マルチメディア通信と分散処理 100−17
（2000. 11. 10）

インターネット トラ フィックデータの特性

T.K.Roy, D. Chakraborty, A. Ashir, G. Mansfield, 白鳥則郎

† 〒980-8577 仙台市青葉区片平 2-1-1
東北大学電気通信研究所 / 情報科学研究科

あらまし　インターネットのトラフィックデータには、長期的なタイムスケールで見た場合に、非常に大きな変化性とバースト性があることが知られており、トラフィックの多少の周期が明確ではない。しかし、トラフィックデータには自己相似性があり、これを用いることによってトラフィックデータの特性抽出を行うことができる。自己相似性は、観測のタイムスケールによって変化する、ある統計学的量により表現される。そこで我々は、本研究室で得られたトラフィックデータをもとに、この自己相似性の尺度である以下の値の計算を行った。i) 分散: 観測の時間周期による分散の減少を見ることにより、自己相似度を示すパラメータ (β) が得られる。ii) 自己相関係数: 自己相関係数はタイムスケールの変化に伴い非常にゆるやかに減衰し、それ自身が自己相似性を示す、自己相似性の尺度の一つである。iii) ハーストパラメータ H: ハーストパラメータは、トラフィックデータに対し rescaled range を行うことにより得られる自己相似性の尺度の一つである。また、トラフィックデータがその部分周期においても相似性を示す場合には、そのデータはフラクタル性をも示す。しかし、インターネットのトラフィックデータに関する自己相似性に関しては様々な研究があるが、フラクタル性についてはほとんど研究されていない。そこで本稿では、フラクタル性についても議論し、自己相似性との関連について述べる。

キーワード　インターネット　トラフィック, 自己相似性, ハースト　パラメータ, 自己相関係数, フラクタル　ディメンジョン

Characteristics of Internet Traffic Data

T.K. Roy, D. Chakrabarty, A. Ashir, G. Mansfield, N. Shiratori

†Research Institute of Electrical Communication
Tohoku University 980-8577, Sendai.

Abstract　The Internet traffic data have been found to possess extreme variability and bursty structures in a wide range of time-scales, so that there is no definite period of busy or silent periods. However, there is a self-similar feature which makes it possible to characterize the data. Self-similarity is expressed in terms of the different statistics varying with the time scale of observation. We give a brief description of those we have calculated to determine the self-similarity of the Internet traffic data obtained in our laboratory [a]. These are i)Variance, the decrease of which with the time scale of observation gives a parameter (β) to specify the degree of self-similarity, ii)Autocorrelation, with a very slow decay rate and itself showing self-similar features and iii)Hurst parameter H, another independent measure from the *rescaled range* of the data. The similarity of the data in a sub-period and its finer intervals leads to the possibility of the data to posses fractal characteristics also. Although extensive works have been done on the self-similar features of Internet traffic data, there has not been much on this aspect, which can exist in both the time and space scales. Here we attempt to provide a description of the fractal characteristics associated with such a self-similarity.

key words　Internet traffic, self-similarity, Hurst parameter, autocorrelation, fractal dimension.

[a]Shiratori Laboratory, RIEC, Tohoku University

1 Introduction

The self-similar nature of Internet traffic data was first proposed by the Leland *et al* in 1993 [lela93] and subsequently established by others in a flood of research works on the subject [lela94, paxs95, will94]. It was then a new concept against the long believed idea of the Poisson traffic. The main observations of the Internet traffic were that the data were found to be highly variable and bursty and did not seem to follow a steady state. The traffic came in starts and fits with lulls in between. The variability existed even in small time scales. This was discovered when an attempt was made to find a time scale of the bursts or lulls or the intervals between them.

The traditional Poison traffic model assumed the variation of data flow to be finite around a mean but the observations on the Internet traffic proved otherwise. It is this large variance of data flow that leads to the self-similar nature. The data show self-similarity almost at all scales of resolution. Such self-similar nature is always associated with a fractal structure of the data. The fractal characteristics can exist both in the temporal and spatial scales. This was indicated by Willinger and Paxson [will98], as due to the extreme variability and the long range dependence in the process.

Presently, one of the main research interests in the field of Internet traffic is that of prediction of data. Before preparing a model of prediction, one of the important tasks is to determine its statistics. Although of stochastic nature, it is still guided by some characteristics which do not or very slowly change with time. Any model to predict the future values will have to preserve these characteristics. Therefore, it is important that the characteristics of the network traffic are determined for a good management and a satisfactory quality of service.

We report in this paper on the nature of Internet data obtained in our Laboratory, along with an attempt to determine if there is any spatial fractal characteristics. After a brief description of the different statistics required to determine self-similarity in a data set, we give a description of the spatial fractal characteristics similar to the temporal, and a method to determine the fractal dimension, similar to that used in nonlinear dynamics. Finally, we conclude with the results.

2 Variance and Autocorrelation in a Self-similar time series

The self-similarity is defined in terms of aggregates from a time series. Experimentally we observe the rate of data flow at an interface either inwards or outwards. We assume that the nature of flow is *stationary*, i.e. the statistics of the process under observation do not change with time. Then if X_i is the record at the ith time resulting in a series

$$X = \{X_1, X_2, ..., X_N\}, \tag{1}$$

the m-aggregated time series $X^{(m)}$ is defined as

$$X_k^{(m)} = \{X_1^{(m)}, X_2^{(m)}, ..., X_M\} \tag{2}$$

where,

$$X_k^{(m)} = \sum_{i=m(k-1)+1}^{km} X_i/m \tag{3}$$

As we move towards higher m, the resolution decreases from the highest that is obtained in experiment to the lowest. The m-aggregated series represents a compression of the data by *m-times*.

If the statistics (for example *mean, variance, correlation* etc.) of a process is preserved with such a compression then it is a ***self similar*** process. The degree of self similarity is expressed by a parameter β, if for all $m = 1, 2, ...$ we have the variance at the mth level of aggregation $Var[X^{(m)}]$ related to the original variance $Var[X]$ as:

$$Var[X^{(m)}] = Var[X]/m^{\beta}, \tag{4}$$

The autocorrelation after k time steps defined as

$$R(k) = \sum_{i=1}^{N} X_i X_{i+k}/N \tag{5}$$

(for $N >> k$) remains a constant at all levels, i.e.

$$R^{(m)}(k) = R(k), \tag{6}$$

for all k and all m, for a perfect self-similar series, with $R^{(m)}(k)$ obtained from $X^{(m)}$ as in Eq. (5). In other words the original time series and that after m-aggregation are same in as far as the autocorrelation is concerned. Experimentally we observe Eq. (4) and Eq. (6) as m becomes large, and β ranging from 0 (for full self-similarity) to 1 (for ordinary data). The autocorrelation $R(k)$ for a self-similar process is found to be long-ranged:

$$R(k) \rightarrow k^{-\beta}(klarge), \tag{7}$$

For noisy data the autocorrelation is zero for $k > 0$ with $\beta = 1$.

3 R/S Statistic

An equivalent characterization is given by a quantity named by H.T.Hurst [hurst65] as the rescaled range (R/S) of the data, defined as:

$$\frac{R(N)}{S(N)} = \frac{RescaledRange(X, N)}{Standard\ deviation(N)}, \tag{8}$$

$$R(N) = maximum\ of\ L_j\ -\ minimum\ of\ L_j, \tag{9}$$

$$L_j = \sum_{i=1}^{j}(X_i - M(N)), \quad 1 \leq j \leq N, \qquad (10)$$

and $M(N)$ is the mean of the data of size N. $S(N)$ is the usual standard deviation.

For self-similarity, R/S follows a power law for large N:

$$R/S \sim (N/2)^H, H > 0.5 \qquad (11)$$

where H is the Hurst parameter. It can be shown that β is related to H as

$$H = 1 - \frac{\beta}{2} \qquad (12)$$

so that for full similarity expressed by $\beta = 0$, the Hurst parameter is 1.

4 Fractal Characteristics of Internet Data

The self similar nature of the Internet traffic data is due to its high or extreme variability in both time and space. Temporal high variability results from the long range dependence which is described as the evolution of a process depending on its state long before. The autocorrelations fall off very slowly with time exhibiting a power law behavior in contrast to the exponential decay as in short range dependent systems. Again, extreme spatial variability is also observed. The underlying distributions for this feature have been found to be those with infinite variance, referred to in the literature as heavy tailed distributions.

Processes with such features, the long range behavior or power law dependence are expected to have fractal characteristics. Indeed the Internet traffic data exhibit the fractal like structure over a long range of time scales. As discussed by Willinger and Paxson [will98] we define a process to possess fractal characteristics, if there exists a relationship of the form:

$$Q(\tau) \propto \tau^{f(D)} \qquad (13)$$

where Q is a certain quantity depending on τ, a resolution in time or space and $f(D)$, a simple, often linear, function of the dimension D of the process, the *fractal* dimension, so defined.

In fact one of the measures of self-similarity is based on such an equation. When Q is taken to be the variance of the data, then $f(D)$ is a simple linear function of the dimension D identified to be the Hurst parameter H:

$$Var[X^{(m)}] \propto m^{-\beta} = m^{2H-2} \qquad (14)$$

Thus equation Eq. (14) describes the fractal behavior of the data in time.

An equivalent description in space will be similar to that of Eq. (14) with the resolution in space. This we borrow from the field of nonlinear dynamical systems [schus95]. Imagine the range of the data to be divided into equal segments of size ϵ, and we count the number of segments that contain the data. Let this be $N(\epsilon)$. Then Q then becomes the number of segments $N(\epsilon)$ of size ϵ required to cover the data:

$$N(\epsilon) \propto \epsilon^{-D} \qquad (15)$$

so that a dimension D at resolution ϵ can be expressed by:

$$D(\epsilon) = -\frac{logN(\epsilon)}{log\epsilon} \qquad (16)$$

Then a $log - log$ plot of $N(\epsilon)$ vs. ϵ is a straight line with a slope, which gives a measure of D. It is to be noted that Eq. (16) is expected to give the dimension for some range of scales depending on the size of the data. For coarse scales (large ϵ) we do not expect to find a description, rather at finer scales, i.e. for small ϵ. This was also the case for Eq. (14). Again at smaller and smaller resolution since we always work with finite size data, we shall find a limit upto which it is valid.

5 Results

Calculations were done with the data [1] from *natori.cysol* and *swan.shiratori* of our Laboratory to test the self-similar nature of Internet traffic. Four sets of data were considered. They are:

1. 2000/2/[1-29]/shiratori/swan/ifInOctets.2 (SSIO2) and
2. 2000/2/[1-29]/shiratori/swan/ifOutOctets.2 (SSOO2), the ingoing and outgoing octets per minute during February 1-29, 2000 at the interface no.2 of swan.shiratori, each of datasize 41415 and,
3. 2000/2/[1-21]/JB/natori.cysol.co.jp/ifInOctets.2 (JBNIO2) and
4. 2000/2/[1-21]/JB/natori.cysol.co.jp/ ifOutOctets.2 (JBNOO2), the ingoing and outgoing octets per minute during February 1-21, 2000 at the interface no.2 of natori.cysol, each of datasize 29759.

A description of the data sets is provided in the following Table:

Table-I (a)

Name	Duration	Type	Size
JBNIO2	2000/2/[1-21]	InOctets	29759
JBNOO2	2000/2/[1-21]	OutOctets	29759
SSOO2	2000/2/[1-29]	OutOctets	41415
SSIO2	2000/2/[1-29]	InOctets	41415

[1] available on request

Table-I (b)

Min	Mean	Max	Variance
512	3954.8	2323200	5.2677E8
256	55103.7	93903360	5.7358E11
0	159020.9	60619648	1.0765E12
14336	31230.3	12975872	4.7012E9

Table-II

Data	β	H	D
SSOO2	0.43	0.75	0.50
SSIO2	0.95	0.62	0.50
JBNIO2	0.35	0.80	0.63
JBNOO2	0.35	0.81	0.71

Except for the first set, the InOctets of swan.shiratori, all the data sets though very much limited in size, show clear signs of self-similarity. This is illustrated by the attached figures and the table below. The slow change of variance with the size of the aggregation (m) gives an estimate of $\beta = 2 - 2H$ from a $log - log$ plot. The rescaled range also shows a power law dependence with N, the size of the data, the power given by H (> 0.5 for self-similarity).

The autocorrelation also shows self-similarity by scaling with the size m of the block. Since the data sizes were finite and not large as in the experiments of Leland *et al* [lela94], it is difficult to see Eq. (6). A real measure of $R(k)$ can be found only if $k << N$ for a long range process. However, we find Eq. (7 to be valid for $k \sim N/100$ [Fig.4], for the original series.

The similar nature of the auto-correlation plots gives evidence that the aggregated time series $X^{(m)}$ is just similar to the previous one, which in turn is similar to its former in previous level, and so on down to the original series at highest resolution. Here, instead of Eq. (6), we have a different indication of self-similarity useful for a small data set, namely, the auto-correlation plots at all levels of aggregation look similar.

We also have an indication that the same features are preserved when we work with some less amount of data (say by $10 - 15$ per cent). This may be of help in a prediction process.

All the data sets show fractal like behavior. The fractal dimension is found to be a constant over an appreciable range of precision in all the cases, as discussed in the previous section. Because the data are self-similar, it is found to give similar results in the respective aggregated series also, though with some smaller ranges. One particular case is interesting, that of SSIO2, which has a convincing fractal characteristic, but there is very less temporal self-similarity ($\beta \simeq 0.95$). It seems that *temporal* and *spatial* fractal behavior are two aspects of Internet traffic data which may be independent of each other. However, this also needs to be tested over many other sets of data obtained from other experiments.

Figure 1: A typical plot of $log\ Var[X^{(m)}]$ $vs.$ $log(m)$, for data SSOO2, with a least square fit line with $\beta = 0.43$

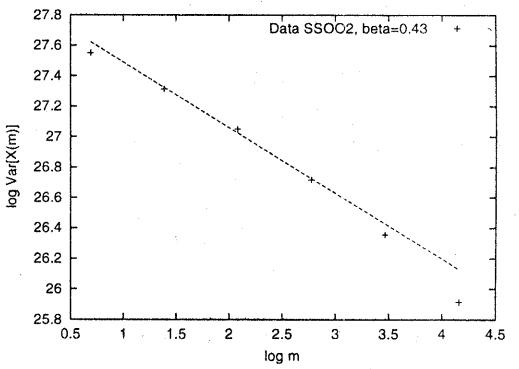
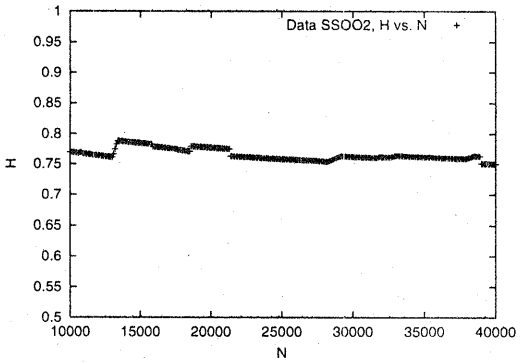


Figure 2: A typical plot of H (Hurst parameter) vs. N (data size)



References

[lela93] W.E. Leland, M.S.Taqqu, W. Willinger and D.V. Wilson, "On the self similar nature of Ethernet traffic", in Proc. ACM Sigcomm 1993, San Francisco, CA, 1993, pp 183-193.

[lela94] W.E. Leland, M.S.Taqqu, W. Willinger and D.V. Wilson, "On the self similar nature of Ethernet traffic",(extended version), IEEE/ACM Trans. on Networking, Feb. 1994.

[paxs95] V.Paxson and S.Floyd, "Wide area traffic: the failure of Poisson Modeling", IEEE/ACM Trans. of Networking, June 1995.

Figure 3: Auto-correlation for different m-aggregates, for data SSOO2

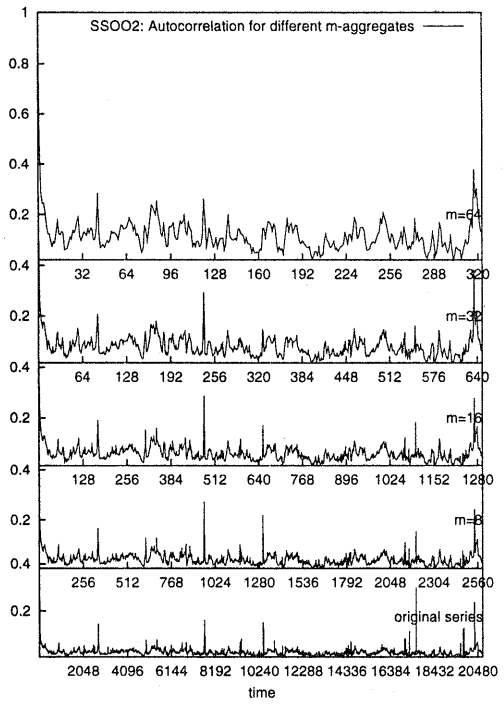
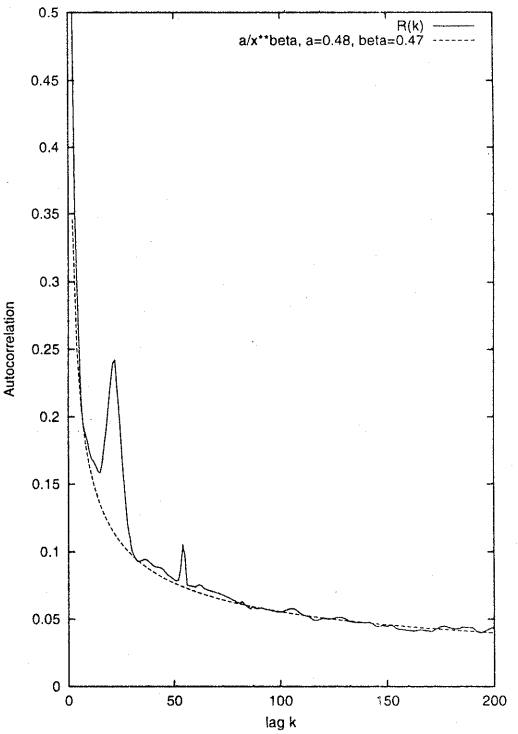


Figure 4: Auto-correlation at large k



[will94] W. Willinger, D.Wilson and M.Taqqu, "Self similar traffic modeling for high speed networks", ConneXions, Nov. 1994.

[hurst65] H.E.Hurst, R. Black and Y.Simaika, "Long term storage: An Experimental Study", London: Constable, 1965.

[will98] W. Willinger and V. Paxson, "Where Mathematics meets the Internet", Notices of the American Mathematical Society, 45 (8), pp 961-970, Sept. 1998.

[schus95] Deterministic Chaos: An Introduction, H.G. Schuster (Third Edition, 1995), VCH, Weinheim (Federal Republic of Germany).

Figure 5: *Log−Log* plot of $N(\epsilon)$ *vs.* ϵ, for data JBNIO2, with a least square line, the slope of which gives the fractal dimension $D = 0.63$

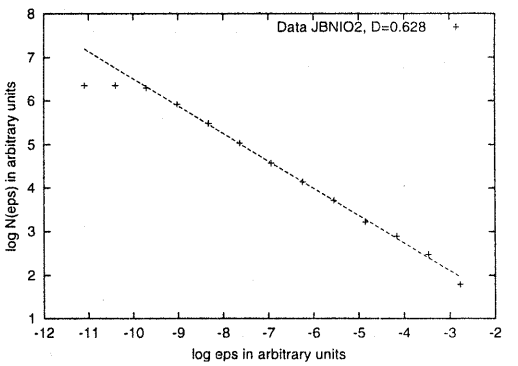


Figure 6: Same as in Fig. (5), for data SSIO2, with $D = 0.50$

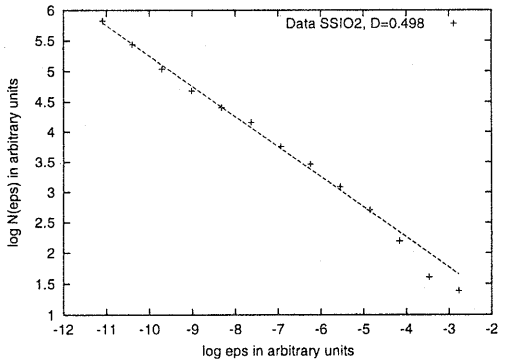


Figure 7: Same as in Fig. (5), for data JBNOO2, with $D = 0.71$

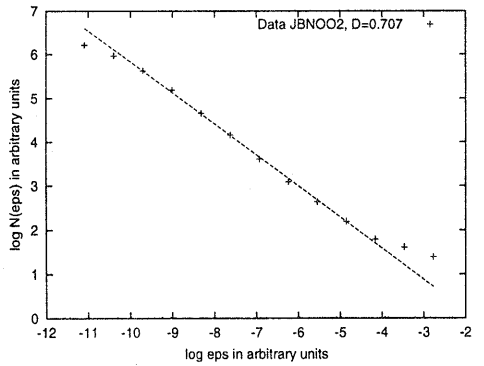


Figure 8: Same as in Fig. (5), for data SSOO2, with $D = 0.50$