Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2013, Article ID 471963, 7 pages
http://dx.doi.org/10.1155/2013/471963

Research Article

Normality of Ethernet Traffic at Large Time Scales

Zhiping Lu,[1] Ming Li,[1] and Wei Zhao[2]

[1] *School of Finance and Statistics, East China Normal University, No. 500 Dong-Chuan Road, Shanghai 200241, China*
[2] *Department of Computer and Information Science, University of Macau, Padre Tomas Pereira Avenue, Taipa, Macau*

Correspondence should be addressed to Zhiping Lu; camping9@gmail.com

Received 19 January 2013; Accepted 4 February 2013

Academic Editor: Carlo Cattani

We contribute the quantitative descriptions of the large time scales for the Ethernet traffic to be Gaussian. We focus on the normality property of the accumulated traffic data under different time scales. The investigation is carried out graphically by the quantile-quantile (QQ) plots and numerically by statistical tests. The present results indicate that the larger the time scale, the more normal the Ethernet traffic.

1. Introduction

The experimental research of the internet traffic (traffic for short), including the Ethernet one, exhibits that fractional Gaussian noise (fGn) may be a model in the sense of unifractal see, for example, [1–3]. This implies that traffic is Gaussian [4]. However, non-Gaussian models, such as stable processes, were also reported; see, for example, [5–8]. Therefore, the normality of traffic is an issue worth investigating.

Research described in [9, 10] revealed a scaling phenomenon of traffic. Taking into account the scales of traffic, we say that whether a traffic trace is Gaussian or not relies on time scales. Paxson and Floyd [10] and Feldmann et al. [9] claimed that traffic is Gaussian at time scales larger than 1 second. That property was qualitatively further confirmed by [11]. Note that real-traffic data used in [9, 10] were recorded in 1980s and 1990s, which are publicly accessible [12]. Thus, one second, as the critical time point, corresponds to the data in [12] and the infrastructure of the internet then.

Though the research exhibits that the statistics of traffic remain the same from the internet last century to the current years [13], the quantity of the critical time point, say one second, may be vague due to the development of high-speed networking. Therefore, when using the same data as those used in [1, 3, 9, 10], we use the concept of packet count, that is, the number of packets within an interval, to represent the number of bytes of packets within an interval.

Let $x(t(i))$ be a sample record of traffic time series, where $t(i)$ $(i = 0, 1, \ldots)$ is the series of time stamps, indicating the time stamp of the ith packet. The series $x(t(i))$ therefore represents the packet size of the ith packet at time $t(i)$. In this research, instead of using $x(t(i))$, we use $x(i)$ representing the packet size of the ith packet. On an interval-by-interval basis, therefore, the accumulated traffic, denoted by $y(n)$, is given by

$$y(n) = \sum_{i=nT}^{(n+1)T} x(i), \tag{1}$$

where T is the interval width, which also has the similar meaning of time scales. Thus, $y(n)$ stands for the accumulated bytes of arrival traffic in the nth interval. The statistics of $y(n)$ may considerably differ when T is small (small time scale) or large (large time scale) [1, 9, 10].

This research utilizes four real-traffic traces, listed in Table 1, which were measured on an Ethernet at the Bellcore Morristown Research and Engineering facility in 1989 [12]. (the originally statistical properties described in the early literature, e.g., [1, 3], turn to be ubiquitous in today's traffic, according to the research stated in [13]. Thus, the traffic trace, BC-Aug89, which was measured in 1989, keeps its value in the description of traffic pattern today).

Figure 1 illustrates four series of real-traffic trace BC-Aug89. Note that the statistics of $x(t(i))$ is consistent with that of $x(i)$, but we may obtain the time scale represented by T in (1), which is irrelevant of the networking speed. Let the interval width be $T = 1024$. Then, Figure 2 indicates $y(n)$ of BC-Aug89 for $T = 1024$.

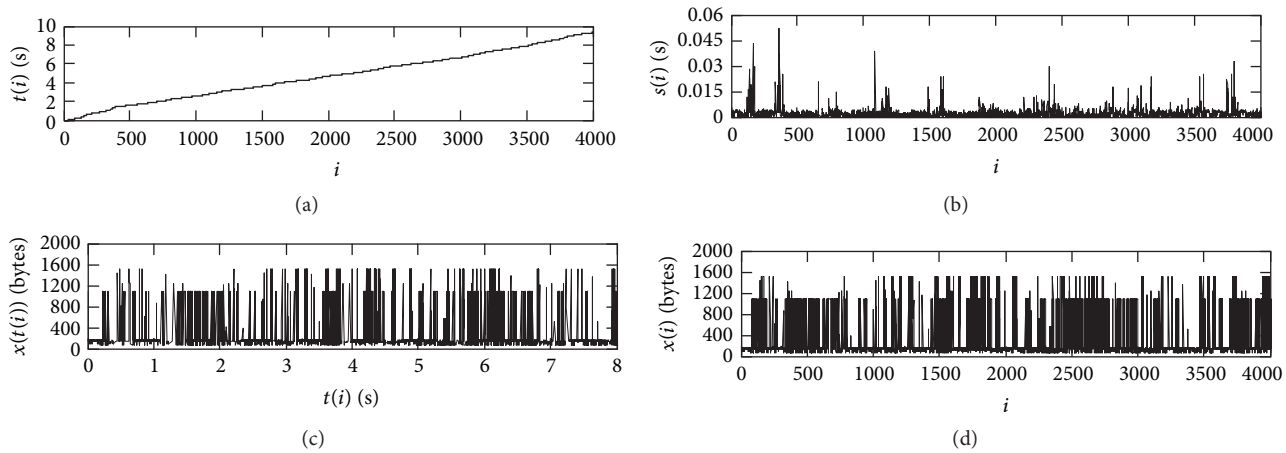

Figure 1: Illustrations of real-traffic trace BC-Aug89. (a) Timestamp series $t(i)$. (b) Interarrival times $s(i)$. (c) Traffic in packet size $x(t(i))$. (d) Traffic in packet size $x(i)$.

Table 1: Four traffic series.

Series name	Starting time	Duration	Series length
pAug.TL	11:25 AM, 29 Aug 89	52 minutes	1 million
pOct.TL	11:00 AM, 05 Oct 89	29 minutes	1 million
OctExt.TL	11:46 PM, 03 Oct 89	34.111 h	1 million
OctExt4.TL	2:37 PM, 10 Oct 89	21.095 h	1 million

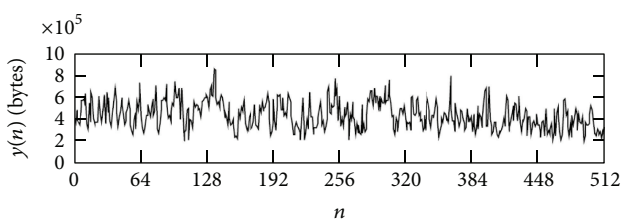

Figure 2: Accumulated traffic of BC-Aug89 with the interval width $T = 1024$.

The paper aims at presenting the quantitatively minimum interval range for the accumulated Ethernet traffic traces to be Gaussian based on the accumulated bytes of the packets within an interval.

The remainder of this paper is organized as follows. In Section 2 we introduce briefly the commonly used normality tests and the idea of the QQ plot. The graphical and numerical results are presented in Section 3, and the discussion of the investigation results is followed in Section 4. Section 5 concludes the paper.

2. Statistical Investigation for Accumulated Traffic

In this section, we discuss the normality tests for the following null and alternative hypotheses:

H_0: the data are sampled from a normal distribution;

H_1: the data are not sampled from a normal distribution.

Many statistical tests have been proposed to find out whether a sample is drawn from a normal distribution or not [14], including the Shapiro-Wilk test, D'Agostino's K^2 test, the Jarque-Bera test, the Anderson-Darling test, the Cramér-Von Mises criterion, the Lilliefors test, the Pearson's χ^2 test, and the Shapiro-Francia test.

The absence of exact solutions for the sampling distributions generated a large number of simulation studies exploring the power of these statistics. A convincing evidence from these studies is that convergence of the sampling distributions to asymptotic results was very slow. The paper [15] concludes that the Shapiro-Wilk test has the best power for a given significance, followed closely by Anderson-Darling test when comparing the Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors, and Anderson-Darling tests. On the other hand, some publications recommend the Jarque-Bera test [16, 17]. But it is not without weakness. It has low power for distributions with short tails. Therefore, we mainly consider three normality test methods listed in the following.

2.1. Shapiro-Wilk Test. The Shapiro-Wilk test tests the null hypothesis that a sample $y(1), \ldots, y(n)$ came from a normally distributed population [18]. The test statistic is

$$W = \frac{\left(\sum_{i=1}^{n} a_i y_{(i)}\right)^2}{\sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2}, \tag{2}$$

where $y_{(i)}$ is the ith order statistic; $\overline{y}_i$ is the sample mean; a_i is given by

$$\left(a_1, \ldots, a_n\right) = \frac{m^T V^{-1}}{\left(m^T V^{-1} V^{-1} m\right)^{1/2}}, \tag{3}$$

where $m = (m_1, \ldots, m_n)$; and $m_1, \ldots, m_n$ are the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution, and V is the covariance matrix of those order statistics. It is worth mentioning that the Shapiro-Wilk test is restricted for the sample size greater than 3 and less than 5000.

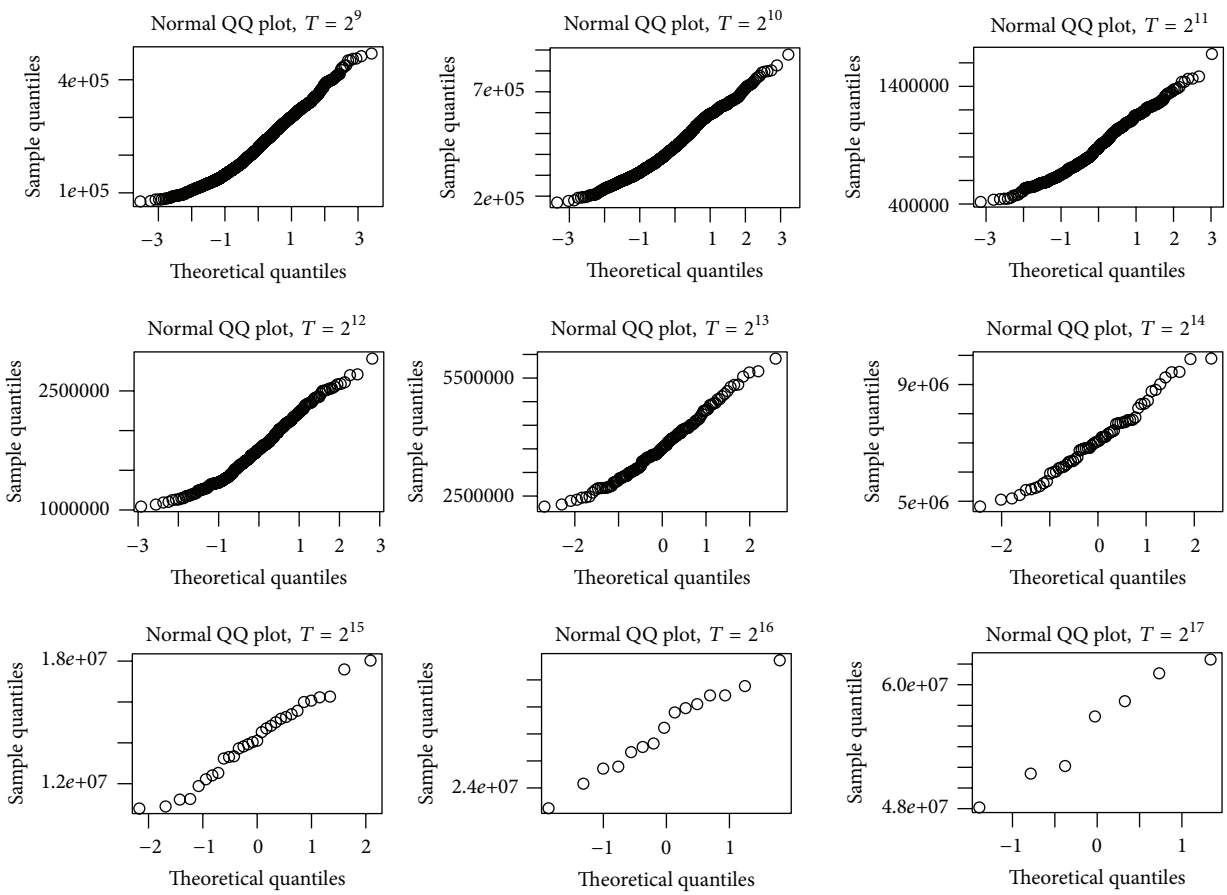

Figure 3: QQ-plot of the pAug.TL data under different time scales.

2.2. Anderson-Darling Test. The Anderson-Darling test is a statistical test of whether a given sample of data is drawn from a given probability distribution [19, 20]. In its basic form, the test assumes that there are no parameters to be estimated in the distribution being tested, in which case the test and its set of critical values are distribution free. When applied to testing if a normal distribution adequately describes a set of data, it is one of the most powerful statistical tools for detecting most departures from normality [21, 22], whereas the sample size needs to be greater than 7.

2.3. Jarque-Bera Test. The Jarque-Bera test is a goodness-of-fit test of whether sample data have the skewness and kurtosis matching a normal distribution [23, 24]. The test statistic JB is defined as

$$\mathrm{JB} = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right), \tag{4}$$

where

$$S = \frac{(1/n)\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^3}{\left((1/n)\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2\right)^{3/2}},$$
$$K = \frac{(1/n)\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^4}{\left((1/n)\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2\right)^{2}}. \tag{5}$$

If the data comes from a normal distribution, the JB statistic asymptotically has a $\chi^2(2)$ distribution, so the statistic can be used to test the hypothesis that the data is from a normal distribution. For small samples, the chi-squared approximation is overly sensitive, often rejecting the null hypothesis when it is in fact true. Thus, JB test only applies to large sample size, at least 7 according to the finite sample study.

Besides statistical tests, we have another informal but powerful tool to assess the normality property of the series, that is, the normal probability plot. This graphical tool is often called the quantile-quantile plot (QQ plot) of the standardized data against the standard normal distribution. The correlation between the sample data and normal quantiles measures how well the data is modeled by a normal distribution. For normal data, the points plotted in the QQ plot should fall approximately on a straight line, indicating high positive correlation.

3. Graphical and Statistical Results

In this section, we present the graphical and numerical results for all the Ethernet traffic series, that is, pAug.TL, pOct.TL, OctExt.TL, and OctExt4.TL data. Figures 3, 4, 5, and 6 are

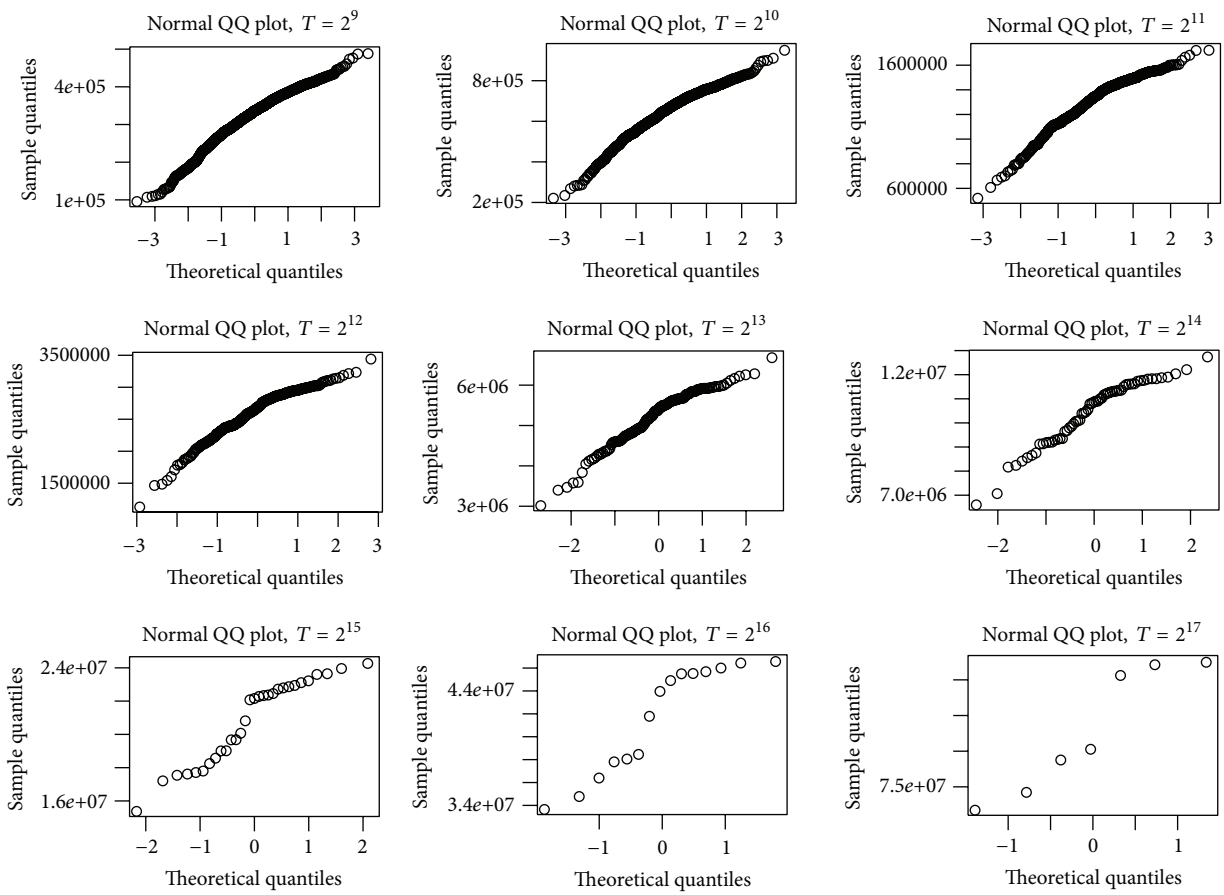

Figure 4: QQ-Plot of the pOct.TL data under different time scales.

the QQ-plot of the four accumulated traffic series under 9 different time scales.

In order to obtain a more complete inference for the series' normality and to be more objective, we finally choose to take advantage of three popular normality tests, that is, the Shapiro-Wilk test, Anderson-Darling test, and Jarque-Bera test to verify the normality property in the application. Based on the software R, we mainly utilize the functions of the packages "fBasics" and "nortest" to realize the statistical tests. The P-value of each test under the time scales $T = 2^n$, $n = 9, \ldots, 17$ are presented in Tables 2, 3, 4, and 5. In particular, since the Anderson-Darling test requires the sample size greater than 7, there is no testing result for the time scale $T = 2^{17}$.

Table 2: Normality test result for pAug.TL series.

Test	Shapiro-Wilk	Anderson-Darling	Jarque-Bera
$T = 512$	$3.222e-16$	$<2.2e-16$	$<2.2e-16$
$T = 1024$	$8.747e-10$	$2.945e-13$	$<2.2e-16$
$T = 2048$	$4.434e-06$	$3.977e-07$	0.004
$T = 4096$	0.0004069	0.001111	0.018
$T = 8192$	0.05681	0.1498	0.086
$T = 16384$	0.2819	0.5117	0.31
$T = 32768$	0.7258	0.8828	0.715
$T = 65536$	0.9528	0.8555	0.716
$T = 131072$	0.6796	N/A	0.524

4. Discussions

Graphically, from Figures 3, 4, 5 and 6, we have some findings listed below.

(i) Comparatively, the pAug.TL series asks for the relatively smallest time scale to be Gaussian among four series.

(ii) The pAug.TL and pOct.TL data seem more likely to be normal than the other two series at each corresponding time scale.

(iii) It is not difficult to observe that the OctExt.TL and OctExt4.TL series exhibit the similar normality behaviors. However, only at quite large time scale, the theoretical normal quantile and the empirical quantile have the high positive correlation.

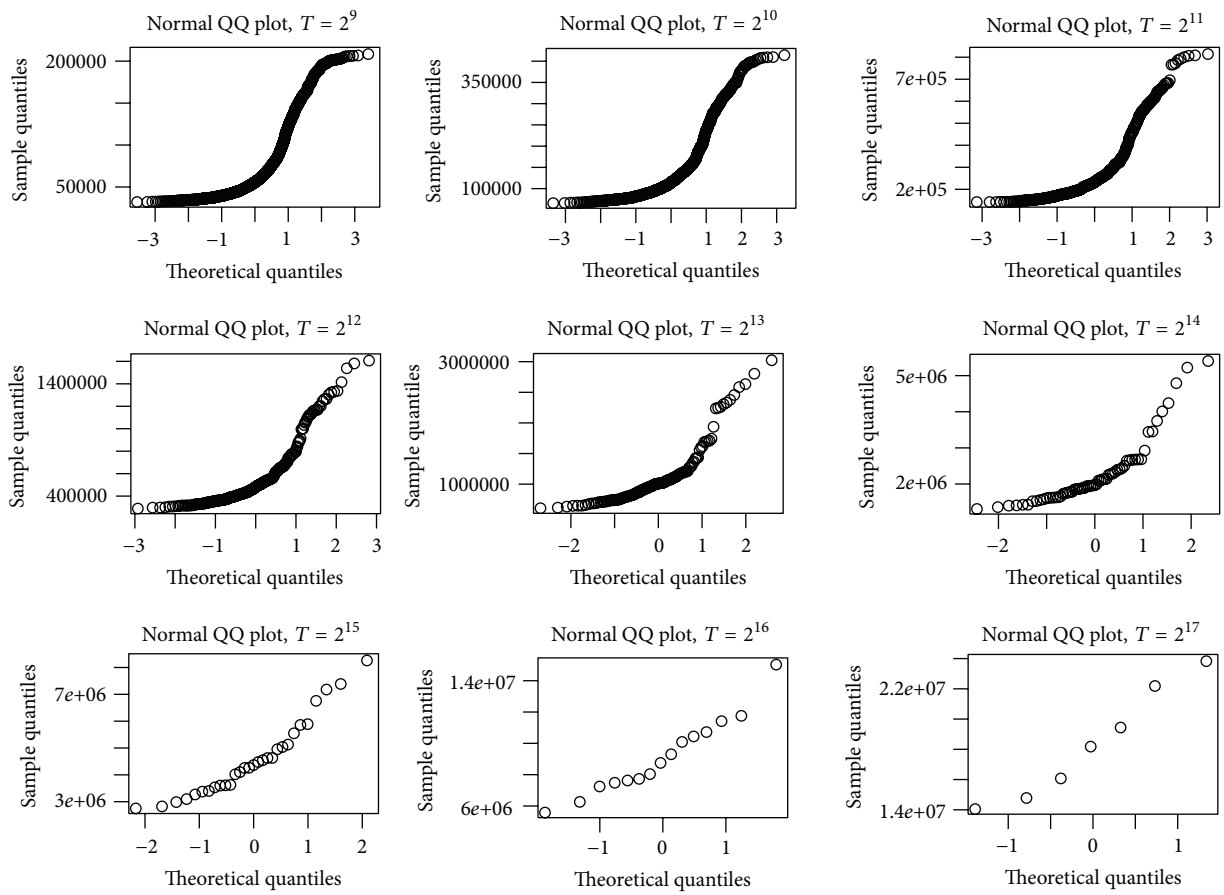

Figure 5: QQ-plot of the OctExt.TL data under different time scales.

Table 3: Normality test result for pOct.TL series.

Test	Shapiro-Wilk	Anderson-Darling	Jarque-Bera
$T = 512$	$<2.2e-16$	$<2.2e-16$	$<2.2e-16$
$T = 1024$	$4.112e-14$	$<2.2e-16$	$<2.2e-16$
$T = 2048$	$3.803e-11$	$2.125e-14$	$<2.2e-16$
$T = 4096$	$2.293e-07$	$2.801e-08$	$<2.2e-16$
$T = 8192$	0.0006598	0.0003278	0.015
$T = 16384$	0.006528	0.003969	0.051
$T = 32768$	0.01494	0.004604	0.118
$T = 65536$	0.02109	0.01588	0.139
$T = 131072$	0.1887	N/A	0.286

Table 4: Normality test result for OctExt.TL series.

Test	Shapiro-Wilk	Anderson-Darling	Jarque-Bera
$T = 512$	$<2.2e-16$	$<2.2e-16$	$<2.2e-16$
$T = 1024$	$<2.2e-16$	$<2.2e-16$	$<2.2e-16$
$T = 2048$	$<2.2e-16$	$<2.2e-16$	$<2.2e-16$
$T = 4096$	$<2.2e-16$	$<2.2e-16$	$<2.2e-16$
$T = 8192$	$6.15e-12$	$<2.2e-16$	$<2.2e-16$
$T = 16384$	$7.654e-08$	$4.916e-10$	$<2.2e-16$
$T = 32768$	0.02127	0.03261	0.048
$T = 65536$	0.5325	0.5585	0.208
$T = 131072$	0.6657	N/A	0.518

(iv) The OctExt4.TL series seems to be even more strict on the time scale. It requires minimum time scale about 65536 to be Gaussian.

Numerically, as could be expected, the testing results given in Tables 2, 3, 4, and 5 provide the evidence that the larger the time scale, the more normal the accumulated traffic series $y(n)$. Specifically,

(i) it is straightforward to see that the normality behavior of pAug.TL data "surpasses" the others according to the P values of the tests; that is, given the significance level $\alpha = 1\%$, the null hypothesis of normality could not be rejected when the time scale is greater than 8192;

(ii) whereas, the pOct.TL and OctExt.TL series possess the comparable normality performance who need the time scale to be at least 32768 in order not to be rejected by the null hypothesis given the significance level $\alpha = 1\%$.

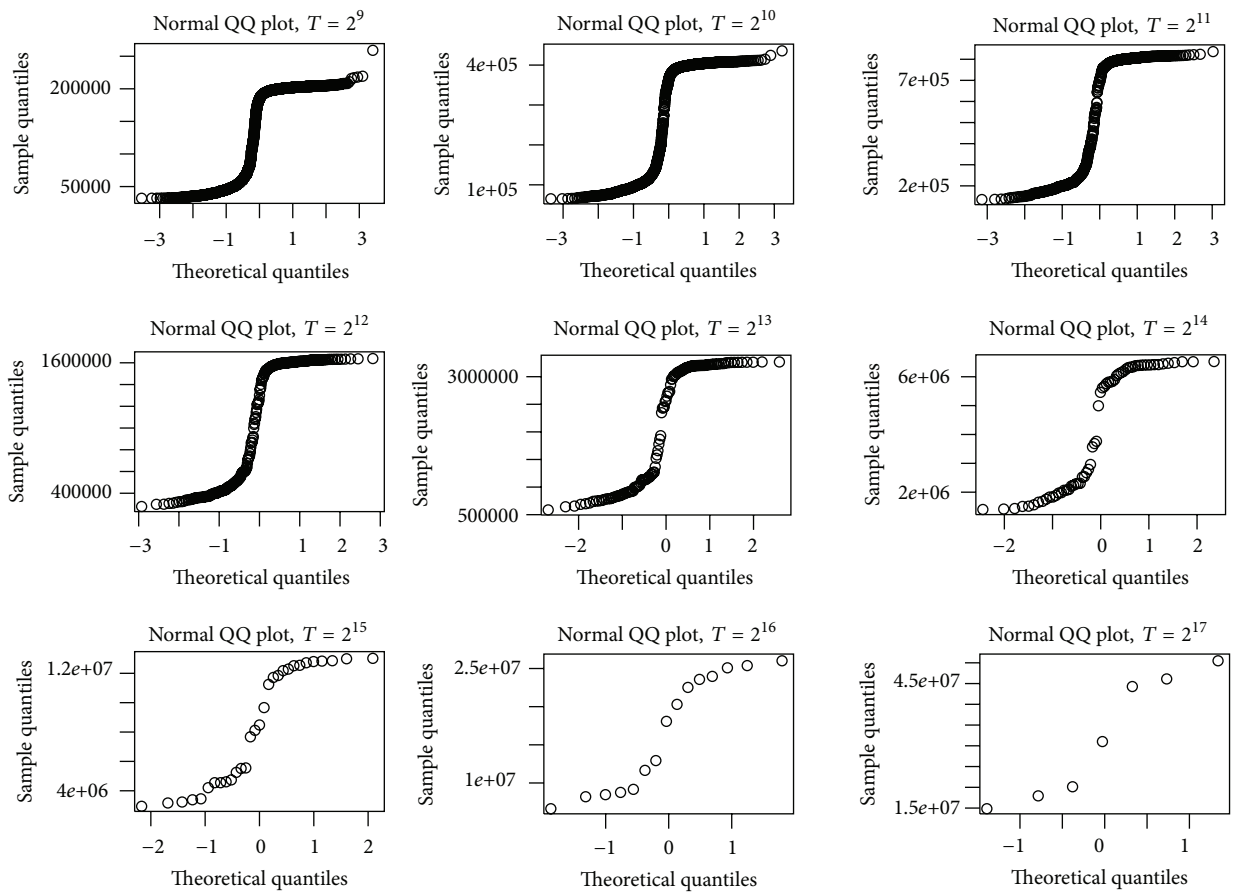

Figure 6: QQ-Plot of the OctExt4.TL data under different time scales.

Table 5: Normality test result for OctExt4.TL series.

Test	Shapiro-Wilk	Anderson-Darling	Jarque-Bera
$T = 512$	$<2.2e-16$	$<2.2e-16$	$<2.2e-16$
$T = 1024$	$<2.2e-16$	$<2.2e-16$	$<2.2e-16$
$T = 2048$	$<2.2e-16$	$<2.2e-16$	$<2.2e-16$
$T = 4096$	$<2.2e-16$	$<2.2e-16$	$<2.2e-16$
$T = 8192$	$6.398e-12$	$<2.2e-16$	0.005
$T = 16384$	$9.832e-08$	$8.577e-13$	0.021
$T = 32768$	0.0002948	$5.561e-05$	0.064
$T = 65536$	0.02198	0.02203	0.121
$T = 131072$	0.227	N/A	0.24

(iii) for the OctExt4.TL series, in order not to reject the null, the time scale should be greater than 65536 given the significance level $\alpha = 1\%$.

The previous discussions are for the Ethernet traffic, but the methods may also be a reference for other types of time series, such as those in [25–28].

5. Conclusions

We have discussed the normality performance of the Ethernet traffic data under different time scales using several normality tests (Shapiro-Wilk test, Anderson-Darling test, and Jarque-Bera test). The graphical results by QQ-plot are consistent with the numerical results, which also provides the evidence for the quantitative results of the large time scales for the normality of the Ethernet traffic traces investigated.

Acknowledgments

This work was in part supported by the 973 plan under the project grant number 2011CB302800, the National Natural Science Foundation of China under the project grant numbers 11101158, 61272402, 61070214, 60873264, and "the Fundamental Research Funds for the Central Universities". We appreciate W. Willinger, W. Leland, and D. Wilson with Bellcore, Morristown, who provided us with their data in this research.

References

[1] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Transactions on Networking*, vol. 2, no. 1, pp. 1–15, 1994.

[2] D. McDysan, *QoS & Traffic Management in IP & ATM Networks*, McGraw-Hill, New York, NY, USA, 2000.

[3] P. Abry and D. Veitch, "Wavelet analysis of long-range dependent traffic," *IEEE Transactions on Information Theory*, vol. 44, no. 1, pp. 2–15, 1998.

[4] W. Willinger and V. Paxson, "Where mathematics meets the internet," *Notices of the American Mathematical Society*, vol. 45, no. 8, pp. 961–970, 1998.

[5] Ph. Barbe and W. P. McCormick, "Heavy-traffic approximations for fractionally integrated random walks in the domain of attraction of a non-Gaussian stable distribution," *Stochastic Processes and Their Applications*, vol. 122, no. 4, pp. 1276–1303, 2012.

[6] R. G. Garroppo, S. Giordano, M. Pagano, and G. Procissi, "Testing α-stable processes in capturing the queuing behavior of broadband teletraffic," *Signal Processing*, vol. 82, no. 12, pp. 1861–1872, 2002.

[7] A. Karasaridis and D. Hatzinakos, "Network heavy traffic modeling using α-stable self-similar processes," *IEEE Transactions on Communications*, vol. 49, no. 7, pp. 1203–1214, 2001.

[8] G. Terdik and T. Gyires, "Lévy flights and fractal modeling of internet traffic," *IEEE/ACM Transactions on Networking*, vol. 17, no. 1, pp. 120–129, 2009.

[9] A. Feldmann, A. C. Gilbert, W. Willinger, and T. G. Kurtz, "The changing nature of network traffic: scaling phenomena," *ACM SIGCOMM Comput Communication Review*, vol. 28, no. 2, pp. 5–29, 1998.

[10] V. Paxson and S. Floyd, "Wide area traffic: the failure of Poisson modeling," *IEEE/ACM Transactions on Networking*, vol. 3, no. 3, pp. 226–244, 1995.

[11] A. Scherrer, N. Larrieu, P. Owezarski, P. Borgnat, and P. Abry, "Non-Gaussian and long memory statistical characterizations for Internet traffic with anomalies," *IEEE Transactions on Dependable and Secure Computing*, vol. 4, no. 1, pp. 56–70, 2007.

[12] http://www.sigcomm.org/ITA/.

[13] P. Borgnat, G. Dewaele, K. Fukuda, P. Abry, and K. Cho, "Seven years and one day: sketching the evolution of internet traffic," in *Proceedings of the 28th Conference on Computer Communications (INFOCOM '09)*, pp. 711–719, Rio de Janeiro, Brazil, April 2009.

[14] H. C. Thode, Jr., *Testing for Normality*, vol. 164 of *Statistics: Textbooks and Monographs*, Marcel Dekker, New York, NY, USA, 2002.

[15] N. Razali and Y. B. Wah, "Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests," *Journal of Statistical Modeling and Analytics*, vol. 2, no. 1, pp. 21–33, 2011.

[16] D. N. Gujarati, *Basic Econometrics*, McGraw-Hill, New York, NY, USA, 4th edition, 2002.

[17] G. G. Judge, R. C. Hill, W. E. Griffiths, H. Lütkepohl, and T. C. Lee, *Introduction to the Theory and Practice of Econometrics*, John Wiley & Sons, New York, NY, USA, 2nd edition, 1988.

[18] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality: complete samples," *Biometrika*, vol. 52, no. 3-4, pp. 591–611, 1965.

[19] T. W. Anderson and D. A. Darling, "Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes," *Annals of Mathematical Statistics*, vol. 23, pp. 193–212, 1952.

[20] T. W. Anderson and D. A. Darling, "A test of goodness of fit," *Journal of the American Statistical Association*, vol. 49, pp. 765–769, 1954.

[21] M. A. Stephens, "EDF statistics for goodness of fit and some comparisons," *Journal of the American Statistical Association*, vol. 69, pp. 730–737, 1974.

[22] M. A. Stephens, "Tests based on EDF statistics," in *Goodness-of-Fit Techniques*, R. B. d'Agostino and M. A. Stephens, Eds., pp. 97–193, Marcel Dekker, New York, NY, USA, 1986.

[23] C. M. Jarque and A. K. Bera, "Efficient tests for normality, homoscedasticity and serial independence of regression residuals," *Economics Letters*, vol. 6, no. 3, pp. 255–259, 1980.

[24] C. M. Jarque and A. K. Bera, "Efficient tests for normality, homoscedasticity and serial independence of regression residuals: Monte Carlo evidence," *Economics Letters*, vol. 7, no. 4, pp. 313–318, 1981.

[25] C. Cattani, G. Pierro, and G. Altieri, "Entropy and multifractality for the myeloma multiple TET 2 gene," *Mathematical Problems in Engineering*, vol. 2012, Article ID 193761, 14 pages, 2012.

[26] C. Cattani, "On the existence of wavelet symmetries in archaea DNA," *Computational and Mathematical Methods in Medicine*, vol. 2012, Article ID 673934, 21 pages, 2012.

[27] C. Toma, "Advanced signal processing and command synthesis for memory-limited complex systems," *Mathematical Problems in Engineering*, vol. 2012, Article ID 927821, 13 pages, 2012.

[28] E. G. Bakhoum and C. Toma, "Specific mathematical aspects of dynamics generated by coherence functions," *Mathematical Problems in Engineering*, vol. 2011, Article ID 436198, 10 pages, 2011.