Statistical Model of Network Traffic

I. Antoniou[1,2], V.V. Ivanov[1,3], Valery V. Ivanov[3,4] and P.V. Zrelov[3]

[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] *Laboratory of Information Technologies, Joint Institute for Nuclear Research, 141980, Dubna, RUSSIA*

[4] *University Scientific Center, Joint Institute for Nuclear Research, 141980, Dubna, RUSSIA*

In [1] we applied a nonlinear analysis to traffic measurements obtained at the input of a medium size LAN. 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 [2] 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 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. This result has been confirmed by application of wavelet filtering and Fourier analysis both to original traffic measurements and individual principal components of original and filtered data [3]. The applicability of the scheme, developed by A. Kolmogorov [4] for the homogeneous fragmentation of grains, to the network traffic is discussed.

[1] 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.

[2] 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.

[3] I. Antoniou, V.V. Ivanov, Valery V. Ivanov and P.V. Zrelov: *Wavelet Filtering of Network Traffic Measurements*, Physica A **324** (2003) 733-753.

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

1. Data acquisition system

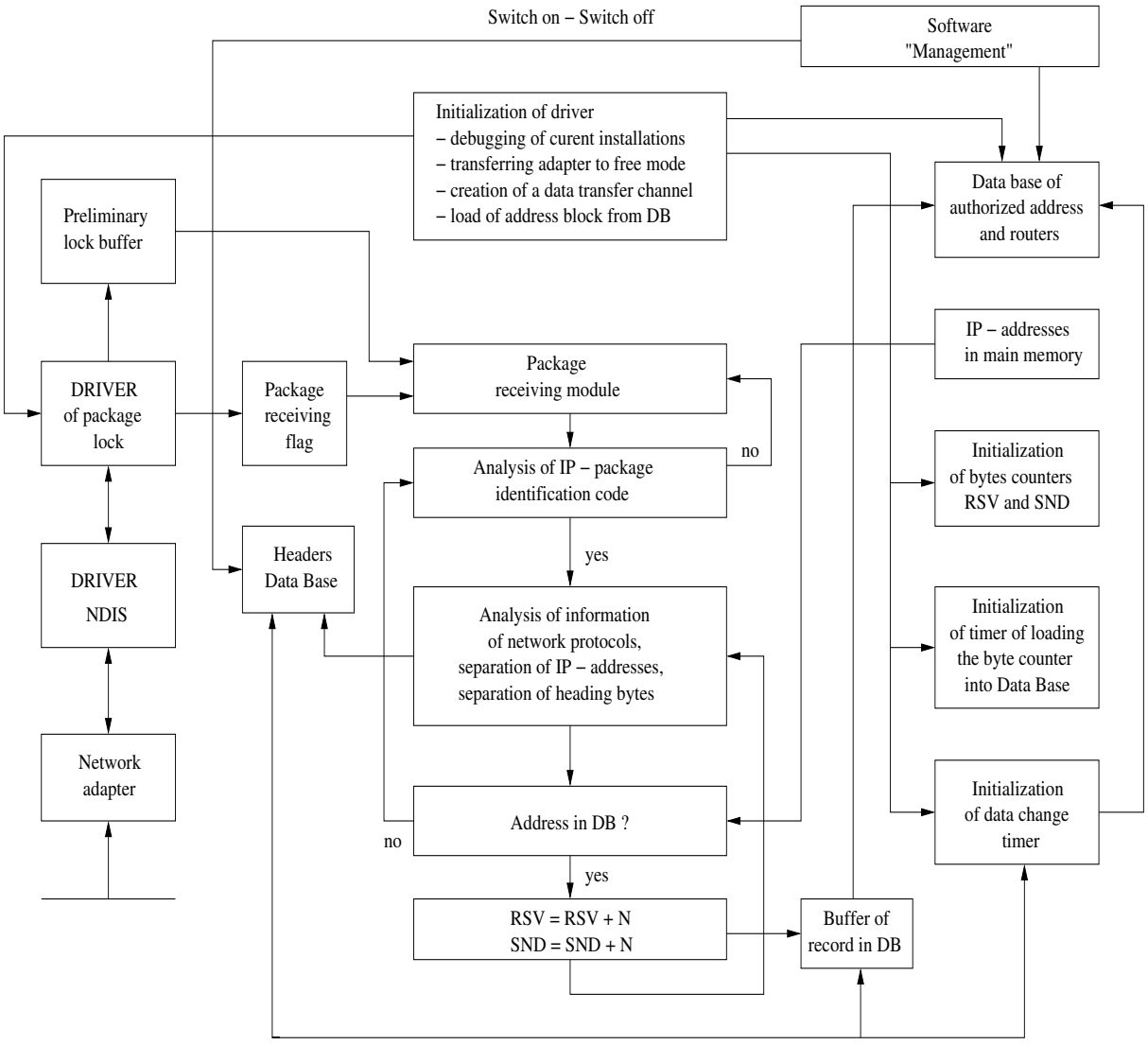

Figure 1: The scheme of data acquisition system

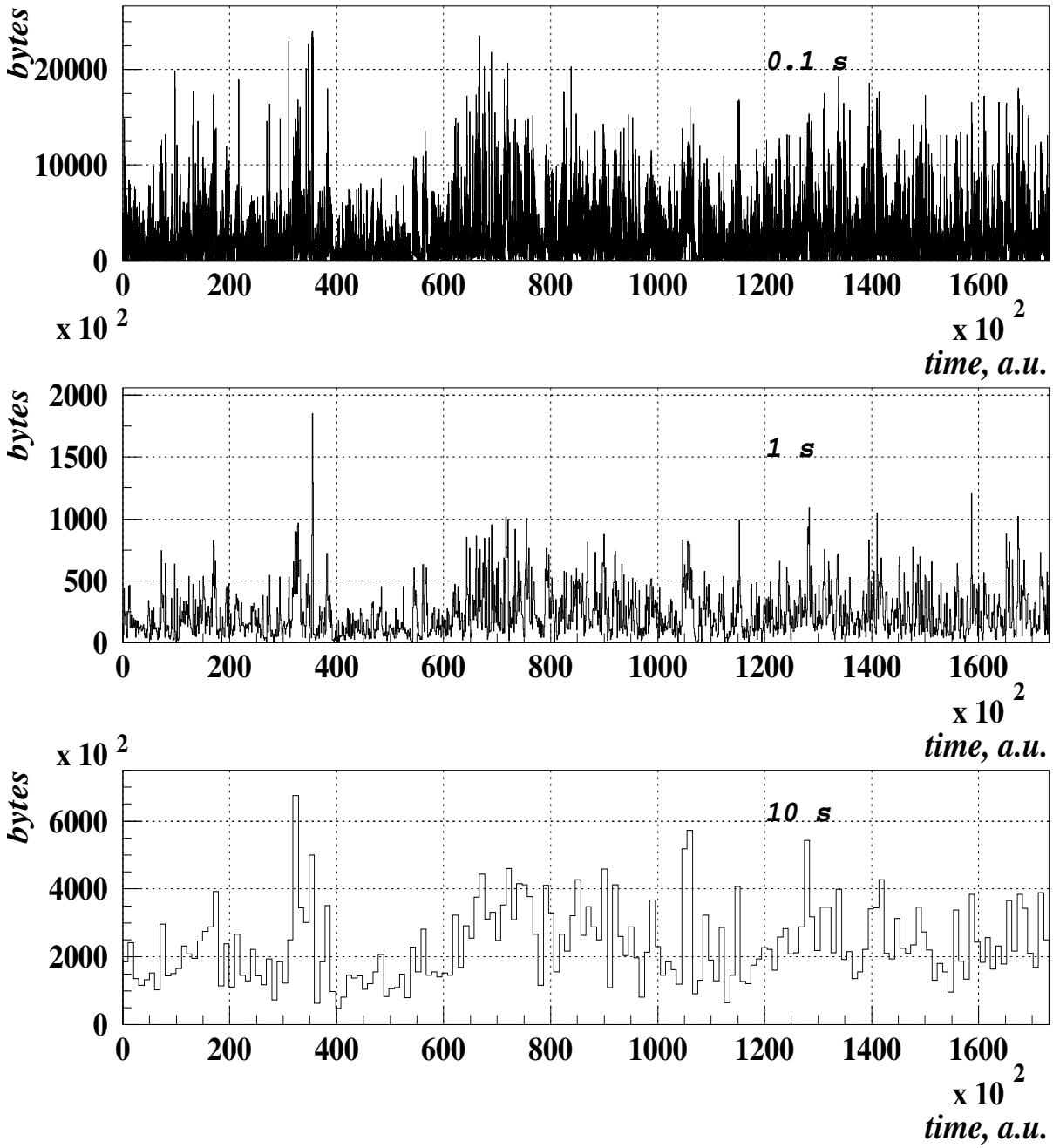

Figure 2: Traffic measurements aggregated with different bin sizes: 0.1 sec, 1 sec, 10 sec

2. Nonlinear analysis of network traffic

In nonlinear analysis the series $\{x_i\}$ is considered as one-dimensional projection of a system operating in space $\vec{y}_i$ of larger dimension:

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

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

The "*phase space reconstruction*" includes three main steps:

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

2.1. Estimating the correlation length

In order to choose independent components from $\{x_i\}$, one may compute the correlation length τ using the auto-correlation function

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

where N is number of points in series. The dependence of the correlation length against the aggregation bin size is presented in Fig. 3. One can see that for bin sizes from 0.1 sec up to 10 sec, τ is in acceptable region: $\tau \sim 10$ sec. Points separated by the time interval τ can be considered as independent.

2.2. Estimating the embedding dimension

m can be estimated applying the Grassberger-Procaccia algorithm:

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

with the distance between two points given by

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

where $\Theta = 1$ if its argument is non-negative and 0 otherwise.

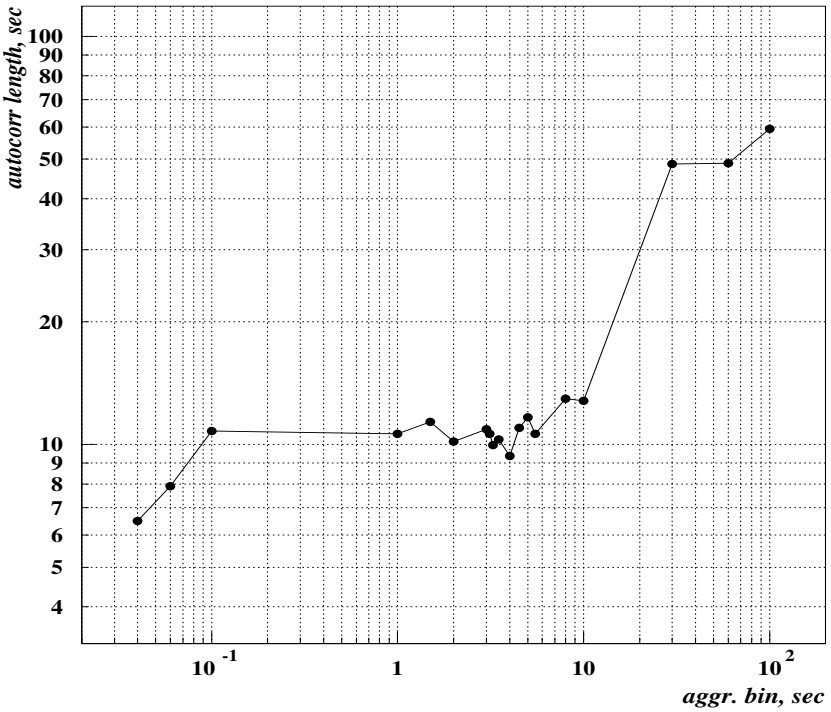

Figure 3: The dependence of the correlation length against the size of the aggregation bin

The slope $\log C_2^m(r)$ vs. $\log r$ gives the estimate of the embedding dimension: see Fig. 4. No saturation of the slope with respect to increasing m was found, which may mean a very high dimension of the time series. We may

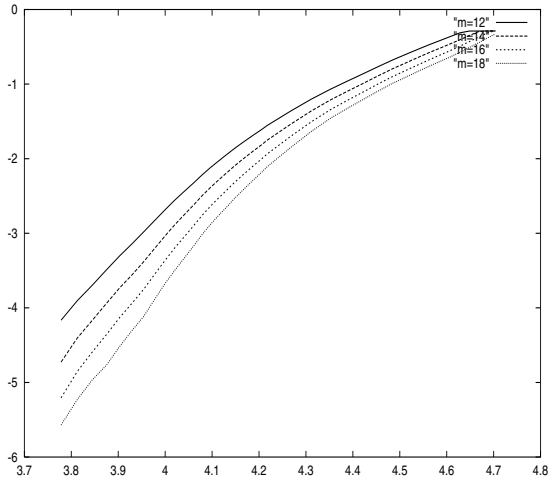

Figure 4: The dependences of $\log C_2^m(r)$ vs. $\log r$ for traffic measurements aggregated with 1 sec bin: $\tau = 10$ sec and m =12, 14, 16, 18

consider traffic measurements as a sum of a regular process and a stochastic part, related to the high frequency *"noise"*, that can be eliminated. In order

to achieve this, we used a discrete wavelet transform based on Daubechies wavelets. As a result we found that m about $16 \div 18$ seems to be close to saturation.

2.3 Reconstruction of underlying system

Reliable values of τ and m provided the application of a layered ANN for reconstruction of the underlying system. We used the ANN with the feed-forward architecture: the input layer with m neurons, two hidden layers with varying number of neurons and one output neuron for getting the predicted value of the ANN model.

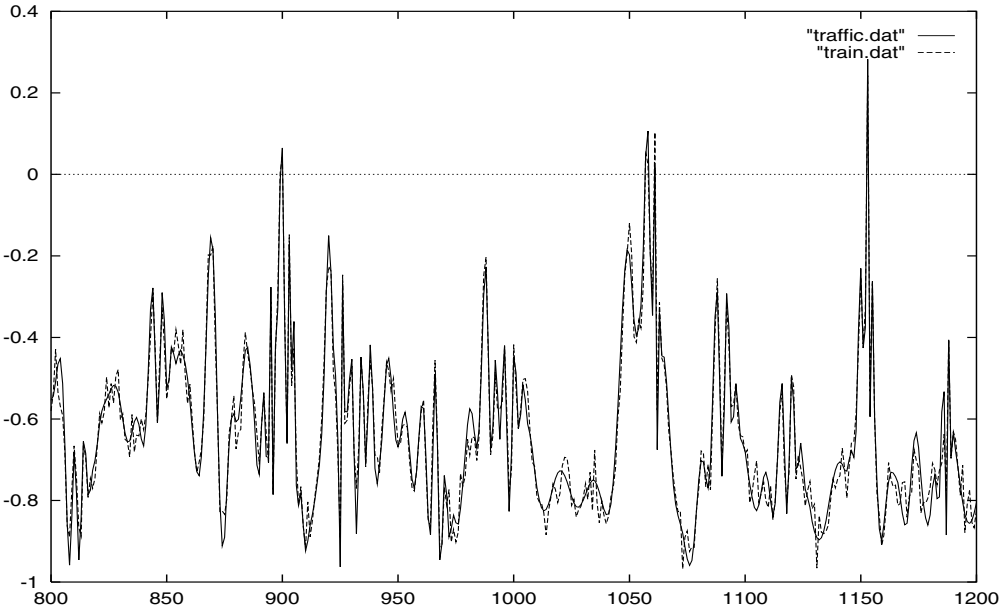

Figure 5: The result of the ANN approximation of the traffic series

Figure 6 demonstrates packet size distributions (normalized to the interval [-1,1]) for original measurements (top figure) and for series generated by the trained ANN (bottom figure). We see that the ANN model quite well reproduces the statistical distribution of real data, which seems to be log-normal.

The distribution of ANN weights between the output neuron and second hidden layer nodes seems quite interesting: see Fig. 7.

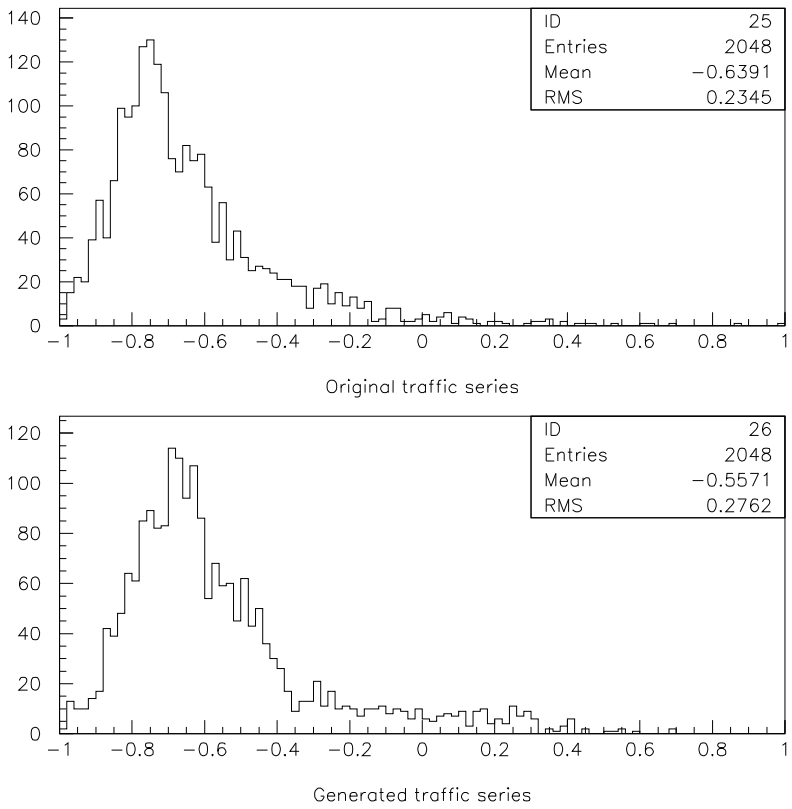

Figure 6: The distribution of packet sizes (normalized to interval [-1,1]) for: a) original traffic measurements, and b) generated by the trained ANN

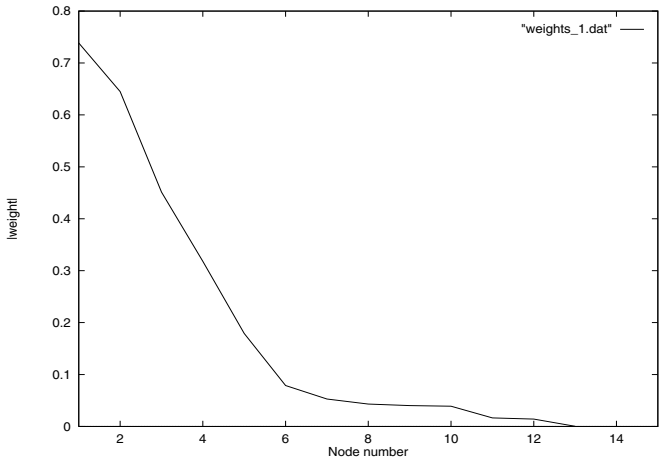

Figure 7: The distribution of absolute values of weights between the output node and second hidden layer nodes of the trained ANN

3. Log-normal distribution of network traffic

Figure 8 shows the packet size distribution for original traffic measurements, while figures 9, 10 and 11 present distributions for measurements aggregated with bin sizes $10\,ms$, $100\,ms$ and $1\,s$, correspondingly.

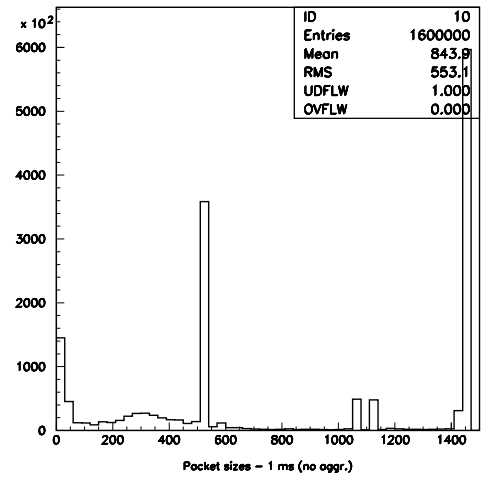

Figure 8: Packet size distribution for original traffic measurements

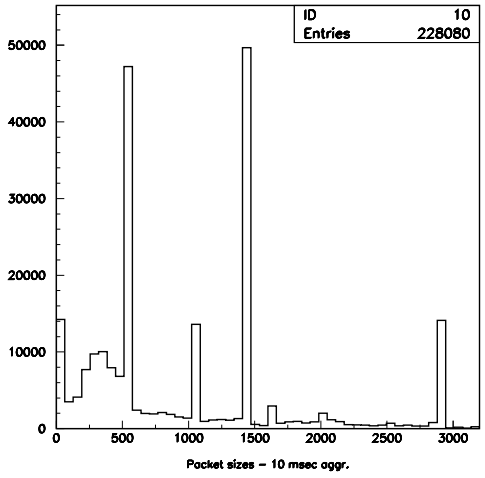

Figure 9: Packet size distribution for traffic measurements aggregated with bin size $10\,ms$

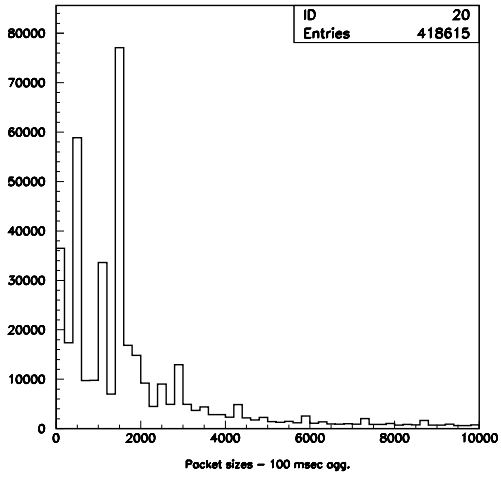

Figure 10: Packet size distribution for traffic measurements aggregated with bin size $100\,ms$

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

corresponding to the bin size 10 s.

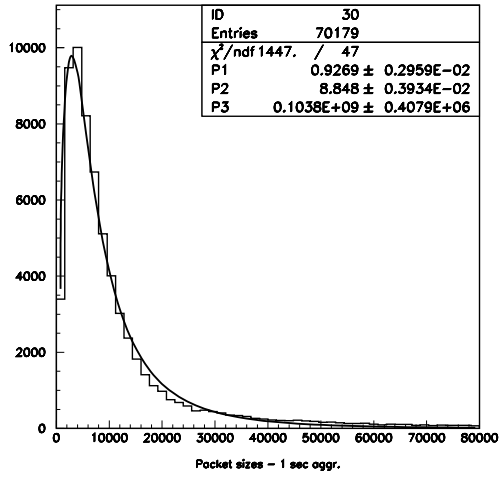

Figure 11: Packet size distribution for traffic measurements aggregated with bin size 1 s: fitting curve corresponds to function (4)

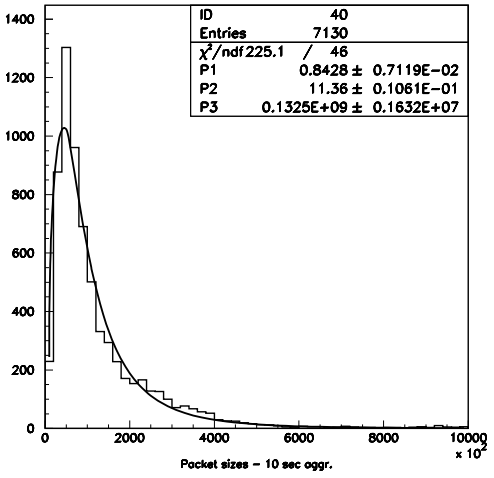

Figure 12: Packet size distribution for traffic measurements aggregated with bin size 10 s: fitting curve corresponds to function (4)

Distributions in figures 11 and 12 are well approximated by the log-normal function

$$f(x) = \frac{A}{\sqrt{2\pi}\sigma}\frac{1}{x}\exp\left[-\frac{1}{2\sigma^2}(\ln x - \mu)^2\right], \qquad (4)$$

x is the variable, σ and μ are parameters and A is the normalizing multiplier. However, they did not pass the χ^2-test.

The reason is that these distributions are based on the whole set of data, which corresponds approximately to 20 hours of continuous measurements. But the traffic series behave differently depending, if measurements were done during working hours or not. In this connection, we tested only the daily traffic. The results of this analysis are presented in Table 1.

Table 1: Results of fitting of daily part of packet size distributions aggregated with different bin sizes by function (4)

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

These results show that the hypothesis (4) is accepted with a high probability: see also Fig. 13. At the same time it must be noted that the influence of the inactive period of LAN does not change seriously the fundamental form of the statistical distribution.

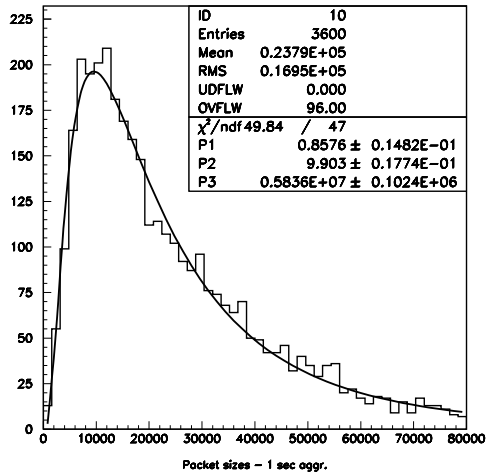

Figure 13: Packet size distribution for daily traffic measurements aggregated with bin size $1\,s$: fitting curve corresponds to the function (4)

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.

4. Principal component analysis of network traffic

The "Caterpillar"-SSA approach can be used for analysis of time series corresponding to a function $f(t), t > 0$ determined in equidistant points. 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

$$x_i = f[i] = f[(i-1)\Delta t], \quad i = 1, 2, \ldots, M \tag{5}$$

into multidimensional one is realized by representing (5) in matrix form:

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

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

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

$$Y = V^T X = (Y_1, Y_2, \ldots, Y_L),$$

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

The "Caterpillar" length C_L has been chosen based on the analysis of the autocorrelation length τ: we used different values of C_L, starting from $C_L = 12$ up to $C_L = 20$.

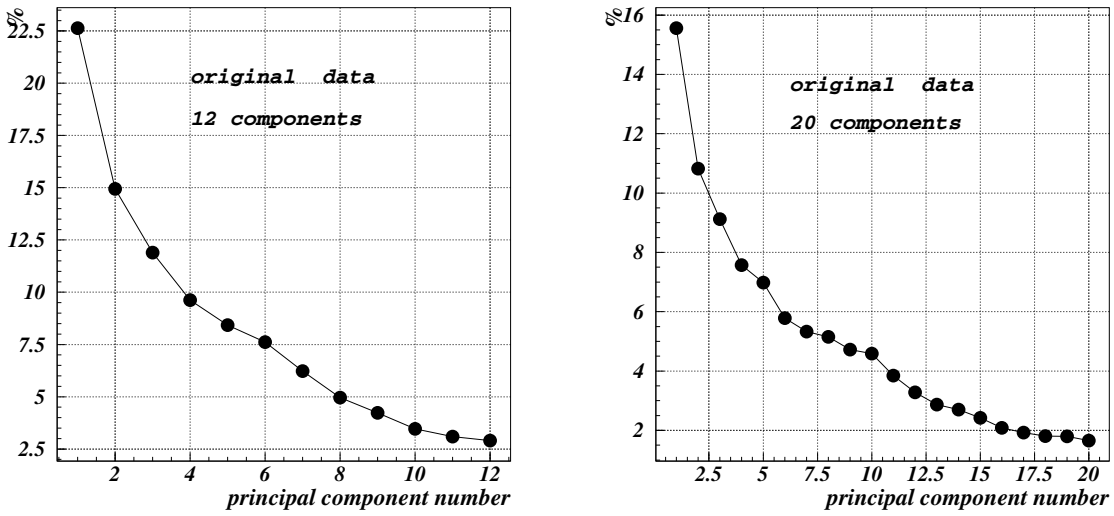

Figure 14: Contributions of eigenvalues in percentages for original traffic data. Results are presented for two cases of the caterpillar length: $C_L = 12$ (left) and 20 (right)

It is reasonable to assume that distributions, corresponding to leading components, may be described by the log-normal distribution. Figure 15 shows the dependence of χ^2/ν versus the number of leading components.

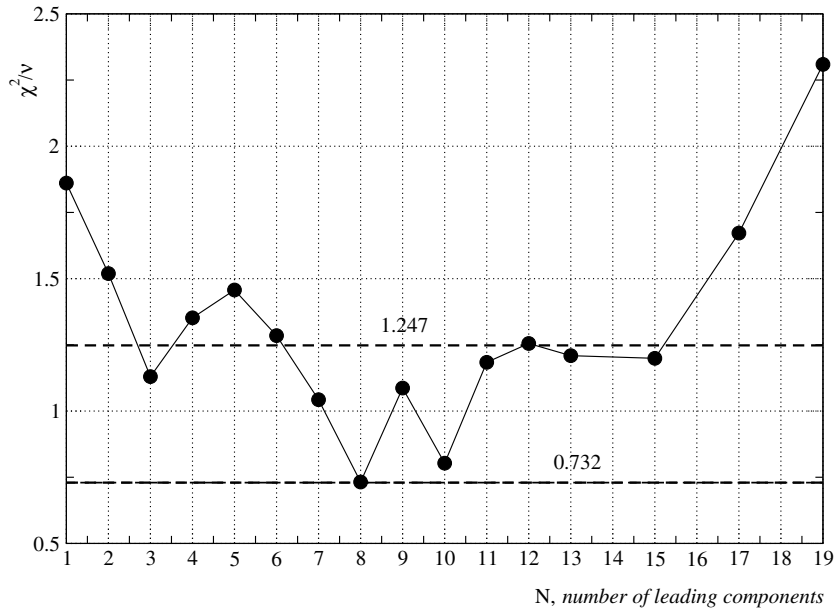

Figure 15: The dependence of χ^2/ν versus the number of leading components

It demonstrates a quit good level of correspondence ($\alpha = 22\%$) of the distribution to the null-hypothesis for $N = 3$: see also Fig. 16.

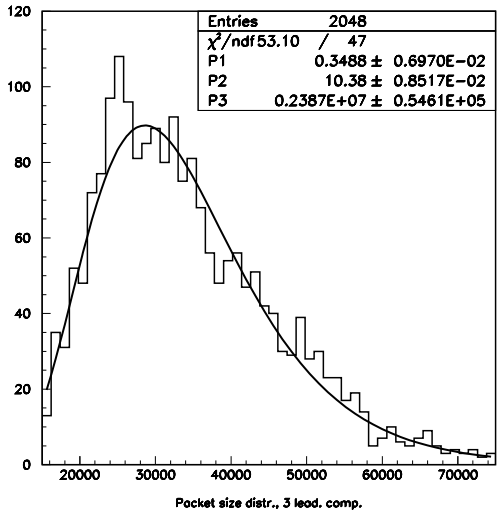

Figure 16: Fitting distribution corresponding to 3 leading components by function (4)

In the region of large N there is a growth of χ^2 especially noticeable at $N \geq$

15: see Fig. 15. Figure 17 shows the series corresponding to the component 20. It looks like a nonstationary process symmetric against zero mean value.

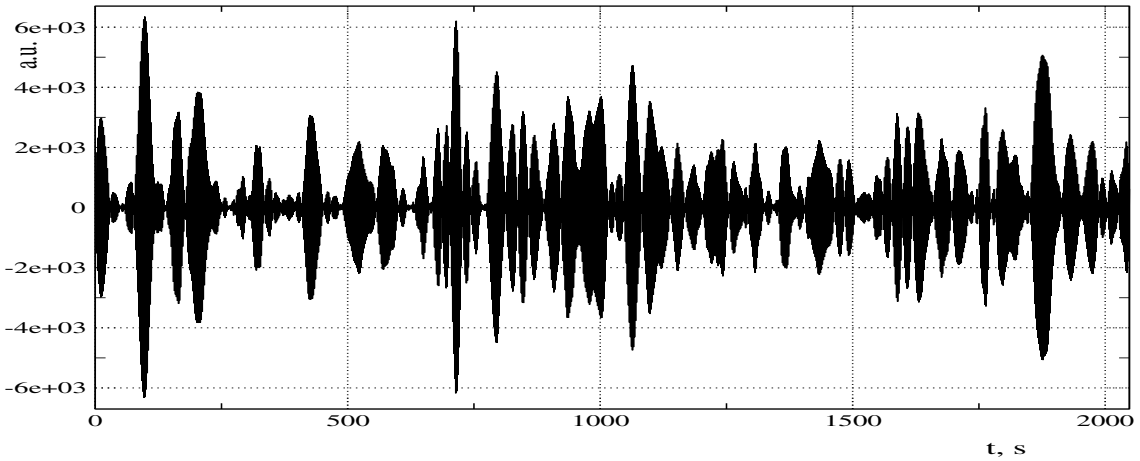

Figure 17: Traffic series reconstructed by the caterpillar method ($C_L = 20$) on the basis of the smallest component

In order to estimate the amount of residual components, which can be eliminated from the original series, we divide all components into two parts:

1. *first part corresponding to the leading components and responsible for the log-normal form of the packet size distribution*,
2. *second part related to residual components, which is described by a symmetric statistical distribution and behaves like a stochastic noise.*

As criterion for selection of second part the sign test has been used for testing the symmetry against zero for residual distributions:

$$\mu = \sum_{i=1}^{n} \Theta(X_i), \tag{6}$$

where $X_1, \ldots, X_n$ are observables, n is the sample size, and

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

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 and $p = 0.5$.

Figure 18 shows that μ exceeds the reliable confidential level, when number of residuals is greater 6 for $C_L = 12$ and 11 for $C_L = 20$.

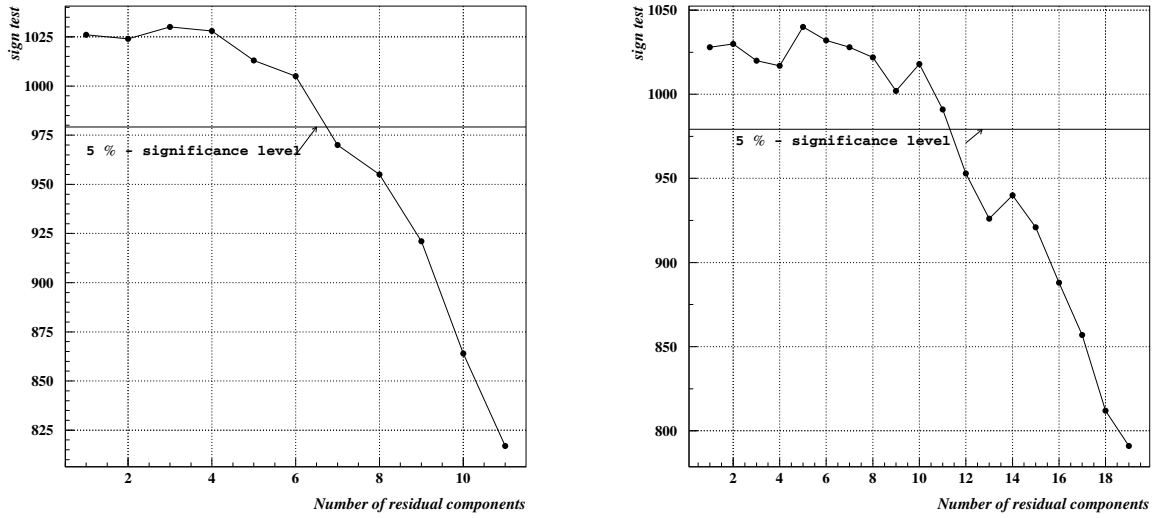

Figure 18: The dependence μ versus the number of residual components for the caterpillar length $C_L = 12$ (left figure) and $C_L = 20$ (right figure)

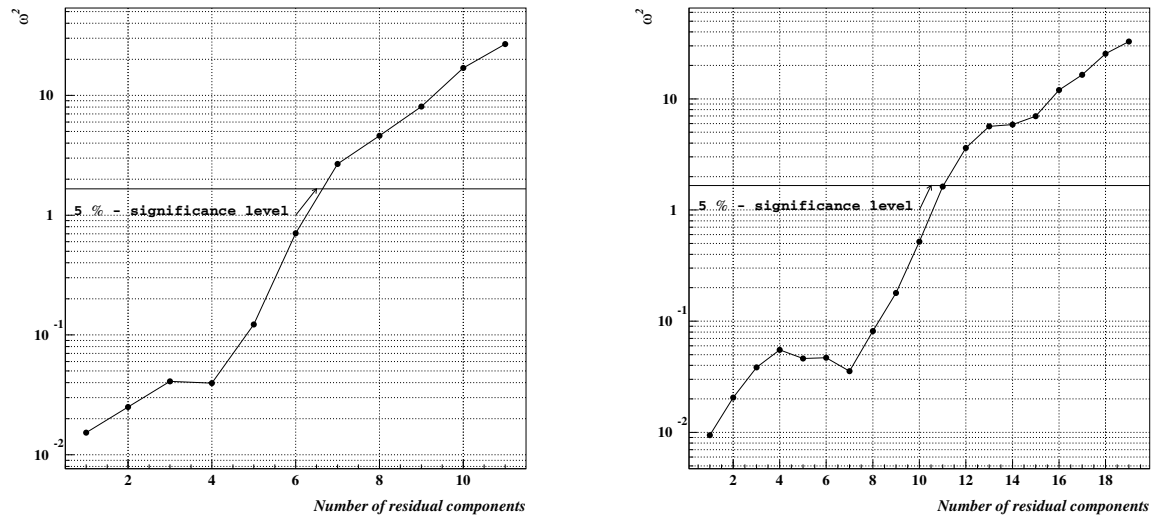

Figure 19: Dependences of ω_n^2 versus the number of residual components for two cases of the caterpillar length: $C_L = 12$ (left figure) and $C_L = 20$ (right figure)

In order to confirm these results, we applied more powerful ω_n^2 criterion. It tests the symmetry of the distribution function $F(x)$ of observables

$X_1, \ldots, X_n$, i.e. the null-hypothesis H_0: $F(x) = 1 - F(x)$:

$$\omega_n^2 = \sum_{j=1}^{n} \Big[F_n(-X_{(j)}) - \frac{n-j+1}{n}\Big]^2,$$

where $X_{(1)} \leq \ldots \leq X_{(n)}$ is the variational series constructed on the basis of observables.

Figure 19 shows that number of residuals $k = 6$ for $C_L = 12$ and $k = 11$ for $C_L = 20$ correspond to the 5% - significance level. This coincides with result for the sign test.

5. Spectral analysis of traffic measurements

In order to estimate the presence or absence of periodic components and to evaluate the viability of stochastic noise in the traffic series, we applied the Lomb spectral method.

The Lomb *normalized periodogram* (spectral power as a function of angular frequency $\omega \equiv 2\pi f > 0$) of one-dimensional time series (5) is defined by

$$P_K(\omega) = \frac{1}{2\pi^2}\left\{\frac{\Big[\sum\limits_{i=1}^{K}(x_i - \bar{x})\cos\omega(t_i - \tau)\Big]^2}{\sum\limits_{i=1}^{K}\cos^2\omega(t_i - \tau)} + \frac{\Big[\sum\limits_{i=1}^{K}(x_i - \bar{x})\sin\omega(t_i - \tau)\Big]^2}{\sum\limits_{i=1}^{K}\sin^2\omega(t_i - \tau)}\right\}, \tag{7}$$

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 α of a peak in the spectrum $P_K(\omega)$, we test the null-hypothesis that the data values are independent Gaussian random values.

Figure 20 shows the result of application of the Lomb method to the traffic measurements. The horizontal dashed and dotted lines correspond (from bottom to top) to $\alpha = 0.5, 0.1, 0.01, 0.001$, respectively. One can see (Fig. 21) 3 highly significant peaks at low frequencies: 0.06, 0.012 and 0.034. For

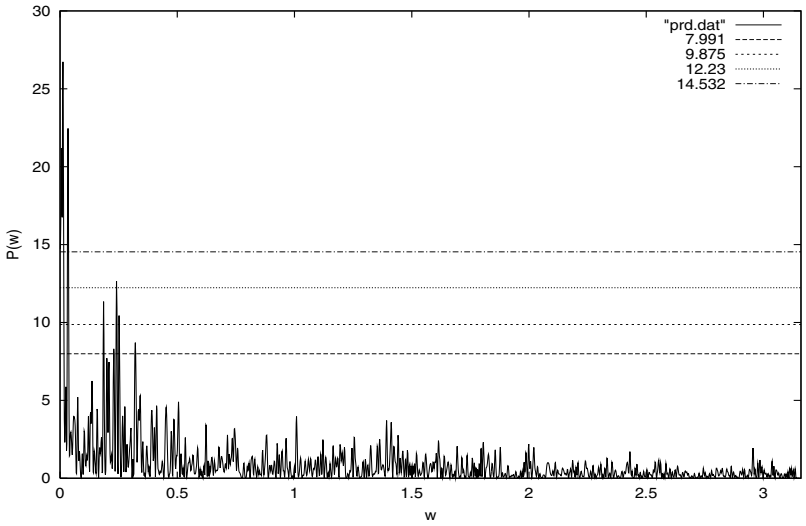

Figure 20: The dependence of $P_K(\omega)$ against the angular frequency ω for traffic measurements: $0 \leq \omega < 2\pi f_c$

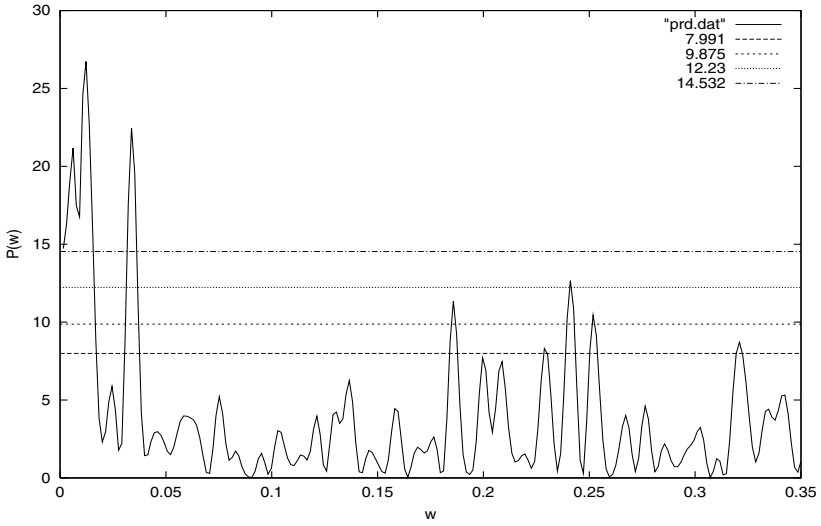

Figure 21: The dependence of $P_K(\omega)$ against the angular frequency ω for traffic measurements: $0 \leq \omega < 0.35$

$\omega > 0.35$ together with ω increase, the amplitude is very quickly decreasing (Fig. 20) without exceeding 5, which corresponds to $\alpha \approx 1$. This means that traffic components related to this part can be interpreted as Gaussian noise.

6. Wavelet filtering of traffic measurements

The function $f(t) \in L_2(R)$ can be represented in terms of shifts and dilations

of a low-pass scaling function $\phi(t)$ and band-pass wavelet $\psi(t)$:

$$f(t) = \sum_k s_k^J \phi(2^J t - k) + \sum_{j \geq J} \sum_{k \in Z} d_k^j \psi(2^j t - k), \tag{8}$$

Figure 22 shows the dependence of ω_n^2 versus the number of rejected coefficients: 1408 coefficients can be eliminated without exceeding the 5%-significance level.

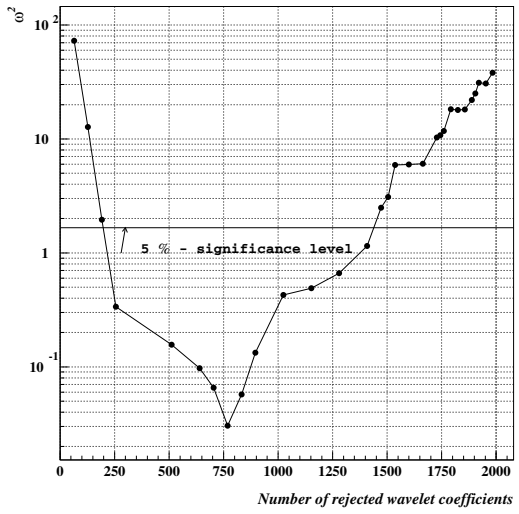

Figure 22: The dependence of ω_n^2 values versus the number of rejected wavelet coefficients

The autocorrelation function can be also used as a criterion for evaluation of the noisy part: Fig. 23 shows that up to $M = 1408$ the rejected part can be considered as noisy.

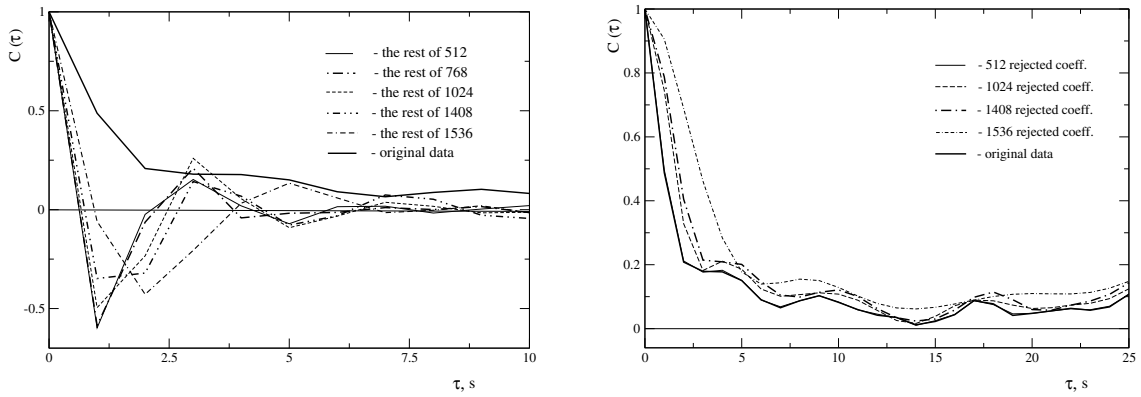

Figure 23: Autocorrelation functions $C(\tau)$ of noisy (left plot) and smooth (right plot) parts corresponding to different number of rejected coefficients

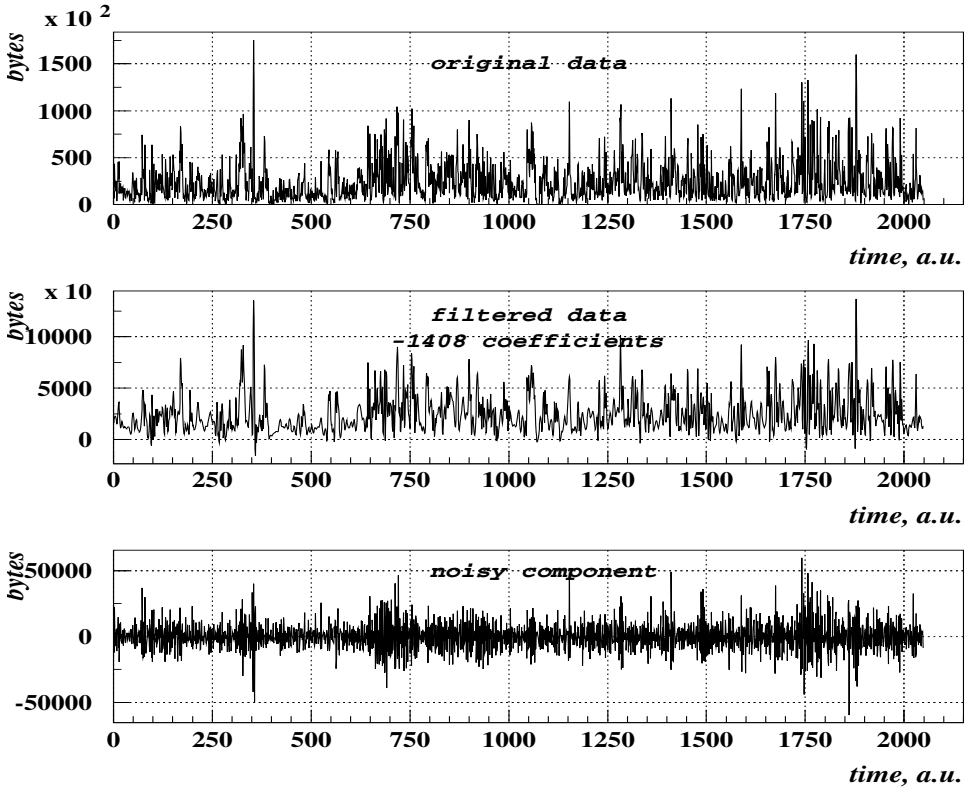

Figure 24: Traffic measurements: 1) original traffic series, 2) filtered signal, 3) noisy component

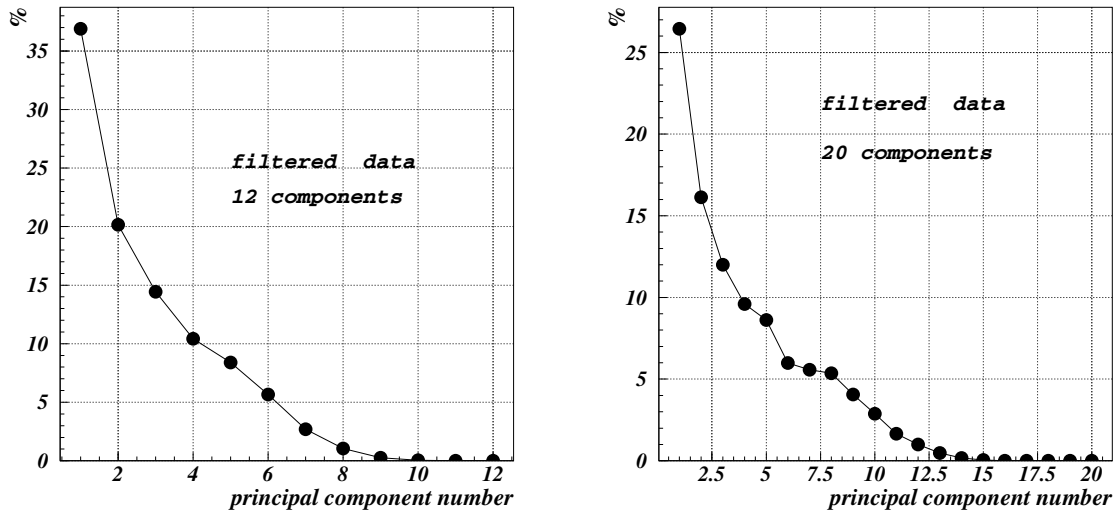

Figure 25: Contributions of eigenvalues in percentages for the traffic data after filtering out the high-frequency part. The results are presented for two cases of the caterpillar length: $C_L = 12$ (left) and 20 (right)

Figure 26 shows the dependence of $P_K(\omega)$ against ω for 3 leading components (continuous curve) and for all components of the filtered signal (dashed curve). This dependence clearly demonstrates that low frequency region of

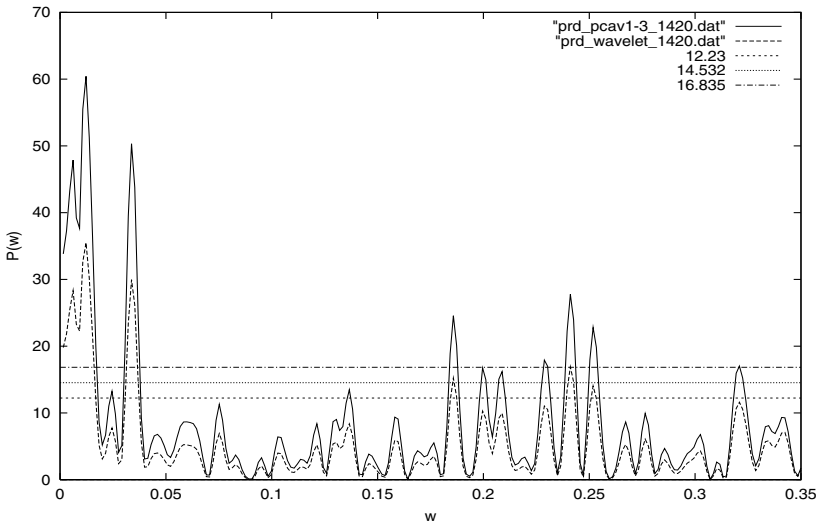

Figure 26: The dependence of $P_K(\omega)$ against the angular frequency ω for 3 first leading components (continuous curve) and for all components of the filtered signal (dashed curve): $0 \leq \omega < 0.35$

traffic series is formed by 3 leading components. At the same time, all frequencies higher $\omega > 0.35$ are suppressed: see Fig. 27.

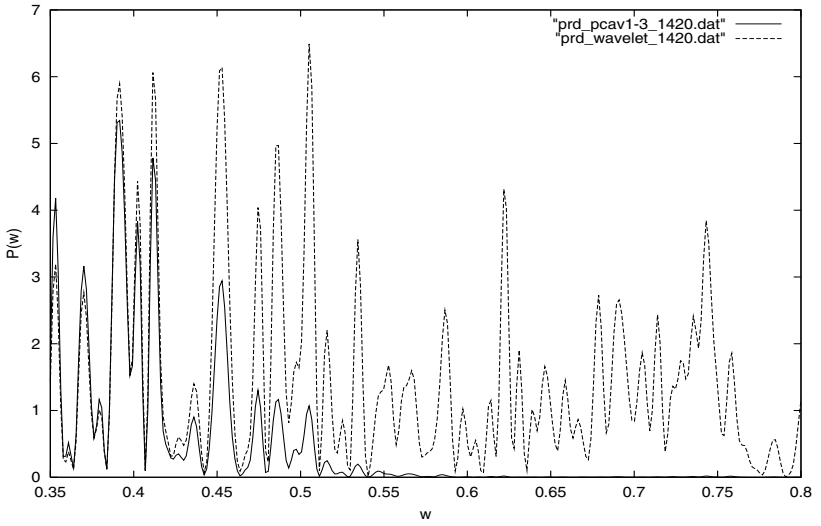

Figure 27: The dependence of $P_K(\omega)$ against the angular frequency ω for 3 first leading components (continuous curve) and for all components of the filtered signal (dashed curve): $0.35 \leq \omega < 0.8$

The dependence (Fig. 28) confirms our previous result concerning the number

of leading components that forms the fundamental part of information traffic.

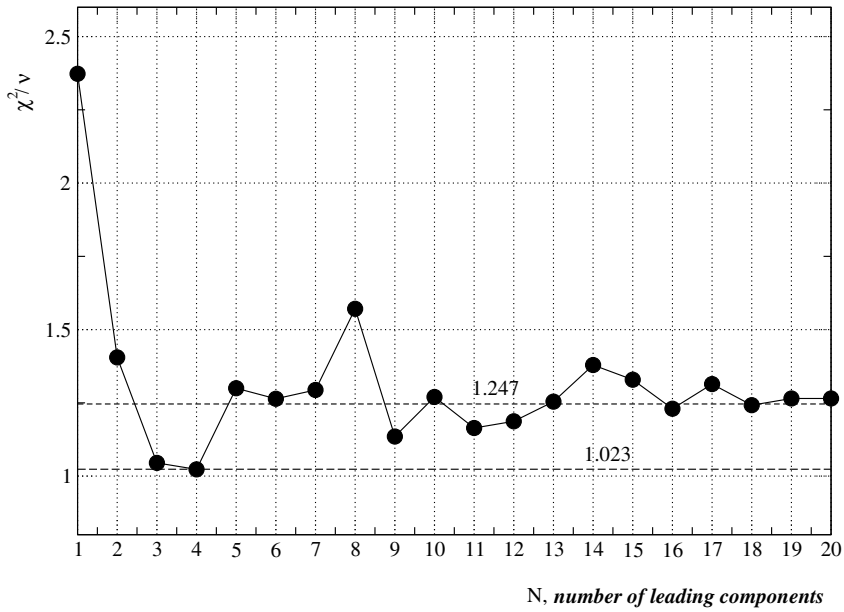

Figure 28: Dependence of χ^2/ν versus number of leading components for filtered data

Figure 29 shows the dependences of the ω_n^2 value versus the number of the residual components for the caterpillar length $C_L = 20$.

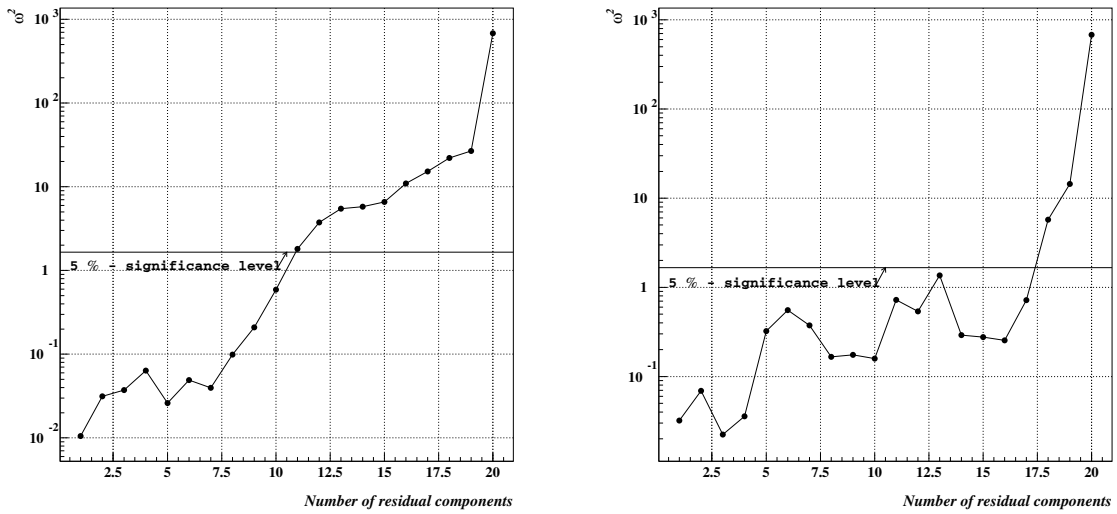

Figure 29: Dependences of ω_n^2 versus the number of residual components for the original and filtered traffic series for $C_L = 20$

7. The Kolmogorov's scheme and network traffic

The log-normal distribution has been first observed by Lucas et al. (1996) for distributions of packet arrivals aggregated at $100\,ms$. Similar inter-arrival time distributions have been observed in cellular telephony (1998).

The theoretical explanation of appearance of the log-normal distribution in natural phenomena was first given by Kolmogorov in 1941. A simplified explanation of Kolmogorov's scheme is the following. Suppose that we have a big rock which crumbles into sand. Then, it can be shown, that the number of grains at the k-th stage of fragmentation must be

$$N_k = \prod_{i=1}^{k} n_i = n_1 n_2 \cdots n_k, \quad \text{or} \quad \log N_k = \sum_{i=1}^{k} \log n_i. \tag{9}$$

The grain sizes S_k are inversely proportional to N_k. Applying the CLT, Kolmogorov found that logarithms of grain sizes are normally distributed, i.e. the distribution of grain sizes is log-normal.

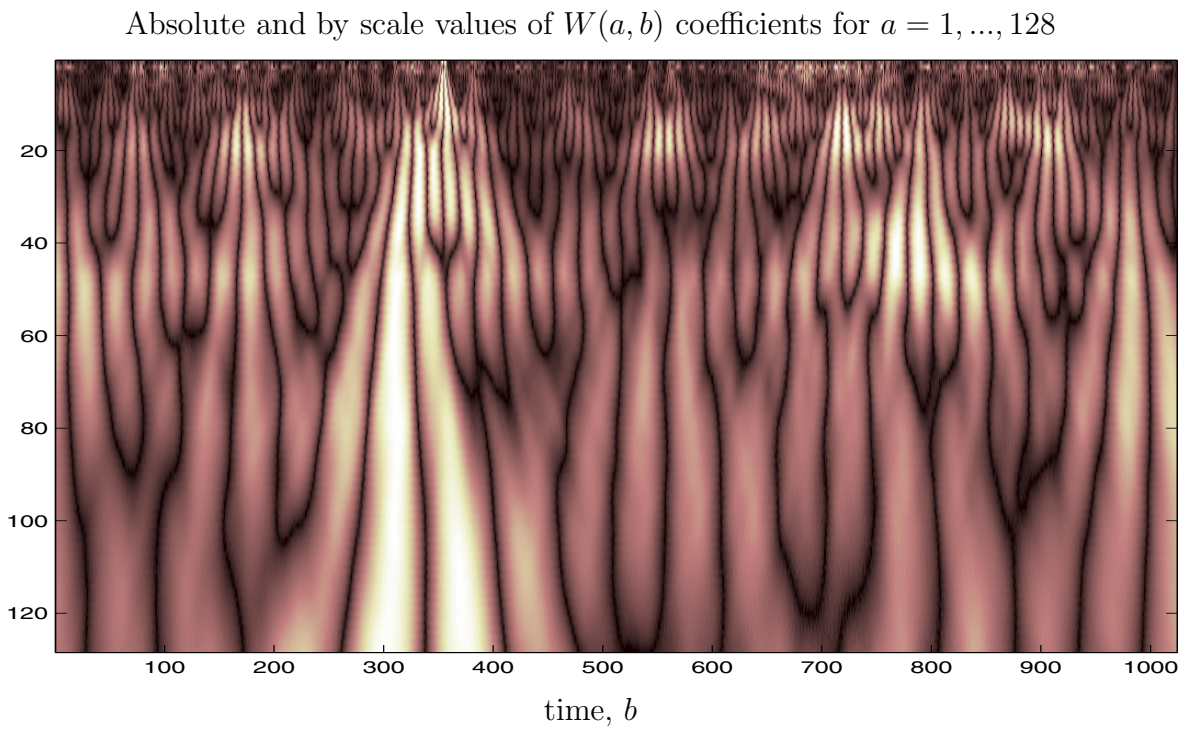

Figure 30: Shade plot of the CWT coefficients for traffic measurements

The tree-like fragmentation structure at different scales shown in Fig. 30 clearly proves the multiplicative character of network traffic. **This result is in agreement with formula (9) and confirms the applicability of**

Kolmogorov's scheme to the description of information traffic.

Conclusion

- ANN trained on traffic measurements reproduced the statistical distribution of real data, which well fits the log-normal form [5].
- Detailed analysis has shown that the reason of this distribution is a simple aggregation of real data [6].
- Applying the "Caterpillar"-SSA approach we demonstrated that few first components already form the fundamental part of network traffic [7].
- This result has been confirmed by application of wavelet filtering both to original traffic measurements and individual components of original and filtered data [8].
- The possibility to apply Kolmogorov's scheme to network traffic is demonstrated.

[**5**] 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.

[**6**] 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.

[**7**] I. Antoniou, V.V. Ivanov, Valery V. Ivanov and P.V. Zrelov: *Principal Component 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; "Particles & Nuclei, Letters" (in press).

[**8**] I. Antoniou, V.V. Ivanov, Valery V. Ivanov and P.V. Zrelov: *On a Statistical Model of Network Traffic*, VIII Int. Workshop on *"Advanced Computing and Analysis Techniques in Physics Research"* - **ACAT'2002**, 24-28 June, 2002, Moscow, RUSSIA, Book of abstracts, p. 177; "Nuclear Instruments & Methods in Physics Research", A502 (2003) 768-771.