Internet Traffic Dynamics: Local Area Network Study

Lecturer Ph.D. Gabriela MIRCEA
Faculty of Economic Sciences, West University of Timişoara

We applied a nonlinear time series approach to the traffic measurements obtained at the input of a medium size local area network. In order to reconstruct the underlying dynamical system, we estimated the correlation length and the embedding dimension of traffic series. The estimated embedding dimension, based on the Grassberger – Procaccia algorithm, is high. In order to extract the regular part from the traffic data and to decrease the system's dimension, we filtered out high-frequency, "noisy" part, applying the wavelet filtering. Using the principal components analysis (PCA), we estimated the number of feature components in the traffic series. The reliable values of the correlation length and the embedding dimension provided the application of a layered neural network for identification and reconstruction of the dynamical system. We have found that the trained neural network reproduces the statistical features of real measurements and confirms the PCA result on the dimension of traffic series.

1 Introduction

A major challenge for the emerging high-speed integrated-services communication networks is to develop models that realistically capture the behavior of network traffic. The performance of the networks depends crucially on the traffic assessment. Complexity is a key issue in network geometry and information traffic. Evidence of traffic complexity appears in any forms, such as the long-range correlations and self-similarities found in the statistical analysis of traffic measurements [1-3]. There is also strong evidence of these phenomena at several time scales [4,5].

The complexity revealed from the traffic measurements has led to the suggestion that network traffic cannot be analyzed in the framework of available traffic models [6]. Alternative reliable traffic models and tools for quality assessment and control [7] should be developed.

Our study is directed towards a deeper understanding of the computer network traffic. We used data at the input of the West University of Timisoara [8] local area network (LAN), including approximately 200-250 interconnected computers.

In Section 2 we describe the data acquisition system of this LAN realized on the basis of a standard IBM PC in order to reconstruct the dynamical system, underlying the traffic measurements. In Section 3 we estimate the system parameters: the correlation length and the embedding dimension of the traffic series. In order to extract the regular component from traffic data and, therefore, decrease the dimension of system, we apply wavelet filtering to traffic measurements. The additional method used for estimation of traffic dimension is the principal components analysis (PCA). In Section 4, based on estimations of the correlation length and the traffic dimension, we reconstruct the underlying dynamical system applying a layered neural network. In Section 5 we analyze the statistical properties of time series generated by the artificial neural work (ANN) and compare them with real traffic measurements. We also estimate the dimension of the traffic data using the distribution of the ANN weights after its training.

2. Data acquisition system

The performance of the data acquisition system is based on realization of an open mode driver [9] (see Fig. 1).

In standard conditions the network adapter of a computer is in a mode of detecting a carrying signal (main harmonic 4-6 MHz). After appearing in the cable bits of the package preamble, the network adapter comes to a mode of 1 bit and 1 byte synchronization with the transmitter and starts receiving the

first bytes of the package heading. As soon as one succeeds in extracting the MAC-address of the shot receiver from the first bytes taken by the adapter, the network adapter compares it to its own. In the case of negative result of the comparison, the network adapter ceases to record the shot's bytes into its internal buffer and cleans its contents and then waits until the next package appears.

In order to provide conditions for receiving and analysis of all packages transmitted over the network, it is necessary to move the adapter devices to a free mode when all possible shots are recorded in the buffer. This operation is executed through the instructions of the NDIS driver.

The free mode driver records the accepted packages in the preliminary capture buffer and displays the flag of receiving the package. Then the receiving package module is activated and analysis of the margin of the margin of the package's type is carried out to extract TCP/IP packages from the whole stream.

After identification it is possible to separate and delete the data block as well as to record the headers to the SQL-server database. The recording is performed together with the time data with the frequency up to 10 kHz. Although the recording is performed with the buffering, the mode of saving the packages' headers requires enormous server's resources, as in this case there is a permanent procedure of recording with small portions to the hard disk. This is why this mode is switched on if required at the management system's instruction.

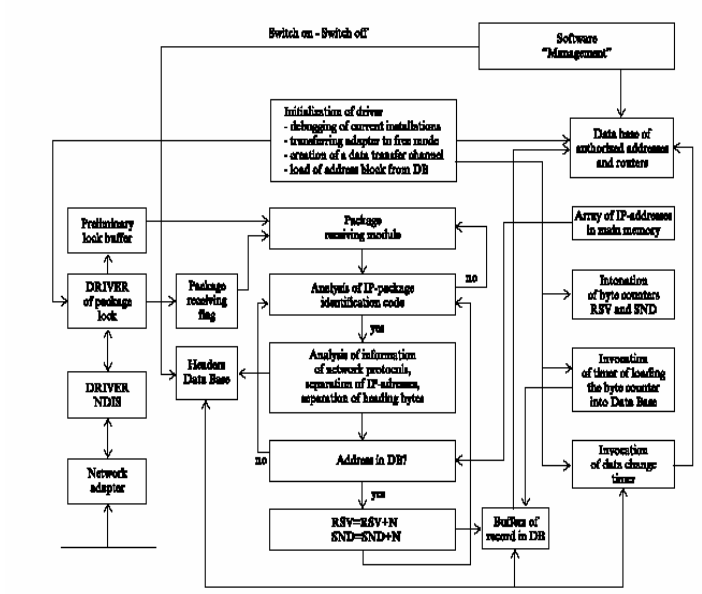

Fig.1. Scheme of a acquisition system (adapted from [35.])

The system also provides control over the external traffic of the local area network on the basis of controlling the records in the router table. Initial information on the legal IP address is saved in the database of the LAN computers from which data on legal addresses are loaded into the main memory array. The users which do not participate in forming the external traffic are not taken into account when calculating the number of transferred and receiving bytes. In order to decrease the number of sessions of recording the information on the external traffic database, a timer of load out of the buffer and a timer of changing a current date have been introduced into the system.

The recorded traffic data correspond approximately to 20 h (1,600,000 records) of measurements. The part of these series corresponding to 1 h of measurements is presented in Fig. 2.

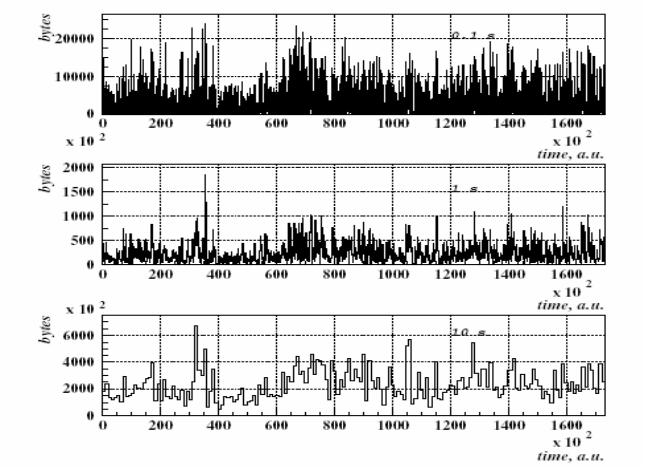

Fig.2. Traffic measurements aggregated with different bin size: 0.1, 1, and 10 s.

3. Estimation of dynamical system parameters

Chaos theory offers a new methodology, nonlinear or *chaotic time series analysis*, to handle irregular time series, such as traffic measurements [10]. First attempts to apply this approach to the network traffic analysis demonstrated serious difficulties as well as some promising results (see [11] and references therein).

In nonlinear time series analysis we view signal $\{x_i\}$ as the one-dimensional projection of a dynamical system operating in a space of vectors $\vec{y}$ of larger dimension [12, 13].

$$\vec{y}_i = (x_i, x_{i+\tau}, ..., x_{i+(m-1)\tau}). \qquad (1)$$

Here m is the dimension of the underlying dynamical system, and τ is a "delay time", or the correlation length of series $\{x_i\}$.

The main steps of this "*phase space reconstruction*" for the traffic measurements presented below.

3.1. Estimating the correlation length

In order to choose the independent components from the traffic data, we computed the correlation length [14, 15], where linear autocorrelation function

$$C(\tau) = \frac{\sum_{i-1}^{N}(x_{i+\tau} - \bar{x})}{\sum_{i-1}^{N}(x_i - \bar{x})^2} \qquad (2)$$

first time crosses the confidence tube corresponding to the Gaussian white noise. Here x_i are the values of traffic measurements, N is the number of points in the analyzed time series and $\bar{x} = \frac{1}{N}\sum_{i-1}^{N} x_i$.

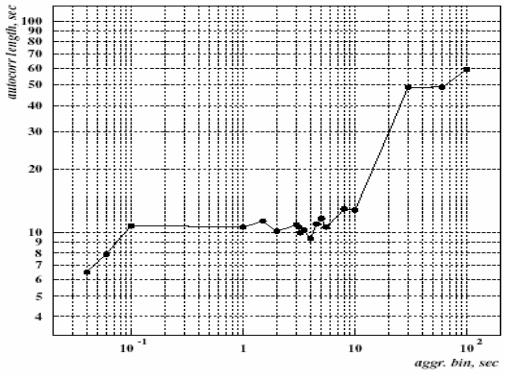

Fig.3. The dependence of the correlation length against the size of the aggregation bin.

The dependence of the correlation length against the aggregation bin size is presented in Fig. 3.

We see that for bin sizes from 0.1 up to 10 s, the correlation length τ is the acceptable region: $\tau \sim 10$ s. The points separated by the time interval τ can be considered as linear independent.

3.2. Estimating the embedding dimension

A set of uncorrelated points may be considered as the components of some m-dimensional vector. The dimension of the underlying process can be estimated by box-counting or neighbor counting methods [10]. To be sure that the dimension counting methods give a reliable result, one must check that starting from a certain value of n

(the dimension of the embedding space), the estimated dimension is not increasing together with the further increase of *m*. If this is the case, the time series can be considered as generated by the finite-dimensional system, which, in principle, can be reconstructed from the original time series.

The dimension counting for aggregated time series has been performed with the Grassberger-Procaccia algorithm [16, 17]. The correlation integral can be estimated by

$$C_n^m(r) = \frac{2}{N(N-1)} \sum_{i \neq j} \Theta(r - |y_i - y_j|) \quad (3)$$

with the distance between two points given by

$$|y_i - y_j| = \max\{|x_i - x_j|, \ldots, |x_{i+(m-1)\tau} - x_{j+(m-1)\tau}|\}.$$

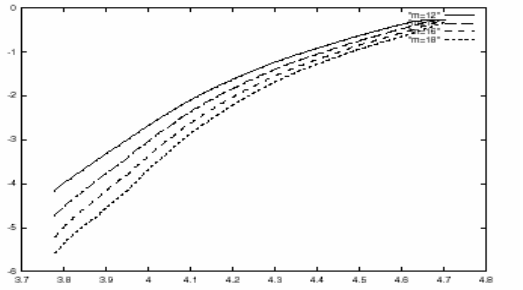

Fig. 4. Correlations integrals $C_2^m(r)$ for traffic measurements aggregated with 1 s bin: $\tau = 10s$ $m = 12, 14, 16, 18$

Here $\Theta = 1$ if this argument is non-negative and 0 otherwise. The value $C_2^m(r)$ is the empirical probability that the randomly chosen pair (y_i, y_j) of points will be separated by a distance of less or equal to r.

To estimate the embedding dimension d_E [16, 18] one computes $C_n^m(r)$ for r ranging from 0 to the largest possible value of $|y_i - y_j|$ and for m increasing from 1 up to the largest possible value. Starting from some *m* in the dependence $\log C_2(r) \approx \beta \log r + \gamma$ if the parameter β does change its value, then the embedding dimension d_E can be estimated from the relation $\beta < d_E < m$.

Thus, the slope of log $C_n^m(r)$ vs. $\log r$ gives the lowest estimate of embedding dimension (see Fig. 4).

For various parts of the time series we have analyzed, no saturation of the slope with respect to increasing m was found. For each given value of *m* in the range of *m* = 2 – 18 the slope β was found to satisfy $m < 2\beta + 1$.(4)

According to the Takens theorem [13], this may imply very high dimension of the studied time series [19].

3.3 Wavelet filtering

As usual we may consider the traffic measurements as a sum of a regular process and a stochastic part, related to the high-frequency "*noise*". The elimination of the *noisy* part may simplify the analyzed time series and reduce the dimension of the underlying dynamical process.

In order to achieve this, we used a discrete wavelet transform based on the Daubechies wavelets [20]. It is known that these wavelets provide high quality of filtering of both high- and low-frequency components of the analyzed signal [21]. After the direct wavelet transform, the decomposition coefficients, corresponding to the high-frequency component, were set to zero, and then, using the inverse wavelet transform, the regular component of the dynamical process has been restored. The difference between the original time series and the filtered signal, corresponds to the noisy component.

Fig. 5 presents part of the original traffic series, the corresponding filtered signal and the noisy component.

The auto-correlation function has been used as a criterion for the rejection of the noisy component: the points corresponding to the noisy component must be uncorrelated.

Fig. 6 gives the correlation length (time lag τ) for the filtered signal corresponding to different levels of threshold. It is clearly seen that the system corresponding to the regular traffic part changes its state at some levels of threshold. At the same time, the points corresponding to the noisy part remain uncorrelated.

After filtering out the noisy component from traffic measurements, we estimated the Grassberger-Procaccia dimension for the remaining regular part. Fig. 7 shows the correlation integrals $C_2^m(r)$ for traffic measurements aggregated with 1 s bin after wavelet

filtering: τ = 10 s and m = 10, 12, 14, 16, 18. We observed that for all curves, the slope of all log-log curves decreased in comparison to the slope calculated for the original (not filtered) data. The dimension about 16-18 seems to be close saturation. This observation shows that wavelet filtering can be useful for the separation of a reasonably regular component from traffic measurements.

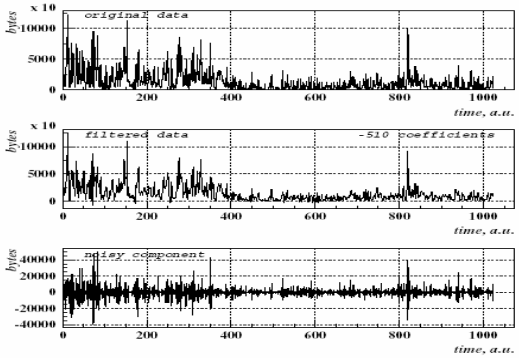

Fig.5. Traffic measurements: (a) original traffic series; (b) filtered signal; (c) noisy component.

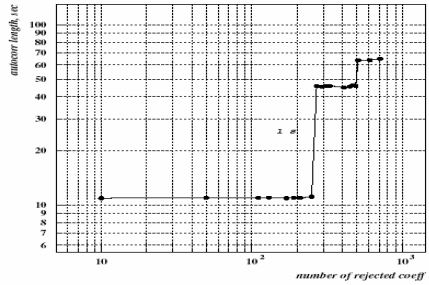

Fig.6. Correlation length (first touch of the auto-correlation function with the confidence tube corresponding to the Gaussian white noise) at different levels of thresholding.

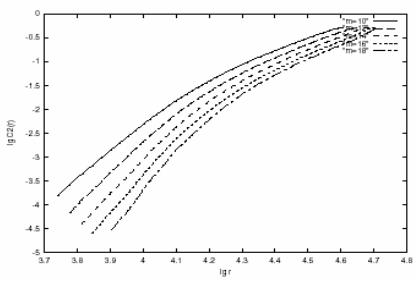

Fig.7. Correlations integrals $C_2^m(r)$ for traffic measurements aggregated with 1 s bin after the wavelet filtering: $\tau = 10\,s$ and $m = 10, 12, 14, 16, 18$.

3.4. Principal components analysis of traffic data

As additional method which has been used for classifications of the dimensionality problem is the PCA, which is also a well-known technique in multivariate data analysis [22-24]. The PCA method (also known as *Karhunen-Loève transformation* in communication theory [25, 26]) consists in applying a linear transformation to the original data space into a *feature space*, where the data set may be represented by a reduced number of "effective" features and yet retain most of the intrinsic information content of the data. In other words, the data set undergoes a dimensionality reduction.

We applied the PCA method to the original traffic data and to the filtered data with the different level of threshold. Fig. 8 shows the results of the PCA analysis.

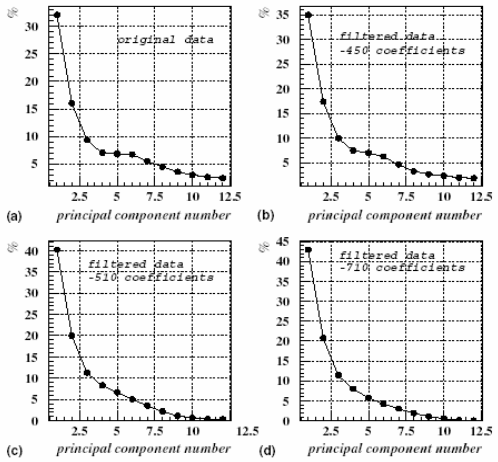

Fig. 8. Principal components: (a) for original data, (b)-(d) for time series filtered at a different thresholding level

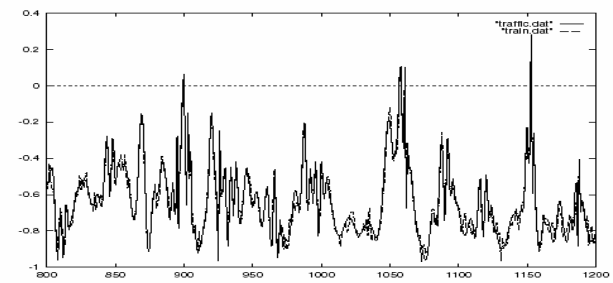

Fig. 9. The result of the ANN approximation of the traffic series after 1000 training epochs.

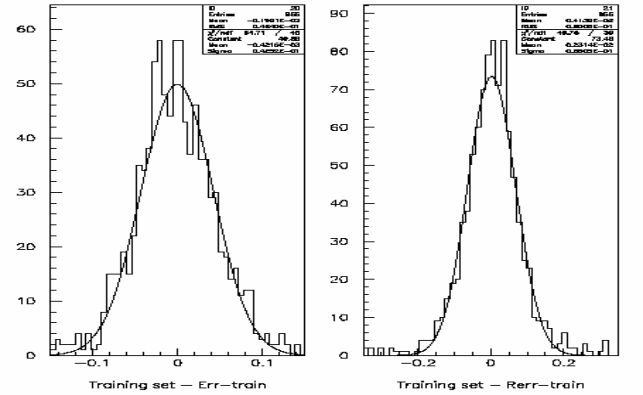

Fig. 10. Distribution of Errors (left) and Errors (right)

We se that the intrinsic information of the analyzed system is accumulated in the few first components and the contribution of this part is increasing as the threshold level increases.

4. Reconstruction of underlying dynamical system

In order to reconstruct the dynamical system corresponding to the traffic measurements, we used an ANN [10, 27, 28]. The major advantages of neural networks are that no prior information is required and the identification of the regular traffic component can be automatically obtained through the ANN training [29, 34, 35]. This is important in our case, not only because the traffic system is very complex, but there is also no information about the contribution of individual components into the system dynamics.

In our study we applied a layered neural network with the feed-forward architecture from the JETNET3 package [30]: the input layer with the numbe4r of neurons corresponding to the embedding dimension of the traffic series, two hidden layers with the varying number of neurons and one output neuron. From the output neuron we obtain the predicted value of the ANN model.

For the ANN training we used a data set corresponding approximately to 34 minutes period and aggregated with time bin 1 s. These data were preliminarily *cleaned* applying wavelet filtering (for the elimination of "noisy" component) and normalized to the interval [-1, 1]. The following parameters were used for the input vector (1) formation: τ = 10 s and d_E = 15-20.

Fig. 9 presents a part of the traffic data (traffic.dat) and the result of the ANN approximation (train.dat) after 1000 training epochs. Fig. 10 shows the statistical distributions of *Errors* (the difference between an actual value and the ANN approximation) and the *Relative Errors* (the ratio error/actual value) after 1000 training epochs. We see that, despite the highly chaotic character of time series, the neural network approximates these data quite well.

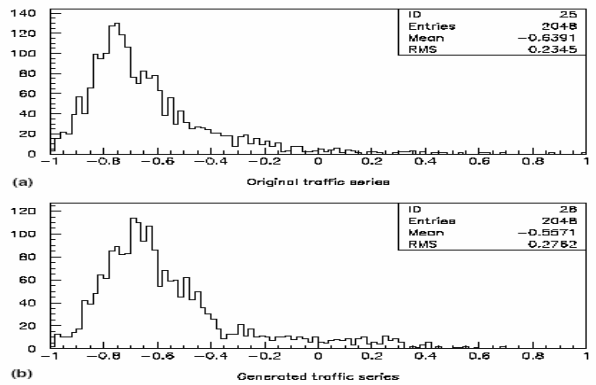

Fig. 11. The distribution of sizes of the traffic packages (normalized to the interval [-1, 1]) for: (a) the original traffic measurements, and (b) the generated by the trained ANN.

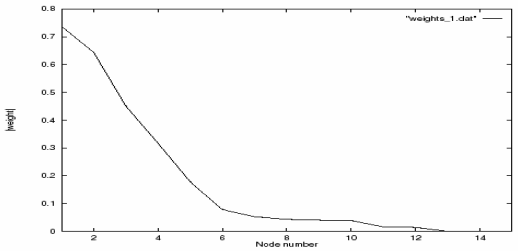

Fig.12. The distribution of the absolute values of weights values of weights between the out put node (neuron) of the ANN and the nodes of the second hidden layer of the trained on the traffic measurements

5. Nonlinear model of network traffic based on ANN

It is obvious that the above described ANN after training can be used as a model of simulation of traffic measurements. The trained ANN can be seen as a non-parametric model, because this approach does not require any assumptions concerning possible parameters of the analyzed time series. The intrinsic information is "extracted" by the ANN from traffic during the training process.

Let us see how this model reproduces the feature characteristics of real measurements. Fig.11 demonstrates the distributions of sizes of traffic packages (normalized to the interval [-1, 1]) for the original traffic measurements (top figure) and for time series generated by the trained ANN (bottom figure). We see that the ANN model reproduces quite well the statistical distribution of real data.

It is known that the ANN training on real data is in general adequate to the solution of the PCA problem [27, 31-33]. In this connection, the distribution of the ANN weights between the output node (neuron) of the ANN and the nodes of the second hidden layer is quite interesting (see Fig. 12).

One can see from Fig. 12 that the indicated distribution of weights reproduces the character of distributions obtained with help of the PCA method (see Fig. 8). Thus, the ANN model provided additional confirmation that the dynamical system underlying the traffic measurements has a dimension ≈ 12.

6. Conclusion

We applied systematically the nonlinear time series analysis approach to the traffic measurements obtained at the input of the intermediate size LAN. We demonstrated that nonlinear techniques can be successfully used for deeper understanding of main features of the traffic data. At the same time, we found that, due to the very complicated character of traffic series, the traditional algorithms of nonlinear analysis do not give reliable estimations of the analyzed time series. For instance, we found that the Grassberger-Procaccia algorithm gives very high dimension for raw traffic measurements. However, after the filtering out of the high-frequency component, which can be considered as a noise, we obtained a more realistic result for

the embedding dimension of the underlying process. This result was supported independently by the PCA method. Using the estimated values for the time lag and the embedding dimension we successfully applied a layered feed-forward neural network for the identification and reconstruction of the dynamical system underlying the traffic measurements. We have found out that the trained ANN reproduces statistical features of real traffic data and confirms the result of the PCA estimation of the traffic dimension.

References

[1.] Leland W, Taqqu M, Willinger W, Wilson D. *On the self-similar nature of Ethernet traffic* (extended version). IEEE/ACM Trans Network 1994; 1-15.

[2.] Lucas MT, Wrege De, Dempsey BJ, Weaver AC. *Statistical characterization of wide-area self-similar network traffic*, University of Virginia Technical Report CS97-04; October 9.

[3.] Crovella ME, Bestavros A. *Self-similary in word web traffic: evidence and possible causes*. IEEE/ACM Trans Network 1997; 5(6):835-46.

[4.] Vishal Misra, Wei-Bo Gong. *A hierarchical model for telegraffic.* Departament of Electrical and Computer Engineering, University of Massachusetts, Amherst MA 01003; 1998

[5.] Jon M.Peha. *Protocals can make traffic apper self-similar*. In Proceedings of the 1997 IEEA/ACM/SCS Communication Network and Distributed Systems Modeling and Simulation Conference.

[6.] Jagerman DL, Melamed B, Wilinger W, *Stochactic modeling of traffic process*. Technical Report 1999.

[7.] Tsunyi Tuan, Kihong Park, *Multiple time scale congestion control for self-similar network traffic.* Network System Lab, Departament of Computer Sciences, Purdue University, West Lafayette, IN 47907, USA, Elsevier Preprint, 2000.

[8.] The West University of Timisoara, http://www.fse.uvt.ro.

[9.] Vasiliev PM, Ivanov VV, Kryukov YuA, Kuptsov SI. *System for acquisition, analysis and management of network traffic for segment of the JINR computer network ,* JINR Communications, Russia, 2001.

[10.] Abarbanel HDI. *Analysis of observed chaotic data.*New York: Springer Verlag 1996.

[11.] Kugiumtzis D, Boundouriedes MA. *Chaotic analysis of internet ping data: just a random number generator?* Contributed paper on the SOEIS meeting at Bielefeld, 1998; March 27-28.

[12.] Packard NH, Crutchfield JP, Farner JD, Shaw RS, *Geometry from a time series*. Phys Rev Lett 1980; 45:712.

[13.] Takens F. *Detecting strange attractors in turbulence*. In: Rand D, Young LS, editors. Dynamical systems and turbulence. Lecture Notes in Mathematics, vol. 898. Berlin: Springer; 1981. p.336.

[14.] Broomhead DS, King GP. *Extracting qualitative dynamics from experimental data.* Physica D 1986; 20:217.

[15.] Albano AM, Muench J, Schwartz C, Mees AI, Rapp PE. *Singular value decomposition and the Grassberger-Proccacia algorithm*. Phys Rev A 1988; 38:3017.

[16.] Grassberger P, Procaccia I. *Characterization of strange attractors*. Phys Rev Lett 1983; 50:346.

[17.] Cutler CD. *A theory of correlation dimension for stationary time series*. Philos Trans R Soc Lond A 1994;348;343.

[18.] Grassberger P, Procaccia I. *Measuring the strangeness of strange attractors*. Physica D 1983;9:189.

[19.] Mircea G, Neamtu M, Opris D. *Dynamic systems in economics, mechanics, biology described by diferential equations with time delay.* Ed. Mirton, Timisoara, 2003.

[20.] Chui CK. In: *An introduction to wavelets*. New York: Academic Press; 1992. p. 1-18.

[21.] Press WH, Teukolsky SA, Vetterling WT, Flanery BP. In: *Numerical recipies in C:* the art of scientific computing. 2end ed. Cambridge: Cambridge University Press; 1988.p. 1992.

[22.] Preizendorfer RW. *Principal component analysis meteorology and oceanography*. New York: Elsevier; 1998.

[23.] Joliffe IT. *Principal component analysis*. New York: Springer; 1986.

[24.] Jackson JE, In: *A user's guide to principal component analysis*. New York: Willey; 1992. p. 26-62.

[25.] Karhunen K. *Uber lineare methoden in der Wahrscheinlichkeitsrechnung*, Annales Academiae Scientiarum Fennicae, series A1:Mathematica-Physica 37, 3-79 (Transl.: RAND corp., Santa Monica, CA, Rep. T-131, 1960).

[26.] Loeve M. *Probability theory* . 3rd ed. New York: Van Nostrand; 1963.

[27.] Haykin S. *Neural networks: a comprehensive foundation*. Englewood Chiffs, NJ: Prentice Hall; 1999.

[28.] Wasserman PD. Neural PD. *Neural computing: theory and practice*. New York: Van Nostrand reinhold; 1989.

[29.] Pham DT, Liu X. *Neural networks for identification, prediction and control.* London: Springer; 1995.

[30.] Peterson C, Rongvaldsson Th. *JETNET-3.0- A versatile artificial neural network package*, Lu Tp 93-29, 1993.

[31.] Oja E. *Data compression, features extraction, and autoassociation in feedforward neural networks.* In: Kohonen T, Makisara K, Simula O, Kangas J, editors. Artificial neural network, vol 1. Amsterdam: North – Holland; 1991. p. 737-46.

[32.] Oja E. Nonlinear PCA: *Algorithms and applications*. In: World Congress on Neural Networks, Portland, OR, vol 2. 1993. p. 396.

[33.] Mircea G, Opris D. *Internet congestion control model with feedback delay,* Economy Informatics, vol III, Number 1/2003, p. 78-84.

[34.] Mircea G, Neamtu M, Opris D. *Internet model with N access link and feedback delay*, Specialization Development & Integration, Cluj- Napoca, 2003, p. 394-400.

[35.] Lapedes A, Farber R. *Nonlinear signal processing using neural networks: prediction and system modeling*. Los Alamos Report LA-UR 87-2662;1987.