Analysis of Power-law Correlation of E-mail Sending Demands

Yoshitsugu Matsubara, Yasuhiro Hieida, and Shin-ichi Tadaki

Computer and Network Center, Saga University

Abstract

Power-law correlations have been observed in the Internet packet flow. Demands for internet services will be one of the origins. We observe sending demands for e-mail services, because they are free from the structure of the Internet. Power-law correlations are observed up to some months by using the detrended fluctuation analysis and the power-spectrum analysis. Power-law correlation in year length is not detected clearly.

1 Introduction

The Internet is one of important communication media in modern societies. The structure of the Internet and dynamical phenomena on it have been attracting research interests. The structure of the Internet has been reported to be scale free [1]. Packet flow in the Internet has been observed to bear power-law fluctuations [2].

There are some candidates of the origins of the power-law fluctuations in the Internet traffic. One is the interaction between the internal dynamics of the Internet and the structure. The other is the external effects of demands for Internet services.

We focus our attention to sending demands for e-mail services. Users in an organization usually send their e-mail requests to their organizational e-mail servers. Those e-mail servers, namely, receive sending demands mostly from their internal network. Therefore the time sequences of those demands are not affected by scale-free properties of the Internet.

E-mail service is one of the most used Internet services. Almost all members of an organization use e-mail services for their daily communication. So large volume of records is available. And e-mail services are operated stably because they are ones of the basic Internet services. So records long enough for analyzing power-law correlations are available.

We analyzed the power-law fluctuations in e-mail sending demands in the previous report [3], where the length of the data was about five months. Shibata and Murakami analyzed correlations between three network services including e-mail sending demands with a data set of some months [4]. The existence of long-range power-law correlation has not been clarified.

In this report, we analyze e-mail sending demands in a data set of three year length for obtaining detailed properties. Two types of data are analyzed. The first is the data flow by sending mails. The second is the number of demands.

2 The Data

We observe long-range correlations in e-mail sending demands for the main mail server in Saga university. Saga university has five faculties with about 2,400 staffs (including faculty members) and 7,300 students. The length of the data is three years from 9-May-2008 to 8-May-2011.

The records contain time stamps and sizes. The temporal resolution is one-second. The shorter range than one-hour contains random originated from originated asynchronous behavior of users and e-mail applications. And the number of sending demands is not large enough. So we use one-hour average of the flow and the number of demands for our analysis.

The data set contains some defects originating from weekly system maintenances, annual power unit maintenances and so on. Those short-time

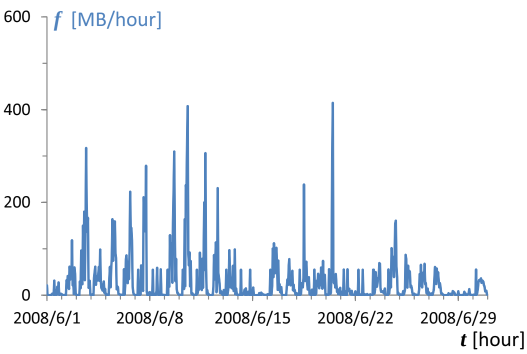

Fig.1: A sample of the data flow $f(t)$, from 1-Jun-2008 to 30-Jun-2008. The horizonal axis denotes time (hour). The vertical axis denotes data flow (Mega Byte per hour).

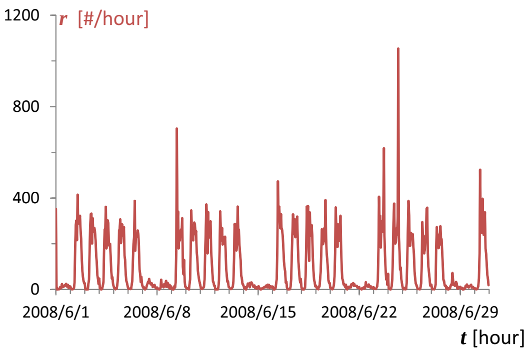

Fig.2: A sample of the sending requests $r(t)$ from 1-Jun-2008 to 30-Jun-2008. The horizonal axis denotes time (hour). The vertical axis denotes the number of sending requests (# per hour).

defects except weekly system maintenances will not affect long-range properties discussed here.

The data flow (Mega Byte per hour) by e-mail sending demands is denoted as $f(t)$. A part of its time sequence is shown in Fig. 1. Periodicity of one-week and one-day periods is recognized. Burst phenomena are also observed.

The number of sending demands per hour for e-mail sending demands is denoted as $r(t)$. A part of its sequence is shown in Fig. 2. Periodic motions and bursts are also recognized.

3 Detrended Fluctuation Analysis

The Detrended Fluctuation Analysis (DFA) is introduced to analyze non-stationary series [5, 6]. The simple form of DFA is described as follows. Consider a temporal data $u(t)$ $(0 \leq t < T)$, where T is the length of $u(t)$. The profile $y(t)$ of $u(t)$ is defined as the accumulated deviation from the average $\langle u \rangle$ of $u(t)$.

The profile $y(t)$ is divided into T/l non-overlapping segument of length l. The local trend $\tilde{y}_n(t)$ in n-th segment is defined by applying the linear least-squares method to $y(t)$ in each segment. So this simplest method is called the 1st-order DFA. The detrend profile $y_l(t)$ is defined as the deviation of $y(t)$ from $\tilde{y}_n(t)$ as $y_l(t) = y(t) - \tilde{y}_n(t)$.

The root mean square deviation $F(l)$ is defined as

$$F(l) = \sqrt{\frac{1}{T}\sum_{t=0}^{T-1} y_l^2(t)}. \qquad (1)$$

If $F(l)$ behaves as $F(l) \sim l^{\alpha}$ $(1/2 \leq \alpha \leq 1)$, the data $u(t)$ contains power-law fluctuations. The power-spectrum $P(k)$ of $u(t)$ obeys $P(k) \sim k^{\beta}$, where $-\beta = 2\alpha - 1$. The value $\alpha = 1/2$ means that $u(t)$ is random and uncorrelated. The sequence with $\alpha = 1$ bears $1/f$ fluctuations.

4 Extracting Periodicity

The power-spectrum of r is shown in Fig. 3. As observed in the time sequence (Fig. 2), peaks corresponding to one-week and one-day appear clearly. In DFA results, those periods appear as bending areas [7]. The DFA result for r is shown in Fig. 4. We can see clear bending features at one-day and one-week lengths. Periodic motions of one-week and one-day appear also in the analysis of f.

Bending shapes in DFA results corresponding to periodicity in the data prevent us to analyze the power-law exponent. We employ the same method for extracting weekly and daily periodicity as used in Ref. [8]. For the case of f, the modified sequence f'' is constructed by extracting the average weekly motion $\bar{f}_{\text{week}}$.

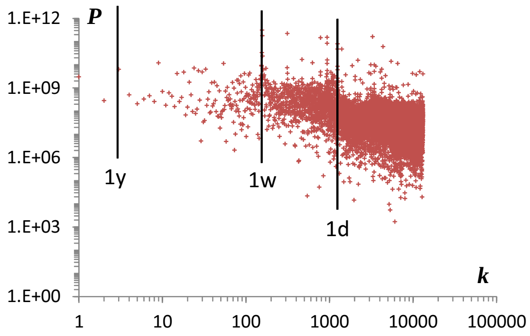

Fig.3: The power-spectrum for r. Three vertical lines correspond to one-year, one-week, and one-day respectively. The horizonal axis denotes the wave number.

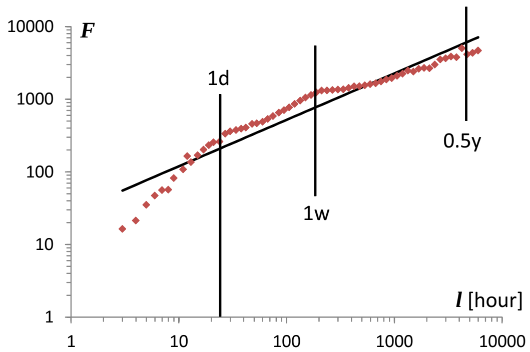

Fig.4: The DFA analysis for r. The horizonal axis denotes the length of segment (hour). The vertical axis denotes the calculated value of DFA. Three vertical lines correspond to one-day, one-week, and half year respectively. There are two bending areas corresponding to one-day and one-week.

$$f''(t) = f(t) - \bar{f}_{\mathrm{week}}(t \bmod T_{\mathrm{week}}), \tag{2}$$

$$\bar{f}_{\mathrm{week}}(\tau_{\mathrm{w}}) = \frac{1}{W}\sum_{w=0}^{W-1} f(w \times T_{\mathrm{week}} + \tau_{\mathrm{w}}), \tag{3}$$

where $0 \le \tau_{\mathrm{w}} < T_{\mathrm{week}}$ ($= 24 \times 7$ hours), and W is a number of weeks in the data. By extracting the average weekly motion, daily periodicity is also removed. We also define r'' as the modified sequence by extracting the average weekly motion from r.

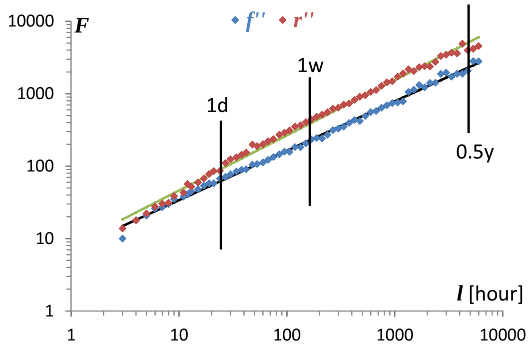

Fig.5: The DFA analysis for f'', r''. The horizonal axis denotes the length of segment (hour). The vertical axis denotes $F(l)$. Three vertical lines correspond to one-day, one-week, and half year respectively. The exponent of f'' is $\alpha \simeq 0.68$. The exponent of r'' is $\alpha \simeq 0.76$.

5 Long-range Correlations

The results of DFA analysis for f'', r'' are shown in Fig. 5. The exponent for f'' is $\alpha \simeq 0.68$ by fitting with the least square method for all points. The exponent for r'' is $\alpha \simeq 0.76$.

Almost all data points of DFA for the flow f'' can be fitted by a straight line. So the power-law correlation certainly exists up to half year for the flow.

For the results of DFA for the demands r'', we can still see a bending area near one-day. And we can see the deviation from the straight line in the long range area. The power-law correlation seems to exist up to some months.

In general, the power-spectrum is used to analyze the correlation of time series. The analysis by using the power-spectrum is described to show the differences from the DFA result.

The power-spectrum of f'' is shown in Fig. 6. The peaks corresponding to one-year and one-week and one-day are not recognizable.

The power-spectrum of r'' is shown in Fig. 7. The peaks corresponding to one-week and one-day are not recognized. There is a peak corresponding to one-year. The spectrum seems to be divided into two area at $k = 156$ corresponding to one-week. From the spectrum in the longer area than a week, the absolute value of the exponent is smaller

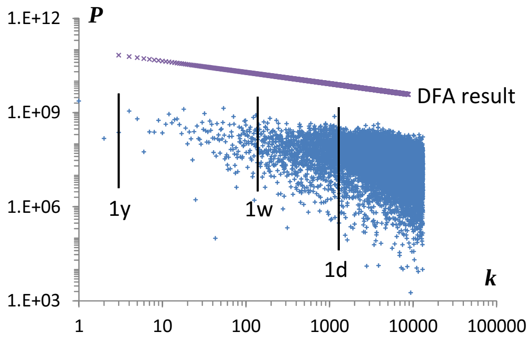

Fig.6: The power-spectrum of f''. Three vertical lines correspond to one-year, one-week, and one-day respectively. The periodicity of one-week is eliminated. The peaks corresponding to one-year and one-week and one-day are not recognized. The line corresponding to the DFA result is plotted to compare the power-spectrum result.

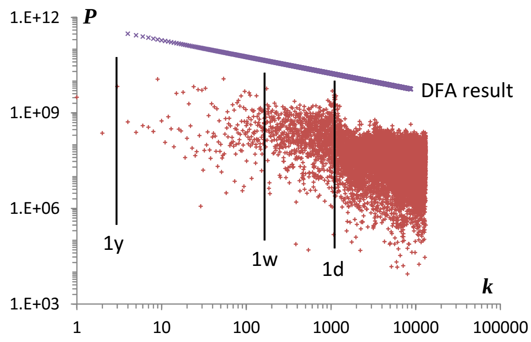

Fig.7: The power-spectrum of r''. Three vertical lines correspond to one-year, one-week, and one-day respectively. The periodicity of one-week is eliminated. The peaks corresponding to one-week and one-day are not recognized. The line corresponding to the DFA result is plotted to compare the power-spectrum result.

than that from the DFA results.

6 Summary

We analyze time sequences of e-mail sending demands for observing behavior of uses for Internet services, because e-mail sending demands are not affected by the structure of the Internet. We analyzed two types of demands: the flow and the number of requests.

We observed power-law correlations in both sequences by DFA. The correlations range up to some months. For sending demands, especially, long range correlations up to one-year is not clarified.

References

[1] M. Faloutsos, M. P. Faloutsos, and C. Faloutsos: ACM SIGCOMM Comp. Commun. Rev. **29** (1999) 251.

[2] K. Fukuda, H. Takayasu, and M. Takayasu: Physica A **287** (2000) 289.

[3] Y. Matsubara, Y. Hieida, and S. Tadaki: Proc. Sym. Sim. Traffic Flow, **14** (2008) 73.

[4] A. Shibata, T. Murakami: Proc. Symp. Sim. Traffic Flow, **16** (2010) 17.

[5] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger: Phys. Rev. E **49** (1994) 1685.

[6] C.-K. Peng, S. Havlin, H. E. Stanley, and A. L. Goldberger: Chaos. **49** (1995) 82.

[7] K. Hu, P. Ch. Ivanov, Z. Chen, P. Carpena and H. E. Stanley, Phys. Rev. E 64 (2001) 011114.

[8] S. Tadaki: J. Phys. Soc. Jpn. **76** (2007) 044001.