2004 International Symposium on Nonlinear
Theory and its Applications (NOLTA2004)
Fukuoka, Japan, Nov. 29 - Dec. 3, 2004

A Self-Organizing Approach to Measuring Long-Range Dependence of IP-Network Traffic based on Detrended Fluctuation Analysis

Masao Masugi†

†NTT Energy and Environment Systems Laboratories, NTT Corporation
3-9-11, Midori-cho, Musashino-shi, Tokyo 180-8585, Japan
E-mail: masugi.masao@lab.ntt.co.jp

Abstract– This paper describes an analysis of IP-network traffic in terms of the time variation of self-similarity. To get a comprehensive view in analyzing the degree of long-range dependence (LRD) of IP-network traffic, this paper used a self-organizing map, which provides a way to map high-dimensional data onto a low-dimensional domain. Also, in the LRD-based analysis, this paper employed detrended fluctuation analysis (DFA), which is applicable to the analysis of long-range power-law correlations or LRD in non-stationary time-series signals. Based on measurements in a real environment, we visually confirmed that the traffic data could be projected onto the self-organizing map in accordance with the traffic properties over time, resulting in a combined depiction of the effects of the LRD and network utilization rates.

1. Introduction

Data set of recent network traffic measurements have shown that the traffic seen in actual IP networks are in fact self-similar, demonstrating fractal-like behaviors [1],[2]. The notion of self-similarity refers to the occurrence of the same patterns at different scales in finite-dimensional distributions of a time-series signal. Previous studies have revealed that aggregated traffic in the real-world network has long-range dependence (LRD), also known as long memory, in which a process is characterized by an autocorrelation function that decays with a lag time. Simulation-based studies have also shown that the LRD in the IP-network traffic can affect the network performance levels in terms of network link bandwidth and buffer responses [3].

Network traffic characteristics in real environments vary randomly over time; that is, the characteristics of the probability distributions of IP packet vary dynamically in the time domain. In [4], for example, Takayasu et al. pointed out that the network traffic behaviors change with the phase transition patterns in the time domain. Furthermore, when considering the effect of IP-network traffic on network systems, other factors, such as network utilization rates, must be taken into account. Namely, the degree of self-similarity is only one aspect of IP-network traffic behavior; we might expect that a higher LRD when network utilization rates are relatively low will have less effect on a network system than a lower LRD when network utilization rates are high. We thus need to develop methods that cover both the effects of LRD and other network factors as an appropriate basis for the analysis.

Incidentally, the previous studies presented so far have been based on the assumption that the target network traffic is stationary or wide-sense stationary. However, a time series signal is only regarded as stationary, if the mean, standard deviation, and higher moments, as well as the correlation functions, are invariant under time translation. This condition is often difficult to find in the real world. The detrended fluctuation analysis (DFA) has been introduced as a way of measuring long-range power-law correlations or LRD of non-stationary time series signals [5], [6]. This method has been successfully applied to evaluate the characteristics of data such as DNA and economic indexes.

This paper describes a DFA-based analysis of measured IP-network traffic in terms of the time variation of LRD. To take the possibility of multiple factors into consideration, this paper used a self-organizing map, which is an effective tool for clarifying the relative relationships in high-dimensional input data [7]. It lets us to depict the nonlinear statistical relationships in high-dimensional data in a two-dimensional space, without losing topological relationships in the input data. Based on measured traffic data, this paper shows that the self-organizing map is effective as a way of displaying how traffic conditions changes over time.

2. Method of analyzing measured traffic

2.1 Definition of detrended fluctuation analysis

Self-similarity in time demonstrates the presence of long-range dependence (LRD), and can affect the performance of network systems. Here, simulation-based studies have indicated that the LRD of network traffic can lower the performance levels of network systems in terms of link bandwidth [3].

This paper employs detrended fluctuation analysis (DFA) [5],[6] to analyze fluctuations in the patterns of measured traffic in terms of LRD. As was stated above, previous studies have been based on the assumption that the target traffic is stationary or at least wide-sense stationary, conditions that are often difficult to establish. The DFA method was introduced as a way of measuring the long-range power-law correlations or LRD of signals that are not necessarily stationary.

Here, let x = {x(i); i = 1, 2, 3… N} be a one-dimensional stochastic process with time i ; we define the following integrated signal y(k):

$$y(k) = \sum_{i=1}^{k} \{ x(i) - \mu \}, \quad (1)$$

where μ is the mean of x(i). Next, we divide the integrated time series y(k) into boxes of equal length n. We then find the least squares line that fits the data in each box of length n. After that, y(k) is detrended by subtracting the local trends $y_n(k)$ in the following way [5]:

$$F(n) = [(1/N) \sum_{k=1}^{N} \{ y(k) - y_n(k) \}^2]^{1/2}. \quad (2)$$

The above computation is repeated across a broad range of scales to characterize the relationship between the average root-mean-square fluctuation F(n) and the box size n. A power-law relationship between them indicates the presence of scaling given by $F(n) \sim n^{\alpha}$, which means that the process obeys the scaling law characterized by the scaling exponent α. When calculating the value of α, we check a trend line in the double-logarithmic plot of F(n) against n. Then, the α value can be defined by the slope of the best fit line in the log-log plot.

The fractal-like nature of the fluctuation is characterized by the scaling exponent α, which represents the long-range power correlation or LRD of the signal. If the target process is similar to white noise, then α is close to 0.5. If the process is correlated or persistent, $\alpha > 0.5$; if the process is anti-correlated or anti-persistent, $\alpha < 0.5$. Namely, α values increasingly greater than 0.5 indicate an increasing degree of LRD for the target time-series signal.

2.2 Analysis of measured traffic

When performing a DFA-based analysis, this paper investigates the time variation patterns of network traffic conditions. In this analysis, the traffic throughput (=the amount of traffic data per second) is measured over time, and a value of α for measured traffic is derived from the N measured samples (one data set). Also, the overlap between the consecutive data sets is M.

As stated above, focusing on LRD alone is not enough when we are considering the effects of fractal-like properties on network systems. Namely, we might expect that a higher LRD with limited network utilization rates will have less effect on a network system than a lower LRD with high network utilization rates. In this sense, it is important to check time-varying properties that significantly affect network utilization rates. Thus, in addition to the value of α, this paper checked the average throughput of measured as described in section 3.

2.3 Training based on self-organizing map

The self-organization algorithm can convert complex, nonlinear statistical relationships among multi-dimensional data into simple geometric relationships in a low-dimensional domain [7]. It calculates multi-dimensional parameters so that they optimally denote the domain in which the relationships of primary data are preserved topologically. In performing an LRD-based analysis, this paper employed a two-dimensional map to map the time variation patterns of measured traffic.

The training of this map is initialized by assigning random values to the weight vector **w** of the units. After the presentation of input vector **z**, the Euclidian distance between the input vector **z** and the weight vector **w** is computed for all units in the neural network. Assuming that i is the unit number of the output layer, the unit with the smallest distance is marked as unit c:

$$\| \mathbf{z} - \mathbf{w_c} \| = \min_i \{ \| \mathbf{z} - \mathbf{w_i} \| \}, \quad (3)$$

where $\mathbf{w_c}$ is the winner, i.e. the unit that best matches **z**. In the next step, all units in some defined spatial neighborhood around unit c are updated through the following training process:

$$\mathbf{w_i}(t+1) = \mathbf{w_i}(t) + h_{c(x),i} (\mathbf{z}(t) - \mathbf{w_i}(t)), \quad (4)$$

where t is the regression step index, and $h_{c(x),i}$ is the neighborhood function. Here, the neighborhood function of the Gaussian type can be given by

$$h_{c(x),i} = a(t) \exp\{ -\| r_i - r_c \|^2 / 2\sigma^2(t) \}, \quad (5)$$

where $0 < a(t) < 1$ is the learning-rate parameter, $r_i \in R^2$ and $r_c \in R^2$ are the vertical locations on the grid, and σ(t) corresponds to the width of the neighborhood function. Also, assuming that T is the total number of training cycles, a(t) and σ(t) can be defined as

$$a(t) = a(0) (1 - t / T), \quad (6)$$

$$\sigma(t) = \sigma(T) + \{\sigma(0) - \sigma(T)\}(1 - t / T). \quad (7)$$

The procedure of this training process is as follows:

i) initialize $\mathbf{w_i}$ to a random value,
ii) input variables to vectors **z**(t),
iii) calculate eq.(3) for all units, and find $\mathbf{w_c}$,
iv) calculate eq.(4) with the aid of eqs.(5)-(7), and
v) repeat the process from ii).

3. Case study: Analysis of measured traffic

3.1 Measurement of IP-network traffic

In this measurement, IP packets entering the NTT R&D center (in Tokyo, Japan) from the Internet were measured via a router at the terminating point of a 17-Mbps least line. Also, to measure the throughput of IP packets, we set the time resolution level of the traffic measuring device to 10 ms from 9:00 to 11:30 on Dec. 3, 2002. Then, in calculating α from sets of data, the number N of sampling points for one data set was 60,000 (10 min.), which covered time period for evaluating network systems. Also, the sampling overlap M between consecutive data sets was 30,000 (5 min.).

Examples of 1000-sample series of measured traffic throughput are shown in Fig. 1, where the amplitude of time-series data is normalized at the maximum peak level in this figure. We see that the bursty nature of IP-network

traffic is visible in the figure, indicating that seemingly self-similar features are present in the fluctuations.

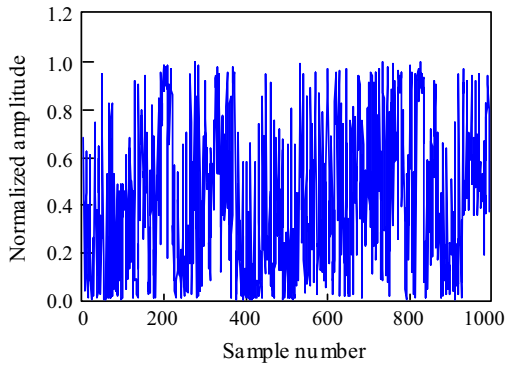


Fig. 1 Example of measured traffic.

3.2 DFA-based analysis of measured traffic

An example of a log–log plot of n vs. F(n) for measured traffic is shown in Fig. 2. As is shown in Fig. 2, different kinds of fluctuation characteristics are seen, which demonstrates that different stochastic structures with respect to scaling behavior are present in a single data set. In the figure, two solid lines represent the best fit line for ranges of around $\log_{10}(n) \leq 2.2$ and $\log_{10}(n) \geq 2.2$ in this figure, resulting in α=0.82 and 1.01 for each range.

Here, when we checked other data sets in the measurements, results confirmed that two α values can be derived from data sets in this case study, as shown in Fig. 2. Since our aim is to grasp the essential scaling tendencies of traffic data and to simplify the DFA-based analysis, the following discussion is concerned with α values defined for different ranges of $\log_{10}(n)$ as follows:

- α_1: $\log_{10}(n)$ is less than around 2.1 - 2.4,
- α_2: $\log_{10}(n)$ is more than around 2.2 - 2.5.

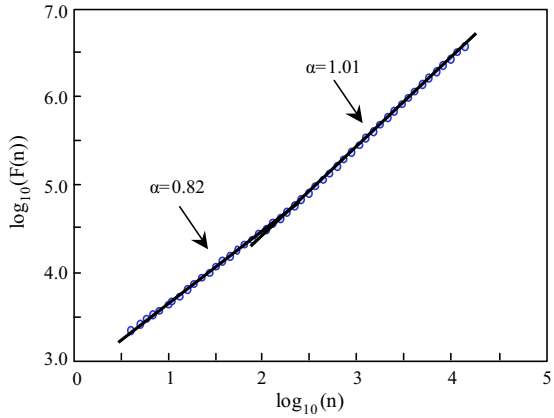
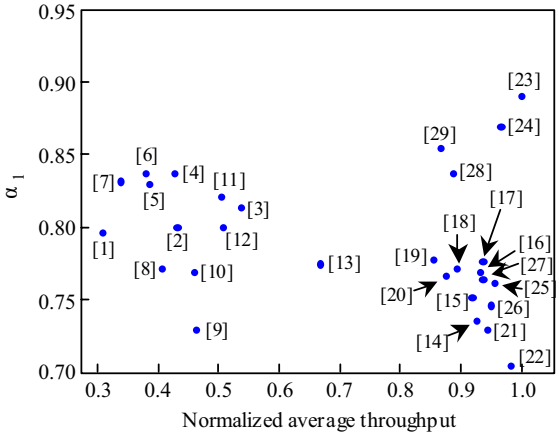


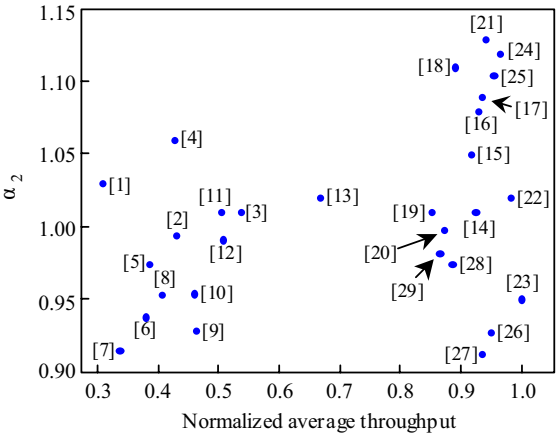
Fig. 2 Example of log-log plot of n vs. F(n).

Relationships between the average throughput and α values for measured traffic are shown in Fig. 3, where the average throughput was normalized at the maximum value in this measurement, and the order of the data set (= Data number) from the beginning is given for each measurement ([ ] represents the order of the data set.).

When we look at results in Fig. 3, we see that α_1 values ranged from around 0.70 to 0.89, while α_2 values were in the range from around 0.92 to 1.13. Namely, α_2 values tended to be greater than α_1 values, corresponding to the slopes in log-log plots of n vs. F(n) for each $\log_{10}(n)$ range. In addition, the results also show that the normalized average throughput tended to increase with passage of time, which suggests that network utilization rates tended to increase over time in this period.



(a) vs. α_1



(b) vs. α_2

Fig. 3 Relationship between normalized average throughput and α values.

3.3 Analysis using self-organizing map

We applied a self-organizing scheme to the measured traffic data and thus evaluated the time variation patterns of IP-network traffic data. In this analysis, we used three parameters of measured traffic data: the α_1 value, the α_2 value, and the normalized average throughput. The α_1 value corresponds to the degree of LRD in the lower range of n in the log-log plots of n vs. F(n), so increases in this value can lead to poorer network performance. The α_2 value corresponds to the degree of LRD in the higher range of n in the log-log plots of n vs. F(n), so increases in this value also can lower network performance. The normalized average throughput refers to the bandwidth of

the IP network; that is, increases in this quantity also lead to loss of performance for the network system.

The data set in Fig. 3 was used in training a topologically rectangular map, and the map size was set to 10 × 10. Also, based on the quantization error defined by the mean of $\| \mathbf{z} - \mathbf{w_c} \|$ [7], this paper set the total number of training for the self-organizing map to 10,000. To raise the efficiency of training, we performed the training in two phases [7]: in the first phase, with training number of 1,000, the initial value of the learning-rate parameter a(0) was set to 0.5, while in the second phase it was set to 0.05. In addition, parameters of the neighborhood radius were also altered: the initial value σ(0) and final value σ(T) in the first phase were set to 5 and 1, while the corresponding values in the second phase were 2 and 1, respectively.

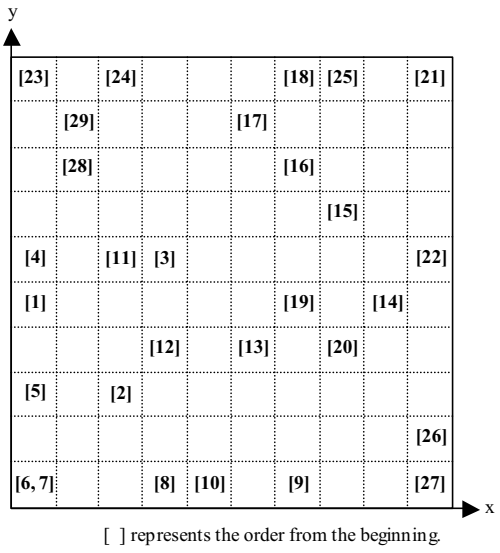


Fig. 4 Visualization results of mapped data.

Figure 4 shows map results projected onto a two-dimensional domain, in which the reference axes are set as x and y, and the order of measured data from the beginning was set. In the map of the time-based sets of traffic data, we see the following points.

• The α_1 value tended to decrease along the x-axis, and the upper-left part is equivalent to the domain where the value of α_1 was highest (resulting in the value of 23th data being the highest). On the other hand, the α_2 value tended to increase along the y-axis, and the upper central or upper-right part is equivalent to the domain where the value of α_2 was highest (resulting in the value of 21th data being the highest). The normalized average throughput tended to increase along the x- and y-axes, with the values in the upper- or right- regions being the highest of all. As a result, we estimate that the upper-part (especially around 24th data) corresponded to the domain where the patterns of measured traffic were most likely to affect the network system in terms of both LRD and network bandwidth.

• The 12 or 13 sets of data starting from 9:00 tended to be located in the lower-left part of the map along the y-axis. Therefore, the results visually demonstrate that the normalized average throughput of network traffic in this time-span was relatively low. The location point of projected data tended to shift from the lower-left part to the other part especially after around 10:00 (around the 13th or 14th data from 9:00). Therefore, we see that the network bandwidth utilization rates after around 10:00 tended to increase.

We thus visually confirmed our technique's ability to projecting traffic data onto a two-dimensional domain in a way that reflects their properties. Our method projects data with multi-dimensional input parameters onto a two-dimensional space, so that we can effectively evaluate effects of both the LRD and network utilization rates over time.

4. Conclusion

In analyzing fractal-based behaviors of actual IP network traffic, this paper applied an integrated approach to evaluating how the properties of network traffic change over time. We measured the degree of long-range dependence (LRD) of measured traffic by applying the detrended fluctuation analysis (DFA), which can measure LRD of signals that are not necessarily stationary. The map produced by the self-organizing algorithm revealed the effects of both the LRD and network bandwidth utilization rates, so we can effectively check the changes of network traffic condition over time.

Future studies will be concerned with analyzing the method's adaptability and precision for various types of traffic, how the LRD extracted by DFA affects IP applications, and so on.

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, Trans. Networking, vol. 2, no. 1, pp. 1-15, 1994.

[2] W. Willinger, M. S. Taqqu, R. Sherman, and D.V. Wilson, "Self-similarity through high-variability: Statistical analysis of Ethernet LAN traffic at the source level, " IEEE/ACM, Trans. Networking, vol. 5, no. 1, pp. 71-86, 1997.

[3] H. Furuya, M. Fukushima, H. Nakamura, and S. Nomoto, "Modeling of aggregated TCP/IP traffic on a bottleneck link based on scaling behavior," IEICE Trans. Commun., vol. E85-B, no.9, pp. 1756-1764, 2002.

[4] M. Takayasu, , H. Takayasu, and K. Fukuda, "Dynamic phase transition observed in the traffic flow," Physica A, vol. 277, pp. 248-255, 2000.

[5] C.-K. Peng, S.V. Buldrev, S. Havlin, M. Simons, H. E. Stanly, and A.L. Goldberger,, "Mosaic organization of DNA nucleotides, " Physical Review E, vol. 49, no.2, pp. 1685-1689, 1994.

[6] M. Masugi, "Self-organizing map-based analysis of IP-network traffic in terms of time variation of self-similarity," IEICE Trans. Fundamentals, vol.E87-A, no.6, pp.1546-1554, 2004.

[7] T. Kohonen, "Self-organizing maps, " Springer, Berlin, Heidelberg New York (Second extended edition), 1997.