E11-2002-223

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

WAVELET FILTERING OF NETWORK TRAFFIC MEASUREMENTS

[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 [3] we applied the Principal Components Analysis, especially the "Caterpillar"-SSA approach [1, 2], to the network traffic measurements. The 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 [3]. The first class includes leading components responsible for the formation of the basic part of network traffic, while the second class involves residual components that play a role of small irregular variations and can be interpreted as a stochastic noise. However, a more detailed analysis of the boundary region between these two classes may provide some additional information on traffic components and, thus, simplify the understanding of traffic dynamics.

In [5] we applied wavelet filtering to network data. The aim was to eliminate a high-frequency, noisy part and decrease the dimension of the dynamical system underlying the network series. The result of this procedure was rather promising [5].

We investigate here the influence of preliminary wavelet filtering both on the characteristics of individual principal components and on the sum distributions of leading and residual components. The aim of this study is to decrease the number of feature components responsible for the formation of the main part of the network traffic. This will be the case, if the wavelet filtering does not influence the main statistical and spectral characteristics of the original traffic series. This is in fact the case here.

In our study we use traffic measurements obtained at the input of the Dubna University [6] Local Area Network (LAN), which includes approximately 200-250 interconnected computers. We describe briefly in Section 1 the data acquisition system of this LAN realized on the basis of the standard IBM PC. In Section 2 we analyze the power spectrum of traffic measurements applying the Lomb periodogram technique. The peculiarities of wavelet filtering of traffic measurements are considered in Section 3. The statistical analysis of traffic components after filtering out of a high-frequency, noisy part is presented in Section 4. In Section 5 we discuss the influence of the filtering procedure on a number of feature components forming the main part of the 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 an open mode driver [7]: see Fig. 1.

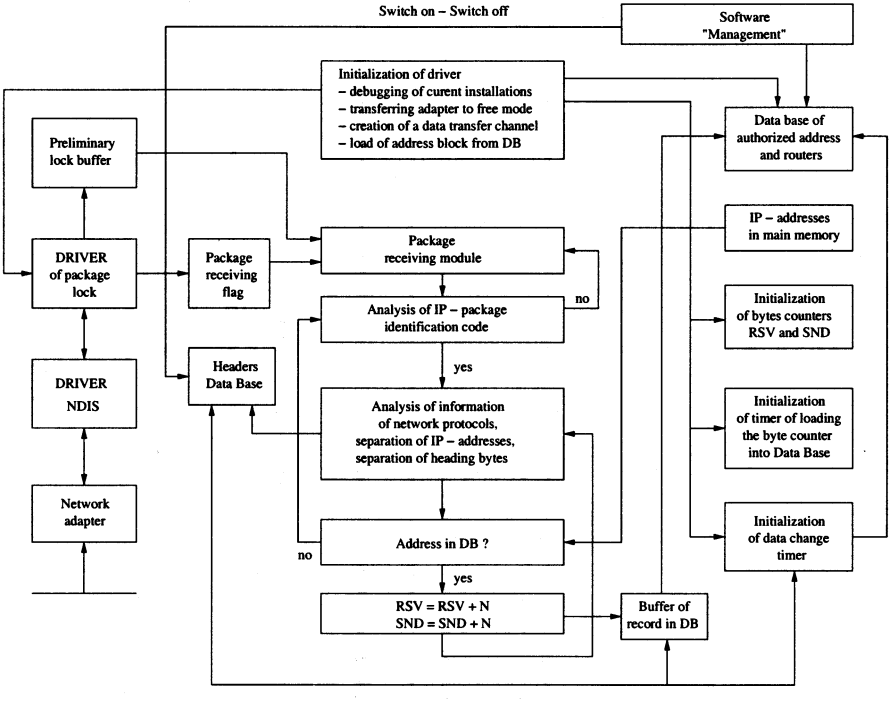

Figure 1: Scheme of a data acquisition system

In standard conditions the network adapter of a computer is in a mode of detecting a carrying signal (main harmonic 4 – 6 MHz). After appearing in the cable bits of the package preamble, the network adapter comes to a mode of 1 bit and 1 byte synchronization with the transmitter and starts receiving first bytes of the package heading. As soon as one succeeds in extracting the MAC-address of the shot receiver from the first bytes taken by the adapter, the network adapter compares it to its own. In the case of negative result of the comparison, the network adapter ceases to record the shot's bytes into its internal buffer and cleans its contents and then waits until the next package appears.

In order to provide conditions for receiving and analysis of all 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

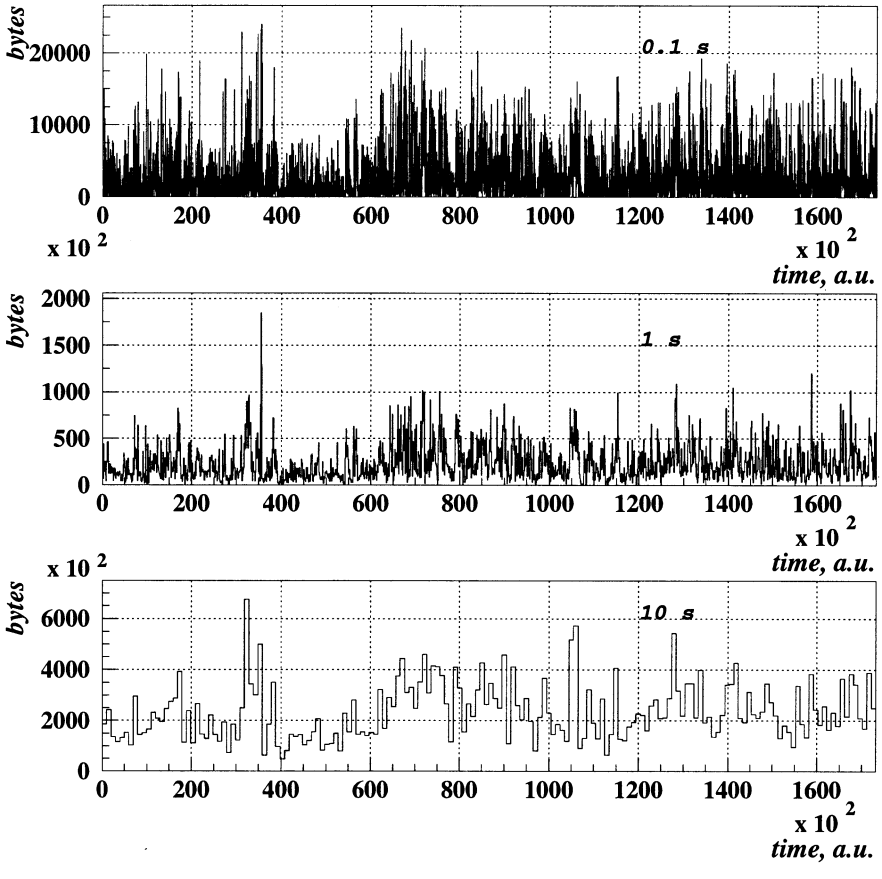

Figure 2: Traffic measurements aggregated with different bin sizes: 0.1 s, 1 s and 10 s

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 net-

work 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. Spectral analysis of traffic measurements

Figure 3 shows the daily part of traffic measurements aggregated with the bin size 1 s, which has been used in this study. The number K of points in this series was $K = 2048 = 2^{11}$, that corresponds approximately to 34 minutes of traffic measurements.

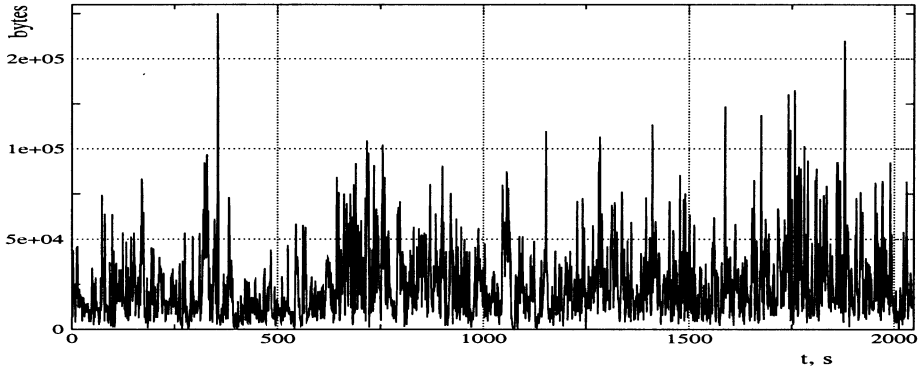

Figure 3: Traffic measurements aggregated with the bin size 1 s

This time series can be represented as

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

where Δt is the sampling interval ($\Delta t = 1$ in our case), whose reciprocal is the sampling rate. A sampled data set (1) contains *complete* information about all spectral components in a signal $y(t)$ up to the Nyquist critical frequency

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

and scrambled or *aliased* information about any signal components at frequencies larger than f_c (see, for example, [12]).

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, [12, 13] 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 (1) is defined by

$$P_K(\omega) = \frac{1}{2\pi^2}\left\{\frac{\left[\sum_{i=1}^{K}(y_i - \bar{y})\cos\omega(t_i - \tau)\right]^2}{\sum_{i=1}^{K}\cos^2\omega(t_i - \tau)} + \frac{\left[\sum_{i=1}^{K}(y_i - \bar{y})\sin\omega(t_i - \tau)\right]^2}{\sum_{i=1}^{K}\sin^2\omega(t_i - \tau)}\right\}, \tag{3}$$

where

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

and τ is defined by the relation

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

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 Gaussian random values. Scargle has shown [14] that, at any ω and when the null-hypothesis is valid, the probability that $P_K(\omega)$ will be between some positive z and $z+dz$, is $\exp(-z)dz$. This means that, if we scan M *independent* frequencies, the probability that none of those gives values larger than z is $(1 - e^{-z})^M$. Thus,

$$p(> z) = 1 - (1 - e^{-z})^M \tag{4}$$

is the false-alarm probability of the null-hypothesis, and it determines the *significance level* α of any peak in the $P_K(\omega)$ spectrum. A small value of $p(> z)$ indicates a highly significant periodic signal at z.

For estimation of the significance level α, we need to know M. Our interest is in the region where α assumes small values, $\alpha \ll 1$, so Eq. (4) can be written as

$$p(> z) \approx Me^{-z}. \tag{5}$$

Relation (5) 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 [15] 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 the frequency f_c.

Figure 4 shows the result of application of the Lomb method to the time series presented in Fig. 3: we used the code **period** from the *Numerical Recipes* library [12]. The figure plots $P_K(\omega)$ against $\omega = 2\pi f$ for the frequency interval starting from 0 up to the frequency 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.

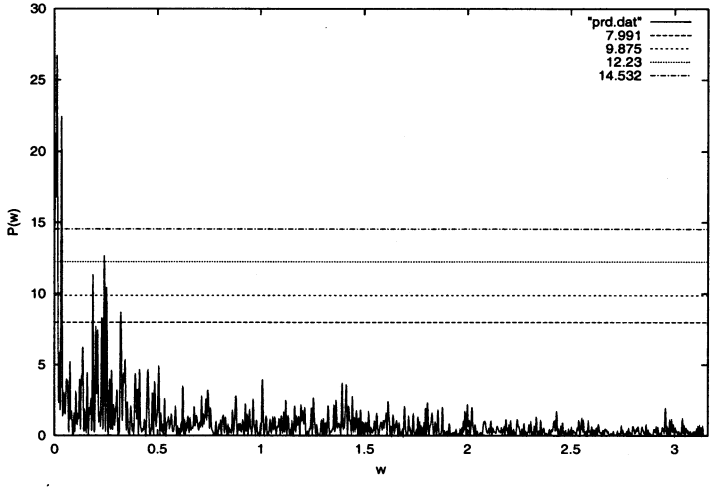

Figure 4: The dependence of $P_K(\omega)$ against the angular frequency $\omega = 2\pi f$ for traffic measurements presented in Fig. 3: $0 \leq \omega < 2\pi f_c$

One can see (Fig. 5) three highly significant peaks at low frequencies: 0.06, 0.012 and 0.034. There are also three other peaks at frequencies 0.186, 0.241 and 0.252, which exceed the 50 % significance level.

For higher frequencies ($\omega > 0.35$), together with the frequency increase, the amplitude of peaks is very quickly decreasing (Fig. 4) and it does not exceed the value 5, which corresponds approximately to the significance level $\alpha \approx 1$. This means that traffic components contributing into this high frequency region can be interpreted as a stochastic Gaussian noise: see further analysis below.

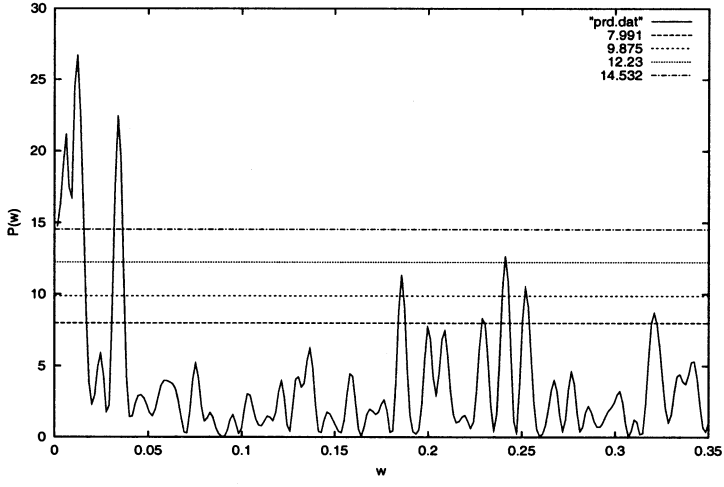

Figure 5: The dependence of $P_K(\omega)$ against the angular frequency $\omega = 2\pi f$ for traffic measurements presented in Fig. 3: $0 \leq \omega < 0.35$

3. Wavelet filtering of traffic measurements

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"*. It has been shown in [5] that the elimination of the noisy part reduces the dimension of the underlying dynamical process and, thus, simplifies the analysis of traffic series.

Wavelet analysis is very suitable for handling 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. This property permits to study localized features of time series, while, for example, Fourier transform provides only general information on the analyzed series in the frequency (scale) domain.

Here we present the main scheme of the wavelet analysis. The details can be found, for instance, in [8]–[11].

The discrete wavelet transform (DWT) of the function $f(t) \in L_2(R)$, given in the form of one-dimensional time series (1), can be represented by the following expansion

$$f(t) = \sum_{j,k \in Z} d_{jk} \psi(2^j t - k). \tag{6}$$

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. (6) can be rewritten in a more

convenient form

$$f(t) = \sum_k s_k^J \phi(2^J t - k) + \sum_{j \geq J} \sum_{k \in Z} d_k^j \psi(2^j t - k), \tag{7}$$

where $\phi(t)$ is the scaling function corresponding to the chosen wavelet function $\psi(t)$ (see, for example, [8]). In (7) the first term describes a smooth part of series (7) restricted by level J, and the second term is related to details, or a high-frequency part of the analyzed series.

The coefficients s_k^j and d_k^j are usually determined with the help of the pyramidal scheme [16] of the fast wavelet transform (see, for instance, [12]) applying the following equations:

$$s_k^{j+1} = \sum_m h_m s_{2k+m}^j, \qquad d_k^{j+1} = \sum_m g_m s_{2k+m}^j, \tag{8}$$

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

We use here the discrete Daubechies wavelets [8, 9], because they provide high quality representation of both high- and low-frequency components of the analyzed signal [12].

Wavelet filtering implies rejection or modification of a 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, [17, 18]). However, the most widespread is the hard thresholding algorithm (see, for example, [12]). In this scheme all coefficients with absolute values less than λ have to be rejected (set to zero).

In all these methods the filtering procedure affects all coefficients, without taking into account their specific resolution level J. Therefore, such a procedure may eliminate both the coefficients $\{d_k^j\}$ which correspond to the high-frequency part of (7) and the coefficients $\{s_k^J\}$ related to the low-frequency part.

In this connection, it is impossible to apply the existing algorithms in our case, because the filtering will affect not only the high-frequency, *noisy* part, but also the 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 runs as follows. Suppose, K is the number of elements in the analyzed series, M is the number of wavelets coefficients that must be rejected and let $M < \frac{K}{2}$. Then, we reject the M smallest of $\frac{K}{2}$ "detailed" coefficients of series (7). If $\frac{K}{2} < M < \frac{3K}{4}$, we eliminate all $\frac{K}{2}$ "detailed" coefficients together with the $M - \frac{K}{2}$ 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 a 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 [19] has been used for the 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.

The ω_n^2 symmetry criterion tests the symmetry against $y = 0$ of the distribution function $F(y)$ of observables $y_1, \ldots, y_n$, i.e. the null-hypothesis H_0: $F(y) = 1 - F(y)$. The corresponding ω_n^2 statistics has the following form:

$$\omega_n^2 = n \int_{-\infty}^{\infty} [F_n(y) + F_n(-y) - 1]^2 \, dF_n(y), \tag{9}$$

where $F_n(y)$ is the empirical distribution function. It is more convenient to calculate the values of statistics (9) using the following formula

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

where $\tilde{y_1} \leq \ldots \leq \tilde{y_n}$ is the variational series constructed on the basis of observables.

Figure 6 shows the dependence of ω_n^2 values versus the number of rejected wavelet coefficients. This dependence has the minimum at $M = 768$. The corresponding packet size distribution (Fig. 7) passes the χ^2-test on the correspondence to the log-normal distribution (see details in Section 4) with the significance level $\alpha = 18\%$. One can also see from Fig. 6 that a possible maximal number of coefficients, that can be eliminated without exceeding the 5%-significance level, is $M = 1408$ (this amounts approximately to 70 % of the whole number K of coefficients).

The autocorrelation function [20]

$$C(\tau) = \frac{\sum_{i=1}^{K} (y_{i+\tau} - \bar{y})(y_i - \bar{y})}{\sum_{i=1}^{K} (y_i - \bar{y})^2}, \qquad \bar{y} = \frac{1}{K} \sum_{i=1}^{K} y_i \tag{10}$$

can be also used as a criterion for the evaluation of the noisy part. The time series corresponding to the noisy part must be uncorrelated. Figure 8 (left plot) presents the dependence of the auto-correlation function for the noisy part corresponding to different number of rejected coefficients M. This plot shows that up to $M = 1408$, the rejected part can be considered as noisy.

In order to monitor the influence of rejection of the noisy part on the main part of the traffic series (from the nonlinear analysis point of view), we also controlled the behavior of the autocorrelation function for the smooth part of the series (7), corresponding to a different number of rejected coefficients: see Fig. 8 (right plot). One can clearly see that the rejection of smallest coefficients up to $M = 1408$ did not influence seriously on the form of the autocorrelation function.

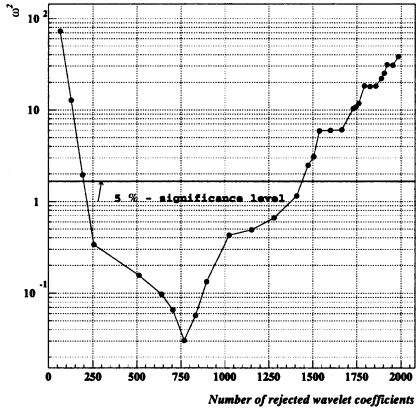

Figure 6: The dependence of ω_n^2 values versus the number of rejected wavelet coefficients

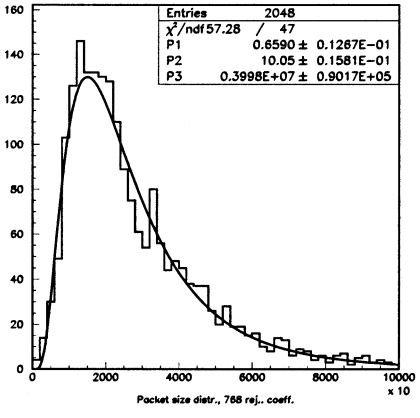

Figure 7: Packet size distribution for series corresponding to $M = 768$ rejected coefficients

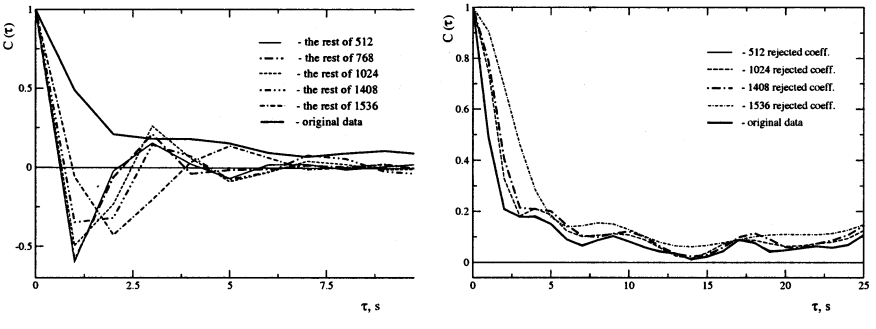

Figure 8: Autocorrelation functions $C(\tau)$ of noisy (left plot) and smooth (right plot) parts corresponding to different number of rejected coefficients

Based on estimations of these two criteria, we came to the conclusion that it is reasonable to assume that $M = 1408$. Figure 9 presents the original traffic series, the filtered signal and the noisy part related to $M = 1408$ of rejected coefficients.

It is also interesting to check, how the filtering procedure influences the spectral characteristics of the analyzed series. Figure 10 shows the dependence of $P_K(\omega)$ against the angular frequency ω for filtered signal (continuous curve) and original

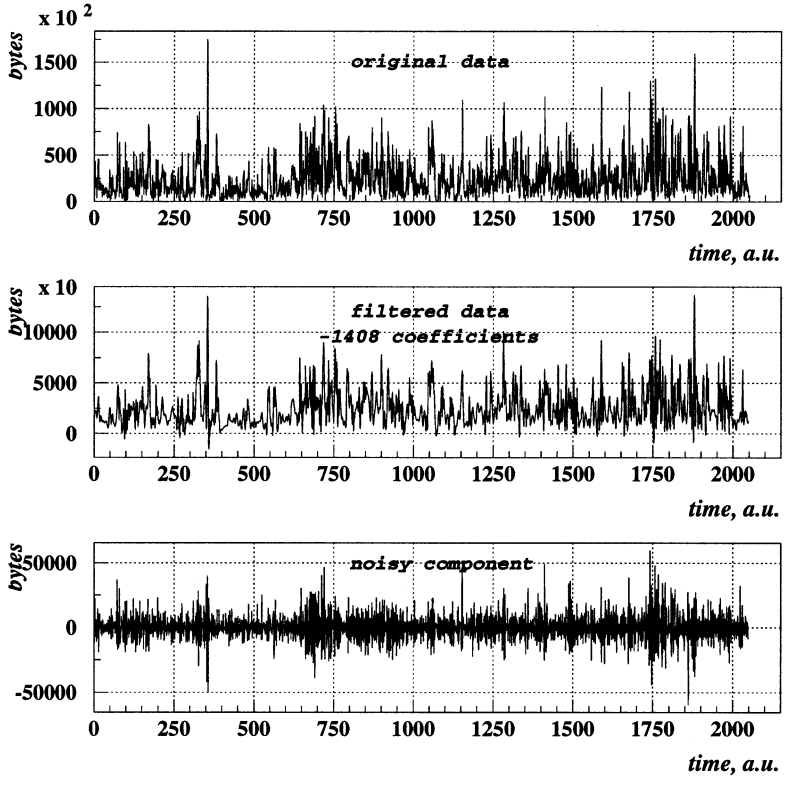

Figure 9: Traffic measurements: 1) original traffic series, 2) filtered signal, 3) noisy part

(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, all high frequencies starting approximately from $\omega = 1.1$ have been significantly suppressed (see also Section 5).

4. Analysis of statistical characteristics of filtered series

In [3] we applied the Principal Components Analysis, especially the "Caterpillar"-SSA approach [1, 2], to the network traffic measurements. This approach permits to obtain the contribution (in decreasing order) of individual components into the analyzed series. Figure 11 shows the corresponding dependence for two cases of the caterpillar length [3]: $C_L = 12$ (left) and 20 (right). This information permits to estimate the number of principal components which effectively contribute to the traffic series.

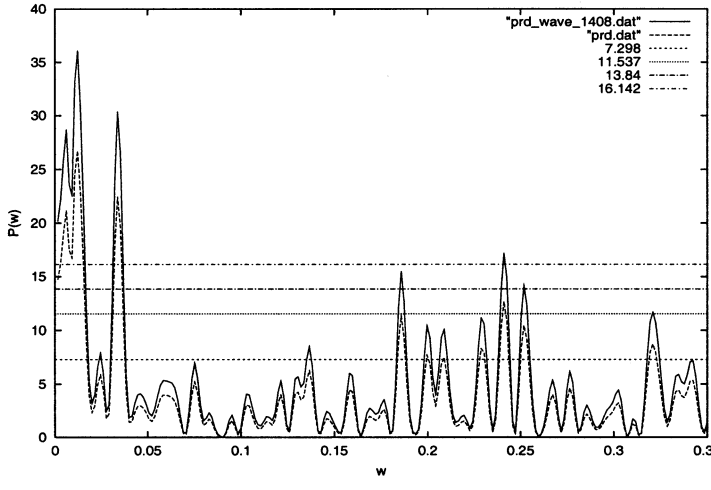

Figure 10: The dependence of $P_K(\omega)$ against the angular frequency $\omega = 2\pi f$ for filtered signal (continuous curve) and for original traffic measurements (dashed curve): $0 \leq \omega < 0.35$

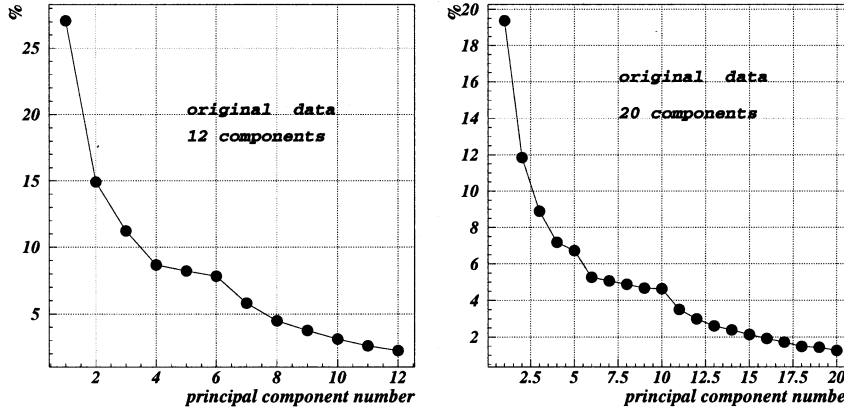

Figure 11: Contributions of eigenvalues in percentages for the original traffic data. The results are presented for two cases of the caterpillar length: $C_L = 12$ (left) and 20 (right)

In Fig. 12 we present similar dependences for traffic data after filtering out the high-frequency part corresponding to $M = 1408$ smallest coefficients. One

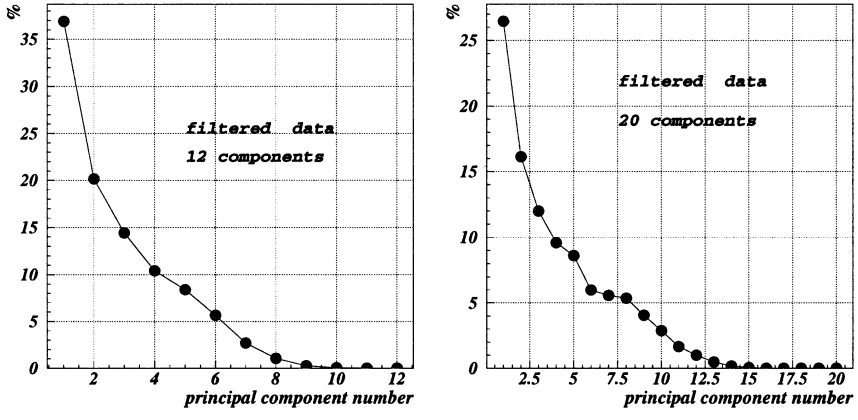

Figure 12: Contributions of eigenvalues in percentages for the traffic data after filtering out the high-frequency part. The results are presented for two cases of the caterpillar length: $C_L = 12$ (left) and 20 (right)

can clearly see that the contribution of residual components noticeably decreased compared to the original traffic data (Fig. 11). At the same time the contribution of leading components significantly increased.

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

We have demonstrated [4] that the aggregation of packet sizes of traffic measurements forms the log-normal distribution. Later, applying the Principal Component Analysis of traffic series [3], we found that just a few first components form the main part of information traffic. The residual components play the role of small irregular variations, which do not fit in the basic component of the network traffic and can be eliminated as stochastic noise.

Taking into account the results of [3] and [4], it is important to analyze how the filtering procedure influences on statistical characteristics of traffic series, namely,

1. does it disturb significantly the packet size distributions, corresponding to leading components, and

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

4.1. PCA of filtered series: analysis of leading components

In order to check the influence of the wavelet filtering on packet size distributions, corresponding to leading components, we applied the same procedure as in [3], i.e. we tested the correspondence of these distributions to the log-normal function [21]:

$$f(x) = \frac{A}{\sqrt{2\pi}\sigma}\frac{1}{x}\exp\left[-\frac{1}{2\sigma^2}(\ln x - \mu)^2\right] \tag{11}$$

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

The fitting procedure has been realized with the help of the MINUIT package [22] in the frame of well-known PAW (Physical Analysis Workstation, see details in [23]). The MINUIT package is conceived as a tool to find the minimum value of a multi-parameter function and to analyze the shape of the function around the minimum [22].

Figure 13 demonstrates the results of fitting of packet size distributions, for the filtered traffic series, corresponding to the sum of a different number N of leading components (results presented here are for the caterpillar length $C_L = 20$, see details in [3]), by the function (11). Here χ^2 is the calculated value of χ^2 and ν is the number of degrees of freedom. Two lines parallel to the abscissa axes show the significance levels (or the probability that the observed chi-square will exceed the value χ^2 by chance *even* for a correct model: see, for instance, [12, 21]) $\alpha = 10\%$ (the top line, $\chi^2/\nu = 1.247$, $\nu = 47$) and $\alpha = 42.9\%$ (the bottom line, $\chi^2/\nu = 1.023$, $\nu = 47$) corresponding to the χ^2-test.

This dependence confirms our previous result, obtained in [3], 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 fits the null-hypothesis (11) with a quite high correspondence level ($\alpha = 39.2\%$): see also Fig. 14.

The dependence of χ^2/ν versus the number N of leading components in Fig. 13 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 near the corridor corresponding to the admissible region for the χ^2-test.

Figure 15 shows the series reconstructed on the basis of the first, second and third leading component, correspondingly, after the subtraction of the caterpillar average value.

These series are very much similar to seria corresponding to the original traffic data (see Fig. 8 in [3]). However, filtered series are visually more smooth compared to original data. Their summary contribution into the analyzed time series is noticeably higher ($\sim 54\,\%$) compared to the original data ($\sim 40\,\%$): see Figs. 11 and 12 for $C_L = 20$.

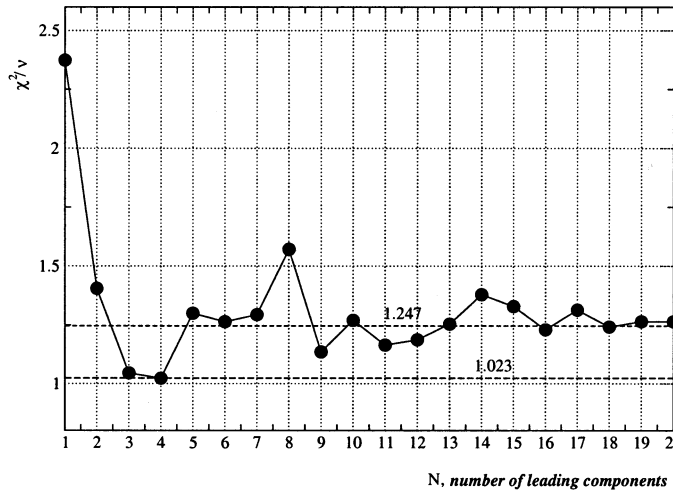

Figure 13: The dependence of χ^2/ν versus the number N of leading components

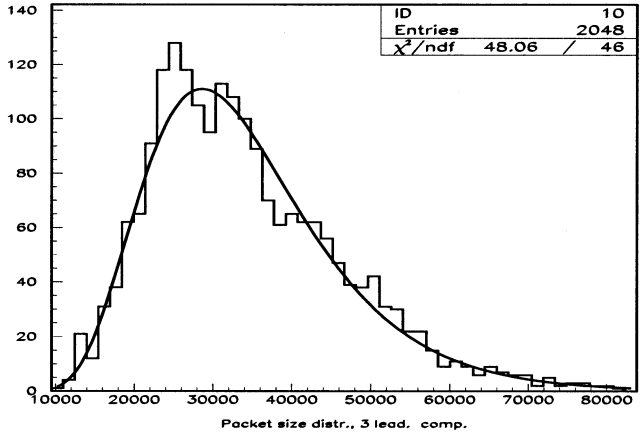

Figure 14: Fitting the distribution corresponding to three leading components by the log-normal function (11)

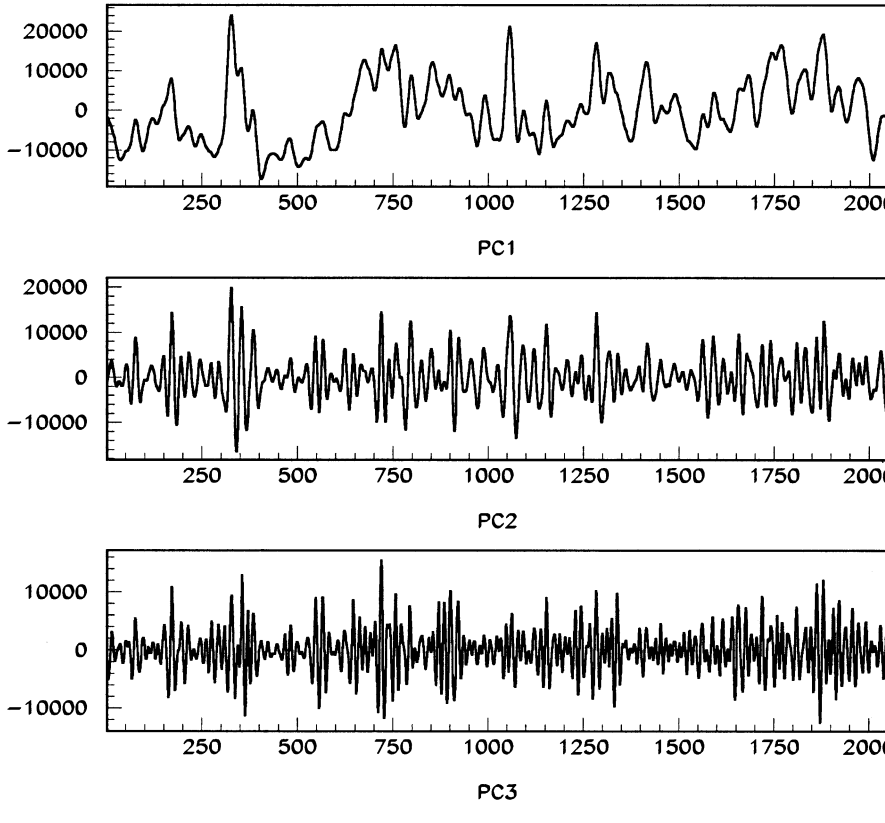

Figure 15: Time seria corresponding to three leading components (after the subtraction the caterpillar average value): the trend component and two first periodic components

4.2. PCA of filtered series: analysis of residual components

Figure 16 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 (see Fig. 11 in [3]).

Figure 17 presents the statistical distribution corresponding to the series in Fig 16. It quite well follows the Gaussian distribution (similar to Fig. 12 in [3]).

At the same time, the amplitude dispersion of the above series (Fig. 16) and the standard deviation of its statistical distribution (Fig. 17) are significantly less compared to the original data: see Figs. 11 and 12 in [3].

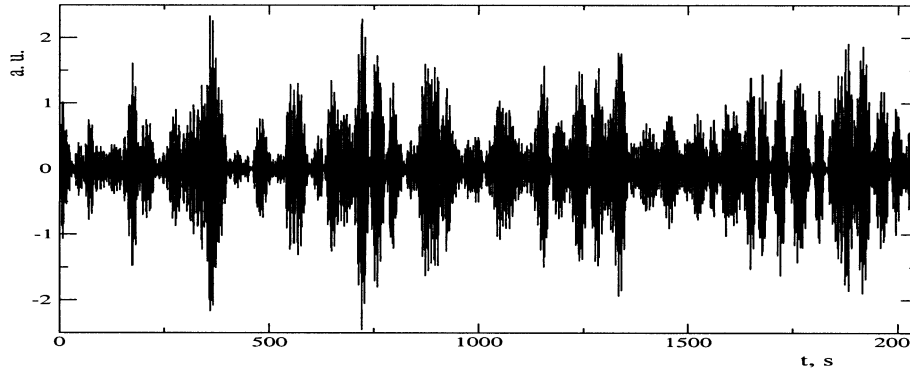

Figure 16: Traffic series reconstructed by the caterpillar method ($C_L = 20$) on the basis of the smallest component

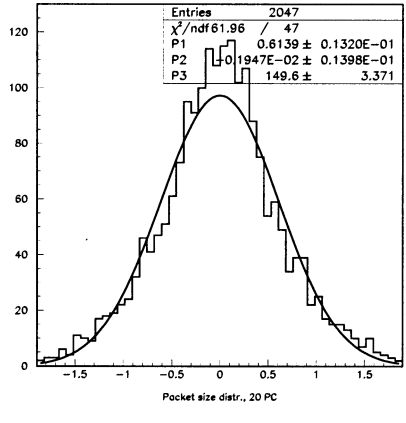

Figure 17: Statistical distribution of the time series presented in Fig. 16; the fitting curve corresponds to the Gaussian distribution

5. Selection of feature components

Together with the increase of a number of residual components, their sum distribution starts gradually to lose the symmetric character. 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 [19, 3].

Figure 18 shows the dependences of the ω_n^2 value versus the number of residual components for original (left figure) and for filtered (right figure) traffic seria for the caterpillar length $C_L = 20$. The horizontal line corresponds to the 5 %-significance level.

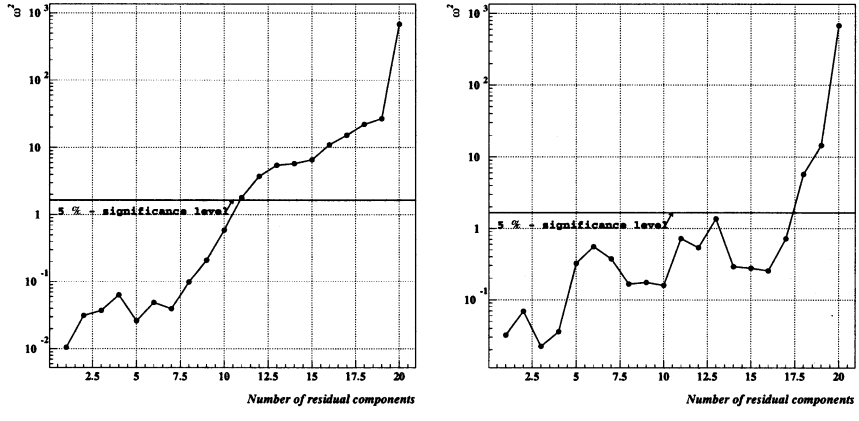

Figure 18: The dependences of the ω_n^2 values versus the number n of the residual components for original (left figure) and for filtered (right figure) traffic seria and for the caterpillar length $C_L = 20$

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 the original traffic measurements and 17 for the filtered series. This result demonstrates that after the wavelet filtering the 17 smallest components can be considered as noisy and can be eliminated from the whole set of principal components. This result is in the agreement with the result obtained in Section 4 when the χ^2-test was used: see Fig. 13.

Figure 19 shows the dependence of $P_K(\omega)$ against ω for three leading components (continuous curve) and for all components of the filtered signal (dashed curve). This dependence clearly demonstrates that a low frequency region of traffic series is formed by the three leading components. At the same time, for the sum of the three leading components the contribution of frequencies higher than $\omega > 0.35$ is significantly suppressed: see Fig. 20.

Conclusion

Applying the "Caterpillar"-SSA analysis, wavelet filtering and statistical χ^2 and ω_n^2 tests, we demonstrated that the main part of network traffic can be described by a minimal number of feature components: three leading components for $C_L = 20$. We

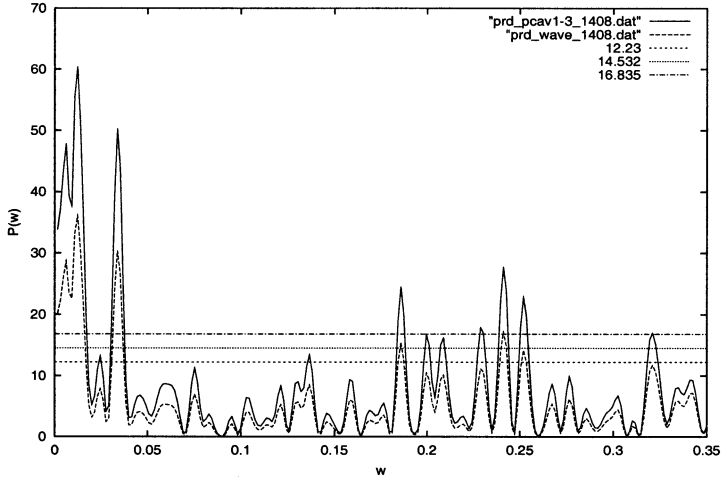

Figure 19: The dependence of $P_K(\omega)$ against the angular frequency ω for 3 first leading components (continuous curve) and for all components of the filtered signal (dashed curve): $0 \leq \omega < 0.35$

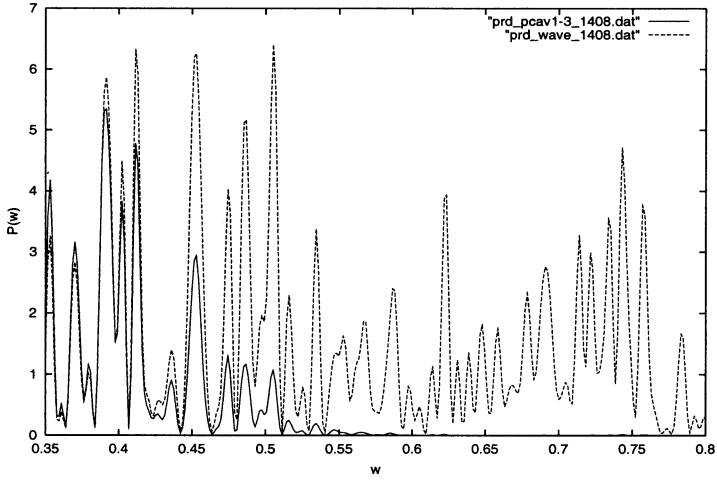

Figure 20: The dependence of $P_K(\omega)$ against the angular frequency ω for 3 first leading components (continuous curve) and for all components of the filtered signal (dashed curve): $0.35 \leq \omega < 0.8$

also show that the series reconstructed on the basis of these components preserves the main spectral characteristics of the original traffic measurements. This suggests that all transformations performed under the original traffic series did not destroy the dynamical characteristics of the traffic.

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

Acknowledgments

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

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

[3] 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).

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

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

[7] 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.

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

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

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

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

[12] 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.

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

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

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

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

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

[18] 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.

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

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

[21] 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.

[22] F. James and M. Roos: *MINUIT – Function Minimization and Error Analysis*, CERN Program Library D506, 1988.

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

Received on November 5, 2002.

Антониоу Я. и др. E11-2002-223
Вейвлет-фильтрация измерений информационного трафика

Анализ измерений информационного трафика на основе подхода «Caterpillar»-SSA и совместного применения статистических критериев χ^2 и ω^2 позволил разбить набор принципиальных компонентов на два класса. Первый класс включает лидирующие компоненты, ответственные за формирование основного вклада сетевого трафика, второй содержит остаточные компоненты, которые можно интерпретировать как шум. Детальный анализ промежуточной области между этими классами может дать дополнительную информацию о компонентах трафика и, таким образом, упростить понимание его динамики. В этой связи мы применили вейвлет-фильтрацию к измерениям трафика и исследовали ее влияние как на отдельные принципиальные компоненты, так и на суммарные распределения лидирующих и остаточных компонентов.

Работа выполнена в Лаборатории информационных технологий ОИЯИ.

Сообщение Объединенного института ядерных исследований. Дубна, 2002

Antoniou I. et al. E11-2002-223
Wavelet Filtering of Network Traffic Measurements

The «Caterpillar»-SSA and statistical analysis of network traffic measurements 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. The first class includes leading components responsible for the main contribution to network traffic, and the second class involves residual components that can be interpreted as noise. More detailed analysis of the boundary region between these two classes may give additional information on traffic components and, thus, simplify the understanding of traffic dynamics. In this connection, we apply wavelet filtering to traffic measurements, and analyze its influence both on the characteristics of individual principal components and on the sum distributions of leading and residual components.

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. 1,5. Уч.-изд. л. 2,34. Тираж 310 экз. Заказ № 53617.

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/