Simulation of Web Data Traffic Patterns Using Fractal Statistical Modelling

Shanyu Tang, Hassan Kazemian

Simulation of web data traffic patterns using fractal statistical modelling

Shanyu Tang and Hassan B. Kazemian
Faculty of Computing, London Metropolitan University
166 – 220 Holloway Road, London N7 8DB, UK.
Email: s.tang@londonmet.ac.uk

Abstract. This paper describes statistical analysis of web data traffic to identify the probability distribution best fitting the connection arrivals and to evaluate whether data traffic traces follow heavy-tail distributions – an indication of fractal behaviour, which is in contrast to conventional data traffic.

Modelling of the fractal nature of web data traffic is used to specify classes of fractal computing methods which are capable of accurately describing the burstiness behaviour of the measured data, thereby establishing web data traffic patterns.

Keywords: data traffic, patterns, fractal modelling

1 Introduction

Data traffic is a vibrant area of queuing research. When modelling network traffic, network arrivals such as packet and connection arrivals are often modelled as Poisson processes for analytical simplicity because such processes have attractive theoretical properties [1]. However, a number of studies have shown that for both local-area [2, 3] and wide-area [4, 5] network traffic, packet interarrivals are clearly not exponentially distributed as expected from the Poisson processes. The distinctly non-Poisson nature of the arrival process has led to a substantial effort in developing statistical models of the traffic. These statistical models differ from the renewal and Markov-chain-type models that formed the basis for traditional queuing models [6], so they inspired new queuing models and solution methods.

Previous wide-area traffic studies were largely focused on FTP, TELNET, NNTP and SMTP (email) traffic [4-7], with little attention being given to web traffic [8]. The aim of this work is to tackle web traffic by providing a view of web data patterns. Since web traffic accounts for a large proportion of the traffic on the Internet, understanding the nature of web traffic is increasingly important. Like other wide-area traffic, web traffic may also have burstiness behaviour; i.e. web traffic is burst on many or all time scales. This burstiness behaviour is analogous to the self-similar or fractal-like behaviour, which exists in many natural phenomena such as Brownian motion, turbulent flow, atmospheric pressure, the distribution of stars and the activity of the stock market [9-13], which are much better characterised by fractal geometry theory than by Euclidean geometry.

Performance evaluation is important for assessing the effectiveness of data traffic pattern prediction, besides monitoring and verifying compliance with network performance goals [14]. Results from performance evaluation are used to identify

existing problems, guide network re-optimisation, and aid in the prediction of potential future problems. Performance evaluation can be conducted in many different ways. The most notable techniques include analytical methods, simulation, and empirical methods based on measurements.

Previous research found that the light- and heavy-tailedness of web data traffic patterns. The tail weight was estimated by using the Hill method in most of the cases presented in [15, 16, 17]. The method is a statistical technique for analysing heavy-tailed phenomena. The less heavy-tailed the data traffic, the more benefit of a buffer [18]. Adding multimedia files to the set of text files led to the increase in the weight of the tail and the distribution of text files might itself be heavy-tailed [16]. For light-tailed input, the delay distribution has an exponential tail; whereas the delay distribution follows a lognormal distribution for heavy-tailed input [19]. A hypothetical example [20] of a normal probability plot for data sampled from a distribution has been illustrated from where the light-tailedness and the heavy-tailedness of the data traffic pattern can be identified by visualization.

2 Characterising heavy-tail behaviour of web data traffic

Several probability distributions including conventional exponential and normal distributions have been chosen to fit the web traffic traces e.g. web document size, request interval and access frequencies collected at the measurement points. Previous work [5, 6] shows that web traffic such as telnet and ftp followed heavy-tail distributions. This means that values are distributed over a very wide range and that the larger values, even if less probable, may still account for a significant portion of the traffic. Several long-tailed distributions that were chosen to fit the traces are a lognormal distribution, an extreme distribution, a log-extreme distribution, a Weibull distribution and a Pareto distribution.

Empirical distribution function (EDF) test was also used to decide whether the data (traces) were from a particular distribution and best fittings were made using the EDF statistics. Parameters (e.g. location parameter, scale parameter, shape parameter) that affect in fitting the curve are highlighted to realize their effects. The analytical expressions of some distributions are detailed in the following sections.

2.1 Log-normal distribution

The lognormal law or log-Gaussian frequency function is defined as

$$\frac{d\phi}{d\log x} = \frac{1}{\sqrt{2\pi}\log\sigma_g}\exp\left[\frac{(\log x - \log x_g)^2}{2\log^2\sigma_g}\right]$$

where ϕ is the general term for the frequency, x is the web file size, x_g is the geometric mean of the distribution, σ_g is the geometric standard deviation, $\log\sigma_g = 0.5\log\left(\frac{x_{84}}{x_{16}}\right)$, and $x_g = \sqrt{x_{84}x_{16}}$, where x_{16} is the size corresponding to the 16% cumulative frequency, and x_{84} is the size corresponding to the 84% cumulative frequency.

2.2 Pareto distribution

The cumulative distribution for the Pareto (simple) random variable is obtained from

$$F(x) = P(X \le x) = 1 - \left(\frac{\alpha}{x}\right)^{\beta}; \quad \alpha, \beta \ge 0, \ x \ge \alpha$$

where α is the location parameter, and β is the shape parameter.

2.3 EDF test

The EDF is a step function calculated from the sample, estimating the population distribution function. EDF statistics are measures of the discrepancy between the EDF and a given distribution function, and they are used for testing the fitting of the sample to the distribution.

Suppose a given random sample of size n is $X_1, \ldots, X_n$ and $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ be the order statistics; suppose further that the distribution of x is $F(x)$. The empirical distribution function, $F_n(x)$, is given by

$$F_n(x) = \frac{number\ of\ observations \le x}{n}; \quad -\infty < x < \infty$$

The calculation can be conducted by using the Probability Integral Transformation (PIT), $Z = F(X)$; when $F(x)$ is the true distribution of x, the new random variable Z is uniformly distributed between 0 and 1. Then Z has the distribution function $F^*(z) = z$, $0 \le z \le 1$. The following formulas can be used for calculating EDF statistics from the Z-values. The formulas involve the Z-values arranged in an ascending order, $Z_{(1)} < Z_{(2)} < \ldots < Z_{(n)}$.

$$\left.\begin{aligned} W^2 &= \sum_{i=1}^{n} \{Z_{(i)} - (2i-1)/(2n)\}^2 + 1/(12n) \\ U^2 &= W^2 - n(\bar{z} - 0.5)^2; \quad where\ \bar{z} = sample\ mean = \frac{\sum_{i=1}^{n} Z_i}{n} \\ A^2 &= -n - \left(\frac{1}{n}\right)\sum_{i=1}^{n}(2i-1)\left[\log Z_{(i)} + \log\left\{1 - Z_{(n+1-i)}\right\}\right] \end{aligned}\right\}$$

where log x represents $\log_e x$ (natural logarithm).

The decision can be made by comparing the calculated value with the tabulated value [3]. The hypothesis is rejected at significance level p if the calculated value is greater than the tabulated value given for level p [21].

3 Data traffic collection

The first step in understanding network traffic is the collection of trace data. The LBL-CONN-7 trace [22] was collected at the Lawrence Berkeley Laboratory (LBL), located in Berkeley, California, containing thirty days' worth of all wide-

area TCP connections between the LBL and the rest of the world. The reduced trace was generated by TCP-reduce. TCP-reduce is a collection of Bourne shell scripts for reducing tcpdump traces to one-line summaries of each TCP connection present in the trace. The scripts are TCP-reduce, which takes a tcpdump trace file as an argument and writes a sorted summary to stdout, TCP-conn (an internal awk script that does all the work) and TCP-summary (an awk script that generates a per-protocol summary of all the TCP connections produced by TCP-reduce). The scripts were written using Bourne shell, tcpdump and the common Unix utilities sed, sort and awk. The trace was written as an ASCII file with one line per connection with the columns such as timestamp, duration, protocol, bytes sent by originator of the connection, bytes sent by responder to the connection, local host, remote host, state that the connection ended in and flags.

The trace ran from midnight, Thursday, 16 September 1993 through midnight, Friday, 15 October 1993 (times are Pacific Standard Time), capturing 606,497 wide-area connections. The tracing was performed on the Ethernet DMZ network over which flows all traffic into or out of the LBL. The raw trace was made using tcpdump on a Sun SPARC station using the BPF kernel packet filter. Fewer than 15 SYN/FIN/RST packets in a million were dropped. Timestamps had microsecond precision. The traffic was filtered to exclude connections with nearby UCB except for nntp.

A special care was taken to check the sanity of the data, as any irregularity or mistakes could set back the entire analysis for an extended period of time. The original data file was divided into smaller file and then the size (byte) of the file was arrayed into different bins for convenience of the compactness of the data file. It is very congenial to array the file size into a bin as thousands of samples can be analysed in an expected range.

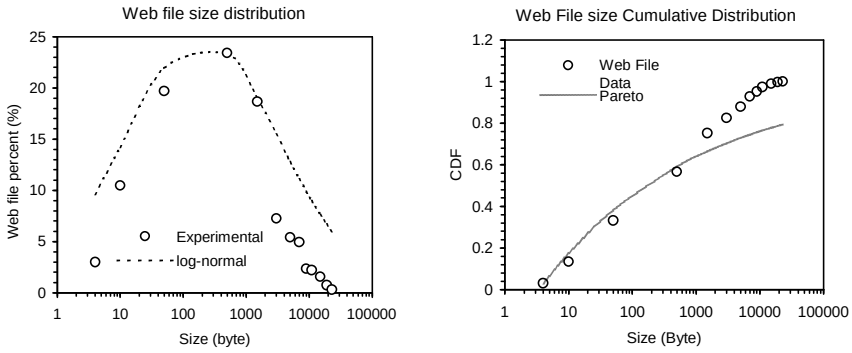
4 Results and Discussion

4.1 Traffic data in byte sent by originator

Fig. 1(a) shows the web file size distribution in percentage. The highest percentage of the file size transferred is 23.4% at 500 bytes whereas the lowest is 0.30% at 23,000 bytes. The lognormal distribution does not follow the trace data satisfactorily. The t-test for one sample (web file size) in two tails distribution found to be 3.4916, which is greater than the tabulated value (Table 1), indicating that the trace data distribution may be tailed. Fig. 1(b) illustrates comparisons of the cumulative distribution between the web file data and the data generated using Pareto model. The model data somewhat fit the web data within the range 10 to 1000 byte, which is an indication of burstiness behaviour.

Table 1. T-test results for web file size data

Traffic	t (Experimental)	Degrees of freedom	t0.05 (tabulated value)
LBL-CONN-7-originator	3.4916	11	2.201
LBL-CONN-7-responder	7.3395	12	2.179



(a) (b)

Fig. 1. Web file size distribution and Cumulative distribution of LBL_CONN_7 (byte sent by originator) data

Fig. 2 compares cumulative probability distributions of the actual trace data, the exponential, Pareto (generalized), Weibull (three parameters), logistic and extreme value distributions. The Pareto and Weibull models show the best fittings between the trace and model curves data, especially the trace data clearly follow Weibull distribution.

The experimental results show good agreements between the web data and Pareto and Weibull models, indicating that the web data distributions are heavily tailed. Such heavy-tailedness of data distributions is an indication of fractal-like burstiness behaviour of web data patterns.

Table 2. Test statistics (A2, W2 and U2) for LBL-CONN-7-originator data

Distribution	Test Statistics (TR = Test Result, α = significance level)		
	A2	W2	U2
Normal	TR = 0.6629 α = 0.05 Accept	TR = 0.1002 α = 0.10 Accept	TR = 0.0904 α = 0.10 Accept
Extreme value	TR = 0.5915 α = 0.10 Accept	TR = 0.0824 α = 0.10 Accept	TR = 0.08 α = 0.10 Accept
Weibull	TR = 2.906 α = OR Reject	TR = 0.1435 α = 0.025 Accept	TR = 0.1218 α = 0.025 Accept
Exponential	TR = 0.6955 α = 0.25 Accept	TR = 0.0807 α = 0.25 Accept	TR = 0.0748 α = 0.25 Accept
Logistic	TR = 0.6127 α = 0.05 Accept	TR = 0.09 α = 0.05 Accept	TR = 0.0851 α = 0.10 Accept
Pareto (generalized)	TR = 2.733 α = 0.001 Reject	TR = 0.1427 α = 0.025 Accept	TR= α = Accept/reject

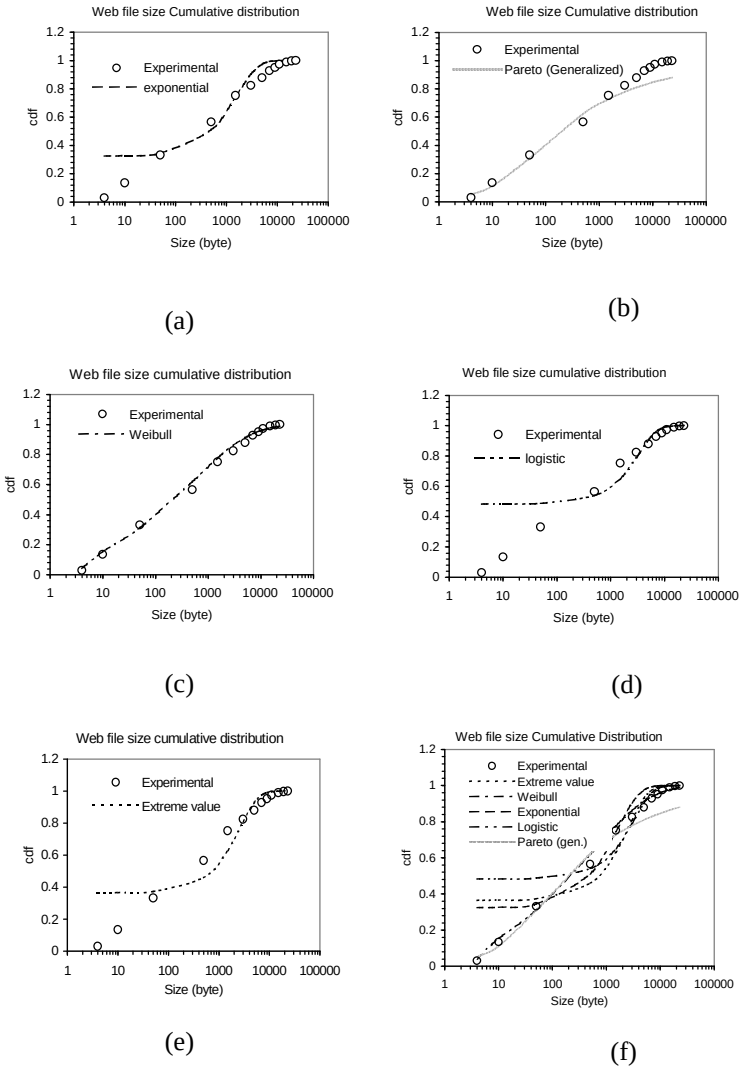
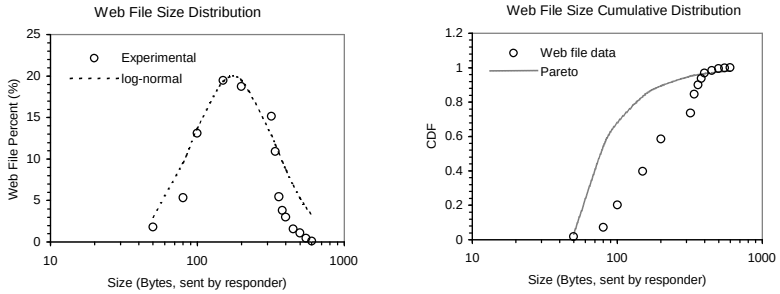


Fig. 2. Web file size cumulative distribution of LBL-CONN-7-originator data

The test statistics (A2, W2 and U2 tests) confirm (Table 2) that the trace data follow the normal distribution. There is contradiction between EDF test statistics and t-test as the t-test does not support the trace data for normality. The A2 test rejected the hypothesis for Weibull distribution whereas W2 and U2 tests show that the data are significant at the level $\alpha = 0.025$ which is consistent with the results shown in Fig. 2(c). The A2 test also rejected the hypothesis for generalized Pareto distribution where the data are significant at the level of 0.001, whereas W2 accepted the hypothesis at the level of 0.025 that makes sense with the graphical results shown in Fig. 2(b). The three test statistics show that the trace data are significant at $\alpha = 0.10$ and $\alpha=0.25$ for extreme value and exponential distributions respectively, also $\alpha = 0.05$ (shown by A2 and W2 tests) and $\alpha = 0.10$ (by U2 test) for logistic distribution. The three distributions could have shown better fitness than others according to significance level. Unfortunately, poor matches were observed as shown in Figs. 2(a), (d) and (e) which is questionable. Fig. 2(f) illustrates the combination of Figs. 2(a), (b), (c), (d), and (e) from where the best fitness can be deemed by comparison.

4.2 Traffic data in byte sent by responder

Fig. 3(a) shows the web file size distribution in percentage. The highest percentage of the file size transferred is 19.48% at 150 bytes whereas the lowest is 0.12% at 600 bytes. The lognormal distribution only partly fits the web data. The t-test for one sample (web file size) in two tails distribution found to be 7.3395, which is greater than the tabulated value (Tables 1), indicating that the trace data do not follow normal distribution.



(a) (b)

Fig. 3. Web file size distribution and Cumulative distribution of LBL_CONN_7 (byte sent by responder) data

Fig. 3(b) compares the cumulative probability distribution of the generated data points using Pareto (simple) model with the original traces. The curve does not match well with the actual data. There is consistent discrepancy between the actual distribution (web file data) and the Pareto model curve. The actual distribution seems to be light-tailed for which a good fittings can not be made as Pareto model follows heavy-tailed distribution.

Table 3. Test statistics (A2, W2 and U2) for LBL-CONN-7-responder data

Distribution	Test Statistics (TR = Test Result, α = significance level)		
	A2	W2	U2
Normal	TR = 0.3079 α = 0.50 Accept	TR = 0.0474 α = 0.50 Accept	TR = 0.0470 α = 0.50 Accept
Extreme value	TR = 0.5028 α = 0.10 Accept	TR = 0.0894 α = 0.10 Accept	TR = 0.0849 α = 0.10 Accept
Weibull	TR = 0.5074 α = 0.15 Accept	TR = 0.0898 α = 0.15 Accept	TR = 0.0836 α = 0.15 Accept
Exponential	TR = 1.1899 α = 0.05 Accept	TR = 0.2072 α = 0.05 Accept	TR = 0.1377 α = 0.10 Accept
Logistic	TR = 0.3321 α = 0.25 Accept	TR = 0.0482 α = 0.25 Accept	TR = 0.0479 α = 0.25 Accept
Pareto (generalized)	TR = 1.18 α = 0.05 Accept	TR = 0.2165 α = 0.05 Accept	TR= α = Accept/reject

Fig. 4 compares cumulative probability distributions of the actual trace data, the exponential, Pareto (generalized), Weibull (three parameters), logistic and extreme value distributions. All models show satisfactory fittings to the data points. Weibull and extreme value distributions illustrate the best fittings of curves as shown in Figs. 4(c) and (e), which is an indication of fractal-like burstiness behaviour of LBL_CONN_7 (byte sent by responder) data patterns.

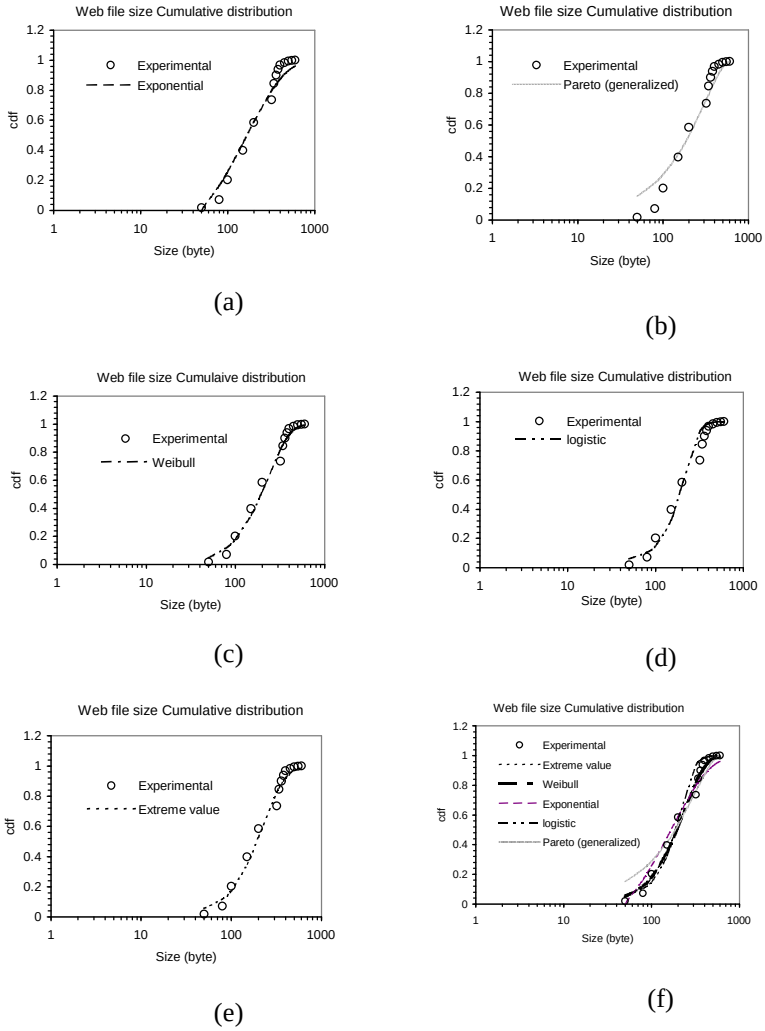


Fig. 4. Web file size cumulative distribution of LBL_CONN_7_responder data

The test statistics (A2, W2 and U2 tests) confirm (Table 3) that the trace data follow the normal distribution. The test statistics also show that the data are significant at the levels of 0.10, 0.15, 0.25, and 0.05 for extreme value, Weibull, logistic and Pareto (generalized) distributions, respectively, which is consistent with the results shown in Figs. 4(b), (c), (d), and (e). For exponential distribution, the A2 and W2 tests show the data are significant at the level $\alpha = 0.05$, and U2 shows at the level of 0.10 and the effect of these significance levels observed in Fig. 4(a) which is understandable. Likewise, there is contradiction between EDF test statistics and t-test as t-test does not support the trace data for normality.

Fig. 4(f) shows the combination of Figs 4(a), (b), (c), (d), and (e) from where comparisons of goodness-of-fit of trace data to different distributions can be made.

5 Conclusions

The T-test and EDF test statistics (A2, W2 and U2) have been used to evaluate how well a particular distribution describes the real web traffic data. The models employed in the work include lognormal, Pareto (simple), exponential, Pareto (generalized), Weibull (three parameters), logistic and extreme value distributions.

There is contradiction between EDF test statistics and t-test; the t-test results do not support the web trace data for normality, but the EDF test accepts the hypothesis that the trace data follow normal distribution. The discrepancy may be due to the different sampling methods, which has to be investigated in the future.

There are satisfactory fittings of curves observed between Weibull and the generalized Pareto model data and the real trace data, which has been confirmed by the EDF test for the two distributions. The results show that Weibull (three parameters) is the most suitable model to approximate the web traffic. In addition, the generalized Pareto model is more suitable for analysing traffic fractal-like behaviour (burstiness behaviour) than simple Pareto model.

Acknowledgements

The authors would like to thank LBL for providing the web traffic data.

References

1. V. Frost, and B. Melamed, "Traffic modeling for telecommunications networks," *IEEE Communication Magazine*, vol. 33, pp. 70-80, March 1994.
2. R. Gusella, "A measurement study of diskless workstation traffic on an Ethernet," *IEEE Trans. Commun.*, vol. 38, no. 9, pp. 1557-1568 , 1990.
3. H. J. Fowler, and W. E. Leland, "Local area network traffic characteristics, with implications for broadband network congestion management," *IEEE J. Select. Areas Commun.*, vol. 9, no. 7, pp. 1139-1149, 1991.
4. P. B. Danzig, S. Jamin, R. Caceres, D. J. Mitzel, and D. Estrin, "An artificial workload model of TCP/IP internetworks," *J. Internetworking: Practice and Experience*, vol. 3, no. 1, pp.1 - 26 , 1992.
5. V. Paxson, and S. Floyd, "Wide area traffic: the failure of Poisson modeling," *IEEE/ACM Trans. Networking*, vol. 3, pp. 226-244, 1995.
6. D. Heyamn, "Some issues in performance modeling of data traffic," *Perform. Eval.*, vol. 34, pp. 227-247, 1998.
7. V. Paxson, "Empirically-Derived Analytic Models of Wide-Area TCP Connections," *IEEE/ACM Trans. Networking*, vol. 2, no. 4, pp. 316-336, Aug. 1994.
8. M. Nabe, M. Murata, and H. Miyahara, "Analysis and modeling of world wide web traffic for capacity dimensioning of internet access lines," *Performance Evaluation*, vol. 34, pp. 249 - 271, 1998.
9. J. Feder, Fractals, Plenum Press, New York, 1988.

10. K. Falconer, Fractal Geometry: Mathematical Foundations and Applications, John Wiley & Sons, Chichester, 1990, pp.146-160.
11. B. H. Kaye, A Random Walk Through Fractal Dimensions, VCH, Weinheim, 1994, pp. 179-188.
12. S. Tang, "Computer simulation of fractal structure of flocs," Encyclopaedia of Surface and Colloid Science, pp. 1162-1168, August 2006. Taylor & Francis, ISBN: 978-0-8493-9615-1.
13. K. Rezaul, S. Tang, T. Wang, and A. Paksta, "Empirical distribution function test for World Wide Web traffic," in *Proc of International Conference on Applied Computing* (*IADIS*), Lisbon, Portugal, 23-26 March 2004.
14. D. Awduche, A . Chiu, A. Elwalid, I. Widjaja, and X. Xiao, "RFC 3272: Overview and Principles of Internet Traffic Engineering," May 2002. Available: http://rfc3272.openrfc.org/
15. L. J. Judith, and J. L. Wang, From Network Management Collection to Traffic performance modelling: Challenges and lessons learned. CAMAD '98, 1998.
16. M. E. Crovella, M. S. Taqqu, A. Bestavros, "Heavy-Tailed Probability Distributions in the World Wide Web". Available: http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.29.4346
17. Bell Labs. Busier Networks create smoother traffic flow. Available: http://www.informationweek.com/story/IWK20010611S0005
18. D. Clark, W. Lehr, and I. Liu. Provisioning for Bursty Internet Traffic: Implications for industry and Internet structure. MIT WISQ. Available: http://www.ana.lcs.mit.edu/anaweb/PDF/ISQE_112399_web.pdf
19. M. S. Squillante, D. D. Yao, and L. Zhang. Internet Traffic: Periodicity, Tail Behaviour, and Performance Implications. Available: http://www.research.ibm.com/mmaa/publication.html
20. PROPHET StatGuide. Goodness of Fit (Chi-square) Test. Available: http://www.basic.nwu.edu/statguidefiles/probplots.html#Heavy-tailed%20Data
21. R. B. D'Agostino, and M. A. Stephens, Tests based on EDF statistics, in *Goodness-of-fit Techniques*, Ch.4, Dekker, New York, 1986.
22. Traces available in the Internet Traffic Archive. Available: http://ita.ee.lbl.gov/html/traces.html