©2006-2015 Asian Research Publishing Network (ARPN). All rights reserved.

INTERNET TRAFFIC ANALYSIS WITH GOODNESS OF FIT TEST ON CAMPUS NETWORK

Murizah Kassim, Mohd Azrul Abdullah and Maizura Mohd Sani
Faculty of Electrical Engineering, Universiti Teknologi MARA, UiTM Shah Alam, Selangor, Malaysia
Email: murizah@salam.uitm.edu.my

ABSTRACT

This paper presents an analysis of internet traffic flows in a campus university network. Internet traffic is collected at the backbones network where two important traffics which are inbound and outbound are collected in 7 days. Statistical analysis is performed to understand the characteristic of traffic population. Empirical cumulative distribution function (CDF) is evaluated and important statistical parameters are characterized. Goodness of Fit (GOF) test with Anderson-Darling (AD) estimation technique is used to identify the best fitted distribution model such as Normal, Lognormal, Weibull and Exponential distribution. Maximum Likelihood Estimator (MLE) is also measured that presents important Weibull distribution characteristics. Results present analysis characteristics on both AD and MLE techniques of the internet traffic are fits to Weibull Traffic model. AD techniques fits presents the value-p probability parameter and MLE fits presents Weibull scale and shape parameters. Correlation results between the two inbound and outbound traffic are presented and discussed. This results help in modeling new algorithms to model new Tele-traffic algorithm based on time to control both inbound and outbound traffic in a network.

Keywords: internet traffic, statistical analysis, cumulative distribution function, maximum likelihood estimation, anderson-darling, normal, lognormal, weibull, exponential.

INTRODUCTION

Traffic analysis is an important task to understand network capabilities and requirements to provide reliable network guarantees. In the past years, numerous traffic models proposed for understanding and analyzing the traffic characteristics of networks but some traffic model do not fit in modeling traffic effectively in the networks. This is due to changes of traffic model [1, 2]. Thus, understanding the characteristic of traffic model and which is best suited with the network implementation is important and it has become a vital task.

There are numerous traffic models that are used widely for traffic modeling with different categories of traffic models. Each model varies significantly from the other and suitable for modeling different traffic characteristics. Gaussian model has shown in [3] that normal distribution can be directly linked to the presence or absence of extreme traffic burst. However, Gaussian distribution is not appropriated to model the traffic demand in large-scale network [4]. It shows that traffic volumes to and from a node in the network are characterized by a lognormal distribution, which has a slower decay than a normal distribution. While in another research lognormal distribution is able to accurately statistical models for flow size and flow duration of traffic application [5]. The techniques and real traffic parameter evaluation, yields changes in network performance. Observation of invariant heavy tails in access traffic patterns of individual users has motivated to investigate traffic transformation or aggregation as it traverses from access to core network [6].

A research has shown that flexible nature of Weibull distribution can capture this transformation at inter-arrival level [7]. How a superposition of heavy-tailed renewal streams models the scaling behavior of traffic at different access networks and tiers of internet hierarchy also is presented in previous study [8]. However, in [9] numerical results of throughput of network show that the network with exponential distributions of link capacities do not able to accommodate much more traffic as it is able for short range traffic. Exponential distribution is more suitable for non-long-tailed traffic data. This is a sharp contrast to commonly made modeling choices that exponential assumptions dominate and show only short-range dependence [10]. The distribution model has its own characteristic and the selection of the traffic model is not only considering on the type of traffic but also depends on the application. Due to this reason, an updates of traffic measures and scaling is very important.

This paper presents an analysis of internet traffic in a campus network for both inbound and outbound traffic flow. Internet traffic is collected at the backbones network in 7 days. Statistical analysis is performed to understand the characteristic of traffic population. Empirical cumulative distribution function (CDF) is evaluated and important parameters are characterized. This research presents Anderson-Darling (AD) estimation technique to identify the best fitted distribution model. AD estimation is also compared with Maximum Likelihood Estimator (MLE) with previous work [1]. Results present Both AD and MLE analyses of the internet traffic are fits to Weibull Traffic model. Weibull important parameters value which are scale and shape are identified and used to model a new tele-traffic algorithm called Time based Traffic Policing and Shaping to control burst traffic in a network. A new mathematical model on the development algorithms is presented which provides QoS in computer network.

MODELING TRAFFIC

Network performance is important for QoS. In order to keep the performance in constant, evaluation on traffic models and parameters should be defined to quantify the model at the optimum approached. The parameter of models defined must be related to the actual performance traffic measures which are to be predicted from the traffic model. Previous research presents various identified model on network traffic such as Poisson [11], Pareto [12], Self-Similar [13], Weibull [6, 14] and many more. Several distribution models are identified to model the measured traffic.

Normal distribution

The normal distribution is the most important and most widely used distribution in statistics. It is sometimes called the "bell curve," as the curve looks bell-shaped curve. It is also called the "Gaussian curve" after the mathematician Karl Friedrich Gauss. The probability density of normal distribution is as Equation 1 [15].

$$f(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad (1)$$

Where μ is the mean and σ is standard deviation. These parameters show the characteristic of normal distribution.

Lognormal distribution

The lognormal distribution has certain similarities to normal distribution. The lognormal distribution is very flexible model that can empirically fit many types of data. The lognormal distribution is used to model continuous random quantities when the normal distribution is skewed curve. The probability density of lognormal distribution with essential parameters μ mean and σ standard deviation is given in Equation 2[16]. The parameter σ is known as the shape parameter while μ is the scale parameter.

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

Weibull distribution

The Weibull distribution is given as in Equation 3 where β ≥ 0 is scale parameter and α > 0 is shape parameter. For The Weibull distributed process is heavy-tailed when β < 1 but it has moment finite. Thus it suitable for convergence modelling in heavy tailed multiplexing area. The Weibull distribution is a kind of exponential and Rayleigh distribution. The Weibull distribution can be used to model devices with decreasing failure rate, constant failure rate, or increasing failure rate. This versatility is one reason for the wide use of the Weibull distribution in reliability [17].

$$f(t) = \alpha\beta^{-\alpha} t^{\alpha-1} e^{-(t/\beta)^{\alpha}} \quad (3)$$

Exponential distribution

Exponential distribution is the probability distribution when modeling the time between independent events that happen at a constant average rate. The probability density function of an exponential distribution is as in Equation 4. Lambda, λ is rate parameter [18]. The exponential distribution is also related to the Poisson distribution and it is widely used to model waiting times.

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (4)$$

The distribution of data plotted using an empirical cumulative distribution function (CDF). The graphs of empirical CDF that is used to evaluate the fit of a distribution to collected data, estimate percentiles, and compare different sample distributions. An empirical CDF plot performs a similar function as a probability plot. However, unlike a probability plot, the empirical CDF plot has scales that are not transformed and the fitted distribution does not form a straight line. CDF is useful to approximate the true CDF if the sample size (the number of data) is large, and the CDF distribution is also helpful for statistical inference.

METHODOLOGY

The internet traffic data flow was monitored in a campus environment. The area of organization campus was selected to characterize the behaviour of internet flow at speed of 16 Mbps. SolarWinds software is setup at gateway router for internet traffic collection. Inbound and outbound internet throughput is collected in MByte. The traffic was measured every 10 minutes inter-arrival times. The transfer rate traffic of internet traffic was collected every day for 7 days. The study was performed on traffic collection with sample size n = 1108. In order to characterize the internet traffic, statistical analysis approach is used. The measurement traffic is analyzed to determine the most suitable analytic model. Several statistical models of probability distributions and estimate parameters based on the measured traffic were analyzed.

Goodness of fit with Anderson-Darling

The goodness of fit (GoF) of a statistical model describes how well it fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question. The GoF test is applied to measure the compatibility of the traffic samples with theoretical probability distribution function or empirical distribution of data population. Research present that there is no clear winner among all the GoF tests considered for all the parent distributions selected [19]. Probably this is related to the fact that the different tests behave better or worse depending on the specific distributions. But in the GoF test [20], Anderson-Darling (AD) statistical parameter is important to determine the fitted model as well as probability level (p-value). Anderson-Darling one of the most powerful Goodness of Fit test as more complex due

©2006-2015 Asian Research Publishing Network (ARPN). All rights reserved.

to introduction of a weight function in test statistic. The parameters of each traffic population were compared and the most suitable distribution is chosen. The distribution model with lower of AD value and higher p-value is selected. The Anderson-Darling test statistic is defined as in Equation 5 and Equation 6.

$$A^2 = -N - S \qquad (5)$$

$$S = \sum_{i=1}^{N} \frac{(2i - 1)}{N}[\ln F(Y_i) + \ln(1 - F(Y_{N+1-i}))] \qquad (6)$$

where F is the cumulative distribution function of the specified distribution. The Yi are the ordered data [17].

Maximum likelihood estimator

The population distribution use maximum likelihood (ML) estimation to estimate the distribution parameters which represent statistically characteristics. The ML estimation aims at determining the certain parameters based on distribution models to maximize the likelihood function. The ML estimation is likelihood function *L(θ)* as a function of *θ* and find the value of *θ* that maximizes it as defined in Equation 7.

$$\ell(\theta|x) = \frac{1}{n}\sum_{i=1}^{n} \ln f(x|\theta), \qquad (7)$$

Research presented an analysis using ML estimation on Internet traffic. Some parameters are modeled with similar accuracy using both Pareto and exponential distribution [21]. The distribution models come with estimate parameters which represent the characteristic of data population. For each distribution, statistical parameters were analyzed to find the best model that represents the data.

ANALYSIS AND RESULT

Data collection of 7 days with 10 minutes interval is analyzed.

Statistical analysis

Figure-1 shows heavy throughputs over time since first day to day 7. The trend of inbound traffic is increasing over time while stable for outbound traffic. The inbound traffic at 1200 MByte represent burst traffic exist in the traffic. It can be identified as bottleneck of the real network. Figure-2 shows the inbound traffic distribution is concentrated at 1000 MByte during day time. During night time between 12am to 8am, fewer throughputs observed. For outbound throughputs as shows in Figure-3, the distribution shows less throughput during night time while high throughput during day time. The inbound traffic can be adequately modeled by the selected distribution. Few types of distributions are tested to fit the collected traffic. For inbound traffic, four types of distribution are normal, lognormal, Weibull and exponential. Table-1 above is the statistics parameters of inbound and outbound traffic. With the number of sample size N is 1108, the inbound traffic distributed at mean μ is 940.69 and standard deviation σ is 420.06. As the distribution is not normal, the throughput is concentrated at 1022.9 MByte. While the outbound traffic distributed at mean μ is 336.1 with deviation σ 177.98.

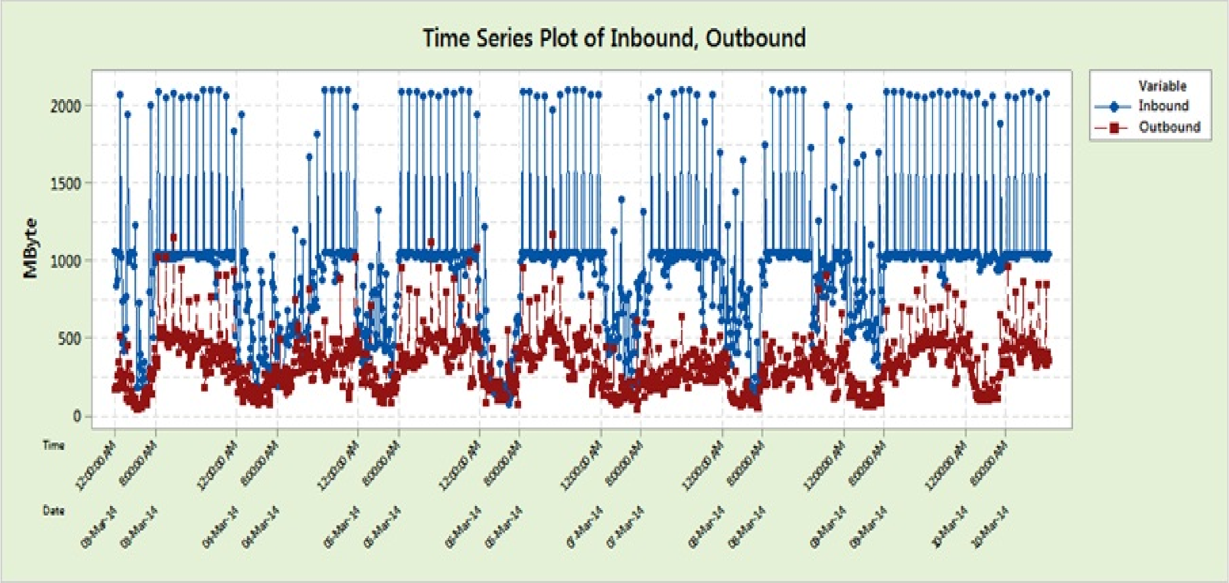

Figure-1. Time series plot of inbound and outbound.

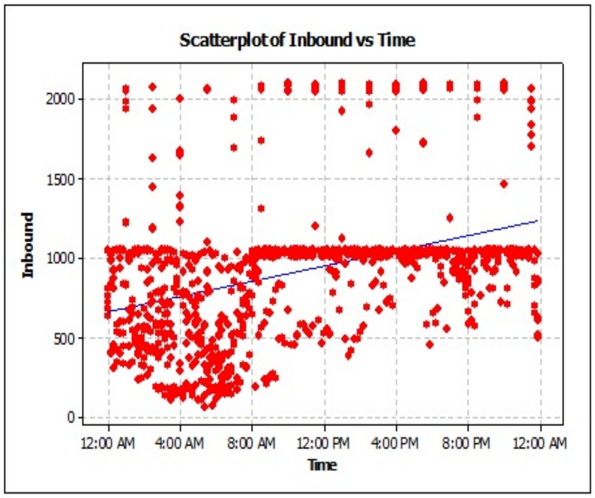

Figure-2. Scatterplot for inbound.

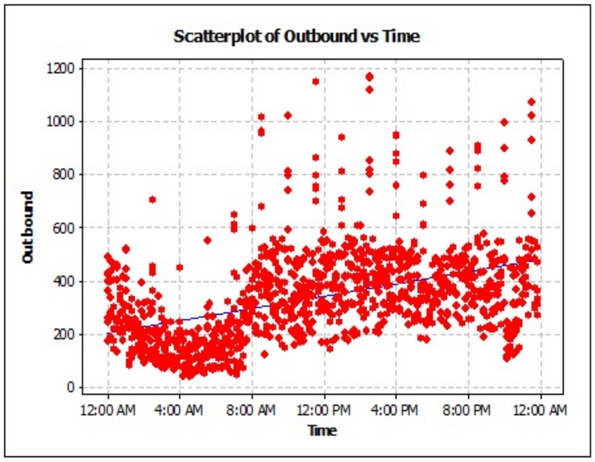

Figure-3. Outbound throughput.

Table-1. Descriptive statistics for inbound and outbound.

Traffic	N	Mean (μ)	StdDev (σ)	Median	Minimum	Maximum
In-B	1108	940.688	420.062	1022.9	69.123	2100.16
Out-B	1108	336.111	177.984	322.086	39.7602	1165.58

Anderson-Darling fit

The collected traffic is then tested with several analytic model distributions to measure how well the traffic follows a particular distribution. The better the distribution fits the traffic, the smaller this statistic p-value. Anderson-Darling statistic is used to compare the fit of several distributions to select the best or to test whether samples of traffic come from a population with specified distribution.

The hypotheses for Anderson-Darling test are:
H_0 : The data follow the specified distribution
H_1 : The data do not follow the specified distribution

Figure-4 shows the probability value (p-value) for the AD for Inbound traffic. The AD test presents lower than the chosen significance level (this case is 0.05), thus it is conclude that H_0 is rejected. It means the data do not follow the specified distribution. As in Figure-4 for Weibull, probability (p-value) of Weibull model is higher than the other distribution model and the distribution with the smallest AD statistic has chosen the Weibull distribution is the closest fit to the traffic. Table-2 shows the differences of P value among the four fitted distributions.

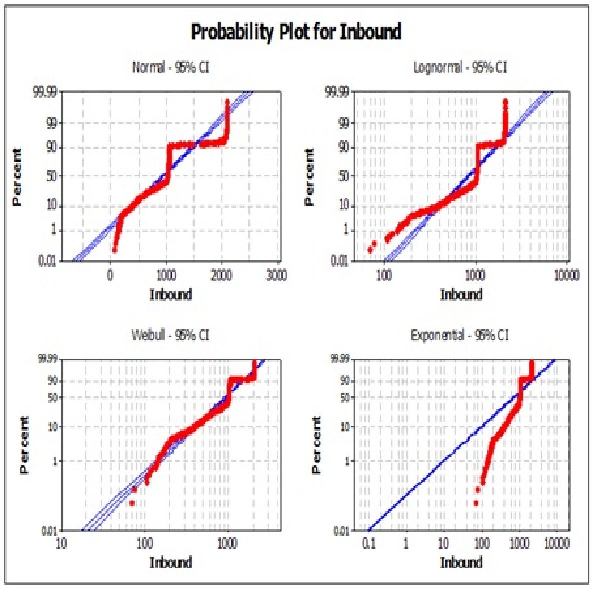

Figure-4. Goodness of fit test for inbound.

Table-2. Inbound probability value (p-value) for the AD.

Distribution	AD	P
Normal	76.170	<0.005
Lognormal	87.901	<0.005
Weibull	73.197	<0.010
Exponential	195.344	<0.003

Figure-5 shows the probability value (p-value) for the AD for Outbound traffic. Here, AD test also shows Weibull is the closest model fit to the outbound traffic with lowest AD is 4.479 and higher p-value <0.01. Table-3 shows the differences of P value among the four fitted distributions for outbound traffic.

©2006-2015 Asian Research Publishing Network (ARPN). All rights reserved.

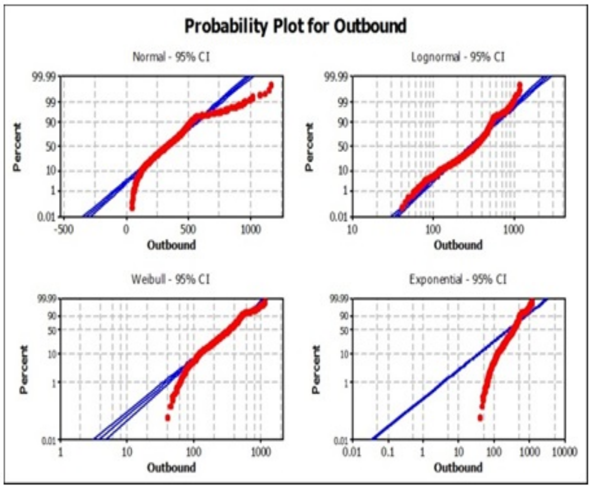

Figure-5. Goodness of fit test for outbound.

Empirical CDF graphs are used to evaluate the fit of a distribution to traffic or to compare different sample distributions. Each distribution provides the estimate parameter to understand the characteristic of traffic. The parameter values determine the location and shape of the curve and each unique combination of parameter values produces a unique distribution curve. Besides using AD to determine the closer distribution, Maximum likelihood estimates (MLE) is also tested to get the best parameter traffic to model one Tele-traffic algorithms called Time based Policing and Shaping.

Table-3. Outbound probability value (p-value) for the AD.

Distribution	AD	P
Normal	10.131	<0.005
Lognormal	14.77	<0.005
Weibull	4.479	<0.010
Exponential	118.548	<0.003

Maximum likelihood estimator fit

MLE is calculated by maximizing the likelihood function. The likelihood function describes, for each set of distribution parameters, the chance that the true distribution has the parameters based on the collected sample traffic. Figure-6 shows that the analytic model distribution fitted to the inbound traffic. At 50th percentile represents the median of the traffic which is at 1000 MByte. Table-2 shows the estimate parameters with the highest MLE. Weibull gives the highest MLE that conclude the closest distribution fit to the traffic. After tested, Table-4 shows the numerical results on Weibull important parameters which are shape and scale for the four best distributions.

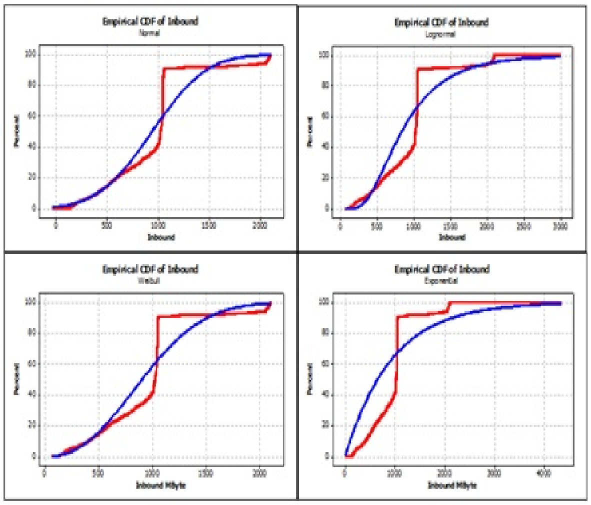

Figure-6. Empirical CDF of inbound.

Table-4. ML estimates of distribution parameter.

Distribution	MLE	Type	Estimate
Normal	-8264.449	Location (μ)	940.6876
		Dispersion (σ)	420.0619
Lognormal	-8358.604	Scale (μ)	6.724411
		Shape (σ)	0.549343
Weibull	-8241.539	Scale (α)	1058.523
		Shape (β)	2.33
Exponential	-8694.045	Scale (θ)	940.6876

Figure-7 shows the empirical CDF for outbound traffic and Table-5 presents the estimated parameters. The results show that Weibull is the best distribution model to fit the traffic for the internet data collection.

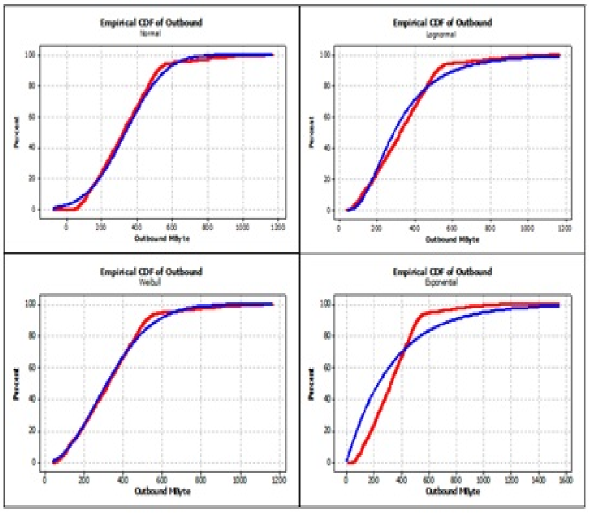

Figure-7. Empirical CDF of outbound.

Table-5. ML estimates of distribution parameter.

Distribution	MLE	Type	Estimate
Normal	-7313.001	Location (μ)	336.1109
		Dispersion (σ)	177.9841
Lognormal	-7258.341	Scale (μ)	5.664868
		Shape (σ)	0.587132
Weibull	-7228.407	Scale (α)	379.8954
		Shape (β)	1.9909
Exponential	-7553.725	Scale (θ)	336.1109

Traffic correlation

Figure-8 shows the correlation between the collected inbound and outbound traffic. The inbound traffic has linear correlation with outbound traffic. Pearson correlation of Inbound and Outbound is 0.668 and P-Value is 0.000. Since the p-value is smaller than 0.05, there is sufficient evidence at a = 0.05 that the correlations are not zero. This means that the two traffic variables for inbound and outbound are dependent with each other.

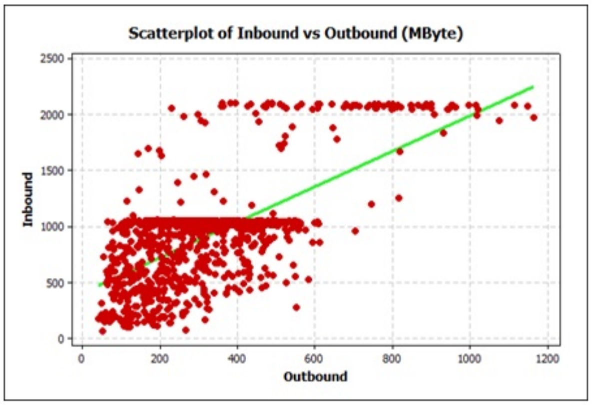

Figure-8. Correlation between inbound and outbound.

CONCLUSION

This paper presents an analysis with statistical measured using Anderson-Darling, Maximum Likelihood Estimator and the correlations between the inbound and outbound traffic for collected internet traffic. The distribution models have presented the analysed traffic on normal, lognormal, Weibull and exponential model. The statistical parameters have been established where mean, standard deviation and MLE are gathered. These characterized parameters have presented traffic population behaviour which is shape and scale. Empirical CDF is shown on real traffic distribution and comparisons are done between samples. Analysed traffic has shown Weibull is the best model fitted to the inbound and outbound traffic. Results also presents that the correlations between both traffic are not zero. This means that the two traffic variables for inbound and outbound are dependent with each other. The analysis is an important subject to understand the behaviour of internet traffic throughput in modeling Tele-traffic engineering algorithms for network communications.

REFERENCES

[1] M. Kassim, M. Ismail and M. I. Yusof. 2015. Statistical analysis and modeling of internet traffic IP-based network for tele-traffic engineering. ARPN Journal of Engineering and Applied Sciences, Vol. 10, pp. 1505-1512.

[2] H. A. H. Ibrahim, S. M. Nor and B. M. Khammas. 2014. Ambiguity and Concepts in Real Time Online Internet Traffic Classification," International Journal of Engineering & Technology (0975-4024), Vol. 6.

[3] R. De O. Schmidt, R. Sadre, N. Melnikov, J. Schonwalder and A. Pras. 2014. Linking network usage patterns to traffic Gaussianity fit," in Networking Conference, 2014 IFIP, pp. 1-9.

[4] K. Fukuda. 2008. Towards Modeling of Traffic Demand of Node in Large Scale Network. IEEE International Conference in Communications, pp. 214-218.

[5] C. Baogang, X. Yong, H. Jinlong and Z. Ling. 2008. Modeling and Analysis Traffic Flows of Peer-to-Peer Application. 3rd International Conference in Innovative Computing Information and Control, ICICIC '08., pp. 383-383.

[6] M. A. Arfeen, K. Pawlikowski, D. McNickle and A. Willig. 2013. The role of the Weibull distribution in Internet traffic modeling. Teletraffic Congress (ITC), 2013 25th International, pp. 1-8.

[7] M. Kassim, M. Ismail and M. I. Yusof. 2014. Adaptive throughput policy algorithm with weibull traffic model for campus IP-based network. Information Technology Journal, Vol. 13, pp. 2632-2644.

[8] M. A. Arfeen, K. Pawlikowski, A. Willig and D. McNickle. 2014. Internet traffic modelling: from superposition to scaling. Networks, IET, Vol. 3, pp. 30-40.

[9] S. Hosoki, S. Arakawa and M. Murata. 2010. A Model of Link Capacities in ISP's Router-Level Topology. Sixth International Conference on Autonomic and Autonomous Systems (ICAS), pp. 162-167.

[10] J.-S. Park, J.-Y. Lee and S.-B. Lee. 2000. Internet traffic measurement and analysis in a high speed network environment: Workload and flow characteristics. Communications and Networks, Journal of, Vol. 2, pp. 287-296.

©2006-2015 Asian Research Publishing Network (ARPN). All rights reserved.

[11] M. Becchi. 2008. From Poisson processes to self-similarity: a survey of network traffic models," ed: Technical Report.

[12] R. Singhai, S. D. Joshi and R. Bhatt. 2009. Offered-load model for Pareto inter-arrival network traffic," in IEEE 34th Conference on Local Computer Networks, LCN 2009. pp. 364-367.

[13] S. Sajeed, D. L. Kabir, M. L. Palash, N. Sultana and S. Rafique. 2010. An approach to measure the Hurst parameter for the Dhaka University network traffic. The 7th International Conference on in Informatics and Systems (INFOS), pp. 1-5.

[14] J. Shuhong, S. Abdalmajeed, L. Wei and W. Ruxuan. 2014. Totally Blind Image Quality Assessment Algorithm Based on Weibull Statistics of Natural Scenes. Information Technology Journal, Vol. 13.

[15] E. Weisstein. 2015. Normal Distribution Function. MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/NormalDistributionFunction.htm.

[16] E. Weisstein. 2015. Log Normal Distribution. MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/LogNormalDistribution.html.

[17] NIST/SEMATECH. 2003. e-Handbook of Statistical Methods: Weibull Distribution, http://www.itl.nist.gov/div898/handbook/eda/section3/eda3668.htm. Available: http://www.itl.nist.gov/div898/handbook/eda/section3/eda3668.htm

[18] E. Weisstein, "Exponential Distribution. 2015. MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/ExponentialDistribution.html.

[19] S. Guatelli, B. Mascialino, A. Pfeiffer, M. G. Pia, A. Ribon and P. Viarengo. 2004. Application of statistical methods for the comparison of data distributions," in Symposium Conference Record of Nuclear Science. IEEE, 2004, pp. 2086-2090 Vol. 4.

[20] A. W. Azim, S. S. Khalid and S. Abrar. 2013. Analysis of modulation classification techniques using Goodness of Fit testing. IEEE 9th International Conference in Emerging Technologies (ICET), pp. 1-6.

[21] A. Tudjarov, D. Temkov, T. Janevski and O. Firfov. 2004. Empirical modeling of Internet traffic at middle-level burstiness," Proceedings of the 12th IEEE Mediterranean Electrotechnical Conference, MELECON. pp. 535-538 Vol.2.