Characterizing Non-stationarity in the Presence of Long-range Dependence

Krishna Kant and M. Venkatachalam
Intel Corporation, Oregon, USA

Abstract— **Based on ours and other studies, it is becoming evident that Internet traffic is often nonstationary. Since such traffic is generally found to be long-range dependent as well, and the two properties can get confused, it is important to have a method for separately characterizing them. This paper proposes such a methodology aimed specifically for traffic scenarios where nonstationarity comes into play at a relatively large time-scale. This methodology allows us to accurately characterize nonstationarity and and long-range dependence and thereby synthetically generate traffic exhibiting it. Such synthetic traffic can be used for performance engineering, dynamic resource allocation and overload control. We illustrate the methodology by applying it to traffic traces from three very different types of commercial web-service environments.**

Keywords: Traffic characterization, long-range dependence, non-stationarity, scaling properties.

I. Introduction

A. Why is non-stationarity important?

It is well established by now that high-speed network traffic, including the web traffic, shows long-range dependence and self-similarity [5]. Although most of the traffic studies look at byte level traffic, we have found that the time series of successive requests to an Internet server shows essentially the same properties as the byte level traffic, as reported in [10]. In this context, long-range dependence can be intuitively explained by the heavy tailed nature of the the user inactivity (think) times in sessions. The traditional tests for self-similarity (e.g., variance-time plot, R/S statistics, etc. [13]), require the stationarity of the time series. In fact, it has been shown that a nonstationary Poisson process can appear to be long-range dependent according to popular tests such as variance time plot [7]. Much of the work on web-traffic analysis assumes that the traffic approximately stationary. Our analysis in the past [10] also indicated that this is a reasonable assumption. However, our recent analysis of traffic logs from several business to business (B2B) and business to consumer (B2C) suggests that "busy periods" (in the traditional telecommunications sense) are often hard to identify, and even over short intervals of 10-20 minutes, the traffic is nonstationary. Unfortunately, a 10 minute timescale is not much larger than the time-scale of user actions or system responses; therefore, such a time-scale could be relevant from a long-range dependence perspective. This complicates matters since nonstationarity could be easily mistaken for long-range dependence and vice-versa. A recent paper [6] has also observed pervasive nonstationarity in Internet traffic, but this characterization is at the TCP level rather than the HTTP request level.

B. Overview of Modeling methodology

Given the importance of non-stationarity in Internet traffic in general and E-commerce traffic in particular, and the fact that the analysis of long-range dependent nonstationary traffic is an open problem, we propose a methodology to decouple the characterization of non-stationarity and long-range dependence. Such a characterization is essential for synthetic load generation and performance engineering purposes. Non-stationarity can be modeled in two flavors: parametric and non-parametric. Parametric models [6] like ARIMA are easy to characterize for nonstationarity since they have well understood parameters, but are of less practical value in the context of traffic engineering. Non-parametric models on the other hand are the most practical models for Internet and e-commerce traffic but are hard to characterize for non-stationarity. Further, the presence of long-range dependence makes it even more complicated. In [16] the authors devise a methodology to characterize long-range dependence in the presence of level-shift non-stationarity. But they make no effort to characterize the nonstationarity itself, which we will see is an important component of traffic modeling.

In order to simplify the problem a little, we make a few assumptions. Firstly, we concern ourselves only with wide sense stationarity, for which only the first two moments of the marginal distribution and the autocorrelation function need be considered. Secondly, we assume, with some justification, that the correlational properties are also stationary. Without this assumption, the inherent long-range dependence properties of the time-series make the analysis intractable. Thus, checking for stationarity is only confined to checking for the stationarity of the mean and variance of the marginal distribution. Also, consistent with our observations regarding the web traffic, we assume that the traffic can be considered reasonably stationary except over intervals (or time-scales) that are much larger than the time-scales at which the measurements are made.

With these assumptions, we obtain the non-stationarity characterization in terms of a non-stationarity time-scale and a non-stationarity profile (NSP). The notion of non-stationarity time-scale (NST), although theoretically imprecise, is introduced to capture the practical observation that traffic characteristics can be considered reasonably "stable" or stationary when viewed over small time intervals, but not so over large time intervals. The concept of NST essentially specifies at what granularity (or time-

scale) we have to worry about non-stationarity. Knowing the NST would allow us to examine the traffic at those time-scales and also allow us to generate the observed traffic reasonably faithfully via a traffic generator.

Our results from the analyzed traces show that the NST is typically 10-20 minutes. While this time-scale is large enough that one doesn't need to consider its queuing consequences directly, it is short enough to affect estimation of long-range dependence. This range of time-scale may also prompt dynamic engineering procedures wherein the resources made available at adjusted dynamically. A study of such procedures would require artificial traffic generation, which is another motivation for the characterization in this paper. A related issue is one of overload protection and control. A characterization of the "stability" properties of the traffic would allow better techniques for overload control. Some basic overload control schemes are described in [8], and they could be enhanced to take advantage of the nonstationarity characterization discussed here.

C. Application to Traffic Traces

We apply our characterization methodology to three traffic traces, henceforth referred to as Trace1, Trace2 and Trace3 respectively. These traces have been obtained from 3 very different sample environments, as described below:
Trace1: Trace1 comes from a large business-to-business (B2B) e-commerce site that conducts multi-billion dollars per month of business. A close study of this site indicates a number of unique characteristics, many of which are believed to be typical of B2B sites. In particular, all transactions (rather than just the sensitive ones) are secure (i.e., HTTPS). Also, such traffic shows a typical day-time busy period pattern (reminiscent of busy-period pattern in traditional telephony systems), although this behavior tends to be washed out a bit because of accesses from several time-zones. Also, an analysis of transaction types shows a lot of serious business being conducted (ordering, order checking, payment, etc.) instead of being dominated by idle "window shopping" (or browsing). The web-pages are found to be rather simple with little in the way of fancy advertising (meaning, not too much dynamic content beyond what comes from the back-end server). Also, B2B requests come exclusively from business customers who typically have high-speed Internet connections. These characteristics imply that there are fewer errors, retries, etc.
Trace2: Trace2 comes from a large business-to-consumer (B2C) e-commerce site. A close study of this site also indicates a number of unique characteristics, many of which are again believed to be typical of B2C sites. In particular, most transactions are non-secure, with HTTPS reserved only for the sensitive parts (e.g., order checkout and payment). Interestingly, the busy-period behavior here is bit hard to identify except for the substantial traffic at nights and weekends as well. As with Trace1, the temporal characteristics are somewhat muted by the fact that orders come in from a variety of time-zones. An analysis of transaction types shows not much serious business being conducted, instead, most users do extensive product browsing, often even going to the step of putting things in the shopping cart only to dump them and exit out at the end. Thus HTTPS transactions don't play that much of a role here, however, the web-pages do contain a significant percentage of images and dynamic information. This combined with primarily low-speed client connections usually results in a lot of abort and retry traffic.
Trace3: Trace3 comes from a non-caching proxy server site serving a large organization. These proxies not only provide access to the content created by the organization itself but also to all external access from within the organization. Thus Trace3 provides an example of traffic that is not limited to a specific native site but rather a mixture of traffic going to a large number popular and not-so-popular sites. Because of its nature, Trace3 traffic is not tied to any particular back-end database or, for that matter, any sort of commercial service per se. The incoming requests in this case were primarily from organizational customers with high-speed LAN access.

Figures 1 – 3 show, respectively, Traces1-3 for a selected week. For uniformity, only the weekday traffic is shown in all cases. One characteristic to note is the lack of stability of traffic pattern for both Trace1 and Trace2 over different days. Although Trace3 also shows some irregularities at the boundary of days 1 and 2, those are caused by special events (e.g., rearrangement of servers). For the most part, the daily pattern appears better defined in Trace3 than in other traces. This could well be a result of the fact Trace3 concerns traffic passing through a proxy server and thus consists of a much broader mix than the traffic going to specific sites (as in Traces 1 and 2).

One point to note from Traces 1 and 2 is that any fine-granularity traffic analysis based on a "typical" day's worth of traffic could be highly erroneous, simply because there is no typical day per se. This is the basic motivation for the characterization of the nonstationarity (or that of lack stability) in the traffic, since without such a characterization, an artificial traffic generator will be unable to produce the kinds of traffic patterns we see in Figures 1 and 2.

The rest of the paper is organized as follows. Section 2 proposes a methodology for decoupling the characterization of long-range dependence and non-stationarity and section 3 deals with its application to the traffic traces considered. Section 4 discusses the synthetic generation of non-stationary long-range dependent traffic based on our methodology. Section 5 concludes the paper.

II. Our Methodology

In this section we propose a methodology to decouple the characterization of non-stationarity and long-range dependence in a general traffic trace.

A. Hypothesis

In order to make the study feasible, we assume the following and attempt to verify them using the data:
1. Correlational properties of the traffic are stationary, i.e., the nonstationarity is present only in the mean of the arrival process.
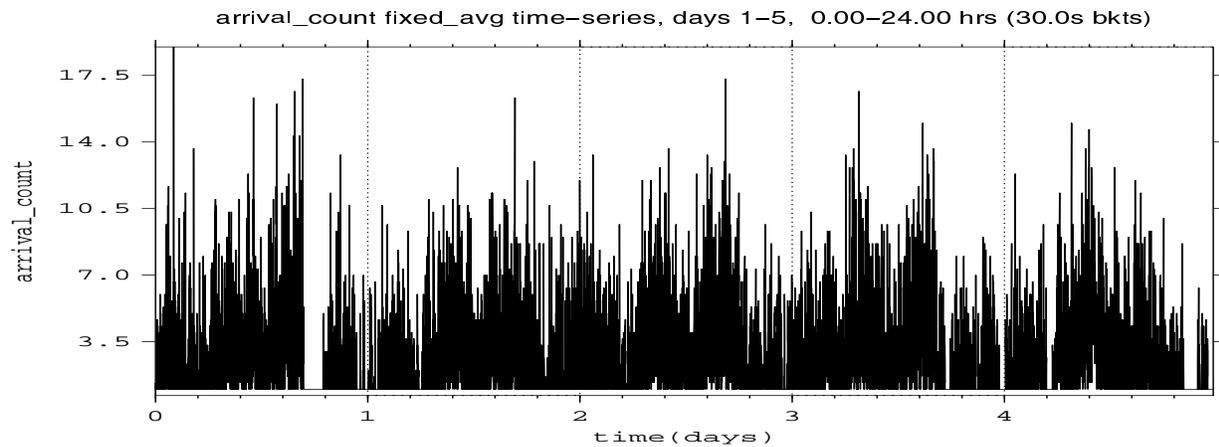
arrival_count fixed_avg time−series, days 1−5, 0.00−24.00 hrs (30.0s bkts)

Fig. 1.   Five days of B2B traffic, bucket-size = 30s

arrival_count fixed_avg time−series, days 1−5, 0.00−24.00 hrs (30.0s bkts)
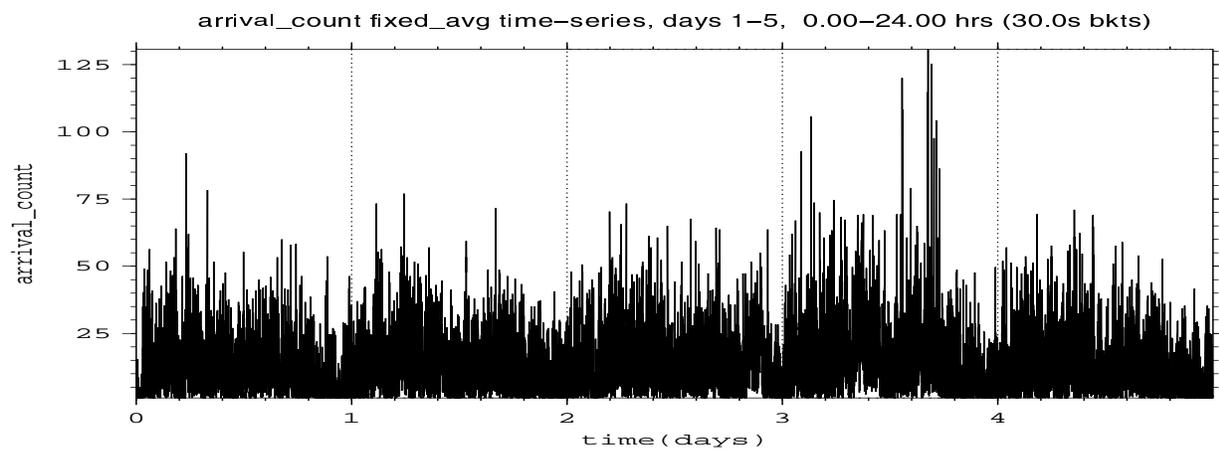
Fig. 2.   Five days of B2C traffic, bucket-size = 30s

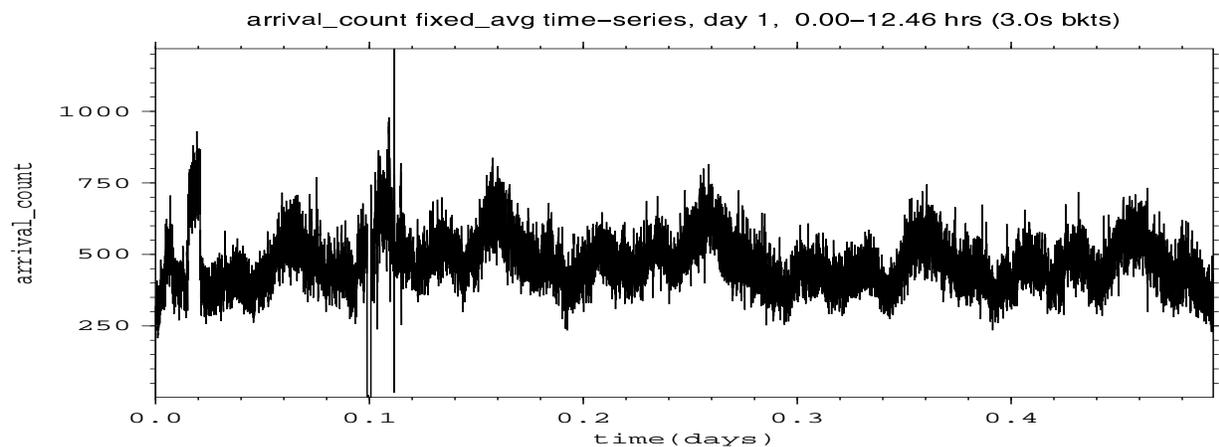**arrival_count fixed_avg time−series, day 1, 0.00−12.46 hrs (3.0s bkts)**

Fig. 3.   Five days of Web-browsing traffic, bucket-size = 30s

2. The timescales at which the non-stationarity has a significant impact are much larger than the timescales at which the measurements are made.

3. The nonstationarity enters into the picture at a specific time-scale, henceforth denoted as τ seconds. That is, the traffic can be considered approximately stationary at time-scales smaller than τ. We call this the non-stationarity time-scale (NST).

Intuitively, assumption (1) amounts to saying that the fundamental nature of user behavior (i.e., heavy-tailed active and inactive periods) does not change over time; only the time between successive requests may change over time.

Scaling properties of such a model have been studied in [16] which shows that the wavelet based estimator of H parameter (or "AV" estimator) works well in such an environment.

Techniques for removing nonstationarity from a time-series are well studied [4]. These techniques attempt to remove trend (i.e., data modulated by a slowly varying function) and seasonality (data modulated by a periodic function). One technique for trend removal is to subtract a moving average from the original time-series. We shall use this approach in this paper.

The assumptions essentially imply that the non-stationarity can be described by a slowly-varying random process operating at the time-scale of τ. Let Δ denote the duration of a bucket over which the original arrival process $\{X_i, i = 1, 2, \ldots, N\}$ is defined. Then, $\tau = \Delta \times J$ for a suitably chosen J. In general, it is difficult and perhaps not meaningful to determine J very precisely; therefore, we shall work with a set of chosen values.

B. Algorithm

The general algorithm for doing this can be cast as follows:
1. Based on the visual inspection of the original arrival time series $\{X_i, i = 1, 2, \ldots N\}$, choose the smallest value of J at which nonstationarity may appear to come into play.
2. Extend the time-series on each end by duplicating the first and last $J/2$ elements.[1] Next, find the moving average of the time-series over J buckets, i.e., construct $Y_n = \frac{1}{J} \sum_{i=n-J/2}^{n+J/2} X_i$, $n = 1, 2, \ldots, N$. We call the resulting process $\{Y_n\}$ as the *nonstationarity profile* of the traffic.
3. Find the residual process $\{R_i = X_i - Y_i, i = 1, 2, \ldots N\}$. If $\{Y_i\}$ truly represents nonstationary component of traffic intensity, the process $\{R_i\}$ must be stationary.
4. Check if $\{R_i\}$ is approximately stationary and has the same scaling behavior as the original process $\{X_i\}$. If so, we have the desired value of J; otherwise, choose another J and go to step 2.
5. Once J is determined, take the corresponding nonstationarity profile $\{Y_i\}$ and characterize it as a random level-shift process.

There are a couple points of interest with the algorithm described above.

Firstly, we do not need more complicated methods of removing non-stationarity (like wavelet techniques etc) since our basic assumption is that the non-stationarity occurs in the marginal distribution of the arrival process and not in the correlational properties. Hence a simplistic moving average technique is justified in order to remove the trend in the arrival time series.

Secondly, we assume that the trend in the variance is negligible. We tried proceeding without this assumption, by basically attempting to transform the residual time series to a zero mean, unit variance time series. With this transformation, we needed the higher moments of the residual

---

[1]It is convenient to use even values of J in this construction.

time series for the stationarity tests described below. This did not yield sensible results and violated the assumption of wide sense stationarity.

C. Stationarity Tests

In step 3 above, we need to check if the residual process $\{R_i\}$ is stationary. Stationarity is often concluded informally by visual inspection of data, since good but general stationarity tests are hard to come by. Generally, there are two types of tests: (a) Parametric and (b) Non-parametric. Parametric tests relate to the situation where a traffic model exists and the stationarity checking amounts to finding out if the parameter values are time-invariant. For example, one could have a ARMA(p,q) (autoregressive moving average) model and check if p and q are time invariant. Similarly, in some cases, the distributions may be known, and we only need to worry about time-invariance of distributional parameters. Parametric models are inappropriate for the problem at hand since the nature of the traffic itself is not well understood.

Nonparametric tests do not require any underlying model and thus can be easily applied to any data. Most tests available in the literature consider the stationarity of stochastic processes that can be modeled via IID random variables under steady state. For example, the point of stationarity is often recognized as the one where the auto-correlation function has decayed to some small value. Such direct assumption of independence is contrary to the situation examined here where we already know that the traffic of interest is long-range dependent. In fact, we know of no test that will work well if the correlations are strong. Although theoretically speaking the long-range dependence extends to infinity, in practice, the correlations tail off at intervals of 10-15 minutes for the web traffic we have examined. We exploit this to consider two well-known stationarity tests in the following. In addition, we also introduce a new metric called "overlap metric" primarily as a backup for these standard tests.

C.1 Reversibility Test

Let $D_1, \ldots, D_N$ denote a set of data points. For $i, j \in 1 .. N$, consider the total number of times that $D_j > D_i$ for $j > i$. Let us denote this number by V. Obviously, V lies in the range $0 .. N(N-1)/2$. For a sequence that does not show any trend or seasonality, only 50% of the total number of values will show the above property. Also, if the values are only weakly correlated, the central limit theorem indicates that as $N \to \infty$, the distribution of V tends to Normal with mean $\mu_V = N(N-1)/4$ and variance $\sigma_V^2 = N(N-1)(2N+5)/72$. Thus, if the original process is stationary, the value of V will be contained in a small interval around μ_V. That is, with $Z = (V - \mu_V)/\sigma_V$ and a given confidence level $1 - \alpha$, we have the usual confidence interval $\text{Prob}(Z_{\alpha/2} < Z \leq Z_{1-\alpha/2}) = 1 - \alpha$.

We found this test to be quite weak, in the sense that at 90% or higher confidence level, all residual time series passed this test irrespective of the nonstationarity time scale J. For this reason, we do not report results of this

test in succeeding sections.

C.2 Runs Test

This test also looks for trends and seasonality, but works with aggregated values rather than the individual samples $X_1, \ldots, X_N$. Let's suppose that we divide the original series into K equal parts and compute some statistical measure Z (e.g., mean, mean-square, variance, etc.) over each part. In particular, taking the example of mean-square, we would define:

$$Z_j = \frac{1}{m} \sum_{i=(j-1)m+1}^{jm} X_i^2, \quad j = 1, \ldots, K \quad (1)$$

where $m = N/K$ is the size of each part.[2] In order to ensure that the successive data points are only weakly correlated, we need to choose m sufficiently large. We compute the median value over all Z_j's, which we denote as M_Z. Now, if X_i^2's do not show any trend or seasonality, Z_i's should be distributed randomly around the median value M_Z. This can be characterized by the "run statistic" W, which is the number of samples for which the sample value stays on one side of the median. For example, if the first five W_j's are less than the median value and next 3 are above, we have $W_1 = 5$, $W_2 = 3$. This gives us a time-series $W_1, W_2, \ldots, W_L$ for some L in the range $1 \, .. \, K$. The lack of trend or seasonality means that the number of runs L will be confined to a small range around the mean of $K/2 + 0.5$ (assuming K is even).

The details of the run-test are bit more complex. If Z_i's are IID, one can write equations for the distribution of L based on *level crossing* arguments [3]. In particular, if the joint distribution of random process $Z(t)$ and its time derivative is Gaussian, the distribution of L can be computed. The Gaussian property again follows from the central limit theorem and requires weak dependence between Z_j's.

We found the runs test to be fairly reliable in indicating stationarity. Compared with the reversibility test, this is perhaps a result of further aggregation of X_i's into Z_j's which leads to lower variance and hence a narrower range for the confidence interval. For the results reported in the next section, we chose a confidence level of 95%. This, along with $K = 50$, gives the confidence interval of 19-32 for the runs test statistic L. The choice of $K = 50$ is somewhat arbitrary but was made after some experimentation to balance out the needs of minimizing the dependencies and having adequate points for a reliable runs test. In particular, m, the size of each subseries was 8000 secs which appears adequate to significantly reduce correlations between successive data points and thereby make the Z_i's approximately IID.

C.3 Overlap Metric

The main motivation for devising this metric was to have a backup for runs test results since the reversibility test

---

[2]For simplicity, we assume that K divides N evenly.

did not provide much information. This metric can be turned into a test, but we instead decided to retain it as an informal measure of "degree of stationarity". In our extensive tests, we found overlap metric to be quite reliable in identifying stationarity.

To define the metric, we again start with the given time series $X_1, \ldots, X_N$. Let $X_f, X_{f+1}, \ldots, X_{f+m}$ denote a subseries of size m starting at offset f. Again, the size m must be long enough so that the correlations between successive data points become insignificant. In particular, we consider L such subseries starting at regularly arranged offsets $f_1, \ldots, f_L$. We take the subseries to be non-overlapping, although small overlaps should not matter. Let Y_j, $j \in 1 \, .. \, L$ denote some statistical measure (e.g., mean, mean-square, etc.) over these L subseries. By definition, for a stationary process, all such estimates should be independent of the starting offsets.

In order to characterize the notion of time-independence, we construct confidence intervals for Y_j's, $j = 1 \, .. \, L$. This is done by assuming that the subseries is long enough so that classical techniques based on central limit theorem become applicable. In particular, we divide the subseries into a small number of parts, obtain the statistical measure for each part, and then construct the confidence interval using t-distribution. This yields L confidence intervals, denoted $\{L_j, U_j\}$ for $j = 1 \, .. \, L$ at some confidence level $1 - \alpha$. Next, we need to construct some overlap metric O that quantifies the extent of overlap between these intervals. Obviously, if the intervals have a good overlap, we can claim that the chosen statistical measure is stationary.

Several potential metrics were considered for O. The most obvious one is the intersection between all intervals, but this is very stringent and intolerant of even one outlier. In fact, for real-life data, even the extent of pairwise intersections is not a very robust metric. Therefore, we considered the following much weaker overlap metric: Consider all $L(L-1)/2$ pair-wise overlaps and count the fraction of overlaps that are non-null. For the same data set, this metric is a monotonically decreasing function of L. Thus to keep the results comparable, we used $L = 50$ in all cases. It is found that with $L = 50$ the artificially generated unif(0,1) IID random numbers result in a metric value of very close to 1.0.

III. APPLICATION TO THE TRAFFIC TRACES

A. Application of Stationarity Tests

In applying the stationarity tests discussed above, we need to choose the statistical measure(s) of interest. Since the residual process $\{R_i\}$ is obtained by subtracting out the moving average, it will have nearly zero sample means; therefore, there is no point in doing the tests on the mean. Thus, it suffices to consider the mean-square statistical measure.

Table 1 lists the results from the stationarity test for Traces1-3. The first row, marked as "orig" shows the test run on the original (presumably nonstationary) time-series. It is seen that the runs test metrics are significantly below the lower bound of 19 for both Trace1 and Trace2 which

time scale	Trace1		Trace2		Trace3	
	runs	ovl	runs	ovl	runs	ovl
orig	14	0.705	16	0.731	19	0.827
150	18	0.782	14	0.824	18	0.818
300	14	0.795	17	0.831	25	0.727
600	16	0.821	14	0.846	24	0.764
700	18	0.821	20	0.846	26	0.782
900	14	0.808	20	0.870	19	0.818
1100	22	0.808	16	0.868	24	0.800
1300	23	0.833	16	0.836	20	0.800
1500	14	0.795	16	0.836	28	0.800
1700	20	0.756	16	0.838	28	0.800
1900	18	0.821	16	0.824	25	0.818
2100	14	0.795	16	0.809	23	0.818
3600	16	0.795	17	0.816	20	0.827
7200	16	0.808	14	0.824	23	0.836

TABLE I

STATIONARITY TESTS FOR TRACES1-3 TRAFFIC

Fig. 4. Scaling properties of Trace1

Fig. 5. Scaling properties of Trace1 during high traffic period

would result in the rejection of stationarity hypothesis. The overlap metric confirms this by returning a rather low value of non-null overlap fraction of 0.705 for Trace1 and 0.730 for Trace2. For Trace3, the runs test metric is right at the lower bound of 19, which means that the original series can be considered marginally stationary. The overlap metric confirms this by returning a rather high value of 0.827. The reason for marginal result for Trace3 appears to be traffic considered over entire day (rather than just the high-traffic periods) which brings the diurnal variations into play.

As a validity check, we also carried out the tests for 40,000 artificially generated unif(0,1) random numbers. The runs test metric in this case was 23 and overlap metric of almost 1, although in some cases the overlap metric was as low as 0.96. This confirms that tests are saying something significant about the data as compared to the IID data.

The remaining rows show the results for the residual process $\{R_i\}$ for τ values ranging from as low as 150s to as high as 2 hours. We do not consider τ values greater than 2 hours since the diurnal nonstationarity comes into play at such timescales. It is clear that for Trace1 the runs test is passed by a significant margin only for τ in the range 1100-1300 secs. Both the runs metric and overlap metric are the highest at $\tau = 1300$ which prompts us to tentatively choose $\tau = 1300$ secs for Trace1. For Trace2, this happens only for τ in the range 700-900 secs. Both the runs metric and overlap metric are the highest at $\tau = 900$ which prompts us to tentatively choose $\tau = 900$ secs for Trace2. Note that the smaller τ for Trace2 supports the intuitive observation that we have made from several studies that B2C traffic is "less stable" than B2C traffic. For Trace3, we find that the runs test is passed in nearly all the cases, which means that there is no longer any small time-scale trend or seasonality. (The results for $\tau \gg 1$ hour will begin to admit daily variations and the metrics will suffer.) The overlap metric appears to confirm this for the most part, although there are some disturbing cases (e.g., 300 sec case where the runs test metric is very good but the overlap metric is not.) In any case, the results do confirm what we suspected from the time-series plots: Trace3 appears stationary up to the
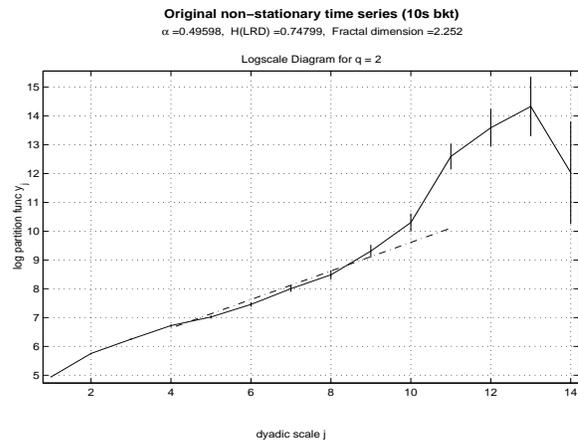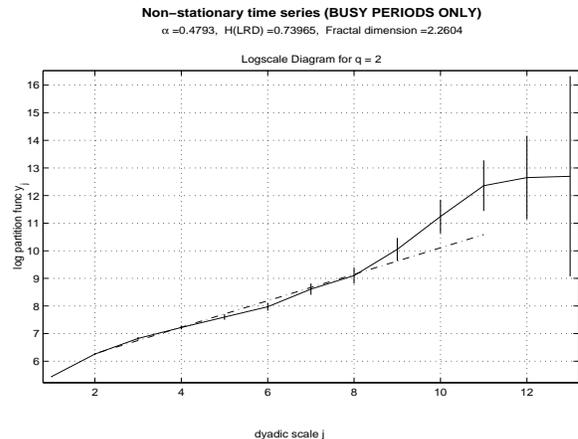
duration of a few hours but the other two do not.

B. Dealing with Long-range dependence

One major difficulty caused by long-range dependence is that all hypothesis tests invoke the central limit theorem on the set of test samples (usually derived by aggregating original data points) to be only weakly interdependent. This is impossible for true long-range dependent traffic. Fortunately, real web traffic shows long-range dependence only up to 10-15 minutes, beyond which weak dependence assumption can be made. That is, the test samples should be obtained by aggregation over time intervals exceeding 30 minutes or so. We have tried to do this in carrying out the tests presented in the last section. Nevertheless the question remains whether the apparent seasonality/trend at small values of τ is really a result of long-range dependence. In order to address this concern, we also need to look at the Hurst parameter H for the residual time-series to ensure that the long-range dependencies are not affected significantly. This is done by comparing the scaling behavior of the original non-stationary time series and that of the residual time series.

To this end, we use the wavelet-based tests developed by

Abry and Veitch [1] (the "AV" estimator), which has several advantages over traditional methods such as variance-time plots. In fact, the "AV" estimator has been shown to be robust with respect to non-stationarity in the form of smooth level shifts [16]. Appendix B includes a brief description of the "AV" estimator. This estimator starts with a log-log plot of the energy in the wavelet coefficients (for the wavelet transform of the arrival time series) vs. the time-scale. Such a plot is known as the *logscale diagram* and shows the scaling properties of the arrival process over the time-scales of interest. If the plot shows a linear behavior over a set of time-scales consistent with long-range dependence, the corresponding Hurst parameter is estimated by a modified linear regression procedure detailed in [1].

Fig. 4 shows the logscale diagram for the original non-stationary Trace1 time-series obtained from the log. The bucket size considered is 10s, and the x-axis labels indicate the dyadic scale-factor in number of buckets. That is, a label (or scale-factor) of i corresponds to the time-scale of 10×2^i sec. It is seen that the energy scales up almost linearly till the scale-factor of 9 or 10, which corresponds to a time-scale of about $\tau_1 \approx 2$ hours, after which the behavior becomes highly nonlinear. In particular, at first there is a sharp increase in the energy, which can be attributed to the usual traffic intensity changes as a function of time of day. At the time-scales exceeding $\tau_2 = 10 \times 2^{13} \approx 1$ day, the energy shows a sharp drop, which, of course, represents the rather periodic nature of the traffic at the time-scale of 1 day. These large time-scale characteristics are bound to be observed in almost any web traffic and are of not much interest here. In particular, at these time-scales, the long-range dependence does not really play any role because the user sessions are at a much smaller time-scale, typically much less than an hour. Also, the non-stationarity of interest here is at lower time-scales than τ_1, where long-range dependence properties also exist. To confirm these observations, Fig. 5 shows the log-scale diagram for the non-stationarity traffic considered only during the busy hours (7:30-11:30am; 1:30-5:30pm). It is seen that the plot here is virtually identical to the one in Fig. 4 up to a time-scale of about 2 hours, and beyond which it is flat. This means that: firstly, long-range dependence properties (i.e., the H parameter) when high traffic periods are considered are the same as for the entire day and secondly, excluding long-time scale behavior does not affect the estimates.

From the above discussion, it is clear that the NSTs of interest here are less than 2 hours. By examining the log-scale diagrams for τ values of 7200, 3600, 1800, 900, 600 and 150 seconds respectively, we find that the H parameter estimate remains almost unchanged up to $\tau = 900\,\text{s}$ but below that, it deteriorates rapidly. In particular, Fig. 6 shows the log-scale diagram for $\tau = 1300\,\text{s}$ and yields a H value of 0.74, which is very close to the H values in Figs. 4 and 5. In contrast, at $\tau = 600\text{s}$, the H value already increases to 0.809.

Earlier, we saw that the stationarity test yielded positive results in the 1300 sec range for Trace1. The analysis in this section confirms that seasonality/trend removal at
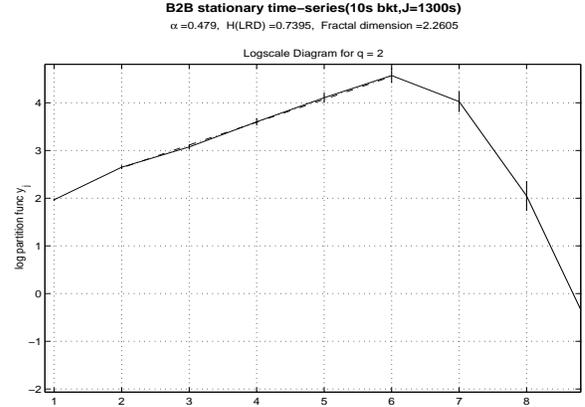

B2B stationary time–series(10s bkt,J=1300s)
$\alpha = 0.479$, H(LRD) =0.7395, Fractal dimension =2.2605

Logscale Diagram for q = 2

Fig. 6. Scaling of residual Trace1 traffic ($\tau = 1300$s)

B2C Non–stationary time–series(10s bkt)
$\alpha = 0.48039$, H(LRD) =0.74019, Fractal dimension =2.2598
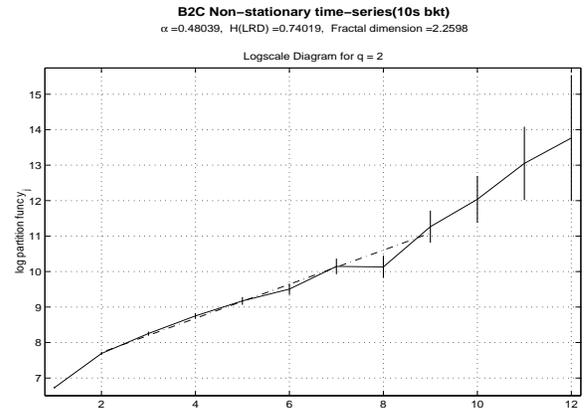
Logscale Diagram for q = 2

Fig. 7. Scaling properties of Trace2

these time scales does not seriously affect the H parameter of the traffic. Thus we can conclude that the variations that we see at these times scales are indeed a result of nonstationarity instead of simply high traffic burstiness.

We now repeat the scaling analysis for Trace2. Fig. 7 shows the scaling behavior of the original non-stationary time series. Fig. 8 shows the scaling behavior of the residual time series obtained using $\tau = 900$s. The scaling behavior is preserved for this time-scale as well as there is a good confidence interval overlap that we see from the stationarity test. As for Trace1, it is found that the log-scale diagrams for smaller time-scales show a significantly different H value, whereas higher time scales do not appear to affect the H value.

C. Characterization of non-stationarity

Having determined τ, the next step is to obtain compact characterization of the non-stationarity profile $\{Y_n\}$. We can view $\{Y_n\}$ as the finite realization of a level-shift process Y. As defined, the sequence $\{Y_n\}$ gives average number of arrivals over successive intervals, which means that the expected value of Y is the overall expected number of arrivals per interval. Let us define a new random vari-
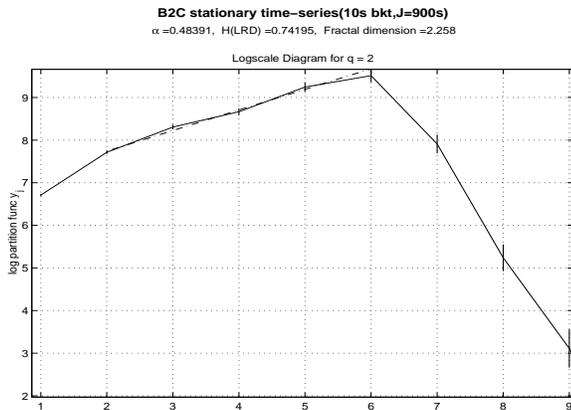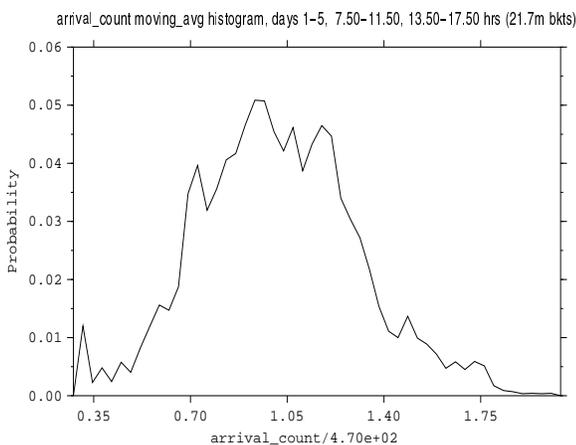
Fig. 8. Scaling of residual Trace2 traffic ($\tau = 900$s)

Fig. 10. Distribution of Z during high-traffic period for Trace2

Fig. 9. Distribution of Z during high-traffic period for Trace1
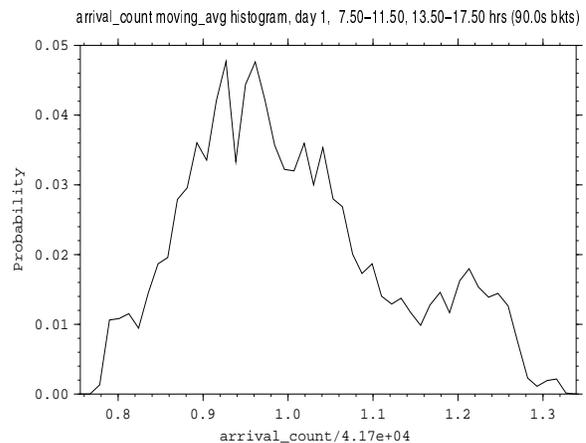
Fig. 11. Distribution of Z during high-traffic period for Trace3

able (RV) $Z = Y/E[Y]$, and the corresponding sequence $\{Z_n = Y_n/\overline{Y}\}$ where $\overline{Y}$ denotes the sample mean of the sequence $\{Y_n\}$. From a traffic characterization, the RV Z is more meaningful as it can be interpreted as the multiplicative factor to the overall arrival rate. Recall that our interest in nonstationarity is only during the high-traffic period; we are not interested in studying the normal diurnal variations. Thus, we can assume that there is no significant deterministic trend in the sequence $\{Z_n\}$. Also, there is little reason to expect a periodic or other well-defined time behavior. Thus, we shall treat $\{Z_n\}$ as a time-series of a stationary level-shift process that stays at each level for τ seconds. We further assume that the successive values taken by this level-shift process are independent. In this case, it suffices to characterize Z by its marginal distribution. Fig. 9 shows the probability distribution for the RV Z for Trace1 at $\tau = 1300$s.

Fig. 10 shows the probability distribution function of the random variable Z for Trace2 with $\tau = 900$ s. The significant difference between Fig. 10 and Fig. 9 is that Trace2 shows variation over a much larger range (0.1 to 2.5) as compared to Trace1 (0.3 to 1.9).
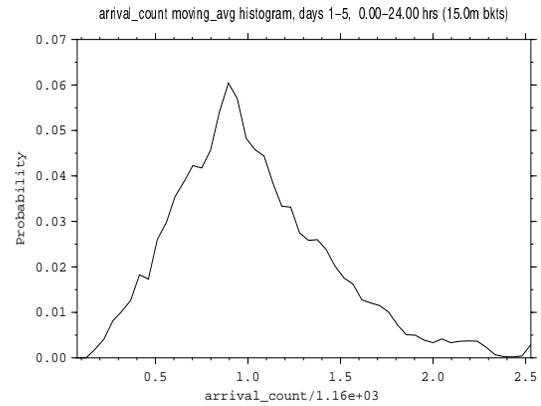
Fig. 11 does the same for Trace3. When compared

against Figs. 9 and 10, we still see a very similar triangular distribution; however, the big difference lies in the range of values around 1.0. Compared with Traces1/2, the range of the distribution is very limited, essentially $\pm 30\%$ around the mean. This again shows that Trace3 traffic is "more stable" than the other two traces.

IV. Generation of non-stationary traffic

As indicated earlier, one of the motivations for the nonstationarity characterization is to enable artificial generation of e-commerce traffic. A similar synthetic traffic generation technique for Trace3 workloads can be found in [2].

We have developed a traffic generator for this purpose [9] which can generate asymptotically self-similar traffic and optionally multifractal properties at intermediate timescales based on the $M/G/\infty$ traffic model [14], [?]. The main motivation for using the $M/G/\infty$ model is that it can generate traffic with not only long-range dependence but also with medium and short range dependence. More discussion on the use of this model may be found in [11], [?]. As such, the $M/G/\infty$ model is driven by a stationary Poisson process and thus generates stationary traffic beyond the initial warm-up period.
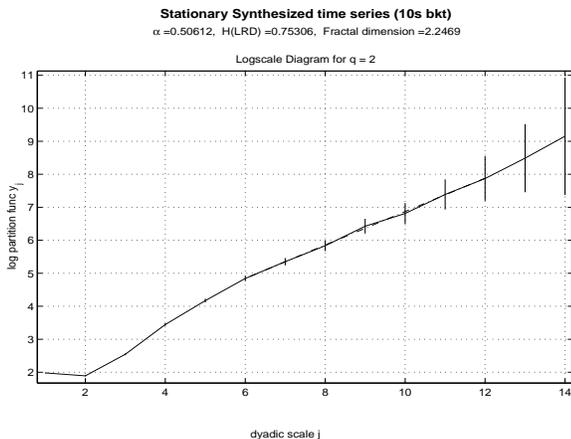
Stationary Synthesized time series (10s bkt)
α =0.50612, H(LRD) =0.75306, Fractal dimension =2.2469

Logscale Diagram for q = 2

Fig. 12. Scaling properties of synthesized stationary traffic

Non–stationary Synthesized time series (10s bkt)
α =0.55315, H(LRD) =0.76658, Fractal dimension =2.2234
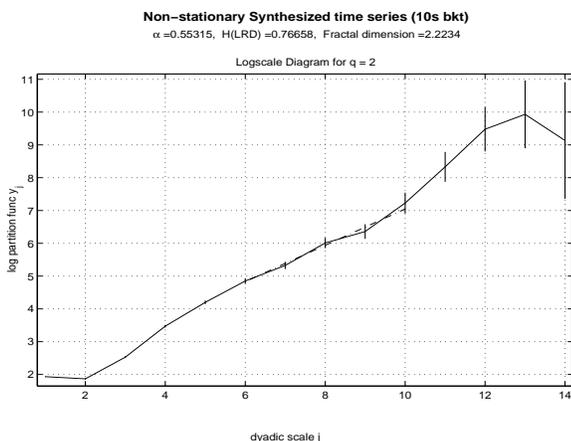
Logscale Diagram for q = 2

Fig. 13. Scaling of synthesized non-stationary traffic

Non-stationarity can be introduced in this model in many ways. A simple approach is to drive the $M/G/\infty$ system with non-stationary Poisson process. The main difficulty with this approach is the lack of simple relationship between the non-stationarity profile of the input Poisson process and that of the resulting $M/G/\infty$ process. Also, the correlational properties of such an $M/G/\infty$ system become very difficult to characterize. Consequently, we take the more direct approach of adding non-stationarity at the output end by modulating the time series with the non-stationarity profile Z. By definition, Z is positive and has a mean of 1.0. We assume that τ is an integer multiple of the slot size Δ with $\tau = J \times \Delta$. Thus a new value for Z is generated every J slots, and the number of arrivals during a slot is multiplied by this value. This approach does not affect the correlational properties of the traffic up to the time-scale of τ seconds and introduces non-stationarity precisely.

Fig. 12 shows the scaling properties of the synthesized stationary traffic, targeted for $H = 0.74799$ corresponding to Trace1 (see Fig. 4 for the scaling properties of the original nonstationary traffic). The traffic has $H = 0.75305$. Note that since the $M/G/\infty$ traffic model is asymptoti-

cally self-similar in nature, the correct scaling behavior of the traffic should be estimated from the large time-scale region, as we have done. Fig. 13 shows the scaling properties of the synthesized non-stationary traffic following our methodology. It is clear that the addition of nonstationarity brings the overall scaling behavior of synthesized traffic much closer to that of the original time-series (even though the H estimate becomes slightly worse).

V. Conclusions

In this paper, we presented a general methodology to simultaneously characterize the non-stationarity and long-range dependence in time series. A good example of such time series would be Internet traffic in general and e-commerce traffic in particular which have been shown to exhibit both long-range dependence and nonstationarity properties. Our application of the characterization methodology to a couple of e-commerce traces show that the Internet server traffic cannot always be assumed to be "reasonably stationary" and a characterization of the non-stationarity allows us to match the scaling properties of the real traffic much better than if the nonstationarity were to be ignored completely. The characterization can be exploited for performance engineering, dynamic resource allocation and overload control.

The analysis presented in this paper makes several simplifying assumptions in order to make the simultaneous characterization of long-range dependence and nonstationarity tractable. It would of interest to find ways of relaxing some of our assumptions. It would also be interesting to apply the developed methodology to a variety of other traces. The applications considered in this paper and a few others not reported here because of lack of adequate data seem to suggest that the traffic measured at the individual e-commerce servers show prominent nonstationarity characteristics and that B2C site traffic is less stable than B2B site traffic. It would be interesting to confirm/refute this; unfortunately, good traces from important sites are generally very hard to come by.

References

[1] P. Abry, P. Flandrin, M.S. Taqqu, and D. Veitch, "Wavelets for the analysis, estimation, and synthesis of scaling data", report available from `www.serc.mit.edu.au/ darryl`.

[2] P. Barford, M.E.Crovella, "Generating Representative Web Workloads for Network and Server Performance Evaluation", Proceedings of Performance '98/ ACM Sigmetrics '98, pp 151-160

[3] J.S. Bendat and A.G. Piersol, "Random data analysis and measurement procedures", 3rd edition, John Wiley, 2000, pp. 105 & 170.

[4] P. Brockwell and R. Davis, *Introduction to Time-Series and Forecasting*, Springer Verlag, 1996.

[5] M. Crovella, A. Bsetavros, "Self-similarity in the World Wide Web", IEEE/ACM Transactions on Networking, Aug 98.

[6] J. Cao, S. Cleveland, et. al., "On the nostationarity of Internet Traffic", Proc. of 2001 ACM Sigmetrics.

[7] D. Heyman, "Nonstationary Poisson process and Long-range dependence", unpublished report, 1996.

[8] R. Iyer, V. Tewari, and K. Kant, "Overload control mechanisms for web servers", Performance and QoS of Next Generation Networks, Nagoya, Japan, Nov 2000, pp 225-244

[9] K. Kant, V. Tewari and R. Iyer, "Geist: A generator of e-commerce and internet server traffic", Proc. of ISPASS, Tucson, AZ, Nov 2001, pp 49-56.

[10] K. Kant and Y. Won, "Server Capacity Planning for Web Traffic Workload", IEEE transactions on knowledge and data engineering, Oct 1999. pp731-747.

[11] K. Kant, "On Aggregate Traffic Generation with Multifractal Properties", proceedings of GLOBECOM'99, Rio de Janeiro, Brazil, pp 1179-1183.

[12] M. Krunz and A. Makowski, "A source model for VBR Video traffic based on M/G/∞ Input", Tech. Report, Univ of Maryland.

[13] W.E. Leland, M.S. Taqqu, W. Willinger and D.V. Wilson, "On the self-similar nature of ethernet traffic", IEEE/ACM trans on networking, Vol 2, No 1, pp 1-15, Fen 1994.

[14] M. Parulekar and A. Makowski, "M/G/∞ input processes: A versatile class of models for network traffic", Proc of IEEE Infocom 97, April 1997.

[15] M. Roughan and D. Veitch, "A study of the daily variation in the self-similarity of real data traffic", SERC technical report, 1998.

[16] M. Roughan and D. Veitch, "Measuring long-range dependence under changing traffic conditions", proceedings of INFOCOM'99, pp1513-1521.

Appendix

I. Self-similar Processes

Informally, self-similarity refers to the fact that the process aggregated over increasing time scales looks "similar", i.e., aggregation does not result in significant smoothing. This is primarily due to long-range dependence (LRD), i.e., significant correlations that persist over very large lags. More formally, let $\{X_i, i = 1, 2, \ldots\}$ denote a covariance stationary process with mean $\mu = E[X_i]$ and autocorrelation function $r(k) = E[(X_i - \mu)(X_{i+k} - \mu)]/\mathrm{Var}(X_i)$. Let $X_i^{(m)}$ denote the aggregated process with block size m, defined as the sample means of blocks of size $m > 1$. More precisely:

$$X_j^{(m)} = \frac{1}{m} \sum_{i=m(j-1)+1}^{mj} X_i \qquad (2)$$

Let $r^{(m)}(k)$ denote the autocorrelation function of the aggregated process $X_i^{(m)}$. The process X_i is called *exactly self-similar* if $r^{(m)}(k) = r(k)$, for all m and k, i.e., aggregation has no effect at all on the correlation structure. It is called *asymptotically self-similar* if $\lim_{m \to \infty} r^{(m)}(k) = r(k)$, i.e., self-similarity holds for large m (i.e., at large time scales).

For an exactly self-similar process, the variance of the aggregated process $X_j^{(m)}$ goes down with m as $m^{-\beta}$ with $0 < \beta < 1$ as the decay rate. Thus, the *variance-time plot*, which plots $\log(\mathrm{Var}[X_j^{(m)}])$ against $\log(m)$ would imply asymptotic self-similarly if for large m the plot is linear [13]. Self-similarity is often characterized using the *Hurst parameter H*, defined as $1 - \beta/2$. H lies in the open interval (0.5,1.0), with H=0.5 for ordinary (i.e., short range dependent) random processes and higher values meaning more bursty traffic.

Self-similarity can be attributed to user behavior, which typically shows a very bursty on-off type of behavior (i.e., high activity periods punctuated by lull periods, both of which are heavy-tailed). It is by now well established that if time-scales of interest are not limited, self-similarity leads to heavy-tailed queue-length distributions at a queuing facility and thereby significantly reduces the utilization levels for which a resource can be safely engineered.

II. Wavelet-based H Parameter Estimation

Wavelets are well suited for studying the scaling (i.e., time vs. frequency) properties of traffic because of their ability to zoom in on the desired range of time and frequency. As in previous sections, the measure of interest here is the number of arrivals during a time period. The wavelet decomposition of such a "signal" characterizes its behavior both over time and over successive time-scales (where the time-scale is synonymous with the aggregate level for the purposes of this paper). Let $\psi(t)$ denote the "mother wavelet", and $\psi_{m,k}(t)$ its scaled and dilated variants for $m = 0, 1, \ldots, K$ and $k = 0, 1, 2, \ldots$, defined as follows:

$$\psi_{m,k}(t) = \sqrt{2^{-m}} \psi(2^{-m}t - k) \qquad (3)$$

where the index m refers to behavior over time-scales of duration 2^m time-slots. Then if $X(t), t = 1, 2, 3, \ldots$ denotes the arrival process, its wavelet decomposition is given by:

$$X(t) = \sum_{\forall m} \sum_{\forall k} d_X(m, k) \psi_{m,k}(t) \qquad (4)$$

where $d_X(m, k)$'s are the coefficients of the wavelet expansion and are given by the following inner product:

$$d_X(m, k) = \sum_{\forall t} X(t) \psi_{m,k}(t) \qquad (5)$$

Given the coefficients $d_X(m, k)$'s, one can study both the time localized behavior of the traffic (by choosing a specific time of interest k) and its global properties (by aggregating over the time parameter k). In this paper, we are concerned with only the global properties of traffic. Considering $d_X(m, k)$'s for a fixed m as distribution of a RV over the time k, we can examine its qth order moment, henceforth called *partition function*:

$$E(q, m) = \frac{1}{n_m} \sum_{k=0}^{n_m-1} |d_X(m, k)|^q \qquad (6)$$

In particular $q = 2$ gives the variance of $d_X(m, .)$'s, or the energy of the signal. It can be shown that if X has stationary increments, $d_X(m, k)$'s (for a given m) are stationary, and have only short-term correlations. In contrast, traditional methods suffer due to strong correlations between $X^{(m)}$'s for different m values. This makes estimators such as $E(q, m)$ more reliable than the corresponding time-domain estimators such as $\mu^{(m)}(q)$ considered in the previous sections. It can be shown that for a self-similar process, $E(q, m)$ shows a power law. That is, for some slope $\alpha(q)$ and intercept $C'(q)$, we have:

$$\log E(q, m) = \alpha(q) \log m + C'(q) \qquad (7)$$

where the function $\alpha(q)$ is given by:

$$\alpha(q) = q(H - 1/2) \qquad (8)$$

Furthermore, for multifractal processes [11], it can be shown that equation (7) holds for small time-scales, but the function $\alpha(q)$ could be arbitrary. Thus plots of $\log E(q, m)$ and $\log m$, called as *Logscale diagrams* in [1], provide another mechanism for studying the scaling properties.