ANNALES UMCS INFORMATICA DOI: 10.17951/AI.2016.16.2.14

VoIP Anomaly Detection - selected methods of statistical analysis

Paweł Dymora
Dept. of Power Electronics and Power Engineering and Complex Systems, Rzeszow University of Technology
Rzeszow, Poland
Pawel.Dymora@prz.edu.pl

Sławomir Jaskółka
Rzeszow University of Technology
Rzeszow, Poland
slawomir.jaskolka@gmail.com

Miroslaw Mazurek
Dept. of Power Electronics and Power Engineering and Complex Systems, Rzeszow University of Technology
Rzeszow, Poland
Miroslaw.Mazurek@prz.edu.pl

***Abstract* —** Self-similarity analysis and anomaly detection in networks are interesting fields of research and scientific work of scientists around the world. Simulation studies have demonstrated that the Hurst parameter estimation can be used to detect traffic anomaly. The actual network traffic is self-similar or long-range dependent. The dramatic expansion of applications on modern networks gives rise to a fundamental challenge to network security. The Hurst values are compared with confidence intervals of normal values to detect anomaly in VoIP.

Keywords — Hurst factor, anomaly detection, self-similarity, long-range dependence.

I. INTRODUCTION

Statistical analysis of network traffic measurements shows a clear presence of the fractal or self-similar properties in computer network [1, 4]. The statistical characteristics of computer network traffic have been on interests to scientists for many years, not least to obtain a better understanding of the factors that affect the performance and scalability of large systems such as the Internet. Research on network anomaly detection is very challenging and has started many years ago. New generation networks are based on Voice over Internet Protocol (VoIP) applications where the transmission quality depends on packet delay parameters. VoIP is crucial for many businesses due to VoIP monitoring and troubleshooting is one of the main tasks in network analyzers [7]. Faults as well other anomalies detection and prediction in VoIP networks using traditional tools is insufficient so the VoIP networks analysis in context of the QoS needs a new traffic models especially based on nonextensive statistics which in this article we propose.

II. SELECTED METHODS FOR DETERMINING THE HURST COEFFICIENT

Stochastic processes are considered as sequences of variables, which can be characterized by using average value, variance, process probability distribution value and higher stochastic moments. A stationary process is a stochastic process for which the process probability distribution value does not change. A characteristic element of a part of stochastic processes is the fact, that the values are mutually dependent in time. A value of a given process in a moment t is dependent on the value the process obtained in a moment preceding the moment t. However, such a process can be a stationary process, because a process probability distribution value in a moment t is determined without any premises about the time preceding the process [3, 6, 7].

The most commonly applied value which characterizes self-similar processes is Hurst exponent. There are many different ways of calculating it and most of those methods of estimation do not need to use the autocorrelation function. The Hurst exponent can take values from the range $(0,1)$. In accordance with the above formula the range of the modulus $(0.5,1)$ relates to the range $(0,1)$ of exponent β value [1, 2]. The Hurst exponent which equals 0.5 is characteristic for the process with independent realizations in different time ranges, which comes from the property $H = 1 - \beta / 2$. Another thing which can be deduced from this property is the fact that Hurst exponent equals 1 for the process with identical realizations in different time ranges. However, when Hurst exponent is <0.5 the realizations for different time ranges are mutually negatively correlated [1, 5].

The oldest method for determining Hurst coefficient is the method implementing rescaled range statistics (R/S method). In details the procedure for determining of the Hurst coefficient with the use of the R/S method is presented in the following algorithm.

ALGORITHM 1. THE R/S METHOD FOR DETERMINING OF THE HURST COEFFICIENT.

[Step 1] For a given finite length of recorded time intervals between the events $t_1, t_2, \ldots, t_n$ to obtain the R/S statistic in the first step the average interval for registered series of time intervals between events is introduced:

$$\hat{t}(n) = \frac{1}{n}\sum_{i=1}^{n} t_i$$

[Step 2] Then the variance *S (n)* of time series is obtained and the moment T_k of the k -th event appearance:

$$S^2(n) = \frac{1}{n}\sum_{i=1}^{n}\left[t_i - \hat{t}(n)\right]^2 \quad T_k = \sum_{i=1}^{k} t_i \,, \quad \text{for } k = 1,2,\ldots,n$$

[Step 3] It should also be taken into account the real deviation T_k of the actual moment of the k -th event appearance in an average moment $k\hat{t}(n)$, where $k = 1,2,\ldots,n: U_k = T_k - k\hat{t}(n)$, for $k = 1,2,\ldots,n$.

[Step 4] In this step the R/S statistic is calculated from the following formula:

$$\frac{R(n)}{S(n)} = \frac{\max(0,U_1,U_2,\ldots,U_n) - \min(0,U_1,U_2,\ldots,U_n)}{\sqrt{S^2(n)}}$$

[Step 5] For the self-similarity phenomena the following relationship can be observed where the variable H is the Hurst exponent:

$$\frac{R(n)}{S(n)} \approx n^H$$

[Step 6] The final step is to create a chart for R/S statistics as a function of time scale the double logarithmic coordinate system should be used, where the following relations will be obtained which is a straight line with a direction coefficient H [1, 4].

$$\log\frac{R(n)}{S(n)} \approx H\log(n) + const$$

Another method for determining the Hurst coefficient is the method of absolute value. In order to determine the Hurst coefficient value with this method the following algorithm is proceeded.

ALGORITHM 2. THE ABSOLUTE VALUE METHOD FOR DETERMINING OF THE HURST COEFFICIENT.

[Step 1] The aggregated time series $X(m)$ is formed by dividing the number of tested observations of length N into blocks of length m and averaging each block:

$$X^{(m)}(k) = \frac{1}{m}\sum_{i=(k-1)m+1}^{km} X_i \,, \quad \text{for } k = 1,2,\ldots,N/m$$

[Step 2] Analyze the n -th element of the time series to obtain $AM_n(m)$ where X is an average of the time series. The aggregated series $X^{(m)}$ for a large values of m asymptotically tends to $Cm^{n(H-1)}$ and $AM_n(m)$ is asymptomatically proportional to $m^{n(H-1)}$.

$$AM_n(m) = \frac{1}{N/m}\sum_{k=1}^{N/m}\left|X^{(m)}(k) - \overline{X}\right|^n$$

[Step 3] Apply the calculated values $AM_n(m)$ in the graph with double logarithmic scale and approximating the obtained points with the least squares method we obtain a straight with a slope to the X axis equal to $H-1$ [2].

The last analyzed method is the aggregate variance method. In order to determine the Hurst coefficient value with this method the following step algorithm is proceeded.

ALGORITHM 3. THE AGGREGATE VARIANCE METHOD FOR DETERMINING OF THE HURST COEFFICIENT.

[Step 1] The aggregated time series $X(m)$ is formed, and then the tested observation time series of length N into blocks of length m and averaging each block.

$$X^{(m)}(k) = \frac{1}{m}\sum_{i=(k-1)m+1}^{km} X_i \,,$$

for $k = 1,2,\ldots,N/m$

[Step 2] For each m value in the range from 2 to $N/2$ the variance is calculated for a sample $X^{(m)}(k)$:

$$VarX^{(m)} = \frac{1}{([N/m]-1)}\sum_{k=1}^{[N/m]}(X^{(m)}(k) - \overline{X})^2$$

[Step 3] In the graph the points $\log(s^2 m)$ and $\log(m)$ are indicated, which for sufficiently high value m, are arranged in a straight line with a negative direction coefficient equal to $2H-2$ [1, 2].

The first of the analyzed parameters is a number of interruptions in a second. To interpret the obtained values of the Hurst exponent we will use two dependencies:

1. If $0.5 < H < 1.0$, long-range dependencies occur, the process has a constant change direction (is persistent) and is stationary.

2. If $H < 0.5$, short-range dependencies occur, the process does not have a constant change direction and is not stationary [4].

III. Network architecture and subject of analysis

Simulation of network traffic in a computer network is modeled in the OPNET Modeler. OPNET Modeler is a program, which offers a variety of tools for virtual modeling and analysis of computer networks. The program allows simulations-ns wide range of different types of networks, interconnected in a manner selected by the user taking into account protocols, used technologies and relationships between devices. Tools built into the program OPNET Modeler enables the analysis of a number of parameters in the modeled networks such as load of devices that support the selected network services, bandwidth, latency between devices or packet loss.

The simulated network is divided into two subnets: the workstations subnet generating network traffic (described in the definition of profiles) and a server that is in another subnet (Fig. 1). Communication between workstations and a server located in another subnet requires the processing of traffic through a number of network devices that generates the delay. Server load and response times for workstations requests for each type of traffic is measured and recorded during the simulation. The program enables defining and testing of all aspects of the computer network behavior and is an excellent tool to eliminate the risk associated with a testing new network solutions for the real computer network.

IV. Framework of network traffic anomaly detection

The scheme of network traffic anomaly detection based on self-similarity may consist of few modules. We suggest scheme based four modules: traffic collection, statistical analysis, Hurst factor estimation and anomaly detection (Fig. 2).

In order to reduce the impact on normal use of network, when collecting LAN traffic, traffic on router is mirrored to traffic collection server. Packets received from router are processed. We can extract some traffic metrics like number of packets, the total length of the packet. The study aimed to observe network traffic and to determine whether there are long-term dependencies in the all network working time and above-hour intervals. In order to carry out the work of all captured packets we isolated ones that had the greatest impact on the network. They were divided in the terms of services and protocols on few main groups. This paper presents the VoIP protocol.

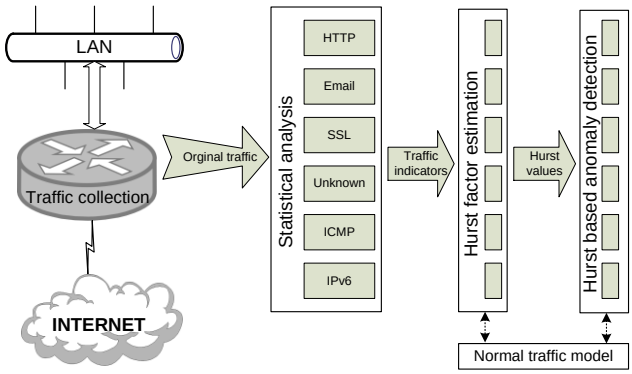

Fig. 1. Network traffic anomaly detection scheme

Next Hurst values of selected traffic metrics are calculated. The values can be used to detect traffic anomaly. The current calculated Hurst value is compared to normal traffic model. If the value deviates from normal model, current traffic is thought to be anomalous and normal otherwise. The assumption is that the Hurst parameter will remain relatively stable.

V. A detailed analysis of VoIP traffic using LABVIEW application

To simulate voice transmission codecs G.723.1 and G711.1 were used which have silence suppression function implemented in order to save bandwidth [11,12]. The simulation results are shown in Fig. 4. The use of G.723.1 codec to simulate the VoIP traffic between the server and the twelve workstations loaded the server with a traffic of variable intensity to 16 Kbit/s. G.711.1 codec transmits high-quality sound what loaded the network bandwidth and generated the traffic at an average of 50 Kbit/s. The uniform distribution of the network delay for the sound transmission is a priority in both cases (except for a short time while creating connection) is fixed and equal to 22 ms.

The application allows to measure the Hurst exponent computation time for each method. In order to isolate the test environment, the calculations were performed on a virtual machine in the "Virtual Box" version 4.3.10 environment, using a two-thread core of the Intel® Core ™ i5-2410M processor, which is clocked at a speed of 2.3 GHz (2.9 GHz Turbo). Virtual machine was assigned with 4 GB of RAM. The results of calculations performed for a small VoIP traffic is shown in Fig. 3 and Fig. 5.

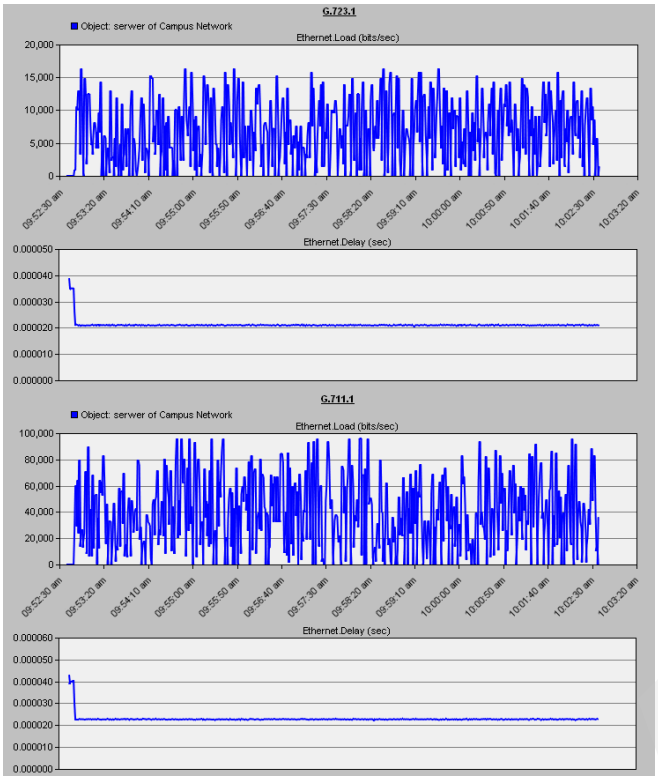

Fig. 2. Statistics of the server load and the network delays during VoIP communication.

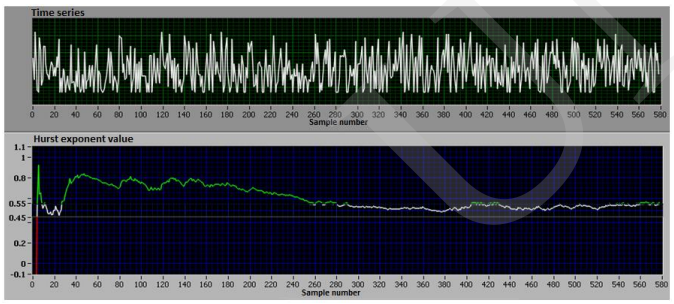

Fig. 3. The charts of the VoIP traffic with a small load and the Hurst exponent value obtained using the absolute value method.

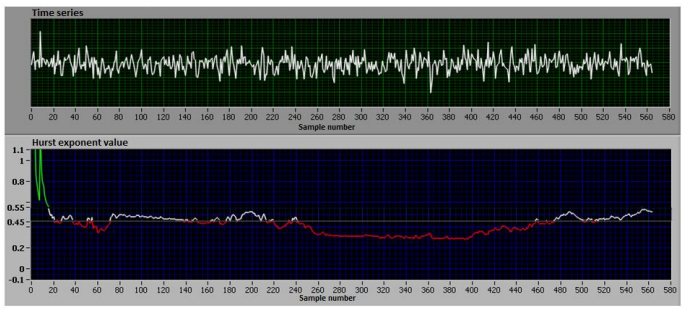

Fig. 4. The charts of the network delays for the VoIP traffic with a heavy load and the Hurst exponent value obtained using the absolute value method.

Hurst exponent for a small VoIP traffic has a significant deviation in the direction of persistent traffic, but the chart is stabilized between 0.47 and 0.59 values, which means that the traffic has little effect of the long memory. Chart of the small VoIP traffic is more similar to Brownian motion or white noise.

The delays for the small VoIP traffic has similar values for the self-similarity parameter as the network traffic itself. The method of the absolute value needed only 16 samples of the signal in order to determine the level of the self-similarity that (except small variations) persists to the end of the measurement. In the chart the deviation index at the direction of antipersistent traffic can be seen and (in the central part of the chart) that at the end of the simulation is back to the previous level by giving results between 0.38 and 0.57.

An analogous test was performed for severe bandwidth load. The results of the performed calculations for this case are shown in Fig. 4 and Fig. 6.

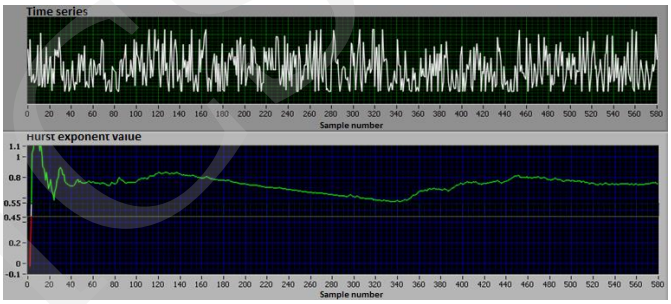

Fig. 5. The charts of the VoIP traffic with a small load and the Hurst exponent value obtained using the absolute value method.

Increasing the intensity of VoIP traffic has increased the value of the Hurst index from 0.59 to 0.81, depending on the selected method and data window width. Similarly as during a low load of the VoIP traffic methods for determining the Hurst index needed a few signal samples to determine the level of self-similarity, which persists to the end of the simulation.

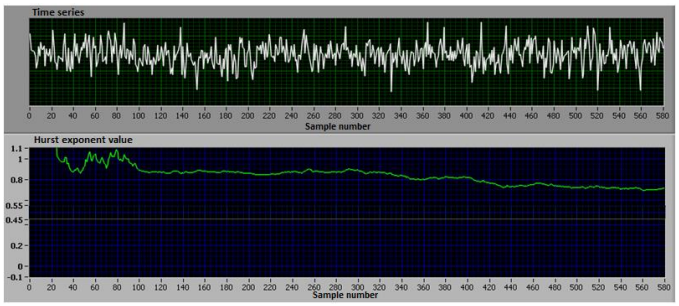

Fig. 6. The charts of the VoIP traffic with a heavy load and the Hurst exponent value obtained using the absolute value method.

Similarly as in the chart of the server load, the chart for a network delays for a heavy VoIP traffic is also increased compared to network delays that occur when modeling the small voice communication traffic. Table 1 provides a summary of the calculations for the VoIP traffic taking into account all types of the carried out studies.

TABLE 1. THE SUMMARY OF THE CALCULATIONS RESULTS FOR VOIP TRAFFIC

VoIP Traffic			Average value of the window width			All (100%)
			20%	30%	40%	
R/S statistic method	Heavy load	Server activity	0.61	0.64	0.69	0.77
		Network delays	0.52	0.58	0.61	0.67
	Small load	Server activity	0.47	0.49	0.54	0.59
		Network delays	**0.38**	0.54	0.59	0.55
Aggregate variance method	Heavy load	Server activity	**0.81**	0.66	0.66	0.71
		Network delays	0.56	0.64	0.67	0.65
	Small load	Server activity	0.59	0.57	0.56	0.57
		Network delays	0.39	0.49	0.46	0.52
Absolute value method	Heavy load	Server activity	0.59	0.65	0.67	0.71
		Network delays	0.48	0.58	0.66	0.67
	Small load	Server activity	0.55	0.59	0.57	0.57
		Network delays	0.41	0.49	0.48	0.53

When calculating the Hurst index for VoIP traffic an increase in the value of self-similarity with increasing traffic can be seen. This relationship also works well for network delays measured during simulation. Large variability of values makes that use of the short data window results in a wide range of Hurst coefficients. The carried out statistical analysis in the application indicates that in most cases the self-similarity of the network traffic regardless of its type, ranges from 0.5 to 1 for the scale of the Hurst factor. It may be noted that the Hurst coefficient becomes higher with increasing filling of the network bandwidth and continuous traffic of low density (eg. VoIP traffic) has a self-similarity comparable to a white noise equal to 0.5.

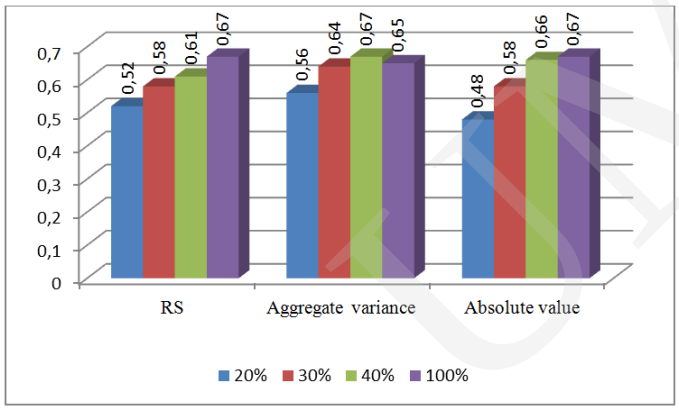

Fig. 7. Comparison of the Hurst coefficient value for a server activity with a heavy load.

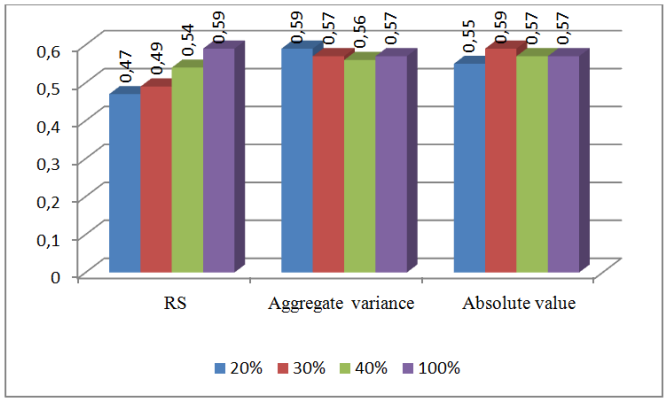

Fig. 8. Comparison of the Hurst coefficient value for a server activity with a small load.

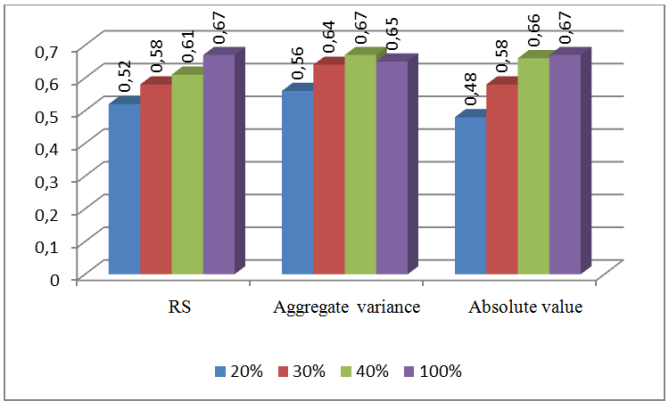

Fig. 9. Comparison of the Hurst coefficient value for a server delay with a heavy load.

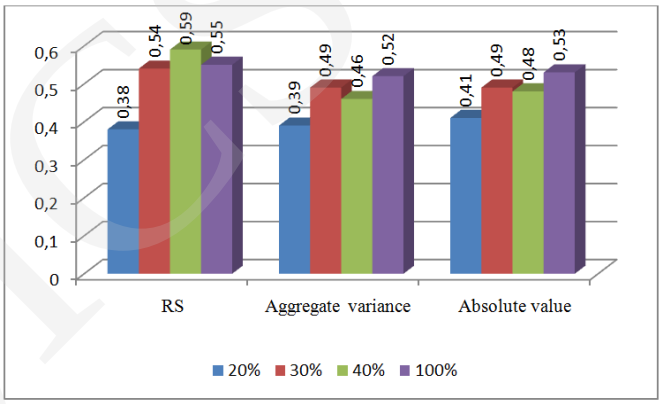

Fig. 10. Comparison of the Hurst coefficient value for a server delay with a small load.

In the case of small traffic reaching the value of the Hurst exponent equal to 0.5 (white noise), it is characterized by a complete randomness and a lack of correlation between packages. The average value of all Hurst coefficients included limited scope takes a similar value as the Hurst coefficient calculated for the entire range of data. The results of calculations for each of methods is usually not differ by more than 15-20%. Differences decreases with the closer the coefficient extremes. It may be noted that at the accuracy of the Hurst coefficient calculations influences the size of the data range. Reducing the data window increases the coefficient sensitivity to signal changes and causes that samples located in front of the scope of the data window do not affect the result of the calculation, which means that each Hurst coefficient is determined on the basis of the same number of signal samples. The beginnings of charts, in which the number of samples is not sufficient to fill the data window are ignored in the calculation. Setting the measuring window limiting the scope of the data to the last 20% of the samples of the entire range caused increase in sensitivity to small changes in the signal. Setting the measurement window constituting 15% of the total input data resulted in large and frequent fluctuations in the value of the Hurst coefficient. The use of a short data window and frequent change in the signal characteristics results in that the average Hurst exponents is in the range from 0.19 to 0.54 identifying

the chart as antipersistent series having tendency for frequent phrases to move on the graph [8].

The average value of the Hurst coefficients designated for the server activity in this simulation is in the range of 0.47 to 0.59 depending on the measuring window and the detection method. As can be seen the value of the Hurst coefficient varies depending on the load, e.g. for the method of the R/S statistic and 100% window width is in the range of 0,59 (low load) to 0.67 (high load) (Fig. 7 and Fig. 8).

Analogously the average value of the Hurst exponents obtained for the network traffic delays in this simulation is in the range of 0.48 to 0.67 at the heavy load (Fig. 9) and from 0.38 to 0.59 at small load (Fig. 10). The obtained result depends on the used measurement window and the detection method. Assuming for the first case (with an average value of 0.55) the greatest discrepancy result was obtained for the R/S statistic method and the window width of 20% which is equal to 0.47.

VI. Summary

The objective of presented research was to investigate the potential of using efficient classifier based on approximation function to estimate the Hurst parameter. By improving the capability of predicting impending network failures, it is possible to reduce network downtime and increase network reliability. The results obtained indicate that in most cases the self-similarity of traffic regardless of its type, ranges from 0.5 to 1 for the scale factor Hurst. It may be noted that the Hurst exponent becomes higher with increasing filling of the network bandwidth and continuous traffic of low density (e.g. a VoIP traffic) has a self-similarity comparable to a white noise equal to 0.5. Limiting the amount of data available for the calculations by setting a fixed number of samples occurring upstream of the calculations (using the data window) causes the rate of self often takes extreme values.

The average value of all the coefficients of the limited scope of Hurst takes a similar value as the Hurst coefficient calculated for the entire range of data. The results of calculation of the various methods is usually not differ by more than 15-20%. The differences decrease as a result of the closer this ratio extremes. The parameter H is larger when network utilization is higher. Low values of the Hurst index at high network load may indicate frequent changes of the transmitted network traffic type. It can be also concluded that the method of R/S statistics has the lowest computational complexity of the three implemented methods for determining the Hurst coefficient and is the least susceptible to slowing down the calculation due to the increase in input data.

The results confirmed that the VoIP has a self-similar nature to the degree of self-similarity in the range of 0.5 to 1 and can be used to detect the anomaly-behaved traffic efficiently.

References

[1] M. Mazurek, P. Dymora, “Network Anomaly Detection Based on the Statistical Self-similarity Factor”, Analysis and Simulation of Electrical and Computer Systems Lecture Notes in Electrical Engineering Volume 324, Springer, pp 271-287, 2015.

[2] M. Mazurek, P. Dymora, “Network anomaly detection based on the statistical self-similarity factor for HTTP protocol”, Przegląd elektrotechniczny, ISSN 0033-2097, R. 90 NR 1/2014, s.127 - 130, 2014.

[3] M. Fernandez-Martinez, M.A. Sanchez-Granero, J.E. Trinidad Segovia, “Measuring the self-similarity exponent in Levy stable processes of financial time series”, Physica A 392, Elsevier, pp 5330-5345, 2013.

[4] J. Cai, W. X. Liu, “A new Method of detecting network traffic anomalies”, Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering, 2013, pp. 2800 – 2803.

[5] P. Dymora, M. Mazurek, D. Strzałka, “Computer network traffic analysis with the use of statistical self-similarity factor”, t.XIII, s.69-81, 2013, Annales Universitatis Mariae Curie-Skłodowska Sectio AI Informatica, z. 2.

[6] H. D. Jeong, J. S. Lee, D. McNickle, K. Pawlikowski, “Self-similar properties of malicious teletraffic”, Int. J. Comput. Syst. Sci. Eng., 28 (1), pp. 1–7, 2012.

[7] H. Sengar, H. Wang, D. Wijesekera, and S. Jajodia, “Detecting VoIP Floods Using the Hellinger Distance”, IEEE Transactions on Parallel and Distributed Systems, Vol. 19, No. 6, pp. 794–805, 2008.

[8] F. Mata, J. Aracil, and J. L. García-Dorado, “Automated detection of load changes in large-scale networks,” in *Proceedings of TMA*, 2009, pp. 34–41.