Internet Traffic Source Based on Hidden Markov Model

Joanna Domańska[1], Adam Domański[2], and Tadeusz Czachórski[1,2]

[1] Institute of Theoretical and Applied Informatics,
Polish Academy of Sciences,
Bałtycka 5, 44-100 Gliwice, Poland
{joanna,tadek}@iitis.gliwice.pl
[2] Institute of Informatics,
Silesian Technical University,
Akademicka 16, 44-100 Gliwice, Poland
adamd@polsl.pl

Abstract. This article shows how to use Hidden Markov Models to generate self-similar traffic. The well-known Bellcore traces are used as a training sequence to learn HMM model parameters. Performance of trained model are tested on the remaining portions of the sequences.Then we can use the HMM trained with the Bellcore data as the traffic source model.

1 Introduction

The necessity of computer modeling appears in many areas of computer networks design and use:

- in the initial design phase of the network mechanisms, to allow a realistic assessment of the quality and comparison the proposed mechanism with the existing solutions,
- at the advanced stage of design, when we know the basic features of the new product, to allow:
  - precisely determine the most profitable parameters of the prototype for the planned tasks,
  - modeling the behavior of a specific mechanism in the specific application (e.g. for wide area network, or for the integration the new network elements with the existing solutions without the need to build a full prototype),
- during the use phase, to adapt the network devices configuration and the network protocols parameters to the specific purposes.

Due to the above-mentioned reasons, for the proper implementation of the planned system it is necessary to create the realistic model of the packets (frames, cells ...) traffic. The number of the parameters of the modeled traffic should be determine in the way that modeling results were comparable to those obtained for the traffic observed directly in the network.

S. Balandin et al. (Eds.): NEW2AN/ruSMART 2011, LNCS 6869, pp. 395–404, 2011.

In the traditional queuing models it is assumed that the input stream of customers (packets, frames, cells ...) is characterized by the interarrival time distribution. The interarrival times are independent and represent the values of the same random variable, hence the generated traffic was characterized by the short-term dependencies. However, the network traffic measurements have shown that these dependencies are long-term. This feature is associated with the self-similarity of the stochastic processes [14]. The problem of self-similarity has been widely described in section 2.

The conventional modeling methods does not take into account the characteristics of self-similarity. Consequently new methods of modeling these sources were founded [5], [24], [2], [4], [7]. Their advantage is undoubtedly a good description of network traffic with low number of parameters. However, they do not offer the possibility of using well-known techniques of the queuing theory to estimate the performance of the computer networks. Therefore we need to develop the methods that allow the use of classical methods of modeling to generate the self-similarity traffic [6], [27], [28].

One of the first attempts to use Markovian modeling, in [27], [28] the authors propose the use of discrete time Markov chains (DTMC) to modulate the packet arrival process. Depending on the value of the model parameters, the traffic generated by the model displays pseudo-LRD characteristics over finite time scales. In [3], the authors use a Batch Markov Arrival Process (B-MAP) generated by a non-ergodic CTMC with an absorbing state and N transient states. The results show a better agreement with the generated traffic compared with the simple Poisson and MMPP generators. The MMPP model proposed in [31] aims at generating traffic with multifractal scaling behavior. The well-known Bellcore traces are fitted with the proposed model, and a number of tests are performed to evaluate the accuracy of the fitting. The MMPP models in [1], [29] are shown to provide good matches of LRD properties under large time scales. The authors of this article used the SSMP self-similar markovian model to generate traffic in a finite time scale [34], [35].

Hidden Markov Models (HMMs) are used in several areas of computer science. Recently the interest in HMM-based models has grown, and HMM models have been proposed as a tool for several network traffic related research problems [17], [18]. In [20], [32] HMM models have been used to model the states of packet channels via corresponding loss probabilities and end-to-end delay distributions. Similar works have been proposed to model wired [21] and wireless [16] packet channels. To the best of our knowledge, only a few modeling works using HMMs to model selfsimilar traffic sources are present in literature.

Section 3 briefly describes the issue of the Hidden Markov Models (HMMs). This section also describes how the Hidden Markov Model is used to create an Internet traffic source. Section 4 presents the analysis of the HMM source in terms of the self-similarity of the generated traffic. Some conclusions are given in section 5.

2 Characteristics of the Internet Traffic

Clasically, the traffic intensity, seen as a stochastic process, was represented in queueing models by short term dependencies [12]. However, the analysis of measurements shows that the traffic has also long-terms dependencies and has self-similar character. It is observed on various protocol layers and in different network structures [2,35,8,9,11].

The term "*self-similar*" was introduced by Mandelbrot [13] in description of proceses in the field of hydrology and geophysics. It means that a change of time scale does not influence statistical properties of the process. A stochastic process X_t is self-similar with Hurst parameter $H(0.5 \leq H \leq 1)$ if for a positive factor g the process $g^{-H}X_{gt}$ has the same distribution as the original process X_t, [14]. Mathematically, the difference between short-range dependent and long-range dependent (self-similar) processes is as follows [15]:

For a short-range dependent process:
- $\sum_{r=0}^{\infty} \mathrm{Cov}(X_t, X_{t+\tau})$ is convergent,
- spectrum at $\omega = 0$ is finite,
- for large m, $\mathrm{Var}\,(X_k^{(m)})$ is asymptotically of the form $\mathrm{Var}\,(X)/m$,
- the aggregated process $X_k^{(m)}$ tends to the second order pure noise as $m \to \infty$;

For a long-range dependent process:
- $\sum_{r=0}^{\infty} \mathrm{Cov}(X_t, X_{t+\tau})$ is divergent,
- spectrum at $\omega = 0$ is singular,
- for large m, $\mathrm{Var}\,(X_k^{(m)})$ is asymptotically of the form $\mathrm{Var}\,(X)m^{-\beta}$,
- the aggregated process $X_k^{(m)}$ does not tend to the second order pure noise as $m \to \infty$,

where the spectrum of the process is the Fourier transformation of the autocorrelation function and the aggregated process $X_k^{(m)}$ is the average of X_t on the interval m:

$$X_k^{(m)} = \frac{1}{m}(X_{km-m+1} + ... + X_{km}) \qquad k \geq 1.$$

There are several methods used to check if a process is self-similar. The easiest one is a visual test: one can observe the behaviour of the process through the scales of time. The other one is the estimation of aggregated index of dispersion IDC or aggregated coefficient of variation CV. The aggregated index of dispersion is equal to the variance of the number of arrivals within the interval m divided by the average number of arrivals during the same interval:

$$IDC(m) = \frac{\mathrm{Var}\,(mX_k^{(m)})}{E(mX_k^{(m)})}$$

and CV is

$$CV(m) = \frac{\sqrt{\mathrm{Var}\,(mX_k^{(m)})}}{E(mX_k^{(m)})}$$

For a self-similar processes, *IDC* increases on several time scales and *CV* is much more than 1 for small time intervals. Estimation of Hurst parameter is the most frequently used method to check if a process is self-similar: for non-self-similar processes $H = 0.5$; for $0.5 < H \leq 1$ process is self-similar; the closer H is to 1, the greater the degree of persistance of long-range dependence. The parameter can be estimated by various methods, among others by the analysis of variance-time plot [14]. The variation of aggregated self-similar process is equal to:

$$\mathrm{Var}\,(X_k^{(m)}) \approx \mathrm{Var}\,(X)\, m^{-\beta}, \quad \text{or} \quad \log \mathrm{Var}\,(X_k^{(m)}) \approx \log \mathrm{Var}\,(X) - \beta \log m$$

so the log-log plot of $\mathrm{Var}\,(X_k^{(m)})$ versus m is a straight line with slope $-\beta$, $0 < \beta < 1$, and $H = 1 - \beta/2$.

Self-similarity of a process means that the change of time scale does not influence the process: the original process and the scaled one are statistically the same. It results in long-range dependence and makes possible the occurrence of very long periods of high (or low) traffic intensity. These features have a great impact on a network performance. They enlarge the mean queue lengths at buffers and increase the probability of packet losses, reducing this way the quality of services provided by a network [19]. According to Stallings [19], "Self-similarity is such an important concept that, in a way, it is surprising that only recently has it been applied to data communications traffic analysis". As mentioned above, many empirical and theoretical researches have shown the self similar characteristics of the network traffic. That is why it is necessary to take into account this feature when you have to create a realistic model of traffic sources.

3 Hidden Markov Model

Hidden Markov Models (HMMs) [22] is a statistical modelling tool for systems with hidden internal states that can be observed and measured only indirectly. These models have numerous applications in computer science. Recently the interest in Hidden Markov Models (HMMs) has grown and HMM-based models have been proposed in several network traffic related research problems.

Hidden Markov Model (HMM) may be viewed as a probabilistic function of a (hidden) Markov chain [22]. This Markov chain is composed of two variables:

- the hidden-state variable, whose temporal evolution follows a Markov-chain behavior ($x_n \in \{s_1, \ldots, s_N\}$ represent the (hidden) state at discrete time n with N being the number of states)
- the observe variable which stochastically depends on the hidden state ($y_n \in \{o_1, \ldots, o_M\}$ and represents the observable at discrete time n with M being the number of observables)

An HMM is characterized by the set of parameters:

$$\lambda = \{\mathbf{u}, \mathbf{A}, \mathbf{B}\}$$

where:

- $\mathbf{u}$ is the initial state distribution, where $u_i = Pr(x_1 = s_i)$
- $\mathbf{A}$ is the $N \times N$ state transition matrix, where $A_{i,j} = Pr(x_n = s_j | x_{n-1} = s_i)$
- $\mathbf{B}$ is the $N \times M$ observable generation matrix, where $B_{i,j} = Pr(y_n = o_j | x_n = s_i)$

Given a sequence of observable variables $y = (y_1, y_2, \ldots, y_L)$ referred to as the *training sequence*, we want to find the set of parameters such that the likelihood of the model $L(\mathbf{y}; \lambda) = Pr(\mathbf{y}|\lambda)$ is maximum. We solved it via the Baum-Welch algorithm, a special case of the Expectation-Maximization algorithm [23], that iteratively updates the parameters in order to find a local maximum point of the parameter set.

We used the well-known Bellcore trace of Internet traffic: OctExt.TL [2]. Each line of this file contains a floating-point time stamp (representing the time in seconds since the start of a trace) and an integer length (representing the Ethernet data length in bytes). We translated the sequence of time stamps into the sequence of inter-arrival times. Then we apply a scheme using *Vector Quantization* (VQ) to translate the obtained sequence of inter-arrival times into a sequence of symbols, and training a HMM for this sequence. The quantization algorithm used is Linde-Buzo-Gray (LBG) algorithm of VQ [25]. Vector Quantization is a clustering technique commonly used in compression, image recognition and stream encoding [26]. It is the general approach to map a space of vector valued data to a finite set of distinct symbols, in a way to minimize distortion associated with this mapping.

We consider an HMM in which the state and the observable variables are discrete. A little portion of the sequences was used as the training sequence to learn model parameters. Performance of trained model are tested on the remaining portions of the sequences.

Then we can use the HMM trained with the Bellcore data as the Internet traffic source model. The Fig. 1 and 2 show respectively the example series of inter-arrival times which are obtained from the Bellcore trace and from our HMM traffic source.

4 Analysis of HMM Traffic from the Self-Similarity Point of View

The analysis presented in this section is based on the data generated by the HMM internet traffic source which is described in section 3.

Figures 3 i 4 confirm the presence of long-term dependencies in the generated traffic. The index of dispersion IDC increases with the time scale while the coefficient of variation CV is much greater than 1 for small time scale. For the comparison the figures present also the aggregate index of dispersion and the aggregate coefficient of variation for a Poisson process which represents the process with short-term dependencies.

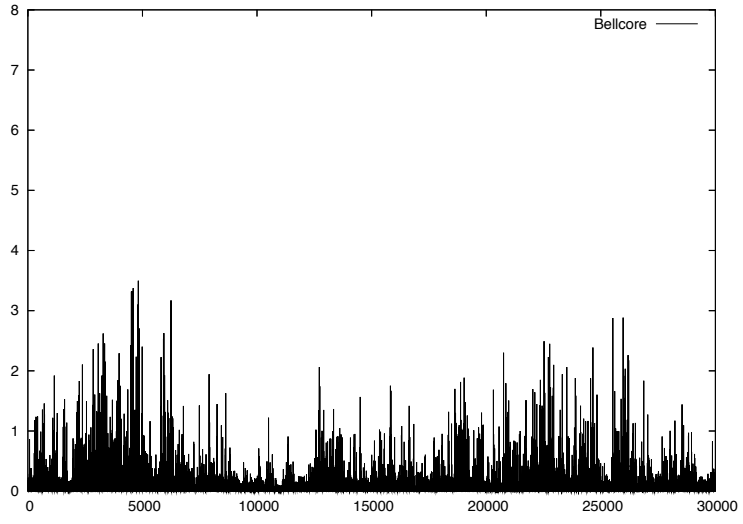

Fig. 1. The sequence of inter-arrival times [s] for Bellcore trace

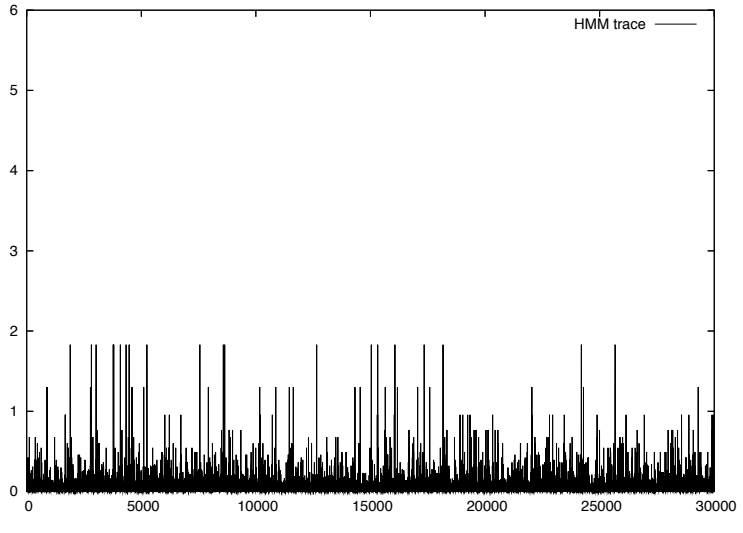

Fig. 2. The sequence of inter-arrival times [s] for HMM traffic source trace

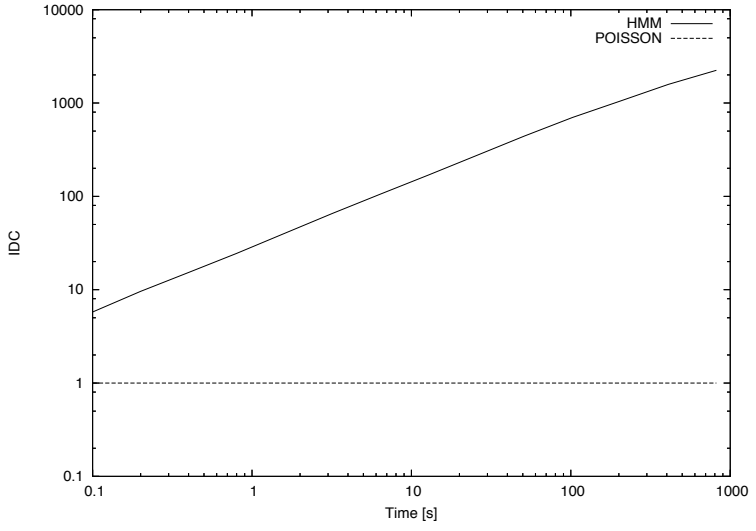

Fig. 3. Index of dispersion for HMM traffic source trace

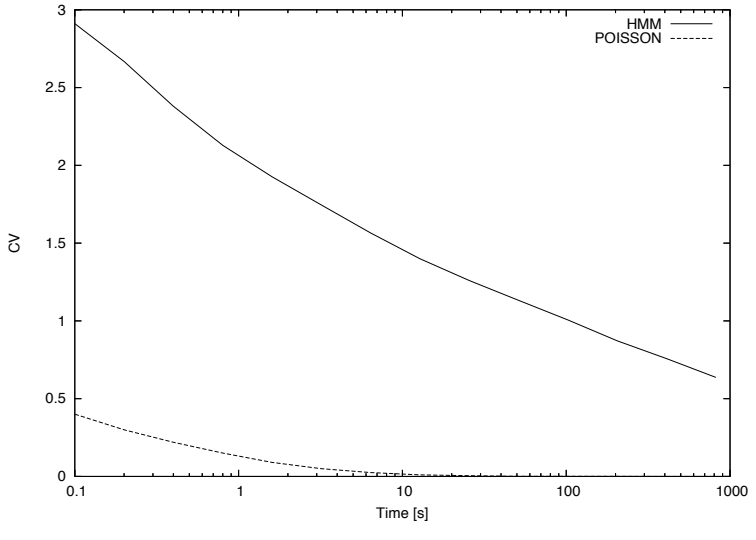

Fig. 4. Coefficient of variation for HMM traffic source trace

The degree of self-similarity of the process, expressed as a Hurst coefficient, has been calculated using the variance method (see section 2). Figure 5 shows the dependence of the variance versus time scale (on a logarithmic scale). The slope of the straight line (estimated by the least squares method) is -0.3, which gives the Hurst coefficient equal to 0.85. For the omparison, we plotted in the same figure the dependence of the variance versus time scale for a Poisson process. As might be supposed, the slope of this straight line is -1, which gives Hurst coefficient equal to 0.5 (that means no self-similarity).

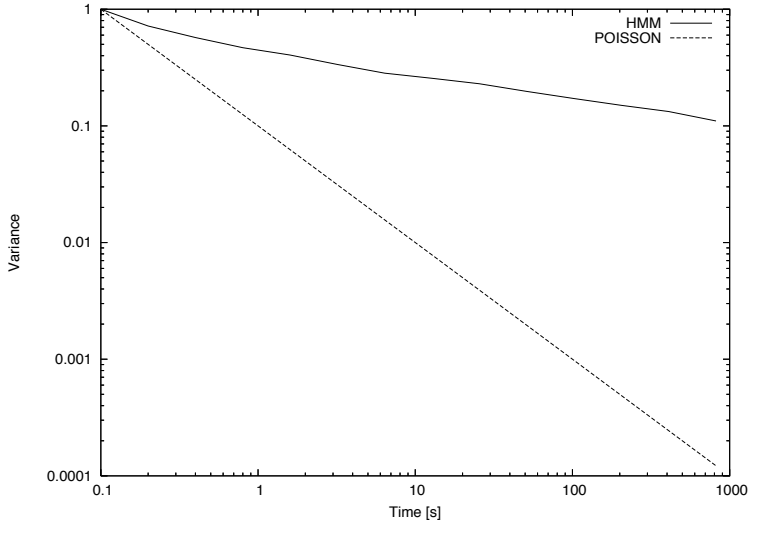

Fig. 5. Variance-time plot for HMM traffic source trace

The analysis presented in this section confirm that the developed by the authors HMM traffic sources can generate the traffic which exhibit the self-similarity.

Our HMM traffic generator can be use not only for modeling, but also for generation the real traffic in network connection – e.g. for hardware tests, as devices can be loaded with traffic having characteristics identical with generated with real application.

5 Summary

This article has demonstrated that it is possible to generate the self-similar traffic using Hidden Markov Model. Our HMM traffic source was created on the base of the well-known Bellcore Internet traffic trace.

The authors further work will focus on developing HMM sources based on the most recent measurements of the Internet traffic, made available to researchers by the CAIDA organizations [30].

Another important direction for further research is to determine how exactly the generated traffic is fitted to real traffic. In the literature one can find some studies to determine which parameters of the input stream have the greatest impact on the occupancy and the loss of packets in buffers of limited capacity. Although the major output of this study is that higher moments of the process (higher than the second) have a small impact on the processes in queues [10], the question remains still open.

Acknowledgements. This research was partially financed by Polish Ministry of Science and Higher Education project no. N N516441438.

References

1. Andersen, A.T., Nielsen, B.F.: A Markovian approach for modeling packet traffic with long-range dependence. IEEE Journal on Selected Areas in Communications 16(5), 719–732 (1998)
2. Willinger, W., Leland, W.E., Taqqu, M.S.: On the self-similar nature of ethernet traffic. IEEE/ACM Transactions on Networking (1994)
3. Klemm, A., Lindemann, C., Lohmann, M.: Modeling IP traffic using the batch Markovian arrival process. Performance Evaluation 54(2), 149–173 (2003)
4. Veitch, D., Abry, P., Flandrin, P., Chainais, P.: Infinitely Divisible Cascade Analysis of Network Traffic Data. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2000), Istanbul, Turkey, vol. 1 (June 2000)
5. Erramilli, A.: Chaotic maps as models of packet traffic. ITC 14 (June 1994)
6. Baiocchi, A., Melazzi, N.B., Listanti, M., Roveri, A., Winkler, R.: Loss Performance Analysis of an ATM Multiplexer Loaded with High-Speed ON-OFF Sources. IEEE-JSAC 9(3) (April 1991)
7. De Vendictis, A., Baiocchi, A.: Wavelet Based Synthetic Generation of Internet Packet Delays. In: Proceedings of International Teletraffic Conference ITC17, Salvador, Brasil (December 2001)
8. Paxson, V., Floyd, S.: Wide Area Traffic: A Failure of Poisson Modeling. IEEE/ACM Transactions on Networking (June 1995)
9. Crovella, M., Bestavros, A.: Self-similarity in World Wide Web Traffic: Evidence and Possible Causes. IEEE/ACM Transactions on Networking (December 1997)
10. Li, S.Q., Hwang, C.L.: Queue response to input correlation functions: continuous spectral analysis. IEEE//ACM Trans. Networking 1(3), 678–692 (1993)
11. Garret, M., Willinger, W.: Analysis, modeling and generation of self-similar VBR video traffic. In: ACM SIGCOMM, London (September 1994)
12. Kleinrock, L.: Queueing Systems, vol. II. Wiley, New York (1976)
13. Mandelbrot, B., Ness, J.V.: Fractional Brownian Motions, Fractional Noises and Applications. SIAM Review 10 (October 1968)
14. Beran, J.: Statistics for Long-Memory Processes. Chapman and Hall, Boca Raton (1994)
15. Cox, D.R.: Long-range dependance: A review. Statistics: An Appraisal (1984)
16. Iannello, G., Palmieri, F., Pescap, A., Salvo Rossi, P.: End-to-end packet-channel Bayesian model applied to heterogeneous wireless networks. In: IEEE GLOBECOM, pp. 484–489 (November 2005)

17. Dainotti, A., Pescape, A., Salvo Rossi, P., Palmieri, F., Ventre, G.: Internet Traffic Modeling by means of Hidden Markov Models. Computer Networks 52(14), 2645–2662 (2008)
18. Colonnese, S., Rinauro, S., Rossi, L., Scarano, G.: H.264 Video Traffic Modeling via Hidden Markov Process. In: 17th European Signal Processing Conference (EUSIPCO 2009), Glasgow, Scotland, August 24-28 (2009)
19. Stallings, W.: High-Speed Networks: TCP/IP and ATM Design Principles. Prentice-Hall, Englewood Cliffs (1998)
20. Salamatian, K., Vaton, S.: Hidden Markov Modeling for network communication channels. In: ACM SIGMETRICS 2001, vol. 29, pp. 92–101 (2001)
21. Salvo Rossi, P., Romano, G., Palmieri, F., Iannello, G.: Joint end-to-end loss-delay Hidden Markov Model for periodic UDP traffic over the Internet. IEEE Transactions on Signal Processing 54(2), 530–541 (2006)
22. Rabiner, L.R.: A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE 77(2) (1989)
23. Bilmes, J.A.: A Gentle Tutorial od the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. University of Berkeley (1998)
24. Norros, I.: On the use of fractional Brownian motion in the theory of connectionless networks. Technical contribution, TD94-33 (September 1994)
25. Linde, Y., Buzo, A., Gray, R.: An Algorithm for Vector Quantizer Design. IEEE Transactions on Communication Com-28 (1980)
26. Romaszewski, M., Głomb, P.: 3D Mesh Approximation Using Vector Quantization. Advances in Soft Computing 57 (2009)
27. Robert, S.: Modélisation Markovienne du Trafic dans Réseaux de Communication. PhD thesis, Ecole Polytechnique Fédérale de Lausanne, Nr 1479 (1996)
28. Robert, S., Boudec, J.Y.L.: New models for pseudo self-similar traffic. Performance Evaluation 30(1-2) (1997)
29. Salvador, P., Valadas, R., Pacheco, A.: Multiscale fitting procedure using markov modulated poisson processes. Telecommunication Systems Journal 23(1-2), 123–148 (2003)
30. The Cooperative Association for Internet Data Analysis, `http://www.caida.org`
31. Horvath, A., Telek, M.: A Markovian Point Process Exhibiting Multifractal Behavior and its Application to Traffic Modeling. In: Proceedings of Fourth International Conference on Matrix-analytic Methods in Stochastic Models, Adelaide, Australia (July 2002)
32. Wei, W., Wang, B., Towsley, D.: Continuous-time Hidden Markov Models for network performance evaluation. Performance Evaluation 49(1-4), 129–146 (2002)
33. Domańska, J.: Procesy Markowa w modelowaniu natężenia ruchu w sieciach komputerowych. PhD thesis, IITiS PAN, Gliwice (2005)
34. Domańska, J., Domański, A., Czachórski, T.: The Drop-From-Front Strategy in AQM. In: Koucheryavy, Y., Harju, J., Sayenko, A. (eds.) NEW2AN 2007. LNCS, vol. 4712, pp. 61–72. Springer, Heidelberg (2007)
35. Domański, A., Domańska, J., Czachórski, T.: The impact of self-similarity on traffic shaping in wireless LAN. In: Balandin, S., Moltchanov, D., Koucheryavy, Y. (eds.) NEW2AN 2008. LNCS, vol. 5174, pp. 156–168. Springer, Heidelberg (2008)