that some property of the object is preserved when scaling space and/or time. Otherwise, they say that there is a repeatability of statistical characteristics of natural time series with the change of scale.

It is known [9, 10, 11, 12, 13, 14] that the processes with self-similarity properties are characterized by the presence of aftereffects due to the factors that cause complex dependencies. The resulting traffic (process in general case) becomes "pulsating": large bursts of intensity are possible at relatively low average rate of arrival of message packets in an information system. Statistical characteristics of such a process are the de-masking features of a particular system. [15, 16, 17, 18].

Today, global networking is growing exponentially, with traffic statistics that mathematically exhibit fractal characteristics: self-similarity and long-range dependence. With these properties, data traffic shows high peak-to-average bandwidth ratios and causes data networks inefficient. These problems make it difficult to predict, quantify, and control data traffic, in contrast to the traditional Poisson-distributed traffic in telephone networks [19, 20].

A commonly used index of self-similarity of a process is the Hurst index H, initially introduced in [21], depending on its values the following conclusions about the processes under study are drawn:

- at $0 \le H \le 0,5$ is a random process, it has no self-similarity;
- at $H > 0,5$ the process has a long memory and is self-similar [22].

The Hearst index calculations use the algorithm for analyzing the adjusted modified range proposed in [23], [24], which consists in the following.

Let the time series be given

$$Z = \{z_i\}, i = 1,2,...,n. \quad (1)$$

in which its initial segments are sequentially allocated

$$Z_\tau = z_1, z_2, ..., z_\tau, \text{ where } \tau = 3,4,...,n, \quad (2)$$

for each of which the current average is calculated

$$\overline{z}_\tau = \frac{1}{\tau}\sum_{i=1}^{\tau} z_i \quad (3)$$

Next, for each fixed $z_\tau, \tau = 3,4,...,n$, calculate the accumulated deviation for each of the segments of length t :

$$X(t,\tau) = \sum_{i=1}^{t}(z_i - \overline{z}_\tau), \text{ where } t = \overline{1,\tau} \quad (4)$$

The main characteristic of a sample of a random process is the normalized range R/S, where

$$R(\tau) = \max_{1\le t\le\tau} X(t,\tau) - \min_{1\le t\le\tau} X(t,\tau) \quad (5)$$

maximum amplitude range of the random process, S is standard deviation of the process

$$S = S(\tau) = \sqrt{\frac{1}{\tau}\sum_{j=1}^{\tau}(z_j - \overline{z}_\tau)^2}, \quad (6)$$

t - discrete time with integer values; τ - duration of the time interval in question.

The normalized R/S spread is described by the empirical relation $R/S = (\tau/2)^H$, where H is the Hurst index.

We obtain the Cartesian coordinates of the trajectory points (x_τ, y_τ) by logarithmization of both parts of this equality (6), whose ordinates and abscissas are, respectively:

$$y_\tau = H(\tau) = \frac{\log(R(\tau)/S(\tau))}{\log(\tau/2)}, x_\tau = \tau. \quad (7)$$

The R/S-trajectory required for the fractal analysis of the series is represented in Cartesian logarithmic coordinates by a sequence of points, the abscissas and ordinates of which are as follows $x_\tau = \log(\tau/2)$, $y_\tau = \log(R(\tau)/S(\tau))$.

Let a sample (dump) of information system traffic X_t for some set of time moments $t \le T$ be obtained at time T. Then the model of prediction of characteristics of information traffic defines a set of output variables $\hat{X}_{T+\tau}$, which can be expressed in vector form (time $\tau > 0$). In general, the expression for the model (regression equation or regression in terms of mathematical statistics) is written as

$$\hat{X}_{T+\tau} = \mathbf{a}_T \mathbf{F}(\tau), \quad (8)$$

where vector $\mathbf{a}_T$ represents the model coefficients derived from the results of traffic dumping up to and including moment T, and matrix $\mathbf{F}$ represents the set of approximating functions.

In most cases, it is necessary to study only the attractor, a compact subset of the phase space to which the evolution trajectories of all points of the system located near this subset are asymptotically "attracted", in order to analyze the behavior of the dynamical system (8). Its dimensionality determines the amount of information required to specify the coordinates of a point belonging to the attractor within the specified accuracy.

The fractal dimension D can be expressed through the Hearst index H by the ratio

$$D = 2 - H. \quad (9)$$

The attractor is related to the fractal dimension through the correlation integral $C(r)$, which is estimated directly for a sequence of points (shows the relative number of pairs of points at a distance not greater than r):

$$C(r) = \lim_{m\to\infty} \frac{1}{m(m-1)} \sum_{i=1}^{m}\sum_{j=1}^{m} \theta(r - \rho(x_i, x_j)), \quad (10)$$

where

$$\theta(\alpha)=\begin{cases}1,\alpha\geq 0,\\ 0,\alpha<0\end{cases} \tag{11}$$

(Haviside function), ρ is the distance between a pair of points in n-dimensional phase space, and m is the number of points x_i on the attractor.

Tackens in [25] showed that for almost every smooth dynamical system it is possible to calculate the correlation integral and the fractal dimension by measurements of only one of the phase coordinates of this system.

The method for synthesizing a mathematical model of a process described in [26] is based on the application of the so-called pseudophase reconstruction.

A pseudophase reconstruction [27] is a mapping that maps the $x(t)$ point of a time series to the $[x(t),x(t+\tau),...,x(t+(m-1)\cdot\tau)]\in R^m$ point, where t is the discrete time $(t=(\overline{(m-1)\tau+1),N}))$, τ is the time delay (in time discretes), and m is the dimension of the embedding space. Thus, it is possible to reconstruct the original attractor in the point space with delay $[x(t),x(t+\tau),...,x(t+(m-1)\cdot\tau)]$ for an initial set of measurements of the phase coordinate $x(1),x(2),...,x(N),$, where N is the number of measurements, so that it preserves the essential topological properties and dynamics of the original attractor. The attractor dimension m is determined by the formula $m\geq 2[d]+1$, where d is the fractal dimension of the attractor.

Consequently, in order to synthesize a mathematical model of the network traffic of the system under study, it is necessary to calculate the Hurst exponent H for one of the parameters of the network traffic dump and then, using the known mathematical models of attractors, select their coefficients so that the Hurst exponent H_S of the synthesized time series coincides with H with an accuracy of some ε.

3. Example of methodology application

Let us consider a traffic dump of 211972 packets from an Internet subscriber point (Table 1). We want to synthesize a mathematical model of this dump to predict, quantify, and control destination ports of data traffic.

Table 1
Sample of a traffic dump from an Internet subscriber point

No.	Time	Source	Destination	Source Port	Dst Port	Protocol	Length
1	0.000000	91.233.219.10	91.210.45.177	80	40914	TCP	1494
2	-0.000739	91.210.45.177	91.233.219.10	40914	80	TCP	66
3	-0.000688	91.210.45.177	91.233.219.10	40914	80	TCP	66
4	-0.000561	185.24.47.1	81.30.125.67	80	38639	HTTP	1494
5	-0.000539	91.214.126.34	91.210.45.179	80	46104	TCP	1494
6	-0.000449	91.214.126.34	91.210.45.179	80	46104	TCP	1494
7	-0.000388	91.214.126.34	91.210.45.179	80	46104	TCP	1494
8	-0.000390	91.233.219.10	91.210.45.177	80	40914	TCP	1494
9	-0.000325	91.233.219.10	91.210.45.177	80	40914	TCP	1494
10	-0.000263	50.7.232.10	91.210.45.182	80	53031	TCP	1494
11	-0.000261	91.233.219.10	91.210.45.177	80	40914	TCP	1494
12	-0.000248	91.210.45.177	91.233.219.10	40914	80	TCP	66
13	-0.000203	91.233.219.10	91.210.45.177	80	40914	TCP	1494
14	-0.000117	50.7.232.10	91.210.45.182	80	53031	TCP	1494
15	0.001378	91.214.126.34	91.210.45.179	80	46104	TCP	1494
16	0.001577	91.210.45.177	91.233.219.10	40914	80	TCP	66
17	0.001628	91.210.45.177	91.233.219.10	40914	80	TCP	66

18	0.001678	91.210.45.177	91.233.219.10	40914	80	TCP	66
19	0.001726	91.233.219.10	91.210.45.177	80	40914	TCP	1494
20	0.001736	91.210.45.177	91.233.219.10	40914	80	TCP	66

We filter out packages by destination port and present them as a time series (Figure 1). Then we calculate R/S spread and Hurts index using the algorithm above.

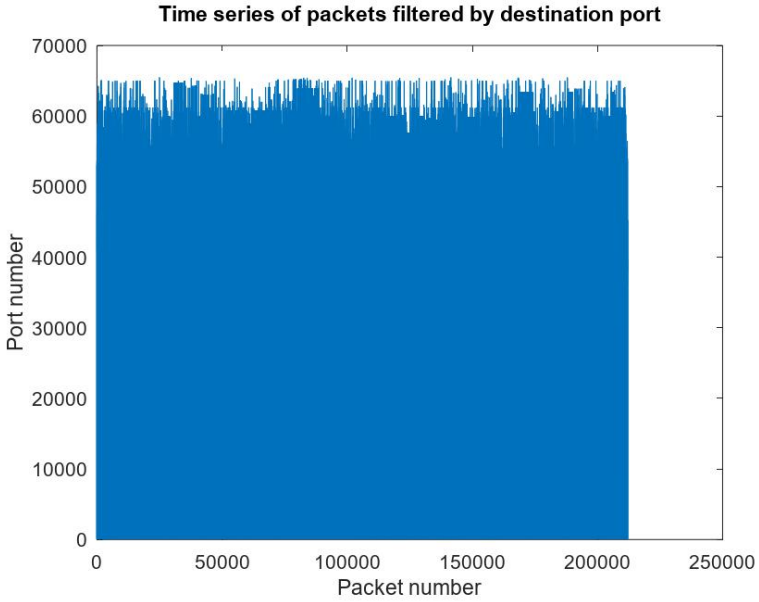

Figure 1: Time series of packets filtered by destination port

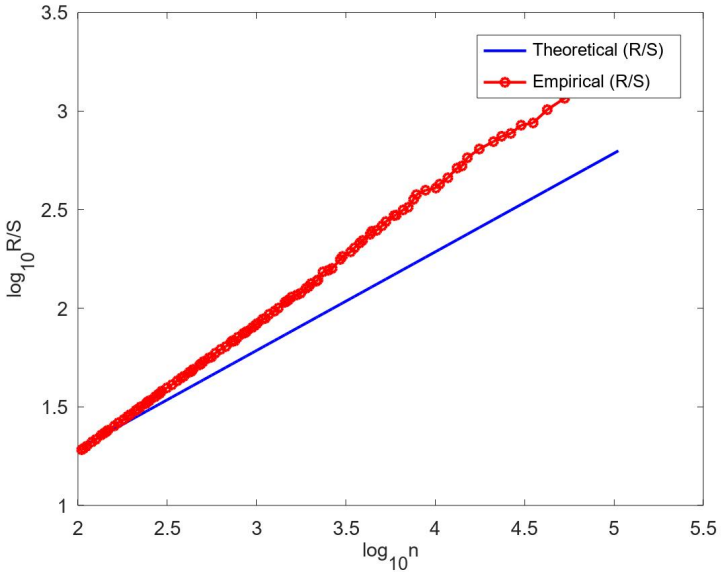

Figure 2: Time dependence R/S in double logarithmic scale and its linear approximation for packets filtered by destination port

The results of calculating the time dependence of the normalized R/S spread in double logarithmic scale and its linear approximation for packets filtered by destination port are shown in Figure 2.

Hearst index for the studied time series $H = 0,67199$, fractal dimension $D = 2 - H = 1,32801$.

Let us approximate the time series under study, for example, by the van der Pol nonlinear oscillator equation [28], 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 $H_S = 0,67199$ to the value of the Hurst index $H = 0,67199$ of the studied time series of packages filtered by destination ports is obtained at $a = 8,514$; $b = 10$.

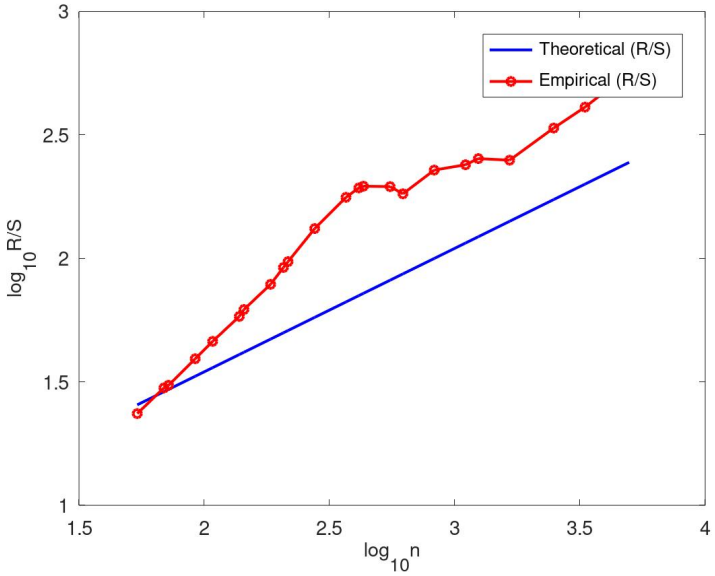

Figure 3: Model R/S dependence in double logarithmic scale and its linear approximation for packets filtered by destination ports (van der Pol)

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 packets with different destination ports, i.e. its mathematical model:

$$\frac{d^2x}{d^2t} - 8,514(1 - 10\frac{dx}{dt}) + x = 0,$$

$$H_S = 0,67199\,, D = 2 - H_S = 1,32801\ .$$

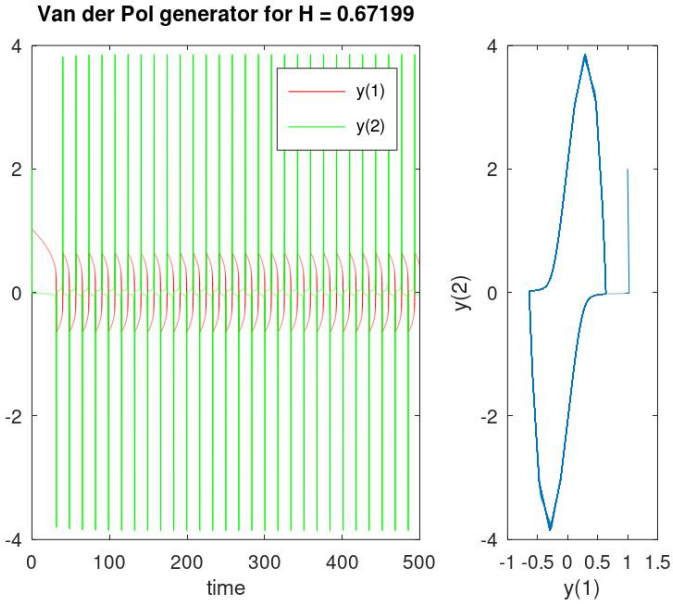

Figure 4: Phase portrait of the Van der Pol oscillator model series with the presented coefficients

Now let's examine packages filtered by source port and present them as a time series (Figure 1). Then we calculate R/S spread and Hurts index.

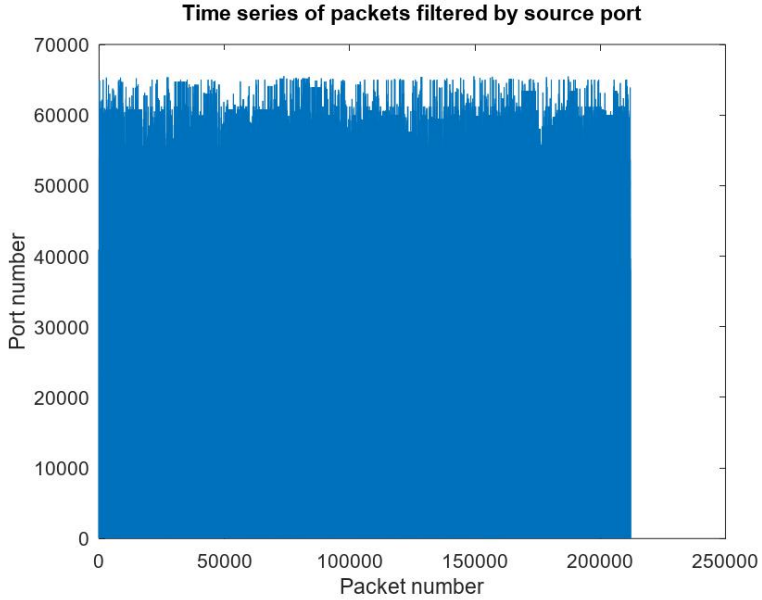

Figure 5: Time series of packets filtered by source port

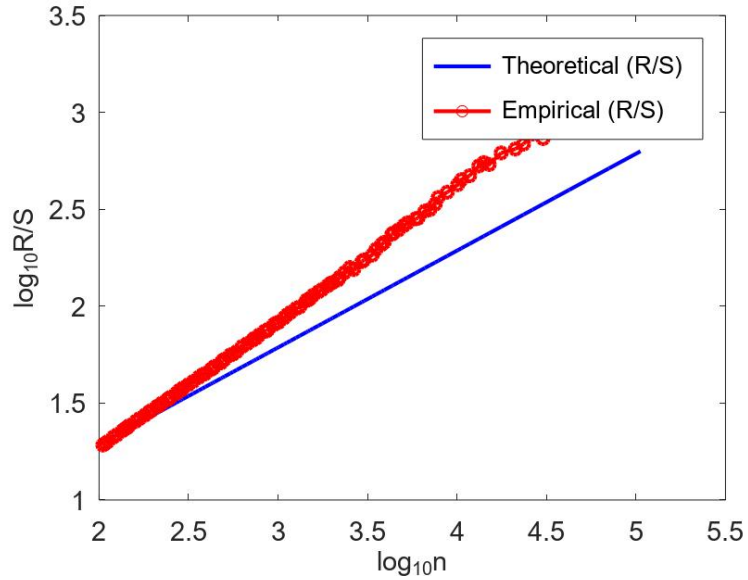

Figure 6: Time dependence R/S in double logarithmic scale and its linear approximation for packets filtered by source port

The results of calculating the time dependence of the normalized R/S spread in double logarithmic scale and its linear approximation for packets filtered by destination port are shown in Figure 6.

Hearst index for the studied time series $H = 0,66513$, fractal dimension $D = 2 - H = 1,33487$.

Let us approximate the time series of source port packages by the Rössler system [29], which has the form

$$\begin{aligned} \dot{x} &= -y - z \\ \dot{y} &= x + ay \\ \dot{z} &= b - cz + xz \end{aligned} \tag{12}$$

Iterate the confidents a, b, and c and solving the system of differential equations with numerical methods using GNU Octave lsode solver we obtain that the closest calculated value of the Hurst index

$H_S = 0,66881$ to the value of the Hurst index $H = 0,66513$ of the studied time series of packages filtered by source ports is obtained at $a = 0,2$, $b = 0,2$, $c = 5,2$.

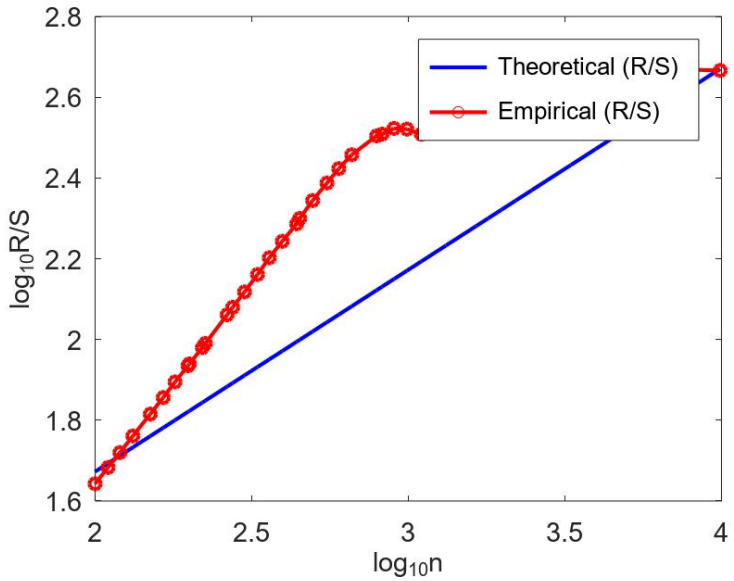

Figure 7: Model R/S dependence in double logarithmic scale and its linear approximation for packets filtered by source ports (Rössler)

Thus, the processing of the Rössler system 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 packets with different source ports, i.e. its mathematical model:

$$\begin{aligned} \dot{x} &= -y - z \\ \dot{y} &= x + 0,5 \cdot y \\ \dot{z} &= 0,5 - 5,2 \cdot z + xz \end{aligned},$$

$H_S = 0,66881, D = 2 - H_S = 1,33119$.

Rossler system strange attractor for H = 0.66881

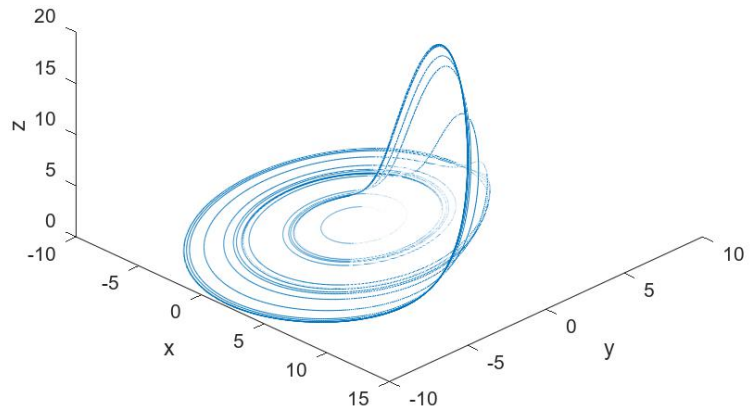

Figure 8: 3D phase portrait of the Rössler system model series with the presented coefficients

The resulting mathematical models can be used to adjust honeypot parameters, allowing the synthesis of false network traffic to different ports that is statistically similar to the reference traffic.

4. Conclusions

Thus, the paper substantiates the possibility of analyzing the information system de-masking features and generation of false network traffic, statistically similar to the traffic of the protected information system, to reduce the effectiveness of network reconnaissance.

The novelty of the developed methodology consists in the application of modified algorithms of fractal analysis to assess the characteristics of network traffic and synthesis of its mathematical model to improve the reliability of false network information objects.

The implementation of the proposed scientific solutions in the information system structure will reduce the availability of its elements and management processes, which will ensure the weakening of the influence or neutralization of network reconnaissance, as well as deprive the attacker the necessary information about the structure of a distributed information system.

5. References

[1] 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.

[2] R. Maksimov, S. Sokolovsky, I. Voronchikhin, A. Gritschin, M. Bodiakin, A. Ignatenko, 2020. Patent No. RU 2726900, Filed November 9th, 2019, Issued July 16th, 2020.

[3] 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.)

[4] 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.)

[5] W. Leland, M. Taqqu, W. Willinger, D. Wilson, On the self-similar nature of Ethernet traffic, IEEE/ACM Transactions on Networking 2(1994) 1-15. doi:10.1109/90.282603

[6] K. Park, W. Willinger, Self-similar network traffic: an overview, in K. Pack, W. Willinger (Ed.) Self-Similar Network Traffic and Performance Evaluation, John Wiley & Sons, New York, 2000, pp. 1. doi:10.1002/047120644X

[7] K. Park, G. Kim, M.E. Crovella, The protocol stack and its modulating effect on self-similar traffic, in K. Pack, W. Willinger (Ed.) Self-Similar Network Traffic and Performance Evaluation, John Wiley & Sons, New York, 2000, pp. 349. doi:10.1002/047120644X

[8] K. Park, G. Kim, M.E. Crovella, On the relationship between file sizes, transport protocols, and self-similar network traffic, in Proceedings of the Fourth International Conference on Network Protocols (ICNP'96), Columbus, OH, 1996, pp. 171-180. doi:10.1109/ICNP.1996.564935

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] L. Yu. Queuing theory with heavy tails and network traffic modeling. 2018. ffhal-01891760f.
[16] L. Yi, S. Tian, L. Le-Jian, Construction of C-ON/OFF network traffic model based on time series, Journal of Intelligent & Fuzzy Systems, 34,2, (2018), 933-943. doi:10.3233/JIFS-169387.
[17] J. Sewall, D. Wilkie, P. Merrell, and M. C. Lin, Continuum traffic simulation, Comput. Graph. Forum, 2010, pp. 439-448.
[18] J.S. Al-Azzeh, M. Al Hadidi, R. Odarchenko, S. Gnatyuk, Z. Shevchuk, Z. Hu, Analysis of Self-Similar Traffic Models in Computer Networks; networks, 1,24, (2017), 20-17 doi:10.15866/iremos.v10i5.12009.
[19] Markov A., Barabanov A., Tsirlov V. Models for Testing Modifiable Systems. In Book: Probabilistic Modeling in System Engineering, by ed. A.Kostogryzov. IntechOpen, 2018, Chapter 7, pp. 147-168. DOI: 10.5772/intechopen.75126.
[20] Markov A., Markov G., Tsirlov V. Simulation of Software Security Tests by Soft Computational Methods. In Proceedings of the VIth International Workshop 'Critical Infrastructures: Contingency Management, Intelligent, Agent-Based, Cloud Computing and Cyber Security' (IWCI 2019). Advances in Intelligent Systems Research, 2019, vol. 169, pp. 257-261. DOI: 10.2991/iwci-19.2019.45.
[21] H. Hurst, R. Black, Y. Simaika, Long-Term Storage: An Experimental Study, Constable, London, 1965.
[22] L.P. Das, S.K. Patra, S. Mishra, Impact of Hurst parameter value in self-similarity behaviour of network traffic, IJRCCT, 5,12, (2017), 631-633
[23] A. Anis, The expected value of the adjusted rescaled Hurst range of independent normal summands, in: A. A. Anis, E. H. Lloyd, Biometrica. No. 63, 1976, pp. 283-298. doi: 10.2307/2335090
[24] E. Peters, Fractal Market Analysis: Applying Chaos Theory to Investment and Economics, Wiley, New York, NY, 1994.
[25] F. Takens, Detecting Strange Attractors in Turbulence, in: F. Takens, Dynamical Systems and Turbulence, volume 898 of Lecture Notes in Mathematics, Springer-Verlag, Berlin, 1981, pp. 366-381
[26] A. Davydov, R. Maksimov, O. Savitsky. Zashchita i bezopasnost' vedomstvennyh integrirovannyh infokommunikacionnyh sistem, Voentelecom, Moscow, 2015.
[27] Matilla-García, M.; Morales, I.; Rodríguez, J.M., Ruiz Marín, M. Selection of Embedding Dimension and Delay Time in Phase Space Reconstruction via Symbolic Dynamics. Entropy 2021, 23, 221. doi:10.3390/e23020221
[28] 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
[29] H. Wang, S. He, K. Sun, Complex Dynamics of the Fractional-Order Rössler System and Its Tracking Synchronization Control, Complexity, vol. 2018, Article ID 4019749, 13 pages, 2018, doi:10.1155/2018/4019749.