APPLICATION OF EXTREME VALUE THEORY TO BURSTS PREDICTION

Abdelmahamoud Youssouf Dahab dahab_tc@hotmail.com
Computer & Information Sciences Department,
Universiti Teknologi PETRONAS,
Bandar Seri Iskandar, Tronoh, 31750 Perak, Malaysia.

Abas bin Md Said abass@petronas.com.my
Computer & Information Sciences Department,
Universiti Teknologi PETRONAS,
Bandar Seri Iskandar, Tronoh, 31750 Perak, Malaysia

Halabi bin Hasbullah halabi@petronas.com.my
Computer & Information Sciences Department,
Universiti Teknologi PETRONAS,
Bandar Seri Iskandar, Tronoh, 31750 Perak, Malaysia.

Abstract

Bursts and extreme events in quantities such as connection durations, file sizes, throughput, etc. may produce undesirable consequences in computer networks. Deterioration in the quality of service is a major consequence. Predicting these extreme events and burst is important. It helps in reserving the right resources for a better quality of service. We applied Extreme value theory (EVT) to predict bursts in network traffic. We took a deeper look into the application of EVT by using EVT based Exploratory Data Analysis. We found that traffic is naturally divided into two categories, Internal and external traffic. The internal traffic follows generalized extreme value (GEV) model with a negative shape parameter, which is also the same as Weibull distribution. The external traffic follows a GEV with positive shape parameter, which is Frechet distribution. These findings are of great value to the quality of service in data networks, especially when included in service level agreement as traffic descriptor parameters.

Keywords: Traffic Bursts, Extreme Value Theory, Prediction, Quality of Service, Self-Similar.

1. INTRODUCTION

Extreme bursts have detrimental effects to network traffic and quality of service. These effects may go far beyond congestion and delay. The Quality of Service (QoS) suffers as well. To remedy, bursts need to be predicted in advance so that proper measures can be taken to mitigate their effect. We need to understand the structure of traffic correctly so that we will be able to predict bursts and enhance the QoS offered.

The discovery of the self-similarity of network traffic is a mile stone in computer networks. All the old method of modeling the traffic may not withstand the self-similarity of the network traffic. New models based on the self-similar properties are being developed [8, 15]. Self-similarity means that traffic is bursty over wide range of scales. Bursts can be defined as high aggregation of data in a relatively very small time interval. This definition is a general one and can be defined more precisely depending on the given contexts.

Traffic prediction problem has been approached using different techniques. These techniques range between stochastic processes based methods [10, 11], Autoregressive moving average (ARMA) with its variants FARIMA, ARIMA to Artificial Neural Network (ANN) and Wavelet Based predictors, recent survey of these methods with references is given in [6]. However, these techniques are using the whole set of data and can be computationally expensive. We suggest the Extreme Value Theory (EVT) as a framework to deal with the bursts prediction problem.

EVT is a branch of knowledge that stems out of statistics. It is analogous to the central limit theorem (CLT). While the CLT deals with the distribution of the sample mean and tells us that it converges asymptotically to a normal distribution, the EVT deals with sample maximum and tells that it converges asymptotically to one of three distributions (Gumbel, Weibull, and Fréchet), these distribution limits are combined in a single representation called Generalized Extreme Value Distribution (GEV). EVT is a rational framework to the problem of burst prediction. It needs only a subset of the data to work on. This will greatly reduce the time space-complexity of the method. Another advantage is that the EVT is being developed for these kinds of problems, i.e. predicting extreme events based on subset of the data. It has been in use in diverse fileds such as Insurance, Finance, and Hydrology etc, see [2, 5].

Recent studies suggested EVT as a framework for modeling different types of traffic [4, 12, 18, 19]. In [19], Masato Uchida used the throughput as a network parameter to be modeled. He argued that throughput, link usage rate, packet loss rate and delay time can be used to predict telecommunication quality. He used the Peak Over Threshold (POT) method for the modeling and fitted a generalized Pareto distribution. He showed that the POT using GPD is better in approximating the unknown part of the data than the previously commonly used lognormal distribution.

In [12], authors also used POT method for the analysis of wireless traffic. They fitted a GPD model and compared their model to the lognormal, Gamma, Exponential models. The computational overhead is clearly reduced when using the EVT model because we need only a subset of the data to work on.

However, these studies are far from complete. They did not include a rigorous check to see the applicability of the EVT analysis. They applied the theory of EVT and in particular the POT method with the assumptions that it holds true. They did not include any preliminary analysis of the data to check whether the theory assumptions hold or not. Such steps are important and crucial for the success of the model. One of the reasons for this is that dependent correlated data do not necessarily converge to the classical form of the three distributions limits. As shown in [9], if the data are highly dependent, then a further parameter called extremal index needs to be carefully introduced.

In a previous work, we applied the EVT to the internal traffic [4]. In this work, we extend our model to the external traffic as well. We applied EVT more faithfully by using EVT based exploratory tools. Our findings are unique. We found that traffic distinguishes itself into external and internal by assuming different signs in GEV model. Using the Block Maxima (BM) method, the external traffic follows a GEV with positive shape parameter, which is analogous to the traffic bursts following extreme Frechet distribution. Internal traffic follows a GEV with a negative shape parameter. This also means that traffic bursts follow extreme Weibull distribution.

The structure of this paper is as follows. First, we present EVT based model selection tools. We apply them to Belcore traffic traces. We then estimate the parameters and show our prediction results, and then we conclude.

Dataset	Type	Size	Source	Parameters
BC-pAug98	Internal Traffic	31429	Belcore	
BC-pOct98	Internal Traffic	15795		
BC-Oct89Ext	External Traffic	759431		
FGN078	Internal Traffic	10000	Simulated FGN	H=0.78
LFSN	External Traffic	10000	Simulated LFSN	H=0.75, Alpha=1.5

TABLE 1 : Datasets

2. MODEL SELECTION TOOLS

Visualizing data is a very important step before a serious decision is to be taken about the underlying process. A great deal of work has been done to illustrate this point. Such analysis bears the name of Exploratory Data Analysis (EDA). In our work, we refer to EDA as EVT based model selection tools. Selecting the right model is important. In this section, we discuss some of EVT based model selection tools. Namely, Records, Maximum to Sum Ratio, Gumbel Plot, Mean Excess plot. Other tools such as QQ-plot and Hill plot will be discussed here. Using these tools, we analyze the Ethernet traffic traces from Belcore data and simulated ones.

Our five data sets consist of three data sets from Belcore Labs [21] and the other two are simulated traces for both internal and external traffic [1, 14, 17]. We transformed Belcore traces into bitrate per 0.1 second using simple MATLAB routines. The simulation of the internal trace is based on Fractional Gaussian Noise (FGN) model suggested by Norros in [14] and the simulated trace of the external traffic is based on Linear Fractional Alpha Stable Noise (LFSN) model, see [20]. A summary of the data under study is given in Table1.

2.1 Records

Records can be used as exploratory tool in distinguishing between i.i.d and non i.i.d data. The number of records of i.i.d data grows very slowly [5]. This fact allows us to use records in our traffic data and check it against expected records in a typically known i.i.d data. If there is a match in the number of records, then we may say that our data can be modeled as i.i.d otherwise we say that our traffic data cannot be modeled as coming from i.i.d random process.

A record occurs if $X_n > M_{n-1} = \max(X_1,...,X_{n-1})$. By definition X_1 is a record. Let I be indicator function, the record counting process N is given by

$$N_1 = 1, N_n = 1 + \sum_{k=2}^{n} I_{X_k > M_{k-1}}, n \geq 2$$

How would Records help? To answer this question, we tabulated the number of expected records in a typical i.i.d dataset [5]. For the purpose of comparison, we computed the number of records in the examined datasets.

A dramatic departure from the expected values in this table suggests clearly rejecting i.i.d assumption. From Table 2, it is clear that disparity exists between the number of records in our data and those expected from a typical i.i.d dataset. For the Aug98 data, it assumes higher values than the average, while the Oct98 data assumes below average values. Nevertheless, they are still within one standard deviation from the mean. So the i.i.d. assumption is a valid one.

$n = 10^k$	EN	$\sqrt{VN}$	BC-pAug98	BC-pOct98	FGN078	LFSN	Oct89Ext
1	2.9	1.2	2	3	1	4	1

2	5.2	1.9	7	4	2	6	2
3	7.5	2.4	10	7	7	9	6
4	9.8	2.8	12	11	7	12	12

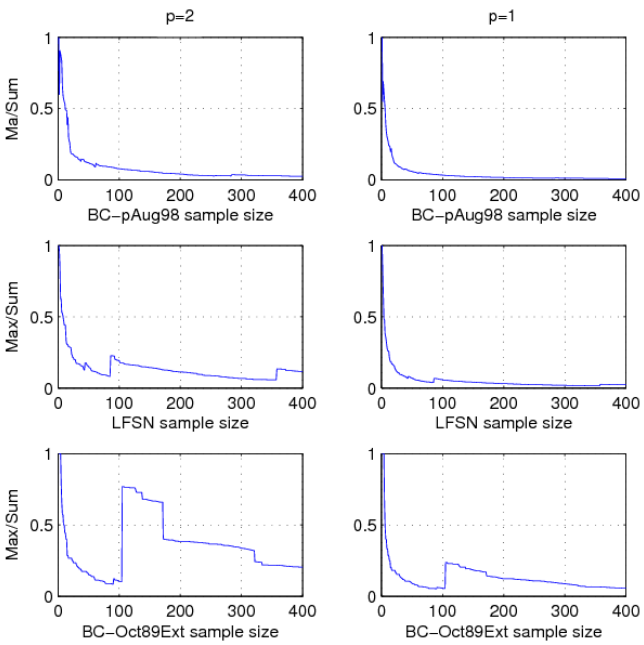

Fig. 1. Maximum to Sum Ratio with p=2, and p=1 for bytes/100 ms datasets.

Ε 2: Records

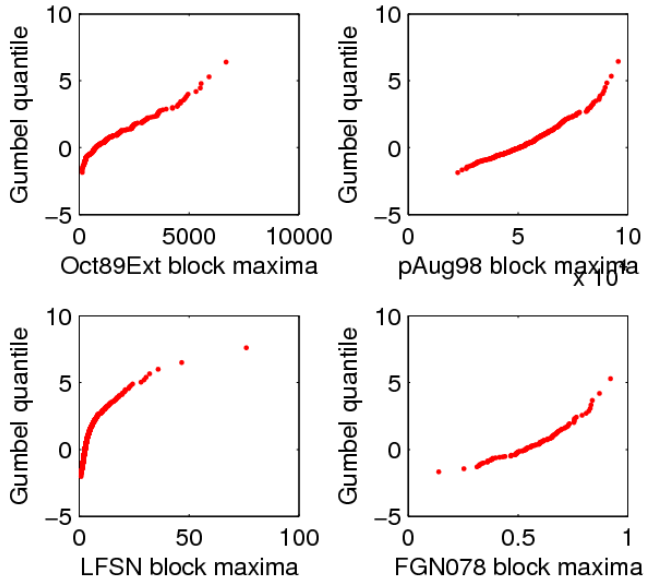

Fig.2. Gumbel plots. Some plots show concave departure from linearity and some show a convex departure from linearity.

2.2 Maximum to Sum Ratio

Maximum to sum ratio M_n / S_n can be used as an explanatory tool to tell about the finiteness (existence) of the moment of given order, say p. It is a standard knowledge that the mean and variance of a given data are their first order and second order moment, respectively.

Using maximum to sum ratio, we will be able to tell whether the variance, which is the second order moment, of data is finite (exists) or infinite (does not exist). The p^{th} order partial sum and p^{th} order maximum are given by $S_n(p) = \sum_{i=1}^{n} |X_i|^p$ and $M_n(p) = \max(|X_1|^p, ..., |X|^p)$, respectively. From [5], we have the following equivalence relation

$$R_n(p) = M_n(p) / S_n(p) \to 0 \Leftrightarrow E|X|^p < \infty$$

This equivalence relation means that p^{th} order maximum to p^{th} order partial sum ratio goes to zero as approaches infinity if and only if the p^{th} order moment exists.

A direct way to exploit the above fact is to plot the max to sum ratio for 1st or 2nd order moment, if the plot for a given moment order goes to zero, then we might say that the moment of that order is finite (exists). It is also to note that a heavy tail distribution with tail parameter $\alpha < 2$ has an infinite variance and if $\alpha < 1$, the mean also is infinite. In the first row first column in Figure 1, we see that internal Ethernet traffic data have a finite variance because the Maximum to Sum Ratio tends to zero very quickly. Notice that we have plotted only the first 400 values, which account for only a fraction of the data. Also from the Figure1, we see that the external traffic (2nd and 3rd row), represented by Oct89Ext and LFSN, have infinite variance but a finite mean.

Now that we have an idea about the different moments and structure of the traffic, it is time to look for the parent distribution that might have produced the data.

2.3 Probability Paper Plot

The idea behind probability paper plot (PPP) is to graphically check whether our sample could have come from the referenced distribution or not. The plot will look linear in case the sample matches the referenced distribution. A departure from linearity is a clear indication that the sample is not well approximated by the suggested distribution. Here we explain how does it work, then we apply it to the samples of traffic data. More details about this method can be found in [5].

2.4 Gumbel Plot

Gumbel plot is probability plot where the reference distribution is the Gumbel distribution. It is one of the most classical methods in extremes. It is a plot of the empirical distribution of the observed data against the theoretical quantiles of the Gumbel distribution. If the data come from Gumbel distribution then the plot will look linear, otherwise the plot shows a convex or concave curvature depending on whether data come from a distribution with a tail heavier than the Gumbel's or lighter, respectively. Gumbel plot method is close in spirit to the QQ and Probability plots. Gumbel plot method is also known as double logarithmic plot.

In Gumbel method, we plot the empirical quantiles versus the quantiles of the theoretical Gumbel distribution. The plot is given as
$\{X_{k,n}, -\ln(-\ln(p_{k,n})), k = 1,...,n\}$, where $p_{k,n} = (n-k+0.5)/n$ are plotting positions.

Four Gumbel plots are shown in Figure 2. These plots are based on the block maximum series of the data. Block maximum of internal LAN (second column) shows a concave curve deviation from straight line (theoretical Gumbel quantiles). This concave deviation suggests block maxima distribution with a lighter tail than Gumbel's. Meanwhile, the two other plots in the first column have a convex shape. This convex departure from linearity suggests block maxima traffic that follows an extreme Frechet distribution. These statements are to be analyzed and confirmed by estimating model parameters. Remember that we are talking about the distribution of block maxima and not the whole traffic data.

2.5 Mean Excess Function

The mean excess function of a probability distribution is given as $e(u) = E(X-u \mid X > u), 0 \le u < x_f$, where u is a given threshold and is the support or right end point of the distribution.

Mean excess function is used under different names in different disciplines. In insurance, it is the expected claim size in the unlimited layer; in finance, it is the shortfall; in reliability, it is the mean residual life.

Mean excess plot (meplot) is based on the mean excess function. It is a useful visualization tool. It helps in discriminating in the tail of traffic data. If traffic data comes from a distribution with a heavier tail than Gumbel's, then the plot will look linearly increasing. If data come from a distribution with a lighter tail than the Gumbel, then the mean excess plot will be linearly decreasing.

As a graphical tool we use the mean excess plot, it is based on the estimate of mean excess function. Suppose that we have $X_1,...,X_n$, the sample mean excess is given as

$$e_n(u) = 1/F(u)\int_u^{\infty} F_n(y)dy$$

The graph $\{X_{k,n}, e_n(X_{k,n})\}$ is called the mean excess plot.

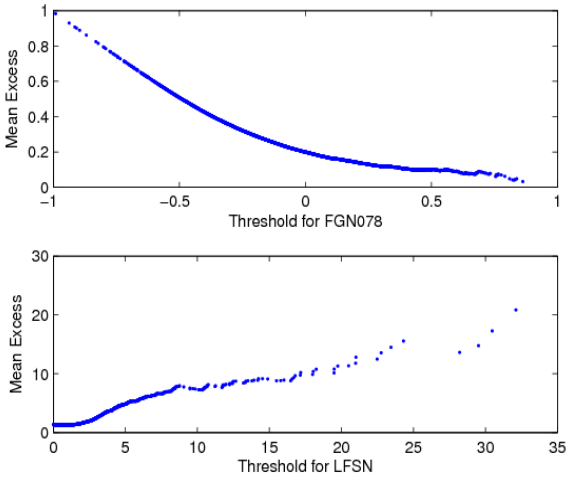

Fig. 3. Mean excess plots for FGN078(upper one) and LFSN

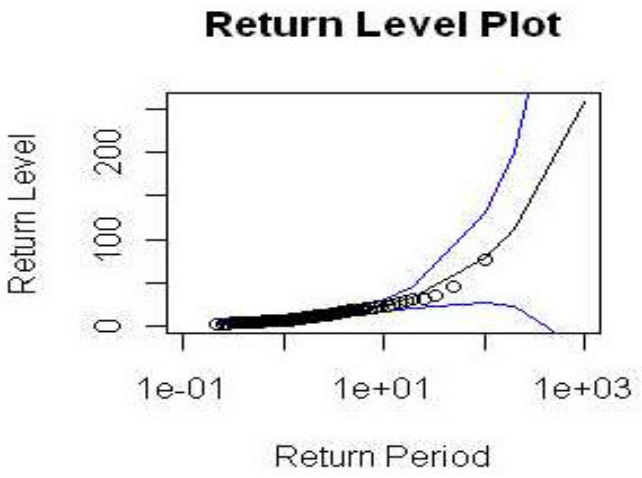

Fig. 4. Return level plot for LFSN.

Putting it together, we listed meplots for both internal and external simulated samples in Figure3. From the Figure, we see that the internal traffic has a decreasing meplot (upper one), while the external traffic has a linear increasing curve. These observations are in conformity with the theoretical Mean Excess function. The decreasing plot suggests a distribution with light tail. In the case of modeling with GEV, this behavior is analogous to a GEV with a negative shape parameter. The increasing plot suggests a distribution with heavier tail than Gumbel's. This increasing meplot is typical for a GEV with a positive shape parameter

3. MODEL PARAMETERS ESTIMATION

In general, to fit a model to the given data we need to estimate the parameters of that particular model, GEV model is no exception [3]. We chose the maximum likelihood (ML) method. ML method is implemented in the function *gev* from EVIM package [7]. We used blocks of size 100; this choice gives us maxima in blocks of 10 seconds each. We estimated model parameters using ML method.

From estimation results, we realized that internal traffic traces have negative shape parameter, while the two external traces have a positive shape parameter. This is an important observation since the shape parameter determines the type of fit distribution. In the internal traffic the negative shape parameter is an indication of the Weibull model. In external traffic case, the positive shape parameter is a clear indication of the traffic block maxima following an extreme Frechet distribution. Norros arrived at a similar conclusion for the internal traffic [14]. He showed analytically that the behavior of a buffer fed with internal traffic, approximated by fractional Gaussian noise, follows Weibull distribution. We arrived at a similar result but using extreme value theory perspective, which is much appropriate to the bursts prediction problem

4. PREDICTION

Apart from the model and its estimated parameters, we may use the powerful return level plot for the purpose of prediction. Return level plot is a plot of return period versus return level. We have produced a return level plot for the LFSN dataset with maxima taken in block of 100 observations. See Figure 4.

5. RESULTS AND ANALYSIS

Using EVT based model selection tools and parameters' estimates, we classify traffic bursts (block maxima) as coming either from a GEV distribution with positive shape parameter or GEV with a negative shape parameter, depending on the type of traffic.

From the EVT-based exploratory tools we notice that records of both kinds of traffic are within the confidence bounds of records from i.i.d data. This allows us to model the data not differently from i.i.d case. The max/sum ration also is a good tool for discriminating the two types of traffic (internal/external). For the external traffic, since the plot for the second moment does not approach zero quickly we say the variance is infinite. From the right column plot, we say that the mean is finite. This is also an obvious observation since the plot of the ratio for the first moment tends to zero quickly. Same thing can be said on the mean of the internal traffic. However, unlike the external traffic, the internal traffic has finite variance.

The internal traffic, which is approximated by fractional Gaussian model, has traffic maxima that follows GEV model with negative shape parameter. Put differently, those block maxima fall in the Weibull domain of attraction.

The external traffic is best approximated by linear fractional stable noise (LFSN). This LFSN traffic model has block maxima that follow a GEV with a positive shape parameter. The same way as for internal traffic, this GEV with positive shape parameter means that bursts or traffic maxima fall in the domain of attraction of Extreme Frechet model.

Using extreme value distribution we are able to model the bursts or traffic maxima. This modeling of bursts is a valuable one. It permits us, in the context of Quality of Service, to define new extreme traffic metrics to be included in the Service Level Agreement (SLA) contracts. This will help service providers to deliver a better service to their clients, and the latter to ask for the service they actually need.

6. CONCLUSION

We used extreme value theory to the bursts prediction problem. In particular, we used Generalized Extreme Value model. Using EVT based model selection tools, traffic is seen to be naturally classified into external and internal traffic, following GEV model with positive and negative shape parameter, respectively. Model selection tools have further improved our confidence in the model. EVT as a framework is better suited than any other analytical model to predict bursts and serious deterioration in network traffic. It is computationally less expensive and requires minimum disk space. In future, we should consider assessing the fit of the model with some quality of fit measures.

7. REFERENCES

1. Patrice Abry and Fabrice Sellan. *"The wavelet-based synthesis for fractional Brownian motion proposed by F. Sellan and Y. Meyer: Remarks and fast implementation"*. Applied and Computational Harmonic Analysis, 3(4):377-383, 1996.

2. N. Balakrishnan, Jose Maria Sarabia, Enrique Castillo and Ali S. Hadi. *"Extreme Value and Related Models with Applications in Engineering and Science"*. Wiley, Wiley, 2004.

3. George Casella and Roger L. Berger. *"Statistical Inference"*. Duxbury, Pacific Grove, CA 93950, USA, 2001.

4. Abdelmahmoud Youssouf Dahab, Halabi Hasbullah and Abas Md Said, *"Predicting Traffic Bursts Using Extreme Value Theory".* In International Conference on Signal Acquisition and Processing, pp.229-233, 2009.

5. Paul Embrechts, Thomas Mikosch and Claudia Kl¨uppelberg. *"Modelling Extremal Events: for Insurance and Finance*". Springer-Verlag, London, UK,1997.

6. Hifang Feng and Yantai Shu. *"Study on network traffic prediction techniques"*. International Conference on Wireless Communications, Networking and Mobile Computing, 2005, 2(23-26 Sept. 2005):1041–1044, September 2005.

7. Ramazan Genay, Faruk Seluk, and Abdurrahman Uluglyagci. *"Evim: A software package for extreme value analysis in matlab*". Studies in Nonlinear Dynamics & Econometrics, 5(3):1080–1080, 2001.

8. A. Karasaridis and D. Hatzinakos. *"Network heavy traffic modeling using alpha-stable self-similar processes*". Communications, IEEE Transactions on, 49(7):1203–1214, Jul 2001.

9. M. R. Leadbetter and Holger Rootzen. "Extremal Theory for Stochastic Processes". The Annals of Probability. 16(2):431-478, 1988.

10. M. Li and S. C. Lim. "*Modeling network traffic using generalized Cauchy process"*. Physica A, 387(11): 2584-2594,15 April 2008.

11. M. Li and W. Zhao. *"Representation of a Stochastic Traffic Bound, accepted for publication"*. IEEE Transactions on Parallel and Distributed Systems, 2009.

12. Chunfeng Liu, Yantai Shu, and Jiakun Liu. "*Application of Extreme Value Theory to the Analysis of Wireless Network Traffic".* In Proceedings of IEEE International Conference on Communications, ICC 2007, Glasgow, Scotland, 24-28 June 2007.

13. Abuagla Babiker Mohd and Sulaiman bin Mohd Nor. "*Towards a Flow-based Internet Traffic Classification for Bandwidth Optimization".* International Journal of Computer Science and Security (IJCSS), 3:2, 146-153. 2009.

14. Ilkka Norros. *"On the use of fractional brownian motion in the theory of connectionless networks*". IEEE Journal of Selected Areas in Communications, 13(6):953–962, 1995.

15. Kihong Park and Walter Willinger. *"Self-Similar Network Traffic and Performance Evaluation".* John Wiley & Sons, New Yotk, USA, 2000.

16. Gennady Samorodnitsky and Murad S. Taqqu. *"Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance"*. Chapman & Hall, London SE1 8HN, 1994.

17. Stilian Stoev and Murad S. Taqqu. *"Simulation methods for linear fractional stable motion and farima using the fast fourier transform"*. Fractals, 12:2004, 2004.

18. Zoi Tsourti and John Panaretos. *"Extreme-value analysis of teletraffic data"*. Computational Statistics Data Analysis, 45(1):85 – 103, 2004. Computer Security and Statistics.

19. M. Uchida. *"Traffic data analysis based on extreme value theory and its applications"*. Global Telecommunications Conference, 2004. GLOBECOM'04. IEEE, 3:1418–1424 Vol.3, Nov.-3 Dec. 2004.

20. Wei Biao Wu, George Michailidis, and Danlu Zhang. *"Simulating sample paths of linear fractional stable motion"*. IEEE Transactions on Information Theory, 50:1086–1096, 2004.

21. Internet traffic Archive, "http://*ita.ee.lbl.gov*".

BIOGRAPHY

1. Abdelmahamoud Youssouf Dahab is a PhD candidate in Computer & Information Sciences Department, Universiti Teknologi PETRONAS.

2. AP Dr Abas Md Said is a lecturer in the Computer and Information Sciences Department at Universiti Teknologi PETRONAS. He obtained his Bachelor and Masters of Science from Western Michigan University, USA. Doctor of Philosophy from Loughborough University, UK. He specializes in Computer Graphics, Visualization, and Networks. His research interests are in Optimized Stereoscopic Rendering for Visualization. He has a number of publications.

3. Dr Halabi Hasbullah is a lecturer in the Computer and Information Sciences Department at Universiti Teknologi PETRONAS. He obtained his Bachelor of Science (Mathematics) from Universiti Malaya, Malaysia. His Masters of Science (Information Technology) from De Montfort University, UK. And PhD (Communications System), National University of Malaysia (UKM), Malaysia. He specializes in Network Communications. His research interests are Bluetooth network, Traffic engineering, Wireless sensor network (WSN), Vehicular ad hoc network (VANET), and IPv6. He has a number of publications.