Surrogate Analysis on the Internet Traffic Time Series

Tohru Ikeguchi† and Mikio Hasegawa††

†Saitama University
255 Shimo-Ohkubo, Saitama, Japan
Phone:+81-48-858-3577, Fax:+81-48-858-3716
Email: tohru@ics.saitama-u.ac.jp,

††Communications Research Laboratory
3-4 Hikarino-oka, Yokosuka, 239-0847, Japan
Phone:+81-468-47-5062, Fax:+81-468-47-5069
mikio@crl.go.jp

Abstract

We apply the method of surrogate data to the Internet traffic time series in order to evaluate the self-similarity. Using the iterated amplitude adjusted Fourier transformed (IAAFT) algorithm, we generate the stochastic time series data sets, which preserve empirical histograms and autocorrelation functions of the original data. Since the self-similarity depends on the autocorrelation structure, we can preserve the self-similarity of the original data. We also analyze network performance by evaluating overflow rates with a simple transmission model. Comparing the results of applying the real world traffic and IAAFT surrogates to the model, we show that both of them have almost the same effects to the transmission model.

1. Introduction

We use the Internet in everyday life. Through the Internet, huge amount of data traffic is transmitted. In case of designing such networks, it is very important to evaluate the statistical features of the network traffic. Although stochastic processes are one of the conventional frameworks for simulating the network traffic, it has been indicated that modeling based on the conventional stochastic process such as the Poisson process may fail to simulate the performance of the networks [1]. The reason is that the network traffic has the self-similarity [2, 3, 4], which cannot be reproduced by the Poisson modeling.

In this paper, we apply the method of surrogate data to the network traffic data, in order to evaluate one of the important characteristics of the Internet traffic data, the self-similarity. The method of surrogate data is one of the analysis framework for investigating possible nonlinear deterministic dynamics. We can produce a surrogate data set, which preserves the autocorrelation structure of the time series. Since the self-similarity depends on autocorrelation, we can preserve the self-similarity of the original data,

We introduce two surrogate data generating algorithms. The first is the random shuffle algorithm [5], which preserves empirical histograms of the original data, but destroys temporal structures. The second is the iterative amplitude adjusted Fourier transform (IAAFT) algorithm [5], which preserves not only the empirical histograms but also autocorrelation functions. Using these two algorithms, we compare the self-similarity index of the given data and the surrogate data sets.

We also apply these data to a simple transmission model, and show the importance of preserving temporal structures of the original traffic data for modeling network architectures.

2. Self-similarity and Surrogate data

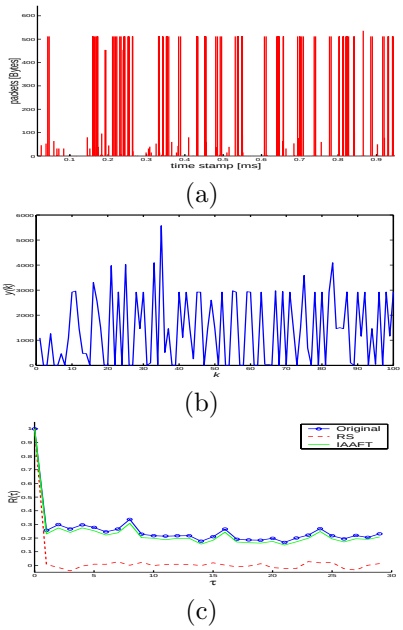
There are many discussions on existence of the self-similarity in the Internet traffic [2, 3, 4, 1]. Since the self-similarity of the time series depends on the autocorrelation structures, if we can use any algorithms which can preserve the autocorrelation structures, we expect that the algorithm also preserve the self-similarity.

For nonlinear time series analysis, the method of surrogate data [5] has been utilized for evaluating uncertainty of obtained results. Surrogate data are stochastic data which have the same statistical property as the original data. There is a wide variety of algorithms for generating surrogate data [5]. One of them, the IAAFT algorithm is an improved version of the amplitude adjusted Fourier transform algorithm [5], which is based on the Fourier transform algorithm that can preserve autocorrelation structures perfectly. Then, we use the IAAFT algorithm for a tool to preserve the self-similarity of the Internet traffic data.

In order to confirm the above concept, we apply the IAAFT algorithm to the real traffic data. Let us denote the observed packet count as $x(t_i)$ where $t_i (i = 1, 2, \ldots, n)$ is the i-th timestamp at time t_i. For the sake of simplicity, we transform $x(t_i)$ to uniformly re-

sampled time series $y(k)(k = 1, 2, \ldots, N)$ with the bin size 0.01[sec]. Namely, the value $y(k)$ is the total amount of packet counts included in the k-th bin.

We show the observed traffic data $x(t_i)$ in Fig. 1(a), and the transformed time series $y(k)$ in Fig. 1(b). We calculate the autocorrelation functions of the original data, and surrogate data produced from two algorithms. We compare the results in Fig. 1(c). The original data (lbl-pkt-4) are available at the Internet cite (http://ita.ee.lbl.gov). In Fig. 1(c), it is clearly shown that an IAAFT surrogate has almost the same autocorrelation function as the original.



(a)

(b)

(c)

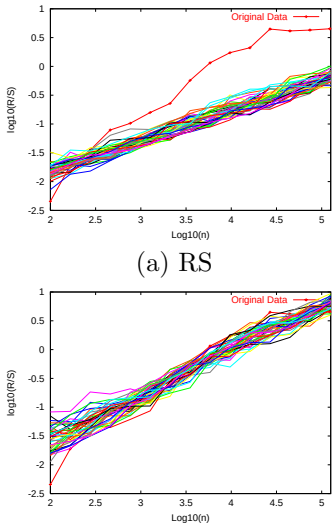
Figure 1: (a) The observed Internet traffic data $x(t_i)$, (b) the uniformly resampled time series $y(k)$, and (c) autocorrelation functions of the original and its surrogates.

We compare the strength of self-similarity of the real traffic time series and those of the surrogate data sets. Figure 2 is the R/S analysis plots which are the rescaled adjusted range plots. We draw the relationships between the number of data points n and $R(n)/S(n)$ in the log-log scale, where $R(n)$ is the adjusted range statistic of the process and $S(n)$ is a sample standard deviation.

Namely, for time series $y(k)(k = 1, 2, \ldots, N)$, the R/S statistic is given by:

$$R(n)/S(n) = [\max(0, W_1, W_2, \ldots, W_n) - \min(0, W_1, W_2, \ldots, W_n)]/S(n), \quad (1)$$

where $W_k = (y(1) + y(2) + \ldots + y(k)) - k\bar{y}$, $\bar{y}$ is a sample mean of $y(k)(k = 1, 2, \ldots, N)$.



(a) RS

(b) IAAFT

Figure 2: The R/S plots of the original traffic data time series and its RS and IAAFT surrogate data sets.

In Fig. 2(a), we compare the R/S plots of the traffic data with those of random shuffled surrogates. We use the same original traffic data as shown in Fig. 1(a)(lbl-pkt-4). The solid line with pluses is the analysis results by the R/S for the original data. Dashed lines are those for random shuffled surrogates. We can clearly see the difference between the original data and surrogates, which means that the random shuffled surrogates could not preserve the self-similarity of the original data. It should be noted that the slope of the original data is steeper than the slopes of the surrogate data sets. It means the Internet traffic has larger strength of self-similarity than the random shuffled surrogates. Namely, we can destroy one of the intrinsic nature of the Internet traffic, the self-similarity, by the random shuffle algorithm.

On the other hand, in case of applying the IAAFT algorithm, the R/S statistics of the original data are completely included in the distribution of those of its surrogate data. Namely, we observe the same slope for the traffic data and its surrogates. With these results, we clarify that the IAAFT surrogate can preserve the self-similarity of the original data.

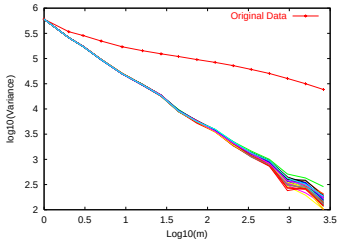
In Fig.3, we show the other results with variance time plots, which is also a popular method for evaluating fractal property of the time series. We use the following local average to the original time series $y(k)$:

$$y^{(m)}(k) = (y(km - m + 1) + \cdots + y(km))/m. \quad (2)$$

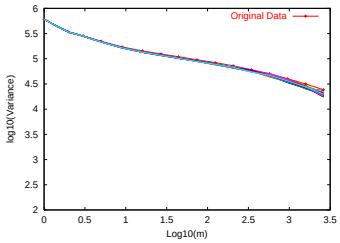
Then we plot the relations m and variances of $y^{(m)}(k)$

in a log-log plot. If the slope of the graph is less steep, the time series data exhibit stronger self-similarity.

From the results of Fig.3, in case of using IAAFT algorithm, we cannot see any differences between the original data and surrogate data sets (Fig.3(b)). In case of applying random shuffle algorithm (Fig.3(a)), the autocorrelation structures are destroyed, then there is no signature on preserving the self-similarity.



(a) RS



(b) IAAFT

Figure 3: The variance time plots of the original traffic data and RS and IAAFT surrogate data sets.

3. Network Performance and Surrogate Data

We introduce a simple transmission model, which has a finite buffer size (Fig. 4). We apply data series to the system from the left side. Then, the data are pushed out from the right side with a fixed transmission rate. When the inputs are larger than the possible transmission rate at the output side, the system keeps the data in the buffer. Once the data is stored in the buffer, they are put out with a constant rate, which corresponds to the possible maximum transmission rate at the output side. If the amount of the inputs are not so large, all the input data are smoothly transmitted through the output. However, if the input data are unexpectedly large

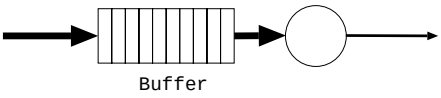
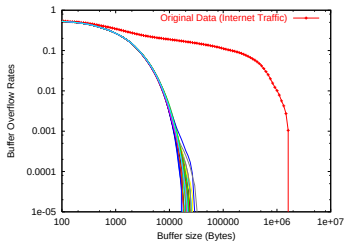


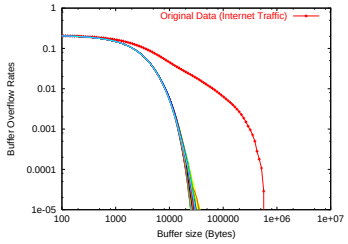
Figure 4: A transmission model.

and continuously coming into the system, the input data will overflow and the traffic may not be correctly transmitted through the system. Changing the buffer size, we evaluate the buffer overflow rates for the real traffic data and the surrogate data sets. Although, we use such a simple model, the essential idea of the input-output relation described by the model is very plausible and widely acceptable as a model for the real transmission system.

First, we compare the buffer overflow rates of the traffic data with random shuffled surrogate. We introduce two data taken at different time and different position, lbl-pkt-4 and dec-pkt-2. The results are shown in Fig. 5. It is clear that results of the random shuffled surrogate show completely different characteristics from the original traffic. We find that buffer overflow rates of the random shuffled surrogate are much smaller than the original data. It means that it is impossible to simulate the same characteristic by preserving only the empirical histogram, the average and the variance. From Fig. 5, we also find that the required buffer size for small buffer overflow rates ($< 10^{-5}$) is 100 times larger for the real traffic data. Therefore, for simulating communication systems, it is not appropriate to use a stochastic model with which the empirical histogram, the average or the variance of the traffic amount are only preserved.



(a) lbl-pkt-4



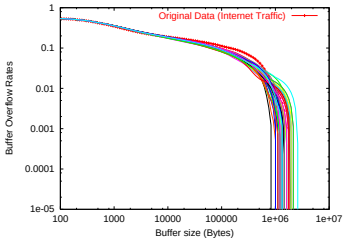
(b) dec-pkt-2

Figure 5: Buffer overflow rates of the two traffic data, and random shuffle surrogates.

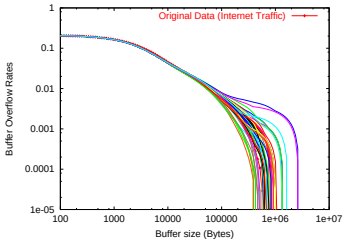
In Fig. 6, results of the traffic data is compared with those of its IAAFT surrogate data sets. In this case, the buffer overflow rates are almost the same for the original data and the surrogates. It means that the au-

tocorrelation structure is very important for analyzing communication systems.

In Fig. 7, we show the relationships between differences on autocorrelation and buffer overflow rates. Here, we evaluate two differences by calculating the distances between the original data and surrogate data sets. From 7, we see that if the differences on autocorrelation between the original and the surrogates are small, there are small differences on buffer overflow rates between them. Namely, preserving the autocorrelation, we can realize almost the same characteristic property as the original traffic. The results also support the suggestions of the conventional researches, which insist the importance of the self-similarity of the Internet traffic data.



(a) lbl-pkt-4



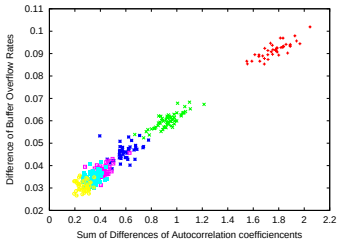
(b) dec-pkt-4

Figure 6: Buffer overflow rates of the original traffic data, and those of its IAAFT surrogates.

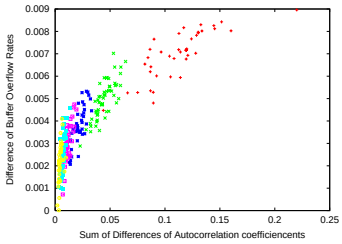
4. Conclusions

In this paper, we show the importance of preserving autocorrelation structures, and the self-similarity, of the Internet traffic data, using the method of surrogate data. Our results are consistent with the conventional results, which show the existence of the self-similarity.

Since the second-order characteristics of the IAAFT surrogates are almost the same as the original traffic, we can have very good candidates for constructing a new traffic model. Even if we only observe a single sample of traffic data, we can generate as many surrogate data sets as we want. Our results also show the plausibility and the reliability of the analysis framework for a traffic modeling with the method of surrogate data, which has been used in the field of nonlinear time series analysis.



(a) lbl-pkt-4



(b) dec-pkt-4

Figure 7: Dscrepancy between the original and the IAAFT surrogate data sets.

The authors would like to thank Professor H. Yashima, Saitama University, for his valuable comments and discussions. The research was partially supported by Grant-in-Aids for Scientific Research (C) from JSPS and for Scientific Research on Priority Areas from MEXT.

References

[1] 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.

[2] W. Leland et al., "On the self-similar nature of Ethernet traffic (extended version)," *IEEE/ACM Trans. Networking*, vol. 2, no. 1, pp. 1–15, 1994.

[3] K. Park et al., "On the effect of traffic self-similarity on network performance," in *Proc. of the SPIE International Conference on Performance and Control of Network Systems*, 1997, pp. 296–310.

[4] Z. Sahinoglu and S. Tekinay, "On multimedia networks: Self-similar traffic and network performance," *IEEE Communications Magazine*, vol. 37, pp. 48–52, 1999.

[5] T. Ikeguchi, T. Yamada, and M. Komuro, *Fundamentals and Applications of Chaotic Time Series Analysis*, K. Aihara, Ed. 2001, Sangyo Tosho Ltd. (in Japanese).