Verification of Web Traffic Burstiness and Self-similarity for Multiple Online Stores

Grażyna Suchacka(✉) and Alicja Dembczak

Institute of Mathematics and Computer Science,
University of Opole, Oleska 48, 45-052 Opole, Poland
gsuchacka@uni.opole.pl, alicja.dembczak@gmail.com

Abstract. Developing realistic Web traffic models is essential for a reliable Web server performance evaluation. Very significant Web traffic properties that have been identified so far include burstiness and self-similarity. Very few relevant studies have been devoted to e-commerce traffic, however. In this paper, we investigate burstiness and self-similarity factors for seven different online stores using their access log data. Our findings show that both features are present in all the analyzed e-commerce datasets. Furthermore, a strong correlation of the Hurst parameter with the average request arrival rate was discovered (0.94). Estimates of the Hurst parameter for the Web traffic in the online stores range from 0.6 for low traffic to 0.85 for heavy traffic.

Keywords: Web traffic · HTTP traffic · Web server · E-Commerce · Web store · Log analysis · Burstiness · Self-Similarity · Hurst parameter · Hurst index

1 Introduction

Numerous studies on the analysis and characterization of Internet traffic have confirmed its specific properties. Two very significant traffic features are its burstiness and self-similarity. *Burstiness* means that the traffic is highly variable, with traffic "bursts" observable on multiple time scales. The resulting time series, which is bursty on a wide range of time scales, may be statistically described as a *self-similar* process [1]. Burstiness and self-similarity have been identified both in the network traffic [2–5] and in Web server workloads [1, 6–9].

In reality, bursts in request arrival rates on the server may lead to transient server overloads and consequently, to a degraded server performance. Even a small amount of the traffic burstiness may degrade the server throughput [10, 11]. That is why this phenomenon has to be taken into account in Web server performance evaluation using the synthetic workload: to achieve reliable results of experiments testing the system performance, it is essential to model and generate bursty Web traffic [12–17].

In this paper we consider an arrival process of HTTP requests on Web servers which host B2C e-commerce websites, i.e., online stores. The motivation for our study was the fact that very few previous Web traffic analyses have been dedicated to e-business sites and the relevant literature lacks the comparative analysis of burstiness and self-similarity factors for multiple e-commerce environments. We obtained 24-hour

L. Borzemski et al. (eds.), *Information Systems Architecture and Technology: Proceedings of 38th International Conference on Information Systems Architecture and Technology – ISAT 2017*, Advances in Intelligent Systems and Computing 655, DOI 10.1007/978-3-319-67220-5_28

access log data for seven various Web stores, differing in the type and size of the store offer, the website structure, and site popularity, and we used them to estimate burstiness and self-similarity factors for the e-commerce sites. To the best of our knowledge, there has not been such a wide study of the e-commerce traffic so far.

The rest of this paper is organized as follows. Section 2 explains a concept of self-similarity and discusses methods used to evaluate the traffic burstiness and self-similarity. Section 3 presents achieved results and Sect. 4 concludes the paper.

2 Approach for Evaluating Self-similarity and Burstiness of the Web Traffic

2.1 Definition of Self-similarity

Self-similarity may be defined in the context of the time series distribution [1]. Let $X = (X_t;\ t = 1, 2, \ldots)$ be a zero-mean, stationary time series. 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. Series X is *H-self-similar* if for all positive m, series $X^{(m)}$ has the same distribution as X rescaled by m^H:

$$X_t = m^{-H} \sum\nolimits_{i=(t-1)m+1}^{tm} X_i \tag{1}$$

for all $m \in N$. H-self-similar series X has the same autocorrelation function: $r(k) = E[(X_t - \mu)(X_{t+k} - \mu)]/\sigma^2$ as the series $X^{(m)}$ for all m.

The degree of self-similarity may be estimated by determining the *Hurst parameter* (*Hurst index*), denoted by H. For a self-similar series this parameter is higher than 0.5. The higher H is, the higher degree of self-similarity is revealed by the series.

Various statistical tests may be applied to assess the Hurst parameter. Popular tests operating in the time domain are the aggregate variance method and the R/S plot method. Other common tests, operating in the frequency domain, include the periodogram-based method, the wavelet-based estimator, and the Local Whittle estimator. Less common methods are multifractal analysis, detrended fluctuation analysis, and the Arby-Veitch estimator.

2.2 Estimation of the Hurst Parameter Using the Aggregate Variance Method

To verify the self-similarity of the traffic, we apply the aggregate variance method, which has been widely applied in previous Internet traffic analyses [1, 2, 5, 8, 9, 11, 13, 17, 18]. This test uses the fact that for a self-similar process variances of the sample mean are decaying more slowly than the reciprocal of the sample size [20].

Based on request timestamps read from log data a time series $X = (X_t;\ t = 1, 2, \ldots, N)$ is created, covering the time interval T. In our case each original series X has a duration of $T = 24$ h $= 86400$ s.

Value of m, i.e., duration of a subinterval, is given in seconds, $m \in [2, N/2]$. Consecutive values of m are generated as a geometric sequence 2^k, $k = 1, 2, \ldots,$

so that $m <= N/2$. The m-aggregated series $X^{(m)}$ are created for consecutive values of m and the variance of series $X^{(m)}$ is determined.

The variance of $X^{(m)}$ is then plotted against m on a log-log plot and approximated by a straight line by using the least squares method. The slope of the line $-\beta$ is computed and used to determine the Hurst parameter, given by:

$$H = 1 - \beta/2. \tag{2}$$

2.3 Estimation of the Burstiness Factor

For each analyzed e-commerce dataset the request arrival data is first plotted on many time scales to visually inspect the traffic burstiness. Then, a more rigorous analysis of a *burstiness factor* is performed in the following way [19].

Let L be the total number of requests that arrived on the Web server in the time interval T. Let λ be the average request arrival rate, given by:

$$\lambda = \frac{L}{T}. \tag{3}$$

Let the time interval T be divided into n equal subintervals of duration m. Let l_k be the number of requests that arrive in subinterval k and λ_k be the arrival rate of requests during subinterval k, given by:

$$\lambda_k = \frac{n}{T} \times l_k. \tag{4}$$

where $k = 1, 2, \ldots, n$. Let l^+ be the total number of requests that arrive in subintervals in which the subinterval arrival rate λ_k exceeds the average arrival rate λ. The *burstiness parameter* b_m is defined as the fraction of time during which the subinterval arrival rate exceeds the average arrival rate:

$$b_m = \frac{Number\ of\ subintervals\ for\ which\ \lambda_k > \lambda}{n}. \tag{5}$$

If the traffic is not bursty, it means that it is uniformly distributed over all subintervals and consequently, $b = 0$. On the other hand, for the bursty traffic $b > 0$.

For each analyzed e-commerce dataset we compute the burstiness parameter, b_m, for $m = 2$, 4, 8, 16, 32, 64, 128, and 256 s. We also determine the mean value of the burstiness parameter, b_{mean}, to compare the burstiness across the multiple datasets.

3 Results

3.1 Empirical Data Description

We analyzed access log data of e-commerce websites obtained from seven online retailers (the identities of the websites are not revealed for confidentiality restrictions).

The analyzed websites vary in the type and size of the store offer, the website structure and the traffic level in terms of the number of HTTP requests received in a 24-hour period. The highest traffic was registered for the online bookstore site (*SiteB*), offering books, films, and multimedia (89,486 requests in total) and for the website offering products and services for elderly people (*SiteE*, 70,352 requests). Two datasets were for the automotive branch e-stores: *SiteA1* (10,472 requests) and *SiteA2* (20,378 requests). Two other datasets, differing significantly in the numbers of samples, were for websites offering tourist equipment and clothes: *SiteT1* (2,616 requests) and *SiteT2* (37,819). The last dataset, *SiteH*, was for the site offering devices and systems for house equipment and contained only 7,666 samples (the corresponding website was not well positioned at that time).

General information on the analyzed Web stores' data is summarized in Table 1. Note that although dates of data collection differ between the individual websites, time samples in each dataset cover the total time interval of 24 h.

Table 1. Basic information on the analyzed e-commerce datasets.

	SiteB	*SiteE*	*SiteT2*	*SiteA2*	*SiteA1*	*SiteH*	*SiteT1*
Branch	Books	For elderly	Tourist	Auto-motive	Auto-motive	For house	Tourist
Date of data collection	Apr 1, 2014	Jan 25, 2016	Mar 29, 2015	Apr 3, 2015	Nov 12, 2016	Apr 10, 2015	Feb 24, 2015
Number of requests, L	89,486	70,352	37,819	20,378	10,472	7,666	2,616
Average request arrival rate, λ	1.04	0.81	0.44	0.24	0.12	0.09	0.03

3.2 Burstiness

Figures 1, 2, 3, 4 and 5 illustrate request arrival rates at different time scales (per subintervals of m = 4, 8, 16, 32, and 64 s) for the most numerous dataset, *SiteB*. Depending on the subinterval duration, data shown in the figures covers various observation windows. For example, Fig. 1 illustrates request arrival rates for 1200 4-second subintervals so the plotted data corresponds to an 80-minute time span. The higher the value of m is, the longer observation window is reflected in a figure.

Data in Fig. 5 corresponds to the whole one day (24 h). In this case a clear diurnal pattern of request arrivals is visible, with the least intensive traffic at night time, the gradually increasing traffic since 5 am till the peak traffic period starting at about 2 pm and lasting till about 10 pm.

A visual inspection of the plots confirm that the Web traffic arriving at *SiteB* is evidently bursty across several different time scales. Plots for other datasets are not presented in the paper due to space limits but they lead to similar conclusions on the traffic burstiness.

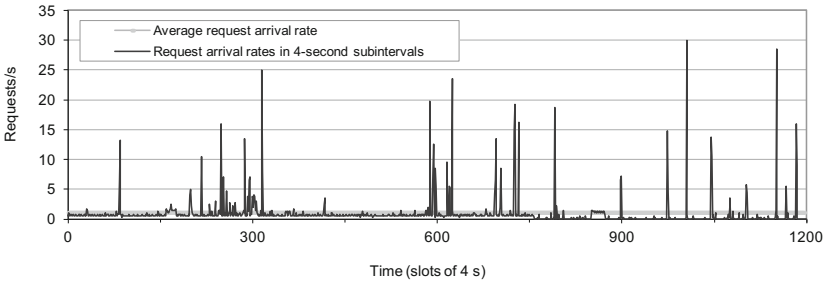

Fig. 1. Burstiness of the Web traffic on *SiteB* in slots of 4 s.

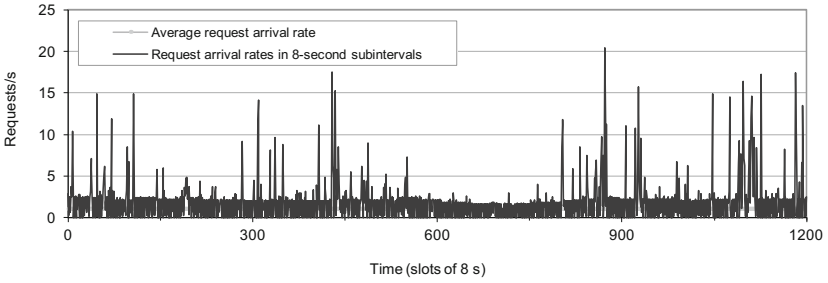

Fig. 2. Burstiness of the Web traffic on *SiteB* in slots of 8 s.

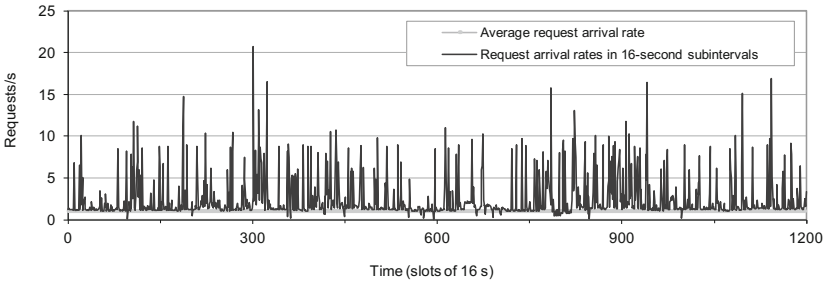

Fig. 3. Burstiness of the Web traffic on *SiteB* in slots of 16 s.

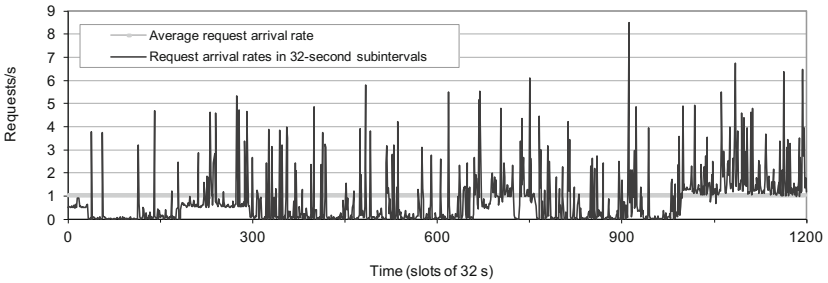

Fig. 4. Burstiness of the Web traffic on *SiteB* in slots of 32 s.

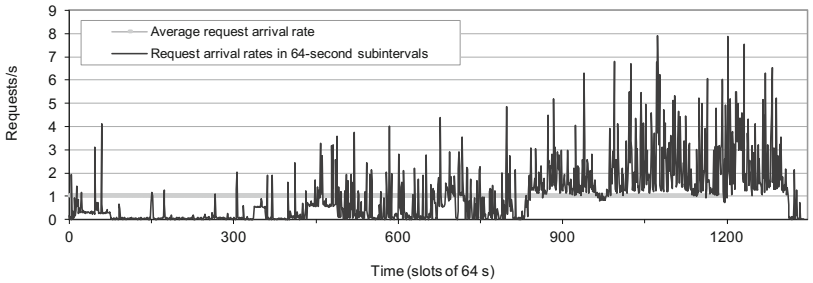

Fig. 5. Burstiness of the Web traffic on *SiteB* in slots of 64 s.

Burstiness visible on the plots of request arrival rates at different time scales is confirmed by estimates of the burstiness parameters, determined according to (5). Table 2 presents burstiness parameter values for different subinterval durations (2, 4, 8, 16, 32, 64, 128, and 256 s) and mean values of the burstiness parameter, denoted by b_{mean}. Figure 6 plots the burstiness parameter vs. subinterval duration for all the datasets. It can be seen that in general the burstiness factor tends to increase with the increase in the time scale. An exception from this tendency is the Web traffic registered for *SiteT1* and *SiteA2*.

Table 2. Burstiness factors for the analyzed datasets.

	SiteB	*SiteE*	*SiteT2*	*SiteA2*	*SiteA1*	*SiteH*	*SiteT1*
b_2	0.12	0.05	0.19	0.32	0.07	0.05	0.02
b_4	0.16	0.08	0.13	0.52	0.11	0.08	0.04
b_8	0.25	0.12	0.12	0.43	0.18	0.12	0.07
b_{16}	0.38	0.10	0.15	0.36	0.16	0.10	0.13
b_{32}	0.39	0.13	0.20	0.31	0.19	0.13	0.23
b_{64}	0.42	0.13	0.25	0.26	0.21	0.13	0.14
b_{128}	0.42	0.17	0.28	0.25	0.24	0.17	0.12
b_{256}	0.41	0.24	0.31	0.26	0.30	0.24	0.15
Mean burstiness (b_{mean})	0.32	0.13	0.20	0.34	0.18	0.13	0.11

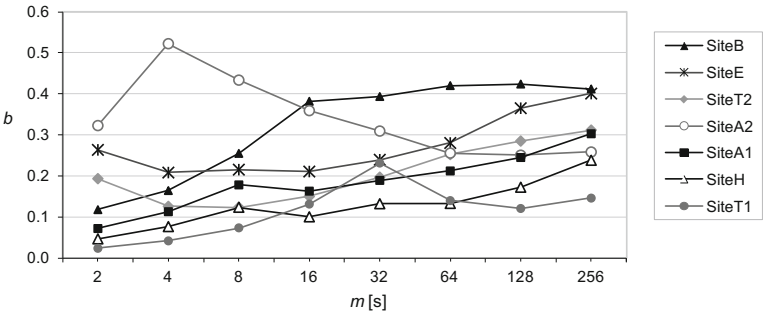

Fig. 6. Burstiness factor vs. subinterval duration.

3.3 Self-similarity

A visual inspection of the traffic burstiness and the determined burstiness factors suggest the presence of self-similarity in the Web traffic on all the analyzed e-commerce sites – even though the traffic stationarity at higher time scales is questionable. This is confirmed by estimates of H, all of which exceed 0.5 (Table 3).

Table 3. Hurst parameter determined for the analyzed datasets.

	SiteB	*SiteE*	*SiteT2*	*SiteA2*	*SiteA1*	*SiteH*	*SiteT1*
H	0.85	0.77	0.69	0.65	0.69	0.66	0.60

We examined the correlation between the H estimates and the burstiness parameters for different durations of a data aggregation subinterval, m. In the case of low values of m (2, 4, 8, 16, and 32) and b_{mean} there was no linear relationship or it was very weak (below 0.4). However, the relationship between H and b_{64} was moderate (0.64) and the relationships between H and b_{128} and b_{256} were quite strong (0.73 and 0.80, respectively).

Figures 7, 8, 9 and 10 show variance-time log-log plots for the analyzed time series. In all cases the shape of the line approximating the data significantly differs from –1, resulting in values of the Hurst parameter ranging from 0.6 to 0.85, depending on a dataset. These estimates confirm previous findings on H for e-commerce traffic, which was estimated as 0.66 for the traffic with the average arrival rate of 0.65 requests/s in the study [7] and ranged from 0.73 to 0.8, depending on the H estimating test, for the peak e-commerce traffic in the study [8].

We observed that higher traffic intensity levels on e-commerce sites correspond to higher values of the Hurst index. It is confirmed by a very high correlation between H and λ, equal to 0.94. This conclusion is also consistent with some previous findings for the non-e-commerce Web traffic [1, 6], stating that although self-similarity is not necessarily an invariant in all Web server workloads, it is evident in heavy workloads. However, in contrast, we identified the self-similarity in all the analyzed e-commerce server workloads, even for the websites subject to very low traffic levels.

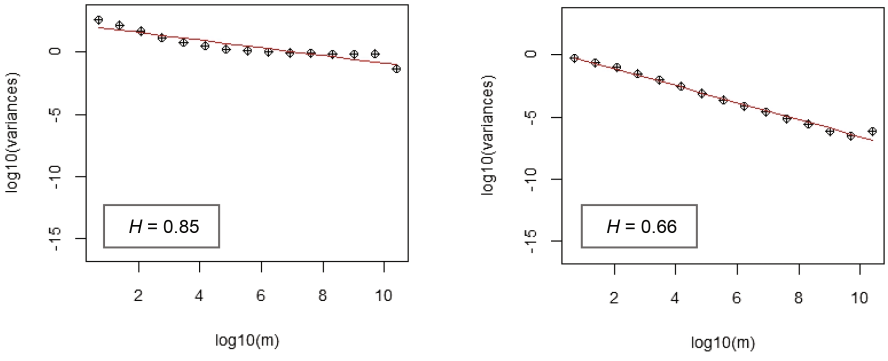

Fig. 7. Aggregate variance plot for *SiteB* (left) and *SiteH* (right).

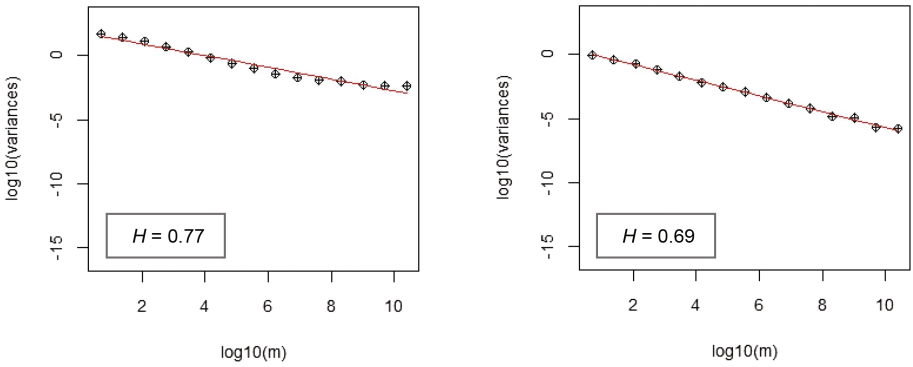

Fig. 8. Aggregate variance plot for *SiteE* (left) and *SiteA1* (right).

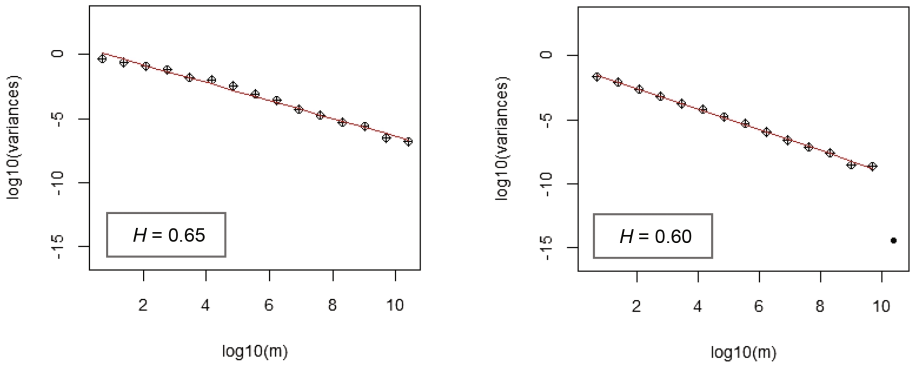

Fig. 9. Aggregate variance plot for *SiteA2* (left) and *SiteT1* (right).

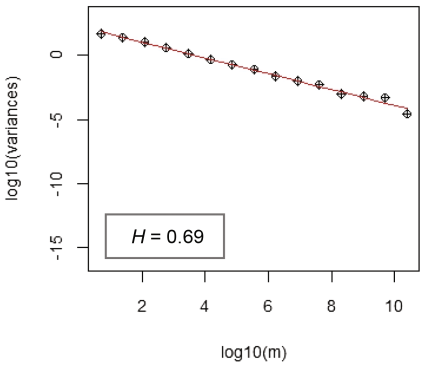

Fig. 10. Aggregate variance plot for *SiteT2*.

4 Concluding Remarks

In our study we verified burstiness and self-similarity of Web traffic on seven different e-commerce sites. The resulting estimates of the burstiness parameters (including mean burstiness factors ranging from 0.11 to 0.32, depending on a site) confirm the very variable character of the workloads on all the analyzed e-commerce servers.

Moreover, in all cases the Hurst parameter exceeds 0.5 which proves the presence of self-similarity in the e-commerce traffic (H estimates range from 0.6 for low traffic level to 0.85 for heavy traffic). Our results are consistent with older reports on H index for e-commerce severs with moderate [7] and heavy [8] traffic levels, estimated as 0.66 and 0.73-0.8, respectively. Our study also confirms some previous conclusions that a degree of self-similarity of Web traffic is a bit higher on e-commerce sites than on other sites [9].

In contrast to some previous related work for non-e-commerce Web traffic, we identified the self-similarity property in all seven analyzed e-commerce datasets, even for the websites subject to low traffic levels. Furthermore, we discovered that the more requests arrive at an e-commerce server, the higher degree of self-similarity is revealed by the traffic – there is a very strong correlation of the Hurst parameter with the average request arrival rate, equal to 0.94.

Our study advances the state-of-the-art on properties of the Web traffic on e-commerce servers. The use of several datasets for online stores differing in the offered products, the website structure, and the site popularity allows us to generalize the results to multiple e-commerce scenarios. Our findings may be useful in developing representative models of e-commerce workloads for Web server performance evaluation.

Acknowledgment. This paper is based upon work from COST Action IC1304 Autonomous Control for a Reliable Internet of Services (ACROSS), supported by COST (European Cooperation in Science and Technology).

References

1. Crovella, M., Bestavros, A.: Self-similarity in World Wide Web traffic: evidence and possible causes. ACM SIGMETRICS Perform. Eval. Rev. **24**(1), 160–169 (1996)
2. 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)
3. Dymora, P., Mazurek, M., Strzałka, D.: Computer network traffic analysis with the use of statistical self-similarity factor. Annales UMCS Informatica AI **13**(2), 69–81 (2013)
4. Domańska, J., Domański, A., Czachórski, T.: A few investigations of long-range dependence in network traffic. In: Czachórski, T., Gelenbe, E., Lent, R. (eds.) ISCIS 2014, Information Sciences and Systems, part III, pp. 137–144. Springer, Cham (2014)
5. Olejnik, R.: Study of the character of APRS traffic in AX.25 network. In: Kwiecień, A., Gaj, P., Stera, P. (eds.) CN 2013, CCIS, vol. 370, pp. 31–37. Springer, Heidelberg (2013)
6. Arlitt, M., Williamson, C.: Web server workload characterization: the search for invariants. ACM SIGMETRICS Perform. Eval. Rev. **24**(1), 126–137 (1996)

7. Vallamsetty, U., Kant, K., Mohapatra, P.: Characterization of e-commerce traffic. Electron. Commer. Res. **3**(1), 167–192 (2003)
8. 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)
9. Suchacka, G., Domański, A.: Investigating long-range dependence in e-commerce Web traffic. In: Gaj, P., Kwiecień, A., Stera, P. (eds.) CN 2016, CCIS, vol. 608, pp. 42–51. Springer, Cham (2016)
10. Banga, G., Druschel, P.: Measuring the capacity of a Web server under realistic loads. World Wide Web **2**(1–2), 69–83 (1999)
11. Hernandez-Orallo, E., Vila-Carbo, J.: Analysis of self-similar workload on real-time systems. In: RTAS 2010, pp. 343–352. IEEE (2010)
12. Borzemski, L., Suchacka, G.: Web traffic modeling for e-commerce Web server system. In: Kwiecień, A., Gaj, P., Stera, P. (eds.) CN 2009, CCIS, vol. 39, pp. 151–159. Springer, Heidelberg (2009)
13. 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)
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, LNCS, vol. 8181, pp. 347–360. Springer, Heidelberg (2013)
15. Suchacka, G., Borzemski, L.: Simulation-based performance study of e-commerce Web server system – results for FIFO scheduling. In: Zgrzywa, A., Choroś, K., Siemiński, A. (eds.) Multimedia and Internet Systems: Theory and Practice, AISC, vol. 183, pp. 249–259. Springer, Heidelberg (2013)
16. 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, Cham (2015)
17. Jakóbik, A.: Big data security. In: Pop, F., Kołodziej, J., Di Martino, B. (eds.) Resource management for big data platforms. Algorithms, modelling, and high-performance computing techniques, Computer Communications and Networks, pp. 241–261. Springer, Cham (2016)
18. Gong, W.-B., Liu, Y., Misra, V., Towsley, D.: Self-similarity and long range dependence on the Internet: a second look at the evidence, origins and implications. Comput. Netw. **48**(3), 377–399 (2005)
19. Menascé, D.A., Almeida, V.: Capacity Planning for Web Services: Metrics, Models and Methods. Prentice Hall PTR, Upper Saddle River (2001)
20. Rose O.: Estimation of the Hurst parameter of Long-Range Dependent time series. Report No. 137. Institute of Computer Science, University of Wurzburg (1996)