Approximation of Internet Traffic in Wavelet Domain

Petar Ćisar* and Sanja Maravić Ćisar**
*Academy of Criminalistic and Police Studies, Belgrade-Zemun, Serbia
**Subotica Tech, Department of Informatics, Subotica, Serbia
petar.cisar@kpa.edu.rs, sanjam@vts.su.ac.rs

Abstract—The paper elaborates the complexity and importance of modeling Internet traffic. In this sense, having in mind the nature of Internet traffic as well as actual information in this sphere, it was emphasized that there is no universally applicable model. This standpoint has initiated the idea of examining the possibility of approximation internet traffic. Because of the rather suitable property, namely that is not necessary to provide stationarity of the input signal, traffic analysis in this paper was performed in wavelet domain, using different types of wavelets, with the aim of determining the convenience of certain wavelets for the analysis of Internet traffic. It was concluded that all examined wavelets have generated approximately similar statistical results, which depend on the selected level of approximation and analyzed traffic curve. For different curves at the same level of approximation, statistical values have been obtained that do not differ significantly from the statistics of the corresponding original signals. It was also found that the discrete approximation of the symlet wavelet sym(2) provided the largest number of statistically accurate results, and in this regard proved to be the most suitable for the analysis of Internet traffic.

I. INTRODUCTION

One of the vital fields of research regarding the context of networking involves traffic model development, with the aim to be applied in various communication networks, and specifically to the Internet. These types of models have the following two advantages. On the one hand, there is need for traffic models as input in network simulations. It is then necessary to perform these simulations so as to be able to study and validate algorithms and protocols to be applied to real traffic, and to analyze the way traffic will react to specific network conditions. Therefore it can be stated that it is vital that the assumed models will significantly mirror the relevant characteristics of the traffic which they are supposed to represent. On the other hand, better understanding might be attained by a good traffic model regarding the characteristics of the network traffic itself. This may be beneficial when designing routers and devices which handle network traffic. As an example one can examine well-validated a model, as it presents some correlation between traffic arrivals, in turn, this information may be implemented in the conception of ad hoc packet handling strategies.

Service providers have a need for accurate traffic models in order to sustain high-level service quality. Many traffic models have been created on the basis of traffic measurement data, thus the need arises for determining precise methods that will identify quick and robust traffic models, which enable the accurate determination of all the essential characteristics of the measured data traces.

This work is written with the goal to analyze the possibility of approximating the internet traffic in the wavelet domain and to evaluate the attained results’ quality which was gained by implementing statistical parameters. Different types of wavelets in Matlab environment were used in the comparative analysis.

There are six sections in this work. The need of modeling Internet traffic and general structural description is given in the introduction. Section 2 will provide an insight into current traffic models and their major characteristics. The following section lays out some of the fundamental concepts of modeling traffic. Section 4 gives an overview of the key properties of wavelet analysis and underlines its benefits. Section 5 presents the practical work on wavelet analysis and discusses the results. The work finally closes with comments and concluding remarks.

II. INTERNET TRAFFIC MODELS

The first traffic model was created based on Poisson processes, within the context of telephony, with call arrivals seen as independent and identically distributed and “holding times” having exponential distribution. While the early implementations were successful and analytically straightforward, it was proven that the Poisson model could not determine data traffic in modern LANs and WANs, where batch arrivals, event correlations and traffic burstiness remain vital aspects to be considered. The application of heavy tailed distributions and of self-similarity gained prominence. It has been pointed out that Internet traffic was long-tailed and frequently modeled by lognormal, Weibull or Pareto distributions theoretically. Nevertheless, these models were challenging to implement directly in traffic analysis and performance evaluation studies because of their complex representations and theoretical characteristics.

Still, data networks are characterized by high or extreme variability [10], [17], [12], [11], [15]. It is statistically, possible to determine temporal high variability by long-range dependences, which means, autocorrelation exhibiting power-law decay. However,
III. TRAFFIC MODELING

One may model Internet traffic as a sequence of arrivals of discrete entities, for example, packets, cells, etc. Mathematically, the usage of two equivalent representations is possible: counting processes and interarrival time processes. A counting process \( \{N(t)\}_{t=0}^{\infty} \) is a continuous-time, integer-valued stochastic process, where \( N(t) \) expresses the number of arrivals in the time interval \((0,t]\). An interarrival time process is a non-negative random sequence \( \{A_n\} \), where \( A_n = T_{n+1} - T_n \) refers to the length of the interval separating arrivals \( n \)-1 and \( n \). The relationship between these two types of processes is represented by the equation given below [2]:

\[
\{N(t) = n\} = \{T_0 \leq t < T_{n+1}\} = \{\sum_{k=0}^{n} A_k \leq t < \sum_{k=0}^{n+1} A_k\} \quad (1)
\]

Regarding compound traffic, arrivals can be either in batches, i.e., a number of arrivals that occur simultaneously \( T_n \). The modeling of this fact may happen by using an additional non-negative random sequence \( \{B_n\}_{n=1}^{\infty} \), where \( B_n \) is the cardinality of the \( n \)-th batch. It is the nature of the chosen stochastic processes \( \{N(t)\} \) and \( \{A_n\} \) which primarily defines the traffic models, to be analyzed in the subsequent sections. One of the determining aspects in the selection of the stochastic process is to what degree it is able to describe traffic burstiness. Specifically, a sequence of arrival times is seen as bursty if the \( T_n \) tend to form clusters, that is, if the corresponding \( \{A_n\} \) sees a mix of relatively long and short interarrival times. In terms of mathematics, traffic burstiness is in connection with short-term autocorrelations between the interarrival times. Still, there is no consensus on the definition of the notion of burstiness [7]; conversely, there are various measures implemented, a number of which do not take into consideration the second order properties of the traffic. A first measure is the ratio of peak rate to mean rate, whose disadvantage is being dependent upon the interval used to measure the rate. A second measure is presented by the coefficient of variation \( c_{\text{VAR}} = \sigma[A_n]/E[A_n] \) of the interarrival times [2]. A metric considering second order properties of the traffic is the index of dispersion for counts (IDC). This is especially the case in ones considers an interval of time \( \tau \), \( \text{IDC}(\tau) = \text{Var}[N(\tau)]/E[N(\tau)] \). Due to the relationship (1), IDC includes in the numerator the effects of the autocorrelation between the \( A_n \).

Bearing in mind how complex Internet traffic modeling is, logically it involves the possibility of its approximation. Approximation can be defined as the quality or state of being close in value to, yet not being the same as the desired value. In the field of science, approximation may be a reference to the use of a simpler process or model when using the correct model poses a difficulty. One of the implementations of an approximate model making calculations easier. Approximations might also be implemented when incomplete information does not allow for using exact representations. In the field of mathematics, approximation theory refers to the best approximation of functions with simpler functions, and to the quantitative characterization of the errors introduced this way. The main goal of a general approximation is to represent non-arithmetic quantities by arithmetic quantities in order for the accuracy to be defined to a desired degree. Wavelet analysis presents one possible method for the mathematical approximation of network traffic.

IV. WAVELET ANALYSIS

Wavelets are mathematical functions that decompose data into frequency components and analyze each of the components with a scale-matched resolution [8]. They are localized functions, continuous in time, which drop to zero instead of long-term oscillation [13]. Since wavelets are able to decorrelate autocorrelated data, the wavelet decomposition method provides a practical way of representing and monitoring data over traditional monitoring techniques [6]. Wavelet methods are highly useful in practice because data from most of the processes are multi-scale in nature due to events occurring at different locations and with different localization in time and frequency. They represent stochastic processes whose energy or power spectrum changes with time and/or frequency. The most important reason for choosing wavelet analysis instead of other techniques, is that for wavelets, the stationarity of the observed signal does not need to be provided, which forms the basis of classical methods (for example, autoregressive moving average). Since the data is frequently analyzed, it becomes appropriate for the purpose of fast detection. However, data that is analyzed frequently, are mostly too noisy and requires special treatment. Basically, wavelet transform becomes more suitable than other methods (for instance, discrete Fourier transform, discrete sine or cosine transform, Hartley transform) for several reasons [5]:

1. Wavelet analysis allows the analysis of a series of data while simultaneously preserving temporal and spatial information at the same time. Other key methods preserve either temporal or spatial information.
2. Wavelet analysis is more flexible in its monitoring of frequently changing data.
3. The majority of software provides a user-friendly environment in order to aid the use of wavelet analysis.

Wavelet transform of a function is the improved version of Fourier transform. Fourier transform is a powerful tool for analyzing the components of a stationary signal. But it is failed for analyzing the non-stationary signal where as wavelet transform allows the components of a non-stationary signal to be analyzed. By using Fourier transform, we lose the time information. Fourier transform cannot locate drift, trends, abrupt changes, beginning and ends of events etc. Also, it uses complex numbers in calculations.
Fourier methods are not always appropriate tools to recapture the signal, particularly if it is highly non-smooth. Too much Fourier information is needed to reconstruct the signal locally. In these cases the wavelet analysis is often very effective because it provide a simple approach for dealing with the local aspects of a signal.

Wavelet domain for the analysis of network traffic is used by several authors, including [16], [9], [3] and [18].

V. ONE-DIMENSIONAL INTERNET TRAFFIC ANALYSIS USING WAVELET TOOLBOX

Internet traffic data used in this paper derive from ISPs. The first analyzed signal is the aggregated traffic in the United Kingdom academic network backbone [4]. It was collected regularly in hourly intervals between 19 November 2004 and 27 January 2005 (1657 values). The subject of examination was one-dimensional (1-D) traffic analysis in wavelet domain. This discrete wavelet analysis consists of several phases such as decomposition of a signal, constructing approximations, regenerating a signal by using the inverse wavelet transform and reconstructing the original signal. For instance, performing a single-level wavelet decomposition of a signal in Matlab generates the coefficients of the level 1 approximation (cA1) and detail (cD1). Using these coefficients, we construct the level 1 approximation and detail.

Using the possibility of the Matlab Wavelet Toolbox for easy selection of the type of wavelet and the level of approximation, the statistical categories were calculated in the form as presented in Figure 1.

![Figure 1: Approximation at level 1 (wavelet: haar) and corresponding statistics](image)

The wavelet family shortened names that were used are listed in the following table [19]:

<table>
<thead>
<tr>
<th>Wavelet Family</th>
<th>Short Name Wavelet Family Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;haar&quot;</td>
<td>Haar wavelet</td>
</tr>
<tr>
<td>&quot;db&quot;</td>
<td>Daubechies wavelets</td>
</tr>
<tr>
<td>&quot;sym&quot;</td>
<td>Symlets</td>
</tr>
<tr>
<td>&quot;coif&quot;</td>
<td>Coiflets</td>
</tr>
<tr>
<td>&quot;bior&quot;</td>
<td>Biorthogonal wavelets</td>
</tr>
<tr>
<td>&quot;rbio&quot;</td>
<td>Reverse biorthogonal wavelets</td>
</tr>
<tr>
<td>&quot;dmey&quot;</td>
<td>Discrete approximation of Meyer wavelet</td>
</tr>
</tbody>
</table>

The first recorded mention of what we now call a “wavelet” seems to be in 1909, in a thesis by Alfred Haar [19]. Haar wavelet is discontinuous, and resembles a step function as presented in Figure 2.

![Figure 2: Haar wavelet](image)

Ingrid Daubechies, invented what are called compactly supported orthonormal wavelets — thus making discrete wavelet analysis practicable. The names of the Daubechies family wavelets are written dbN, where N is...
the order, and \( db \) the “surname” of the wavelet. The \( db1 \) wavelet is the same as Haar wavelet [19].

![Figure 3: Daubechies family wavelet [19]](image)

The symlets are nearly symmetrical wavelets proposed by Daubechies as modifications to the \( db \) family. The properties of the two wavelet families are similar [19].

![Figure 4: Symlets [19]](image)

Coiflets, the wavelet function has \( 2N \) moments equal to 0 and the scaling function has \( 2N-1 \) moments equal to 0. The two functions have a support of length \( 6N-1 \) [19].

![Figure 5: Coiflets wavelet [19]](image)

Biorthogonal wavelets, this family of wavelets exhibits the property of linear phase, which is needed for signal and image reconstruction. By using two wavelets, one for decomposition (on the left side) and the other for reconstruction (on the right side) instead of the same single one, interesting properties are derived [19].

![Figure 6: Bior1.3 wavelets [19]](image)

Reverse Biorthogonal Wavelet Pairs \([n, m]\), represents a reverse biorthogonal spline wavelet of order \( n \) and dual order \( m \) [20].

![Figure 7: Bior2.2 wavelets [19]](image)

‘dmey’ wavelet is a FIR based approximation of the Meyer wavelet, allowing fast wavelet coefficients calculation using Discrete Wavelet Transform.

![Figure 8: ‘dmey’ wavelets [20]](image)

In wavelet analysis the choice of a wavelet is crucial from an analysis point of view [21]. By changing the type of wavelets, as well as the level of approximation, the following tabular data were obtained. The upper part of the table refers to the approximation level 1, while the lower part, to the level 3.

<table>
<thead>
<tr>
<th>Original signal –</th>
<th>Approximation level – 1</th>
<th>Approximation level – 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (dB)</td>
<td>Median (dB)</td>
<td>Mode (dB)</td>
</tr>
<tr>
<td>haar</td>
<td>4.64</td>
<td>4.21</td>
</tr>
<tr>
<td>coif (1)</td>
<td>4.64</td>
<td>4.176</td>
</tr>
<tr>
<td>db (1)</td>
<td>4.64</td>
<td>4.246</td>
</tr>
<tr>
<td>sym (2)</td>
<td>4.64</td>
<td>4.208</td>
</tr>
<tr>
<td>bior (1,1)</td>
<td>4.64</td>
<td>4.265</td>
</tr>
<tr>
<td>rbi (1,1)</td>
<td>4.64</td>
<td>4.187</td>
</tr>
<tr>
<td>‘dmey’</td>
<td>4.64</td>
<td>4.155</td>
</tr>
</tbody>
</table>

**TABLE II.** Statistical data for two approximation levels (1 and 3) and different wavelet families – Original signal – 1
The shaded values are closest to the values that correspond to the original signal’s statistics (Row 1 in Table 1). In that sense, a smaller difference means a wavelet that generates a more precise approximation. Analyzing the tabular values, it can be concluded that the increase in the level of approximation (from level 1 to 3) deteriorated the accuracy of the calculated statistical values compared to the values of original signal. However, the discrete approximation of the symlet wavelet (sym(2)) provided the largest number of statistically accurate results, and in this regard proved to be the most suitable for the analysis of Internet traffic.

Similar to the above described case, another Internet traffic was analyzed. It derives from a private ISP with centers in 11 European cities [4]. The data correspond to a transatlantic link and were collected every hour from 7 June to 31 July 2005 (1231 values).

### TABLE III.  
**Statistical data for original signal – 2**

<table>
<thead>
<tr>
<th>Approximation</th>
<th>Mean (10^6)</th>
<th>Median (10^6)</th>
<th>Mode (10^6)</th>
<th>Maximum (10^6)</th>
<th>Minimum (10^6)</th>
<th>Range (10^6)</th>
<th>Standard dev. (10^6)</th>
<th>Median abs. dev. (10^6)</th>
<th>Mean abs. dev. (10^6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original signal - 2</td>
<td>4.574</td>
<td>1.35</td>
<td>2.055</td>
<td>10.09</td>
<td>1.524</td>
<td>8.765</td>
<td>2.581</td>
<td>1.458</td>
<td>2.211</td>
</tr>
<tr>
<td>Approximation</td>
<td>Mean (10^6)</td>
<td>Median (10^6)</td>
<td>Mode (10^6)</td>
<td>Maximum (10^6)</td>
<td>Minimum (10^6)</td>
<td>Range (10^6)</td>
<td>Standard dev. (10^6)</td>
<td>Median abs. dev. (10^6)</td>
<td>Mean abs. dev. (10^6)</td>
</tr>
<tr>
<td>Wavelet (level=1)</td>
<td>5.074</td>
<td>1.37</td>
<td>1.842</td>
<td>9.818</td>
<td>1.417</td>
<td>8.501</td>
<td>2.579</td>
<td>1.494</td>
<td>2.207</td>
</tr>
<tr>
<td>Wavelet (level=2)</td>
<td>4.574</td>
<td>1.35</td>
<td>2.055</td>
<td>10.09</td>
<td>1.524</td>
<td>8.765</td>
<td>2.581</td>
<td>1.458</td>
<td>2.211</td>
</tr>
<tr>
<td>Wavelet (level=3)</td>
<td>4.574</td>
<td>1.35</td>
<td>2.055</td>
<td>10.09</td>
<td>1.524</td>
<td>8.765</td>
<td>2.581</td>
<td>1.458</td>
<td>2.211</td>
</tr>
<tr>
<td>Wavelet (level=4)</td>
<td>4.574</td>
<td>1.35</td>
<td>2.055</td>
<td>10.09</td>
<td>1.524</td>
<td>8.765</td>
<td>2.581</td>
<td>1.458</td>
<td>2.211</td>
</tr>
</tbody>
</table>

Even though another input signal is used, statistical values of approximate accuracy have been obtained, as in the previous case (at the same level of approximation). It is interesting to note that the statistical characteristics of the corresponding two different observed traffics differ by an estimated 10 percent.

### VI. CONCLUSIONS

Several conclusions can be drawn based on the research described in the paper. First, it was concluded that all the examined wavelets have generated approximately similar statistical results, which depend on the selected level of approximation and analyzed traffic curve. For different internet traffic curves at the same level of approximation, statistical values have been obtained that do not differ significantly from the statistics of the corresponding original signals. Also, it was found that the discrete approximation of the symlet wavelet sym(2) provided the largest number of statistically accurate results, and in this regard, proved to be the most suitable for the analysis of Internet traffic.

### REFERENCES


