Modeling Web Request and Session Level Arrivals

Xuan Wang and Katerina Goseva-Popstojanova
Lane Department of Computer Science and Electrical Engineering
West Virginia University, Morgantown, WV, 26506-6109, USA
E-mails: xwang@mix.wvu.edu, Katerina.Goseva@mail.wvu.edu

Abstract

This paper is focused on modeling Web request and session level arrival processes. We propose a statistically rigorous approach which includes testing for non-stationarity and Gaussianity, and uses model selection criterion. Furthermore, a goodness of fit test is applied to each candidate model – ARMA, ARIMA, FARIMA, and FGN – and for validation purpose real data is compared with data simulated from the models. The results based on data extracted from six Web servers with different workload intensities show that (1) there is no one-fits-all solution and (2) servers with high workloads have both request and session traffic modeled well with FARIMA model which is capable of capturing both long-range and short-range dependence.

1. Introduction

Network traffic analysis and modeling play a major role in areas such as workload generation and performance evaluation and prediction. Models that accurately capture the characteristics of the traffic are useful for analysis and simulation, and they aid design and control. Numerous studies have found that different types of network traffic exhibit self-similar behavior [17], [21], [7], [4], [22]. Of special interest in traffic analysis are the asymptotically second-order self-similar processes, also called long-range dependent processes, which are characterized by a hyperbolically (i.e., power-law) decaying autocorrelation function. These findings were in contrast to the classical traffic models such as Poisson process or Markov models and led to using traffic models capable of capturing long-range dependence, such as for example Fractional Gaussian Noise and Fractional ARIMA. This, so called descriptive or 'black-box' approach of traffic modeling has potential to provide traffic description at a vary fine granularity. However, application of time series models to real Internet and especially Web traces often lacked rigor and even more importantly rarely included goodness of fit tests or any way of model validation. In addition, Web sessions – a characteristic of Web workloads which much better represents the users view on the server's performance than individual requests – were disregarded.

Web sessions are defined as a sequence of requests from the same user during a single visit to the Web site. For example, placing an order through the e-commerce site involves requests related to selecting a product, providing shipping information, arranging payment, and receiving confirmation. So, for a customer trying to place an order or a retailer trying to make a sale, the real measure of a Web server success is its ability to process the entire sequence of requests needed to complete the transaction. While modeling request-based workload may provide basis for important tasks such as capacity planning for example, without modeling Web sessions it would be difficult to capture different users behaviors and their effect on the Web servers performance or, for example, to develop admission policies that will increase the chances of sessions completion.

Some more recent modeling efforts of Web and Internet traffic belong to so called constructive or 'white-box' approach which takes into account Web sessions (or IP flows) and their characteristics such as session duration and number of request (or number of packets). For example, in [19] session arrivals were modeled with a Poisson process, and then the number of request per session were described with either inverse Gaussian or Pareto distributions. Similar approach was taken in [11] for modeling the TCP traffic. Although long-range dependence was established for TCP flows, based on semi-experimental approach it was concluded that medeling flow arrivals with a Poisson process does not affect the long-range dependence of the packet level traffic. Our previous work [10] considered session level Web traffic, establishing that session arrivals are long-range dependent for Web servers with high intensity workloads, as well as that intra-session characteristics such as session duration, number of request per session and especially bytes transferred per session are modelied well with Pareto distribution and often are heavy-tailed.

In this paper we go a step further, focusing on modeling request and session level arrivals with time series models capable of capturing short-range dependence, long-range dependence, or both. Specifically,

- We propose well defined, statistically rigorous approach for modeling Web arrival traffic on both request level and session level. Previously published papers on modeling either Internet or Web arrival traffic were missing some important steps, and in some cases used statistical methods that are not appropriate in the specific context.

1550-445X/09 $25.00 © 2009 IEEE
DOI 10.1109/AINA.2009.114

- We consider four different time series models – ARMA, ARIMA, FARIMA, and FGN – which have different ability to capture short-range dependence, long-range dependence, or both. Related work in most cases considered one or at most two models. To our surprise, FARIMA model has never been used for modeling Web traffic despite the fact that it is very flexible and can capture both long-range dependent and short-range dependent behavior.
- We apply the proposed approach on Web traffic from six servers which allows for generalization of the results. We study the effect of the traffic intensity on model choice, and validate the accuracy by comparing the simulated values from the models with the real Web server traces. Very few papers in the past applied time series models to real Web traffic. Thus, in [13] ARMA was used to model one hour of request level traffic of the 1998 Nagano Olympic Games Web site, while in [19] Poisson process was used to model session arrivals extracted from log files of four Web servers collected between 1995-1999.
- The results showed that the request level traffic of the three servers with higher workloads is modeled with FARIMA model. Two of these servers have session level traffic modeled with FARIMA, while the third is modeled with FGN. Both the request and session level traffic of the remaining three servers, which have lower traffic intensities, are modeled well with either only ARMA or with both ARMA and ARIMA models. Combining these results on the session arrival process with distributions of intra-session attributes, such as session duration, number of requests per session, and bytes transferred per session from our previous work [10] yields to a constructive model of Web traffic.

Searching for invariants and exploring the parameter space, as suggested in [8], are important strategies in resolving the difficulties of modeling and simulating the Internet. In [8] it was further suggested that the invariants should be derived from the empirical observations. We believe that as Internet traffic changes and the workload intensity of Web servers increases, these invariants have to be revisited and revised if necessary. Thus, our results show that Web session arrivals of servers with even moderate traffic intensity are long-range dependent, which basically means that one of the invariants given in [8] stating that "Network user session arrivals are well-described using Poisson process" derived on traces of WAN traffic dating 1989-1991 needs to be revised. Note that although it may be possible to disregard the long-range dependent nature of session arrivals when it comes to its effect on the request arrival process as it was suggested in [11], there are applications, such as for example session based admission control, for which it is important to account for the long-range dependence of session arrivals, combined with the heavy-tailed distributions of intra-session attributes.

It should be noted that the approach and analysis presented in this paper are not restricted to Web workloads; they can be used for analyzing other types of Internet traffic. For example, it can be used for modeling the Cellular Digital Packet Data (CDPD) used for mobile data networks which has been shown to be long-range dependent [14].

The rest of the paper is organized as follows. The related work is discussed in Section 2, while a brief overview of the four time series models used in this paper is presented in section 3. In section 4 we present the steps of our approach for modeling the Web traffic, including the specific statistical tests being used. The analysis of the real data from six Web servers and modeling the Web traffic at request and session levels are presented in Section 5. Finally, the paper is concluded in section 6.

2. Related work

Following the pioneering work of Leland, Taqqu, Willinger and Wilson [17] which established that Ethernet LAN traffic is self-similar in nature, in [21] it was shown that the Poisson process cannot be used for modeling different types of WAN traffic due to their long-range dependent nature. These results led to developing traffic models that account for newly discovered phenomena. The earliest models were focused on packet, i.e., request level arrivals and belonged to so called descriptive approach. Thus, [18] suggested that fractional ARIMA (FARIMA) model could be applied on Internet traffic, but did not fit the model to a real network traffic. In [20] it was shown that FARIMA model is better fit than ARIMA and FGN based on the Mean Square Error goodness of fit test and comparison of the real data to the data simulated using the models. The traces of Ethernet traffic used in [20] were 20 seconds long, so the authors did not test for stationarity. In addition, the FGN model was fitted to the data without first testing the traces for Gaussianity. Another paper focused on modeling the Ethernet traffic [12] discussed the parameters estimation of FARIMA model and used the goodness of fit test given in [1]. That work used several minutes long trace and did not test for stationarity. In addition, model selection was not done. In [26], ARIMA and FARIMA models were compared based on the Akaike's Information Criterion. However, the traffic was not tested for stationarity and no goodness of fit test was used in [26].

Similarly to LAN [17] and WAN traffic [21], the analysis of Web traffic at request level showed that the busiest hours are well described as self-similar [7]. The request level traffic from the 1998 Nagano Olympic Games Web site was analyzed in [13] and a piecewise ARIMA model was fitted into four different phases within one hour of Web arrivals. Long-range dependence, formal model selection, and goodness of fit test were not considered in [13]. In a closely related work [25], ARMA model was suggested as arrival process for the G/G/1 performance model of a Web server. In [27], FGN was proposed as a good model to capture the long-range dependence of the request level traffic of commercial Web sites. However, Gaussian distribution assumption was not tested and no goodness of fit test or any other verification methods was used.

The more recent work on modeling and simulating the Internet and Web traffic follows the constructive approach which takes white-box view by first considering TCP flows or Web sessions, and then accounting for the flows or sessions internal structure [19], [11], [3]. Thus, in [19] first the session arrivals were modeled using Poisson process and then the number of request (i.e., clicks) of each session were modeled with inverse Gaussioan or Pareto distribution depending on the server. The model was used to build a synthetic traffic generation tool WAGON, which in most cases produced self-similar request level traffic [19]. This paper did not include detailed analysis of the potential long-range dependence at session level and did not address the effects of non-stationarity. It should be noted that this approach fits into the so called immigration-death process suggested by Cox [5] as a structural way to construct long-range dependent process.

The constructive approach to modeling TCP packet traffic taken in [11] follows the same basic idea of [5]. First, it was established that both packet arrival and flow arrival processes are long-range dependent. Then, based on the so called semi-experimental approach which consisted of several manipulations on the original traces (e.g., modifying aspects of the flow arrival process, the internal dynamics of flows, and the number of packets per flow) it was concluded that modeling the flow arrivals as a Poisson process, with packets within flows following finite GR distribution and heavy-tailed flow volume leads to long-range packet arrival process. This basically means that for the purpose of modeling the packet arrivals, the long-range dependence of flow arrivals can be neglected while keeping the heavy tailed nature of the number of packets in a flow. The traces analyzed in [11] were collected from lightly loaded links at four different locations during 1999-2002. The work presented in the paper was based on two hour long excerptions, which were assumed to be stationary.

Another model for generating synthetic HTTP traffic from the network rather than server perspective was presented in [3]. In that work Web traffic was represented as a collection of independent TCP connections, and then each TCP connection was represented by one or more request-response exchanges between a client and server pair. The TCP connection arrivals were modeled by Fractional Sum-Difference (FSD) model. The model was built and validated based on packet traces from two links collected in 2000. (In [3] the traffic non-stationarity was resolved by breaking the measurements into 5 minutes time blocks.)

In our earlier work [9] we introduced several attributes which collectively describe Web workload in terms of sessions. Then, in [10] we studied in detail different characteristics of both request level and session level Web workloads. For example, all four Web servers considered [10] had a long-range dependent request arrival processes. We also showed that Web session arrivals are long-range dependent on longer periods. In addition, we showed that Pareto distribution cannot be rejected for the three attributes of Web sessions: session duration, number of request per session, and bytes transferred per session. Session duration was heavy-tailed (with infinite variance) for some time periods, number of requests per session was boarder line between finite and infinite variance, while the bytes transferred per session had the heaviest tail (in some cases with both infinite mean and variance). Combining these results related to session attributes from [10], with the results on session arrival process presented in this paper leads to a hierarchical model of Web traffic which follows the recent constructive approach to modeling network traffic. It should be noted that similarly to [19], our focus is the server side model which represents the workload from multiple clients (i.e., 'client cloud') directed to a single Web server. In that respect our work is complementary to the recent models of network traffic [11], [3] which are more appropriate for network traffic simulations, that is, model the traffic over an access link that connects a 'client cloud' with a 'server cloud'.

3. Overview of the time series models

In this section, we present a brief overview of ARMA, ARIMA, FARIMA and FGN time series models.

Autoregressive Moving Average process ARMA(p, q) is a combination of a p^{th} order Autoregressive process, AR(p), and a q^{th} order Moving Average process, MA(q) [2]. ARMA is a linear model which generates a short-range dependent time series.

The Autoregressive Integrated Moving Average (ARIMA) model can be used for modeling non-stationary time series which shows a homogeneous variation about a local trend [2]. ARIMA(p, d, q) model is just an ARMA(p, q) model non-seasonally differenced d times. The difference operator d is assumed to have an integer value. ARIMA is a short-range dependent model since its autocorrelation function (ACF) decays exponentially.

FARIMA(p, d, q) model is the same as ARIMA(p, d, q) model, except the fact that the parameter d is allowed to take nonintegeral values. Obviously, FARIMA model also reduces to ARMA(p, q) model when $d = 0$. FARIMA may fit a long-range dependent time series with $d = H - 1/2$ and $0.5 < H < 1$, where H is the Hurst exponent estimate. FARIMA model is flexible and can capture both long-range dependent and short-range dependent behavior. However, it has high computational complexity and long procedure involved.

Brownian motion consists of steps in a random direction with increments that are independent random variables. Fractional Brownian motion differs from the Brownian motion in the fact that the increments are no longer independent. Fractional Brownian motion is a non-stationary process, but its increments form a stationary Fractional Gaussian Noise (FGN). For $0.5 < H < 1$ the increments tend to display long-range dependence. Due to its Gaussianity, FGN(H) lends itself to a rigorous analytical studies of queueing behavior. Unfortunately, FGN is unrealistic model for bursty non-Gaussian traffic.

4. Proposed approach for Web traffic modeling

The main steps of our approach are as follows:

Step 1.Test the assumptions of all candidate models: ARMA, ARIMA, FARIMA, and FGN.

Step 2.If the time series is non-stationary, remove the trend and periodicity.

Step 3.Estimate the Hurst exponent.

Step 4.Use a formal model selection criterion for ARMA, ARIMA, and FARIMA models. Estimate models' parameters.

Step 5.Perform a goodness of fit test on the candidate models to choose the best model for specific Web server.

Step 6.Validate the results by comparing the autocorrelation function of the actual traffic with the autocorrelation function of the data simulated from the models.

Next, we provide the details of each step, including the specific statistical test and methods being used.

Test models assumptions. Although this seems to be an obvious step, in the past different time series models were applied on Internet and Web traffic data without first testing the corresponding assumptions. To test the assumption of stationarity, common to ARMA, FARIMA and FGN, we use the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) [16] method which has been confirmed to work well on both short-range and long-range dependent time series. Most of the related work either ignored this issue [19], [26], [27] or assumed that the traffic is stationary over a short time periods [3], [11], [12], [20]. To test whether the arrival process follows the Gaussian distribution, which is an assumption of the FGN model, we use the method discussed in [12]. It should be emphasized that without Gaussianity test, fitting FGN model to the data series may not work well [12]. Nevertheless, FGN model was applied in [20], [27] without testing for Gaussianity.

Decomposition of the non-stationary time series. When the time series is non-stationary, before fitting ARMA, FARIMA, and FGN, it is necessary to decompose the signal, that is, remove the trend, seasonal component and other non-stationary factors. To remove the non-stationary factors we use STL (Seasonal-Trend Decomposition of Time Series based on LOESS)[6]. The biggest advantage of LOESS over many other methods is that it is flexible and does not require specification of a function to fit to all of the data in the sample.

Estimation of the Hurst exponent. A predominant way to quantify the self-similarity and long-range dependence is through the Hurst exponent H. For a long-range dependent process $0.5 < H < 1.0$; as H increases from 0.5 to 1.0, the degree of long-range dependence increases. Hurst exponent is a parameter of FARIMA and FGN models. It is important to note that non-stationary factors affect the estimation of the Hurst exponent and may lead to erratic analysis [15]. As we have shown in our earlier work [10], using non-stationary time series often leads to overestimating the Hurst exponent. This means that the Hurst exponent values for the FGN models in [20], [27] and the FARIMA models used in [12], [20], [26] may not be accurate since they were estimated without testing the stationarity. In other words, Hurst exponent has to be estimated after removing the trend and seasonality. In this paper, for both request-based and session-based arrival processes, we estimate Hurst exponent values using Whittle and Abry-Veitch methods.

Model selection and parameter estimation. To select the order of the Autoregressive part (AR) and Moving Average part (MA) of ARMA, ARIMA, and FARIMA models we use Akaike's Information Criterion (AIC) [2], which performs well in model comparison and selection. We build ARMA, ARIMA and FARIMA models combining AR and MA parts with orders from 1 to 10 each (usually 10 is large enough for the orders of AR and MA) and create a 10 x 10 matrix. Finally, based on the lowest value of AIC we choose the best model from these models. Then, for the best model, we estimate the parameters using Maximum Likelihood Estimation.

Goodness of fit test. To test if the model is a good fit to the data set we use the Jan Beran's goodness of fit test [1] which, unlike the Mean Square Error, works well on the long-range dependent data series. We use significance level $\alpha = 0.05$ when testing the null hypothesis, that is, we reject with 95% confidence the null hypothesis that the true spectral density is identically equal to the spectral density of the model if p-value is less than $\alpha = 0.05$. It should be noted that most of the related papers focused on Internet and Web traffic modeling did not use any goodness of fit test.

Model validation based on simulation. For the purpose of validation, we compare the autocorrelation function (ACF) of the actual data series and the autocorrelation functions of data simulated by the models. The simulations were done using the R packages [24]: *fracdiff.sim* for FARIMA, *arima.sim* for ARIMA and with $d = 0$ for ARMA, and

SimulateFGN for FGN.

5. Data analysis and modeling

In this paper, we use empirical data from six Web servers: Web server of the West Virginia University (WVU), Web server of the Lane Department of Computer Science and Electrical Engineering (CSEE), and four Web servers at the NASA Independent Verification and Validation Facility[1]. As the other research papers that considered sessions, we define a session as a sequence of requests issued from the same IP address with the time between requests less than some threshold value [19]. Considering each unique IP address in the access log to be a distinct user clearly is not always true [23]. For example, if a proxy server exists between the user and the server, the IP address in the Web access log will be the address of the proxy, rather than the address of the originating machine. However, in spite of the inaccuracies, we believe that using the IP address provides a reasonable approximation of the number of distinct users. Based on the empirical study of eleven different Web servers presented in our earlier work [9], we adopt a threshold value of 30 minutes.

For each server, we analyze the Web traffic data for four weeks period summarized in Table 1. Note that the Web servers in all tables in this paper are sorted by the total number of requests in descending order.

Next, we apply the approach presented in section 4 on request per second and session per second traffic of the Web servers listed in Table 1.

	Start time	Requests	Sessions
WVU	Feb. 2, 2005	14,856,151	188,056
CSEE	Feb. 2, 2005	481,627	34,325
NASA-Pub2	Sept. 18, 2005	131,058	14,331
NASA-Pvt3	Sept. 18, 2005	61,377	3,400
NASA-Pub1	Sept. 18, 2005	23,896	4,757
NASA-Pub3	Sept. 18, 2005	15,160	2,696

Table 1. Summary data for four weeks period

5.1. Request-based analysis

In this section, we apply the steps of our approach on the Web traffic at the request level. Figure 1 shows the time series plot of the number of requests per second for four weeks period of the WVU raw data. The existence of the trend and seasonal component are obvious from this figure. Using the periodogram method we found that all data sets have slight trend and daily (day/night) and weekly (Monday to Sunday) periodicity. Therefore, for each Web server we estimate and eliminate the trend and cycles from the request level time series using STL [6]. KPSS test [16] run on residuals proved that the time series are stationary. We fit ARMA, FARIMA and FGN in the stationary time series, after eliminating the trend and cycles. We fit ARIMA in the traffic data without cycles only.

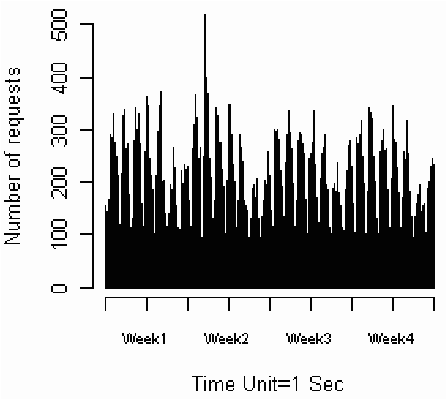

Figure 1. Requests per second - WVU raw data

The test for Gaussian distribution [12] on the request-based time series failed for all Web servers. Nevertheless, we still apply FGN model on all data sets to illustrate by means of goodness of fit test and model validation process that traffic modeling, when it is not done carefully, may lead to inaccurate and often misleading results.

For each server, we estimate the Hurst exponent of the stationary time series using the Abry-Veitch method (see Table 2). It is obvious that the request level workloads of WVU, CSEE and NASA-Pub2 are long-range dependent with the degree of long-range dependence increasing with the workload intensity, which is consistent with the related work [7], [10], [17]. On the other side, NASA-Pvt3, NASA-Pub1, and NASA-Pub3 are unlikely to be long-range dependent since the Hurst exponent values are close or even smaller than 0.5. We believe that this is a result of the fact that NASA-Pvt3, NASA-Pub1, and NASA-Pub3 servers have much lower traffic intensity than WVU, CSEE and NASA-Pub2; this belief is supported by the findings in [17], [7] which showed that many less busy hours in their traces do not show self-similar characteristics.

The best ARMA, ARIMA, and FARIMA models chosen based on the Akaike's Information Criterion are identified in Table 3. The fact that FARIMA cannot be used for modeling the NASA-Pub3 request arrivals since $H < 0.5$ and $d < 0$ is annotated with NA in Table 3.

Based on the p-values of the goodness of fit test given in Table 4 we draw the following conclusions:

- FARIMA is the only model that cannot be rejected for WVU, CSEE, and NASA-Pub2 which shows that their traffic traces exhibit both long-range and short-range dependence (i.e., $d \neq 0$ and $p, q \neq 0$ simultaneously). The fit is better for Web servers with higher Hurst

[1] The Web logs of the NASA IV&V servers were sanitized, that is, IP addresses were replaced with unique identifiers.

	H
WVU	0.66
CSEE	0.60
NASA-Pub2	0.57
NASA-Pvt3	0.51
NASA-Pub1	0.51
NASA-Pub3	0.47

Table 2. Hurst exponent of requests per second

	ARMA	ARIMA	FARIMA
	(p,q)	(p,d,q)	(p,d,q)
WVU	3,4	2,1,4	3,0.16,3
CSEE	3,4	3,1,3	3,0.10,2
NASA-Pub2	7,3	5,1,6	7,0.07,8
NASA-Pvt3	3,1	4,1,7	5,0.01,6
NASA-Pub1	2,4	5,1,4	7,0.01,5
NASA-Pub3	1,4	1,1,3	NA

Table 3. Model identification at request level

	ARMA	ARIMA	FARIMA	FGN
WVU	0.0001	0.0149	0.3390	0.0071
CSEE	0.0037	0.0221	0.1141	0.0179
NASA-Pub2	0.0242	0.0382	0.0944	0.0045
NASA-Pvt3	0.0813	0.0048	0.0065	0.0001
NASA-Pub1	0.0749	0.0565	0.0011	0.0001
NASA-Pub3	0.2012	0.0013	NA	NA

Table 4. p-value at request level

estimate, that is, higher traffic intensity.

- ARMA model cannot be rejected for NASA-Pvt3 and NASA-Pub3 servers, while both ARMA and ARIMA cannot be rejected for NASA-Pub1. These three Web servers have Hurst exponents very close or even lower than 0.5, which results in a good fit with ARMA or ARIMA models which capture short-range dependence.
- FGN is strongly rejected for all servers based on p-values which are significantly lower than 0.05. This result is consistent with the fact that the assumption of Gaussianity failed on all Web servers.

At last, we present the results of model validation based on simulation. Figures 2 to 5 present the comparison of the Autocorrelation function (ACF) of the real data and the data simulated from models for WVU, CSEE, NASA-Pub2, and NASA-Pub1 servers. Based on Figures 2 to 5, the following observations can be made:

- FARIMA model is better fit than ARMA and ARIMA for WVU and CSEE request arrivals. The fit is especially good for WVU request traffic. Choosing the best model among FARIMA, ARMA or ARIMA for NASA-Pub2 based on Figure 4 is not easy, which is due to the fact that the request traffic is only slightly long-range dependent ($H = 0.57$). However, based on $p = 0.09$, the goodness of fit test selects the FARIMA model. Obviously, without statistical test, visualization may not always identify the best model.
- Figures 2 to 5 clearly show that FGN does not capture the characteristics of the request-based traffic of any Web server considered in this paper. These results are not consistent with the work presented in [27] which suggested using FGN to model the request traffic of commercial Web sites. In our case, FGN model was formally rejected for all six Web servers, including the servers with $H > 0.5$.
- ARMA and ARIMA fit most closely the real traffic of NASA-Pub1 server (see Figure 5). Somewhat smoother and more stable ACF of the ARMA model is consistent with slightly larger p-value (see Table 4). ARMA is also the best model for the request traffic of NASA-Pub3 and NASA-Pvt3 which are not shown in figures due to space limitation. It should be noted that the dynamic request traffic of the 1998 Nagano Olympic Games presented in [25] was modeled using ARMA$(2, 1)$ process, without including a goodness of fit test or any way to validate the proposed model.

5.2. Session-based analysis

As in case of request arrivals, session arrival processes for all Web servers considered in this paper are non-stationarity, that is, have a slight trend and daily/weekly seasonal components. Before estimating the Hurst exponent and fitting ARMA, FARIMA, and FGN models we remove the trend and periodicities. Since ARIMA model assumes non-stationary time series, we eliminate only the seasonal component. Of all six servers considered in this paper, only NASA-Pub2 session traffic follows Gaussian distribution.

Hurst exponent, which is a parameter of FARIMA and FGN models, is estimated on the stationary session-based time series. For the session workload we used Whittle method since it has been shown to have desirable statistical properties for Gaussian processes [17]. We also completed the model fitting process for FARIMA and FGN models with the values of Hurst exponent estimated using the Abry-Veitch method. From Table 5 we conclude that Hurst exponents of session arrivals are lower than the corresponding Hurst exponents of the request arrivals. In addition, in most cases Abry-Veitch method provides slightly higher value of H than Whittle method which is consistent with the results presented in [10], [15].

As with request traffic, session traffic of WVU, CSEE, and NASA-Pub2 is long-range dependent, while session level traffic of NASA-Pub1, NASA-Pub3, and NASA-Pvt3 is unlikely to be long-range dependent. The best ARMA, ARIMA, and FARIMA models for each server chosen in terms of the AIC are summarized in Table 6.

The p-values of the goodness of fit test are given in Table 7. Based on these results we make the following observations.

- FARIMA is the only model that cannot be rejected for the session arrival processes of WVU and CSEE.

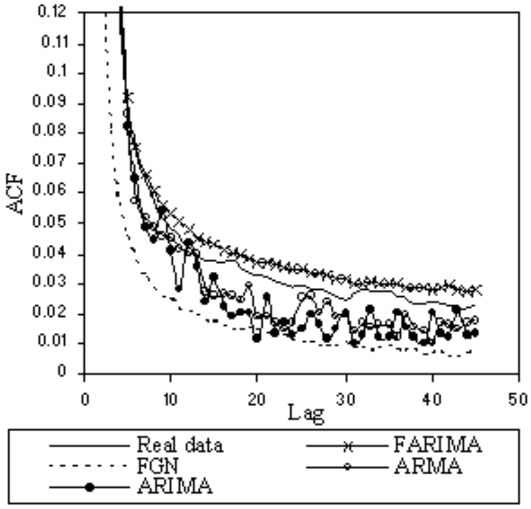

Figure 2. Request time series-WVU

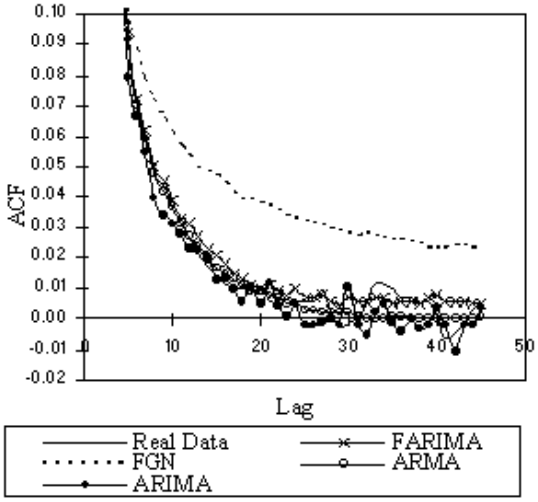

Figure 4. Request time series–NASA-Pub2

	Whittle	Abry-Veitch
WVU	0.59	0.61
CSEE	0.54	0.55
NASA-Pub2	0.54	0.57
NASA-Pvt3	0.47	0.49
NASA-Pub1	0.52	0.50
NASA-Pub3	0.48	0.51

Table 5. Hurst exponent of sessions initiated per second

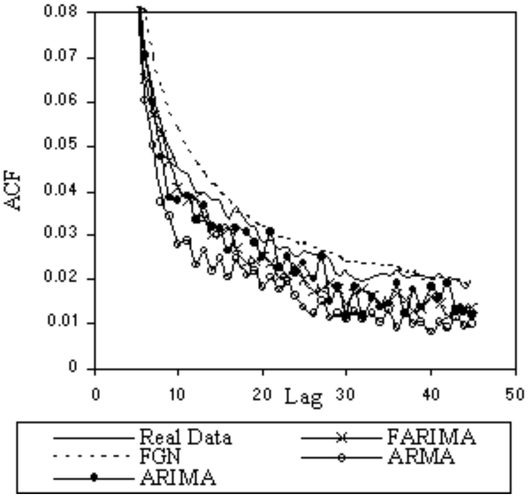

Figure 3. Request time series–CSEE

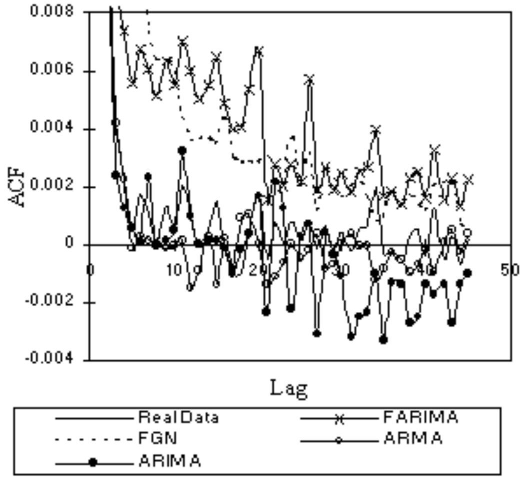

Figure 5. Request time series–NASA-Pub1

	ARMA (p,q)	ARIMA (p,d,q)	FARIMA (p,d,q)
WVU	4,4	3,1,5	4,0.09,7
CSEE	4,2	2,1,3	3,0.04,4
NASA-Pub2	8,7	6,1,7	9,0.04,6
NASA-Pvt3	5,4	3,1,2	NA
NASA-Pub1	2,4	1,1,3	6,0.02,4
NASA-Pub3	3,2	3,1,3	NA

Table 6. Model identification at session level

Similarly as the request arrival process, WVU – the server with the higher Hurst exponent (and higher traffic intensity) – has larger p-value.

- FGN cannot be rejected only for the NASA-Pub2 server, which is the only server that has Gaussian distributed session arrival process.
- ARMA and ARIMA cannot be rejected for NASA-Pvt3 and NASA-Pub1 session level traffic. ARMA model has somewhat better fit than ARIMA model. ARMA model is the only model that cannot be rejected for the NASA-Pub3 session level traffic.

It should be emphasized that we repeated the model fitting process for FARIMA and FGN with the Abry-Veitch estimates of the Hurst exponent for the session arrivals of WVU, CSEE, and NASA-Pub1 servers which are non-

	ARMA	ARIMA	FARIMA	FGN
WVU	0.0001	0.0112	0.2084	0.0142
CSEE	0.0105	0.0247	0.0976	0.0010
NASA-Pub2	0.0022	0.0088	0.0423	0.1298
NASA-Pvt3	0.0697	0.0591	NA	NA
NASA-Pub1	0.1840	0.0904	0.0217	0.0001
NASA-Pub3	0.4362	0.0139	NA	NA

Table 7. p-value at session level

Gaussian distributed. The results given in Tables 6 and 7 are not significantly different and the above observations made about the best model for each server remain the same.

As in case of the request arrival traffic, we validated the best fitted model by comparing the autocorrelation function of the real session traffic and simulated time series models. The figures could not be shown due to space limitation, but we briefly summarize the observations.

- FARIMA model is very consistent for WVU and to large extent for CSEE, the two servers with the highest Hurst exponent estimates.
- Both FGN model and FARIMA seem to be better fit than ARMA and ARIMA to the actual session arrival process of the NASA-Pub2 server. However, only the hypothesis that the session arrival process is FGN cannot be rejected based on the p-value.
- Based on the plot of the ACF it is hard to tell whether ARIMA or ARMA is better model for the NASA-Pub1. ARMA model, however, tends to be smoother and more stable than ARIMA, which is consistent with larger p-value that indicates stronger degree of acceptance.

The results presented in this section are continuation of our earlier work [10] which showed that, unlike TELNET and FTP traffic, piecewise Poisson process can only be used to model Web sessions in a few intervals under low to moderate workloads. The results presented in this paper go further, showing that depending on the nature of the session level traffic, either ARMA, ARIMA, FARIMA, or FGN model can be the best fit to the actual data, with a note that FARIMA model fits well the servers with the highest workload intensity. Along these lines, it should be noted that the server with the highest load in our sample has two to ten times higher traffic intensity than the Web servers with traces dated 1995-1999 used in [19], which were modeled with a Poisson process. This confirms the fact that as the Internet traffic evolves and the intensity of servers workloads increases, there is a need to revisit and revise as needed the invariants established based on older empirical data [8], [19].

6. Conclusion

In this paper we have presented a well defined, statistically rigorous approach for modeling the Web traffic at both request level and session level. The empirical results are based on the data extracted from the access logs of six real Web servers.

With respect to the methods used for modeling of both request and session level traffic the important points are as follows. (1) The stationarity of the arrival process has to be tested, not just as an assumption of models such as ARMA, ARIMA or FARIMA, but also for more accurate estimate of the Hurst exponent which is used as a parameter of both FARIMA and FGN models. (2) The assumption of Gaussian distribution has to be tested before applying the FGN model on the data. Our results show that this assumption is not valid for most Web servers, which consequently means that FGN will results in an inaccurate model. (3) Although visualizing the comparison of the actual data with data simulated from the fitted models is useful, for some Web sites the best model cannot be chosen without a formal goodness of fit test.

A brief summary of the main findings with respect to the traffic characteristics is as follows. (1) Both the request and session arrival processes of the Web servers with the highest traffic, WVU and CSEE, are best described with FARIMA models which capture well both short-range and long-range dependence. The fit is better for the server with higher degree of long-range dependence, that is, higher traffic intensity. (2) The session arrivals of NASA-Pub2 is the only process that is modeled fairly well with FGN. This means that this session arrival process shows only long-range dependence. As in case of WVU and CSEE, the request arrivals of NASA-Pub2 are modeled well with FARIMA model. (3) Both the request and session arrival processes of the remaining three servers which have at least one order of magnitude lower traffic intensity are modeled well with ARMA and/or ARIMA models which typically fit well into data sets with short-range dependence.

The models of the Web arrival traffic presented in this paper, combined with the distributions of session duration, number of request per session and bytes transferred per session [10] provide basis for building an empirically based constructive model of Web traffic which has parameters with clear physical meaning and can be used for developing workload generating tool or modeling and simulation of Web server performance.

Acknowledgements

This work is funded by the National Science Foundation under CAREER grant CNS-0447715.

References

[1] J. Beran, *Statistics for Long-Memory Processes*, Chapman-Hall, New York, 1994.

[2] G. E. P. Box, G. M. Jenkins and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, Third Edition, Prentice-Hall, 1994.

[3] J. Cao, W. S. Cleveland, Y. Gao, K. Jeffay, E. D. Smith, M. Weigle, “Stochastic Models for Generating Synthetic HTTP Source Traffic”, *IEEE INFOCOMM*, 2004, pp. 1546-1557.

[4] I. Cevizci, M. Erol and S. F. Oktug, “Analysis of Multi-Player Online Traffic based on Self-similarity”, *ACM SIGCOMM Workshop on Network and System Support for Games (NetGames’06)*, 2006.

[5] D. R. Cox, “Long-range Dependence: A Review”, in: H. A. David, H. T. David (eds.), *Statistics: An Appraisal*, Iowa State University Press, Ames, IA, 1984, pp. 55-74.

[6] R. B. Cleveland, W. S. Cleveland, J.E. McRae and I. Terpenning, “STL: A Seasonal-Trend Decomposition Procedure Based on LOESS”, *Journal of Official Statistics*, Vol.6, 1990, pp. 3-73.

[7] M. E. Crovella and A. Bestavros, “Self–Similarity in World Wide Web Traffic: Evidence and Possible Causes”, *IEEE/ACM Transactions on Networking*, Vol.5, No.6, Dec. 1997, pp. 835-846.

[8] S. Floyd and V. Paxon, “Difficuties in Simulating the Internet”, *IEEE/ACM Transactions on Networking*, Vol.9, No.4, Aug. 2001, pp. 392-403.

[9] K. Goseva-Popstojanova, A. Singh, S. Mazimdar and F. Li, “Empirical Characterization of Session-based Workload and Reliability for Web Servers”, *Empirical Software Engineering Journal*, Vol.11, No.1, Jan. 2006, pp. 71-117.

[10] K. Goseva-Popstojanova, F. Li, X. Wang and A. Sangle, “A Contribution Towards Solving the Web Workload Puzzle”, *36th Annual IEEE/IFIP International Conference on Dependable Systems & Networks (DSN 2006)*, 2006, pp. 505-514.

[11] N. Hohn, D. Veitch and P. Abry, “Cluster Processes: A Natural Language for Network Traffic”, *IEEE Transactions on Signal Processing*, Vol.51, No.8, 2003, pp. 2229-2244.

[12] J. Ilow, “Parameters Estimation in FARIMA Processes with Applications to Network Traffic Modeling”, *10th IEEE Workshop Statistical Signal and Array Processing*, 2000, pp. 505-509.

[13] A. K. Iyengar, M. S. Squillante and L. Zhang, “Analysis and Characterization of Large–Scale Web Server Access Patterns and Performance”, *World Wide Web*, 1999, pp. 85-100.

[14] M. Jiang, M. Nikolic, S. Hardy, and L. Trajkovic, “Impact of Self-similarity on Wireless Data Network Performance”, *IEEE International Conference Communications (ICC’01)*, 2001, pp. 477-481.

[15] T. Karagiannis, M. Faloutsos and R. H. Riedi, “Long–Range Dependence: Now You See It, Now You Don’t!”, *IEEE Globecom*, Vol.3, 2002, pp. 2165-2169.

[16] D. Kwiatkowski, P. Phillips, P. Schmidt and Y. Shin, “Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root: How Sure are We that Enconomic Time Series have a Unit Root?”, *Journal of Econometrics*, Vol.54, Oct/Dec 1992, pp. 159-178.

[17] W. Leland, M. S. Taqqu, W. Willinger and D. V. Wilson, “On the Self-Similar Nature of Ethernet Traffic”, *IEEE/ACM Transactions on Networking*, Vol.2, No.1, Feb. 1994, pp. 1-15.

[18] J. Li, A. Wolisz and R. Popescu-Zeletin, “Modelling and Simulation of Fractional ARIMA Processes based on Importance Sampling”, *ACM Symp. Applied Computing*, 1998, pp. 453-455.

[19] Z. Liu, N. Niclausse and C. Jalpa-Villanueva, “Traffic Model and Performance Evaluation of Web Servers”, *Performance Evaluation*, Vol.46, 2001, pp. 77-100.

[20] J. Liu, Y. Shu, L. Zhang, F. Xue and O. W. W. Yang, “Traffic Modeling Based on FARIMA Models”, *IEEE Canadian Conf. Electrical and Computer Engineering*, Vol.1, 1999, pp. 162-167.

[21] V. Paxson and S. Floyd, “Wide–Area Traffic: The Failure of Poisson Modeling”, *IEEE/ACM Transactions on Networking*, Vol.3, No.3, June 1995, pp. 226-244.

[22] K. M. Rezaul and V. Grout,“An Overview of Long-Range Dependent Network Traffic Engineering and Analysis: Characteristics, Simulation, Modeling and Control ”, *2nd International Conference on Perforamnce Evaluation Methodologies and Tools (ValueTools’07)*, Oct. 2007.

[23] M. Rosenstein, “What is Actually Taking Place in Web Sites: E-Commerce Lessons from Web Server Logs”, *2nd ACM Conference on Electronic Commerce (EC’00)*, Minneapolis, MI, Oct. 2000, pp. 38-43.

[24] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2005, http://www.R-project.org.

[25] M. S. Squillante, D. D. Yao and L. Zhang, “Web Traffic Modeling and Server Performance Analysis”, *38th IEEE Conference on Decision and Control*, Vol.5, 1999, pp. 4432-4439.

[26] Y. Takahashi, H. Aida and T. Saito, “ARIMA Model’s Superiority over f-ARIMA Model”, *International Conference on Communication Technology*, 2000, pp. 66-69.

[27] C. H. Xia, Z. Liu, M. S. Squillante, L. Zhang and N. Malouch, “Traffic Modeling and Performance Analysis of Commercial Web Sites”, *ACM SIGMETRICS Performance Evaluation Review*, Vol.30, No.3, Dec 2002, pp. 32-34.