*2012 International Conference on Computer Technology and Science (ICCTS 2012)*
***IPCSIT vol. 47 (2012) © (2012) IACSIT Press, Singapore***
*DOI: 10.7763/IPCSIT.2012.V47.33*

# Long Range Dependence (LRD) in the Arrival Process of Web Robots

Derek Doran and Swapna S. Gokhale

Department of Computer Science and Engineering
University of Connecticut, Storrs, CT, 06269
{derek.doran,ssg}@engr.uconn.edu

**Abstract.** There is strong evidence to suggest that a significant proportion of traffic on Web servers, across many domains, can be attributed to Web robots. With the advent of the Social Web, widespread use of semantic Web technologies, and development of service-oriented Web applications, it is expected that this proportion will only rise over time. One of the most important distinctions between robots and humans is the pattern with which they request resources from a Web server. In this paper, we examine the arrival process of Web robot requests across Web servers from three diverse domains. We find that, regardless of the domain, Web robot traffic exhibits long range dependence (LRD) similar to human traffic. We discuss why, at least in some cases, LRD in robot traffic may not be generated by heavy-tailed response sizes as in the case of human traffic.

**Keywords:** Web robots, Web traffic, long range dependence

## 1. Introduction

Traffic on the Internet from Web robots has increased significantly as they have grown in importance and number. These agents are a critical component of many applications, services, and information repositories that must continually collect data off of the Web. A Web robot can be defined as an autonomous system that sends requests for resources to Web servers across the Internet. The robot then analyses these resources to gain some knowledge in order to fulfil a specific purpose in a broader context.

The proliferation of social networking sites and features has led to a dramatic increase in the volume of time-sensitive and dynamic information posted to the Web. Many organizations believe that this information is very valuable because it can contain users' opinions about products and services or thoughts about current events. This increase in the volume and value of information has led to the development of contemporary Web robots that take on specialized functionality [7], make frequent visits to Web sites [14,6], and employ advanced algorithms to ensure thorough crawls [1,15]. For example, modern Web robots crawl news and social networking sites and user blogs to harvest emotional thoughts and feelings [13].

Given the ever-increasing importance of Web robots, it is necessary to understand the characteristics of their traffic across servers in different domains. Robots and humans may differ in the pattern of their resource requests to a Web server [1]. A robot may send requests at a constant rate to retrieve many different, possibly unrelated resources. A human, however, may deliberately visit a site to retrieve specific information and will request resources according to the behaviour of a Web browser. Thus, despite the significant evidence that human requests exhibit long range dependence (LRD) [4,10], we cannot assume that arrivals of Web robot requests also exhibit such long range dependence.

In this paper, we examine the arrival process of Web robot requests to Web servers from three different domains. The analysis reveals that, similar to humans, robot traffic does exhibit long range dependence (LRD). Because LRD in Web traffic is associated with the heavy-tailed nature of either the response size or

inter-arrival time distributions, we investigate these possibilities for Web robots. Whereas LRD in human traffic is associated with the response size distribution, we find evidence to suggest that in certain domains, LRD in Web robot traffic may be related to the distribution of inter-arrival times.

This paper is organized as follows: Section 2 describes the data. Section 3 provides an overview of LRD. Section 4 examines LRD in robot traffic. Section 5 evaluates the mechanism underlying LRD in robot traffic. Section 6 offers conclusions and directions for future work.

## 2. Data Description

In this section, we summarize the robot and human traffic measured from three Web servers from different domains: a university bookstore E-business (Coop) server, the Roper Center for Public Opinion Research (Roper Center) server, and the University of Connecticut School of Engineering (SoE) academic server. The Coop server hosts a Web application that allows visitors to browse items in a catalogue of merchandise without requiring them to sign-up or register. The Roper Center server provides a large volume of data about public opinion polls and also hosts an application that requires a login to search for and download public opinion datasets. Finally, the SoE server hosts traditional Web pages offering information about the school, faculty and students.

We collected http requests that were logged on these servers between August 1st and September 17th 2009. These logs contain a single record for each http request which includes the sender's IP address and user-agent field, resource requested, request time, and response code. Each request was classified as human or robot by comparing its user-agent field against a database of regular expressions representing well-known Web robots. We choose this simple technique over many robot detection algorithms [8,9,16] because it effectively obtains a sizable sample of robot requests to permit meaningful statistical inferences.

Table 1 summarizes robot and human requests over this six-week period. The demands on each Web server vary widely. For example, the Coop server receives between 3 and 4 times as many requests as the other two servers. However, although the SoE server services the smallest total number of requests, it consumes over twice the total bandwidth of the other two servers. The difference between robot and human activity across these servers may be exacerbated because of the different access levels of the Web applications hosted on the Roper Center and Coop Web servers. Whereas any robot can access the Coop Web application, it cannot access the Roper Center application without authorization.

Table 1: Summary of human and robot requests (Coop; Roper Center; SoE)

| *Statistic* | *Aggregate* | *Human Traffic* | *Robot Traffic* |
|---|---|---|---|
| Total Requests | 4,586,201; 1,409,734; 1,087,826 | 4,439,685; 1,339,803; 947,524 | 146,515; 69,931; 140,302 |
| Average Req/day | 95,545; 29,369; 22,663 | 92,493; 27,912; 19,740 | 3,052; 1,457; 2,923 |
| Bytes Transferred (GB) | 24.22; 15.32; 81.84 | 20.65; 12.27; 66.3 | 3.57; 3.06; 15.55 |
| Average Bytes/day (GB) | 0.5; 0.32; 1.71 | 0.43; 0.26; 1.38 | 0.07; 0.06; 0.33 |
| Unique Resource Requests | 5,014; 69,213; 74,061 | 4,980; 64,396; 34,516 | 2,555; 5,744; 58,115 |

## 3. Long Range Dependence (LRD): An Overview

Long Range Dependence (LRD) is a property of time series processes that are *self-similar*, that is, where the correlation between measurements increasingly farther away does not appreciably diminish. Because of these very long-range correlations, a plot of the number of events per unit time does not smooth out as measurements become coarser. LRD analysis is critical to design high-performance Web servers. For example, Web server performance models incorporating LRD behaviour are more accurate [12] and energy efficient [5]. Thus, given the increasing proportion and sophistication of Web robots, learning if their traffic exhibits LRD is critical for Web servers to achieve high performance.

Self-similarity and LRD are formally defined as follows. Given a stochastic process $X = \{X_t \mid t = 1,2,3,\dots\}$ where each $X_t$ is sampled over the same time interval, define the $m$-aggregated time series $X^{(m)} = \left\{X_t^{(m)} \middle| t = 1,2,3,\dots\right\}$ by summing the values of the original process over non-overlapping consecutive ranges of size $m$ ($m$ is also referred to as the *lag* of the time series). Then, $X$ is *m-self-similar*

with parameter $H$ if $X_t = m^{1-H}X_t^{(m)}$ for all $t$, that is, if the $m$-aggregated time series is proportional to the original process at a finer scale. Furthermore, $X$ is *long-range dependent* if both processes $X$ and $X_t = m^{1-H}X^{(m)}$ have the same asymptotic variance and autocorrelation as $m \to \infty$.

If the process is LRD, as the lag $m \to \infty$ the autocorrelation function will behave as $\rho(m) = \frac{\mathrm{E}[(X_t-\mu)(X_{t+m}-\mu)]}{\sigma^2} \sim cm^{-\alpha}$ where $\mu$ is the mean value of the process, $\sigma^2$ is its variance, and $0 < \alpha < 1$. The impact of this power-law decay can be realized by summing across all the lag $m$ autocorrelations. When the process is not LRD, $\sum_{m=0}^{\infty} \rho(m) < \infty$, thus beyond some lag $m$ the autocorrelation will drop to zero as future measurements become independent of the past ones taken from longer than some lag period ago [3]. When the process is LRD, however, the sum is infinite. This means that in a LRD process the autocorrelation depends on an infinite number of previous measurements.

## 4. LRD Analysis

A statistical approach to determine if a process is LRD is to estimate its Hurst parameter [3], which is defined as $H = (1 - \alpha)/2$. Because of the restrictions on $\alpha$, $H$ must fall between 0.5 and 1 if and only if the process is LRD. Many proposed techniques to estimate $H$ can produce inconsistent results even when evaluated using synthetic data [3]. Different estimators may also be biased against large or small values of $H$ and can perform poorly depending on the periodicity, non-stationarity, and noise of the time series. LRD is unlikely to exist, however, if several estimators cannot consistently estimate of $H$. Therefore, to ensure accuracy, we verify consistency in the estimates of $H$ provided by a suite of estimators. This suite includes the R/S plot, the periodogram, wavelet analysis, and the Whittle estimator.

We plot the R/S statistic and periodogram of robot requests in Figures 1 and 2. A R/S plot shows R/S statistic evaluated at different lags and the lower and upper dotted lines correspond to boundaries where $H$ = 0.5 and $H$ = 1. The R/S estimate is the slope of the straight line that best fits the R/S statistic. Similarly, the slope of the best fitting straight line in the periodogram in Figure 2 corresponds to the periodogram estimate of $H$. The agreement between these estimates suggests that Web robot traffic is LRD across all three servers. In Table 2, we additionally include the Hurst parameters computed through wavelet analysis, and provide a 95% confidence interval for the Whittle estimate. The wavelet estimate closely agrees with the periodogram estimate, and nearly all estimates fall within the confidence interval of the Whittle estimate. Thus, our analysis strongly suggests that, regardless of the domain, Web robot traffic does exhibit LRD.

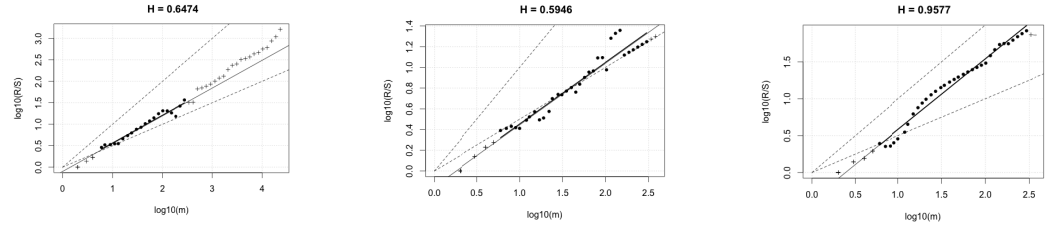


Figure 1: R/S estimation (Coop; Roper Center; SoE)

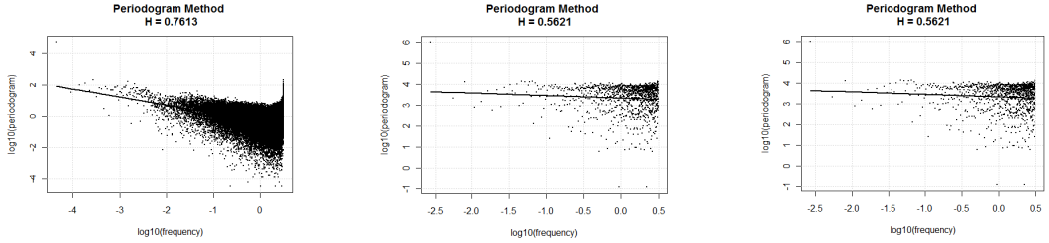


Figure 2: Periodogram estimates (Coop; Roper Center; SoE)

Table 2: Summary of Hurst parameter estimates

| Server | R/S | Periodogram | Wavelet | Whittle [95% CI] |
|---|---|---|---|---|
| Coop | 0.65 | 0.76 | 0.74 | 0.73 [0.72, 0.73] |
| Roper Center | 0.59 | 0.56 | 0.51 | 0.53 [0.49, 0.57] |
| SoE | 0.96 | 0.95 | 0.89 | 0.99 [0.96, 1.04] |

# 5. LRD Generation

A well-accepted model that explains the generation of LRD traffic considers streams of Web traffic that alternate between ON and OFF periods, where the distribution of ON times is governed by the heavy-tailed response sizes with parameter $\alpha_{on}$ and OFF times are associated with the heavy-tailed inter-arrival times of requests with parameter $\alpha_{off}$ [4]. The aggregation of many such streams forms a self-similar fractional Gaussian noise process with Hurst parameter $H = (3 - \min(\alpha_{on}, \alpha_{off}))/2$. Since the distributions of ON times consistently show a heavier tail ($\alpha_{on} < \alpha_{off}$), generally, LRD in Web traffic is associated with the distribution of response sizes.

The distribution of response sizes for human traffic can be related to the heavy-tailed distribution of files hosted on a Web server. If we assume that all Web clients have a sufficiently large cache that is seldom emptied, the distribution of the sizes of all responses should approach the distribution of resources hosted on the server as the number of clients considered increases. Web robot traffic, however, is distinct because each robot may or may not take advantage of a client-side cache to store resources. Furthermore, robots may be unable to interact with Web applications the same way that humans do, preventing them from accessing the same type of information. Thus, it is likely that the LRD in robot traffic could instead be associated with the OFF time distribution (the inter-arrival times of requests) and not with the heavy-tailed distribution of the requested resource sizes.

To examine this further, Table 3 lists the values of $\alpha_{on}$ and $\alpha_{off}$ for robot traffic for each server. These estimates were obtained using the maximum likelihood method across a range of values at which the heavy-tail can start, choosing α that best fits the empirical data according to the Kolmogorov-Smirnov goodness-of-fit statistic [2]. Since $\alpha_{on} < \alpha_{off}$ for the Roper Center and SoE Web servers, LRD in robot traffic on these servers is associated with the response size distribution. On the Coop server, however, we find that $\alpha_{off} < \alpha_{on}$. Inother words, LRD in robot traffic over an e-commerce server may be generated by the distribution of inter-arrival times.

Table 3: Heavy-tail parameters for ON/OFF periods

| Server | $\alpha_{on}$ | $\alpha_{off}$ |
|---|---|---|
| Coop | 3.88 | 3.68 |
| Roper Center | 2.25 | 3.33 |
| SoE | 2.28 | 3.69 |

# 6. Conclusions and Future Work

This paper examined whether Web robot traffic exhibits LRD across Web servers in three different domains. The estimates of the Hurst parameter obtained using several different methods are all in agreement, suggesting that Web robot traffic is indeed LRD, regardless of the domain. Our further analysis revealed that in some cases this LRD may be generated by the distribution of inter-arrival times and not by the distribution of response sizes as in the case of human traffic. Our future research will investigate why LRD in robot traffic across e-commerce servers depends on inter-arrival times, and will assess how this LRD in robot traffic impacts Web server performance.

**Acknowledgements**

The authors would like to thank Rohit Mehta from the UConn School of Engineering, Marc Maynard and Darlene Hart from the Roper Center for Public Opinion Research, and Michael Jean from the UConn Coop for making their Web server logs available for this study.

# References

[1] Anbukodi, S., Manickam, K. Reducing web crawler overhead using mobile crawler. *Proc. Of Emerging Trends in Electrical and Computer Technology*, pg. 926-932, 2011.

[2] Clauset, A., Shalizi, C. R., Newman, M. Power-law Distributions in Empirical Data. *Technical Report, arXiv:0706.1062v2 [physics.data-an]*, 2009

[3] Clegg, R. A practical guide to measuring the Hurst parameter. *Technical Report CSTR-916, 21st UK Performance Engineering Workshop, University of Newcastle,* 2006.

[4] Crovella, M., Bestavros, A. Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes. *IEEE/ACM Transactions on Networking*, 5(6):835-846, 1997.

[5] Deng Y., Meng X., Zhou J. Self-similarity: Behind workload reshaping and prediction. *Future Generation Computer Systems,* 28(2):350-357

[6] Dikaiakos, M. D., Stassopoulou, A., Papageorgiou L. An investigation of Web crawler behavior: characterization and metrics. *Computer Communications*, 28:880-897, 2005

[7] Doran, D., Gokhale, S. Discovering New Trends in Web Robot Traffic Through Functional Classification. *Proc. Of Seventh IEEE Intl. Symposium on Network Computing and Applications,* pg. 275-278, 2008.

[8] Doran, D., Gokhale, S. Web Robot Detection Techniques: Overview and Limitations. *Data mining and Knowledge Discovery*, 22(1):183-210, 2010.

[9] Guo, W., Ju, S., Gu, Y. Web robot detection techniques based on statistics of their requested URL resources. *Proc. Of 9th Intl. conference on Computer Supported Cooperative Work in Design,* pg. 302-306, 2005.

[10] Gupta, H., Mahanti, A., Ribeiro, V. Revisiting Coexistence of Poissonity and Self-Similarity in Internet Traffic, *Proc.of 17th IEEE/ACM Intl. Symposium on Modelling, Analysis, and Simulation of Computer and Telecommunication Systmes,* pg. 1-10, *2009*

[11] Hayati, P., Potdar, V., Talevski, A., Smyth, W. Rule-Based On-the-fly Web Spambot Detection Using Action Strings, *Proc. Of Seventh annual Collaboration, Electronic messaging, Anti-Abuse, and Spam Conference,* 2010

[12] Hernandez-Orallo E., Vila-Carbo J. Web server performance analysis using histogram workload models. *Computer Networks*, 53(15):2727-2739

[13] Horowitz, D., Kamvar, S. D. We feel fine and searching the emotional web. *Proc. Of Fourth Intl. Conference on Web Search and Web Data Mining,* pg. 117-126, 2011.

[14] Huntington, P., Nicholas, D., Jamali H. R. Web robot detection in the scholarly information environment. *Journal of Information Science*, 34:726-741, 2008

[15] Qureshi, M., Younus, A., Rojas, F. Analyzing the web crawler as a feed forward engine for an efficient solution to the search problem in the minimuim amount of time through a distributed framework. *Proc. Of Information Scient and Applications*, pg. 1-8, 2010

[16] Stassopoulou, A., Dikaiakos, M. D. Crawler detection: A Bayesian approach. *Proc. Of Intl. Conference on Internet Surveillance and Protection*, pg. 16-21, 2006.