Detection of cyber threats to network infrastructure of digital production based on the methods of Big Data and multifractal analysis of traffic[a]

Daria Lavrova[1*], *Maria* Poltavtseva[1], *Anna* Shtyrkina[1], and *Pyotr* Zegzhda[1]

[1] Peter the Great St. Petersburg Polytechnic University, Institute of Computer Sciences and Technologies, 195251 Polytechnicheskaya st. 29, Russian Federation

Abstract. The article offers an approach to analyzing data security of network infrastructure of digital production providing for contraction of network traffic size and detecting anomalies in the network traffic on the basis of multifractal analysis. The contraction of data size will be provided due to extraction of significant parameters from the network packets and dropping the rest data, as well as due to application of such Big Data method as aggregation. The experimental investigations on contracting data size on analyzing security have proven the operability and efficiency thereof. The method of contracting data size has demonstrated a possibility of traffic volume contraction from hundreds of Gbit to several Mbyte. The suggested approach to security analysis using the assessment of width of multifractional spectrum as a criterion of anomaly presence has detected both simulated attacks of denial of servicing SYN-flood and smurf. Thus, the suggested approach can be efficiently used for analyzing big volumes of dissimilar traffic of network infrastructure of digital production.

1 Problem of providing security of digital infrastructure

The digitization of key and vital field of human activity has resulted in that the ability of data exchange and the works with the use of Internet became the most important demand of the modern community. The use of M2M technology and incorporating an Internet of things concept into production branches have triggered the emergence of cyber-physical systems (CPS) consisting of interrelated sensors, networks, cloud systems of data storage, applications and devices. CPS differ from information and computer systems by the principles of work organization: a collective functioning of the elements of cybernetic and

[a] With financial support from the Ministry of Education and Science of the Russian Federation in the framework of the Federal targeted program "Investigations and developments in the priority field of development of Russian science and technology complex for 2014-2020", Agreement No. 14.578.21.0231, agreement unique identifier RFMEFI57817X0231).
[*] Corresponding author: lavrova@ibks.spbstu.ru

physical environments integrated with computational resources, with minimal participation of an individual or without him is provided.

CPS implements a technological concept of digital environment bringing all the field of country's activity to the new level of competitive ability [1]. At that, according to source [2], more than half of the modern CPS have been integrated with critical field of life-sustaining activity, such as power engineering, defense industry, communication, public health services and transport. It makes the problem of cyber security of such systems highly relevant.

The structural intricacy of CPS, high heterogeneity and low computational capability of the components thereof significantly obstruct a development of a unified methodology of ensuring cyber security of digital infrastructure. The most promising approach to detecting cyber threats in complex intellectual systems according to sources [3, 4] is the analysis of security of infrastructure network data, since the implementation of production processes in CPS takes place in the way of network data exchange by the components thereof, including commands and informing/status messages.

The network infrastructure of digital production systems is highly diversified, it includes both traditional network devices and components of systems of Internet of things (sensors and actuators, "smart" devices). The availability of a great number of "smart" devices brings about heterogeneity of the used network protocols and an essential growth of the network traffic volume.

The cyber security of digital production implies the analysis of security of super-high data volumes, since the development of cloud, mobile, sensor technologies and technology of Internet of things integrated with the digital production has resulted in a vigorous growth of volume of the global IP-traffic. According to data of Cisco corporation the volume of global IP-traffic will increase 3 times by 2021, attaining 3.3 ZettaByte. At that, the share of traffic of devices of Internet of things and devices interacting with respect to M2M technologies (which is characteristic for digital production) will amount to about 5% of the global IP-traffic, while a share of connections of M2M supporting Internet of things will constitute more than half of the total number of devices and connections constituting 27.1 billion devices. A super-high volume of heterogeneous network traffic influences negatively the quality of detecting security threats, since the "surges" caused by abnormal traffic are much less distinguishable in big data volumes. In view of this fact a risk of omitting cyber attacks and compromising digital infrastructure increases.

In order to timely detect cyber threats and counteract them, the analysis of network traffic security shall be performed in the real-time mode, which is complicated by the enormous traffic volumes. Taking into account the above peculiarities of the network infrastructure of digital production, the following is required for solving the task of effective analysis of its data security:

1) provide contraction of size of super-high volume of network traffic for increasing the rate of security analysis [5];

2) develop methods of network traffic security analysis able to detect anomalies under conditions of big data scales with high accuracy.

This article offers an approach to detecting anomalies in the super-high volumes of network traffic based on the assessment of multifractal spectrum of traffic parameter time series. It also describes an approach to preliminary network traffic processing with the use of Big Data technology providing for traffic size contraction on the basis of extracting essential traffic parameters "on-the-fly" and aggregation of the extracted values with the use of hierarchically-dependable windows.

2 Methods of contracting size of super-high network traffic volumes with the use of hierarchically-dependable windows of aggregation

Big Data corresponds to a combination of methods and tools of processing big volumes of heterogeneous data with the aim of extracting information important from the point of view of the task to be solved. The Big Data methods include the methods of data aggregation, normalization, filtration, classification and clusterization.

This article offers to use such Big Data method as aggregation combining values of parameters extracted from the network traffic into a unified indicator [5]. It is proposed to perform total contraction of traffic size in two stages: an extraction of traffic parameters important from the point of view of further security analysis takes place at the first stage, the aggregation of values of parameters with the use of hierarchically-dependable windows is performed at the second stage. The following traffic network parameters have been distinguished for security analysis:

1) IP-addresses of sender and receiver;
2) ports of sender and receiver;
3) timestamp;
4) network packet size;
5) number of network packets in a flow;
6) type of network protocol of transportation level;
7) number of network packets of protocols of every type;
8) number of outgoing and incoming connections for a host.

This list can be expanded, besides, these values shall be a subject to certain statistical processing in future in order to make it possible to handle such statistics as an average, maximum and minimum number of packets, packet size, flow length, etc.

The rate of analysis increases significantly and the data size decreases due to extraction of these parameters, it is related to the fact that there is no need to analyze the payload network packets – all the required data get obtained as quick as possible in the way of analyzing predominantly the packets of the network and transportation levels.

However, such size contraction will be insufficient under conditions of intensively arriving network traffic and a necessity of providing quick data processing. Therefore, a second stage is introduced – an aggregation of extracted values of parameters being implemented with the use of hierarchically-dependable windows [6]. Every window corresponds to a structure describing aggregated data over a certain period of time with the aim of calculating statistical indicators. An upper-level window is the aggregation interval for the inferior windows taking into account the aggregation coefficient α expressed as a positive integer. The window is described by a finite sequence: $Window\langle Id, Dt, N, Parent, Fr, \alpha, \alpha_{cur}, Par\rangle$, where Id is a window identifier, Dt is an aggregation time period, N is a length of numerical rows of parameters the same for all window parameters *(100 < N < 1000),* N_{cur} is a number of values in the window time rows, $Parent$ is a parent window located higher in the hierarchy of windows (for all window except a minimal one), Fr are related methods of calculation of statistical parameters applied to data of window under consideration, $Par = \{Qp_1, \ldots, Qp_m\}$ are the time rows or waiting lines of parameters. The window structure is shown in accordance with Fig. 1.

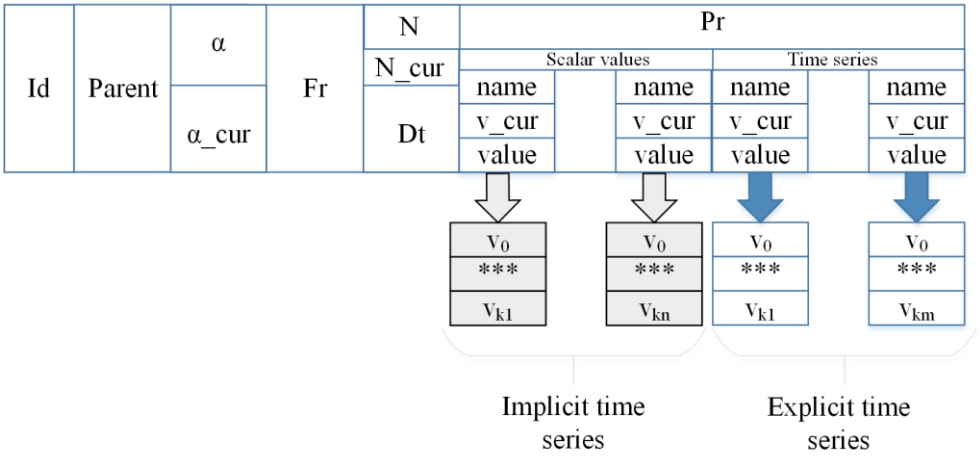

Fig. 1. Structure of hierarchically-dependable aggregation window.

Parameters α and α_{cur} are used in the course of shifting time rows of windows and organizing hierarchic dependence thereof. α is a coefficient of dependence between windows, α_{cur} is a number of accumulated shifts of the superior window. Upon attaining $\alpha_{cur} = \alpha$ a counter of current shifts is reset, while the current window shifts by one value in the aggregation row.

The shaping of hierarchy of dependable time windows is implemented in the way of selecting windows of analysis in accordance with specified values, selecting the preceding host window for every time window, shaping a tree of windows and shaping a "virtual" window of tree root not taking part in the analysis. General diagram of functioning method of size contraction is shown in Fig. 2.

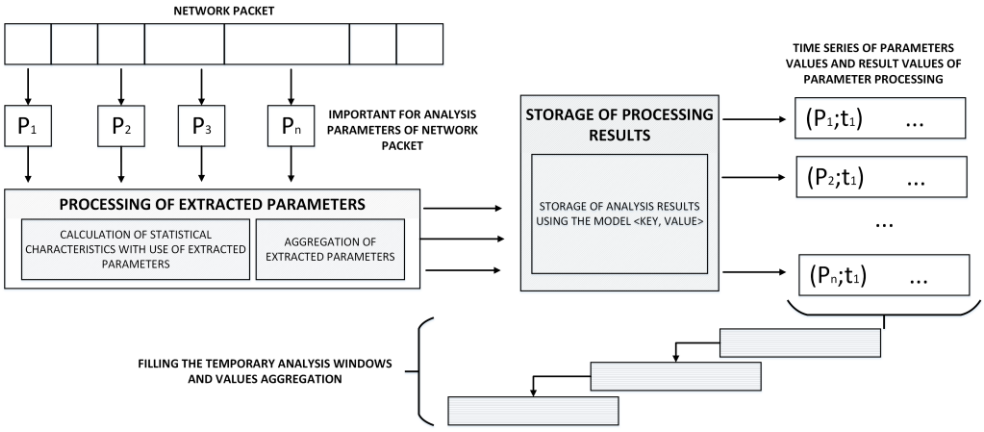

Fig. 2. Diagram of functioning method of size contraction of super-high volumes of network traffic.

A general assessment of contracting data size with the use and support of dependable aggregation windows will be determined in the following way (1):

$$V = N*P*B*m \ , \tag{1}$$

where:

1) N – number of aggregation windows;
2) P – number of parameters in every window;
3) B – size of one value of one parameter;
4) m – length of time row for window parameters.

Thus, the developed method provides for contraction of direct size of the stored information for getting an opportunity of its prompt processing and analyzing without losing significance.

3 Analysis of network traffic security with the use of multifractal characteristics

In order to detect anomalies under conditions of big volumes of the network traffic, it is necessary to use metrics and characteristics providing for high accuracy of detecting even insignificant deviations of traffic behaviour from a normal one. A big volume of traffic according to the data of many investigations features a self-similarity property, i.e., it looks qualitatively in the same way with any rather irresistible proportions of time axis [7]. The self-similarity property is used for simulation of the network traffic and for solving tasks of information security after revealing anomalies [8].

The fractal methods, which are gaining the ever-growing popularity in the recent times, are built just on this peculiarity. The main focus in the fractal analysis of traffic is made on calculating Hurst exponent, which a measure of stability of statistical phenomenon or measure of duration of a long-time process dependence [9]. However, the investigations show that under conditions of real networks the traffic not always possesses fractal properties, more frequently the network data exhibit the multifractal properties [10]. At that, the self-similarity property can be seen on the big scales of traffics, while the multifractal properties are characteristic within smaller time sections of the network. Thus, the property of scale invariance, or fractality reflects the long-term signal behaviour, while multifractal metrics reflect its momentary behaviour.

The authors suggest in this article to calculate the multifractal heuristics providing for detection of anomalies invisible in the big traffic volumes and making it possible to describe the network traffic in the form of a multifractal. The multifractal is assigned by several algorithms changing in sequence, each of them generates a pattern with its fractal size.

It is offered in this article to use a value of multifractal spectrum width as a criterion of whether the network traffic is normal or abnormal [11].

In order to build the spectrum, one can suppose that there is a certain fractal object. Let us assume that its surface is covered with cubes measuring ε, then a probability of hitting a cell i with an arbitrary point will be designated as ρ_i. In this case the spectrum of fractal sizes will be expressed by formula (2):

$$D_q = \frac{\tau(q)}{q-1} = \lim_{\varepsilon \to 0} \frac{1}{q-1} \left(\frac{\ln\left(\sum_{i=1}^{N(\varepsilon)} \rho_i (\varepsilon)^q \right)}{\ln \varepsilon} \right), \tag{2}$$

where, $\tau(q)$ is referred to as a scaling function.

In practice it is convenient to use variables $f(\alpha)$ and α, which are acquired from $\tau(q)$ and q by means of Legendre transformation [11]. The value $f(\alpha)$ is referred to as multifractal Legendre spectrum, while the variable α – is referred to as a local Lipschitz-Helder indicator. In order to build the Legendre spectrum, a row of process of incrementations $Z_1, Z_2, \ldots, Z_n$ is assigned and an aggregated sequence corresponding to it $\left\{Z^{(m)}\right\}$ is determined at the level of aggregation m (3):

$$Z_k^{(m)} = Z_{(k-1)m+1} + Z_{(k-1)m+2} + \ldots + Z_{km}, \tag{3}$$

where, $k, m = 1, 2, \ldots$. Further, a sum of separation is determined (4):

$$S_m^Z(q) = \sum_{k=1}^{N/m} \left(\overline{Z_k}^{(m)}\right)^q, \tag{4}$$

where, Z – vector of data, for which a multifractal spectrum is built, while $\overline{Z_k}^{(m)} = \sum_{l=1}^{m} Z_{(k-1)n+1}$ – discretization of measure μ on scale $\delta_m = \frac{m}{N}$, $m = 1, 2, 2^2, \ldots, 2^n$ – size of summation unit [11].

It is widely thought that Z_i is a multifractal, if $\log S_m^Z(q)$ with approximation linearly depends on $\log(m)$. The inclination of approximation is designated $\tau(q)$ and calculated with the use of linear approximation (5):

$$\log S_m^Z(q) \approx \tau(q) \log m + c(q), \tag{5}$$

The multifractal spectrum is the Legendre transformation from separation function $\tau(q)$ (6):

$$f_L(\alpha) = \inf_{q \in R} \left(\alpha q - \tau(q)\right), \tag{6}$$

Thus, the multifractal Legendre spectrum will be built in the course of performing experimental investigations [11]. From the point of view of security analysis one can assume that the width of multifractal spectrum will either decrease or increase during network attacks implementation. The increase of width of multifractal spectrum testifies to the fact that the process under investigation became more heterogeneous. Regarding the network traffic and its parameters it can denote the emergence of the new traffic and new dependencies in the traffic testifying to an attempt of attack implementation. The decrease of width of multifractal spectrum testifies to the fact that the process became more homogeneous. Regarding the network traffic, it can denote the increased percentage of the identical traffic (e.g., attacked with the repeating packets).

4 Conducting experimental investigations

The conducted experimental investigations concerned the assessment of contracting size of the network traffic and detection of network attacks on the basis of multifractal analysis. The intensively arriving network traffic of super-high volumes has been simulated in the work framework. The speed of traffic under investigation equals 100 Gbit/s, i.e., during 1 hour of observations the volume of accumulated data equals 100 Gbit×3 600=360 000 Gbit=360 Tbit. Hence, that during one day only more than 8 Pbit of information are to be subject to information, where, 1 Pbit = 10^6 Gbit = 10^{15} bit.

The first stage of method of data size contraction consists in the extraction of key parameters from the network traffic. The average size of network packet equals 1,500 byte. Let us assume that around 50 parameters get extracted from every packet taking into account derived parameters from the flows and packets of higher level. If 10 bytes are allotted to every parameter, the total volume required for one packet storage is 500 byte. Thus, this method helps keep 500 byte out of 1,500 byte of data only in the course of the first stage, i. e., the saving of space in data storage will amount to 66%.

Figure 3 presents results of experimental investigations on data size contraction following the first stage of method functioning.

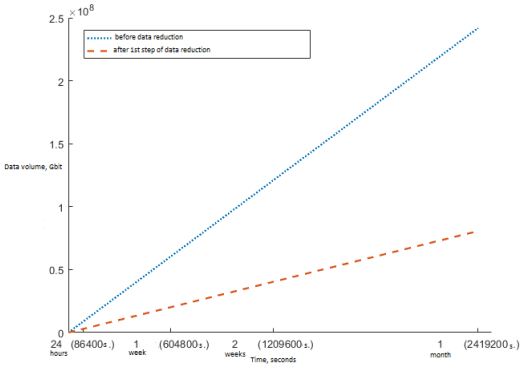

Fig. 3. Result of Functioning method of size contraction at the first stage for different time intervals.

In order to provide for greater contraction of data size at the second stage, the aggregation of acquired values of parameters will be carried out with the use of hierarchically-dependable windows. The efficiency of decreasing data volume at this stage depends on the parameters selected by investigator. 10 aggregation windows corresponding to time intervals from seconds to a month have been used in the conducted experiment.

The volume of "clean" data for traffic under investigation without taking into account auxiliary structures has amounted to about 5 Mb of memory for an average length of time row of $m = 1000$ elements, number of aggregation windows $N = 10$, number of parameters $P = 50$, and $B = 10$ byte required for storage of one parameter. Taking $N = 5$ aggregation windows, about 2.5 Mb of memory will be required for data storage.

Consideration has been given to dump network traffic without attacks, dump with attack SYN-flood and dump with attack smurf in the course of conducting experimental investigations related to detection of anomalies in the network traffic [12, 13].

A parameter characterizing the number of TCP-packets of definite type over a period of time of Δt=5 seconds has been used when conducting attack SYN-flood or shaping time row. A change of spectrum in the course of conducting attack SYN-flood is demonstrated in Fig. 4 a) and b).

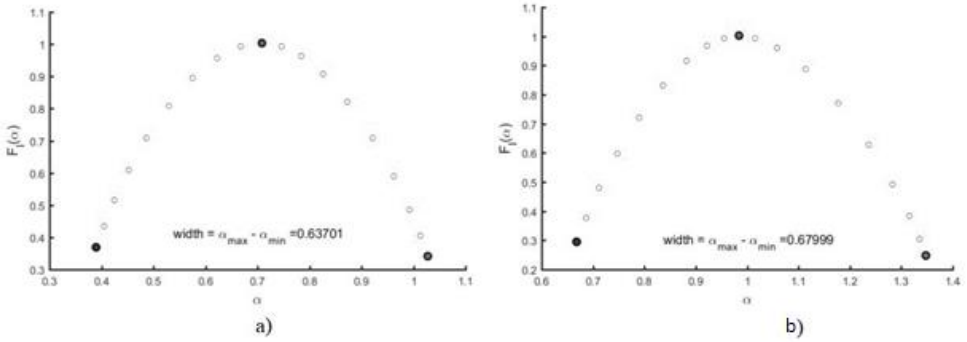

Fig. 4. Multifractal legendre spectrum for time rows of network packets of traffic without attack (a), with attack SYN-flood (b).

The increase of width of multifractal Legendre spectrum by more than 0.4 is observed in this experiment. Such a phenomenon is explained by the fact that the number of TCP-packets in the normal traffic did not vary and remained approximately at the same level. In the course of attack the number of TCP-packets began to increase gradually, as a result of which the network traffic became more heterogeneous over the investigated period of time, which resulted in manifestation of more multifractal properties.

In the course of conducting attack smurf the other network traffic parameter has been investigated – an average size of network packets. Figure 5 presents multifractal spectra for the cases of attack smurf and without it.

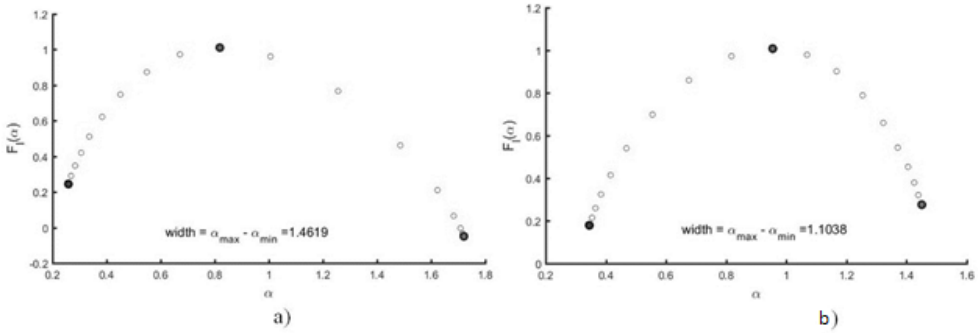

Fig. 5. Multifractal legendre spectrum for time rows of average size of traffic network packets without attack (a), with attack smurf (b).

Apparently, in case of availability of attack of smurf type a great number of identical ICMP-packets appear in the traffic. As a result, the network traffic gets saturated with identical packets and becomes less heterogeneous, therefore, a degree of time row multifractality (of width of multifractal spectrum) decreases.

5 Conclusions

This article offers an approach to detecting cyber threats in the network infrastructure of digital production taking into account specific features of application environment. The network infrastructure of digital production features high volume and heterogeneity of its composition, which entails a necessity of super-high volumes of network traffic. At that, the traffic shall be processed as quick as possible for efficient detection of security threats and immediate counteracting them, which is an unconventional problem under conditions of big volumes of intensively arriving data.

The approach offered in the article includes a method of contracting size of super-high volumes of network traffic on the basis of Big Data technology and method of detecting anomalies in the network traffic based on multifractal analysis.

The contraction of traffic size takes place in two stages, an extraction of parameters important from the point of view of security analysis takes place at the first stage, the aggregation of the extracted values of parameters on the basis of hierarchically-dependable windows is performed at the second stage. The performance of this method has been proven by the contraction of data size approximately by 2/3 after the first stage of contraction. The efficiency of data volume contraction at this stage depends on parameters selected by the investigator. The experimental investigations have demonstrated the contraction of traffic volumes down to several Mbyte, which testifies to high efficiency of using hierarchically-dependable windows of aggregation.

In order to analyze security of network traffic, an assessment of such traffic characteristic as multifractal spectrum has been performed. The assessment of multifractal properties of network traffic required a significantly lower data volume, since the multifractality is distinguishable particularly at relatively small volumes. From the point of view of security analysis it is reasonable to monitor the multifractal traffic properties, the multifractal describes the network traffic with more accuracy demonstrating dependencies, which are not distinguishable at big data volumes. It is especially relevant under conditions of big traffic volumes from the objects of network infrastructure of digital production.

The conducted experimental investigations have proven the functionality of the proposed approach to security analysis: the value of width of multifractal spectrum appeared to be sensitive to the network traffic changes. At that, the use of different characteristics of network traffic is possible when building the multifractal spectrum. In particular, the medium size of network packets and a number of TCP-packets in the traffic have been used, both characteristics have reflected the emergence of anomalies.

A further direction of investigations will be devoted to establishing boundaries of values of spectrum width for shaping a certain confidential interval. The system of security analysis will not need to generate a signal on the possible cyber attack, if it enters this interval. Besides, further investigations will be dedicated to searching other network traffic characteristics capable of reflecting presence of anomalies in the network traffic with high accuracy.

References

1. R. Seiger, S. Huber, P. Heisig, U. Assmann, LNBIP, **248** (2016)
2. Y. S. Vasiliev, P. D. Zegzhda, D. P. Zegzhda, Aut. Cont. and Comp. Scien., **63** (2016)
3. D. S. Lavrova, A. I. Pechenkin, IJCNIS, **7** (2015)
4. D. S. Lavrova, Aut. Cont. and Comp. Scien., **50** (2016)
5. M. A. Poltavtseva, A. I. Pechenkin, D. S. Lavrova, Soft. & Sys., **2** (2016)
6. N. Karthick , X. Agnes Kalarani, IJCTT, **17** (2014)

7. D. E. Sokolov, N. G. Trenogin, SibSUTIS, **34** (2001)
8. D. E. Sokolov, N. G. Trenogin, Mod. probl. of inf. in tech. and technologies., **10** (2004)
9. R. Yan, Y. Wang, IT Journ, **11**, (2012)
10. F. H. T. Vieira, G. R. Bianchi, L.L. Lee, J. High Speed Netw, **17** (2010)
11. A. N. Pavlov, V. S. Anishchenko, Physics-Uspekhi, **50** (2007)
12. N. Wattanapongsakorn, INC, **98** (2011)
13. C. Fachkha, E. Bou-Harb, M. Debbabi, Wireless Com. and Mob. Comp., **15** (2015)