

Scaling and memory in recurrence intervals of Internet traffic

SHI-MIN CAI^{1,2}, ZHONG-QIAN FU¹, TAO ZHOU^{2,3(a)}, JUN GU¹ and PEI-LING ZHOU¹

¹ *Department of Electronic Science and Technology, University of Science and Technology of China
Hefei Anhui, 230026, PRC*

² *Department of Physics, University of Fribourg - Chemin du Musée 3, 1700 Fribourg, Switzerland*

³ *Department of Modern Physics, University of Science and Technology of China - Hefei Anhui, 230026, PRC*

PACS 89.75.-k – Complex systems
PACS 89.75.Da – Systems obeying scaling laws
PACS 89.20.Hh – World Wide Web, Internet

Abstract – By studying the statistics of recurrence intervals, τ , between volatilities of Internet traffic rate changes exceeding a certain threshold q , we find that the probability distribution functions, $P_q(\tau)$, for both byte and packet flows, show scaling property as $P_q(\tau) = \frac{1}{\bar{\tau}} f(\frac{\tau}{\bar{\tau}})$. The scaling functions for both byte and packet flows obey the same stretching exponential form, $f(x) = A \exp(-Bx^\beta)$, with $\beta \approx 0.45$. In addition, we detect a strong memory effect that a short (or long) recurrence interval tends to be followed by another short (or long) one. The detrended fluctuation analysis further demonstrates the presence of long-term correlation in recurrence intervals.

Many complex systems are characterized by heavy-tailed distributions, such as power-law distributions [1], lognormal distributions [2], and stretched exponential distributions [3]. These distributions imply a nontrivial probability of the occurrences of the extreme events. Statistical laws on these extreme events provide evidence for the understanding of the mechanism that underlies the dynamical behaviors of the corresponding complex systems. Recently, some typical complex systems, such as earthquakes [4–8], financial markets [9–12] and many other natural hazards [13], have been widely investigated. Taking the earthquakes for example, excluding the well-established *Omori Law* [14] and *Gutenberg-Richter Law* [15], the scaling laws for temporal and spatial variability of earthquakes have been observed by Bak *et al.* [4] and Corral [5–7], and the memory effect in the occurrence of the earthquakes is revealed by showing the statistics of the recurrence times above a certain magnitude [8].

The Internet has been viewed as a typical complex system that evolves in time through the addition and removal of nodes and links, and empirical evidence has demonstrated its small-world and scale-free structural properties [16,17]. One of the research focuses, the Internet traffic, has been widely studied by computer scientists, physicists and beyond. For instance, Leland *et al.*

first found the self-similar nature and long-range dependence of Ethernet traffic that have serious implications for the design, congestion control, and analysis of computer communication networks [18]. After that, several traffic models are proposed to understand the underlying mechanism for information transport and congestion control of the Internet traffic [19–21], especially, those models (see refs. [19,20] about the models and the ref. [21] about the time series analysis) can, to some extent, reproduce the self-similar nature of the Internet traffic, which indicates the existence of burstiness of traffic and the large volatility of traffic rate changes.

Herein we are interested in the large volatility that implies the suddenly drastic changes of traffic rate. In previous studies, by analyzing a set of time series data of round-trip time, Abe and Suzuki [22,23] reported that the drastic changes, named of *Internet quakes*, are characterized with the Omori Law and Gutenberg-Richter Law. By statistical analysis on recurrence interval τ between the volatility of traffic rate changes exceeding a certain threshold q , this letter reports that:

- i) the probability distribution functions (pdfs) $P_q(\tau)$ for both byte and packet flows, rescaled by the mean recurrence interval $\bar{\tau}$, yield scaling property that the scaling function $f(x)$ follows a stretching exponential form $f(x) = A \exp(-Bx^\beta)$ with $\beta \approx 0.45$ for all data;

^(a)E-mail: zhutou@ustc.edu

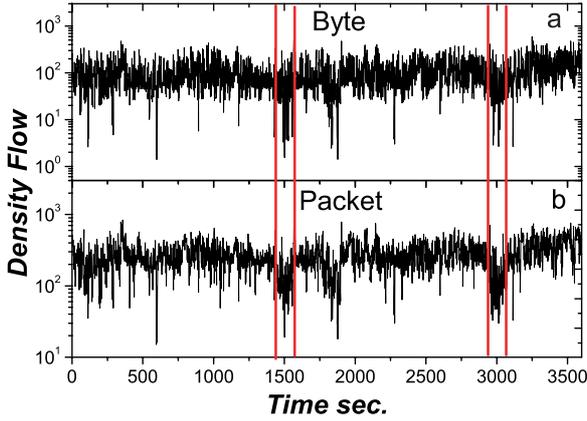


Fig. 1: (Color online) The density flows of both bytes and packets in a second resolution. Between red lines, the clustering behaviors are observed.

- ii) a short/long recurrence interval tends to be followed by another short/long recurrence interval, implying a strong memory effect.

The data used in this letter are part of Ethernet traffic set collected at Bellcore. They correspond to one “normal” hour’s worth of traffic, collected every 10 milliseconds, hence resulting in a time series with a length of 360000. There are two types of measurements, recording the number of bytes and packets per unit time, respectively. The data and information can be found at the Internet traffic archive [24]. These data have been widely used and become the most important benchmark data in relevant areas. We firstly integrate the time series into a second resolution, that is, each integrated data point is an average of 100 original points. As shown in fig. 1, the time series exhibit the clustering phenomenon that results from the traffic congestion in the Internet. For both byte and packet flows, we use the absolute value of changes, $|\Delta S_i| = |S_i - S_{i-1}|$, where S_i denotes the data point at time i , to quantify the volatility. It has been demonstrated that the pdf of ΔS_i decays in an asymptotic power-law form and the volatilities are long-term correlated [25]. We then normalize the volatility time series by the standard deviation $(\langle |\Delta S_i|^2 \rangle - \langle \Delta S_i \rangle^2)^{1/2}$. In this way, the threshold q are in units of the standard deviation of the volatility. An illustration of recurrence interval τ is shown in fig. 2.

For the normalized data, the smaller value of threshold, q , does not suggest the recurrence interval between the large volatilities (extreme events), therefore we concentrate on the cases with $q \geq 1$. As shown in fig. 3, for both byte and packet flows, the number of recurrence intervals decays exponentially fast as the increasing of q . For large q , the results are inaccurate and incredible for the low statistics, and thus we mainly discuss the statistics for a very limited range of q . Figures 4(a) and (b) present the behaviors of the pdfs, $P_q(\tau)$, for both byte and packet flows with different q , which are obviously broader than the

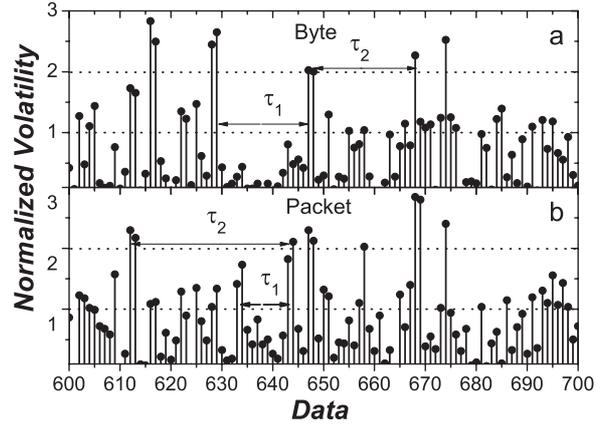


Fig. 2: Illustration of recurrence intervals, τ_q , of normalized volatility time series with $q = 1$ and $q = 2$.

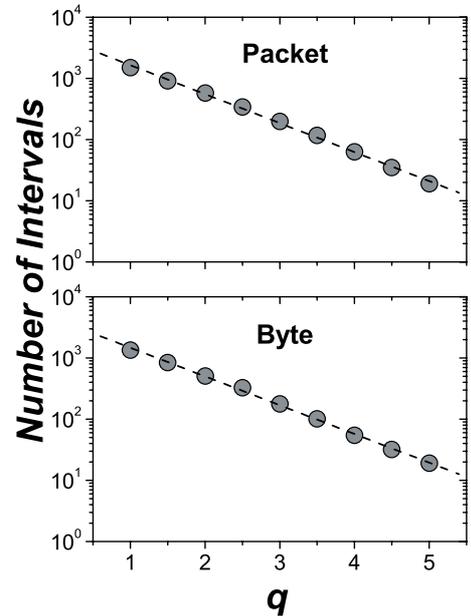


Fig. 3: Number of intervals *vs.* the threshold q in linear-log plots. Both of the two fitting lines are of slopes 0.474.

Poisson distributions as for uncorrelated data, and the pdf for larger q decays slower than that for smaller q . To understand how $P_q(\tau)$ depends on q , in figs. 4(c) and (d), we show the rescaled pdfs, $P_q(\tau)\bar{\tau}$, for byte and packet flows as functions of the rescaled recurrence intervals $\frac{\tau}{\bar{\tau}}$ (see also a similar method to uncover the scaling property by using τ_{max} as the normalized factor instead of $\bar{\tau}$ [26]). The data collapse to a single curve, indicating a scaling relation

$$P_q(\tau) = \frac{1}{\bar{\tau}} f\left(\frac{\tau}{\bar{\tau}}\right), \quad (1)$$

which suggests that the scaling function does not directly depend on the threshold q but only through $\bar{\tau} = \bar{\tau}(q)$. Furthermore, as shown in figs. 4(c) and (d), the scaling functions for both byte and packet flows follow the same stretching exponential form, $f(x) = A \exp(-Bx^\beta)$ with $\beta \approx 0.45$, indicating a possibly universal scaling property

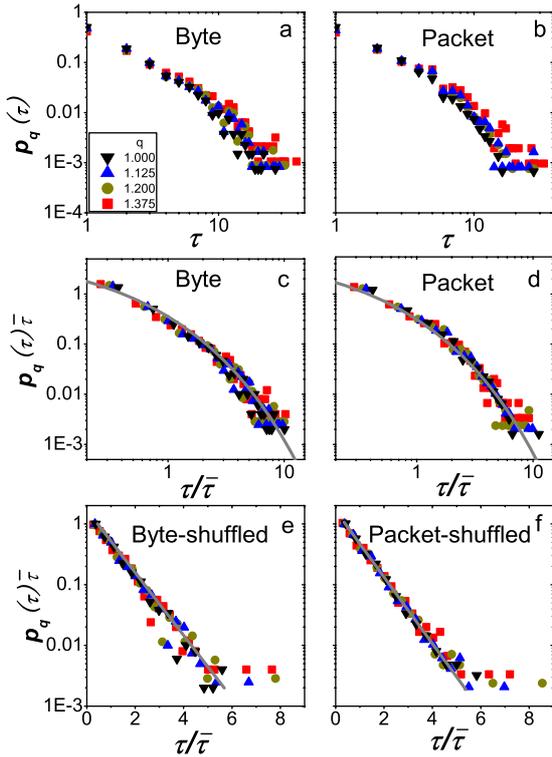


Fig. 4: (Color online) (a) and (b): distributions of recurrence intervals between two consecutive volatilities of traffic rate changes above some thresholds. (c) and (d): rescaled distributions $P_q(\tau)\bar{\tau}$ vs. rescaled recurrence intervals $\frac{\tau}{\bar{\tau}}$. The solid lines denote a stretching exponential function $f(x) = A\exp(-Bx^\beta)$ with $\beta \approx 0.45$. (e) and (f): rescaled distributions $P_q(\tau)\bar{\tau}$ vs. $\frac{\tau}{\bar{\tau}}$ for shuffled data, which obey an exponential form $f(x) = A\exp(-Bx)$. The rescaled distributions of real data have heavier tails than the ones of shuffled data.

in the recurrence intervals of Internet traffic data (see also a similar scaling in the intertrade time in financial markets [27]). Therefore, one can estimate $P_q(\tau)$ for an arbitrary q with the knowledge of $P_{q'}(\tau)$ for a certain q' . This scaling property is particularly significant for the understanding of the statistics of large- q case where the number of data is usually very small. Similar analysis for very large q has also been done, as shown in fig. 5, the rescaled distributions get much closer to each other than the original distributions. However, it is hard to tell whether these curves collapse a single master curve since the number of intervals for large q is very small. Hereinafter, we focus on the statistics for $q \in [0, 1.375]$.

The stretching exponential distributions of rescaled recurrence intervals suggest the existence of correlation of volatilities. In contrast, the recurrence intervals for uncorrelated time series are expected to follow a Poisson distribution, as $\log f(x) \sim -x$. To confirm this expectation, the volatilities are shuffled to remove the correlations, and the resulting distributions, as shown in figs. 4(e) and (f), decay in an exponential form, which is remarkably different from that of the real time series. Furthermore, the very short and very long recurrence intervals occur more

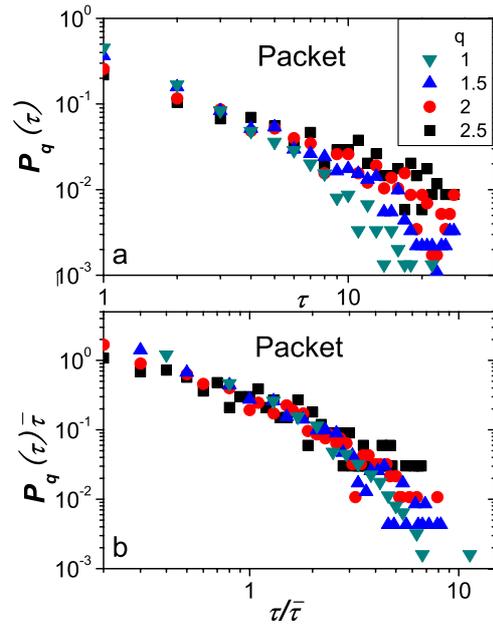


Fig. 5: (Color online) Original distribution (a) and the rescaled distribution (b) of recurrence intervals between two consecutive volatilities for packet flow. Here, the range of threshold q is much larger than that in fig. 4. The case for byte flow is almost the same, thus it is omitted here.

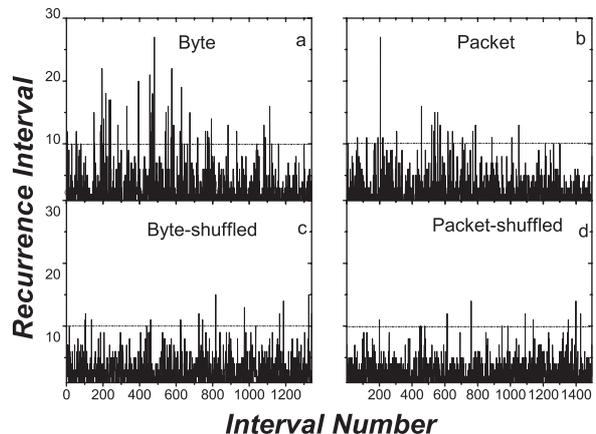


Fig. 6: (a) and (b): typical examples of recurrence intervals for byte and packet flows with $q=1$. (c) and (d): same as (a) and (b), except that the original volatility time series are shuffled. The horizontal line is used to indicate the cluster of large recurrence intervals.

frequently in the real data (see figs. 4(c)–(f)), indicating a burstiness of Internet traffic, similar as observed in many other complex systems [28].

The scaling property of $P_q(\tau)$ of recurrence intervals only indicates the long-term correlations of volatility time series of traffic rate changes, but does not tell if the recurrence intervals are themselves correlated. To answer this question, we next investigate the memory effect in the recurrence intervals. Before that, we show a typical example of recurrence intervals for both byte and packet flows in figs. 6(a) and (b), as well as the corresponding

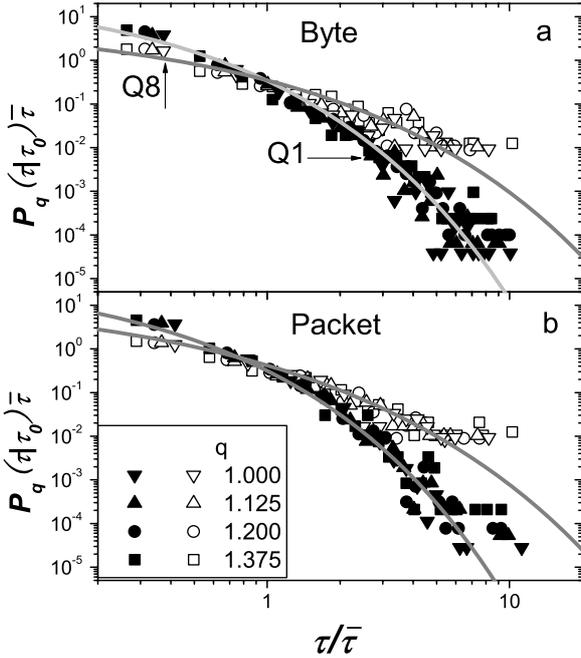


Fig. 7: Conditional pdfs, $P_q(\tau|\tau_0)$, of recurrence intervals of byte and packet flows for τ_0 in Q_1 (full symbols) and Q_8 (open symbols) as functions of $\frac{\tau}{\bar{\tau}}$. The fitting curves follow the stretched exponential forms.

shuffled sequences in figs. 6(c) and (d). Compared with the shuffled ones, the original data exhibit the clustering of large intervals, which indeed indicates the memory effect that a short (or long) recurrence interval tends to be followed by another short (or long) one.

To quantify the memory effect, we study the conditional pdf $P_q(\tau|\tau_0)$, representing the probability a recurrence interval, τ , immediately follows a recurrence interval, τ_0 . If no memory effect exists, $P_q(\tau|\tau_0)$ will be identical to $P_q(\tau)$ and independent to τ_0 . Therefore, we study $P_q(\tau|\tau_0)$ not for a specific τ_0 , but for a range of τ_0 values. Analogous to the analysis of daily volatility return intervals [11], the data set of recurrence intervals are sorted in increasing order and divided into eight subsets, Q_1, Q_2, \dots, Q_8 , so that each subset contains $1/8$ of the total data. It makes the $N/8$ lowest recurrence intervals belong to Q_1 , whereas the $N/8$ largest ones belong to Q_8 , where N denotes the total number of data points. Figures 7(a) and (b) show $P_q(\tau|\tau_0)$ for byte and packet flows. The distribution corresponding to Q_1 is obtained by recording all the τ values (they form a distribution) if their predecessor, τ_0 is no less than the smallest interval in Q_1 and no more than the largest interval in Q_1 (see ref. [11] for more details). As shown in fig. 7, the rescaled pdfs, $P_q(\tau|\tau_0)\bar{\tau}$ for different q , collapse into a single curve that can also be fitted by stretching exponential functions. The remarkable difference between the distributions with τ_0 in Q_1 and Q_8 clearly demonstrated the existence of memory effect.

To check whether the memory effect is limited only to the neighboring recurrence intervals, we use the detrended

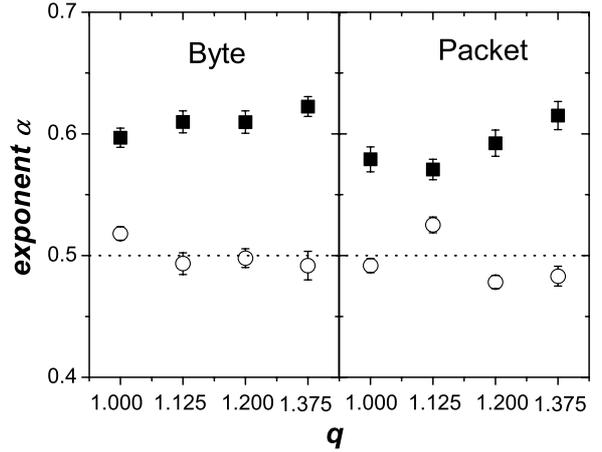


Fig. 8: Long-term memory in recurrence intervals of byte and packet flows. The full and open symbols represent the exponent α for the real data and the shuffled data, respectively. The mean value of α for the real data is 0.600 with mean error 0.009, while for the shuffled data it is 0.497 with mean error 0.007.

fluctuation analysis (DFA), which is a benchmark method to quantify long-term correlations (see ref. [29] for the original method, as well as ref. [30] and ref. [31] for the effects of trends and nonstationarities, respectively). The fluctuation $F(l)$ of a time series and window of l seconds, computed by DFA, follows a power-law relation as $F(l) \sim l^\alpha$. Uncorrelated time series corresponds to $\alpha = 0.5$, while the larger (or smaller) α indicates long-term correlation (or anti-correlation). Figure 8 shows the values of α for recurrence intervals, which are all larger than 0.5 and whose mean value is 0.600 with mean error 0.009, indicating the presence of long-term correlations in recurrence intervals. Furthermore, for the shuffled recurrence intervals, the long-term correlations are absent with $\alpha \simeq 0.5$ (the mean value is 0.497 with mean error 0.007).

In summary, we have investigated the scaling and memory properties in recurrence intervals of the Internet traffic. The empirical pdfs $P_q(\tau)$ for byte and packet flows, respectively, can fall into a single curve by rescaling with the mean recurrence intervals $\bar{\tau}$, as shown in eq. (1). The scaling function has a stretching exponential form, as $f(x) = A \exp(-Bx^\beta)$ with $\beta \approx 0.45$ for both byte and packet flows. This scaling property can be used to predict the occurrence probability of rare events that correspond to large q . We also detected the memory effect that a short (or long) recurrence interval tends to be followed by another short (or long) one, which is further demonstrated by the empirical results that the conditional pdf, $P_q(\tau|\tau_0)$, is strongly dependent on τ_0 . Further more, by using the DFA method, we found that the recurrence intervals are indeed long-term correlated. Some recently reported empirical studies show that the Internet-based human activities exhibit burstiness and memory in temporal statistics, such as the web accessing [32,33] and on-line

entertainment [34]. All those activities contribute some to the Internet traffic, and thus we think the analysis of the burstiness and memory of the Internet traffic itself can be considered as a valuable complementary work. More interestingly, the results suggest that the Internet shares some common properties with other complex systems like earthquake and financial market [9–12], which gives support to the possibly generic organizing principles governing the dynamics of apparently disparate complex systems, as dreamed by Goh and Barabási [28].

This work is supported by the National Science Foundation of China under Grant Nos. 70671097, 60874090 and 10635040. S-MC acknowledges the financial support of State 863 Projects (2008AA01A318).

REFERENCES

- [1] NEWMAN M. E. J., *Contemp. Phys.*, **46** (2005) 323.
- [2] CROW E. L. and SHIMIZU K., *Lognormal Distributions: Theory and Applications* (Marcel Dekker Inc., New York) 1988.
- [3] LAHERRERE J. and SORNETTE D., *Eur. Phys. J. B*, **2** (1998) 525.
- [4] BAK P., CHRISTENSEN K., DANON L. and SCANLON T., *Phys. Rev. Lett.*, **88** (2002) 178501.
- [5] CORRAL A., *Phys. Rev. E*, **68** (2003) 035102(R).
- [6] CORRAL A., *Phys. Rev. Lett.*, **92** (2004) 108501.
- [7] CORRAL A., *Phys. Rev. E*, **71** (2005) 017101.
- [8] LIVINA V., HAVLIN S. and BUNDE A., *Phys. Rev. Lett.*, **95** (2005) 208501.
- [9] WANG B.-H. and HUI P.-M., *Eur. Phys. J. B*, **20** (2001) 573.
- [10] LILLO F. and MANTEGNA R. N., *Phys. Rev. E*, **68** (2003) 016119.
- [11] YAMASAKI K., MUCHNIK L., HAVLIN S., BUNDE A. and STANLEY H. E., *Proc. Natl. Acad. Sci. U.S.A.*, **102** (2005) 9424.
- [12] WANG F. Z., YAMASAKI K., HAVLIN S. and STANLEY H. E., *Phys. Rev. E*, **73** (2006) 026117.
- [13] BUNDE A., EICHNER J. F., KANTELHARDT J. W. and HAVLIN S., *Phys. Rev. Lett.*, **94** (2005) 048701.
- [14] OMORI F., *J. Coll. Sci. Imp. Univ. Tokyo*, **7** (1894) 111.
- [15] GUTENBERG B. and RICHTER C. F., *Bull. Seismol. Soc. Am.*, **34** (1944) 185.
- [16] PASTO-SANTORRAS R. and VESPIGNANI A., *Evolution and Structure of the Internet: A Statistical Physics Approach* (Cambridge, Cambridge University Press) 2004.
- [17] ZHANG G.-Q., ZHANG G.-Q., YANG Q.-F., CHENG S.-Q. and ZHOU T., *New J. Phys.*, **10** (2008) 123027.
- [18] LELAND W. E., TAQQU M. S., WILLINGER W. and WILSON D. V., *IEEE/ACM Trans. Netw.*, **2** (1994) 1.
- [19] TADIĆ B., THURNER S. and RODGERS G. J., *Phys. Rev. E*, **69** (2004) 036102.
- [20] YAN G., ZHOU T., HU B., FU Z.-Q. and WANG B.-H., *Phys. Rev. E*, **73** (2006) 046108.
- [21] CAI S.-M., YAN G., ZHOU T., ZHOU P.-L., FU Z.-Q. and WANG B.-H., *Phys. Lett. A*, **366** (2007) 14.
- [22] ABE S. and SUZUKI N., *Europhys. Lett.*, **61** (2003) 852.
- [23] ABE S. and SUZUKI N., *Physica A*, **319** (2003) 552.
- [24] The specific description of the measurement of Ethernet traffic sets is given at <http://ita.ee.lbl.gov/html/contrib/BC.html>.
- [25] ZHOU P.-L., CAI S.-M., ZHOU T. and FU Z.-Q., *Proceedings of International Conference on Communication Circuit And Systems 2006* (IEEE Press) 2006, pp. 1744–1748.
- [26] IVANOV P. CH., ROSENBLUM M. G., PENG C.-K., MIETUS J., HAVLIN S., STANLEY H. E. and GOLDBERGER A. L., *Nature*, **383** (1996) 323.
- [27] IVANOV P. CH., YUEN A., PODOBNIK B. and LEE Y., *Phys. Rev. E*, **69** (2004) 056107.
- [28] GOH K.-I. and BARABÁSI A.-L., *EPL*, **81** (2008) 48002.
- [29] PENG C. K., BULDYREV S. V., HAVLIN S., SIMONS M., STANLEY H. E. and GOLDBERGER A. L., *Phys. Rev. E*, **49** (1994) 1685.
- [30] HU K., IVANOV P. CH., CHEN Z., CARPENA P. and STANLEY H. E., *Phys. Rev. E*, **64** (2001) 011114.
- [31] CHEN Z., IVANOV P. CH., HU K. and STANLEY H. E., *Phys. Rev. E*, **65** (2002) 041107.
- [32] DEZSÖ Z., ALMAAS E., LUKÁCS A., RÁCZ B., SZAKADÁT I. and BARABÁSI A.-L., *Phys. Rev. E*, **73** (2006) 066132.
- [33] GONCALVES B. and RAMASCO J. J., *Phys. Rev. E*, **78** (2008) 026123.
- [34] ZHOU T., KIET H. A.-T., KIM B.-J., WANG B.-H. and HOLME P., *EPL*, **82** (2008) 28002.