An Investigation of the Performance of Weibull Self-Similar Internet Traffic Model

J. Popoola[1]; W.B.Yahya[2]; J.O. Popoola[3]; O.R. Olaniran[4]

[1,2,4]Department of Statistics,
University of Ilorin, Ilorin, Nigeria.

[3]Department of Computer Engineering,
University of Ilorin, Ilorin, Nigeria
E-mail: jmkbalpop@unilorin.edu.ng[1]

Abstract **— The Markovian-memoryless assumption in classical traffic modeling most times is not applicable in traffics with concurrent arrivals. This is often evident in bursty Internet traffic, which leads to self-similarity or long-range dependency problems measured by the Hurst index. Modeling Internet traffic data in this situation requires the use of heavy-tailed distributions. In most reviewed work, the analysis in modeling the Internet traffic is based on simulation studies. In this paper, some of the existing distributions for fitting self-similar Internet traffic data on arrival and transmission times were reviewed, specifically with a real-life application study using a general form of the Weibull Internet traffic model. The empirical performance of the model was observed via real-life Internet traffic data of a corporate organization. Performance analysis of the proposed model alongside the light-tailed conventional Poisson/ Exponential model; heavy-tailed distribution of Gamma, Log-normal, Pareto; and another light-tailed distribution of Erlang was carried out using R/S statistics to estimate the Hurst index. The results from the analyses establish the Weibull distribution model as a possible and suitable distribution for fitting Internet traffic data.**

Keywords - *long range dependency, Self-similarity, heavy-tailed distribution, burstiness.*

I. Introduction

Internet traffic with self-similarity property occurs when packets of the same burst length arrive at the same time or when packets burst at the same inter-arrival period on the server with infinite variance syndrome. A process shows self-similarity if it is indistinguishable from its scaled versions, obtained by averaging the original process within different observation time scales.

A description of self-similarity can be summarised mathematically as follows; Assume an increment process $X_i (i = 1,2, \dots)$ and another process $X_j^m (i = 1,2, \dots)$ which is obtained by averaging the values in non-overlapped blocks of size m in X_i, i.e

$$X_j^m = \frac{1}{m}(X_{jm-m+1}, +X_{jm-m+2} + . \ . \ . + X_{jm}) \qquad (1)$$

The process X_i is said to be self-similar if

$$X_j^m \sim m^{H-1} X_i$$

which implies that

$$var(X_j^m) = m^{2H-2} var(X_i) \qquad (2)$$

where m $(m \geq 1)$ is the scale parameter whereas $H, (0.5 < H \leq 1)$ is the Hurst parameter. The Hurst parameter is used to measure the burstiness of a process.

A self-similar process can also contain a property of long-range dependence [1]. Long-range dependence explains the memory effect, where a present value strongly depends upon the past values of a stochastic process, and it is characterized by its autocorrelation function.

For $0 < H < 1, H \neq 1/2$ the autocorrelation function $r(k)$ for lag k is;

$$r(k) = H(2H-1)k^{-2H-2} \qquad (3)$$

The autocorrelation function in equation (3) decays hyperbolically, as k increases, which means that the autocorrelation function is not summable [2]. Short and long-range dependence have a common relationship with the value of the Hurst parameter of the self-similar process [3] [4]:

i. $0 < H < 0.5$ implies Short Range Dependence with exponential autocorrelation function decay
ii. $0.5 < H < 1$ implies Long-Range Dependence with hyperbolical autocorrelation function decay.

II. Research Methodology

A. *A Review of Some Existing Distributions for Modeling Self-similar Internet Traffic Model.*

(i) The Exponential Distribution

The probability density function of exponential distribution given by [5] is:

$f(x, \theta) = \theta e^{-\theta x}, x > 0$ (4)

and the Cumulative Distribution Function (CDF) is:

$F(x) = 1 - e^{-\theta x}$ (5)

where the θ is the intensity parameter of the process.

(ii) The Gamma Distribution

The Gamma distribution is also used in publications for the modeling of Internet traffic as explained by [6]. The PDF of the Gamma distribution is defined as follows:

$f(x, k, \theta) = \frac{x^{k-1}\theta^{k}e^{-\theta x}}{\Gamma_k}, k > 0\,; x > 0$ (6)

and the cumulative distribution function (CDF) on the support of X is:

$F(x, k, \theta) = \frac{\Gamma(\ k,\ \theta x\ )}{\Gamma_k}, k > 0\,; x > 0$ (7)

where $\Gamma(s, x) = \int_0^x t^{s-1}e^{-t}dt$

For $s > 0\,; x > 0$ is the incomplete Gamma function, k is the shape and θ is the intensity parameter of the distribution, and Γ is the Gamma function which has the formula;

$$\Gamma(s) = \int_0^\infty t^{s-1}e^{-t}dt$$

for $s > 0$ is the Gamma function

(iii) The Erlang Distribution

Thus, the PDF of Erlang distribution described by [7] is;

$f(x, k, \theta) = \frac{x^{k-1}\theta^{k}e^{-\theta x}}{(k-1)!}, k = 1,2,3, \ldots\ ; x > 0$ (8)

and the cumulative distribution function (CDF) is:

$F(x, k, \theta) = 1 - \sum_{i=0}^{k-1} \frac{e^{-\theta x}(\theta x)^i}{i!}, x > 0$ (9)

(iv) The Weibull Distribution

The Weibull PDF is determined by the parameters k and θ. For k = 1, Weibull distribution is identical to the exponential distribution [2]:

$f(x; k, \theta) = k\theta(\theta x)^{k-1}e^{-(\theta x)^k}; k > 0\,; x > 0$ (10)

and the cumulative distribution function (CDF) is:

$F(x; k, \theta) = 1 - e^{-(\theta x)^k}$ (11)

The parameters k and θ are the shape and the intensity parameter of the distribution respectively

(v) The Pareto Distribution

The Pareto distribution is heavy-tailed, it follows a power law over its entire range. The PDF of the Pareto distribution is given by [5] as :

$f(x) = ka^k x^{-k-1}, 0 < a \leq x$ (12)

and the cumulative distribution function (CDF) is:

$(x) = 1 - (a/x)^k$ (13)

where in equations (12) and (13), the constant a , represents the smallest possible value of the random variable, x, and, k is the shape parameter of the distribution.

(vi) The Log-normal Distribution

The PDF of the Log-normal distribution as described by [2] is given by:

$f(x; k, \theta) = \frac{1}{x\theta\sqrt{2\pi}}e^{-\frac{(Inx-k)^2}{2\theta^2}}\ ; x > 0$ (14)

There is no close form for the CDF of the lognormal distribution. Most statistical packages often provide approximate result.

The letter k and θ are the location and scale parameters of the distribution respectively.

B. Statistical Tests for Self-similarity

(i) The variance-time plot: It relies on the slowly decaying variance of self-similar tests. The variance of $X^{(m)}$ is plotted against m on a log-log plot; a straight line with slope (-β)>-1 is indicative of self-similarity, and the parameter H is given by the Hurst parameter (H=1-β/2).

(ii) The R/S plot: It uses the fact that for a self-similar dataset, the rescaled range or R/S statistic grows according to a power law with exponent H as a function of the number of points included (n). The plot of R/S against n on a log-log plot has a slope which is an estimate of H.

(iii) The Periodogram method: It uses the slope of the power spectrum of the series as frequency approaches zero. On a log-log plot, the periodogram slope is a straight line with a slope β = 1-2H close to the origin.

(iv) Whittle estimator: It is more enhanced in the sense that it provides confidence interval, but has the drawback that the form of the underlying stochastic process must be supplied. The most commonly used forms are fractional Gaussian noise and fractional ARIMA.

C. Data Description for Real-life Application

TCP network flow of University of Ilorin, Ilorin, with IP address written as 140.105.47.3 and aliased as www.unilorin.edu.ng was monitored for a month, as a sample network flow for this research. The sample flow generated secondary data for both packet arrival and transmission times, which were recorded in form of TCP connection format. A total of 1,255,981 request packets generated from several IP connectivities of users were observed between the periods of 19th April 2016 to 20th April 2016. The periods chosen represent periods of normal activities on the UNILORIN server that will not yield over-estimated or under-estimated results of the various analyses carried out on the arrival and transmission data in this study. The data was further sub-divided into peak and off-peak periods of usage of the website. The off-peak period data covered the time period of 6:27 pm of 19th April 2016 to 7:57 am of 20th, April 2016. A total of 169,628 request packets were collected for the off-peak period analysis. Similarly, the peak period data covered the time period of 8:27 am to 3:57 pm of 20th April 2016. A total of 1,086,363 request packets were collected for the peak period analysis.

III. Results And Discussion

A. Results

To examine the self-similarity nature, the Hurst index (H) was estimated using R/S statistics [8] The goodness-of-fit result in terms of log-likelihood, AIC and BIC were used for analysing both the arrival and transmission processes.

(i) Descriptive Statistics

Table 1: Descriptive statistics of arrival and transmission processes.

Period	Process	n	mean	sd	min	Max
Peak 8:27am – 3:57pm	Arrival	1086363	2.645489	21.50063	0	1801
	Transmission	1086363	1.424029	3.013069	0.000411	100.4988
Off peak 6:27pm – 7:57am	Arrival	169628	2.819626	249.0301	0	86400
	Transmission	169628	1.289142	2.445542	0.000391	100.6715

where;
n: numbers of observations;
sd: standard deviation;
min: minimum time between successive arrival and transmission;
max: maximum time between successive arrival and transmission;
Time unit: Seconds.

(ii) Hurst Index Result

Table 2: The Hurst index estimate and their Standard error (in parentheses) for the real-life data of Arrival and Transmission of internet packets.

Period	Number of Observations of request-packets	Arrival process	Transmission process
Peak 8:27am – 3:57pm	1,086,363	0.8636*** (0.2271)	0.7786*** (0.1584)
Off peak 6:27pm – 7:57am	169,628	0.7022*** (0.1154)	0.6814*** (0.0761)

*** Significant at 5% level.

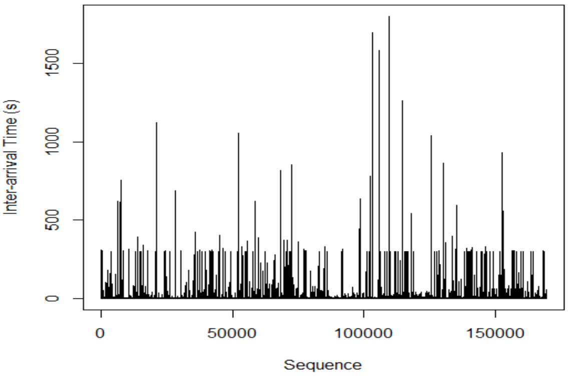

Fig. 1: Peak period Internet request-packet inter-arrival time showing self-similarity

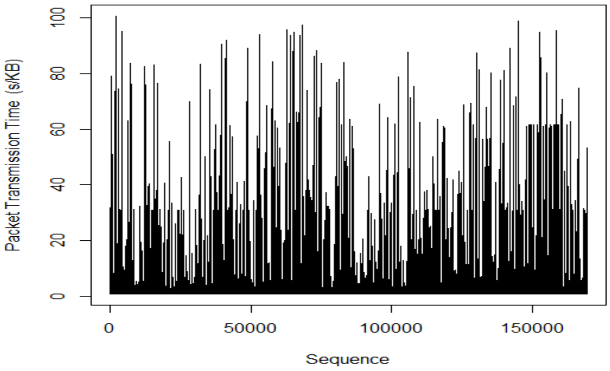

Fig. 2: Peak period Internet request-packet transmission time showing absence or little self-similarity

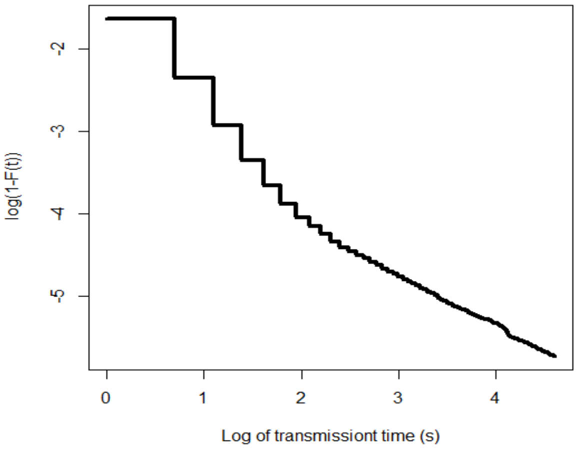

Fig. 4: Peak period log of complementary density function for Internet request-packet transmission time. The fast decay over time shows the absence or weak self-similarity as well as the absence of Long-Range Dependency in the transmission time.

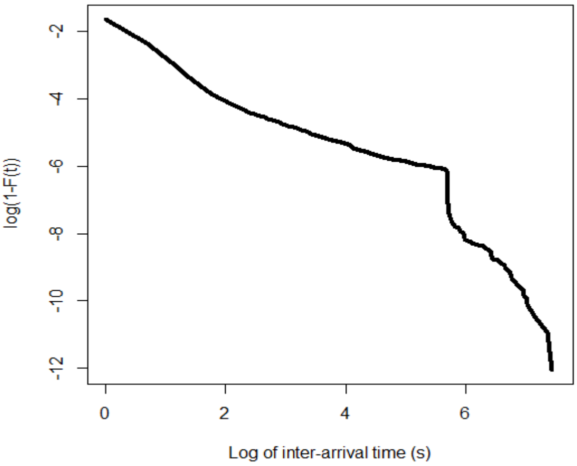

Fig. 3: Peak period log of complementary density function for Internet request-packet inter-arrival time. The slow decay over time shows the presence of self-similarity as well as Long Range Dependency in the inter-arrival time.

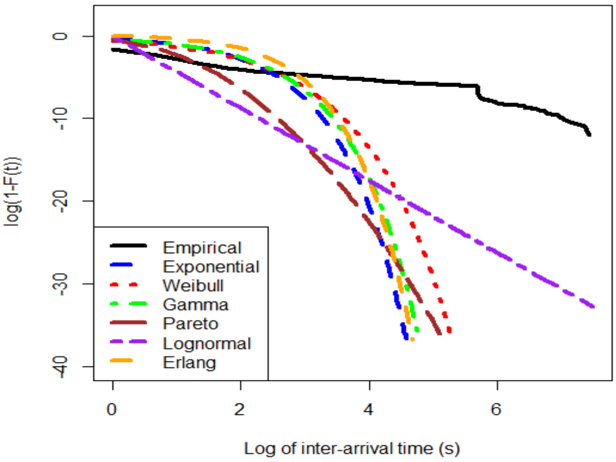

Fig. 5: Peak period empirical log of complementary density function for Internet request-packet inter-arrival time and possible light and heavy-tailed distributions complementary density function (Exponential, Weibull, Gamma, Pareto, Log-normal and Erlang)

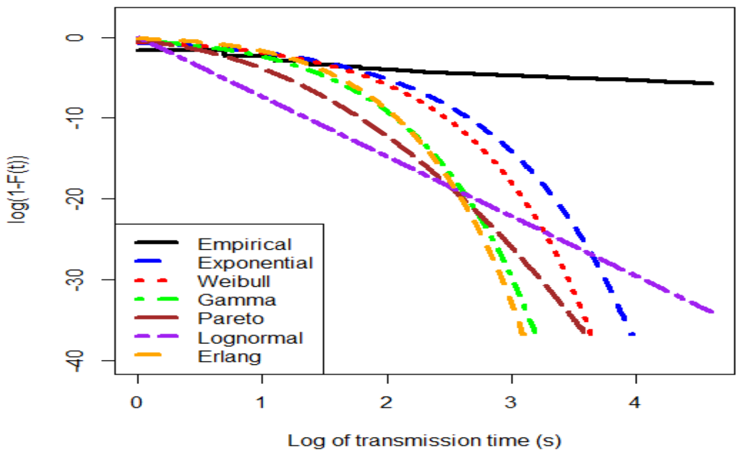

Fig 6: Peak period empirical log of complementary density function for Internet request-packet transmission time and possible light and heavy-tailed distribution complementary density function (Exponential, Weibull, Gamma, Pareto, Log-normal and Erlang)

Table 3: Goodness -of-fit measure for the peak period data

	Distribution	Parameters estimate	Log-likelihood	AIC	BIC
Arrival Process	Exponential	$\hat{\theta} = 0.3780$	-334651.6	669305.2	669315.2
	Weibull	$\hat{\gamma} = 0.7705$ $\hat{\theta} = 1.8533$	-303393.3	606790.6	606810.6
	Gamma	$\hat{\alpha} = 0.7976$ $\hat{\beta} = 0.3015$	-331547.9	663099.7	663119.8
	Log-normal	$\hat{\mu} = 0.2285$ $\hat{\sigma} = 0.6002$	**-192866.7**	**385737.3**	**385757.4**
	Pareto	$\hat{a} = 1$ $\hat{b} = 4.3762$	−420100.03	840180.05	840080.01
	Erlang	$\hat{k} = 1$ $\hat{\theta} = 0.3781$	-334651.6	669305.2	669315.2
Transmission Process	Exponential	$\hat{\theta} = 0.7022$	−229589.8	459181.7	459191.7
	Weibull	$\hat{k} = 1.1158$ $\hat{\theta} = 1.5010$	−226365.6	452735.1	452755.2
	Gamma	$\hat{k} = 2.4504$ $\hat{\theta} = 1.7207$	−199126.3	398256.5	398276.6
	Lognormal	$\boldsymbol{\hat{k} = 0.1358}$ $\boldsymbol{\hat{\theta} = 0.42\ 13}$	**−117100**	**2 342 04**	**2 342 2 14**
	Pareto	$\hat{a} = 1.0004$ $\hat{k} = 7.3873$	−246556.1	493110.2	493100.1
	Erlang	$\hat{k} = 3$ $\hat{\theta} = 2.1067$	−201216.9	402435.7	402445.8

Table 4: Goodness-of-fit measure for the off-peak period data.

	Distribution	Parameters estimate	Log-likelihood	AIC	BIC
Arrival Process	Exponential	$\hat{\theta} = 0.0846$	-3769420	7538843	7538855
	Weibull	$\hat{\gamma} = 1.03411$ $\hat{\theta} = 12.1100$	−3765044	7530092	7530116
	Gamma	$\hat{\alpha} = 4.0281$ $\hat{\beta} = 0.3408$	-3372155	6744314	6744338
	Log-normal	$\boldsymbol{\hat{\mu} = 2.34051}$ $\boldsymbol{\hat{\sigma} = 0.2\ 02\ 7}$	**-2350495**	**4700993**	**4701017**
	Pareto	$\hat{a} = 1$ $\hat{b} = 4.3762$	-743360.79	1486750.6	1486870.5
	Erlang	$\hat{k} = 1$ $\hat{\theta} = 0.4230$	-3386869	6773739	6773751
Transmission Process	Exponential	$\hat{\theta} = 0.7757$	-1362274	2724549	2724561
	Weibull	$\hat{k} = 1.2025$ $\hat{\theta} = 1.4016$	-1304011	2608027	2608051
	Gamma	$\hat{k} = 3.5114$ $\hat{\theta} = 2.7239$	-1024462	2048927	2048951
	Log-normal	$\boldsymbol{\hat{k} = 0.10487}$ $\boldsymbol{\hat{\theta} = 0.3374}$	**-475326.7**	**9 506 57.3**	**950681.1**
	Pareto	$\hat{a} = 1.0004$ $\hat{k} = 9.570435$	− 12534450	2506889	2506877
	Erlang	$\hat{k} = 1$ $\hat{\theta} = 3.1028$	−1029709	2059420	2059432

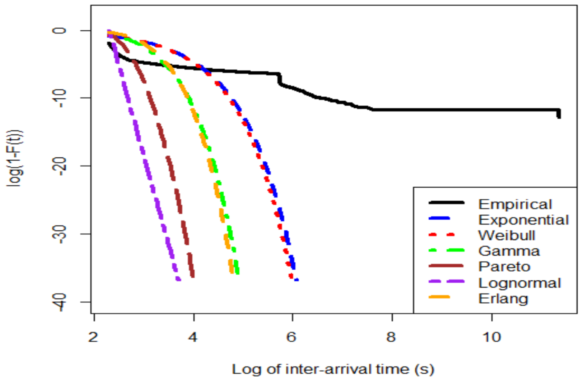

Figure 7: Off-peak period empirical log of complementary density function for Internet request-packet inter-arrival time and possible light and heavy tailed distribution complementary density function (Exponential, Weibull, Gamma, Pareto, Log-normal and Erlang)

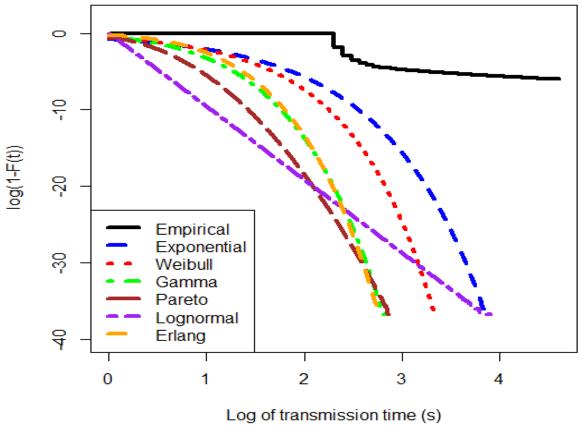

Figure 8: Off-peak period empirical log of complementary density function for packet transmission time and possible light and heavy tailed distribution complementary density function (Exponential, Weibull, Gamma, Pareto, Log-normal and Erlang).

.

IV. Discussion and Conclusion

A. Discussion

(i) The Peak Period Analysis: Table 1 below shows the descriptive summary in terms of mean, standard deviation, the minimum and maximum time between successive arrival and transmission. The minimum value, 0, in the arrival process indicates the possibility of concurrent arrivals; that is, some of the request packets arrived on the network at the same time, which suggests self-similarity in the arrival process. The absence of this structure in the transmission process indicates the possibility of a memoryless process. The validity of the data can be observed from the fact that irrespective of the partitioning, the maximum processing time is approximately 100s. Also, the stability of the network can be observed from the mean of the inter-arrival time being greater than mean transmission time. Hence, the network process monitoring is possible. The standard deviation, relatively larger than their respective means, suggests adequacy of skewed distribution for the arrival and transmission process.

To further examine the self-similarity nature, the Hurst index (H) was estimated using R/S statistics [8]. The Hurst estimates results in Table 2 reveals that the arrival process at the peak period is more self-similar than the transmission process while at the off-peak period, the self-similarity is low for both the arrival and transmission processes. The peak and off-peak periods partitioning used here represent low and high traffic. Tables 1 and 2 present the modeling results for the off-peak and peak periods performances for the traffic models at low intensity and wide differences in the performances at high intensity.

Table 3 shows the goodness-of-fit result in terms of log-likelihood, AIC and BIC for both arrival and transmission processes; the closest distribution to the empirical is Log-normal, giving the highest values of log-likelihood and lowest values of AIC and BIC but shows close values of Weibull competing well next to it and then other distributions considered in this study.

In Fig.5, the important point to note is the decay property; the closest to the empirical result in terms of decay is Log-normal. This implies that Log-normal is a highly probable distribution for fitting the self-similar *inter-arrival time* for University of Ilorin network flow in the peak period, Weibull is also a possible suitable distribution after Gamma followed by other distributions considered in this study.

In Fig.6, the important point to note is the decay property, the closest to the empirical result in terms of decay is Log-normal distribution but Exponential distribution also compete with it. Although, Log-normal and Exponential distributions are highly probable distributions for fitting the self-similar *transmission time* for University of Ilorin network flow at peak periods with a high inflow of users' request, Exponential is also a possible distribution at peak periods with a low inflow of users' request with Weibull as its competing and suitable distribution.

(ii)The Off-Peak Period Analysis: Table 4 shows that the closest distribution to the empirical is Log-normal, giving the highest values of log-likelihood and lowest values of AIC and BIC but shows values of Weibull distribution competing well next to Gamma followed by other distributions considered in this study.Fig.7 and 8 show that for both inter-arrival and transmission times, the closest to the empirical result in terms of decay is Exponential distribution and Weibull also competes next to it after Gamma, followed by other distributions considered in this study.

B. Conclusion

In this research work, based on the results obtained for the peak period, it was observed that Weibull, Gamma, Log-normal distributions efficiently modeled both the transmission and the arrival times of the Packet-switched Network with self-similarity and Long Range Dependency properties. Although Gamma and Lognormal distributions are more probable for the modeling in this study, Weibull distribution also performed well. .On the other hand, the off-peak period analyses have Exponential distribution as the most appropriate distribution for modeling the inter-arrival and transmission processes.

Further examinations of the behaviours of these statistical distributions on the real-life data for transmission and arrival times of Packet-switched Network, using Kolmogorov-Smirnov test re-affirmed the goodness-of-fit of the distributions on the arrival and transmission processes. All the graphical results from the cumulative distribution and the Q-Q plots of the empirical and theoretical distributions of the transmission and arrival times of Packet-switched Network gave more credence to the appropriateness of Lognormal and Gamma distributions with Weibull distribution as also a competing and suitable distribution for modeling the arrival and transmission processes in this study. Thus, Weibull distribution has been found to adequately fit both the arrival and transmission processes of packet-switched networks.

Acknowledgment

The authors appreciate the anonymous reviewers for their comments on the original draft of this manuscript.

References

[1] O. Sheluhin, S. Smolskiy and A. Osin, Self-similar processes in telecommunications, Hoboken, New Jersey: John Wiley & Sons, 2007.

[2] M. Fras, J. Mohorko and Ž. Čučej, "Limitations of a Mapping Algorithm with Fragmentation Mimics (MAFM) when modeling statistical data sources based on measured packet network traffic," *Computer Networks,* pp. 57(17), 3686-3700., 2013.

[3] K. Park and W. Willinger, Sef-similar Traffic and Performace Evaluation, NY: John Wiley & Sons, 2000.

[4] H. Yılmaz, "IP over DVB: management of self-similarity," Doctoral dissertation., Boğaziçi University, 2002.

[5] O. H. Alakiri, A. Oladeji, C. B. Benjamin, C. C. Okolie and M. F. Okikiola, "The desirability of pareto distribution for modeling modern internet traffic characteristics.," *International Journal of Novel Research in Engineering and Applied Sciences (IJNREAS),* 2014.

[6] P. Tran-Gia, D. Staehle and K. Leibnitz, "Source traffic modeling of wireless applications," *AEU-International Journal of Electronics and Communications,* pp. 55(1), 27-36., 2001.

[7] E. Brockmayer, "Table of Erlang's loss formula, The life and works of AK Erlang.," *Transanctions of the Danish Academy of Technical Sciences ,* pp. 268-275, 1948.

[8] B. B. Mandelbrot, "A multifractal walk down Wall Street. Scientific American, 280(2), 70-73," *Scientific American,* pp. 70-73, 1999.