Mirosław MAZUREK , Paweł DYMORA

Politechnika Rzeszowska, Zakład Systemów Rozproszonych

Network anomaly detection based on the statistical self-similarity factor for HTTP protocol

Streszczenie. *Analiza samopodobieństwa i wykrywania nieprawidłowości działania sieci stanowi interesujący problem dla naukowców na całym świecie. W artykule pokazano wykorzystanie współczynnika Hursta, jako parametru na podstawie którego można wykryć wszelkie anomalia pracy sieci. Odchylenia od wartości bazowej parametru Hursta w czasie pracy mogą sygnalizować nieprawidłowości działania. Badania mogą obejmować dowolny typ ruchu np. usługi HTTP. (**Wykrywanie nieprawidłowości działania sieci w oparciu o wykorzystanie statystycznego współczynnika samopodobieństwa na przykładzie protokołu HTTP**).*

Abstract*. Self-similarity analysis and anomaly detection in networks are interesting field 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 Hurst values are compared with confidence intervals of normal values to detect anomaly in few kinds of traffic: HTTP protocol, email, SSL.*

Słowa kluczowe: współczynnik Hursta, wykrywanie anomalii, samopodobieństwo, zależności długozasięgowe.
Keywords: Hurst factor, anomaly detection, self-similarity, long-range dependence.

doi:10.12915/pe.2014.01.30

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. Network traffic is inherently fractal or long-range dependent (LRD). The normal network behavior is used to detect network anomalies. That fact leads to question the extent to which the results of these studies are applicable in practice. Is it possible to diagnose network traffic and provide congestion risk. At the time being, there is mounting evidence that LRD is of fundamental importance for a number of engineering problems, such as traffic measurements [2, 5] and queuing behavior [7]. The similar processes have been observed and analyzed in a number of other areas like, for instance hydrology, economics, biophysics. A self-similar phenomenon represents a process displaying structural similarities across a wide range of scales of a specific dimension. Recent measurements of network traffic have shown that traffic exhibits variability at a wide range of scales [4, 10, 12].

The approaches used to address the anomaly detection problem are dependent on the nature of collected data. Network data can be obtained at multiple levels of granularity such as end-user-based or network-based. End-user-based data contains information which characterize the end application (refers to the transmission control protocol TCP and user datagram protocol UDP). Network-based data describe the functioning of the network devices and includes information gathered from network equipment's (routers, switches). Traffic counts obtained from both types of data can be used to generate time series to which statistical signal processing techniques can be applied. Statistical data analysis makes it possible to quantify network behavior and, therefore, describe network anomalies [8].

Self-similarity statistical factor

Self-similarity and fractals are notions pioneered by Benoit B. Mandelbrot. Self-similarity can be associated with "fractals", which are objects with unchanged appearances over different scales. Estimating the Hurst exponent for a data set provides a measure of whether the data is a pure white noise random process or has underlying trends. Brownian walks can be generated from a defined Hurst exponent. If the Hurst exponent is $0.5 < H < 1.0$, the random process will be a long memory process. Data sets like this are sometimes referred to as fractional Brownian motion. Fractional Brownian motion can be generated by a variety of methods, including spectral synthesis using either the Fourier transform or the wavelet transform. In the case of statistical fractals it is the probability density that repeats itself on every scale. On the other hand a dynamical fractal is generated by a low-dimensional dynamical system with chaotic solutions. Self-similarity means network traffic displays shape similarity in a very wide time scale. Hurst parameter (abbreviated as H) is an important index to measure the similarity which is often used in traffic congestion control and access control. Therefore, to estimate Hurst parameter accurately and rapidly has significance in network management and control.

A self-similar time series has the property that when aggregated (leading to a shorter time series in which each point is the sum of multiple original points) the new series have the same autocorrelation function as the original [3, 4, 5].

The main advantage of using models of self-similar patterns of the time series is that the degree of self-similarity of the series is expressed by only one parameter. The parameter expresses the speed of decay series autocorrelation function. The fractal dimension is directly related to the Hurst exponent for a statistically self-similar data set. A small Hurst exponent has a higher fractal dimension and a rougher surface. A larger Hurst exponent has a smaller fractional dimension and a smoother surface. For historical reasons, the parameter used is the *Hurst* parameter $H = 1 - \beta/2$. For self-similar series, $1/2 < H < 1$. As $H \rightarrow 1$ the degree of self-similarity increases. Thus, the main criterion for a series of self-similarity reduces the question of whether H is significantly different from the 1/2.

There are many ways determine the variance. We can use the variance-time plot, relies on the slowly decaying variance of a self-similar series. The variance of $X^{(m)}$ is plotted against m on a log-log plot; a straight line with slope (β) greater than -1 is indicative of self-similarity, and the parameter H is given by $H = 1 - \beta/2$. We can use R/S method. The R/S plot, uses the fact that for a self-similar dataset, the rescaled range or R/S statistic grows

according to a power law with exponent H as a function of the number of points included (n). Thus the plot of R/S against n on a log-log scale has a slope which is an estimation of H. The last approach, the periodogram method, uses the slope of the power spectrum of the series as frequency approaches zero. On a log-log plot the periodogram slope is a straight line with slope $\beta - 1 = 1 - 2H$ close to the origin [9].

Network architecture and subject of analysis

The scope of this paper is to describe the problem of IP network anomaly detection in a single administrative domain, in a private computer network (small company). The collected data were the result of the normal operation of programming between the hours of 10 am to 10 am the following day. The company has an eight-hour working time in two hourly intervals from 7:00 to 15:00 and from 9:00 to 17:00. It is possible that employees remained after hours. At night, all computers should be turned off, but this is not strictly adhered to.

The network uses 19 computers and network devices. All the computers were located in the same subnet and were connected via a switch to one of the server ports. The Server was connected to the Internet through a router. The analyzed network topology was shown in Fig. 1.

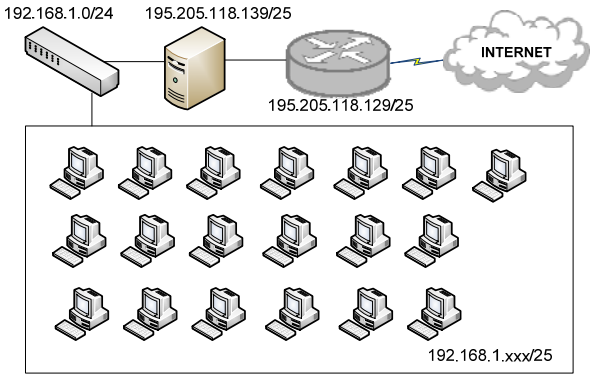

Fig. 1. Analyzed network architecture model

To collect the data we used one of the sniffer programs to capture packets – Wireshark. This program can save the movement from the level of the data link layer [4]. Captured traffic samples contain information such as the location of the file, its size, format, type of encapsulation and packet size limit, time of the first packet that is the start time of the test procedure and its completion, the total length of the work. In addition, provides information about the number and type of packets. Through 24 hours the analyzed network recorded 7 818 848 packets, the average speed was 90.447 packet per second.

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 only the HTTP protocol.

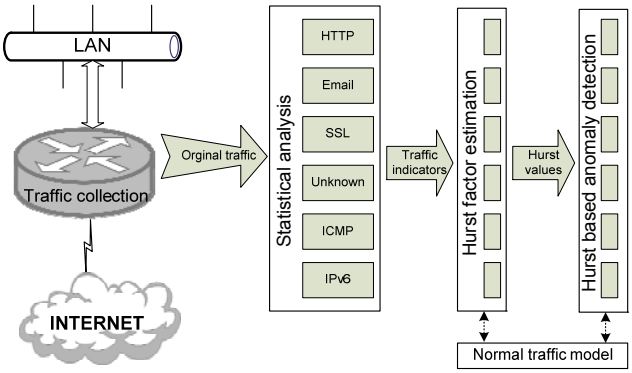

Fig. 2. Network traffic anomaly detection scheme

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.

Anomaly Traffic Detection

Network anomalies typically refer to circumstances when network operations deviate from normal network behavior. Network anomalies can rise due to various causes such as network overload, failure network devices, distributed denial of service attacks (DDoS), network intrusions, SYN floods, etc. These anomalous events will disturb the normal behavior of analyzed network [2, 6].

In this paper we propose to identify self-similarity anomalies based on the multi-time scaling nature and heavy-tailed distribution. Multi-time scaling nature means that the time series X and its time scaled version $x^{(m)}$, after normalizing must follow the same distribution.

The self-similar process also obeys the heavy-tailed distribution, under which it exhibits extreme variability. In the other word, a heavy-tailed distribution gives rise to very large values with a distribution results in content of the values being small but with a few samples having very large values [8, 9]. We note that each data set in the self-similar aggregated traffic appears to have different burstiness characteristics.

Under normal circumstances network traffic shows daily pattern. To reduce the impact on Hurst estimation caused by periodicity of network traffic whole time series is divided into 24 data sets. For each data set, the histogram of 24 equivalent time bins is computed. For each group number of packets, the total length of the packet and the average packet length in hourly intervals was calculated. There were also the largest and the smallest size packet. The next step was to calculate the Hurst factor by earlier estimation of β using the method of Benoit and Power Spectrum.

In practice, firstly a day of normal traffic is collected and traffic metrics mentioned before are extracted. When starting real time detection, the calculated Hurst factor is compared to normal traffic model for each metrics.

Exemplary power-spectrum density for the selected metrics at the same time operation of the network (14:00 – 5:00) shown in Fig. 3.

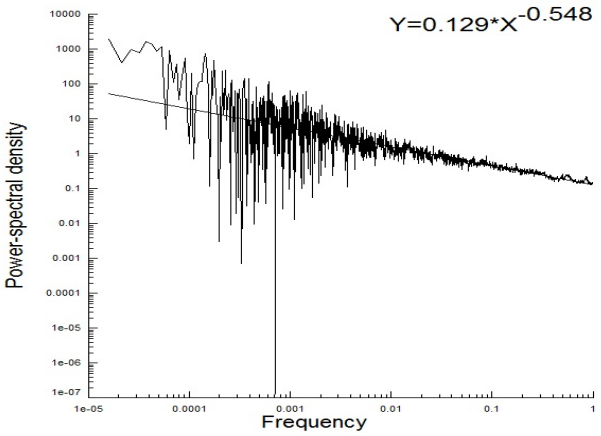

Fig. 3. Power-spectrum density of HTTP traffic for hours14:00 - 15:00

Results of investigations

In the following analysis, we collected a real trace from the private network with duration of 24 hours. The remaining flows during network (company) working time are treated as the "normal" traffic. First we studied the HTTP protocol. The HTTP is the most commonly used protocol that supports the Internet. The data are used only with TCP protocol and the default configuration uses port 80. Each object (e.g. website, video, audio) downloaded from a Web server sends through a single session.

Detailed test results for the HTTP protocol are shown in Table 1 and Fig.4. The gray in the table indicates the company's work time.

Table 1: Simulation of the http protocol

Time clock	Number of packets	The total length of the packets [b]	β	H
7:00 - 8:00	108370	133943057	-0,575	0,7875
8:00 - 9:00	100086	121579982	-0,559	0,7795
10:00 -11:00	123748	154772543	-0,532	0,766
11:00 -12:00	99778	121724775	-0,523	0,7615
12:00 -13:00	83546	98390225	-0,593	0,7965
13:00 -14:00	78579	41128195	-0,573	0,7865
14:00 -15:00	187423	249432486	-0,548	0,774
15:00 -16:00	48763	53698411	-0,633	0,8165
16:00 -17:00	17940	21589341	-0,743	0,8715
17:00 -18:00	6064	3951190	-0,164	0,582
18:00 -19:00	8211	6451138	-0,596	0,798
19:00 -20:00	10754	9900788	-0,576	0,788
20:00 -21:00	17989	18185640	-0,609	0,8045
21:00 -22:00	11229	8600740	-0,577	0,7885
22:00 -23:00	5430	3136815	-0,327	0,6635
23:00 -00:00	4958	3100426	-0,247	0,6235
00:00 -1:00	4982	3131088	-0,287	0,6435
1:00 - 2:00	4933	3086660	-0,178	0,589
2:00 - 3:00	4990	3151512	-0,021	0,5105
3:00 - 4:00	4855	2974499	-0,084	0,542
4:00 - 5:00	4865	3017560	-0,046	0,523
5:00 - 6:00	21652	27088116	-0,747	0,8735
6:00 -7:00	33960	41251848	-0,741	0,8705

An essential element of communication and business users is the e-mail. Messages that are send may use traffic spam. We send a text messages using standard protocols (POP3 and SMTP) which does not significantly affect network traffic. However, if the message contains a large attachment, sending and retrieving it may have a significant impact on the operation of the entire network.

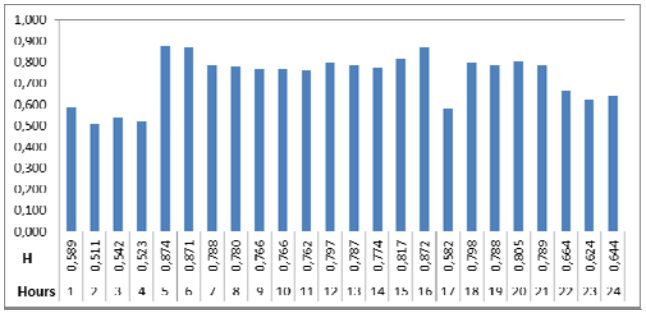

Fig. 4. Hourly detailed distribution of Hurst factor for HTTP protocol

Table 2: The average values of HTTP metrics of Network traffic

	Total H_T	Network busy time H_W	Free time H_F
The average Hurst value	0,729	0,791	0,686

The obtained results of this experiment show that in contradiction to so far existing belief the traffic can have a self-similar nature. As it can be seen such a property is well visible. Power spectral density can be used not only for proposal of analytical model but also to confirm the validity of proposed model basing on experiment.

We compare the Hurst factor estimation of each data set with the Hurst factor of the normal reference model (H_T, H_W – Table 2).

Table 3: Detailed analysis of variation for HTTP protocol

Hour	HTTP				
	H	H_T –H	% H_T	H_W -H	% H_W
1	**0,589**	0,140	19,20	0,202	27,71
2	**0,511**	0,219	29,97	0,281	38,48
3	**0,542**	0,187	25,65	0,249	34,16
4	**0,523**	0,206	28,26	0,268	36,76
5	**0,874**	-0,145	19,82	-0,083	11,32
6	**0,871**	-0,142	19,41	-0,080	10,91
7	**0,788**	-0,059	8,02	0,004	0,48
8	**0,780**	-0,051	6,93	0,012	1,58
9	**0,766**	-0,037	5,01	0,026	3,50
10	**0,766**	-0,037	5,08	0,025	3,43
11	**0,762**	-0,033	4,46	0,030	4,05
12	**0,797**	-0,068	9,26	-0,005	0,75
13	**0,787**	-0,058	7,89	0,005	0,62
14	**0,774**	-0,045	6,17	0,017	2,33
15	**0,817**	-0,088	12,00	-0,026	3,50
16	**0,872**	-0,143	19,55	-0,081	11,04
17	**0,582**	0,147	20,16	0,209	28,67
18	**0,798**	-0,069	9,47	-0,007	0,96
19	**0,788**	-0,059	8,09	0,003	0,41
20	**0,805**	-0,076	10,36	-0,014	1,85
21	**0,789**	-0,060	8,16	0,003	0,34
22	**0,664**	0,066	8,98	0,128	17,49
23	**0,624**	0,106	14,47	0,168	22,98
24	**0,644**	0,086	11,73	0,148	20,23

The analysis of selected protocols shows that network traffic is self-similar. The degree of self-similarity Hurst exponent is specified by the range 0.5 to 1. The shorter the average length of packet including Hurst exponent tends to 0.5 (white noise).

Any deviation from the mean value demonstrates some anomalies in the operation of the network (equipment failure, collisions) or anomalies in the behavior of users who are responsible for the generation of network traffic (absence of employees, slowness of employees,

purposefully running services at night, etc.). The bottom line of the analysis is the value of allowable deviations from the reference value. In this paper we assume a single acceptable percentages, as shown in Table 3.

The table shows the deviation of the total value and the value in working hours. As we can see in operation time (7:00 – 18:00) the deviations are of the order of 5-8%, while outside working hours significantly increased. It can be seen as an anomaly in the operation of the network, but on the other hand the situation is normal because nobody is working after 18:00. Anomaly are increasing factor H during the night, where the network should not work. This may indicate a network intrusion attempts, or other hazards.

Examination of the results shows high degree of H parameter stability during operation of the company. Approximately at 10 am we can see a significant increase in the metric labeled Unknown. This can be data sent by well-known protocols, but with changed by the administrator port numbers.

Conclusion

The objective of presented research was to investigate the potential of using efficient classifier based on approximation function to estimate the Hurst parameter for network traffic self-similarity measurements. The results confirmed that the analyzed traffic 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. The parameter H is larger when network utilization is higher. The degree of self-similarity is dependent of the volume of traffic and the type of service. In the case of minor traffic approaching the value of the exponent equal to 0.5 (white noise), is characterized by the complete randomness and lack of correlation between packets. By improving the capability of predicting impending network failures, it is possible to reduce network downtime and increase network reliability. The results of analytical considerations and experiment shows that self-similarity factor can be successfully used in the computer network traffic analysis, and eventually lead to more efficient methods for both failure and intrusion detection.

The future work is to define some parameters for differentiate abnormal communities from normal communities. In addition, we simulate some specific types of network anomalies in a laboratory environment and trace them.

REFERENCES

[1] Strzałka B., Mazurek M., Strzałka D., *Queue Performance in Presence of Long-Range Dependencies – an Empirical Study*, International Journal of Information Science, 2012, 2(4), pp. 47-53

[2] Yan R., Wang Y., *Hurst parameter for security evaluation of LAN traffic,* Information technology Journal 11 (20, 2012, pp. 269-275

[3] Dietmar S., *Algorithms for random fractals*, Chapter 2 of *The Science of Fractal Images* by Barnsley et al, Springer-Verlag, 1988

[4] Dymora P., Mazurek M., Strzałka D., *Computer network traffic analysis with the use of statistical self-similarity factor*, Annales UMCS Informatica, 2013, in printing.

[5] Idris M.Y., Abdullah A. H., Maarof M. A., *Self-similarity measurement methods for network traffic anomaly detection,* Preceedings of the Postgraduate Annual Research Seminar, 2005, pp. 244 – 248

[6] Thottan M., Ji Ch., *Anomaly detection in IP networks,* IEEE Transactions on signal processing, Vol. 51, No.8, 2003, pp. 2191 – 2204

[7] Dymora P., Mazurek M., Strzałka D., *Long-range dependencies in memory pages reads during man-compute system interaction*, Annales UMCS Informatica XII (2) 2012, pp. 49-58

[8] Kettani H., Gubner J. A., *A novel approach to the estimation of the Hurst parameter in self-similar traffic,* Proceedings of the 27th Annual IEEE Conference on Local Computer Networks, 2002, pp. 160 – 165

[9] Dymora P., Mazurek M., Strzałka D., *Influence of batch structure on cluster computing performance - complex systems approach*, Annales UMCS Informatica XII (1) 2012, pp. 57-66

[10] Strzałka D. *Non-extensive statistical mechanics – a possible basis for modeling processes in computer memory system,* Acta Physica Polonica A 117(4), 2010, pp. 652–657

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

[12] Dymora P., Mazurek M., Strzałka D., *Statistical mechanics of memory pages reads during man–computer system interaction*, Metody Informatyki Stosowanej, vol. 1/2011 (26), 2011, pp. 15-21

***Authors**: dr inż. Paweł Dymora, Politechnika Rzeszowska, Zakład Systemów Rozproszonych, ul. W. Pola 2, 35-959 Rzeszów, E-mail: pawel.dymora@prz.edu.pl; dr inż. Mirosław Mazurek, Politechnika Rzeszowska, Zakład Systemów Rozproszonych, ul. W. Pola 2, 35-959 Rzeszów, E-mail: miroslaw.mazurek@prz.edu.pl .*