The recurrence plot as a tool in the analysis of network traffic anomaly detection

Aida Saeed-Bagińska[1], Romuald Mosdorf[2]

[1]*The University of Finance and Management*
Ciepla 40, 15-472 Bialystok, Poland

[2]*The University of Finance and Management*
Ciepla 40, 15-472 Bialystok, Poland

aida.sb@op.pl, mosdorf@gmail.com

Abstract

In this paper the dynamical properties of ARP data for anomaly of internet traffic detection are presented. The local network traffic dynamics for single workstation has been observed experimentally on networks with 42 workstations.

In the paper RQA (Recurrence Quantification Analysis) measures have been analyzed. Three of them have been explored using RP windowing method: recurrence rate (RR), divergence (DIV) and the longest vertical line (VMAX). The RP windowing algorithm samples the explored ARP time series, as a result each RP window characterized the disintegration of recurrence plots. The RR and VMAX defined the low activity of ARP data flow. Inversely, the DIV identified the high activity of ARP time series.

The authors have proved that both, the recurrence plot technique and the RP windowing method as the tools in the analysis of network traffic anomaly detection are effective and accurate.

1. Introduction

The anomaly of internet traffic is one of the known ways to detect the security infringement in computer networks. Network traffic anomaly defines any departure from the established trend that is unique [1]. The fact of the huge number of data being collected during network traffic shows that long-term analysis should be taken into consideration.

Most of time series investigations formed by collected network packets in constant time intervals indicate series of multifractal properties [2]. This problem can be analyzed using non-linear methods of dynamics which identify the multifractal nature of time series [1]. The algorithms based-on multifractal characteristic of a time series database have been shown in works [13].

The authors in previous work [1] had analyzed the dynamics of the ARP time series database of a local Ethernet traffic based-on attractor reconstruction method. This approach allowed to visualization the dynamics of explored system in n-dimensional space.

One of the attempts in the analysis of network traffic anomalies is a signal analysis, presented in [14]. The authors' results have shown that wavelet filters are quite effective at exposing the details of both ambient and anomalous traffic. These results indicate that traffic anomaly detection mechanisms, based on deviation score techniques may be effective, however further development is necessary [14].

Another approach to identify anomalies in network traffic, the TARZAN algorithm, which detects surprising patterns in a time series database in linear space and time, has been presented in [12]. The authors' algorithm uses a suffix tree to efficiently encode the frequency of observed patterns and uses a Markov model to predict the expected frequency of previously unobserved patterns. Their work only considered only one feature extraction technique, based on local slopes [12].

In this paper, among others, the recurrence plot analysis is considered.

This method allows for a two-dimensional vision for tested network multi-dynamics and for the evaluation of nonlinear system aperiodicity [3].

The analysis presented in the paper should be considered as preliminary research and future analysis required in creating the systems to detect signs of network intrusions.

978-1-4244-7818-7/10/$26.00 ©2010 IEEE

2. The methods of nonlinear analyzes and results

The trajectories of nonlinear dynamical system in the phase space form objects called strange attractors of the structure resembling the fractal [3]. The analysis of strange attractor gives us information about the properties of dynamical system such as system complexity and its stability. Nonlinear analysis starts from attractor reconstruction. In nonlinear analysis the reconstruction of attractor in certain embedding dimension has been carried out using the stroboscope coordination. In this method subsequent co-ordinates of attractor points have been calculated basing on the subsequent samples, between which the distance is equal to time delay τ. The time delay is a multiplication of time between the samples. The image of the attractor in n-dimensional space depends upon time-delay τ. When the time-delay is too small, the attractor gets flattened, that makes further analysis of its structure impossible. The selection of time-delay value is of great significance in the analysis of the attractor properties. Therefore the analysis of the experimental data is initiated by determining the time-delay. For that purpose the autocorrelation function is calculated. Autocorrelation function allows identification of correlation between the subsequent samples. In case of chaotic data the value of autocorrelation function rapidly decrease when τ increase. Value of the time-delay τ is determined from the condition $C(\tau)\approx 0.5*C(0)$ [3].

Recurrence plot (RP) visualizes the recurrence of states x_i in a phase space. The RP enables us to investigate the recurrence of state in m-dimensional phase. The recurrence of a state at time i at a different time j is marked within black dots in the plot, where both axes are time axes. From the formal point of view the RP can be expressed as [4]:

$$R_{i,j} = \Theta\left(\varepsilon_i - \left\| x_i - x_j \right\|\right), \qquad x_i \in \mathcal{R}^n, \qquad i,j = 1 \ldots N \qquad (1)$$

where N is the number of considered states x_i, ε_i is a threshold distance, $\|\ \|$ a norm and Θ the Heaviside function.

Homogeneous RPs are typical for stationary systems in which relaxation times are short in comparison with the time of system investigation. Oscillating systems have RPs with diagonal oriented, periodic recurrent structures. For quasi-periodic systems the distances between the diagonal lines are different. The drift is caused by systems with slowly varying parameters which cause changes of brightens of the RP's upper-left and lower-right corners. Abrupt changes in the dynamics as well as extreme events cause white areas or bands in the RP [4].

The recurrence rate (*RR*) represents the percentage of recurrence points in RP [4]:

$$RR = \frac{1}{N^2} \sum_{i,j} R_{ij} \qquad (2)$$

The recurrence rate corresponds to the correlation sum [4]. The factor gets the values from 0 (0% no recurrence points) to 1 (100% all points are recurrent) [5].

A diagonal line occurs in the RP when a segment of the trajectory runs parallel to another segment and the distance between trajectories is less than ε. The length of this diagonal line is determined by the duration of this phenomenon. Determinism (DET) is a percentage of recurrence points which form diagonal lines [4,8]:

$$DET = \frac{\sum_{l=l\min}^{N} lP(l)}{\sum_{l=1}^{N} lP(l)} \qquad (3)$$

P(l) is a histogram of lengths *l* of diagonal lines.

Ratio RATIO is a ratio between DET and *RR* [4]. The coefficient RATIO Has the following form:

$$RATIO = N^2 \frac{\sum_{l=l_{\min}}^{N} lP(l)}{\left(\sum_{l=1}^{N} lP(l)\right)^2} \qquad (4)$$

The length of the longest diagonal line is calculated according to the following formula [4]:

$$L_{\max} = \max\left(\{L_i ; i = 1 \ldots N_l\}\right) \qquad (5)$$

The inverse of $L_{\max}$ is defined as [4]:

$$DIV = \frac{1}{L_{\max}} \qquad (6)$$

and it is related with the *KS* entropy of the system, i.e. with the sum of the positive Lyapunov exponents [4]. The positive Lyapunov exponent is a measure of instability of system trajectory. The shorter Lmax gets the more signal becomes unstable.

$$LAVG = \frac{\sum_{l=l_{\min}}^{N} lP(l)}{\sum_{l=l_{\min}}^{N} P(l)} \qquad (7)$$

LAVG is an average time in which two segments of trajectory stay close to one another. LAVG may be interpreted as expected time.

The longest vertical line VMAX is a measure of the longest time, in which a system state is steady:

$$V_{\max} = \max(\{V_i; i = 1 \ldots N_v\}) \tag{8}$$

Entropy defines a complexity of deterministic system structures [7]:

$$p(l) = \frac{P(l)}{\sum_{l=l\min}^{N} P(l)} \tag{9}$$

where *P(l)* is a histogram of lengths *l* of diagonal lines.

3. Analyzes and tests

Experimental network consisted of 42 workstations connected through several network switches and one router with Internet connection [11]. The local network traffic dynamics for single workstation have been observed. We capture frames transmitted from or into single workstation plus broadcast communication of local network traffic. For identification of changes in time the obtained series set of ARP data (over 2000 samples) was observed. Experimental data was collected during a time interval of one month from 28th January to 28th February 2008. The time series of a number of ARP frames counted in one minute interval was analyzed. The frames were captured using *tshark* software [6].

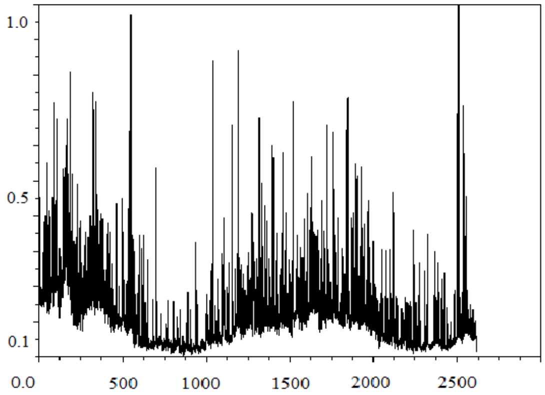


Fig. 1 The time series of ARP data

In this paper a time series classification relating to selected time periods by the detection of their differences or similarities is shown.

The experimental time series of recorded ARP data is presented in Fig. 1. The OX axis presents the number of ARP data samples while the OY axis contains normalized ARP frames activity.

3.1. Attractor reconstruction and recurrence plots

The time series of ARP data has been part for the sections. Each section had about 500 samples.

In Fig. 2 four parts of ARP time series with attractor reconstruction in 3-dimensional surface have been presented. A set of attractor points, both of the first section and the third is equal to 492. It has been presented in Fig. 2a) and 2c). In Fig. 2b) the activity intensification of ARP data is close to 1.0, which is adequate to the second part of ARP time series.

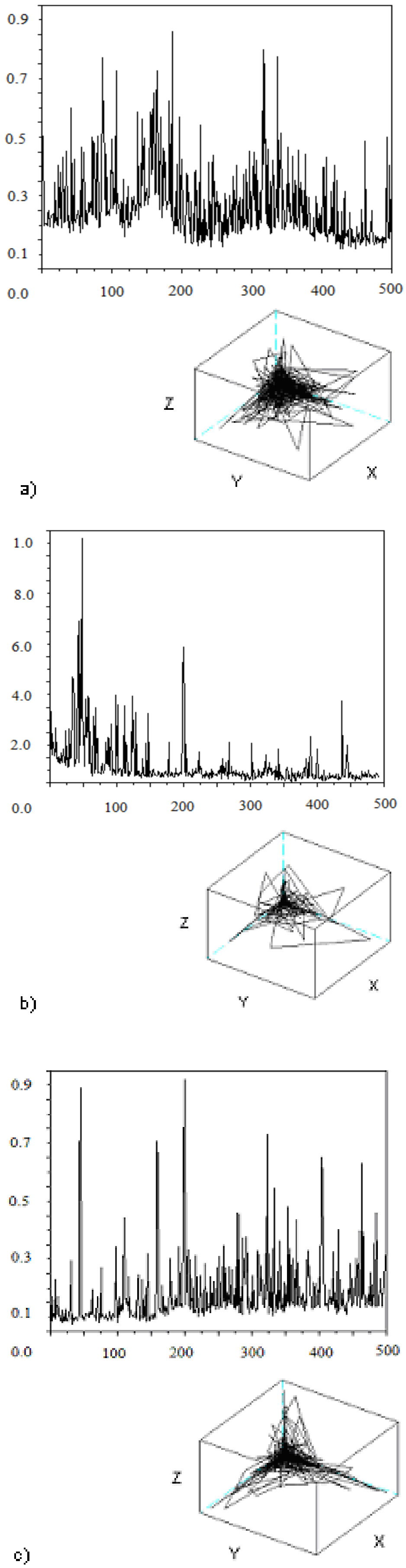
The results of attractor reconstructions presented in Fig. 2 a), b), c) and d) shown the chaotic character of ARP data flow.

One of the main advantages of RP method is possibility to use unstationary and short-range data. In Fig. 3 the recurrence plots obtained for different sections of ARP time series presented in Fig. 2a), b), c) and d) have been shown.

The activity intensification of ARP data is close to 1.0 in two different ranges of the graph (Fig. 1) and then decreases. It refers to the last days of months. The state is illustrated in Fig. 3b) as the black areas or bands. The same cases are presented in Fig. 3a) on the upper-right corner. Fig. 3c) introduces black areas uniformly distributed. Abrupt changes in experimental data cause white bands in the RP presented in Fig. 3. The horizontal or vertical lines in the recurrence plots note that system states changes very slow.

The activity of ARP data flow defines rather chaotic character. Presented recurrence plots if Fig. 3a), b), c) and d) have no parallel diagonal lines, with a same distance between them, which identify the time series period.

The results of the analysis shown in Fig. 2 and Fig. 3 point that in different terms of tested month, the procedures of data processing transferred by ARP protocol cause the generation of different number of ARP frames.



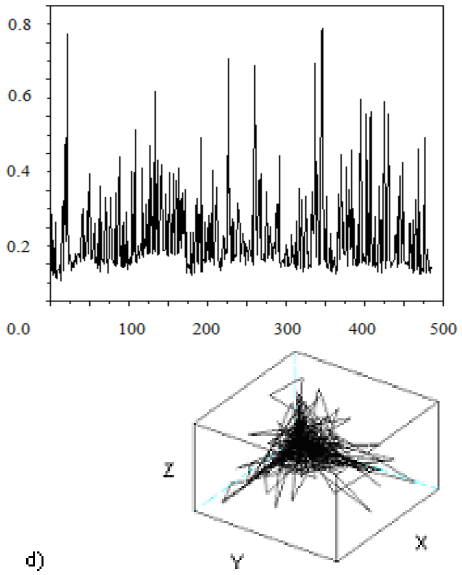


Fig. 2 Short parts of explored ARP time series with attractors reconstructed in 3-dimensional surface: a) the first section of ARP data (5-503 samples), b) the second (500-992), c) the third (1001-1499), d) the fourth (1517-2002)

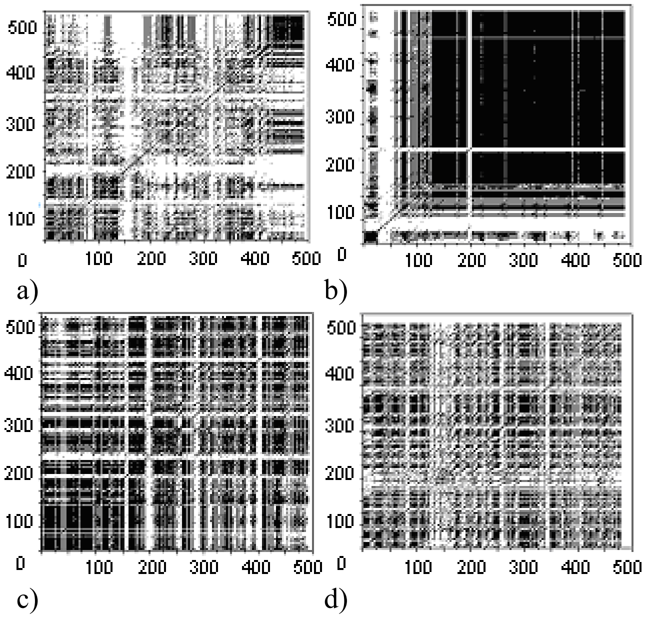


Fig. 3 Recurrence plots of explored ARP data series: a) the first section of ARP data (5-503 samples), b) the second (500-992), c) the third (1001-1499), d) the fourth (1517-2002)

3.2. RQA analysis

RQA (Recurrence Quantification Analysis) works on the basis of quantity and repetition time of dynamic system analysis [8,9,10]. RQA analysis contains selected measures, which are calculated to define a complexity of investigated time-series. Measured items, as RR, Ratio, DIV, ENT, etc. are shown in Table 1.

Table 1. The results of ARP time series analysis

	Recurrence quantification analysis (RQA)			
	dimension (dim) = 3, time delay (t) = 3, eps = 0.1			
The sections of ARP time series	5-503	500-992	1001-1499	1517-2002
The number of series samples	498	492	498	485
A set of attractor points	492	486	492	479
Recurrence rate (RR)	0,13455	0,53301	0,29529	0,19747
The longest vertical line (VMAX)	24	84	24	19
RATIO	7,43211	1,87612	3,38645	5,06391
The average length of a diagonal line (LAVG)	1,23929	3,58361	1,54141	1,28817
The longest diagonal line (LMAX)	23	81	24	14
Divergence (DIV)	0,04347	0,01234	0,04166	0,07142
Entropy (ENT)	0,5334	1,62778	0,91198	0,62714

The data has been reconstructed with the following parameters: dimension = 3, time τ = 3 and ε = 0.1. The recurrence rate (RR) represents the percentage of recurrence points in RP. The highest value of RR is a result of increase of complexity of attractors and is equal to 0.53301. Both, the values of the longest diagonal line and the longest vertical line are almost the same, particularly at the first sections of ARP data. The length of the longest diagonal line is not constant, in consequence the DIV changes.

In the paper the authors have presented the RP windowing analysis. The main aim of this method is sampling the explored ARP time series. The ARP data is sampling with a defined size and sampling step. Then stepwise a known sample (window) is being shifted by the whole ARP data. Each RP window characterized the disintegration of recurrence plots and the values are presented in the graphical illustrations.

The graphical results of RQA analysis are presented in Figures 4, 5 and 6:

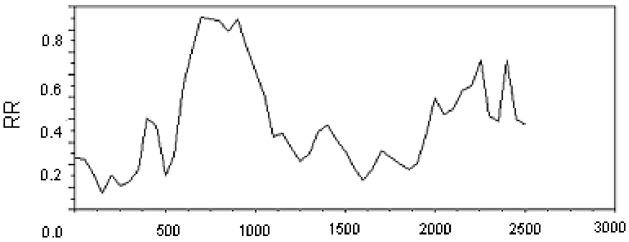


Fig. 4 The value of recurrence rate (RR) of ARP time series with a window size equals 100

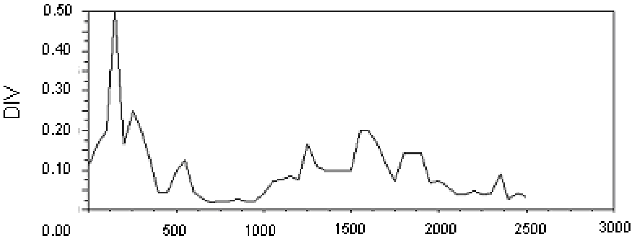


Fig. 5 The value of divergence (DIV) of ARP time series with a window size equals 100

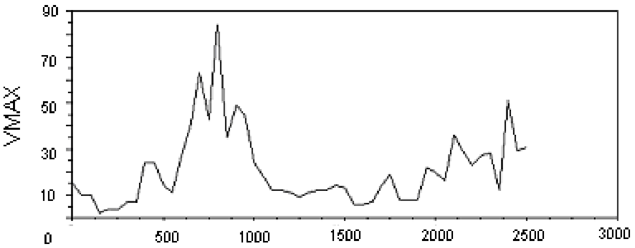


Fig. 6 The value of the longest vertical line (VMAX) of ARP time series with a window size equals 100

Figure 4. shows the value of recurrence rate (RR) of ARP time series. Fig. 5 presents the inverse of L_{max}, the value of divergence (DIV). The longest vertical line (VMAX) of ARP data has been presented in Fig. 6. It has been shown that, both, Fig. 4 and Fig. 6 are close to one another. Obtain results indicates, that the number of recurrence points decreases with decrease of ARP data flow. The same case is observed with the value of VMAX. It has been presented in Figures 4 and 6 in the section from 500-1000 samples. The recurrence plot shown in Fig. 3b) illustrated the vertical and horizontal black bands. The first section of ARP data, with 250-350 samples, has shown a lot of individual dots, means the process is random. The value of divergence of ARP time series appears inversely. The number of recurrence plots increase with increase of ARP data flow. It has been shown in Fig. 5. Received results suggested that in obtained cases, the ARP data flow have a chaotic character.

4. Conclusions

The main goal of this research was to explore the dynamical properties of Address Resolution Protocol (ARP) data for anomaly of internet traffic detection. In this paper the local network traffic dynamics for single workstation has been observed. Our experimental network consists of 42 workstations connected through several network switches and one router with Internet connection. In this article the recurrence plot (RP) analysis has been considered. The RP method used unstationary and short-range data. The RP has been made for 3D attractor reconstruction with ε equal to 0.1. It has been found that different sections of ARP data presented variety standards of RP. Horizontal and vertical lines in the recurrence plots indicate that the system states changes very slow. White areas or bands illustrated abrupt dynamic system changes. The activity intensification of ARP data have been shown in two principal points of ARP time series. It has been referred to the last days of months, when the activity of ARP data was the smallest. The RP illustrated also the states when the ARP data flow is distributed uniformly in a particular time. Black dots or clusters define the mentioned state.

In the paper it has been shown that the activity of ARP data flow has rather chaotic character. The recurrence plots have no parallel diagonal lines with the same distance that may identify the character of presented ARP time series.

The authors have introduced a technique called RP windowing analysis. The algorithm samples the explored ARP time series. In the result each RP window characterized the disintegration of recurrence plots. It has been proved that both, the recurrence plot technique and the RP windowing method as the tools in the analysis of network traffic anomaly detection are effective and accurate. The identification accuracy of anomaly detection we obtained is adequate to the selected RP window size.

The analysis presented in the paper should be considered as future analysis required in creating the systems to detect signs of network intrusions.

Acknowledgment

The authors are grateful to Dr. Oskar Świda for his help in accessing his whole data base on the subject of the ARP data flow.

5. References

[1] N. Siemieniuk, R. Mosdorf: Zastosowanie technologii informacyjnych do wspomagania zarządzania procesami gospodarczymi, pp. 277-286, Wyd. WSFiZ, Bialystok 2008 (in Polish)

[2] J. F. Muzy, E. Bacry, A. Arneodo, Multifractal formalizm for fractal signals: The structure function approach versus the wavelet-transform moduluj-maxima metod. Physical review E 47:875-884, 1993; A. Barth, G. Baumann, T. F. Nonnenmacher, J. Phys. A: Math. Gen. 25 381, 1992; J. F. Muzy, E. Bacry, A. Arneodo, Phys. Rev. Lett. 67 3515, 1991; Z. R. Struzik, Local Effective Holder Exponent Estimation on the Wavelet Transform Maxima Tree. Centre for Mathematics and Computer Science (CWI) Kruislaan 413, 1098SJAmsterdam – THENETHERLANDS email: Zbigniew.Struzik@cwi.nl

[3] H. G. Schuster: Deterministic chaos, An introduction, PWN, Warsaw 1993 (in Polish)

[4] N. Marwan, M. C. Romano, M. Thiel, J. Kurths: Recurrence Plots for the Analysis of Complex Systems, Physics Reports. 438(5-6), pp. 237-329, 2007 Recurrence Plots And Cross Recurrence Plots (*www.recurrence-plot.tk*)

[5] K. Ramirez-Amaro, J. Figueroa-Nazuno: *Recurrence Plot Analysis and its Application to Teleconnection Patterns*, 15th International Conference on Computing (CIC '06), 2006

[6] *http://www.wireshark.org*

[7] N. Marwan: Encounters With Neighbours – Current Developments Of Concepts Based On Recurrence Plots And Their Applications, Ph.D. Thesis, University of Potsdam 2003, ISBN 3-00-012347-4

[8] N. Marwan, M. Carmen Romano, M. Thiel, J. Kurths: Recurrence Plots for the Analysis of Complex Systems. Nonlinear Dynamics Group, Institute of Physics, University of Potsdam, Potsdam 14415, Germany 2007

[9] T. Grabowski: Zastosowanie Metody Recurrence Plots w analizie danych pomiarowych. Elektrotechnika i elektronika, Wyd. AGH, Kraków 2007

[10] N. Marwan: A historical review of recurrence plots. European Physical Journal: Special Topics, 2008

[11] O. Świda, A. Saeed-Bagińska, R. Mosdorf: Frequency and Fractal Analysis of Address Resolution Protocol Traffic. International Conference IEEE-CISIM'08, Ostrava, Czech Republic, 2008

[12] E. Keogh, S. Lonardi and W. Chiu: Finding Surprising Patterns in a Time Series Database In Linear Time and Space. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. July 23-26. Edmonton, Alberta, Canada, 2002. pp 550-556 (http://www.cs.ucr.edu/~eamonn/sigkdd_tarzan.pdf)

[13] A. Barth, G. Baumann, T. F. Nonnenmacher, J. Phys. A: Math. Gen. 25 381, 1992; J. F. Muzy, E. Bacry, A. Arneodo, Phys. Rev. Lett. 67 3515, !991; Z. R. Struzik, Local Effective Holder Exponent Estimation on the Wavelet Transform Maxima Tree. Centre for Mathematics and Computer Science (CWI) Kruislaan413, 1098SJAmsterdam – THENETHERLANDS email: Zbigniew.Struzik@cwi.nl

[14] P. Barford, J. Kline, D. Plonka, A. Ron: A Signal Analysis of Network Traffic Anomalies, In proceedings of ACM SIGCOMM Internet Measurement Workshop 2002