E11-2002-222

I. Antoniou[1,2], V. V. Ivanov[1,3], Valery V. Ivanov,
P. V. Zrelov

STATISTICAL MODEL OF NETWORK TRAFFIC

[1]International Solvay Institutes for Physics and Chemistry, CP-231,
ULB, Bd. du Triomphe, 1050, Brussels, Belgium
[2]Department of Mathematics, Aristoteles University of Thessaloniki,
54006 Thessaloniki, Greece
[3]Permanent address: Laboratory of Information Technologies,
Joint Institute for Nuclear Research, 141980, Dubna, Russia

Introduction

In [1, 2] we applied systematically the nonlinear time series analysis approach [7] to the traffic measurements obtained at the input of the intermediate size Local Area Network (LAN). We have demonstrated that nonlinear techniques can be successfully used for a deeper understanding of main features of the traffic data. In order to reconstruct the underlying dynamical system, we estimated the correlation length and the embedding dimension of 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 packet size distribution of real measurements, which follows the log-normal distribution [2].

The log-normal distribution has been first observed, to our knowledge, by Lucas et al. [8] for the empirical probability distributions of packet arrivals aggregated at $100\,ms$. Later they developed the background traffic model, or (M,P,S) model [9], which realistically generated the aggregated traffic flows for a large campus network. The log-normal distributions for packet arrivals have been observed at different stream scales [9]. Similar inter-arrival time distributions for channel arrivals have been observed in cellular telephony [10]. However, there was no a reliable explanation of reasons, which may cause the appearence of such distribution.

In our work [3], based on the detailed analysis of traffic measurements, we demonstrated that the reason of this distribution may be a simple aggregation of real data. In fact, we show that the aggregation of traffic measurements forms (starting from some threshold value of the aggregation window) a stable statistical distribution, which does not change its form with further increase of the aggregation window. Applying the χ^2-test we proved that with a high significance level this distribution corresponds to the log-normal distribution.

Later in [4] we proved that the Principal Components Analysis, especially the "Caterpillar"-SSA approach [11, 12], is very efficient for understanding main features of terms forming the network traffic. The statistical analysis of leading components demonstrated that a few first components already form the fundamental part of the information traffic [4]. The residual components play a role of small irregular variations, which do not fit in the basic part of the network traffic and can be interpreted as a stochastic noise.

In order to further decrease the dimension of the dynamical system underlying the network traffic, we applied the wavelet filtering to traffic measurements [5]. The analysis of influence of this preliminary filtering on characteristics of individual prin-

cipal components and on summary distributions of leading and residual components gave additional arguments for the correctness of results obtained in [4]. The Fourier analysis of original traffic measurements and individual principal components both for original and filtered data confirmed that the fundamental part of information traffic is formed by a few first leading components.

Applying the continious wavelet transform to traffic measurements, we found that the corresponding series has a multifractal, multiplicative character. This circumstance together with the log-normal distribution of traffic data confirms the applicability of the Kolmogorov's scheme [6] to the description of network traffic.

The aim of this work is to summarize the results of obtained in [1, 2, 3, 4, 5], to formulate main characteristics of the background statistical model of network traffic and to emphasize possible directions for further studies.

In our work we used traffic measurements collected at the input of Dubna University [13] LAN, which includes approximately 200-250 interconnected computers.

In Section 1 we describe the data acquisition system of this LAN, realized on the basis of a standard PC. In Section 2 we present first results of application of the nonlinear analysis to the traffic measurements. We show that the dynamical model based on the neural network reproduces the packet size distribution of real measurements, and that this distribution fits in the log-normal form. In Section 3 we explain that the reason of this distribution may be caused by a simple aggregation of real data. In Section 4, applying the Principal Components Analysis, we demonstrate that a few first components already form the basic part of the network traffic, while the residual components play a role of small irregular variations. In Section 5 we analyze the spectral characteristics of traffic measurements applying the Lomb periodogram technique. The peculiarities of the wavelet filtering of traffic series are considered in Section 6. Section 7 is devoted to the analysis of statistical and spectral characteristics of the filtered traffic series. In Section 8 we show that the main part of network traffic can be efficiently described by a minimal number of feature components. In Section 9 we show that the traffic measurements have a multifractal, multiplicative character, and discuss the applicability of the Kolmogorov's scheme to the description of network traffic.

1. Data acquisition system

Two protocols are used in the "Dubna" LAN. The NetBEUI protocol is applied only for internal exchanges, and the TCP/IP for external communications. The measurements of network traffic have been realized at the external side of the input lock of LAN.

The performance of the data acquisition system is based on realization of an open mode driver [14]: 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

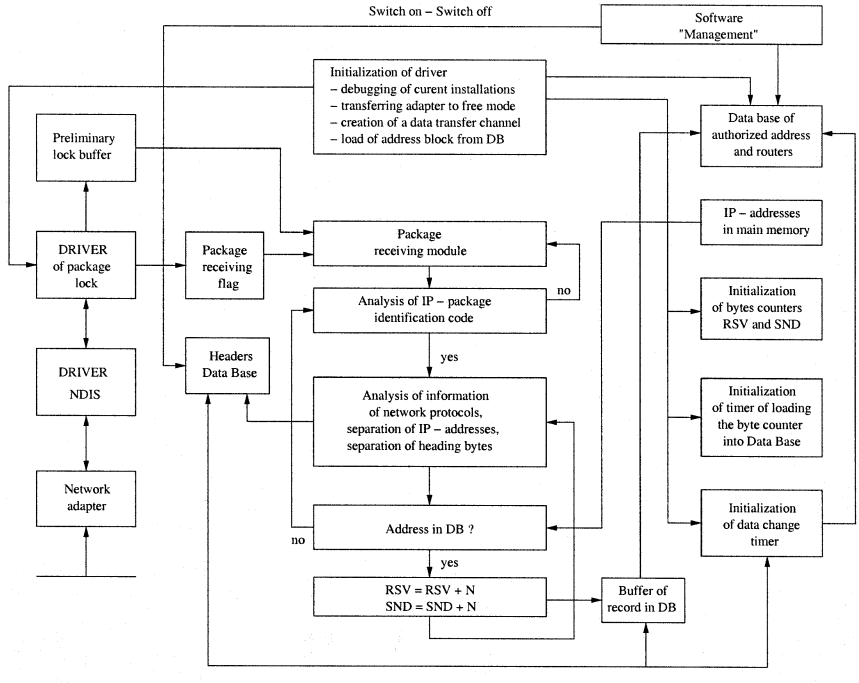

Figure 1: Scheme of a data acquisition system

of the package preamble, the network adapter comes to a mode of 1 bit and 1 byte synchronization with the transmitter and starts receiving 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 a 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 the 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 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 a frequency up to 10 kHz. Although the recording is performed with 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. That is why this mode is switched on if required at the management system's instruction.

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 addresses 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 received bytes. In order to decrease the number of sessions of recording the information on the external traffic in the 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 hours (1600000 records with a frequency up to 10 kHz, which corresponds to 1 ms bin size) of measurements. The part of this series corresponding approximately to 1 hour of measurements and aggregated with different bin sizes is presented in Fig. 2.

The contribution of the NetBEUI traffic has been estimated around 1-6 packages per second during daily working hours. This is negligibly small compared to the TCP/IP traffic. In this connection, we may neglect the influence of non-IP traffic on the TCP/IP traffic.

2. Nonlinear analysis of network traffic

Chaos theory offers a new methodology, nonlinear or *chaotic time series analysis*, to handle irregular time series, such as traffic measurements [7]. First attempts to apply this approach to the network traffic analysis demonstrated serious difficulties as well as some promising results (see [15] and references therein).

In nonlinear time series analysis we view the signal $\{x_i\}$ as the one-dimensional projection of a dynamical system operating in a space of vectors $\vec{y}_i$ of larger dimension [16, 17]:

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

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 include three main steps:

1. Estimation of the correlation length τ,
2. Estimation of the embedding dimension m,
3. Reconstruction of underlying dynamical system.

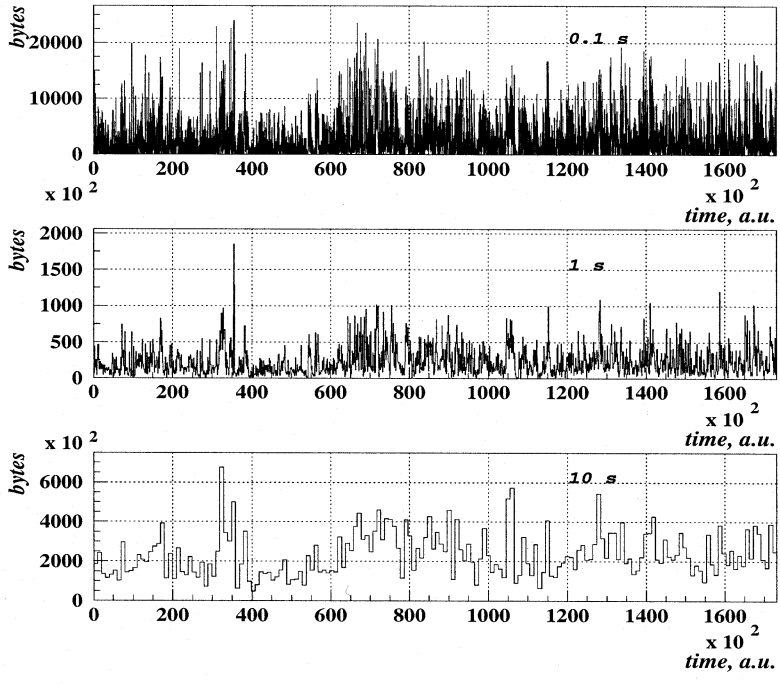

Figure 2: Traffic measurements aggregated with different bin sizes: $0.1\,s$, $1\,s$ and $10\,s$

2.1. Estimating the correlation length

In order to choose the independent components from the traffic data, we may compute the correlation length [18, 19], where the linear auto-correlation function

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

first time crosses the confidence tube corresponding to Gaussian white noise. Here x_i are the values of traffic measurements, N is the number of points in the analyzed time series and

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

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

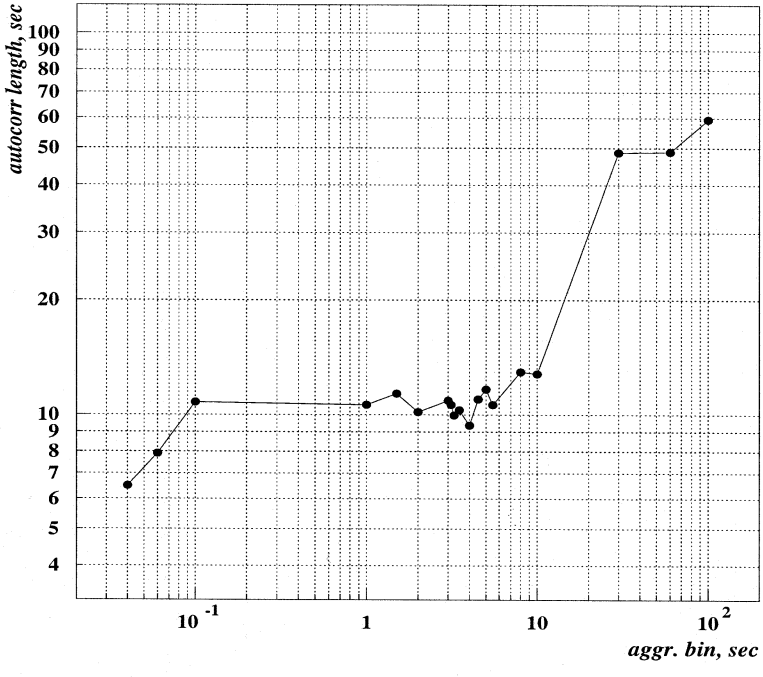

Figure 3: The dependence of the correlation length against the size of the aggregation bin

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

2.2. 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 [7]. To make 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 further increase of m. If this is the case, the time series can be considered as generated by a 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 [20, 21]. The correlation integral can be estimated by

$$C_2^m(r) = \frac{2}{N(N-1)} \sum_{i \neq j} \Theta(r - |\mathbf{y_i} - \mathbf{y_j}|), \tag{3}$$

with the distance between two points given by

$$|\mathbf{y}_i - \mathbf{y}_j| = \max \left\{ |x_i - x_j|, ..., |x_{i+(m-1)\tau} - x_{j+(m-1)\tau}| \right\}.$$

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

To estimate the embedding dimension d_E [20, 22], one computes $C_2^m(r)$ for r ranging from 0 to the largest possible value of $|\mathbf{y}_i - \mathbf{y}_j|$ 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 the $\log C_2^m(r)$ vs. $\log r$ gives the lowest estimate of the embedding dimension: see Fig. 4.

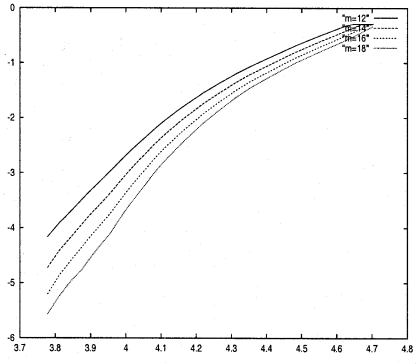

Figure 4: The dependences of $\log C_2^m(r)$ vs. $\log r$ for traffic measurements aggregated with 1 sec bin: $\tau = 10$ sec and m =12, 14, 16, 18

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 $m = 2 \div 18$ the slope β was found to satisfy

$$m \leq 2\beta + 1. \tag{4}$$

According to the Takens theorem [17], this may imply very high dimension of the studied time series.

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 applied the filtering based on a discrete wavelet transform: the details of wavelet filtering are discussed in Section 6.

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 $16 \div 18$ seems to be close to saturation.

2.3 Reconstruction of underlying dynamical system

In order to reconstruct the dynamical system corresponding to the traffic measurements, we used an artificial neural network (ANN) [7, 25, 26]. The major advantages of neural networks are that no prior information is required and the identification of the regular traffic component can be obtained automatically through the ANN training [27, 28, 29]. 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 number of neurons corresponding to the embedding dimension of the traffic series, two hidden layers with varying number of neurons and one output neuron. From the output neuron we get 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 *sec*. These data were preliminary *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: $\tau = 10\,sec$ and $d_E = 15 \div 20$.

Figure 5 presents part of the traffic data (traffic.dat) and the result of the ANN approximation (train.dat) after 1000 training epochs. We see that, despite the highly chaotic character of time series, the neural network approximates these data quit well.

Figure 6 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, which seems to be the log-normal.

It is known, that the ANN training on real data is in general adequate to the solution of the PCA problem [25, 31, 32, 33]. In this connection, the distribution of

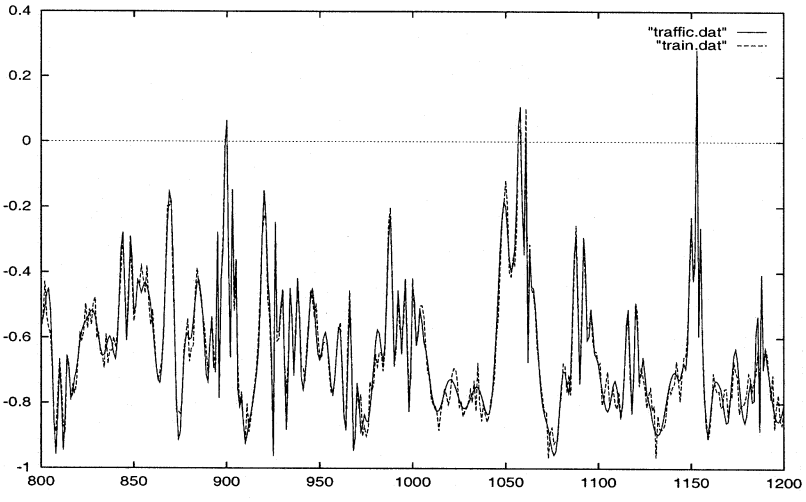

Figure 5: The result of the ANN approximation of the traffic series after 1000 training epochs

the ANN weights between the output node (neuron) of the ANN and the nodes of the second hidden layer is quite interesting: see Figure 7. We will see below, in Section 4, that this distribution of weights reproduces the character of the eigenvalues distribution obtained with the help of the PCA method.

3. Log-normal distribution of network traffic

Having available traffic data measured at high-frequency (each arriving packet has been recorded independently, see Section 1), we obtained the possibility to analyze the influence of the aggregation bin on the form of the packet size distribution. Figure 8 shows the packet size distribution for original traffic measurements, while figures 9, 10 and 11 present the distributions for measurements aggregated with bin sizes $10\,ms$, $100\,ms$ and $1\,s$, correspondingly.

One can clearly see that for the aggregation with small bin sizes the packet size distributions have rather chaotic and non-systematic character. However, when the aggregation bin size approaches $1\,s$ (see Fig. 11) the distribution assumes a stable form that does not change with further increase of the aggregation bin: see, for example, Fig. 12 corresponding to the aggregation with the bin size $10\,s$.

The distributions in figures 11 and 12 are well approximated by the log-normal function [3]

$$f(x) = \frac{A}{\sqrt{2\pi}\sigma}\frac{1}{x}\exp\left[-\frac{1}{2\sigma^2}(\ln x - \mu)^2\right], \tag{5}$$

where x is the variable, σ and μ are the parameters of log-normal distribution and A is the normalizing multiplier.

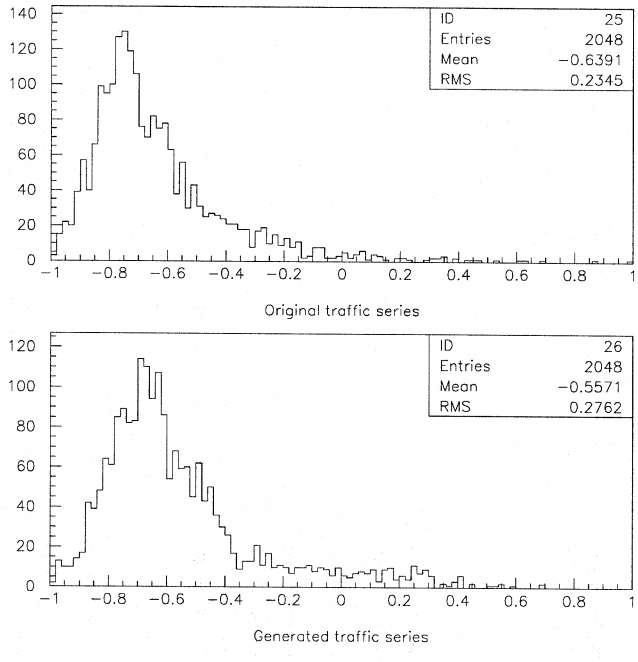

Figure 6: The distribution of sizes of the traffic packages (normalized to the interval [-1,1]) for : a) the original traffic measurements, and b) the generated by the trained ANN

The fitting procedure was realized with the help of the MINUIT package [35] in the frame of the well-known PAW (Physical Analysis Workstation, see details in [36]).

As we mentioned above, the fitting curves corresponding to the log-normal distribution approximate experimental distributions with a reliable accuracy on all regions of the analyzed distributions. However, they did not pass the χ^2-test [3].

The main reason is that the distributions presented in figures 11 and 12 are based on the whole set of data, which corresponds approximately to 20 hours of measurements. But the traffic series, as well as corresponding statistical distributions, behave differently depending, if the measurements were done during working hours or not.

In this connection, we tested the correspondence of experimental distributions to the null-hyphothesis (5) applying the χ^2 goodness-of-fit criterion using only the daily traffic. The results of this analysis are presented in Table 1.

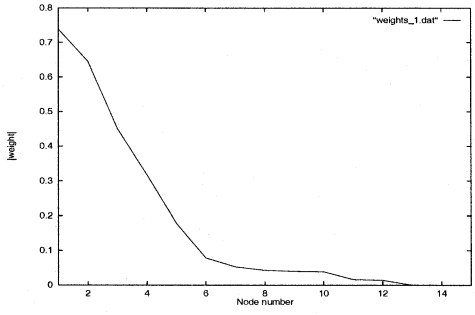

Figure 7: The distribution of the absolute values of weights between the output node (neuron) of the ANN and the nodes of the second hidden layer of the ANN trained on traffic measurements

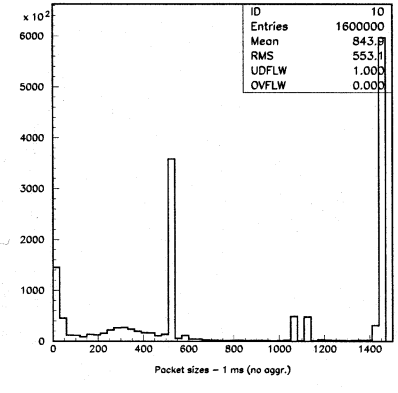

Figure 8: Packet size distribution for traffic measurements

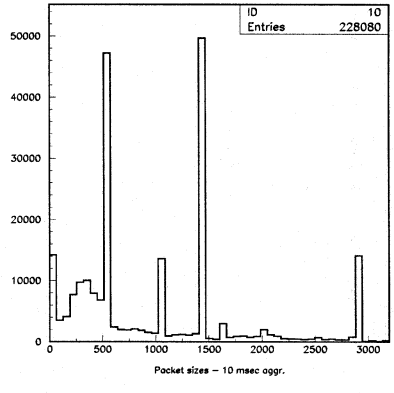

Figure 9: Packet size distribution for traffic measurements aggregated with bin size $10\,ms$

Here α is the probability (in %) that the observed chi-square will exceed the value χ^2 by chance *even* for a correct model: see, for instance, [34, 37]. These results show that the hyphothesis (5) can be accepted with a high probability: see also Fig. 13. At the same time it must be noted (see figures 11 and 12) that the influence of the inactive period of LAN does not change significantly the fundamental form of the statistical distribution.

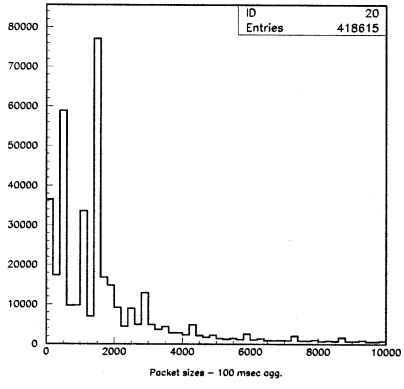

Figure 10: Packet size distribution for traffic measurements aggregated with bin size 100 ms

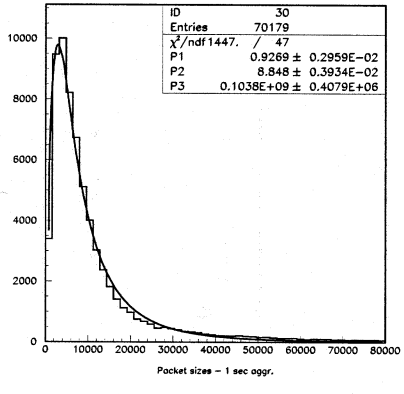

Figure 11: Packet size distribution for traffic measurements aggregated with bin size 1 s

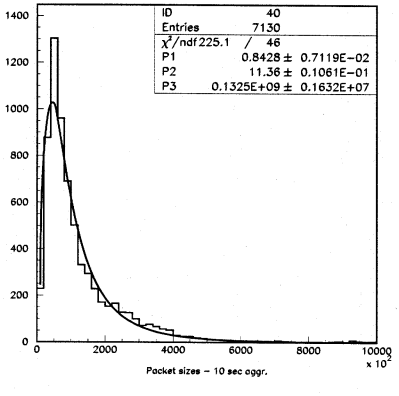

Figure 12: Packet size distribution for traffic measurements aggregated with bin size 10 s: fitting curve corresponds to the function (5)

We conclude, therefore, that

- the aggregation of traffic measurements forms (starting from some threshold value of the aggregation window) a statistical distribution, which does not change its form with further increase of the aggregation window;
- this distribution is approximated with high accuracy by the log-normal distribution.

Table 1: Results of fitting of daily part of packet size distributions aggregated with different bin sizes by the function (5)

Bin, sec	ν	χ^2	α,%
1	47	49.84	32.30
2	47	44.76	52.51
3	47	41.53	65.98

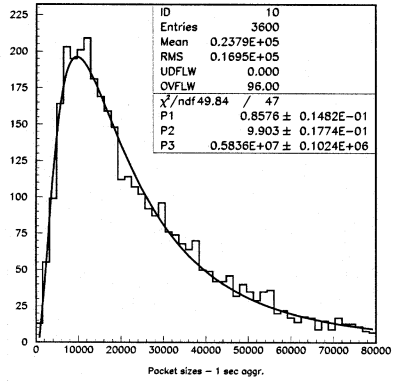

Figure 13: Packet size distribution for daily traffic measurements aggregated with bin size $1\,s$: fitting curve corresponds to the function (5)

4. Principal component analysis of network traffic

The "Caterpillar"-SSA approach [11, 12] can be used for analysis of time series corresponding to any arbitrary function $f(t), t > 0$ determined in equidistant points:

$$x_i = f[t_i] = f[(i-1)\Delta t], \quad i = 1, 2, \ldots, K, \tag{6}$$

where Δt is the sampling interval (in our case $\Delta t = 1$), whose reciprocal is the sampling rate.

The basic "Caterpillar"-SSA scheme includes four main steps:

- transformation of one-dimensional series into multidimensional form,
- singular value decomposition of multidimensional series,
- principal components analysis and selection of feature components,

- reconstruction of one-dimensional series on the basis of selected components.

The transformation of one-dimensional series (6) into multidimensional one is realized by representing 6 in matrix form:

$$X = (x_{ij})_{i,j=1}^{k,L} = \begin{pmatrix} x_1 & x_2 & x_3 & \dots & x_L \\ x_2 & x_3 & x_4 & \dots & x_{L+1} \\ x_3 & x_4 & x_5 & \dots & x_{L+2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_k & x_{k+1} & x_{k+2} & \dots & x_K \end{pmatrix}, \tag{2}$$

where $L < M$ is called the caterpillar or window length and $k = K - L + 1$.

Then the eigenvalues λ_i, $i = 1, 2, \dots, L$ and eigenvectors $\vec{V}_i$, $i = 1, 2, \dots, L$ of the covariance matrix $C = \frac{1}{k} XX^T$ are determined. The matrix of eigenvectors V is used for transition to the principal components

$$Y = V^T X = (Y_1, Y_2, \dots, Y_L), \tag{3}$$

where Y_i $(i = 1, 2, \dots, L)$ are rows of k elements.

The equality

$$\sum_{i=1}^{L} \frac{\lambda_i}{L} = \sum_{i=1}^{L} \alpha_i = 1$$

permits to estimate the contribution α_i (in decreasing order) of the i-th principal component into the analyzed series.

The "Caterpillar" length (or window) C_L has been chosen based on the analysis of the autocorrelation function for traffic measurements [2]. In this study we used different values of C_L, starting from the minimal value $C_L = 12$ up to $C_L = 20$.

Figure 14 shows the daily part of traffic measurements aggregated with the bin size $1\,s$, which has been used in this study. The number of points in this series $K = 2048$, that corresponds approximately to 34 minutes of traffic measurements.

One of the main results of the application of the "Caterpillar"-SSA technique to the analyzed series is presented in Fig. 15. It shows the contribution of eigenvalues in percentages for $C_L = 12$ and 20. This information permits to estimate the number of principal components, which effectively contribute into the analyzed series.

Taking into account [3], it is reasonable to assume that the packet size distributions, corresponding to leading components, may be described by the log-normal distribution.

In Fig. 16 we present the results of fitting of the packet size distributions, corresponding to different number N of leading components (the results presented here are for $C_L = 20$), by function (5). Here χ^2 is the calculated value of χ^2, corresponding to the testing distribution and ν is the number of degrees of freedom.

This dependence demonstrates that for $N = 3$ there is quite a good level of correspondence ($\alpha = 22\%$) of the distribution to the null-hypothesis (see also Fig. 17).

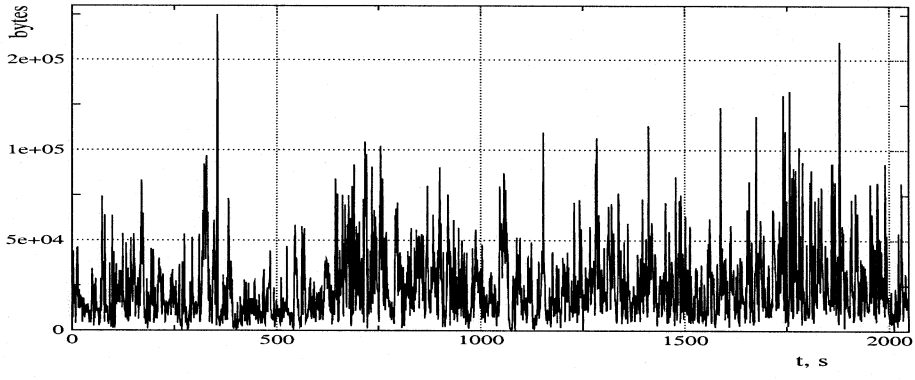

Figure 14: Traffic measurements aggregated with the bin size $1\,s$

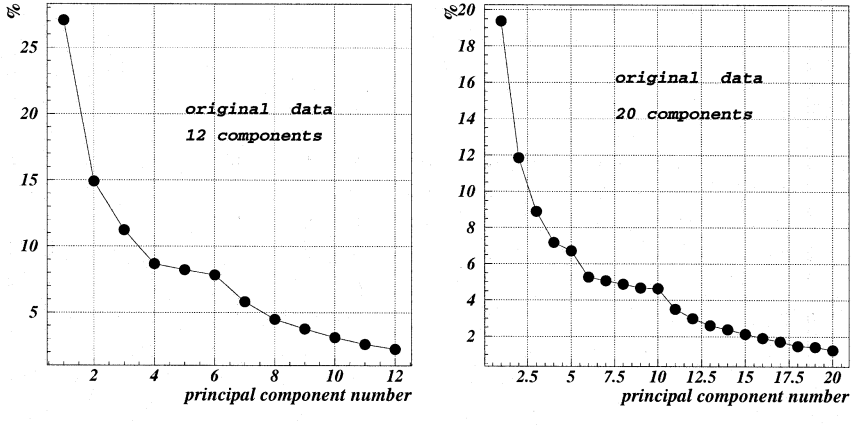

Figure 15: Contributions of eigenvalues in percentages for the original traffic data. The results are presented for two cases of the caterpillar length: $C_L = 12$ (left) and 20 (right)

This result is of great interest, because only 3 first components (of 20) already form the fundamental part of the information traffic. Their summary contribution into the general dispersion is around 40 % (see Fig. 15 for $C_L = 20$).

The value of χ^2/ν reaches its record minimal value 0.732 for $N = 8$. The corresponding statistical distribution is presented in Fig. 17. It demonstrates both a very good level of correspondence of the reconstructed distribution to the null-

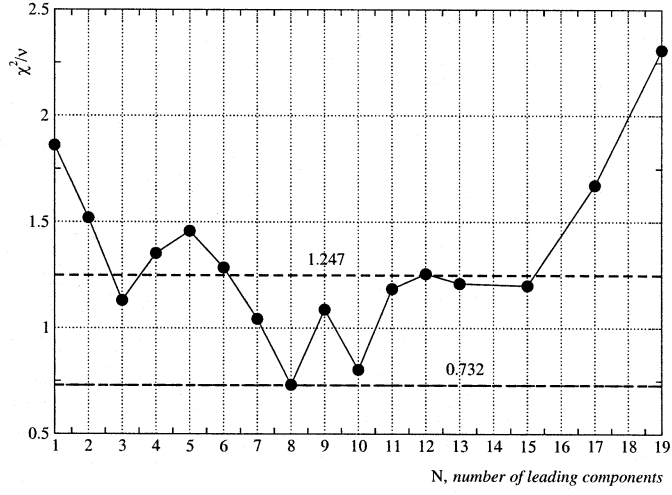

Figure 16: The dependence of χ^2/ν versus the number of leading components

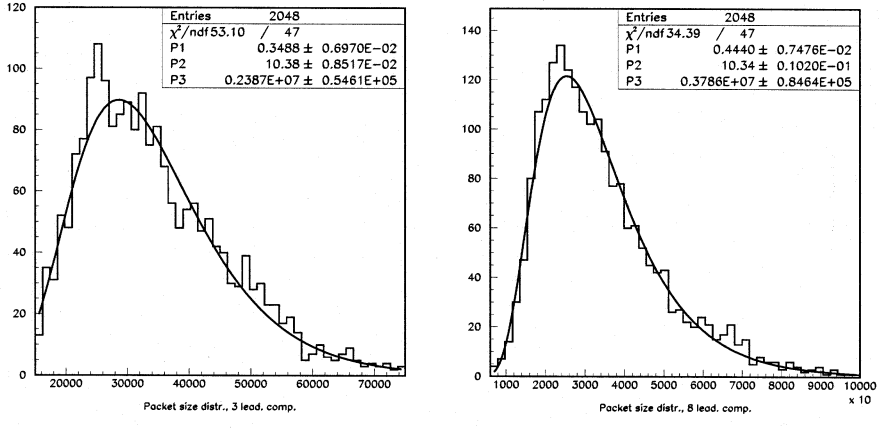

Figure 17: Fitting distributions corresponding to 3 (left figure) and 8 (right figure) leading components by function (5)

hypothesis ($\alpha = 89.5\,\%$) and a reliable accuracy of approximation on all regions of the analyzed distribution. The summary contribution of 8 leading components into the general dispersion is around 66 %.

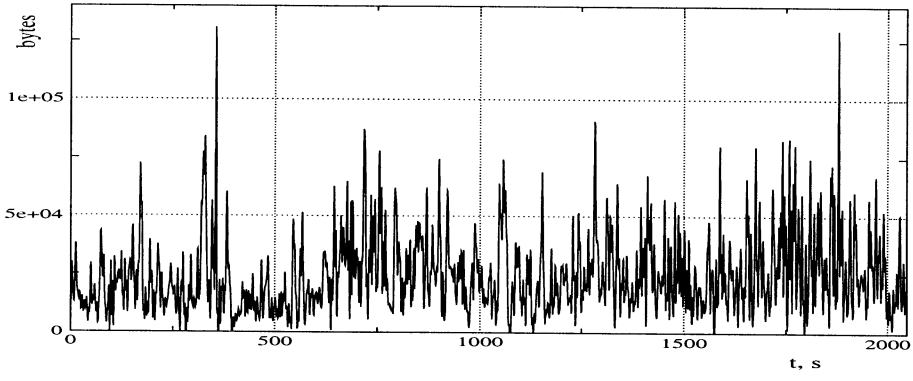

Figure 18: Traffic series measurements reconstructed by the caterpillar method (for $C_L = 20$) on the basis of 8 leading components

Figure 18 shows the series reconstructed by the Caterpillar method (for $C_L = 20$) on the basis of eight leading components. One can clearly see that it reproduces characteristic features of the original series presented in Fig. 14.

In the region of large N there is a growth of χ^2 especially noticeable at $N \geq 15$: see Fig. 16. Such tendency is caused by the influence of the residual components related to small irregular variations, which do not fit in the basic model of network traffic (5) and can be interpreted as a stochastic noise (see Section 5).

Figure 19 shows the series reconstructed on the basis of the smallest residual component, namely, the component 20. One can clearly see that this series is of significantly different character as compared to the original traffic measurements. It looks like a nonstationary dynamical process symmetric against zero mean value.

Figure 20 shows the statistical distribution corresponding to the series presented in Fig 19. It quite well follows the Gaussian distribution that is confirmed by the χ^2-test (see Fig. 20). The autocorrelation function of the corresponding series shows that it behaves like noise.

However, when increasing the number of residual components, their summary distribution starts to gradually lose the symmetric form together with growth of correlations between the series terms.

In order to estimate the amount of residual components, which can be eliminated from the original time series without the influence on its fundamental part, we divide all principal components into two parts:

1. first part corresponding to the leading components and responsible for the log-normal form of the packet size distribution,

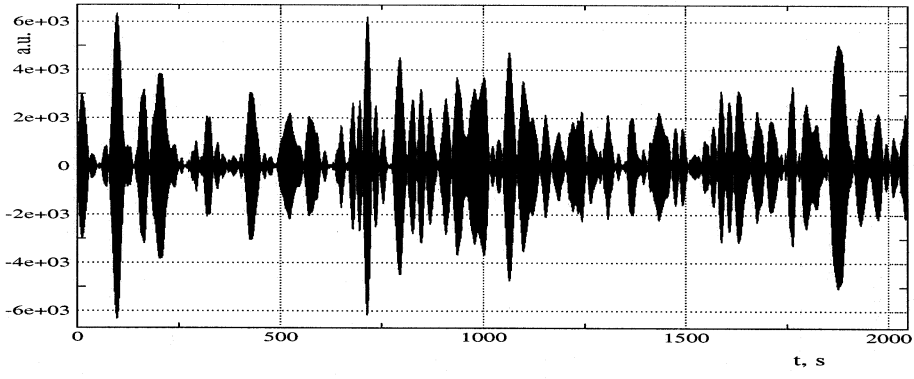

Figure 19: Traffic series reconstructed by the caterpillar method ($C_L = 20$) on the basis of the smallest component

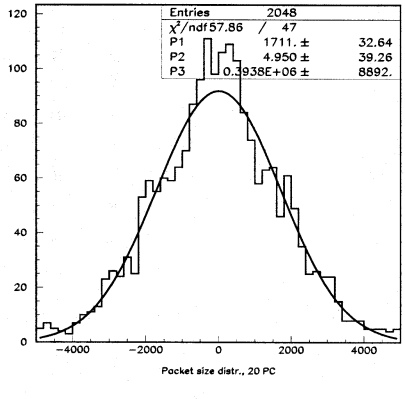

Figure 20: Statistical distribution of the time series presented in Fig. 19; the fitting curve corresponds to the Gaussian distribution

2. second part related to residual components, which is described by a symmetric statistical distribution and behaves like a stochastic noise.

As the criterion for selection of the second part we used the "moment" of the symmetry violation for the series corresponding to the residual components. The well-known sign test has been used for testing the symmetry against zero of residual distributions. The sign test has the following form:

$$\mu = \sum_{i=1}^{n} \Theta(x_i), \tag{7}$$

where $x_1, \ldots, x_n$ are observables, n is the sample size, and Θ is the Heaviside function:

$$\Theta(x) = \begin{cases} 1, & x > 0 \\ 0, & x \leq 0. \end{cases}$$

When the null-hypothesis is true, the μ distribution is approximated (in case of large n) by

$$P\{\mu \leq m \mid n, p\} \approx \Phi\left(\frac{m - np + 0.5}{\sqrt{np(1-p)}}\right),$$

where Φ is the distribution function of the normal distribution, $p = 0.5$ and $n = 2048$ (in our case).

Figure 21 shows the dependence of μ value versus the number of the residual components (for caterpillar lengths 12 and 20). It is clearly seen that the μ value exceeds the reliable confidential level, when the number of residual components is greater than 6 for $C_L = 12$ and 11 for $C_L = 20$.

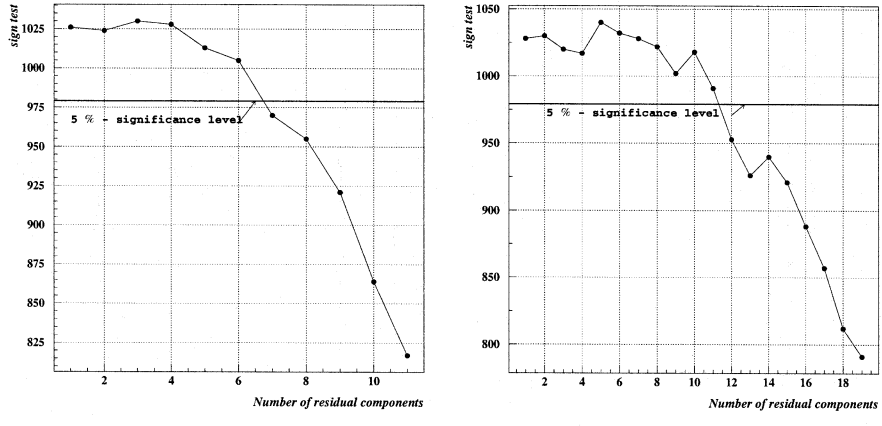

Figure 21: The values of sign test μ versus the number of the residual components for the caterpillar length $C_L = 12$ (left figure) and $C_L = 20$ (right figure)

In order to confirm the results obtained by the sign test, we applied more powerful criterion based on the ω_n^2 statistics [40]. This criterion tests the symmetry against $x = 0$ the distribution function $F(x)$ of the observables $x_1, \ldots, x_n$, i.e. the null-hypothesis H_0: $F(x) = 1 - F(x)$. The corresponding ω_n^2 statistics has the following form:

$$\omega_n^2 = n \int_{-\infty}^{\infty} [F_n(x) + F_n(-x) - 1]^2 \, dF_n(x), \tag{8}$$

where $F_n(x)$ is the empirical distribution function. It is more convenient to calculate the values of statistics (8) using the following algebraic formula

$$\omega_n^2 = \sum_{j=1}^{n} \left[F_n(-\tilde{x_j} - \frac{n-j+1}{n} \right]^2,$$

where $\tilde{x_1} \leq \ldots \leq \tilde{x_n}$ is the variational series constructed on the basis of observables.

Figure 22 shows the dependences of the ω_n^2 value versus the number of the residual components for two cases of the caterpillar length: $C_L = 12$ and 20.

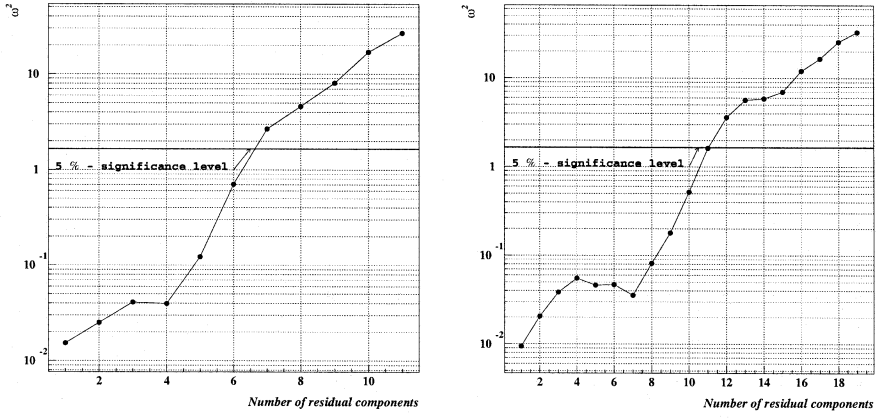

Figure 22: The dependences of the ω_n^2 values versus the number of the residual components for two cases of the caterpillar length: $C_L = 12$ (left figure) and $C_L = 20$ (right figure)

One can see from Fig. 22 that the number of residual components $l = 6$ for $C_L = 12$ and $l = 10$ for $C_L = 20$ corresponds to the 5% - significance level for the ω^2-criterion. This coincides with the result obtained for the sign test: see Fig. 21.

The dependences presented in Fig. 22 have distinct characteristic features at $l = 4$ for $C_L = 12$, and $l = 7$ for $C_L = 20$ (one can see that the number of such components approximately equals to one third of the caterpillar length), after which, when l is increasing, there is a quick rise of ω_n^2. This means that the residual series looses its symmetric character, because in the second part are involved the components responsible for the fundamental property of the system – the log-normality.

5. Spectral analysis of traffic measurements

A sampled data set (6) contains *complete* information about all spectral components in a signal $x(t)$ up to the Nyquist critical frequency

$$f_c = \frac{1}{2\Delta t}, \tag{9}$$

and scrambled or *aliased* information about any signal components at frequencies larger than f_c (see, for example, [37]).

In order to estimate the presence or absence of periodic components and to evaluate the viability of stochastic noise in the traffic series, we apply here the Lomb spectral method: see, [37, 41] and references therein.

The Lomb *normalized periodogram* (spectral power as a function of angular frequency $\omega \equiv 2\pi f > 0$) of one-dimensional time series (6) is defined by

$$P_K(\omega) = \frac{1}{2\pi^2}\left\{\frac{\left[\sum\limits_{i=1}^{K}(x_i-\bar{x})\cos\omega(t_i-\tau)\right]^2}{\sum\limits_{i=1}^{K}\cos^2\omega(t_i-\tau)} + \frac{\left[\sum\limits_{i=1}^{K}(x_i-\bar{x})\sin\omega(t_i-\tau)\right]^2}{\sum\limits_{i=1}^{K}\sin^2\omega(t_i-\tau)}\right\}, \tag{10}$$

where

$$\bar{x} = \frac{1}{K}\sum_{i=1}^{K} x_i, \qquad \sigma^2 = \frac{1}{K-1}\sum_{i=1}^{K}(x_i-\bar{x})^2$$

and τ is defined by the relation

$$\tan(2\omega\tau) = \sum_{i=1}^{K}\sin 2\omega t_i \Big/ \sum_{i=1}^{K}\cos 2\omega t_i.$$

In order to estimate the significance of a peak in the spectrum $P_K(\omega)$, we have to test the null-hypothesis that the data values are independent of Gaussian random values.

Scargle has shown [42] that for the normalized Lomb peridogram (10) at any ω and when the null-hypothesis is valid, $P_K(\omega)$ has an exponential probability distribution with unit mean. This means that the probability that $P_K(\omega)$ will be between some positive z and $z+dz$ is $\exp(-z)dz$. If we scan some M *independent* frequencies, the probability that none give values larger than z is $(1-e^{-z})^M$. Thus,

$$p(>z) = 1-(1-e^{-z})^M \tag{11}$$

determines the false-alarm probability of the null-hypothesis, and it shows the *significance level* α of any peak in the $P_K(\omega)$ spectrum.

For estimation of the significance level α, we need to know M in the region where α assumes small values, $\alpha \ll 1$, and Eq. (11) can be represented as

$$p(>z) \approx Me^{-z}. \tag{12}$$

The relation (12) shows that the significance level changes linearly with M. In practice, an error of even $\pm 50\%$ in the evaluated significance is often tolerable, which means that our estimation of M need not to be very accurate.

Horne and Baliunas [43] have found that M is very nearly equal to K when the data points are equally spaced, and when the sampled frequencies "fill" the frequency range from 0 up to f_c.

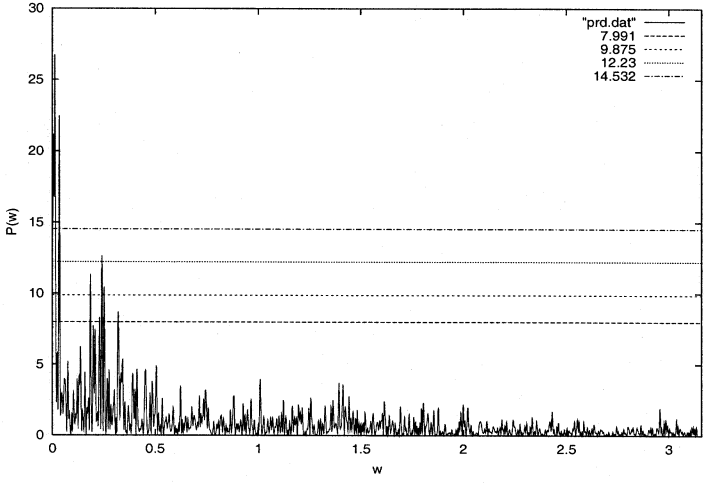

Figure 23: The dependence of $P_K(\omega)$ against the angular frequency $\omega = 2\pi f$ for traffic measurements presented in Fig. 14: $0 \leq \omega < 2\pi f_c$

Figure 23 shows the result of application of the Lomb method to the time series shown in Fig. 14: we used the code **period** from the *Numerical Recipes* library [37]. The figure plots $P_K(\omega)$ against the angular frequency $\omega = 2\pi f$ for the frequency interval starting from 0 up to f_c. The horizontal dashed and dotted lines correspond (from bottom to top) to the significance levels 0.5, 0.1, 0.01, 0.001, respectively.

One can see (Fig. 24) three highly significant peaks at low frequencies: 0.06, 0.012 and 0.034. There are three other peaks at frequencies 0.186, 0.241 and 0.252, which also exceed the 50 % significance level.

For frequencies higher $\omega > 0.35$ together with the frequency increase, the amplitude of peaks is very quickly decreasing (Fig. 23) and does not exceed the value 5. This amplitude corresponds to the significance level $\alpha \approx 1$. This may mean that the traffic components related to this high frequency part can be interpreted as a stochastic Gaussian noise.

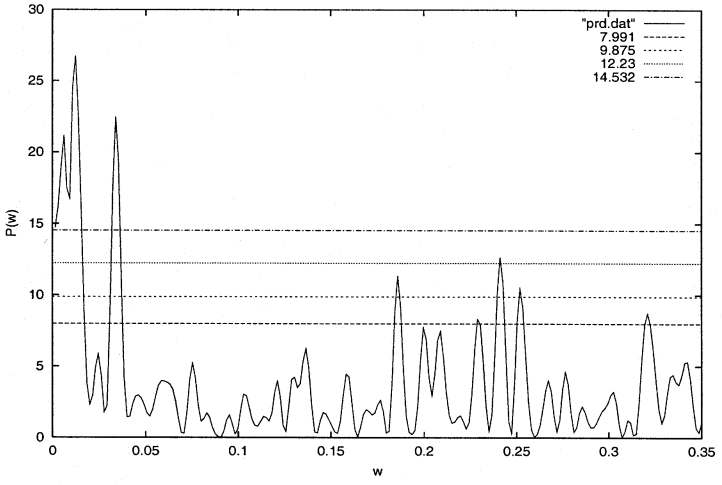

Figure 24: The dependence of $P_K(\omega)$ against the angular frequency ω for traffic measurements presented in Fig. 14: $0 \leq \omega < 0.35$

6. Wavelet filtering of traffic measurements

The wavelet analysis is the most suitable approach to handle irregular time series, such as traffic measurements, because it permits to focus on localized signal structures along with a zooming procedure that progressively reduces the scale parameter: see, for instance, [53, 54].

The discrete wavelet transform (DWT) of the function $f(t) \in L_2(R)$ given in form of one-dimensional time series (6) can be represented by the following expansion

$$f(t) = \sum_{j,k \in Z} d_{jk} \psi(2^j t - k). \tag{13}$$

Here the set of basis functions (wavelets) $\{\psi_{jk}(t) = \psi(2^j t - k), \quad j, k \in Z\}$ is obtained from a single "mother" wavelet function $\psi(t) \in L_2(R)$ applying the binary dilation 2^j and the dyadic translation $k/2^j$.

Following the multiresolution wavelet analysis, Eq. (13) can be rewritten in a more convenient form

$$f(t) = \sum_k s_k^J \phi(2^J t - k) + \sum_{j \geq J} \sum_{k \in Z} d_k^j \psi(2^j t - k), \tag{14}$$

where $\phi(t)$ is the scaling function corresponding to the chosen wavelet function $\psi(t)$ (see, for example, [23]). In (14) the first term describes a smooth part of series (14) restricted by level J, and the second term is related to details, or a high-frequency

part of the analyzed series. We use here the discrete Daubechies wavelets [23, 24], because they provide high quality representation of both high- and low-frequency components of the analyzed signal [37].

The coefficients s_k^j and d_k^j are usually determined with the help of the pyramidal scheme [44] of the fast wavelet transform (see, for instance, [37]) applying the following equations:

$$s_k^{j+1} = \sum_m h_m s_{2k+m}^j, \qquad d_k^{j+1} = \sum_m g_m s_{2k+m}^j, \tag{15}$$

where h_m and g_m are the coefficients of low pass and high pass filters, respectively.

The wavelet filtering implies rejection or modification of part of expansion coefficients with absolute values less of a preassigned threshold value λ. There exist several different wavelet filtering algorithms specified as *hard*, *soft*, *quantile* and *universal thresholding* (see, for example, [38, 39]). However, the most widespread is the hard thresholding algorithm (see, for example, [37]). In this scheme all coefficients with absolute values less than λ have to be rejected (set to zero).

In all methods mentioned above the filtering procedure affects all coefficients, without taking into account their belonging to some resolution level J. Therefore, such a procedure may eliminate both the coefficients $\{d_k^j\}$ which correspond to the high-frequency part of (14) and the coefficients $\{s_k^J\}$ related to the low-frequency part.

In this connection, it is impossible to apply the existing algorithms to our case, because the filtering will affect not only a high-frequency, *noisy* part, but also a regular part, which should not be touched.

To overcome this problem, we modified the *hard thresholding* scheme in such a way that the groups of coefficients corresponding to different levels of wavelet decomposition are filtered in a successive order. The modified algorithm performs as follows. Suppose K is the number of elements in the analyzed series and $M < \frac{K}{2}$. Then, M smallest of $\frac{K}{2}$ "detailed" coefficients of series (14) have to be rejected. If $\frac{K}{2} < M < \frac{3K}{4}$, then we eliminate all $\frac{K}{2}$ "detailed" coefficients together with $M - \frac{K}{2}$ smallest coefficients corresponding to a lower level of accuracy (the whole number of such coefficients is $\frac{K}{4}$), etc.

Compared to the traditional filtering procedure, the modified scheme provides more effective elimination of the high-frequency component from such highly irregular time series as traffic measurements.

After the DWT, the selected M coefficients are set to zero, and then, using the inverse wavelet transform, the regular part of the traffic series is reconstructed. The difference between the original time series and the filtered signal, is considered as a noisy component.

The symmetry test based on the ω_n^2 statistic [40] has been used for estimation of a possible number of wavelet coefficients related to the noisy part. The result of the ω_n^2 test has been independently checked by analyzing the autocorrelation function behavior for the rejected part.

Figure 25 shows the dependence of ω_n^2 values versus the number of rejected wavelet coefficients. This dependence clearly shows the minimal value of ω_n^2 at $M = 768$. One can also see in Fig. 25 that a possible maximal number of coefficients that can be eliminated without exceeding the 5%-significance level is $M = 1408$. This corresponds to approximately 70 % of 2048 coefficients.

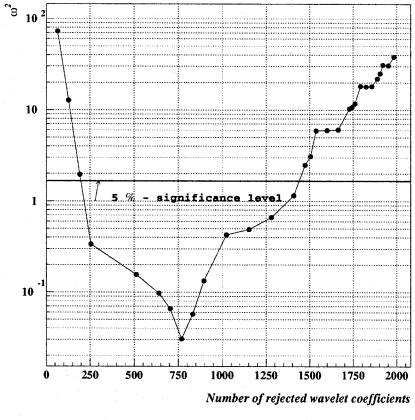

Figure 25: The dependence of ω_n^2 values versus the number of rejected wavelet coefficients

The autocorrelation function can be also used as a criterion for evaluation of the noisy part. The time series corresponding to the noisy part must be uncorrelated. Figure 26 (left plot) presents the dependence of the auto-correlation function for the noisy part corresponding to different number of rejected coefficients M. This figure shows that up to $M = 1408$ the rejected part can be considered as noisy.

Based on estimations of these two criteria, we came to the conclusion that it is reasonable to assume $M = 1408$. Figure 27 presents the original traffic series, the filtered signal and the noisy part that may be rejected.

In order to monitor the influence of the rejected part on the main part of traffic series (from the nonlinear analysis point of view), we also controlled the behavior of the autocorrelation function of the smooth part of series (14) for different number of rejected coefficients: see Fig. 26 (right plot). One can clearly see that the rejection of the smallest coefficients up to $M = 1408$ did not influence seriously the form of the autocorrelation function.

It is also interesting to check the influence of the filtering procedure on spectral characteristics of the analyzed series. Figure 28 shows the dependence of $P_K(\omega)$ against the angular frequency ω for filtered signal (continuous curve) and original (dashed curve) traffic measurements.

This plot shows that the filtering procedure increased the power of all frequencies contributing into low frequency region. At the same time, higher frequencies starting approximately at $\omega = 1.1$ have been significantly suppressed.

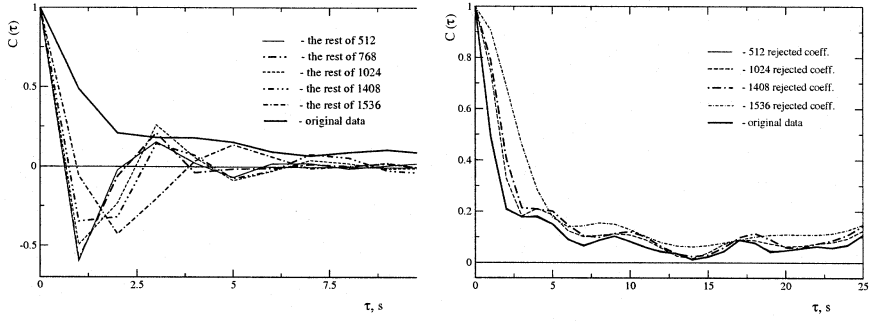

Figure 26: Autocorrelation functions $C(\tau)$ of noisy (left plot) and smooth (right plot) parts corresponding to different number of rejected coefficients

7. Analysis of statistical characteristics of filtered series

In Fig. 29 we present the contribution of individual components into the analyzed series for traffic data after filtering out the high-frequency part corresponding to $M = 1408$ smallest coefficients. One can clearly see that the contribution of the residual components noticeably decreased compared to the original traffic data (Fig. 15). At the same time the contribution of the leading components significantly increased.

This result may play a very important role for decreasing the dimension of the system describing the information traffic, but this may be the case, if the filtering procedure does not seriously disturb the statistical and dynamical characteristics of traffic series.

Taking into account the results of Sections 5 and 6, it is important to see how the filtering procedure influences the statistical characteristics of traffic series, namely,

1. if it disturbs seriously the packet size distributions, corresponding to leading components, and

2. how this procedure influences the residual components, whose contribution have been significantly suppressed by the filtering procedure.

In order to check the influence of the wavelet filtering on the packet size distributions of leading components, we applied the same procedure as in Section 4, i.e. we tested the correspondence of these distributions to the log-normal form.

Figure 30 shows the results of fitting of the packet size distributions (for the filtered traffic series), corresponding to the sum of a different number N of leading components (the results presented here are for $C_L = 20$), by function (5).

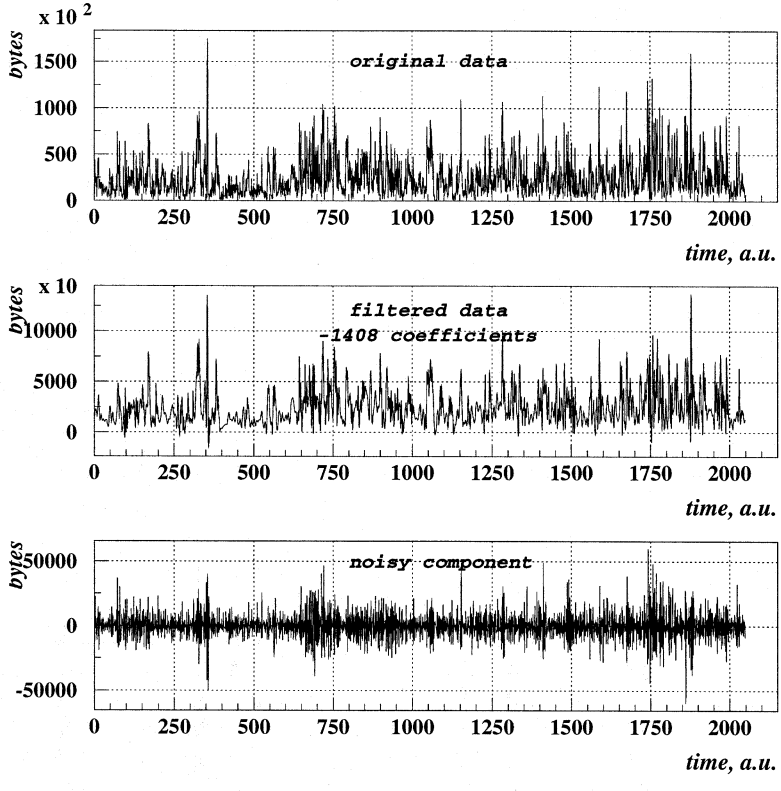

Figure 27: Traffic measurements: 1) original traffic series, 2) filtered signal, 3) noisy part

Here the top and bottom lines correspond to significance levels $\alpha = 10\%$ ($\chi^2/\nu = 1.247$) and $\alpha = 42.9\%$ ($\chi^2/\nu = 1.023$) for $\nu = 47$, correspondingly.

This dependence confirms the result of Section 4 (Fig. 16) concerning the number of leading components that form the main part of information traffic. One can clearly see that three leading components form the distribution that follows the null-hypothesis (5) with a quite high correspondence level ($\alpha = 39.2\%$): see also Fig. 31.

The dependence of χ^2/ν versus the number N of leading components in Fig. 30 shows that

1. the maximal significance level of the χ^2-test corresponds to the sum of 3-4 first leading components;

2. this dependence is compactly distributed around the corridor corresponding to the admissible region for the χ^2-test.

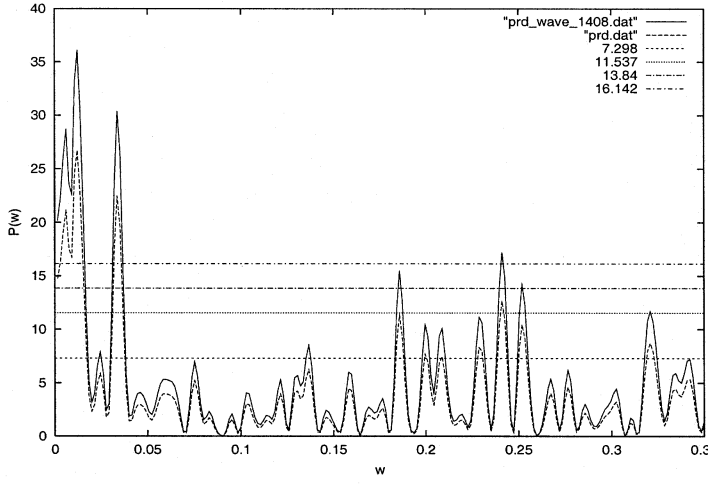

Figure 28: The dependence of $P_K(\omega)$ against the angular frequency $\omega = 2\pi f$ for filtered signal (continuous curve) and for original traffic measurements (dashed curve)

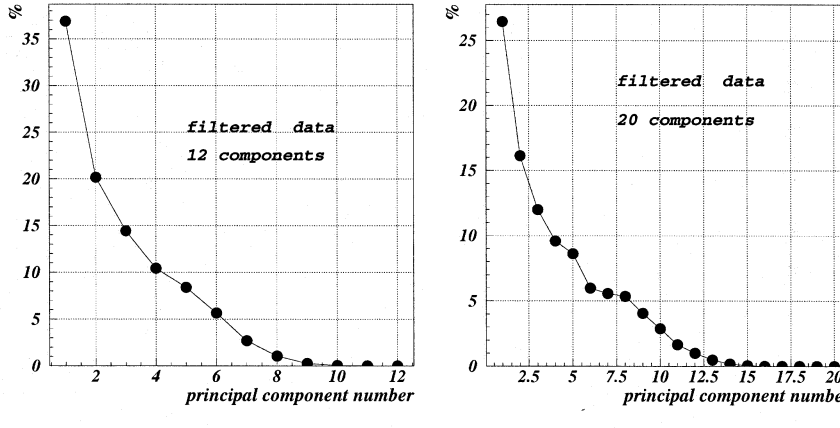

Figure 29: Contributions of eigenvalues in percentages for the traffic data after filtering out the high-frequency part. The results are presented for two cases of the caterpillar length: $C_L = 12$ (left) and 20 (right)

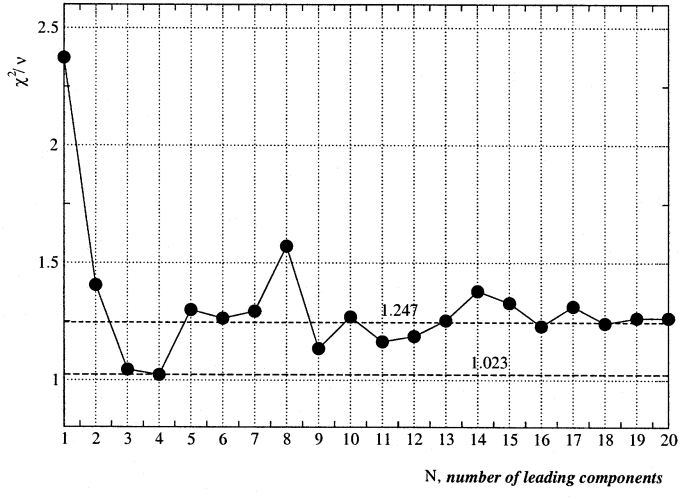

Figure 30: The dependence of χ^2/ν versus the number of leading components for filtered series

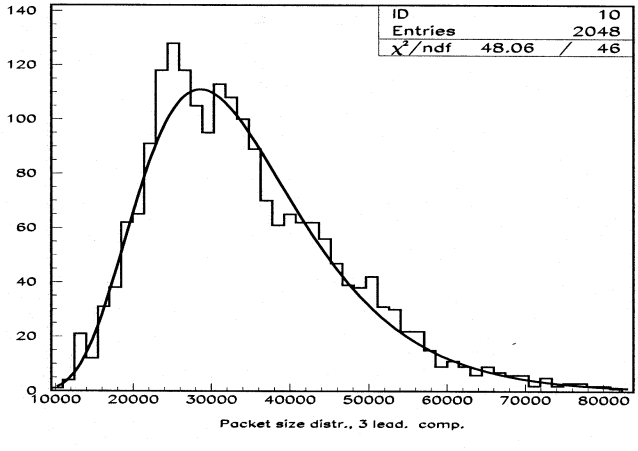

Figure 31: Fitting the distribution corresponding to three leading components by the log-normal function (5)

Figure 32 shows the seria reconstructed on the basis of first, second and third leading component, correspondingly, after the subtraction of the caterpillar average value.

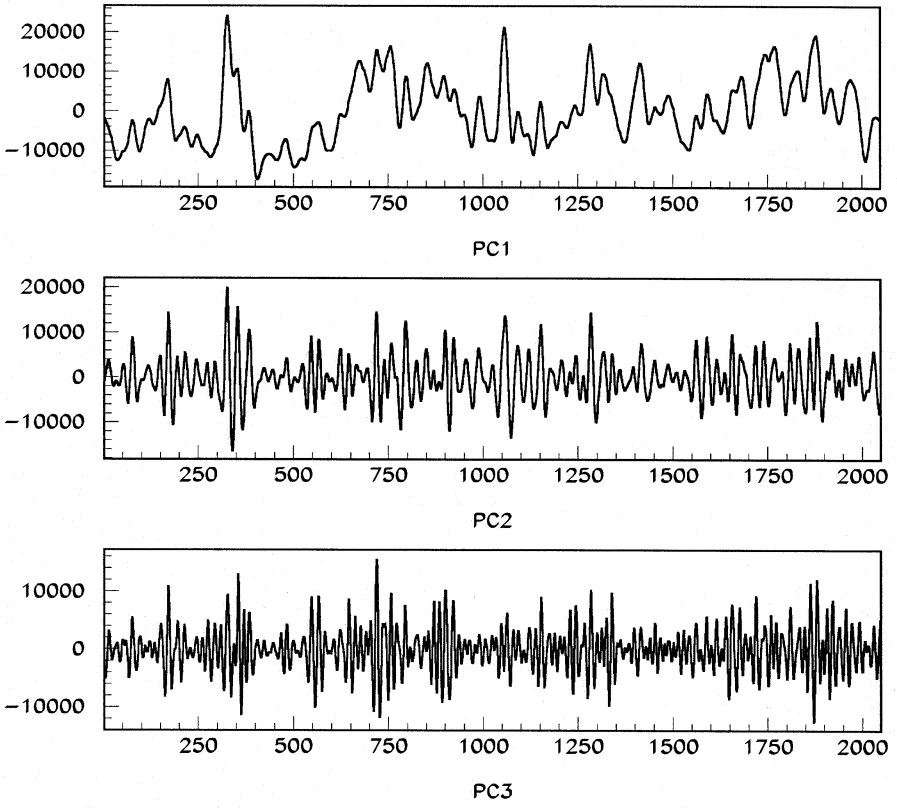

Figure 32: Time seria corresponding to three leading components (after the subtraction the caterpillar average value): the trend component and two first periodic components

These seria are very much similar to the seria corresponding to the original traffic data (see Fig. 8 in [4]). However, the filtered seria are visually more smooth if compared to the original data. Their summary contribution into the analyzed time series is noticeably higher ($\sim 54\,\%$) if compared to the original data ($\sim 40\,\%$): see Figs. 15 and 29 for $C_L = 20$.

Figure 33 shows the series reconstructed on the basis of the smallest residual component, namely, the component 20. It looks very similar to the same component of the original traffic measurements (Fig. 19). The statistical distribution

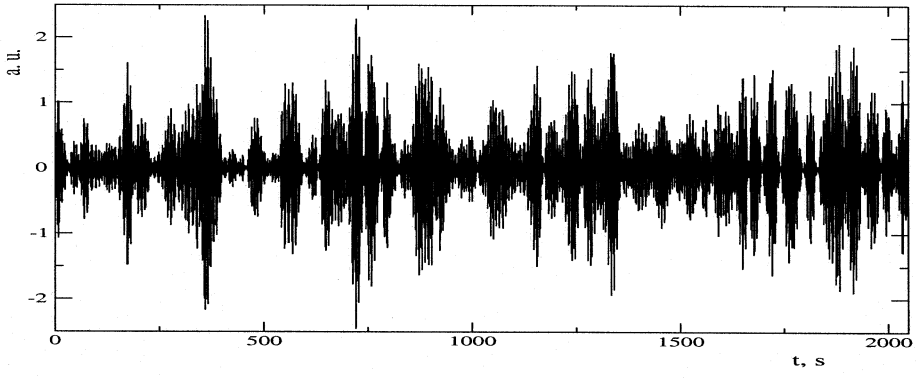

Figure 33: Traffic series reconstructed by the caterpillar method ($C_L = 20$) on the basis of the smallest component

corresponding to this series quite well follows the Gaussian distribution (the same as Fig. 20).

At the same time, the amplitude dispersion and the standard deviation of this series are significantly less if compared to the original data: see Figs. 19 and 20.

8. Selection of feature components

In order to estimate the number of residual components that can be eliminated from the filtered time series without influence on its main part, we applied here the statistical criterion of symmetry based on the ω_n^2 statistic: see Section 4.

Figure 34 shows the dependences of the ω_n^2 value versus the number of residual components for original (left figure) and filtered (right figure) traffic seria for the caterpillar length $C_L = 20$. The horizontal line corresponds to the significance level 0.05.

It is clearly seen that the ω_n^2 value exceeds the reliable confidential level (corresponding to the 5 %-significance level), when the number n of residual components exceeds 10 for original traffic measurements and 17 for the filtered series. This result demonstrates that after the wavelet filtering 17 smallest components can be considered as noisy and can be eliminated from the whole set of principal components. This confirms the result of Section 4 obtained by the χ^2-test: see Fig. 30.

Figure 35 shows the dependence of $P_K(\omega)$ against the angular frequency $\omega = 2\pi f$ for three leading components (continuous curve) and for all components of the filtered signal (dashed curve). This dependence clearly demonstrates that the low frequency region of traffic series is formed by three leading components. This plot also shows that the powers of all frequencies contributing into this frequency domain have been increased if compared to the powers of the series corresponding

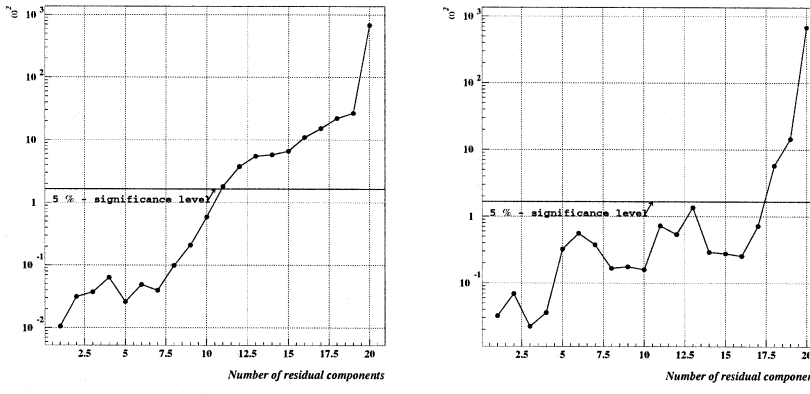

Figure 34: The dependences of the ω_n^2 values versus the number of the residual components for the original (left figure) and filtered (right figure) traffic series and for the caterpillar length $C_L = 20$

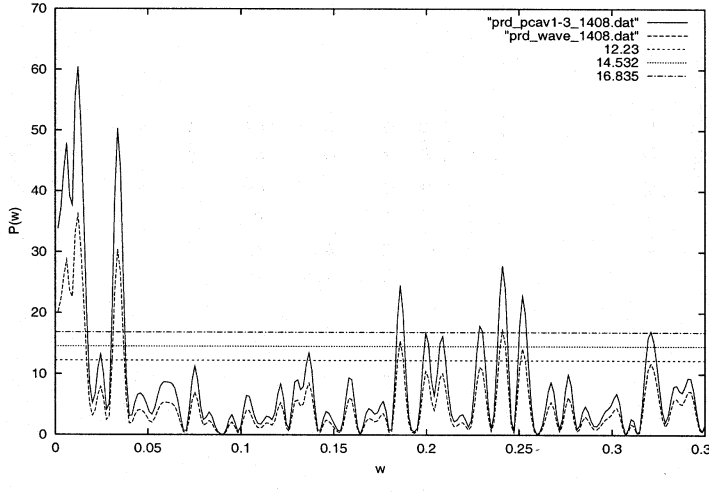

Figure 35: The dependence of $P_K(\omega)$ against the angular frequency $\omega = 2\pi f$ for 3 first leading components (continuous curve) and for all components of the filtered signal (dashed curve): $0 \leq \omega < 0.35$

to all components of the filtered signal. At the same time, in the case of 3 leading components all frequencies higher $\omega > 0.35$ are suppressed: see Fig. 36.

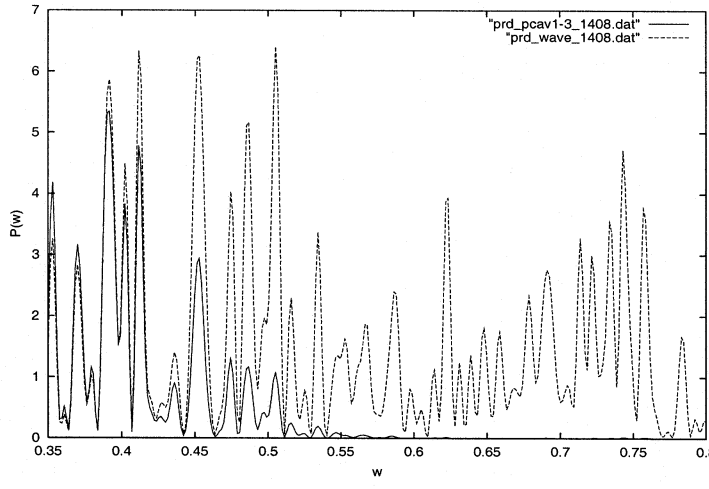

Figure 36: The dependence of $P_K(\omega)$ against the angular frequency $\omega = 2\pi f$ for 3 first leading components (continuous curve) and for all components of the filtered signal (dashed curve): $0.35 \leq \omega < 0.8$

9. A.Kolmogorov's scheme and log-normal distribution of network traffic

It has long been observed that in a large variety of physical phenomena, where self-similar processes take place, the logarithms of dynamical variables are normally distributed. This holds for grain sizes in crust fragmentation [45], for energy released in seismic events [46, 47], for the distribution of topographic contours, tree rings, leaves, rivers: see, for example, [48].

The theoretical explanation of appearance of the log-normal distribution in nature was first given, to our knowledge, by Andrei N. Kolmogorov in 1941 in a "small paper" [6] not well-known in the Western literature. Kolmogorov proposed a general scheme of a random process of the homogeneous fragmentation of grains.

A simplified explanation of Kolmogorov's result, see p. 206 in [46], is the following. Suppose that we have a big rock which crumbles into sand. If the environmental stresses are the same whatever the size of the rock, the probability that a given piece of rock is fragmented into n_i smaller rocks is independent of the stage i of the fragmentation process. Therefore, if we start out with a single rock ($n_0 = 1$), in the next stage we have n_1 smaller rocks, in the next stage each of these smaller rocks is

fragmented into n_2 still-smaller rocks, and so on. As the n_i are independent random variables, the number of grains at the k-th stage of fragmentation must be

$$N_k = \prod_{i=1}^{k} n_i = n_1 n_2 \cdots n_k, \tag{16}$$

or

$$\ln N_k = \sum_{i=1}^{k} \ln n_i. \tag{17}$$

The grain sizes S_k are inversely proportional to the number of grains N_k. Applying a variant the Central Limit Theorem, Kolmogorov found that the logarithms of the grain sizes were normally distributed [6], i.e. the distribution of grain sizes was log-normal.

The basic feature of log-normality is the power law or self-similarity. Let X and Y be two random variables. Then if X is log-normal and if

$$Y = aX^d, \tag{18}$$

Y is also log-normal. The parameter a is called the *scale factor* and the exponent d is the *fractal dimension.* Power laws such as (18) are known as *self-similarity relations.* Conversely, if both X and Y are known to be log-normal, there must exist a self-similarity relation, such as (18), between them. Kolmogorov invoked this property to deduce that, if the distribution of grain sizes of sand is log-normal, so are the grain volumes and the fractions by weight retained in sieves of different mesh size.

In [49] the wavelet transform has been applied to the self-similar stochastic processes, which Kolmogorov used in his theory of turbulence [6]. For such processes, after suitable re-scaling, the wavelet transform at predetermined position becomes a stationary random function of the logarithm of the scale argument in the transform [49]. The re-scaling depends on the scaling component.

Unfortunately, the approach of Vergassola and Frisch [49] can not be directly applied to network traffic measurements, because they have significantly a more complex structure [50, 51, 52].

However, the wavelet transform, being very powerful technique for extracting specific information from a given data [23, 24, 37], may provide additional information necessary for understanding the log-normality of traffic measurements. It has been shown (see, for instance, [54]) that the local signal regularity is characterized by the decay of the wavelet transform amplitude across scales. Singularities and edges are identified by following the wavelet transform local maxima at fine scales. All these features appear in complex signals like multi-fractals. The wavelet transform takes advantage of multi-fractal self-similarities, in order to compute the distribution of the singularities of the signals.

In order to reveal the self-similarity of traffic measurements at different scales, we applied the Continuous Wavelet Transform (CWT) to traffic measurements (Fig. 6).

Figure 37 shows the shade plot of the CWT, based on the biorthogonal spline wavelets, of the time series analyzed. The self-similar, multi-fractal character of traffic measurements is clearly shown in the tree-like fragmentation structure.

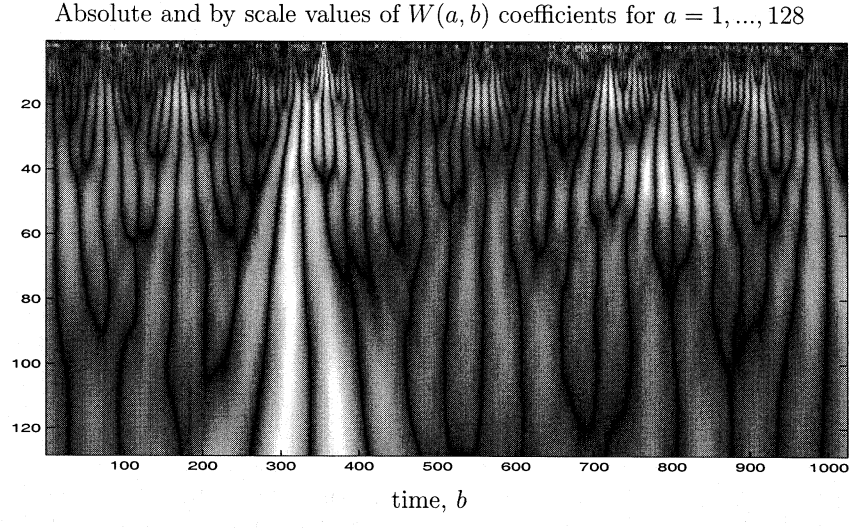

Figure 37: Shade plot of the CWT coefficients for traffic measurements aggregated with 1 s window

Figure 37 clearly demonstrates the multiplicative character of traffic measurements. This result is in agreement with formula (16) and confirms the applicability of the Kolmogorov's scheme to the description of network traffic.

Conclusion

Applying a nonlinear analysis to network traffic measurements and using a layered neural network for identification and reconstruction of the underlying dynamical system, we found that the trained neural network reproduced the statistical distribution of real data, which well fits the log-normal form [2]. Based on detailed traffic measurements we demonstrated that this distribution is caused by a simple aggregation of real data [3]. The "Caterpillar"-SSA [11, 12] and statistical analysis based on the joint utilization of χ^2 and ω^2 tests provided the possibility to divide the whole set of components into two classes [4]. The first class includes the leading components responsible for the main contribution to network traffic, and the second class involves residual components that can be interpreted as a stochastic noise. A detailed analysis of the boundary region between these two classes, based on the "Caterpillar"-SSA analysis, wavelet filtering and statistical χ^2 and ω_n^2 methods, demonstrated that the main part of the network traffic can be described by

a minimal number of feature components: three leading components for $C_L = 20$. We also found that the time series reconstructed on the basis of these components preserves main spectral characteristics of original traffic measurements. This may mean that all transformations realized on the original traffic series did not disturb its dynamical characteristics.

We hope that such simplification of a very complicated structure of the original traffic series may open additional possibilities for development of a more realistic dynamical model of network traffic and serve as a basis for elaboration of efficient Quality of Service (QoS) tools.

Acknowledgements

We are grateful to Prof. I. Prigogine and Prof. V. G. Kadyshevsky for encouragement and support.

This work has been partly supported by the European Commission in the frame of the Information Society Technologies program, the IMCOMP (IST-2000-26016) project.

References

[1] P. Akritas, I. Antoniou and V.V. Ivanov: *Internet Traffic Investigation*, Preprint SOLVAY 01-4, 2001, 93 pp.

[2] P. Akritas, P.G. Akishin, I. Antoniou, A.Yu. Bonushkina, I. Drossinos, V.V. Ivanov, Yu.L. Kalinovsky, V.V. Korenkov and P.V. Zrelov: *Nonlinear Analysis of Network Traffic*, "Chaos, Solitons & Fractals", Vol. **14(4)**(2002) pp.595-606.

[3] I. Antoniou, V.V. Ivanov, Valery V. Ivanov and P.V. Zrelov: *On a Log-Normal Distribution of Network Traffic*, Physica D **167** (2002) 72-85.

[4] I. Antoniou, V.V. Ivanov, Valery V. Ivanov and P.V. Zrelov: *Principal Components Analysis of Network Traffic: the "Caterpillar"-SSA Approach*, VIII Int. Workshop on Advanced Computing and Analysis Techniques in Physics Research, ACAT'2002, 24-28 June 2002, Moscow, Russia, Book of abstracts, p. 176 (submitted to Physica D).

[5] I. Antoniou, V.V. Ivanov, Valery V. Ivanov and P.V. Zrelov: *Wavelet Filtering of Network Traffic Measurements*, V Int. Congress on Mathematical Modeling, September 30-October 6, 2002, Dubna, Moscow region, Russia, Book of abstracts, p. 120, (to be submitted).

[6] A.N. Kolmogorov: Über das logarithmisch normale Verteilungsgesetz der Dimensionen der Teilchen bei Zerstückelung, *Dokl. Akad. Nauk SSSR*, **31**, pp. 99-101, 1941.

[7] Henry D.I. Abarbanel: *Analysis of Observed Chaotic Data*, 1996 Springer-Verlag New York, Inc.

[8] M.T. Lucas, D.E. Wrege, B.J. Dempsey, and A.C. Weaver: *Statistical Characterization of Wide-Area Self-Similar Network Traffic*, University of Virginia Technical Report CS97-04, October 9, 1996.

[9] M.T. Lucas, B.J. Dempsey, D.E. Wrege and A.C. Weaver: *(M,P,S) - An Efficient Background Traffic Model for Wide-Area Network Simulation*, Department of Computer Science, University of Virginia, Technical Report, 1997.

[10] J.I. Sánchez, F. Barceló and J. Jordán: *Inter-arrival Time Distribution for Channel Arrivals in Cellular Telephony*, in: Proc. 5-th Int. Workshop on Mobile Multimedia Communication, *MoMuc'98*, October 12-14 1998, Berlin.

[11] D.L. Danilov and A.A. Zhigljavsky, Eds.: *Principal Components of Time Series: Caterpillar Method*, St. Petersburg University Press, 1997.

[12] N. Golyandina, V. Nekrutkin, and A. Zhigljavsky: *Analysis of time series structure: SSA and related techniques*, Chapman & Hall/CRC, 2001.

[13] The State University "Dubna": http://www.uni-dubna.ru.

[14] P.V. Vasiliev, V.V. Ivanov, V.V. Korenkov, Yu.A. Kryukov and S.I. Kuptsov: *System for Acquisition, Analysis and Control of Network Traffic for the JINR Local Network Segment: the "Dubna" University Example*, JINR Communications, D11-2001-266, JINR, Dubna, RUSSIA, 2001.

[15] D. Kugiumtzis and M.A. Boudourides: *Chaotic Analysis of Internet Ping Data: Just a Random Number Generator?*, Contributed paper on the SOEIS meeting at Bielefeld, March 27-28, 1998.

[16] N.H. Packard, J.P. Crutchfield, J.D. Farmer and R.S. Shaw: *Geometry from a time series*, Phys. Rev. Lett. **45** (1980), 712.

[17] F. Takens: *Detecting strange attractors in turbulence* in "Dynamical Systems and Turbulence", edited by D. Rand and L.S. Young, Lecture Notes in Mathematics **898** (Springer-Verlag, Berlin, 1981), 366.

[18] D.S. Broomhead and G.P. King: *Extracting qualitative dynamics from experimental data*, Physica **20D** (1986), 217.

[19] A.M. Albano, J. Muench, C. Schwartz, A.I. Mees, and P.E. Rapp: *Singular value decomposition and the Grassberger Procaccia algorithm*, Phys. Rev. **A38** (1988), 3017.

[20] P. Grassberger and I. Procaccia: *Characterization of strange attractors*, Phys. Rev. Lett. **50** (1983), 346.

[21] C.D. Cutler: *A theory of correlation dimension for stationary time series*, Phil. Trans. Roy. Soc. Lond. **A348** (1994), 343.

[22] P. Grassberger and I. Procaccia: *Measuring the strangeness of strange attractors*, Physica **9D** (1983), 189

[23] C.K. Chui: *An Introduction to Wavelets.* Academic Press: New York, 1-18(1992).

[24] I. Daubechies: *Wavelets*, Philadelphia: S.I.A.M., 1992.

[25] S. Haykin: *Neural Networks: A Comprehensive Foundation*, Prentice-Hall, Inc., 1999.

[26] P.D. Wasserman: *Neural Computing: Theory and Practice*, Van Nostrand Reinhold, 1989.

[27] Duc Truong Pham and Liu Xing: *Neural Networks for Identification, Prediction and Control*, Springer-Verlag London Limited, 1995.

[28] A. Lapedes and R. Farber: *Nonlinear Signal Processing using Neural Networks: Prediction and System Modeling*, Los Alamos Report LA-UR 87-2662, 1987.

[29] P. Akritas, I. Antoniou and V.V. Ivanov: *Identification and Prediction of Discrete Chaotic Maps Applying a Chebyshev Neural Network*, Chaos, Solitons and Fractals **11** (2000) 337-344.

[30] C. Peterson and Th. Rongvaldsson: *JETNET-3.0 – A Versatile Artificial Neural Network Package*, LU Tp 93-29, 1993.

[31] E. Oja: *Data Compression, Feature Extraction, and Autoassociation in Feedforward Neural Networks*, In "Artificial Neural Networks" (T. Kohonen, K. Mäkisara, O. Simula and J. Kangas, eds.), Vol. **1**, pp. 737-746, Amsterdam, North-Holland, 1991.

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

[33] P. Baldi and K. Hornik: *Neural Networks and Principal Component Analysis: Learning from Examples Without Local Minimum*, Neural Networks, vol. **1**, pp. 53-58, 1989.

[34] W.T. Eadie, D. Dryard, F.E. James, M. Roos and B. Sadoulet: *Statistical Methods in Experimental Physics*, North-Holland Pub.Comp., Amsterdam-London, 1971.

[35] F. James: *MINUIT – Function Minimization and Error Analysis*, Reference manual, version 94.1, CERN Program Library D506, 1998.

[36] R. Brun, O. Couet, C. Vandoni and P. Zanarini: *PAW - Physics Analysis Workstation*, CERN Program Library Q121, 1989.

[37] W.H. Press, S.A. Teukolsky, W.T. Vetterling and B.P. Flannery: *Numerical Recipes in C: The Art of Scientific Computing*, II-d Edition, Cambridge University Press 1988, 1992.

[38] D. Donoho, I. Jonhstone, G. Kerkyacharian and D. Picard: *Density Estimation by Wavelet Thresholding*, Technical report, Department of Statistics, Stanford University, 1993.

[39] G.P. Nason and B.W. Silverman: *The discrete wavelet transform in S*, Journal of Computational and Graphical Statistics, vol. 3, pp. 163-191, 1994.

[40] G.V. Martinov: *Omega-squared criteria*, Moscow, "Nauka", 1978 (in Russian).

[41] N.R. Lomb: *Astrophysics and Space Science*, vol. **39**, 1976, pp. 447-462.

[42] J.D. Scargle: *Astrophysical Journal*, vol. **263**, 1982, pp. 835-853.

[43] J.H. Horne and S.L. Baliunas: *Astrophysical Journal*, vol. **302**, 1986, pp. 757-763.

[44] S.G. Mallat: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 674-693, 1989.

[45] N.K. Razumovsky: On a Distribution Character of Metals Contents in Ore Fields, *Dokl. Akad.Nauk SSSR*, **28**, pp. 815-817, 1940 (in Russian).

[46] Cinna Lomnitz: *Fundamentals of Earthquake Prediction*, John Wiley & Sons, Inc. 1994.

[47] V.I. Keilis-Borok: *Symptoms of Instability in a System of Earthquake-Prone Faults*, Physica D, **77**, pp. 193-199, 1994.

[48] J. Aitchison and J.A.C. Brown: *The Lognormal Distribution*, Cambridge, Univ. Press, Cambridge, 176 pp., 1957.

[49] M. Vergassola and U. Frisch: *Wavelet Transforms of Self-Similar Processes*, Physica D **54** (1991) 58-64.

[50] M.S. Taqqu, V. Teverovsky and W. Willinger: *Is Network Traffic Self-Similar or Multifractal?*, Fractals, 1996.

[51] G. Taubes: *Fractals Reemerge in the New Math of the Internet*, Science, Vol. **281**, pp. 1947-1948, 1998.

[52] R.H. Riedi, M.S. Crouse, V.J. Riberio and R.G. Baraniuk: *A Multifractal Wavelet Model with Application to Network Traffic*, IEEE Trans. on Information Theory, Vol. **45**, No. 3, 1999.

[53] A.K. Louis, P. Maa*b* and A. Rieder: :*Wavelets. Theory and Applications*, John Wiley & Sons, 1997.

[54] S. Mallat: *A Wavelet Tour of Signal Processing*, Academic Prees, 1999.

Received on November 5, 2002.

Антониоу Я. и др. E11-2002-222
Статистическая модель информационного трафика

В [1,2] мы применили нелинейный анализ к измерениям информационного трафика, полученным на выходном шлюзе локальной сети среднего размера. Реалистичные величины временного сдвига и вложенной размерности обеспечили возможность применения прямоточной нейронной сети для идентификации и реконструкции лежащей в его основе динамической системы. Обученная на этих данных нейронная сеть воспроизвела статистическое распределение агрегированных пакетов реальных данных, которое хорошо фитируется логнормальным распределением. Детальный анализ измерений трафика [3] показал, что такое распределение возникает в результате агрегации реальных данных. Анализ принципиальных компонентов измерений трафика продемонстрировал, что уже несколько лидирующих компонентов формируют фундаментальную часть сетевого трафика, в то время как остаточные компоненты играют роль небольших нерегулярных вариаций, которые могут быть интерпретированы как стохастический шум [4]. Этот результат был поддержан применением вейвлет-фильтрации и фурье-анализа как к исходным измерениям трафика, так и к отдельным принципиальным компонентам оригинальных и отфильтрованных данных [5]. Логнормальное распределение агрегированных измерений и мультипликативный характер временной серии трафика подтверждает применимость схемы, разработанной А. Колмогоровым [6] к однородной фрагментации крупинок, также и для сетевого трафика.

Работа выполнена в Лаборатории информационных технологий ОИЯИ.

Сообщение Объединенного института ядерных исследований. Дубна, 2002

Antoniou I. et al. E11-2002-222
Statistical Model of Network Traffic

In [1,2] we applied a nonlinear analysis to traffic measurements obtained at the input of a medium size local area network. The reliable values of the time lag and embedding dimension provided the application of a layered neural network for identification and reconstruction of the underlying dynamical system. The trained neural network reproduced the statistical distribution of real data, which well fits the log-normal form. The detailed analysis of traffic measurements [3] has shown that the reason of this distribution may be a simple aggregation of real data. The principal components analysis of traffic series demonstrated that a few first components already form the fundamental part of network traffic, while the residual components play a role of small irregular variations that can be interpreted as a stochastic noise [4]. This result has been confirmed by application of the wavelet filtering and Fourier analysis to both the original traffic measurements and individual principal components of original and filtered data [5]. The log-normal distribution of traffic measurements and a multiplicative character of traffic series confirms the applicability of the scheme, developed by A. Kolmogorov [6] for the homogeneous fragmentation of grains, also to the network traffic.

The investigation has been performed at the Laboratory of Information Technologies, JINR.

Communication of the Joint Institute for Nuclear Research. Dubna, 2002

Макет *Т. Е. Попеко*

Подписано в печать 15.11.2002.
Формат 60 × 90/16. Бумага офсетная. Печать офсетная.
Усл. печ. л. 2,68. Уч.-изд. л. 3,97. Тираж 310 экз. Заказ № 53616.

Издательский отдел Объединенного института ядерных исследований
141980, г. Дубна, Московская обл., ул. Жолио-Кюри, 6.
E-mail: publish@pds.jinr.ru
www.jinr.ru/publish/