Investigating Long-Range Dependence in E-Commerce Web Traffic

Grażyna Suchacka[1](✉) and Adam Domański[2]

[1] Institute of Mathematics and Informatics,
Opole University, ul. Oleska 48, 45-052 Opole, Poland
gsuchacka@uni.opole.pl

[2] Institute of Informatics, Silesian Technical University,
Akademicka 16, 44-100 Gliwice, Poland
adamd@polsl.pl

Abstract. This paper addresses the problem of investigating long-range dependence (LRD) and self-similarity in Web traffic. Popular techniques for estimating the intensity of LRD via the Hurst parameter are presented. Using a set of traces of a popular e-commerce site, the presence and the nature of LRD in Web traffic is examined. Our results confirm the self-similar nature of traffic at a Web server input, however the resulting estimates of the Hurst parameter vary depending on the trace and the technique used.

Keywords: Long-range dependence · Self-similarity · Hurst parameter · Hurst index · H index · Web server · Web traffic · HTTP traffic

1 Introduction

Analysis and modelling of Web traffic has been a hot research issue in recent years. HTTP requests' arrival times at a Web server may be easily observed and analyzed. In reality, a request arrival process on a Web server has been proven to reveal significant variance (burstiness): peak request rates can exceed the average request rate even tenfold and surpass the server capacity, resulting in the poor quality of Web service [1,2]. When this process is bursty on a wide range of time scales, it may have a feature of *self-similarity*. As a consequence of burstiness on many time scales, the arrival process may show *long-range dependence* (LRD), which means that values at any instant are non-negligibly positively correlated with values at all future instants [3]. Although the concepts of self-similarity and long-range dependence are not equivalent, in the literature they are often used interchangeably which may be attributed to the fact that the presence of both self-similarity and LRD may be estimated with the Hurst parameter (Hurst index), denoted as H.

Self-similarity has been discovered not only in Web server workload [2–4] but also in computer network traffic [5–9] or Web query traffic [10]. The synthetic self-similar traffic can be constructed by multiplexing a large number of

P. Gaj et al. (Eds.): CN 2016, CCIS 608, pp. 42–51, 2016.
DOI: 10.1007/978-3-319-39207-3_4

on/off sources characterized by heavy-tailed on and off period lengths. Analysis of the Web traffic [3] showed that the self-similarity feature of such traffic can be attributed to several factors, including heavy-tailed distributions of Web document sizes and user "think times", the effect of caching, and the superimposition of many such transfers in the network.

Self-similarity may have a significant negative impact on system performance and scalability [11]. That is why taking into consideration this Web traffic feature is essential when developing a synthetic workload model used to test the server system capacity – otherwise system performance may be overestimated. A number of traffic models and synthetic traffic generators implementing self-similarity and burstiness have been proposed [12–15].

Very few studies have investigated self-similarity and LRD of the arrival process at e-commerce websites so far [2,4]. The main impediment for this fact is a difficulty in obtaining traffic traces from online retailers, mainly due to e-business profitability and e-customer privacy concerns. In this paper, we investigate LRD in traffic arriving on a popular e-commerce Web server. The additional motivation for our study was a huge increase in popularity of online marketing and Web analytics in recent years, which could induce changes in Web traffic patterns at e-commerce servers, mainly due to the increased share of bot-generated traffic.

The paper is organized as follows. Section 2 presents background information on self-similary, LRD, and some methods for investigating these phenomena in time series. Section 3 presents datasets analyzed in our study and discusses the results of LRD intensity estimation. Section 4 concludes the paper.

2 Background

In this section notions of self-similarity and long-range dependence are briefly presented and some methods for estimating these phenomena are introduced. For detailed discussion on these issues refer e.g. to [16].

2.1 Self-similarity and Long-Range Dependence

Self-similarity may be defined in terms of the process distribution as follows. A stochastic process $Y(t)$ is *self-similar* with a self-similarity parameter H if for any positive stretching factor c, the distribution of the rescaled process $c^{-H}Y(ct)$ is equivalent to that of the original process $Y(t)$ [17].

A self-similar process shows *long-range dependence* if its autocorrelation function follows a power law: $r(k) \sim k^{-\beta}$ as $k \to \infty$, where $\beta \in (0,1)$ [3] (it is worth noting that LRD can be also defined for non self-similar processes).

A presence and a degree of self-similarity and long-range dependence is expressed by the *Hurst parameter*, H. When H is in the range of 0.5 and 1, one can say that a process is self-similar [18] and the higher H is, the higher degree of self-similarity and LRD is revealed by the series [2] (although a process can be self-similar even if $H \leq 0.5$, e.g., for the special case of Fractional Brownian motions).

2.2 Selected Methods for Estimating the Hurst Parameter

We apply five popular methods for assessing self-similarity and LRD of the Web traffic [3,8,12,19]. Four of them are graphical methods: aggregate variance method, R/S plot, periodogram-based method, and wavelet-based method. The last method is Local Whittle estimator.

The aggregate variance method and the R/S plot method are in the time domain. Let us consider a time series $X = (X_t; t = 1, 2, \ldots, N)$. In the *aggregate variance method*, the m-aggregated series $X^{(m)} = (X_k^{(m)}; k = 1, 2, \ldots)$ is defined by summing the time series X over nonoverlapping blocks of length m. The variance of series $X^{(m)}$ is plotted against m on a log-log plot and the points are approximated by a straight line, e.g., by using the least squares method. Then, the slope of the line, $-\beta$, is established and the Hurst parameter is computed as $H = 1 - \beta/2$. For a self-similar series variance decays slowly so $-\beta$ is greater than -1, which gives H higher than 0.5.

In the *R/S plot method*, the rescaled range, i.e., the R/S statistic, is plotted against m (which has been traditionally denoted by d in this method) on a log-log plot. For a self-similar series, R/S grows according to a power law with exponent H as a function of d and the plot has slope which is an estimate of H.

Other three methods are in the frequency domain. In the *periodogram-based method*, a periodogram of a time series X is defined by:

$$I_N(\lambda) = \frac{1}{2\pi N} \left| \sum_{t=1}^{N} X_t e^{i\lambda t} \right|^2 , \tag{1}$$

where $i = \sqrt{-1}$. Usually it is evaluated at the Fourier Frequencies $\lambda_{j,N} = \frac{2\pi j}{N}$, where $j \in [0, n/2]$. The estimation of H is based on the slope γ of a log-log plot $I_N(\lambda_{j,N})$ versus $\lambda_{j,N}$ as frequency approaches zero. The relationship between the periodogram slope and the Hurst parameter is given by the formula $\gamma = 1 - 2H$.

Local Whittle estimator is a non-graphical method based on periodograms. This method assumes that the spectral density $f(\lambda)$ of the series can be approximated by the function:

$$f_{c,H}(\lambda) = c\lambda^{1-2H} \tag{2}$$

for frequencies λ as frequency approaches zero. The Local Whittle estimator of H is defined by minimizing:

$$\sum_{j=1}^{m} \log f_{c,H}(\lambda_{j,N}) + \frac{I_N(\lambda_{j,N})}{f_{c,H}(\lambda_{j,N})} \tag{3}$$

with respect to c and H; I_N is defined in (1) and $f_{c,H}$ is defined in (2).

In the *wavelet-based estimator* of the Hurst parameter, wavelets are considered as a generalisation of Fourier transform. For the series X the wavelet coefficients are determined; based on their values a time average μ_j is performed at a given scale (for the j-th octave). The relationship between μ_j and H is given by the formula:

$$E \log_2(\mu_j) \sim (2H - 1)j + C, \tag{4}$$

where E means the average, C depends only on H. Using this relationship, H may be determined based on the slope of an appropriate weighted linear regression.

Some other methods for determining the Hurst parameter in time series have been also proposed, e.g. detrended fluctuation analysis (DFA) [20] or multifractal analysis [21]. We do not discuss them in the paper due to space limitations.

3 Estimation of the Hurst Parameter for E-Commerce Traffic

3.1 Data Collection

The main goal of our analysis was to investigate LRD in e-commerce Web traffic. The analysis was done for data recorded in Web server log files obtained from an online retailer trading car parts and accessories. HTTP description lines were converted into time series reflecting the request arrival process at the Web server during the successive 14 days. 14 one-day traces were separately analyzed (traces are named with dates of traffic collection).

To verify the results obtained for the e-commerce traces, we decided to perform an additional LRD analysis of traffic at an actual non e-commerce server. To this end, we used seven traces from a server hosting a specialized mailing list.

The number of samples (i.e., the number of HTTP requests) in each trace is presented in Table 1.

Table 1. Cardinality of the analyzed data sets

E-commerce trace		Non e-commerce trace	
Trace (date)	Number of samples	Trace (date)	Number of samples
01.12.2015	13 643	10.01.2016	12 151
02.12.2015	56 284	11.01.2016	13 832
03.12.2015	9 642	12.01.2016	13 640
04.12.2015	17 842	13.01.2016	13 552
05.12.2015	25 082	14.01.2016	14 010
06.12.2015	16 092	15.01.2016	15 438
07.12.2015	15 860	16.01.2016	1 765
08.12.2015	16 138		
09.12.2015	190 934		
10.12.2015	170 529		
11.12.2015	41 249		
12.12.2015	9 758		
13.12.2015	14 594		
14.12.2015	17 453		

Package R [22] was used to estimate the Hurst parameter for both sets of traces with the application of the five methods described in Subsect. 2.2.

3.2 Results and Discussion

Figures 1, 2, 3, 4 provide examples of the application of four graphical methods to analyze two e-commerce traces, collected on 1 and 5 December 2015. Traffic in the 01.12.2015 trace is characterized by rather low LRD intensity compared to other e-commerce traces. On the other hand, for traffic registered in the 05.12.2015 trace, the highest mean H estimate was achieved in our analysis. Thus, in Figs. 1, 2, 3, 4 one can compare plots for Web traffic characterized with a moderate and a high level of long-range dependence.

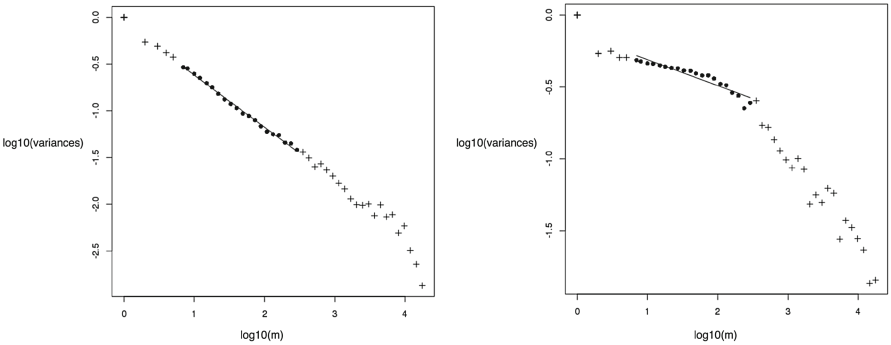

Fig. 1. Aggregate variance plot for the 01.12.2015 trace (left) and the 05.12.2015 trace (right)

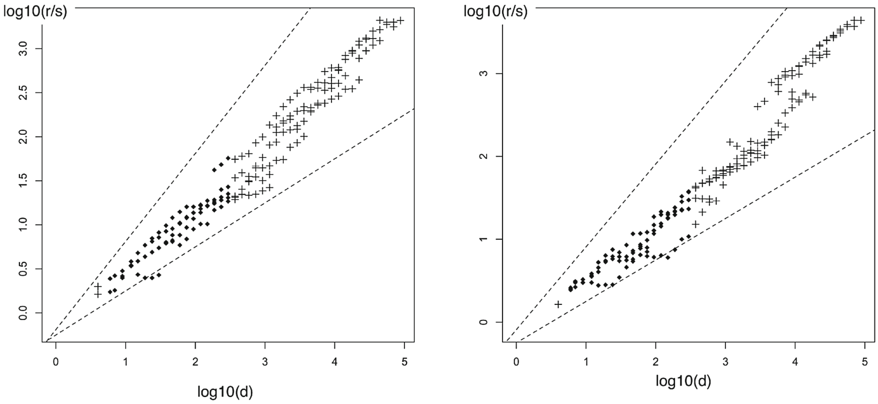

Fig. 2. R/S plot for the 01.12.2015 trace (left) and the 05.12.2015 trace (right)

Figure 1 shows the aggregate variance plots. One can observe that the linear plots are characterized by a slope clearly different from -1 which confirms the self-similarity of the analyzed time series. The slope of the plot for 01.12.2015 data (left) was estimated as -0.56, giving an estimate for the Hurst parameter of 0.72. The slope estimated for a 05.12.2015 data plot (right) is -0.18 which results in H of 0.91.

The R/S plots in Fig. 2 have an asymptotic slope between 0.5 and 1 (the corresponding lines are shown for comparison). The slope, being an estimate of H, was determined using regression as 0.65 for the 01.12.2015 trace and 0.54 for the 05.12.2015 trace.

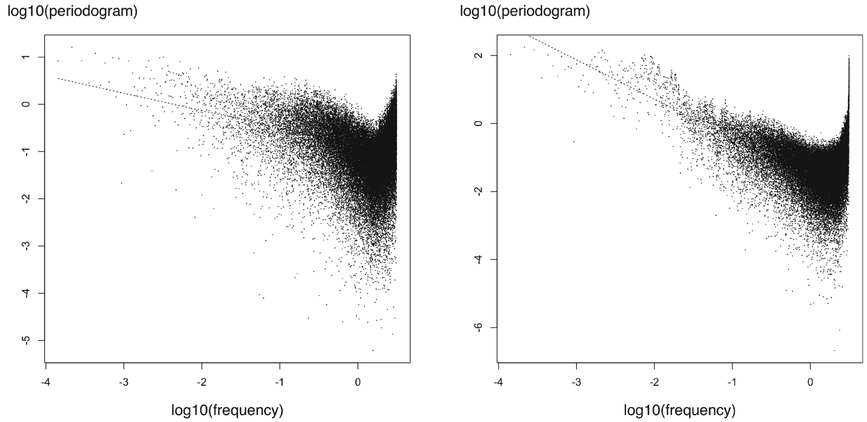

Fig. 3. Periodogram for the 01.12.2015 trace (left) and the 05.12.2015 trace (right)

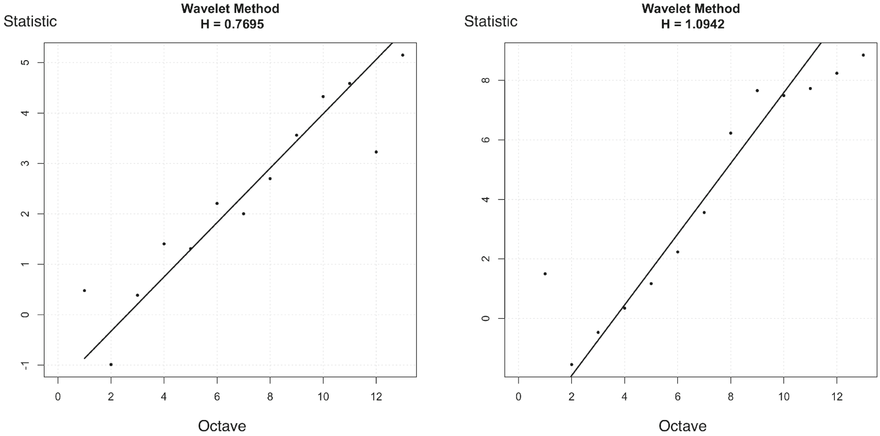

Fig. 4. R/S plot for the 01.12.2015 trace (left) and the 05.12.2015 trace (right)

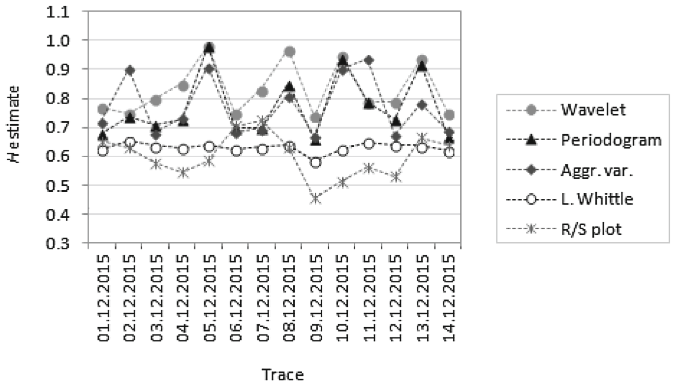

Fig. 5. Comparison of H estimates for the e-commerce traces

Figure 3 presents example results achieved using the periodogram-based method. Regression lines for periodogram plots have a slope of −0.36 and −0.5, giving the estimates of H as 0.68 and 0.75 for the 01.12.2015 and 05.12.2015 traces, correspondingly.

Figure 4 shows results of application of the wavelet-based estimator of the Hurst parameter to the two example e-commerce traces. The corresponding H of 0.77 and 1.09 were estimated. For H determined with this method confidence intervals are provided (Table 2).

Table 2 summarizes the results of our study across the different methods for all 14 e-commerce traces. In general, the H estimate exceeds 0.5 which indicates the self-similar character of the traffic. Only H estimated for the 09.12.2015 trace using the R/S plot method was 0.46. Other values of the Hurst parameter exceed 0.5 and they vary significantly, ranging from 0.51 to even 0.98.

Mean H values estimated for each e-commerce trace (the last column) show significant fluctuations in LRD intensity depending on a day, with H ranging from 0.6 for the 09.12.2015 trace to 0.8 for the 05.12.2015 trace. The last row of Table 2 shows even bigger differences in H estimates depending on the method applied.

Fluctuations in H estimates depending on the trace and the method applied are graphically presented in Fig. 5. One can observe that for the Local Whittle method, the estimate of H stays relatively consistent across all 14 analyzed datasets (with the mean value of 0.63). On the other hand, for the graphical methods it varies greatly. The wavelet-based method tends to give the highest H estimates (with the mean of 0.85) whereas H estimates for the R/S plot method are the lowest (with the mean of 0.6). We cannot give reasons for such a big variance of H estimates across various methods. However, such variance is not uncommon - it has been also obtained in some previous studies, e.g., for network traffic [8,12] and MPEG-1 encoded video sequences2 [23].

Table 3 presents estimates of the Hurst parameter for the non e-commerce traces and Fig. 6 illustrates fluctuations in these estimates depending on the

Table 2. H estimates for the e-commerce traces

Trace	Aggregate variance method	R/S plot	Periodogram-based method	Local Whittle estimator	Wavelet-based method	MEAN
01.12.2015	0.72	0.65	0.68	0.62	0.77 ± 0.05	0.67
02.12.2015	0.90	0.63	0.74	0.65	0.75 ± 0.04	0.73
03.12.2015	0.68	0.58	0.71	0.64	0.80 ± 0.06	0.65
04.12.2015	0.73	0.54	0.73	0.63	0.85 ± 0.07	0.66
05.12.2015	0.91	0.59	0.98	0.64	0.98 ± 0.03	0.80
06.12.2015	0.68	0.70	0.70	0.62	0.75 ± 0.05	0.68
07.12.2015	0.70	0.72	0.70	0.63	0.83 ± 0.06	0.69
08.12.2015	0.81	0.63	0.85	0.64	0.97 ± 0.03	0.74
09.12.2015	0.67	0.46	0.66	0.58	0.74 ± 0.03	0.60
10.12.2015	0.90	0.51	0.94	0.63	0.95 ± 0.07	0.74
11.12.2015	0.94	0.56	0.79	0.65	0.79 ± 0.05	0.73
12.12.2015	0.67	0.53	0.73	0.64	0.79 ± 0.05	0.64
13.12.2015	0.78	0.67	0.92	0.64	0.94 ± 0.03	0.75
14.12.2015	0.69	0.64	0.67	0.62	0.75 ± 0.05	0.65
Mean	0.77	0.60	0.79	0.63	0.85	

Table 3. H estimates for the non e-commerce traces

Trace	Aggregate variance method	R/S plot	Periodogram-based method	Local Whittle estimator	Wavelet-based method	MEAN
10.01.2016	0.60	0.65	0.62	0.62	0.66 ± 0.02	0.63
11.01.2016	0.61	0.62	0.64	0.62	0.66 ± 0.01	0.63
12.01.2016	0.59	0.64	0.60	0.62	0.65 ± 0.02	0.62
13.01.2016	0.68	0.60	0.74	0.63	0.75 ± 0.04	0.68
14.01.2016	0.73	0.62	0.73	0.65	0.73 ± 0.03	0.69
15.01.2016	0.61	0.61	0.61	0.62	0.65 ± 0.02	0.62
16.01.2016	0.79	0.74	0.73	0.68	0.71 ± 0.02	0.73
Mean	0.66	0.64	0.66	0.63	0.69	

trace and the method used. For this traffic the Hurst parameter (with the mean of 0.66) seems to be a little lower than the one for the e-commerce traffic (with the mean of 0.7). At the same time, H estimates for non e-commerce traffic are much more consistent across days and methods applied (c.f. Fig. 5).

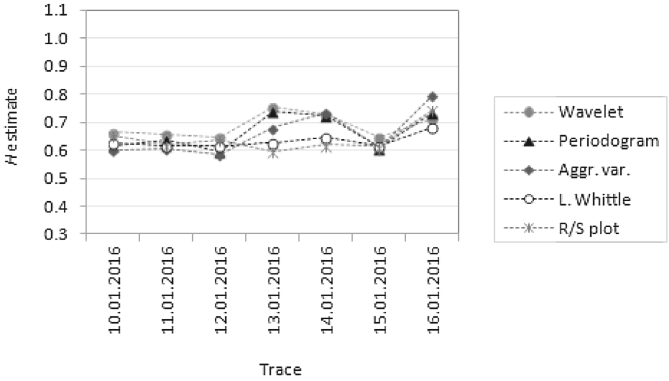

Fig. 6. Comparison of H estimates for the non e-commerce traces

4 Conclusions

Application of five popular Hurst parameter estimators to e-commerce traffic shows that this traffic reveals a significant level of long-range dependence. The mean H estimate ranges from 0.6 to 0.8 depending on a day. This result is consistent with results for request arrival process at other e-commerce sites: 0.66 in [2] and 0.73–0.8 in [4]. Furthermore, our study confirms previous findings that one cannot rely on a single method to estimate the Hurst parameter since different methods usually give different results. In our case, the mean H estimate for the e-commerce traffic ranges from 0.6 to 0.85 depending on the method. For the non e-commerce traffic, analyzed in the paper for comparative purposes, these fluctuations are much smaller and range from 0.63 to 0.69.

A coarse analysis of our results shows that the Hurst parameter determined for 24-hour intervals does not depend on the number of HTTP requests arrived on the server within these intervals. It also does not depend on the share of Web bot requests in the intervals. A deeper LRD analysis, performed for intervals shorter than 24 hours, is being planned to investigate these possible relationships. Furthermore, we plan to use traces from multiple Web servers to inspect if it is possible to use the Hurst parameter to distinguish between e-commerce and non e-commerce source.

References

1. Kant, K., Venkatachalam, M.: Transactional characterization of front-end e-commerce traffic. In: IEEE Global Telecommunications Conference (GLOBECOM 2002), vol. 3, pp. 2523–2527. IEEE Press, New York (2002)
2. Vallamsetty, U., Kant, K., Mohapatra, P.: Characterization of e-commerce traffic. Electron. Commer. Res. **3**(1), 167–192 (2003)
3. Crovella, M., Bestavros, A.: Self-similarity in world wide web traffic: evidence and possible causes. IEEE/ACM Trans. Networking **5**(6), 835–846 (1997)
4. Xia, C.H., Liu, Z., Squillante, M.S., Zhang, L., Malouch, N.: Web traffic modeling at finer time scales and performance implications. Perform. Eval. **61**(2–3), 181–201 (2005)

5. Domańska, J., Domański, A., Czachórski, T.: A few investigations of long-range dependence in network traffic. In: 29th International Symposium on Computer and Information Sciences, Kraków, Poland. Information Sciences and Systems, part III, pp. 137–144. Springer, Heidelberg (2014)
6. Dymora, P., Mazurek, M., Strzałka, D.: Computer network traffic analysis with the use of statistical self-similarity factor. Ann. UMCS Informatica AI **13**(2), 69–81 (2013)
7. Leland, W.E., Taqqu, M.S., Willinger, W., Wilson, D.V.: On the self-similar nature of ethernet traffic. IEEE/ACM Trans. Networking **2**(1), 1–15 (1994)
8. Park, C., Hernández-Campos, F., Le, L., Marron, J.S., Park, J., Pipiras, V., Smith, F.D., Smith, R.L., Trovero, M., Zhu, Z.: Long-range dependence analysis of internet traffic. J. Appl. Stat. **38**(7), 1407–1433 (2011)
9. Olejnik, R.: Charakter Ruchu HTTP w Lokalnych Sieciach Bezprzewodowych (Title in English: The nature of HTTP traffic in wireless local area networks). Metody Informatyki Stosowanej **4**, 175–180 (2011)
10. Balakrishnan, R., Kambhampati, S.: On the self-similarity of web query traffic: evidence, cause and performance implications. Technical report, Arizona State University (2009)
11. Hernandez-Orallo, E., Vila-Carbo, J.: Analysis of self-similar workload on real-time systems. In: 16th IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS 2010), pp. 343–352. IEEE Press, New York (2010)
12. Domańska, J., Domański, A., Czachórski, T.: Estimating the intensity of long-range dependence in real and synthetic traffic traces. In: Gaj, P., Kwiecień, A., Stera, P. (eds.) CN 2015. CCIS, vol. 522, pp. 11–22. Springer, Heidelberg (2015)
13. Kaur, G., Saxena, V., Gupta, J.P.: Characteristics analysis of web traffic with Hurst index. Lect. Notes Eng. Comput. Sci. **2186**(1), 234–238 (2010)
14. Lu, X., Yin, J., Chen, H., Zhao, X.: An approach for bursty and self-similar workload generation. In: Lin, X., Manolopoulos, Y., Srivastava, D., Huang, G. (eds.) WISE 2013, Part II. LNCS, vol. 8181, pp. 347–360. Springer, Heidelberg (2013)
15. Suchacka, G.: Generating bursty web traffic for a B2C web server. In: Kwiecień, A., Gaj, P., Stera, P. (eds.) CN 2011. CCIS, vol. 160, pp. 183–190. Springer, Heidelberg (2011)
16. Beran, J.: Statistics for Long-Memory Processes. Monographs on Statistics and Applied Probability. Chapman and Hall, New York, NY (1994)
17. Cavanaugh, J.E., Wang, Y., Davis, J.W.: Locally self-similar processes and their wavelet analysis. Handbook Stat. **21**, 93–135 (2003)
18. Stolojescu, C., Isar, A.: A comparison of some Hurst parameter estimators. In: 13th International Conference on Optimization of Electrical and Electronic Equipment (OPTIM 2012), pp. 1152–1157. IEEE Press, New York (2012)
19. Clegg, R.G.: A practical guide to measuring the Hurst parameter. Int. J. Simul. Syst. Sci. Technol. **7**(2), 3–14 (2006)
20. Krištoufek, L.: Rescaled range analysis and detrended fluctuation analysis: finite sample properties and confidence intervals. AUCO Czech Econ. Rev. **4**, 315–329 (2010)
21. Sanchez-Ortiz, W., Andrade-Gómez, C., Hernandez-Martinez, E., Puebla, H.: Multifractal Hurst analysis for identification of corrosion type in AISI 304 stainless steel. Int. J. Electrochem. Sci. **10**, 1054–1064 (2015)
22. The R Project for Statistical Computing. https://www.r-project.org
23. Cano, J.C., Manzoni, P.: On the use and calculation of the Hurst parameter with MPEG videos data traffic. In: 26th Euromicro Conference, vol. 1, pp. 448–455. IEEE Press, New York (2000)