ISSN (e): 2250 – 3005 || Vol, 04 || Issue, 5 || May – 2014 ||
International Journal of Computational Engineering Research (IJCER)

Experimental Studies of the Statistical Properties of Network Traffic Based on the BDS-Statistics

Alexey Smirnov [1,] Dmitriy Danilenko[2]

1 Professor in the Department of Software, Kirovohrad National Technical University, Kirovohrad, Ukraine,
2 Graduate student in the Department of Software, Kirovohrad National Technical University, Kirovohrad, Ukraine

Abstract:
Experiments, which outcome in the results of correlation analysis of network traffic based on the BDS-test, which may be used as part of an analytical component of modern anti-virus systems, have been conducted. In addition, the correlation analysis of network traffic may be used for organizing of one of the main elements of the system for monitoring a network activity as a touch subsystem (sensors to collect traffic information), as well as an analytical part (decision-making module component). A *unit to assess the significance differences of two or more samples (series)of independent observations of network traffic (Wilcoxon criterion) is used to solve the problem of detection of individual services of a telecommunications system following the observed network traffic. It allows to set different data streams belonging to the same general totality with a given accuracy and reliability..*

Keywords: *telecommunication systems and networks, system of intrusion detection and prevention, BDS-statistics*

I. INTRODUCTION

To ensure the safety of modern telecommunication networks the so-called intrusion detection system (IDS) and intrusion prevention systems (IPS) are used [1-14]. At the heart of their operation there is the collection, analysis and processing of information about the events related to the security perimeter of the telecommunications network, the accumulation of the collected data, monitoring of network activity of individual services, deciding on the status of the protected system, as well as identifying and countering possible unauthorized use of information and communication resources [2]. One of the directions in improving systems for intrusion detection and prevention is the study of anomalies (Anomaly-Based Intrusion Detection and Prevention Systems – AB IDPS) in telecommunication systems, which is based on statistical analysis of network traffic [2]. Within this approach, IDPS defines a "normal" network activity of individual information services of a telecommunication system, then all the traffic that is not covered under the definition of "normal" is marked as "anomalous".

The analysis of correlation methods for the identification of objects showed that one of the most effective approaches for identifying dependencies in data traffic is BDS-statistics, which is constructed based on the BDS-tests (BDS-methods). BDS-tests are effective methods to identify dependencies in the time series. Their aim is to test the null hypothesis H_0 about the independence and the identical distribution of the time series' values $\vec{\xi} = (\xi_1, \xi_2, ..., \xi_N)$, using for it a criterion of significance. According to this criterion, for accepting the hypothesis H_0 it is necessary to choose a critical domain G_α satisfying the condition of $P(g \in G) = \alpha$, where $g(\xi_1, \xi_2, ..., \xi_N)$ – is the observation statistics, and α – is an adjustable level of significance [17-20].

II. BDS-STATISTICS DESCRIPTION

BDS-test is based on the statistic value of $w(\vec{\xi})$ (BDS-statistics) [17-20]:

$$w_{m,N}(\varepsilon) = \sqrt{N - m + 1}\,\frac{C_{m,N}(\varepsilon) - C_{1,N-m}(\varepsilon)^m}{\sigma_{m,N}(\varepsilon)},$$

where $C_{m,N}(\varepsilon) - C_{1,N-m}(\varepsilon)^m$ – (numerator BDS-statistics) is determined by the correlation integrals $C_{m,N}(\varepsilon)$, $C_{1,N}(\varepsilon)$ for the dimension m; ε – is the radius of the hypersphere; $\sigma_{m,N}(\varepsilon)$ – is a standard deviation of the difference $C_{m,N}(\varepsilon) - C_{1,N-m}(\varepsilon)^m$; N – a number of elements of the time series.

A number of studies [17-20] have proposed "simplified" algorithms of BDS-statistics estimation. In them, for the calculation of $C_{m,N}(\varepsilon)$ $(m>1)$, it is necessary to perform "embedding" of the time series of m-dimensional pseudo-phase space, the elements of which, by the theorem of Takens [19], are the points $\xi_i^m = (\xi_i, \xi_{i+1}, ..., \xi_{i+m})$ with the coordinates $\{\xi_{i+k}\}_{k=1}^m$ given by m successive values of the original time series. Correlation integral determines the frequency of contact of any pair of phase space points in the hypersphere of ε radius:

$$C_{m,N}(\varepsilon) = \frac{2}{(N-m+1)(N-m)} \sum_{s=m}^{N} \sum_{t=s+1}^{N} \prod_{j=0}^{m-1} I_\varepsilon(\xi_{s-j}^m, \xi_{t-j}^m),$$

$$I_\varepsilon(\xi_i^m, \xi_j^m) = \begin{cases} 1, & \left\| \xi_i^m - \xi_j^m \right\| \le \varepsilon \\ 0, & \left\| \xi_i^m - \xi_j^m \right\| > \varepsilon \end{cases},$$

$$\{\xi_i\}_{i=1}^N \quad 0 \le i \le N \text{ and } 0 \le j \le N,$$

where $I_\varepsilon(\xi_i^m, \xi_j^m)$ is the Heaviside function for all pairs of values i and j.

The value of the correlation integral approaches a definite limit as ε decreases. The analysis of studies [17-20] showed that there is a range of ε values, which allows performing calculations with the specified accuracy coefficient. This range depends on the number of elements of the time series N. If ε is too small, there will not be enough points to capture the statistical structure; if ε is too large, there will be too many points.

The studies [17-20] recommend to choose ε so that $\varepsilon = 0.5\sigma \div 2\sigma$, where σ – is a standard deviation of the process $\{\xi_i\}_{i=1}^N$. In accordance with the theory of statistics, the dependence of the correlation integral from ε is as follows:

$$C_{m,N}(\varepsilon) \sim \varepsilon^{D_c},$$

where D_c is a correlation dimension of the time series.

For $m=1$ we have:

$$C_{1,N}(\varepsilon) = \frac{2}{N(N-1)} \sum_{s=1}^{N} \sum_{t=s+1}^{N} I_\varepsilon(\xi_s, \xi_t).$$

The studies performed has shown, that when $N \to \infty$, the correlation integral $C_{m,N}(\varepsilon) \Rightarrow C_{1,N}(\varepsilon)^m$, and the value $\left(C_{m,N}(\varepsilon) - (C_{1,N}(\varepsilon))^m\right) \cdot \sqrt{N-m+1}$ is asymptotically normally distributed random variable with a mean zero and standard deviation $\sigma_{m,N}(\varepsilon)$, which is defined as:

$$\sigma_{m,N}(\varepsilon) = 2\sqrt{k^m + 2\sum_{j=1}^{m-1} k^{m-j} \cdot (C_{1,N}(\varepsilon))^{2j} + (m-1)^2 \cdot (C_{1,N}(\varepsilon))^{2m} - m^2 k (C_{1,N}(\varepsilon))^{2m-2}},$$

where

$$k = \frac{1}{(N-1)(N-2)N} \left\{ \sum_{t=1}^{N} \left[\sum_{s=1}^{N} I_\varepsilon(\xi_t, \xi_s) \right]^2 - 3\sum_{s=1}^{N} \sum_{t=s+1}^{N} I_\varepsilon(\xi_t, \xi_s) + 2N \right\}.$$

BDS-statistics $w(\vec{\xi})$ is a normally distributed random variable with the proviso that the estimate $\sigma_{m,N}(\varepsilon)$ is close to its theoretical value $\sigma_{m,N}(\varepsilon)$.

The problem of detection of chaotic signal is considered as a non-parametric verification of one of the two hypotheses:

1) H_0 – the observed data (data traffic) $\vec{\xi} = (\xi_1, \xi_2, ..., \xi_N)$ are independent and identically distributed, i.e. the density (function) distribution is factored $F_N(\xi_1, \xi_2, ..., \xi_N) = \prod_{i=1}^N F(\xi_i)$.

2) H_1 - the obtained as the result of the experiment data (traffic information) have a certain relationship (a process is structured).

According to the hypothesis H_0 the statistics $w(\vec{\xi})$ is asymptotically distributed as N(0,1), if a number of observations asymptotically approaches the infinity. A number of studies [17-20] settle the hypothesis about the need for the pilot study of more than 500 observations. Such number of experiments will allow to argue about the reliability of the received results.

The studies have shown that the criterion of the hypothesis validity H_0 (in the absence of any data traffic dependencies) is the inequality:

$$\left|w_{m,N}(\varepsilon)\right| \le 1{,}96 .$$

For a value of the statistic $w_{m,N}(\varepsilon)$ the given value corresponds to the level of significance $\alpha = 0{,}05$ (probability of a 1st type error), and when the above inequality is the true hypothesis H_0 (I.I.D.) is accepted with the probability $P_{H_0} \approx 0{,}95$.

In the case where the alternative hypothesis H_1 is true, a distribution of the statistic criterion $w(\vec{\xi})$ is changed. Therefore, when checking statistical hypotheses it is insufficient to focus on the value of the significance level α.

A power of the criterion $1-\beta$ or probability of error of the second kind β should be determined when considering the alternative hypothesis H_1, which implies dependence (possibly nonlinear) of a time series, if the first difference of the natural logarithms have been taken. A power of the criterion is the probability of considering the alternative hypothesis H_1 in applying the criterion $w(\vec{\xi})$ with the proviso that it is true, that is its ability to detect the existing deviation from the null hypothesis. Obviously, for a fixed error of the 1st kind (we set it ourselves, and it does not depend on the criterion properties) the criterion will be the better, the more its power is (i.e., the smaller is the error of the 2nd kind). For calculating the power of the criterion $1-\beta$ ($\beta = p\left(w(\vec{\xi}) \in G_\alpha \mid H_1\right)$), G_α – is a critical area at a given level of significance α it is necessary to know the conditional density distribution $p\left(w(\vec{\xi}) \mid H_1\right)$. The power of the criterion (test) is determined empirically.

For the experiment and improving the reliability of the results, it is necessary to choose such embedding dimension m, whereby the phase space reconstruction is neither "too rare" nor "too crowded." A number of studies [ ] recommends m = 6 in experiments.

Thus, the conducted analysis of different approaches in the statistical testing showed that the BDS-test gives a possibility to detect various types of deviations from the independence and identical distribution and can serve as a general model test of the processes classification (time series) $\vec{\xi}$, especially in the presence of nonlinear dynamics.

The nonparametric nature of BDS-testing may be considered to be its main feature. This is reflected in the fact that BDS-test uses a nonlinear function $w(\vec{\xi})$ as the statistics from observations, the distribution of which is independent of the observed values $\vec{\xi}$ distribution. In this case, we are able to get some information about the multidimensional function (density) of distribution $F_N(\xi_1,\xi_2,...,\xi_N)$ analyzing a dimensional empirical function of distribution $p(w)$ of the statistics w.

BDS-test calculation may be carried out by various methods, many implementations are known, this technique proposes using the implementation suggested by the authors of the BDS-test (by W. A. Brock, W. D. Dechert and J. A. Sheinkman). Fast algorithms for calculating BDS statistics has been also proposed [17-20]. Below are the results of experimental studies of the network traffic statistical properties based on the correlation analysis of time series (by BDS-testing). The results include the data matching the most popular network protocols and services.

III. EXPERIMENTAL STUDIES OF THE NETWORK TRAFFIC STATISTICAL PROPERTIES BASED ON THE CORRELATION ANALYSIS OF TIME SERIES

The sample size for research based on the statistical properties of the correlation analysis of time series is N = 1000, the calculations has been performed with different parameters (the results are shown in the respective tables). The analysis of HTTP traffic experimental data shows that the phase portrait of network traffic for data transmission using HTTP protocols shows having some dependency, consisting in grouping the most points in a certain area. Table 1 shows the values of BDS-test with different sets of parameters. HTTP is an application level of protocol data (initially – in the form of hypertext documents). HTTP is based on the "client-server" technology, that assumes existing consumers (clients) who initiate a connection and send a request, and providers (servers) that are waiting for connection request, produce the necessary actions and return a message back with the result.

Table 1 – The results of BDS-test for experimental data of HTTP traffic with different parameters of m and ε

	m	ε = 0.5 σ	ε = σ
HTTP	4	13,52	12,69
	5	13,97	11,554
	6	13,87	10,65
	7	13,58	9,92

Thus, as it follows from the experimental data obtained for this kind of HTTP traffic, the characteristic value of the BDS-test is as follows:

- for the radius ε = 0.5 σ when m = 4..7 the values range from ≈ 15 to ≈ 14;
- for the radius ε = σ there exists a greater variation of values ranging from 9.92 to 12.69.

These values may be used as test (reference) values at the detection of the traffic type and the corresponding network service.

Let's process the obtained experimental data of FTP traffic. FTP is a protocol used to transfer files via computer networks. It allows connecting to FTP servers, browsing the contents of directories and downloading files from the server or uploading them to the server; moreover, a mode of transmission of files between the servers is possible. FTP protocol refers to the application layer protocols and uses the TCP transport protocol for data transfer. The commands and data, in contrast to most other protocols, are transmitted on different ports. The outbound port 20 opened on the server side is used to transmit data, and the port 21 for sending commands. The port for receiving the customer data is determined in the matching dialogue. If a file transfer is interrupted for any reason, the protocol provides a means to download the rest of the file, which is very useful when transferring large files.

In the above snippet of experimental data of network traffic using FTP protocols the transition from a small number of packets (transmission of protocol commands, retrieving a list of directories, etc.) and the beginning of active download files at a certain level of speed can be clearly seen. As the server used had a load speed limit which was significantly less than bandwidth network capabilities, the "bursts" which rapidly decrease to the threshold speed limit are sometimes observed.

A points' grouping that indicate the presence of dependencies in the source data are observed in the phase portrait. Table 2 shows the values of BDS-test with different sets of parameters.

The data of Table 2 may be used as reference values for the FTP traffic in a system of intrusion detection of telecommunications systems and networks.

Table 2 – Values of BDS-test for experimental data of FTP traffic with different parameters of m and ε

	m	ε=0.5 σ	ε= σ
FTP	4	19,5	17,69
	5	17,81	16,13
	6	16,49	14,91
	7	15,437	13,95

Let's process the obtained experimental data of Skype traffic. Skype is free software with a closed code which enables encrypted voice communications over the Internet between computers (VoIP), as well as paid services for calls to mobiles and landlines. Skype app also allows making conference calls, video calls, and also provides text messaging (chat) and file transfer. There is an opportunity to transmit an image from the screen instead of an image from a webcam.

Two areas around which the majority of the points are formed are observed in the phase portrait for «skype» traffic, which indicates the presence of dependencies in the original sequence. Table 3 shows the values of BDS-test for Skype traffic with different sets of parameters.

Table 3 – Values of BDS-test of experimental data for Skype traffic with different parameters of m and ε

	m	ε=0.5 σ	ε= σ
Skype	4	16,49	180,11
	5	15,05	163,77
	6	13,95	150,94
	7	13	140,58

The results of the Skype traffic studies obtained listed in Table 2 may be used as reference values in intrusion detection systems of communication systems and networks.

Let's process the obtained experimental data of streaming video traffic. Streaming video is multimedia data which is continuously received by a user from a provider of stream broadcasting. Currently, this service is very popular and traffic volume is about half of the transmitted traffic on the Internet.

On the corresponding fragment of network traffic many bursts and downs characteristic for streaming video is observed. Two areas may be identified in the phase portrait. The first has the distribution close to random. However, in another area of the explicit the points are grouped with sufficiently high accuracy, indicating the presence of dependencies in the original sequence. Table 4 shows the values of BDS-test with different sets of parameters for experimental data of streaming video.

Table 4 – Values of BDS-test for experimental data of stream broadcasting with different parameters of m and ε

	m	ε=0.5 σ	ε= σ
Streaming video	4	32,5	18,2
	5	32,254	19,28
	6	38,45	20,39
	7	41,42	28,49

The data analysis of Table 4 shows that the values of the BDS-test for streaming video are in a sufficiently large range, but can be used as a reference.

Thus, the results of experimental studies of the statistical properties of network traffic using correlation analysis of time series show that the specific values of the BDS-test can be "meterized" for different services of telecommunication systems and networks. The calculations confirm the theoretical assumptions that different types of traffic the result of BDS-test gives different values that can be taken as a reference.

Thus, Table 6 shows the average values corresponding to the different traffic types for different values of ε, which allows identifying the network services. For example, for the values of ε = 0.5 σ and ε = σ the averaged values of the BDS-tests for each type of traffic can be distinguished, and on this basis a network activity detection process of a separate service of a telecommunications network can be organized.

It should be noted that in a real telecommunication network popular services may be used simultaneously, which would change the values of the BDS-test. However, the implementation of unauthorized network intrusion will change the statistical properties of the network traffic and the corresponding values of the BDS-statistics. Furthermore, even in the combined operation, as a rule, some specific service prevails, which allows to select it. And traces of the traffic generated by malicious software, viruses, etc., modify the meaning of the test to a level sufficient to generate solutions about suspicious traffic, i.e. detection of possible traces of the virus traffic in the general flow.

Table 5 – Average values of BDS-test for different services of a telecommunication network

Average values of the BDS-tests		
Service type	ε=0.5 σ	ε=σ
HTTP	13,7	11,2
Skype	14,6	171,9
Multiservice traffic	11,5	13,5
FTP	17,3	15,7
Streaming video	36,2	21,6
Malicious software	40,7	28,5

Thus, the experimental data confirm the theoretical assumption about the possibility of using the values of the BDS-test to detect traces of malicious software in the network traffic.

On this basis, the correlation analysis of the network traffic based on the BDS-testing can be used as follows. First, as part of the analytical component of modern anti-virus systems. Second, the correlation analysis of the network traffic can be used for the organization of one of the main elements of the system for monitoring of a network activity as a touch subsystem (sensors to collect information about the traffic) and the analytical part (decision component module).

When building a system for monitoring of a network activity it is necessary to solve the problem of detection of individual services and telecommunications system from the observed network traffic. Let's consider the problem of assessing the significance of differences of two or more samples (series) of independent observations of the network traffic in order to establish (with a given accuracy and reliability) of their belonging to the same general totality. For this, let's use the mathematical formalism of statistical research and a criterion of belonging of the two samples to one and the same general totality (Wilcoxon criterion).

IV. EXPERIMENTAL STUDIES OF BELONGING OF TWO SAMPLES OF NETWORK TRAFFIC TO ONE AND THE SAME GENERAL TOTALITY

In the study of complex technical systems the problem of assessing the significance of differences in two or more samples (series) of independent observations often arises, i.e. it is necessary to install (with a given accuracy and reliability) their belonging to the general totality. Let's suppose there are two samples:

$$x_1, x_2, \ldots, x_{N1}, \quad (1)$$

and

$$y_1, y_2, \ldots, y_{N2} \quad (2)$$

of random variables X and Y, with distribution $P_X(t)$ and $P_Y(t)$, respectively.
Let's assume that the observed x_i and y_i give different values of sample means

$$x^* = (x_1 + x_2 + \ldots + x_{N_1}) / N_1, \quad (3)$$

$$y^* = (y_1 + y_2 + \ldots + y_{N_2}) / N_2, \quad (4)$$

$$x^* = y^*,$$

and/or sample dispersions (variances)

$$\sigma_X^2 = \frac{1}{N_1}\sum_{i=1}^{N_1}(x_i - x^*)^2, \quad (5)$$

$$\sigma_Y^2 = \frac{1}{N_2}\sum_{i=1}^{N_2}(y_i - y^*)^2, \quad (6)$$

$$\sigma_X^2 = \sigma_Y^2.$$

The solution of assessing the significance of differences in the observed values of x_i and y_i is reduced to testing the null hypothesis H_0, consisting in the distribution functions $P_X(t)$ and $P_Y(t)$ are identical for all t. An alternative hypothesis is formulated in the form of the inequality $P_X(t) < P_Y(t)$.

The criterion of belonging of two samples to one and the same general totality (Wilcoxon's criterion) is based on counting the number of inversions. For this, the observed values of x_i and y_i are located in the general sequence of ascending order of their values. The resulting non-decreasing sequence contains $N_1 + N_2$ of the values and if the hypothesis $P_X(t) = P_Y(t)$ is correct, the values of both sequences $x_1, x_2, \ldots, x_{N1}$ and $y_1, y_2, \ldots, y_{N2}$ are well mixed. The degree of mixing is determined by the number of inversions of the first sequence members relatively to the second. If the overall ordered sequence one certain value of x is preceded by one value of y, which means that there is one inversion. If some values of x are preceded by k values of y, then this value of x has k inversions.

Let's denote the number of inversions for the value of x_i by u_i relatively to antecedent values of y. Then the total number of inversions (for all values of the sequence $x_1, x_2, \ldots, x_{N1}$ relatively to the values from the sequence $y_1, y_2, \ldots, y_{N2}$) will be determined by the sum

$$u = u_1 + u_2 + \ldots + u_{N_1}.$$

The null hypothesis H_0 is rejected if the number u exceeds the selected in accordance with the level of significance of the boundary, determined from the calculation that with the sample sizes of $N_1 > 10$ and $N_2 > 10$ the number of inversions u is approximately normally distributed with center:

$$M_u = \frac{N_1 N_2}{2}, \quad (7)$$

and the variance:

$$D_u = \frac{N_1 N_2}{12}(N_1 + N_2 + 1). \quad (8)$$

At a significance level of q and the normality of distribution of inversions number, the probability of not getting the value of u into the critical area (which means no refutation of the null hypothesis) is [ ]:

$$P(|M_u - u| \le \varepsilon) = 1 - q = 2\Phi\left(\frac{\varepsilon}{\sigma_u}\right), \quad (9)$$

where ε sets the value of the maximum deviation of the resulting estimate from the true value, i.e. ε represents the absolute value of the error in determining the values of the desired characteristics. At the same time, the confidence probability

$$P_д = P(|M_u - u| \le \varepsilon)$$

indicates the probability of achieving the specified accuracy ε.

Let's fix the value of the confidence probability $P_д$, the values of the left and right critical boundaries will be, respectively, equal to:

$$u_1 = M_u - t_\alpha \sigma_u, \quad (10)$$

$$u_2 = M_u + t_\alpha \sigma_u, \quad (11)$$

where: $\sigma_u = \sqrt{D_u}$ – is the standard deviation of the number of inversions, t_α – is the root of the equation $2\Phi(t_\alpha) = P_д$,

$$t_\alpha = \Phi^{-1}\left(\frac{1-q}{2}\right).$$

The less is the significance level of q, the less likely is the possibility to reject the tested hypothesis when it is true, i.e., to make a mistake of the first kind. But with a decrease in the level of significance the range of admissible errors expands, which leads to the increase in the probability of making the wrong decision, i.e., committing type II errors. Typically, the significance level is selected following the considerations that the relevant events in the present situation of research are (with some risk) "practically impossible" (q = 10%, 5%, 2%, 1% etc.).

Using the discussed above criterion of belonging of the two samples to the same general totality, let's carry out experimental studies of the network traffic properties of telecommunication systems and networks.

Let's carry out an experimental study of belonging of two samples of network traffic to the same general totality. For this, let's form the sample (1) – (2) at 100 time reference of randomly selected network traffic segments corresponding to different telecommunication and information services (YouTube (720p), YouTube (360p), Skype (voice), Skype (video), E-mail, HTTP, FTP), thus $N_1 = N_2 = 100$. Let's estimate selective average (3) – (4) and the dispersion (5) – (6), and expectation (7) and the dispersion (8) of the number of inversions:

$$M_u = \frac{N_1 N_2}{2} = 5000,\ D_u = \frac{N_1 N_2}{12}(N_1 + N_2 + 1) = 167500,\ \sigma_u \approx 409{,}3.$$

Let's suppose the significance level of q = 10%. Using (9) to (10) – (11) let's calculate the values of the left and right critical borders:

$$P_д = 1 - q = 0{,}9,$$

$$t_\alpha = \Phi^{-1}\left(\frac{1-q}{2}\right) = \Phi^{-1}(0{,}45) \approx 1{,}65,$$

$$u_1 = M_u - t_\alpha \sigma_u \approx 4324{,}7,$$

$$u_2 = M_u + t_\alpha \sigma_u \approx 5675{,}3.$$

Let's consider the first case when the observed data of network traffic YouTube (720p) act as the sample $x_1, x_2, \ldots, x_{100}$, and the data of network traffic YouTube (720p), YouTube (360p), Skype (voice), Skype (video), E-mail, HTTP and FTP alternatively act as the sample $y_1, y_2, \ldots, y_{100}$.

The obtained results of the studies of belonging of the respective samples of network traffic to one and the same general totality are given in Table 7. The studies have been conducted for the different ways of representing of the network traffic, as in the form of packets' number transmitted per time unit, so as the number of bits transmitted per second.

The results of experimental studies provided in Table 6 indicate that the network traffic of YouTube (720p) service differs in its statistical properties from the network traffic of other services of a telecommunication system. By the criterion of belonging of the network traffic samples to one and the same general totality (Wilcoxon criterion) for different ways of representing of the network traffic (packet/s, bit/s) the number of the observed inversions u exceeds the selected in accordance with the level of significance border and the null hypothesis H_0 is rejected. At the same time, the Wilcoxon criterion gives a reliable mechanism for

detecting of the network traffic YouTube (720p), as a number of inversions u for the corresponding service is, as expected, close to the theoretical value $M_u = 5000$, the null hypothesis is rejected.

Table 6 – Results of research of belonging of network traffic samples to one and the same general totality (network traffic YouTube (720p))

Traffic type (service)	Representation of traffic as a packet/s		Representation of traffic as a bit/s	
	Inversions' number u	The decision on testing the hypothesis	Inversions' number u	The decision on testing the hypothesis
YouTube (720p)	5005	Accepted	5006	Accepted
YouTube (360p)	542	Rejected	572	Rejected
Skype (voice)	302	Rejected	300	Rejected
Skype (video)	2140	Rejected	364	Rejected
E-mail	54	Rejected	47	Rejected
HTTP	1732	Rejected	1011	Rejected
FTP	9539	Rejected	9960	Rejected

Let's consider the second case when the observed data of network traffic YouTube (360p) act as the sample $x_1, x_2, \ldots, x_{100}$, and the data of other network traffic alternatively act as the sample $y_1, y_2, \ldots, y_{100}$. The obtained results are shown in Table 8.

The data shown in Table 7 confirm the correctness of detection of network traffic YouTube (360p), as the results of studies of the belonging of samples to one and the same general totality in different presentation forms the number of the observed inversions u is close to the theoretical value $M_u = 5000$ ($u = 5006$ and $u = 5008$, respectively), the null hypothesis is not rejected. It should be also noted that by the Wilcoxon criterion, one should statistically differentiate the network traffic YouTube (360p) from the traffic services of most other services. HTTP traffic is the exception, which in each of the studied methods of presentation (packet/s and bits/s) is statistically indistinguishable from the traffic YouTube (360p), by Wilcoxon criterion belonging of these network traffic samples to the same general totality is not rejected.

Table 7 – Results of research of belonging of network traffic samples to one and the same general totality (network traffic YouTube (360p))

Traffic type (service)	Representation of traffic as a packet/s		Representation of traffic as a bit/s	
	Inversions' number u	The decision on testing the hypothesis	Inversions' number u	The decision on testing the hypothesis
YouTube (720p)	9555	Rejected	9525	Rejected
YouTube (360p)	5006	Accepted	5008	Accepted
Skype (voice)	7364	Rejected	2900	Rejected
Skype (video)	9502	Rejected	7542	Rejected
E-mail	1066	Rejected	590	Rejected
HTTP	5333	Accepted	4766	Accepted
FTP	989	Rejected	998	Rejected

The results of the study of the statistical properties of network traffic Skype (voice) and testing of the hypothesis that supplies statistical data to the same general totality are shown in Table 8.

Table 8 – Results of research of belonging of network traffic samples to one and the same general totality (network traffic Skype (voice))

Traffic type (service)	Representation of traffic as a packet/s		Representation of traffic as a bit/s	
	Inversions' number u	The decision on testing the hypothesis	Inversions' number u	The decision on testing the hypothesis
YouTube (720p)	9795	Rejected	9797	Rejected
YouTube (360p)	2764	Rejected	7171	Rejected
Skype (voice)	5003	Accepted	5008	Accepted
Skype (video)	10000	Rejected	10000	Rejected
E-mail	333	Rejected	212	Rejected
HTTP	3742	Rejected	5684	Rejected
FTP	10000	Rejected	10000	Rejected

The data presented in Table 9 show the difference in the average sense of the properties of the network traffic Skype (voice) and the properties of other services' traffic of a telecommunication system. The observed number of inversions lies in the critical area, indicating that the test samples belong to different general totalities. In other words, using the Wilcoxon criterion allows detecting Skype (voice) traffic with high reliability and distinguishing it from other network traffics.

The results of studies of the network traffic Skype (video) properties are shown in Table 10. The analysis of the obtained results shows that the traffic Skype (video) is statistically distinguishable from other traffic services, and using of the Wilcoxon criterion allows the detection of the corresponding service of a telecommunication system.

Table 10 shows the results of studies of network traffic samples belonging to one and the same general totality, obtained by analyzing statistical properties of the E-mail traffic network. Tables 11 and 12 show the results of similar studies of HTTP and FTP network traffic, respectively. The analysis shows that the network traffic of E-mail, and FTP services are statistically different by the Wilcoxon criterion from other traffic services, and for them there is no similarity observed in any of the investigated samples. At the same time, the statistical properties of the HTTP traffic are similar to the traffic observed in the operation of the service YouTube (360p). This is also confirmed by the data of Table 7.

Table 9 – Results of research of belonging of network traffic samples to one and the same general totality (network traffic Skype (video))

Traffic type (service)	Representation of traffic as a packet/s		Representation of traffic as a bit/s	
	Inversions' number u	The decision on testing the hypothesis	Inversions' number u	The decision on testing the hypothesis
YouTube (720p)	7932	Rejected	9733	Rejected
YouTube (360p)	604	Rejected	2586	Rejected
Skype (voice)	0	Rejected	0	Rejected
Skype (video)	5009	Accepted	5009	Accepted
E-mail	3	Rejected	0	Rejected
HTTP	2699	Rejected	3108	Rejected
FTP	9901	Rejected	10000	Rejected

Table 10 – Results of research of belonging of network traffic samples to one and the same general totality (network traffic E-mail)

Traffic type (service)	Representation of traffic as a packet/s		Representation of traffic as a bit/s	
	Inversions' number u	The decision on testing the hypothesis	Inversions' number u	The decision on testing the hypothesis
YouTube (720p)	947	Rejected	961	Rejected
YouTube (360p)	9016	Rejected	9483	Rejected
Skype (voice)	9774	Rejected	9890	Rejected
Skype (video)	999	Rejected	10000	Rejected
E-mail	5001	Accepted	5001	Accepted
HTTP	8095	Rejected	8320	Rejected
FTP	10000	Rejected	10000	Rejected

Table 11 – Results of research of belonging of network traffic samples to one and the same general totality (network traffic HTTP)

Traffic type (service)	Representation of traffic as a packet/s		Representation of traffic as a bit/s	
	Inversions' number u	The decision on testing the hypothesis	Inversions' number u	The decision on testing the hypothesis
YouTube (720p)	8314	Rejected	9056	Rejected
YouTube (360p)	4734	Accepted	5288	Rejected
Skype (voice)	6297	Rejected	4323	Rejected
Skype (video)	7327	Rejected	6923	Rejected
E-mail	1990	Rejected	1770	Rejected
HTTP	5003	Accepted	5009	Accepted
FTP	9893	Rejected	998	Rejected

Table 12 – Results of research of belonging of network traffic samples to one and the same general totality (network traffic FTP)

Traffic type (service)	Representation of traffic as a packet/s		Representation of traffic as a bit/s	
	Inversions' number u	The decision on testing the hypothesis	Inversions' number u	The decision on testing the hypothesis
YouTube (720p)	559	Rejected	137	Rejected
YouTube (360p)	29	Rejected	7	Rejected
Skype (voice)	0	Rejected	0	Rejected
Skype (video)	195	Rejected	0	Rejected
E-mail	2	Rejected	0	Rejected
HTTP	204	Rejected	4	Rejected
FTP	5007	Accepted	5007	Accepted

The final results of the conducted experimental studies are summarized in Tables 13, 14, which list the number of inversions and the results of testing the hypothesis of homogeneity of network traffic for different forms of representation (packet/s and bits/s, respectively).

Table 13 – The number of inversions and the results of testing the hypothesis of homogeneity of network traffic (packet/s)

	YouTube (720p)	YouTube (360p)	Skype (voice)	Skype (video)	E-mail	HTTP	FTP
YouTube (720p)	5005 «+»	9555 «–»	9795 «–»	7932 «–»	947 «–»	8314 «–»	559 «–»
YouTube (360p)	542 «–»	5006 «+»	2764 «–»	604 «–»	9016 «–»	4734 «+»	29 «–»
Skype (voice)	302 «–»	7364 «–»	5003 «+»	0 «–»	9774 «–»	6297 «–»	0 «–»
Skype (video)	2140 «–»	9502 «–»	10000 «–»	5009 «+»	999 «–»	7327 «–»	195 «–»
E-mail	54 «–»	1066 «–»	333 «–»	3 «–»	5001 «+»	1990 «–»	2 «–»
HTTP	1732 «–»	5333 «+»	3742 «–»	2699 «–»	8095 «–»	5003 «+»	204 «–»
FTP	9539 «–»	989 «–»	10000 «–»	9901 «–»	10000 «–»	9893 «–»	5007 «+»

Table 14 – The number of inversions and the results of testing the hypothesis of homogeneity of network traffic (bits/s)

	YouTube (720p)	YouTube (360p)	Skype (voice)	Skype (video)	E-mail	HTTP	FTP
YouTube (720p)	5006 «+»	9525 «–»	9797 «–»	9733 «–»	961 «–»	9056 «–»	137 «–»
YouTube (360p)	572 «–»	5008 «+»	7171 «–»	2586 «–»	9483 «–»	5288 «+»	7 «–»
Skype (voice)	300 «–»	2900 «–»	5008 «+»	0 «–»	9890 «–»	4323 «–»	0 «–»
Skype (video)	364 «–»	7542 «–»	10000 «–»	5009 «+»	10000 «–»	6923 «–»	0 «–»
E-mail	47 «–»	590 «–»	212 «–»	0 «–»	5001 «+»	1770 «–»	0 «–»
HTTP	1011 «–»	4766 «+»	5684 «–»	3108 «–»	8320 «–»	5009 «+»	4 «–»
FTP	9960 «–»	991 «–»	10000 «–»	10000 «–»	10000 «–»	998 «–»	5007 «+»

The following designations are applied in tables 13, 14:

«+» – the hypothesis of network traffic homogeneity is not rejected,

«–» – the hypothesis of network traffic homogeneity is rejected,

The results of the hypothesis testing shown in Tables 13 and 14 are symmetrical relatively to the main diagonal, which confirms the reliability of the obtained results in each specific experiment.

The analysis of the data provided in Tables 13 and 14 shows that in the process of the study of various telecommunication services network traffic samples in the majority of cases, the hypothesis of homogeneity is rejected, i.e., there is a correct decision on the belonging of the samples to various processes. This provision

may be made the basis of one of the elements of the system for network activity monitoring, i.e. by using the Wilcoxon criterion an initial detection of telecommunications service can be carried out.

One of the most important characteristics of a random variable is the index of dispersion, which allows comparison of the random parameters of the studied process to determine the significance of differences or matching their characteristics. The following section provides a variance analysis of individual services network traffic and TCS services based on an assessment of relations the variances sample, the statistical hypotheses about the homogeneity of the simulation results on this main index are checked.

CONCLUSIONS

1. The results of experimental studies of the statistical properties of network traffic using the correlation analysis of time series confirm the theoretical assumptions that for different types of traffic the result of the BDS-test gives different values that can be taken as a reference, i.e. and on this basis a network activity detection process of a separate service of a telecommunications network can be organized. Thus, the experimental data confirm the theoretical assumption about the possibility of using the values of the BDS-test to detect traces of malicious software in the network traffic.
2. The obtained results of the correlation analysis of the network traffic based on the BDS-test is recommended to be used, first, as part of the analytical component of modern anti-virus systems. Second, the correlation analysis of the network traffic may be used for the organization of one of the main elements of the system for monitoring network activity as a touch subsystem (sensors to collect traffic information) and as the analytical part (decision module component).
3. To solve the problem of individual telecommunications system's services detection in the observed network traffic, an evaluation unit of the significance differences of two or more samples (series) of independent observations in the network traffic (Wilcoxon criterion) is used. It allows to set different data streams belonging to the same general totality with a given accuracy and reliability.
4. Analysis of the obtained experimental research data shows that during the study of samples of network traffic of various telecommunication services in the majority of cases, the hypothesis of homogeneity is rejected, i.e., there is a correct decision on the belonging the samples to different processes. This position is recommended to be used as one of the elements of a system for monitoring network activity, i.e. through the use of the Wilcoxon criterion is offered to produce a primary detection of a telecommunication service.
5. A promising direction for further research is the assess of the relationship of the sample variances, statistical hypothesis testing in the uniformity of simulation results in terms of dispersion. The results of these studies are designed to perform a comparison of the random parameters of the tested process to determine the significance of differences or matching in their characteristics.

REFERENCES

[1] Kharpuk N. M. The Statistic Analysis of the Network Traffic. Electronic Library of the Belarusian State University – 2008. – P. 116-119. [Electronic Resource]. Access regime: http://elib.bsu.by/bitstream/123456789/7401/1/6.pdf

[2] NIST Special Publication 800-94. Guide to Intrusion Detection and Prevention Systems (IDPS). – Computer Security Division Information Technology Laboratory National Institute of Standards and Technology, Gaithersburg. – 127 pages (February 2007)

[3] Brian Caswell, Jay Beale, Andrew Baker. Snort Intrusion Detection and Prevention Toolkit. – Syngress Media, U.S. 2006. http://www.lehmanns.de/ shop/sachbuch-ratgeber/21797174-9780080549279- snort-intrusion-detection-and-prevention-toolkit#drm1

[4] Ushakov D. V. Development of the Principles of Functioning of Network Intrusion Detection Systems Based on the Model of a Protected Distributed System: Ph. D Dis.: 05.13.19 Moscow, 2005 175 p.

[5] Zapechnikov S. V., Miloslavskaya N. G., Tolstoy A. I., Ushakov D. V. Information Security in Open Systems. Textbook for high schools. In 2 volumes. – M., 2008. – V. II: Network Protection Tools. – 558 p.

[6] Comparison of Firewall, Intrusion Prevention and Antivirus Technologies. http://www.juniper.net/solutions/literature/white_papers /200063.pdf

[7] Intrusion Prevention Systems (IPS). http://www.securecomputing.com/ pdf/Intru-Preven-WP1-Aug03-vF.pdf

[8] Intrusion Prevention Systems (IPS). http://hosteddocs.ittoolbox.com/ BW013004.pdf

[9] State of the Practice of Intrusion Detection Technologies. http://www.sei.cmu.edu/pub/documents/99.reports/pdf/99tr028.pdf

[10] Wireless Intrusion Detection and Response. http://users.ece.gatech.edu/~owen/Research/ Conference%20Publications/wireless_IAW2003.pdf

[11] Anomaly Detection in IP Networks. http://users.ece.gatech.edu/~jic/ sig03.pdf

[12] Design and Implementation of an Anomaly Detection System: an Empirical Approach. http://luca.ntop.org/ADS.pdf

[13] Host-Based Intrusion Detection Systems. http://staff.science.uva.nl/~delaat/snb-2004-2005/p19/report.pdf

[14] Olifer V. G., Olifer N. A. computer Networks. Principles, Technologies, Protocols. Spb.: Piter, 2010. – 944 p.

[15] Smirnov N. V., Dunin-Barkovski I. V., Course on Probability Theory and Mathematical Statistics for Technical Applications. Ed. 2. – M.: Nauka, 1969.-512 p.

[16] Sheffe G., Variance Analysis: Trans. From Eng. Ed. 2. M.: Nauka, 1980. – 512 p.

[17] Kuznetsov A. A. Method of Structural Identification of Information Flows in Telecommunication Networks Based on BDS-test / A. A. Kuznetsov, S. G. Semenov, S. N. Simonenko, E. V. Masleshko// Scientific and technical journal "Science and Technology of the Air Force of Ukraine". Publication 2 (4) – Kharkiv: KhAFU. – 2010. – P. 131 - 137.

[18] Semenov S. The method of processing and identification of telecommunication traffic based on BDS-tests / S. Semenov, A.Smirnov., E.Meleshko // The book of materials International Conference «Statistical Methods of Signal and Data Processing (SMSDP-2010)» – Kiev, Ukraine, National Aviation University "NAU-Druk" Publishing House, October 13-14, 2010. – C.166-168. – engl.

[19] B. LeBaron "A Fast Algorithm for the BDS Statistic", Studies in Nonlinear Dynamics and Econometrics. 1997. Vol. 2. No. 2. P. 53-59.

[20] D. Chappell J. Padmore and C. Ellis. "A note on the distribution of BDS statistics for a real exchange rate series", Oxford Bulletin of Economics and Statistics, 58, 3, 561- 566, 1996.