Advanced Engineering Forum *Online: 2012-09-26*
ISSN: 2234-991X, Vols. 6-7, pp 836-842
doi:10.4028/www.scientific.net/AEF.6-7.836

Research on Network Traffic Modeling and Applications

Yongli Ma[1, a] , Zongjue Qian[2,b] ,Guochu Shou[2,c],Yihong Hu[2,d]

[1]Beijing GFA E-commerce Security CA Co.,Ltd.,China

[2]Beijing University of Posts and Telecommunications, China

[a]hafmn@263.net, [b]lzongjueqian@126.com, [c]gcshou@bupt.edu.cn, [d]yhhu@bupt.edu.cn

Keywords: Network Traffic, Modeling, Self-Similar.

Abstract. With the popularity of internet and the growing of applications in recent years, network traffic characteristics are also undergone a great change. The traditional flow models are unable to meet the current traffic. Therefore, it is done to study the law of current traffic model, to propose these models of service traffic characteristics and to explore the applications of these models in practice.

Introduction

In recent years, internet obtained rapid development; it has changing people's daily lives. It always keeps changing from service content to the service form of network services, from simple text, images to complex audio, video; terminal from an ordinary PC to mobile phones, tablet PCs and even televisions, refrigerators and other traditional household appliances. As a result, the size of internet is becoming increasingly large, the structure has become increasingly complex, the traffic has become more and more huge, and the characteristics make the traffic modeling, analysis and control to become very important and extremely difficult.

The internet is essentially an open, heterogeneous computer networks, it connects a variety of computer networks and systems together that be distributed around the world through the TCP / IP protocol. In recent years, many research indicate that the internet service flow is a self-similar process [1],it has Long-range correlation characteristics [2], the flow of the service in the WAN shows multiple fractal characteristics [3][4].

The self-similarity of network traffic can produce many adverse effects to network performance. The self-similarity is stronger, the average queue length of the output buffer of network switching node is longer, the average delay of the network is bigger; and the convergence of service flow will increase their sudden. It directly leads to reduce bandwidth utilization in output node, and have a bad effect on service quality (including delay, jitter, and packet loss ratio). Poisson model in the telephone network don't not considered generally self-similarity that was successfully used for many years, and thus find to describe traffic model of self-similarity characteristic has become an important research problem.

Self-similarity of traffic is usually described using the Hurst parameter, as follows: (1), when the Hurst parameter is between 0.5 to 1, the service flow has obvious self-similarity, and the Hurst parameter is larger, the degree of similarity is higher; (2), when H = 1/2, the random process is not self-correlation, that it does not affect the future, such as the common white noise sequences. (3), when $0 < H < 1/2$, the random process exists only short-range correlation.

It can be seen from above analysis that Hurst parameter can characterize the self-similarity of the random process. In the paper, flow sequence is mapped to a random series; the research on the characteristics of flow time series is mainly to estimate the Hurst parameter.

Traffic Modeling

Analysis of Data Source. In order to capture network traffic at different time and location , we tested operation broadband network to collect more than 5 million valid data which include DNS, VoIP, Web, stream media, game, and ftp.

In order to estimate self-similarity Hurst parameter of a variety of services such as stream media based on P2P technology, online game, VoIP, web browsing, we select the following six representative data sets from repeatedly collected data, as follows:

Table 1: Data Source Analysis

Dataset name	software name	Service type	period	TCP (%)	UDP (%)	Packets (Pieces)	Bytes (byte)
Dataset Ⅰ	PPLive	stream media	10'9"	30.09	68.75	257830	107340524
Dataset Ⅱ	Feidian	stream media	23'33"	1.42	94.18	135184	101923673
DatasetⅢ	Rytv	stream media	46'6"	95.59	1.55	457180	286568035
DatasetⅣ	Star	game	33'35"	15.01	69.21	94908	19930138
Dataset Ⅴ	NGN	VoIP	64'41"	36.76	62.24	573586	269157196
DatasetⅥ	http	Web	45'12"	84.46	10.72	130552	65903249

Self-Similarity of Traffic analysis. Based on MATLAB platform, self-similarity Hurst parameter of streaming media service, VoIP service, online games service and Web browser service is estimated respectively by absolute moment method, Aggregated Variance method[5], Modified Period gram method[6], Higuch method[7], Differential variance method[5],Variance of Residuals method[8], and R/S method[9] based on six datasets. The average Hurst parameter of each service traffic is calculated apart, as shown in Table 2.

The self-similarity Hurst parameter for the different service flow is different that can be found from Table 2, the Hurst parameter is higher for Web service, VoIP services, and game service respectively, while the Hurst parameter of the stream media service based on P2P technology is lower. The reasons analyzed as the following:

(1) Hurst parameter is related to the operating mechanism of a protocol which supports this service,

(2) Hurst parameter of a service is related to the user's habits that are network behaviors,

(3) From the form of expression, and the scale of the observed time scales is smaller, the self-similarity is stronger usually.

Table 2 Hurst parameter of traffic

Dataset name	Absolute Moment method	Aggregated Variance method	Modified Period gram method	Higuch method	Differenced Variance method	Variance of Residuals method	R/S method	Hurst Parameter mean value
Dataset Ⅰ	0.503	0.445	0.396	0.992	0.51	0.492	0.472	0.544
Dataset Ⅱ	0.517	0.495	0.474	1	0.519	0.499	0.539	0.577
Dataset Ⅲ	0.39	0.402	0.397	0.999	0.649	0.491	0.375	0.529
Dataset Ⅳ	0.851	0.631	0.624	0.999	0.516	0.692	0.73	0.721
Dataset Ⅴ	0.37	0.483	0.721	1	0.374	0.319	0.335	0.689
Dataset Ⅵ	0.941	0.882	0.767	0.999	0.567	0.825	0.89	0.839

Analysis of Traffic Models

Model Fit Test

In this paper, the Kolmogorov- Smirnov test method is selected, it is more accurater than χ^2 [10], It can test that the empirical distribution is subject to distribution of the theory, but also can test whether two samples is from the same population.

In MATLAB toolbox, K-S test has the parameter H, P, K, the CV, for example, H is the results of the KS test, H = 1 implies to deny the assumption, H = 0 represents to accept the assumptions; P is a P-value of the KS test, that ,randomized trial results is greater than the probability of the sample statistics; K is statistic calculated results of KS test; the CV is corresponds to a significant level of alpha sub-sites (Kolmogorov-Smirnov critical value), the default significance level a is 5%.

Stream Media Service

Streaming media services is from network TV, also known as IPTV, Overall, network television according to the terminal is divided into three forms, namely the PC platform, TV (STB) platform and the mobile phone platform (mobile network). Variety of network TV software based on P2P technology include PPLive, PPStream, FeidianTV, etc. Different network TV software uses different technology of transport layer, and some service use the TCP protocol to establish a signaling connection, using the UDP protocol for media transmission; some service use the UDP protocol to establish the signaling connection, use the TCP protocol for media transmission, while others is mix, detailed analysis is in Table 1.

Fig.1, Fig.2, Fig.3, Fig.4, Fig.5, Fig.6, respectively, is the CDF (cumulative probability distribution function) diagram and PDF (the probability density distribution function) diagram of the Dataset Ⅰ, the Dataset II, the Dataset III. As follows:

Dataset Ⅰ (PPLive video traffic) submits to mean μ = 7.5896, standard deviation σ = 1.9150 of the lognormal distribution, KS test parameter p = 0.8718, statistical amount of k = 0.0071, the fractile cv =0.0162.

Dataset II (Feidian video traffic) obeys the shape parameter k=1.0292, scale parameter σ=0.7362, the threshold parameters θ=0.7362 of the pareto distribution, KS test parameter p=0.8098, statistical amount k-= 0.0103, the fractile cv = 0.0219.

Dataset III (Rytv video stream) complies with mean μ = 7.7026, standard deviation σ = 2.0876 of the lognormal distribution, the KS test parameters p = 0.5480, statistic k = 0.0089, the fractile cv =0 .0152. As follows:

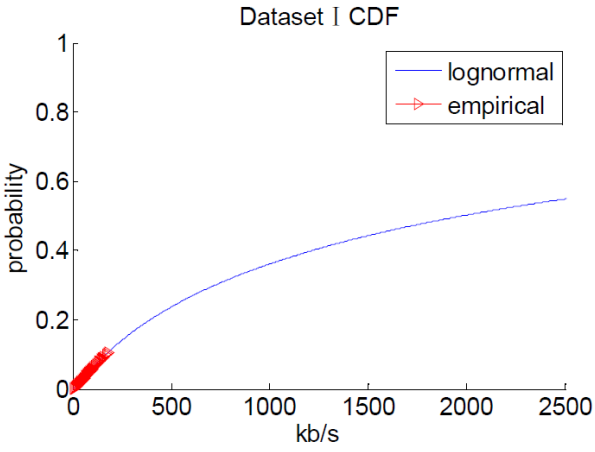

Fig.1 Dataset Ⅰ CDF

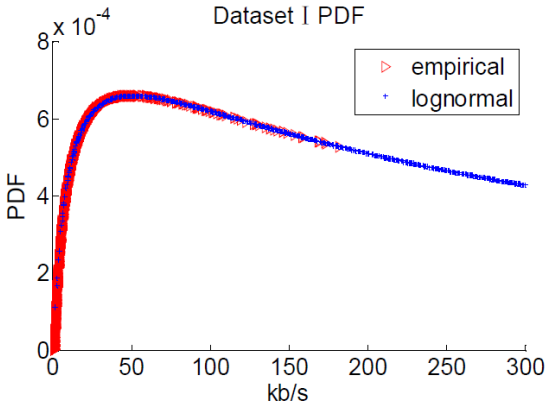

Fig.2 Dataset Ⅰ PDF

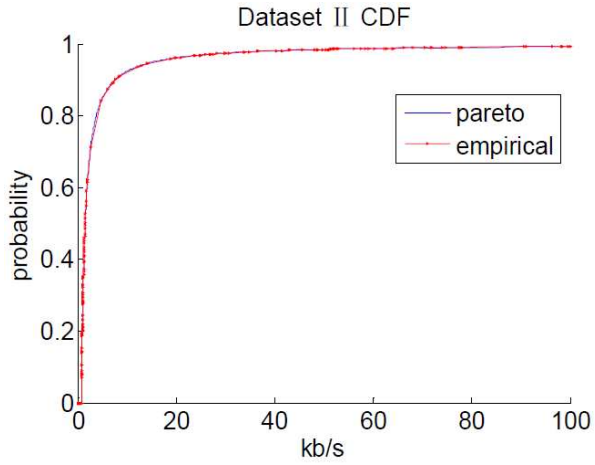

Fig.3 Dataset II CDF

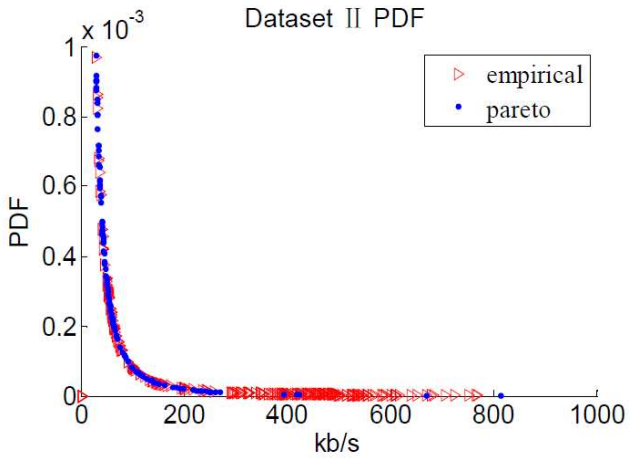

Fig.4 Dataset II PDF

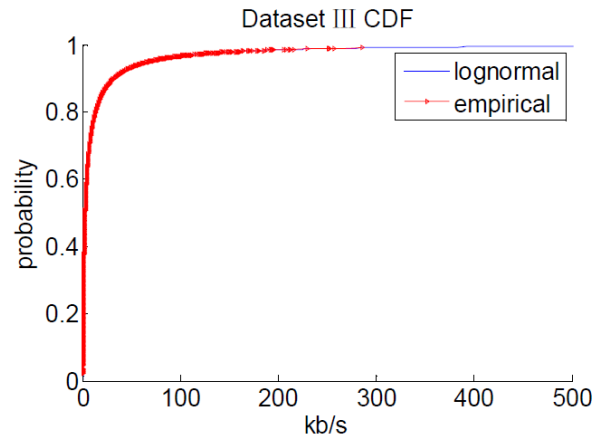

Fig.5 Dataset III CDF

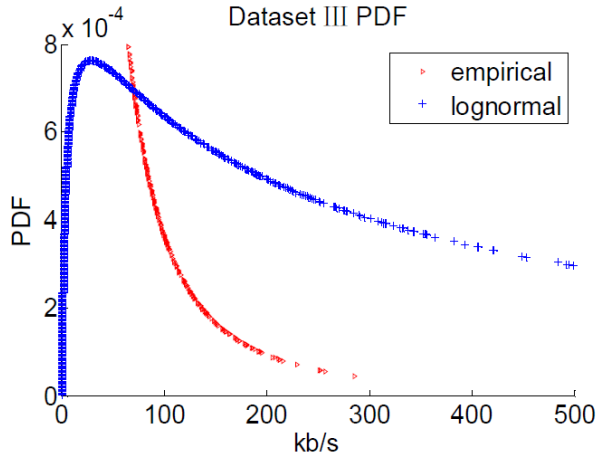

Fig.6 Dataset III PDF

Game Service

The online game is a new game project based on internet, many people can also participates in the interaction for the purpose of communication, entertainment and leisure. Therefore, as the users increase, the growing traffic generated by online games, the paper collected the flow sample of StarCraft game, that is the Dataset Ⅳ, MATLAB tools were used to fit the distribution function, and finished the KS test, the KS test results show h = 0; p = 0.9163; statistic k = 0.0094; the fractile cv =0.0229, it obeys the parameters α = 0.02, β = 13152 of gamma distribution. As follows:

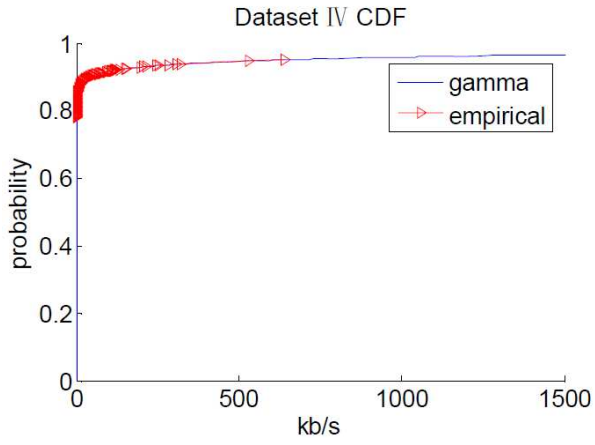

Fig.7 Dataset Ⅳ CDF

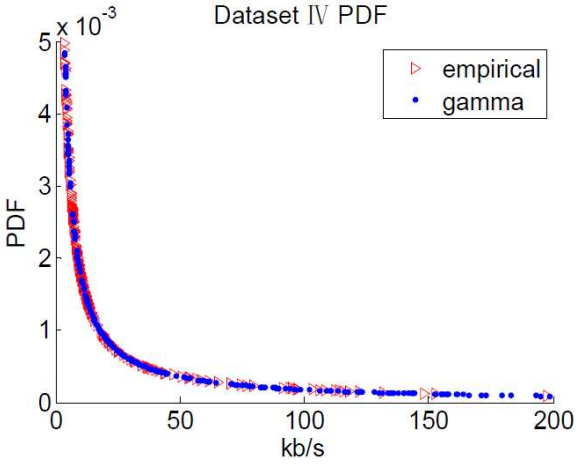

Fig.8 Dataset Ⅳ PDF

VoIP Service

With the continuous development of Internet technology, VoIP (Voice Over IP) is used widely. At present, leading VoIP protocol stack includes H.323, SIP, MEGACO and MGCP, and so on.

The distribution function fitting and the KS test are done by MATLAB tool for data sets Dataset V. KS test results show h = 0; p = 0.9883; statistic k = 0.0101; the fractile cv = 0.0305.

It belongs to the weibull distribution of parameters $\alpha = 20211$, $\beta = 0.49$. As follows:

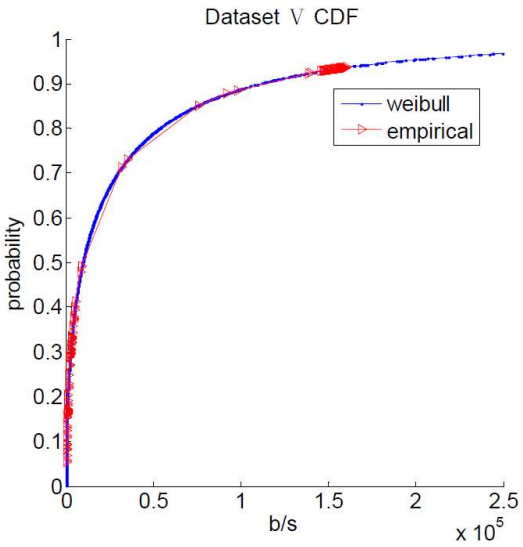

Fig.9 Dataset Ⅴ CDF

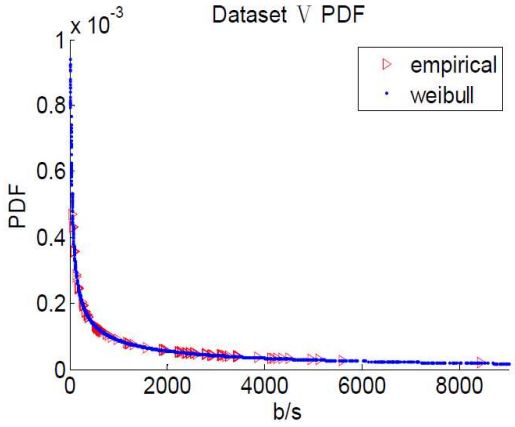

Fig.10 Dataset Ⅴ PDF

Web Browser Service

Web browser service has outburst features, the typical Web browser service likes this: firstly, click on a hyperlink from the browser, secondly a hypertext document request is sent to the server. The server received the request, according to the content of hypertext documents, related to the master file, the embedded image file, audio files, video files, and so on, sequentially sent to the browser.

The distribution Function fitting and KS test are done by MATLAB tools for the Dataset Ⅵ, KS test results show for h = 0; p = 0.2213; statistic k = 0.0086; the fractile cv = 0.0112. It belongs to the lognormal distribution with mean $\mu = 5.2828$, the standard deviation $\sigma = 1.2878$.

Fig.11 and Fig.12 represent respectively the CDF curve and PDF curve of the flow sample. it can be seen from Figure 2-5b, most of the time, the flow rate is concentrated in the 10Kbps following such low speed range, high-speed transmission continues only a small period of time. It is correspond to the characteristics of typical Web browsing. Self-similarity Hurst parameter is about 0.839, the sample of flow has a strong self-similarity. As follows:

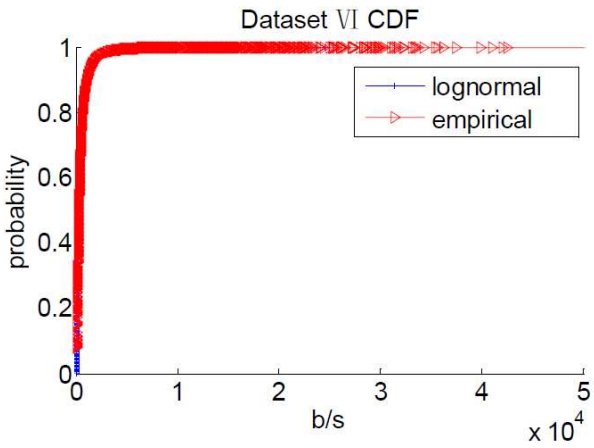

Fig.11 Dataset Ⅵ CDF

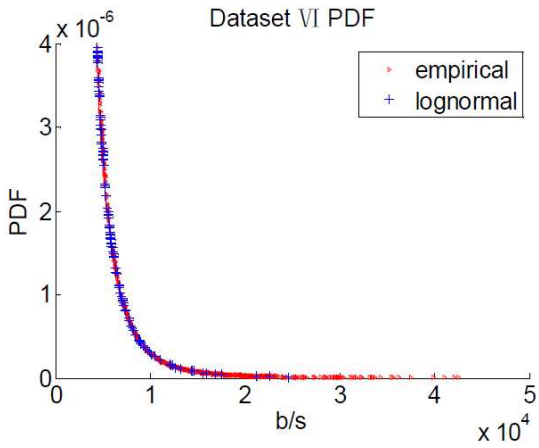

Fig.12 Dataset Ⅵ PDF

Analysis results of lots of typical traffic samples show that the different service characteristics lead up to different types traffic, there are great difference in the probability density distribution characteristics, but also self-similarity level is different. The probability density of the traffic is various which traffic generated by the same type service. From the analysis results of the sample, the traffic generated by the games service and Web browsing has a stronger self-similarity; self-similarity of the traffic generated by the media service is relatively weak or even no self-similarity.

Traffic Model Application

It is the key to network traffic model from the laboratory to industrial applications to be used in the actual network design and network equipment design. At present, the application of network flow model includes the following aspects:

Cache Design. When the rate of the device port is determined in the design of the network equipment (for example, is set to 100Mbps or 1Gbps), network traffic characteristics become a decisive factor to determine the device cache size. In such applications, the model of network traffic is basis as the computing device cache size required.

For example, we use multi-scale queuing (MSQ, Multi scale Queue) model [11] [12] Proposed by Rudolf.H Riedi et al. to estimate the size of the cache, a brief description of such applications for the following: According to this model, the cache size that network equipment ports whose maximum rate is c need can be estimated by the following formula:

$$P[Q > b] \approx MSQ[b] \tag{1}$$

$$MSQ[b] = 1 - \prod_{i=1}^{n} P[k_{2^{k-n}} < b + c2^{n-i}] \tag{2}$$

In the above formula, Q is the queue size, b is the cache size, n is a positive integer, k_r is the arriving traffic during the time -r+1 to 0, $P[Q > b]$ represents the congestion probability when the queue size is Q, cache size is b. From Eq.2, when the probability density distribution of network traffic is determined, the cache size b can be determined based on the congestion probability which the design permits. Vice versa, it can estimate the congestion probability according to the cache size designed. The probability density distribution which required in the estimate process can be obtained from the network flow model which has been established.

Link bandwidth design. At present, the network operators take link flow mean measured 2-3 times as link bandwidth, and expect to protect the QoS metrics of traffic during peak periods by providing a well-off bandwidth.

The advantage of this method is simple. The disadvantage is that usually bring two negative consequences: First, providing redundant bandwidth is too much. Although the bandwidth is redundancy, it can't meet the need during the flow peak to happen congestion.

Anomaly Detection. Currently, the network traffic anomaly detection technology is the primary technology means to solve the problem of abnormal network traffic. In order to identify abnormal traffic, it is necessary to collect and analyze network traffic under normal conditions, finish the normal network traffic modeling based on time. When the network anomaly traffic detection system is online, it creates a dynamic flow benchmark on the monitored network traffic for each time period, variety of protocols in the system database. If a modeling of traffic generated by a protocol does not match its current base on particular time, it will be given an exception alarm, and the alarm will escalate over time. Then, it will take targeted measures such as tracking of abnormal traffic, finding an abnormal source and filtering traffic on the basis of analyzing the alarm, to achieve the protection of the network.

Summary

The article firstly analyzes the self-similarity characteristics of network traffic, and estimates the Hurst parameter of the streaming media service, the online game service, VoIP services, as well as the Web browser service. It finds the Hurst parameter of streaming media service based on P2P technology closes to 0.5, and is very weak self-similarity. Secondly, the article finishes modeling of the stream media service, the online game service, VoIP services and the Web browser service based on the measured data and the probability density function, and accomplishes the K-S testes for these modeling. Finally, the article researches application based on traffic modeling. It has opened up a new way for planning, management and control of broadband network.

References

[1] W.Leland, M.Taqqu,W.Willinger, et al. On the self-similar nature of Ethernet traffic (extended version).IEEE/ACM Trans.On Networking.1994, 2(1):1-15.

[2] J.Beran,R.Sherman,M.S.Taqqu,et al. Long-range dependence in variable-bit-rate video traffic. IEEE Trans.On Communications.1995, 43(2):1566-1579.

[3] A.Feldmann, A.C.Gilbert, W.Willinger, etal.The changing nature of network traffic: Scaling phenomena. ACM SIGCOMM Computer Communication Review.1998, 28(2): 5-29.

[4] A.Feldmann, A.C.Gilbert, W.Willinger. Data networks as cascades: Investigating the multifractal nature of Internet WAN traffic. Proc.of the ACM/SIGCOMM'98, Vancouver, B.C, 1998:25-38.

[5] V.Teverovsky and M.S.Taqqu.Testing for long range dependence in the presence of shifting means or a slowly declining trend using a variance type estimator.Preprint.1995.

[6] M.S.Taqqu,V.Teverovsky and W.Willinger.Estimators for long-range dependence:an empirical study.Fractals.1995.3(4).pp785-798

[7] T.Higuchi.Approach to an irregular time series on the basis of the fractal theory.Physica D,31.1998.pp277-283

[8] C.K.Peng,S.V.Buldyrev,M.Simons,H.E.Stanley,and A.L.Goldberger. Mosaic organization of DNA nucleotides.Physical Review E.1994.pp1685-1689

[9] M.S.Taqqu and V.Teverovsky.Estimating long-range dependence in finite and infinite variance series.In R.Adler,R.Feldman and M.S.Taqqu, editors. A Practical Guide to Heavy Tails: Statistical Techniques for Analyzing Heavy-Tailed Distributions. Boston 1996.

[10] Y.T.Zhu,C.J.Wang,M.X.Zhao. Mathematical Statistics, Northwestern Polytechnical University Press Co.Ltd.XIAN,1999

[11] H.R.Riedi, S.M.Crouse, J.V.Ribeiro, and G.R.Baraniuk. A Multifractal Wavelet Model with Application to Network Traffic. IEEE Transactions on Information Theory, April 1999.

[12] J.V.Ribeiro,H.R.Riedi,S.M.Crouse, and G.R.Baraniuk. Multiscale Queuing Analysis of Long-Range-Dependent Network Traffic, Proceedings of IEEE INFOCOM'00,March 2000.