A Packet Distribution Traffic Model for Computer Networks

Ewerton Castro*[†], Ajey Kumar*[‡], Marcelo S. Alencar*[‡], Iguatemi E.Fonseca[⋆]
*Federal University of Campina Grande, Department of Electrical Engineering, Iecom, Campina Grande, PB, Brazil
[†]E-mail: ewerton.castro@ee.ufcg.edu.br
[‡]E-mail: {ajeykumar,malencar}@iecom.org.br
[⋆]Universidade Federal Rural do Semi-Árido, Mossoró, RN, Brazil E-mail: iguatemi@ufersa.edu.br

***Abstract*—The traditional traffic models which are based on conventional telephone traffic are not suitable for modeling self-similar traffic on computer networks. Therefore, emphasis has been put on self-similarity characteristics. This paper presents a model which can be used to compare, simulate and estimate the packet traffic distribution on computer networks.**

***Index Terms*—Self-similar, Computer Network, Internet Traffic, File Size Distribution**

I. Introduction

The modeling of network traffic is generally based on traditional telephone traffic models [1], [2]. They are the pure Poisson or Poisson-related model such as Poisson-batch or Markov-Modulated Poisson processes, packet-train models or fluid flow models. But, measurements of real traffic indicate that the commonly assumed models for voice traffic are not suitable for modeling data traffic, i.e. traffic on computer networks. Understanding the nature of network traffic is critical in order to properly design and implement computer networks and network services. Self-similar methods are used to model that type of traffic [3].

Kihong showed that the degree to which file sizes are heavy-tailed, can directly determine the degree of traffic self-similarity at the link level [4]. The relationship between self-similar traffic and file sizes was suggested by Crovella [5]. Paxson shows that the file size distribution (in *bytes*) using FTP (*File Transfer Protocol*) fits a Pareto distribution with $0.9 \leq \alpha \leq 1$ [2]. Pustisek made a statistical analysis of the flow of traffic data and, using some parameters from measurements, he found that the flow size, measured in number of packets, is modeled by a Pareto distribution [6].

Rastin found that 90% of the UDP packets are smaller than 500 *bytes* and that most packets are transmitted via TCP, with 40 *bytes* of *Acknowledgment* and 1500 *bytes* of Ethernet *Maximum Transmission Unit* (MTU) [7].

Tafvelin showed a bimodal traffic distribution, in which 40% of the packets are smaller than 44 *bytes* (first peak) and another 40% packets are between 1400 *bytes* and 1500 *bytes* (second peak) [8]. These results are similar to Rastin's, who found a bimodal packet size distribution with 43% of the packets having a length of 40 *bytes* and 30% of the packets contain 1500 *bytes* of information. That behavior is demonstrated by graphs, of size distributions for the packets, presented later in this paper. This paper summarizes some results from the literature regarding the packet size distribution. Using the results a mathematical model for the packet distribution is presented and compared with actual measurements of packet sizes.

The remaining of the paper is organized as follows. Section II describes how and under what conditions the data were collected. Section III shows a summary of measured data. Section IV presents a mathematical model to estimate the size distribution of the packets. Section V compares the measured values with the proposed mathematical model and Section VI concludes the paper.

II. Description of the Measurement Process

The measurements are divided into two types: performed measurements and measurements obtained from other sources. For the performed measurements, the program IPTRAF was used on a desktop computer with Linux. IPTRAF collects the packet sizes at the input and output of the network during a specified time. This information is then saved periodically in a *log* file. Data sets were obtained from the Ville Mattila site [9].

A. Performed measurements

To perform the measurement, the following scenarios of data traffic were chosen:

Data Set I: *One computer in a lab* – First, the data collection was performed taking packets directly from the Internet at one of the Iecom labs. The main objective was obtain different packet sizes from a computer that had access to Internet content. To do this, the information accessed during the collection period as diverse as possible, including Brazilian news sites, *blogs* sites, Brazilian portals, videos site, *webmail*, *download* of videos, programs and CD images (see Table I).

Data Set II: *A computer accessing sites with video content only* – The second data set was obtained in a situation in which a computer had access to video content. Several videos were opened, with different durations, from the YouTube website.

Data Set III: *A computer downloading files using Torrent (p2p)* – The third data set was divided into two subsets, A and B. Both were obtained when a computer downloaded content from the Internet using programs such as *Torrent*. In subset A, several downloads of files of varying sizes were made (5MB, 10MB, 12MB, 15MB e 17MB). Then, in subset B, the image of a 2.1 G*bytes* DVD was downloaded.

Data Set IV: *A computer downloading files using FTP* – The fourth collection was assembled using the traffic data of one computer downloading the contents from the Internet using FTP. In this experiment, an image of a DVD with 1.8 Gbytes was downloaded using FTP.

Data Set V: *Collection of all traffic passing through a server in a laboratory of the Department of Computer Science* – The fifth data set was collected from server in the Department of Computer Science at UFCG. This server is connected with 56 computers, divided in 3 classrooms. In first room has 10 desktops (LAN) and 16 notebooks (WLAN).The second room has 10 desktops and the third has 20 desktops.

Data Set VI: *Traffic from a server of a packaging industry* – The traffic data were obtained from the gateway server in a packaging industry. This gateway server is connected to an ADSL modem running at 1 M*bit/s*, to provide Internet access of 80 computers divided into 5 rooms.

B. Obtained measurements

Data Set VII: *Data traffic measurements available on the Internet*. – The final data sets were obtained from the Ville Mattila website [9] (see Table II).

III. Size and frequency of occurrence of the packets

The first two columns of Table I show the values in the *log* file obtained using IPTRAF and the remaining columns are derived from it. The first line shows 582,510 packets with sizes ranging from 1 to 75 (*bytes*). Using the concept of intervals, class limits and the midpoint of a class from statistical theory, the fourth column of Table I, Medium Size Package (TMP_i), is obtained. TMP for the interval i, is given by

$$TMP_i = \frac{(Vm_i + VM_i)}{2} \qquad 1 \leq i \leq 20, \tag{1}$$

in which, Vm_i and VM_i are the lower values and higher value of i-th interval, respectively. Column ”T” in Table I. T shows the packet size (in *bytes*) and FP is the frequency of occurrence of the packets.

The TMP standard value (TMP_s) shown in Table I is obtained from Equation (2) dividing the value of TMP_i by MTU (1500), standard for Ethernet networks. In the last column of the same table, the values of FP standard or (FP_s) are obtained dividing the number of occurrences of the packet size by the total number of packets, given by Equation (3).

$$TMP_s = \frac{TMP_i}{1500} \qquad 1 \leq i \leq 20 \qquad 1 \leq s \leq 20 \tag{2}$$

$$FP_s = \frac{FP_i}{\sum_{i=1}^{20} FP_i} \qquad 1 \leq i \leq 20 \qquad 1 \leq s \leq 20 \tag{3}$$

i	T	FP	TMP_i	TMP_s	FP_s
1	1 - 75	582510	38	0.02533	0.3941991128
2	76 - 150	11559	113	0.07533	0.0078222649
3	151 - 225	5471	188	0.12533	0.0037023628
4	226 - 300	9506	263	0.17533	0.0064329484
5	301 - 375	5056	338	0.22533	0.0034215219
6	376 - 450	3203	413	0.27533	0.0021675504
7	451 - 525	6548	488	0.32533	0.0044311957
8	526 - 600	19331	563	0.37533	0.0130817721
9	601 - 675	5007	638	0.42533	0.0033883624
10	676 - 750	4722	713	0.47533	0.0031954957
11	751 - 825	5114	788	0.52533	0.0034607719
12	826 - 900	4666	863	0.57533	0.0031575991
13	901 - 975	3353	938	0.62533	0.0022690591
14	976 - 1050	3166	1013	0.67533	0.0021425115
15	1051 - 1125	3144	1088	0.72533	0.0021276236
16	1126 - 1200	2604	1163	0.77533	0.0017621920
17	1201 - 1275	3965	1238	0.82533	0.0026832149
18	1276 - 1350	2257	1313	0.87533	0.0015273685
19	1351 - 1425	10148	1388	0.92533	0.0068674059
20	1426 - 1500+	786375	1463	0.97533	0.5321596665
	Total	1,477,705	15,010		

TABLE I
Data collected using IPTRAF and standardized for Data Set I.

i	T	FP	TMP_i	TMP_s	FP_s
1	1 - 16	0	8.5	0.00566	0
2	17 - 32	45	24.5	0.01633	0.0000018679
3	33 - 48	1936144	40.5	0.02700	0.0803666205
4	49 - 64	6645143	56.5	0.03766	0.2758305611
5	65 - 80	266130	72.5	0.04833	0.0110466829
6	81 - 96	820938	88.5	0.05900	0.0340759844
.	.	.	.	.	.
.	.	.	.	.	.
91	1441 - 1456	2427	1448.5	0.96566	0.0001007414
92	1457 - 1472	2321	1464.5	0.97633	0.0000963415
93	1473 - 1488	12575	1480.5	0.98700	0.0005219706
94	1489 - 1504	12084780	1496.5	0.99766	0.5016222597
	Total	24,091,395	70,735		

TABLE II
Data obtained using IPTRAF and standardized for Data Set VII-b.

It is important to notice that the values presented in Tables I and II were obtained from measurements and the Internet as mentioned in Section II, but the quantity of tables are too big to be inserted in this paper. So, snapshots are included to present a rough idea to the readers about the data base used.

IV. Mathematical Model

The mathematical model is based on the analysis of Table I, data input and data output of the system, the concept of system identification, and using the Maple and Matlab programs to adjust the parameters of the curve to approximate the measured data.

First consider the sine tone probability density function (pdf), given by [10]

$$p_X(x) = \frac{1}{\pi\sqrt{V^2 - x^2}}, \qquad | x | < V \tag{4}$$

in which V is the maximum sinusoidal amplitude. The probability density function (pdf) are shown in Figure.

The curve represents the traffic behavior for an idealized system. For a real network the measured traffic is asymmetrical, as shown in Figure 2. A new mathematical model is proposed, as follows.

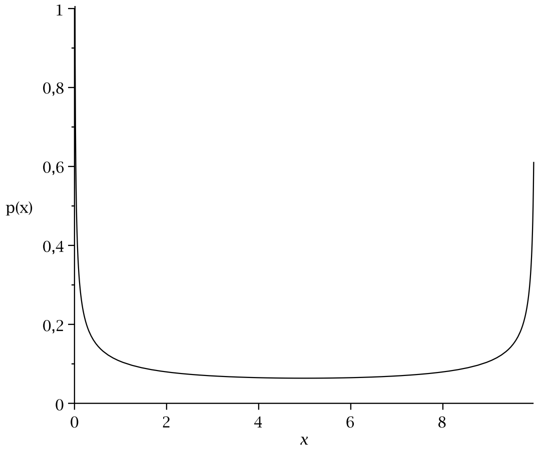

Fig. 1. Probability density function.

In one of the Tafvelin papers, he observed a bimodal traffic distribution, 40% of the packets are smaller than 44 *bytes* (first peak) and another 40% of the packets are between 1400 *bytes* and 1500 *bytes* (second peak) [8]. This behavior is verified in the pdf, in Figure 1, and in the packet sizes measurement, in this paper. This means that they are similar. The CDF of the packet sizes were shown in Rastin's paper [7] and the CDF from Figure 1 presents the same behavior.

The following equation is proposed to model the distribution of network traffic

$$p(w) = \frac{1}{2\sigma\sqrt{2\pi}}.\frac{e^{-\frac{[\arccos(2w-u)^2]}{2\sigma^2}}}{\sqrt{u^2-(2w-u)^2}}, \tag{5}$$

in which, u is the normalized MTU (*Maximum Transmission Unit*), w is the packet size in the interval i, $p(w)$ is the probability density function of a packet of size w and σ is a parameter of the distribution function related to the traffic type.

V. Results

A. Comparison between the actual measurement and the proposed model

Figure 2 shows two distinct graphs. The bar-graph shows the measurements from Table I in each interval. The second, a continuous line, represents $p(w)$, adjusted by the least squares method to find the lowest value of σ, with $\sigma > 1$, which is a requisite of the model. The graph of $p(w)$ presents a peak near the origin. Using the same fitting procedure for each of the data sets mentioned in Section II, leads to the graphs illustrated in Figures 2 up to 6. It is important to notice that, in Figure 6, the bar-graph is thinner due of the small interval i of data sets (see Table II). The results are summarized in Table III.

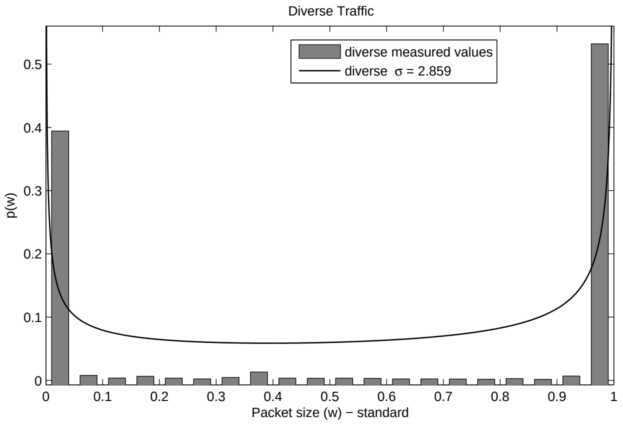

Fig. 2. Measurements and $p(w)$ for Data Set I.

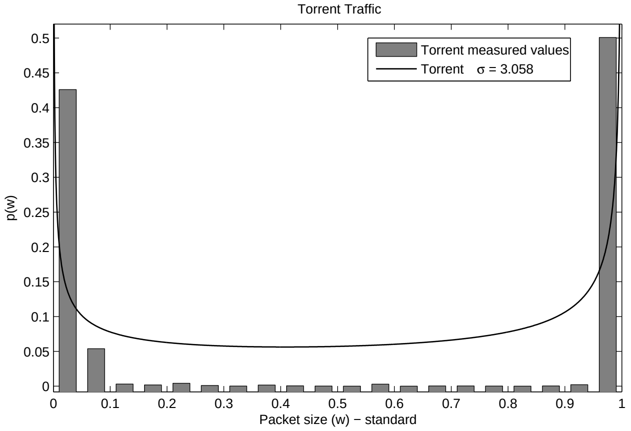

Fig. 3. Measurements and $p(w)$ for Data Set III-a.

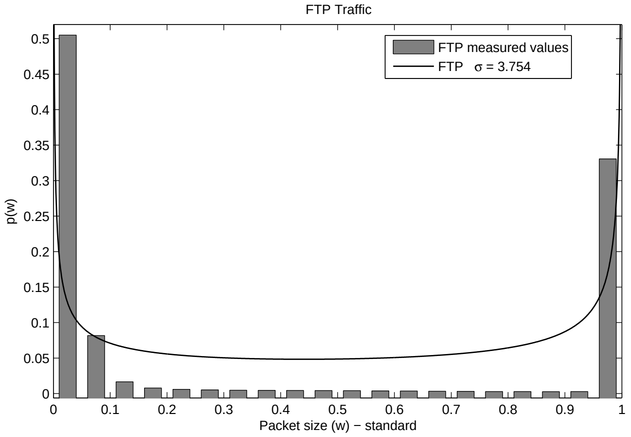

Fig. 4. Measurements and $p(w)$ for Data Set IV.

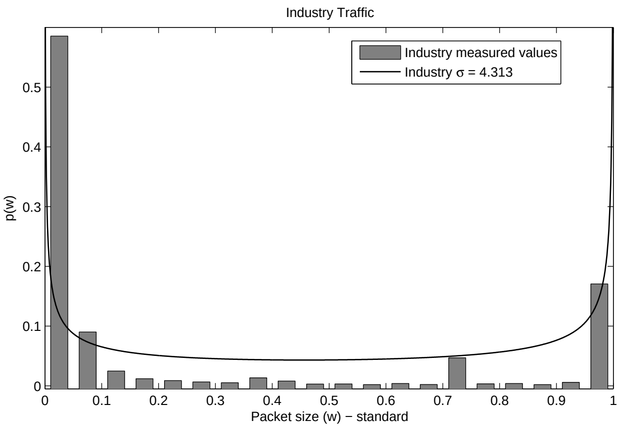

Fig. 5. Measurements and $p(w)$ for Data Set VI.

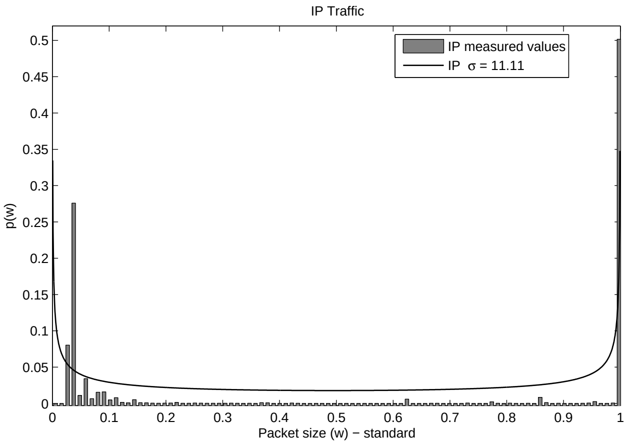

Fig. 6. Measurements and $p(w)$ for Data Set VII-B.

In Table III, "Internet" represents data obtained from the Internet [9]. The total number collected of packets is N. Using The values of σ, in column 3, control the shape of $p(w)$, as shown in Figures 2 to 6. The SSE is the *Sum of Squares due to Error* and RMSE is *The Root Mean Square Error*. For both, values that are close to zero indicate a good fitting.

Based on the data from Table III and Figures 2 to 6 one concludes that:

- The values of σ can be separated into two blocks. For the first block, obtained from the measurements, $2.69 \leq \sigma \leq 4.305$. For the values obtained from the Internet, $10.72 \leq \sigma \leq 36.48$.
- Curves for $p(w)$ are those that are close to the values of the packet sizes, with values of σ in column 3 of Table III. The values of SSE and RMSE show approximation errors.
- The $p(w)$ curves show peaks at the left and right ends. This confirms the results of Rastin [7] and Tafvelin [8], but the error increases in the middle. This reflects in higher values of SSE and RMSE presented in the last two columns of Table III.

Type	N packets	σ	SSE	RMSE
*Torrent*2	2,618,212	2.69	0.2909	0.1237
YouTube	203,764	2.756	0.288	0.1231
Diverse	1,477,705	2.859	0.2535	0.1155
Torrent	214,105	3.06	0.253	0.1154
FTP	3,606,361	3.753	0.2202	0.1077
Comp	40,903,828	3.758	0.1165	0.07832
Industry	10,149,954	4.305	0.254	0.1156
Internet TCP	23,007,226	10.72	0.2457	0.0514
Internet IP	24,091,395	11.11	0.223	0.04897
Internet HTTP	622,582	15.68	0.2307	0.04981
Internet SMTP	112,859	24.09	0.1812	0.04414
Internet SSH	279,991	31.25	0.3154	0.05824
Internet Domain	22,029	34.19	0.1918	0.04541
Internet ICMP	290,800	34.67	0.9228	0.09961
Internet UDP	332,804	36.48	0.2755	0.05443

TABLE III
TABLE WITH σ VALUES.

- The values of σ are lower for heavy traffic applications. For example: a Youtube video uses a higher data transmission rate, for the available link, for a longer period of time, than opening an HTTP (*Hypertext Transfer Protocol*) site, this means that the value of σ is lower for the YouTube video then for the HTTP site.
- The last five rows of Table III, represent the values obtained from Internet traffic for SMTP (*Simple Mail Transfer Protocol*), SSH (*Secure Shell*), DNS (domain), ICMP and UDP. The traffic, observed for each of those applications or protocols, is characterized by a high number of packets of small sizes. In the case of traffic generated by the DNS (*Domain Name System*) or UDP, all packets have sizes that are less than 600 *bytes*. This characteristic of UDP was highlighted by Rastin [7].
  The ICMP case is similar, many of the packet sizes are equal to zero and there is a large number of packets of small sizes, less than 400 *bytes*. Due to the peculiarities of SMTP, its traffic has three peaks, one near the origin, a second peak at about 550 *bytes* and the third of the end of the scale. However, that specific behavior is hardly noticed in the IP graph (Figure 6), because the IP protocol incorporates all data in one set and due to the total number of packets distributed in all size ranges, the imperfections of the model are minimized when analyzed for each specific protocol. For SSH, the behavior of the graph is similar to that illustrated in Figures 2 and 5.

B. Analysis of the influence of the parameter σ

Figure 7 illustrates the behavior of $p(w)$, parametrized by σ, for each set of measured data. In this figure and in the following ones, $p(w)$ scale axis was adjusted to shown that the differences between the curves depend on the value of σ. Each value of σ depends on the type of data traffic on the network. It is observed that applications p2p (*Torrent2*) and video (*YouTube*), which require a high transmission rate for a time period longer, present a lower value of σ than the (*Web*) sites and *E-mail* server (industry data).

Figure 8 illustrates the behavior of $p(w)$, with $\sigma > 1$, for different values of σ, for data sets obtained from the Internet. The figure shows that the difference from $p(w)$ depends on the value of σ. Each value of σ depends on the type of data traffic on the network. The main difference in this data set is that the data were sorted by protocols. It is observed that the values of σ for IP, TCP and HTTP are smaller than the others.

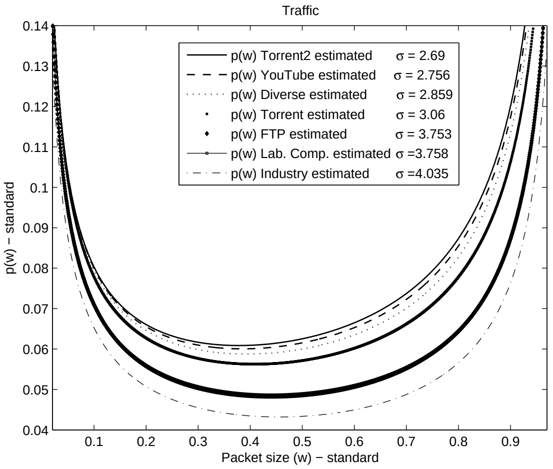

Fig. 7. Comparison of curves $p(w)$ for the set of measured data.

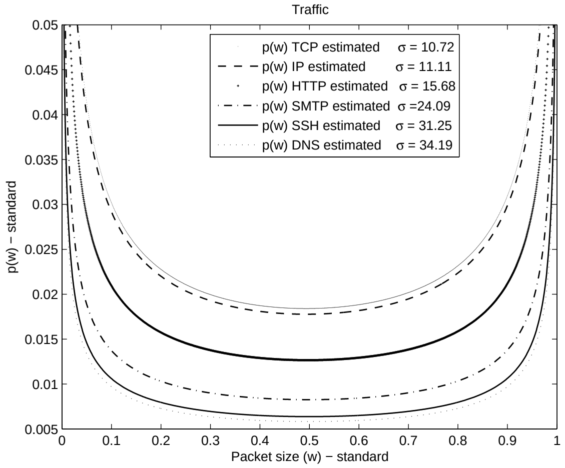

Fig. 8. Comparison of $p(w)$ curves, with emphasis on the differences between the curves for each type of traffic.

Figure 9 shows some results for data sets obtained from the Internet and the values measured in the laboratory for $p(w)$ and with $\sigma > 1$. The data obtained from the Internet present a greater granularity for the packet sizes. This is reflected, graphically, in the proximity of the curve $p(w)$ with the values of intermediate-sized packets, in Figure 6.

VI. Conclusions

This paper presents a packet size distribution model. This model can be used to estimate the distribution of packet traffic on computer networks.

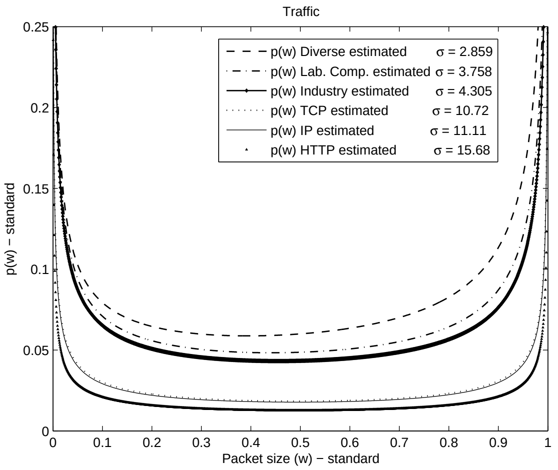

Fig. 9. Comparison of $p(w)$ with the set of measured data and the set of obtained data.

Data was collected under various situations and from many sources, and a mathematical model to estimate the size distribution of the packets was proposed.

A comparison has been provided for the obtained results, and it was observed that the values of σ are usually low for applications with heavy traffic. It was also observed that the behavior of the graph of $p(w)$, with peaks near the origin and near the MTU, are similar to the results obtained by Rastin and Tafvelin.

References

[1] W. E. Leland, M. S. Taqqu, W. Willinger and D. V. Wilson, *On The Self-Similar Nature of Ethernet Traffic*, IEEE ACM Trans. on Networking, 1994, vol.2, n.1,pg 1-15, February.

[2] V. Paxson and S. Floyd, *Wide-Area Traffic: The Failure of Poisson Modeling*, IEEE/ACM Transactions on Networking, 1995, vol. 3, n. 3, pg 226-244, Jun.

[3] Mark E. Crovella and Azer Bestavros, *Explaining World Wide Web Traffic Self Similarity*, Computer Science Department, 1995, Technical Report TR-95-015, Boston University.

[4] Kihong Park, Gitae Kim and M. Crovella, *On the relationship between file sizes, transport protocols, and self-similar network traffic*, International Conference on Network Protocols, 1996, pg.171-180, Oct.

[5] Mark E. Crovella and Azer Bestavros, *Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes*, IEEE/ACM Trans. on Networking, December, 1997, v. 5, n. 1, pg.835-846.

[6] M. Pustisek, I. Humar and J. Bester, *Empirical Analysis and Modeling of Peer-to-Peer Traffic Flows*, The 14th IEEE Mediterranean Electrotechnical Conference, MELECON, May 2008, pg. 169-175.

[7] R. Pries, F. Wamser, D. Staehle, K. Heck and P. Tran-Gia, *Traffic Measurement and Analysis of a Broadband Wireless Internet Access*, IEEE 69th Vehicular Technology Conference. VTC Spring 2009, pg.1-5, April.

[8] W. John and S. Tafvelin, *Analysis of Internet Backbone Traffic and Header Anomalies Observed*, IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, New York, NY, USA, 2007, pg. 111-116.

[9] Traffic Analysis, A review of Internet traffic packet size distributions, http://poliisi.iki.fi/˜ville/sekalaiset/Internet/traffic_ analysis/packet_ size_ distribution/data/, accessed February 2010.

[10] M. S. Alencar, *Probabilidade e Processos Estocásticos*. Editora Érica Ltda, 2009. ISBN 978-85-365-0216-8, So Paulo, Brasil.