# Predicting Information Flows in Network Traffic

**Melvin J. Hinich**
*Research Faculty, Applied Research Labs, Austin, TX 78712. E-mail: hinich@mail.la.utexas.edu*

**Robert E. Molyneux**
*National Commission on Libraries and Information Science, 1110 Vermont Avenue NW, Suite 820, Washington, DC 20005-3552. E-mail: bmdyneux@nclis.gov*

**In optimizing information flows in networks, it would be useful to predict aspects of the network traffic. Yet, the notion of predicting network traffic does not appear in the relevant literature reporting analysis of network traffic. This literature is both well developed and skeptical about the value of traditional time series analysis on network data. It has consistently reported three "traffic invariants" in the analysis of network and Internet traffic. This study uses such time series analysis on a day's worth of Internet log data and finds poor support for one of the invariants. In the preliminary analysis, evidence of nonlinearity was discovered in these data and the analysis presented here examines this question further. This study posits that nonlinear events may be a traffic invariant although this hypothesis would have to be investigated further. The appearance of nonlinear structures is important to the question of predicting network traffic because there are currently no methods to predict time series with nonlinear structures. The discovery of nonlinear structures, then, may mean that developing a predictive model is impossible with current techniques. On the other hand, these nonlinearities may result from interactions from other OSI Layers than the one studied.**

## 1. Network Traffic Measurement

A recently published review of Internet measurement (Molyneux and Williams, 2001) concluded that studies of Internet traffic comprise a systematic investigation of the characteristics of that traffic. An end of such a literature is to understand the phenomena being investigated and to predict them. This survey of the Internet traffic literature concluded it lacks only attempts at forecasting the traffic it examines to be scholarly in the defined sense. Prediction is a method for testing the accuracy of the analysis of causes leading to phenomena we wish to control or alter and prediction is normally a part of scholarly literature as a discipline and test of models of processes.

It would, clearly, be useful, for example, to predict traffic bursts or information flows on networks before they arrive rather than having to adjust for them afterwards. So, in addition to being a test of the models of processes, prediction is also of practical value because it is an important matter in managing information networks.

A brief review of an important set of conclusions from that literature will serve to introduce this article and the analytic method used here. The literature describes three major "traffic invariants": "heavy tails," "self-similarity," and "long-range dependence." "Traffic invariants" are results that are reported consistently in network traffic studies. Any analysis of network traffic data can expect to see these invariants, and two of them figure in the analysis presented here.

### *1.1 Heavy Tails*

Given that the treatment here of one of these invariants, "heavy tails" is encountered more in the statistical literature than in the network traffic literature, a brief discussion of heavy tails is appropriate in order to provide a comparison of raw and treated data. "Heavy tails" refers to the fact that most traffic on the Internet consists of small connections in which little data are exchanged but from time to time, large connections with much data flow. A histogram of such a distribution would show many observations of these small connections, with low byte counts, and a few observations of very large byte counts.

Figure 1 is a histogram of the first three quartiles of one variable in a set of network traffic data that is analyzed in this study. This set of data is discussed further below in Section 4, but for now this histogram represents traffic flows sorted by the size of the connection. In this distribution, there are 512,694 transactions, the smallest of which is 0 bytes while the largest is over 64 million bytes. The third

Received February 18, 2002; revised July 15, 2002; accepted July 15, 2002

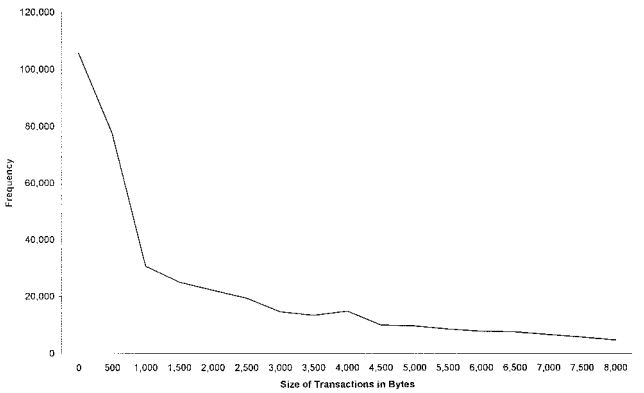


FIG. 1. Responder bytes (first three quartiles).

quartile boundary is at 8,188 bytes. In fact, this histogram is typical of a heavy-tailed distribution in that it has many small transactions and a few large transactions. The amount of data in the upper quartile of this distribution (with the small numbers of large flows) is greater than the amount of data in the three lower quartiles. These are the heavy tails. It is a problem to represent this type of distribution graphically when the highest values are so much larger than the bulk of the distribution; that is why only the first three quartiles are used here.

There are 384,522 transactions in these three quartiles for a total of 725 million bytes. In contrast, there are 326 transactions of over a million bytes, and these transactions have a total of 1.4 billion bytes.

The skewness and kurtosis of this distribution are 158 and 37,193, respectively. Skewness is a measure of how asymmetrical a distribution is about the mean value and kurtosis is a measure of how peaked a distribution is. Skewness and kurtosis of gaussian "normal" distributions are zero. Skewness is calculated by the formula:

$$\gamma = \frac{E(X - \mu)^3}{\sigma^3},$$

where $\mu = EX$ and the variance, $\sigma^2 = E(X - \mu)^2$

Kurtosis is calculated by:

$$\mathrm{K} = \frac{E(X - \mu)^4}{\sigma^4} - 3$$

The mean transaction size is 11,216 bytes, which is higher than the third quartile boundary. Variance is $4.12 \times 10^{10}$, although the variance of a raw heavy-tailed distribution is not a useful statistic. The variance and standard deviation are measures of central tendency of a distribution. The standard deviation is the square root of the variance, but the higher either number is, the more deviation from the mean value. The coefficient of variation (CV) is another measure of central tendency that is calculated by dividing the standard deviation by the mean. The purpose for making this calculation is that the larger the population values, the larger the standard deviation, so to provide a means of comparison of standard deviations from different populations, the coefficient of variation is often useful. The CV of this population is 1,811.

Heavy-tailed distributions are common, and an early problem faced by the discipline of Statistics was how to deal with them. These distributions occur in rainfall, income distributions, stock prices to name but a few areas. In the network traffic literature, the literature dealing with heavy tails seems to indicate that this phenomenon is a function of file sizes being requested. That is, the distribution of the sizes of files stored on Internet servers also shows heavy tails (Woodruff et al., 1996, Crovella et al., 1998). There are, however, no generalizations about the characteristics of the distributions of file sizes. A work on heavy-tailed distributions (Adler, et al. 1998) has discussions of these kinds of distributions from several fields including authors who have written on network traffic.

### *1.2. Long Range Dependence*

"Dependence" as used in traffic studies occurs when events in one time are correlated with events in a previous time. "Long range dependence" occurs when the correlation function does not go to zero rapidly enough as the lags increase. It is occasionally referred to as "long memory." As of this writing, there appears to be no theory to explain this relationship of traffic over time. We present evidence against long-range dependence in Section 6.

### *1.3. "Black Box" Methods*

In addition to these "traffic invariants," a belief that time series analysis cannot be applied to network traffic is a common view. Mukherjee (1994) is frequently cited but more recently in response to Resnick (1997), Willinger and Paxson (1997), and later Willinger et al. (1998) argue that "black box" methods—methods which do not look into each packet to analyze them—do not contribute to the understanding of network traffic. The contrasting methods are referred to as "structural" because they look at the individual packets to analyze network traffic. This study uses conventional time series analysis on aggregate network traffic data—that is, a "black box" method.

A time series is nothing more than data indexed by time, and the analysis of such data is common in many fields. Techniques that help explain underlying phenomena should be used no matter how they might be characterized.

### *1.4. Transforming Raw Data*

Another criticism of Resnick (1997) was made by Adler (1997) who asked why Resnick did not transform the data he worked with.

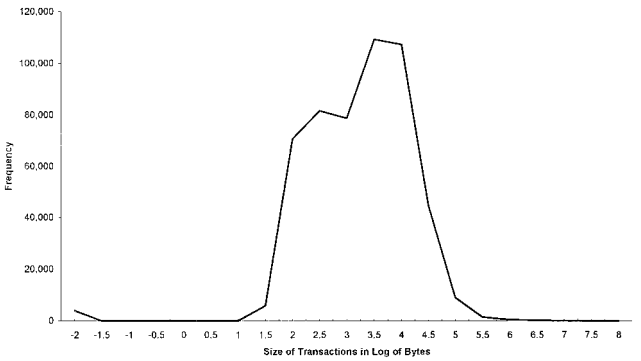


FIG. 2. Log of responder bytes (complete distribution).

Adler is aware that transforming data is a technique with a long pedigree because it can be dated back at least to Francis Galton (1879) and McAlister (1879) at the dawn of Statistics. Subsequently, Edgeworth (1898), Johnson (1949), and Aitchison and Brown (1957), among others, have contributed to our understanding of this technique for examining data. It is one method for manipulation of raw data to create a distribution of a known type, better behaved, and with many tools to analyze it.

There are distributions that cause spurious results when common population parameters are calculated—the previously mentioned variance of a heavy-tailed distribution is one example of such a distribution. Taking the logarithms of individual values in a distribution in certain cases—for example, heavy-tailed distributions—results in parameters that more accurately reflect the characteristics of the parent distribution. This study uses a logarithmic transformation of data from a Web log file as a part of its analysis.

The value of this kind of transformation is illustrated in Figure 2. This figure is a histogram of the distribution discussed in Figure 1 but here, the bytes in the transactions have been "transformed" by taking their logarithms. As a related but incidental result, the whole distribution can now be presented in a graph. (Zero) does not have a logarithm so .01 has been added to all numbers as a method to allow analysis that includes the 0 values. It is a compromise but the .01 is small and its effects on the entire distribution are not great. Note the cluster at $-2$, the log of .01—these are the 0's in the untransformed data. Note also the effect of this transformation: to pull the higher values in towards the center and create a derived distribution that is easier to analyze.

The 64 million-byte transaction, with the log transform, is about 7.8. Table 1 reports the skewness and kurtosis of the raw and transformed distributions; the variance of this distribution is .89. This number is lower than the variance reported for the raw data ($4.12 \times 10^{10}$), but a large part of that difference is the different scales of the numbers. The coefficient of variation is a number that takes the scale into account. As mentioned earlier, the coefficient of variation (CV) for the raw data is 1,811 while the CV for the log transformed data is 29, indicating considerably less spread about the mean for this new distribution.

Transforming data, therefore, is an accepted technique and one almost as old as the discipline of Statistics itself.

## 2. Time Series Analysis

Before proceeding to the examination of the data, the methods used to examine them are considered. Detecting nonlinearity in time series data has become an important area of statistical and econometric research in the last decade. A number of new methodologies have been developed to test for the presence of nonlinearity as a consequence of the increasingly widely held view that many systems are nonlinear, so testing for nonlinearity is a prudent exercise when approaching a new set of data. The original impetus for this study was to focus on examining the data for methods that might be used to predict traffic, and the consistently reported traffic invariants also offered a logical place to start.

In examining the Web log data analyzed for this study, the tests for nonlinearity found strong evidence of nonlinear structures. We were also able to replicate analysis that leads to the "invariants" but decided to focus on the topic of nonlinearity first because there is no accepted mechanism for prediction of nonlinear series. Thus our investigation of a mechanism for prediction has hit a snag. Although we only have a single day's traffic, it seems possible that nonlinearity is another traffic invariant and tests for its presence could be included in other analysis to examine this possibility. The bulk of this paper deals with this topic before returning first to one of the invariants in Section 6, then to the matter of prediction in the conclusion of this paper.

Standard time series analysis employs linear system theory. A linear system is scalable and additive. If x(t) denotes the input to a linear system and y(t) denotes its output, then a system is scalable if the output for c x(t) is c y(t) for any scale value c. A system is additive: the output for an input $x_1(t) + x_2(t)$ is $y_1(t) + y_2(t)$.

A nonlinear system is neither additive nor scalable. For example, the output from an impulse of magnitude ten can be totally different from the output from an impulse of magnitude one. The output y(t) is the observed time series in most cases, whereas the input is not observed. The value of the output depends on its past pattern, and this dependency is more complicated than the dependence owing to the correlation function $c_y(t) = E\, y(t)y(t + d)$, which only depends on the time lag d. The analysis reported here shows

TABLE 1. Comparison of the raw and transformed distributions in figures 1 and 2.

| | Raw | | | Transformed | | |
|---|---|---|---|---|---|---|
| | Skewness | Kurtosis | CV | Skewness | Kurtosis | CV |
| Responder | 158 | 37,193 | 1,811 | −1.23 | 5.63 | 29 |

signs of nonlinearity in the network traffic data examined for this study.

As mentioned, after the initial examination of the data, there was reason to believe that nonlinearity exists in network traffic and that it is important to detect nonlinear episodes in such traffic data as a key to understanding the events that drive them and to use that understanding in the process of moving from observation to prediction. As a result, this study turns to the means of detecting these kinds of events.

In order to simplify the exposition of the tests on the data, this discussion next turns to the Hinich Portmanteau Bicorrelation Test. In Section 4, the data analyzed by the Hinich test are discussed with results. Section 5 presents a selection of plots of key parameters to expose aspects of the structure of the data. Section 6 discusses long-range dependence and Section 7 concludes the paper.

## 3. Hinich Portmanteau Bicorrelation Test

Hinich (1996) suggests a modified version of the Box-Pierce (1970) portmanteau $Q$-statistic for autocorrelation and a third order portmanteau statistic, which can in a sense be viewed as a time domain analog of the bispectrum test. A full theoretical derivation of the test statistics and a number of Monte Carlo simulations to assess their size and power are given in Hinich (1996), and Hinich and Patterson (1995).

Let $x(t)$ denote a time series that is sampled at a fixed rate. As is the custom in the non-engineering time series literature, the time unit is suppressed and $t$ is an integer. In this paper, the time series will be from log files of network traffic. The observed series was broken into equal length windows and a number of statistics was calculated in each window, generating a multivariate time series of window statistics that are then used to detect events depending on the algorithm used.

Let $t_p$ denote the time of the first observation in the $p$th frame whose length is $T$. Thus the $(p + 1)th$ frame begins at $t_p + T$. The data in each frame are standardized by subtracting the sample mean of that frame and dividing by the frame's standard deviation. Let $z_p(t)$denote the standardized data in the $p$th frame.

The two test statistics we use for each frame are portmanteau test statistics. The statistic $C_p = \sum_{r=1}^{L} (\mathrm{T} - \mathrm{r})^{-1} \rho_{\mathrm{p}}^2(\mathrm{r})$, $\rho_{\mathrm{p}}(\mathrm{r}) = \sum_{t=1}^{T-r} z_p(t) z_p(t + r)$, is a slightly modification of the $Q$ test statistic for autocorrelation. If the zs are independently distributed, the distribution of this statistic is approximately chi square with $L$ degrees-of-freedom for large $T$.

The bicorrelation test statistic introduced by Hinich (1996) for detecting third order correlation in a time series is:

$$H_p = \sum_{r=2}^{L} \sum_{s=1}^{r-1} (T - s)^{-1} B_p^2(r, s)$$

$$\text{where } B_p(r) = \sum_{t=1}^{T-r} z_p(t) z_p(t + r) z_p(t + s)$$

The distribution of $H_p$ is approximately chi square with $L(L - 1)/2$ degrees of freedom for large $T$ if $L = T^c$ $(0 < c < 0.5)$. The parameter $c$ is under the choice of the user. Based on the results of Monte Carlo simulations, the use of $c = 0.4$ is recommended in order to maximize the power of the test while ensuring a valid approximation to the asymptotic theory even when $T$ is small. Simulations for the size of this test statistic presented by Hinich and Patterson show that the test is conservative for small sample sizes.

The test is of a null of pure white noise against an alternative that the process has $m$ non-zero correlations or bicorrelations in the set $0 < s < r \leq L$; i.e., that there exists second or third order dependence in the data generating process, and relies on the property of pure noise that it has zero bicovariance. The test is particularly useful in detecting nonlinear dependencies, since it has much better small-sample properties, and does not have such stiff data requirements as many of its competitors, such as the BDS test (Brock et al., (1987); see Brock, et al. (1991) for a useful survey).

Rather than reporting $C$ and $H$ as chi square variates, the T23 program written by Hinich (2002) reports the statistics as $p$-values using the appropriate chi square cumulative distribution value to transform the computed statistic to a $p$-value. The assumptions behind the test are that the observations are independently distributed and with finite moments. There is no assumption of a "normal" distribution.

## 4. The Data

The analysis presented here is based on one day's traffic on a Web server at Berkeley National Labs, from midnight (PDT) October 22, 1998 through midnight October 23, 1998. During this time, 700,893 transactions occurred. The traffic data were collected and aggregated using Bro (Paxson, 1998)—a program developed by Vern Paxson, who supplied the data.

The data are of cumulated flows of the Transmission Control Protocol (TCP) a part of the Internet suite of protocols. This suite of protocols has come to dominate networking and is often referred to as "TCP/IP" where the "IP" refers to the Internet Protocol. TCP is roughly equivalent to the Office System Interconnect Reference Model (OSI RM) Layers 4 and 5. Layer 4 is responsible for ensuring delivery of network traffic, but TCP also has Layer 5 functions in that it manages network sessions by providing for setting up, managing, and tearing down connections when sessions end. TCP connections involve a series of exchanges to set

TABLE 2. Skewness and kurtosis of the data For "SF" transactions.

| | Raw | | Transformed | |
|---|---|---|---|---|
| | Skewness | Kurtosis | Skewness | Kurtosis |
| Originator | 428 | 256,844 | .56 | −1.67 |
| Responder | 255 | 100,580 | .64 | −1.46 |

up connections and to end them. During a session, TCP, in effect, ensures that all the packets are received and requests ones that are missing.

TCP involves two-way communications between the originator of the request and the responder.

Originator and responder bytes were summed in 10ths of a second slices for the whole day on all the transactions reported in the server log and for those established transactions with "FIN" handshake for completion (referred to by Bro as "SF" transactions). FIN is the TCP method for ending a session and we might regard these sessions as complete sessions. Bro also tracks sessions that failed to initiate or that did not end with the FIN, sessions that were rejected (REJ), and other anomalies.

There are 512,694 SF transactions in the dataset as we have seen in Figures 1 and 2 and the related discussion. Calculations below are performed both on SF transactions and "All" transactions; that is, all transactions, even if incomplete or the result of failed connections, in order to analyze two major ways of looking at the data. Other kinds of transactions reported by Bro were examined but are not separately treated here.

As has been indicated, the examination of these server log data for nonlinearity involves a logarithmic transformation of the data. Before turning to the analysis itself, logarithmic transforms and their effect on these specific data are examined. The results are similar to those observed above.

### *4.1. Logarithmic Transformation of the Data and Time Slices*

In addition to transforming the data, the SF transactions for the day were divided into slices of one tenth of a second. The data (originator and responder bytes) were then summed in each of these slices and next transformed with base 10 logarithms. Slices with no traffic have 0 bytes. In order to illustrate the effect of performing these transformations, skewness and kurtosis for the raw and transformed data are compared in Table 2. As we saw in Figure 2, in order to calculate a logarithm for 0, .01 was added to all 864,000 numbers and then the logarithm of the new values calculated.

"SF" refers to established transactions with normal FIN handshake for completion. Basic calculations were done with SAS© (SAS, 2002). Univariate procedure on raw and logarithmic transformed data summed by tenths of a second for the day. The SAS calculation of the Kolmogorov Smirnov D for the logarithmic transformed data for both the originator and responder bytes at .409 and .405 respectively, the probability that either of these distributions is "normal" is less than .01 based on the Kolmogorov Smirnov test.

As expected from long experience, logarithmic transforms on these data provide a means of simplifying the analysis of this distribution. As mentioned above, these results were presaged by Adler (1997) and also in the previously cited work by Woodruff et al (1996). In the latter article there are no generalizations about the distributions of file sizes but the fourth graph in Figure 1 (logarithmic scale of file sizes) looks suspiciously as if the file sizes she found in her sample are more nearly lognormal than the raw data. Figure 2 in this article shows a similar result.

In any case, the purpose of this analysis is not to examine the data to see if they are lognormal—they clearly are not—but to deal with the heavy tails and symmetrize the data. As mentioned above, this procedure is a well recognized one and suitable to the problem at hand, which is to examine a set of data by appropriate methods and without imposing structures on them, that may not be there.

### *4.2. Tests for Nonlinearity*

Table 2 reports an analysis of the transformed data for originator and responder bytes for SF transactions and "All." The Hinich Portmanteau Bicorrelation was conducted on these four sets of data with two separate sets of significance thresholds of .005 and .0001. The implications of testing at these levels is that at the first, .5% of the time the test would conclude a window's value was nonlinear when it was not. If there were no nonlinear values, then, the test would indicate nonlinearity for about 43 (43.2) of the 8,640 windows and we would infer that the tests for nonlinearity failed. In the second threshold, .0001, there would be about 1 (.86) windows indicated as nonlinear by the test and one could infer that the test for nonlinearity failed. If the numbers of windows is greater than these numbers at the respective thresholds, the tests would indicate nonlinear events.

Non-overlapping moving windows of 100 consecutive transformed data values were created and computed correlations and bicorrelations and other statistics were calculated for each window. Windows are used to isolate the bursts for examination and those used in this analysis are 10 seconds long. There are 8,640 windows.

Hinich's bicorrelation test was used to determine if there was episodic nonlinearity in the data. In Table 3, the results are reported for the residuals of an autoregressive (AR) fit with a lag of 5 to each window.

The null hypothesis of this test is that the windows are "white"; that is, the numbers in the windows are serially independent. The initial run through the data tested these 10-second windows for correlation and found that at the .0001 level, 1.64% (142) of the originator windows had serial correlation, and of the responder windows 1.48% (128) had serial correlation, thus rejecting the null. An AR fit of 5 reduced the number of windows with significant

TABLE 3. Number of significant windows (and percent) at 0.01% and 0.5% thresholds AR = 5.

| | .01% | | .5% | |
|---|---|---|---|---|
| | SF | All | SF | All |
| Originator | 59 (0.68%) | 36 (0.42%) | 168 (1.94%) | 143 (1.66%) |
| Responder | 78 (0.90%) | 45 (0.52%) | 215 (2.49%) | 161 (1.86%) |

Non-overlapped windows of length 100 with the number of windows = 8,640. Test statistics calculated using T23[20].

correlation to zero, thereby removing the correlation. Thus, the H—the test statistic discussed in Section 3—reported in Table 3 is calculated with the AR(5) on each window to remove the correlations, leaving just the bicorrelations—sometimes called "triple correlations." The reader will recall that the H is reported by T23 as a *p* value.

As Table 3 demonstrates, the Hinich test rejected the null hypothesis of linearity well above the preset false alarm rates that were used to develop the test thresholds for all sets of data. For example, the expected value for the test of nonlinearity on SF originator bytes at .01% would be .86 in a sample of this size (n = 8,640), while the result in these data is 59, allowing clear rejection of the white noise hypothesis—and similarly through all values of all variables reported in the table. Thus, the analysis reported gives clear indication of nonlinearity in these data.

## 5. Plots

The plots here are chosen to illustrate structures in the data. Rather than present all plots generated, only plots for SF responder transactions at AR(5) data are provided as a sample. The AR(5) fit removes correlations within windows and thereby clarifies the non-linear structures in the data. It also removes the correlations expected if the data were to show long-range dependence. We take up the discussion of removing the correlations again in Section 6.

Figures 3 through 5 use the same x axis so it is useful to discuss these axes briefly. The data are in order by time and there are 10-second windows, or 360 windows per hour. These x axes show ticks at each 360th window—or one hour—interval for each of the 24 hours for this day.

Figure 3 shows the H statistic for each window indicates nonlinear bursts in the traffic. If only Hs above .99 probability are examined, there is a pattern indicating that nonlinearities in this sample traffic are more common from midnight to 6 a.m. (windows 1–2161) than they are from 9

TABLE 4. Coefficients and adjusted R-square for AR = 5 model. AR(5) parameters/t values.

| *p* value threshold for the iterative AR | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1- | 0.13 | 2- | 0.11 | 3- | 0.11 | 4- | 0.12 | 5- | 0.11 |
| | 124.58 | | 106.13 | | 105.97 | | 109.30 | | 102.05 |
| Adjusted R-square: 0.132 | | | | | | | | | |

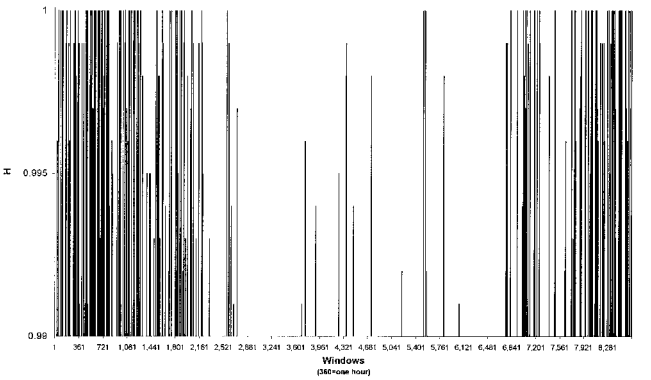


FIG. 3. H Statistic of SF Transactions.

a.m. (3241–6481) to 6 p.m. but overall, the bursts are episodic throughout the day.

T23 calculates a standard deviation for each of the windows, and then is scaled by dividing that figure by the sample standard deviation. Figure 4 is a plot of the window standard deviations for the day. The standard deviations vary more than the standard error, indicating the nonstationary nature of these data.

Figure 5 plots means of the windows, scaled by dividing these window means by the mean of the sample. Compared with the standard error of the estimates of the means for each window, the means are not significantly different from 0. The mean for all SF windows = −.301. Recall that in order to calculate a logarithm for 0, .01 was added to each number. Given that over half the windows have zero bytes, this mean figure is lower than it would be without the correction necessary for calculating the logarithms. This plot indicates further evidence of nonstationarity in these data.

Figure 6 is a plot of the events in three consecutive windows (33–35) of SF originator bytes as analyzed here. Two, 33 and 34 have H statistics that are significant, while the H statistic for 35 is not significant.

There are 100 10th-of-a-second slices in each of these windows so this plot is of 30 seconds of traffic. The points at −2 indicate no traffic with bursts on the logarithmic scale

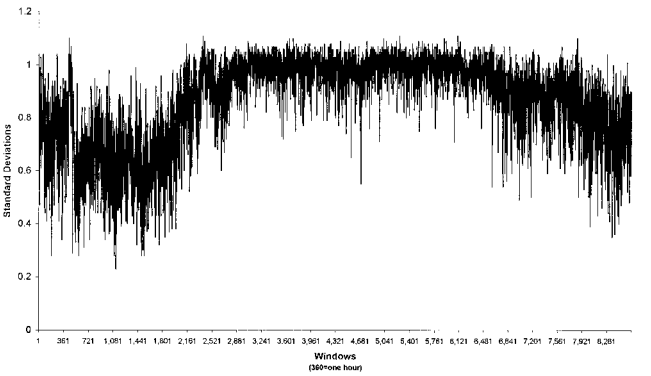


FIG. 4. Standard deviation of SF transactions.

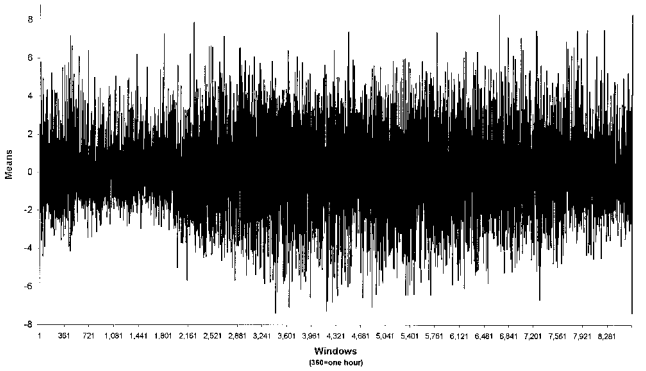


FIG. 5. Mean of log SF transactions.

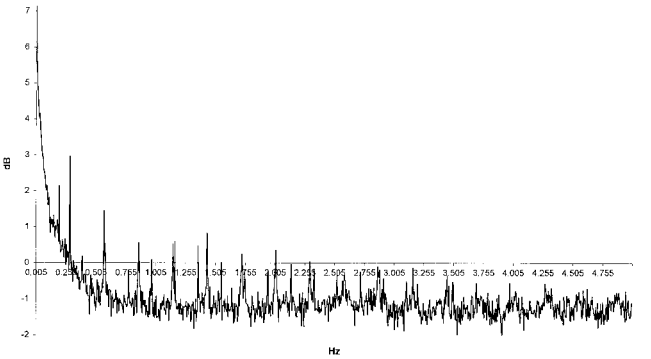


FIG. 7. Spectrum of log of mean originator bytes, REJ = No, standard error = 0.21.

at 3 indicating that traffic around 1,000 bytes was reported by Bro during this slice. This plot indicates that it is not the size of the bursts but the pattern that mark the traffic in the first two windows as nonlinear.

## 6. Long Range Dependence

If there is long-range dependence, the values of the correlations between windows do not go to zero over long periods of time. This conclusion has been observed in the literature but a cause for this phenomenon has apparently not been cited.

The power spectrum is the Fourier transform of the correlation function. If a process has long-range dependence, then the power spectrum will go to infinity as the frequency goes to zero. Another implication is that a low order autoregressive (AR) model will not remove the low-frequency bulge in the spectrum.

The statistical theory and methods for fitting time series with AR models and for spectral analysis can be found in Anderson (1971), Box and Jenkins (1970), Fuller (1976), Hamilton (1994), and Priestley (1981).

We use a standard spectral analysis method to compute the sample spectrum and use the AR fit on the entire distribution.

Figure 7 is the spectrum of the mean for each window of log transformed Originator bytes for all transactions. The horizontal axes of both spectra is the frequency in Hertz.

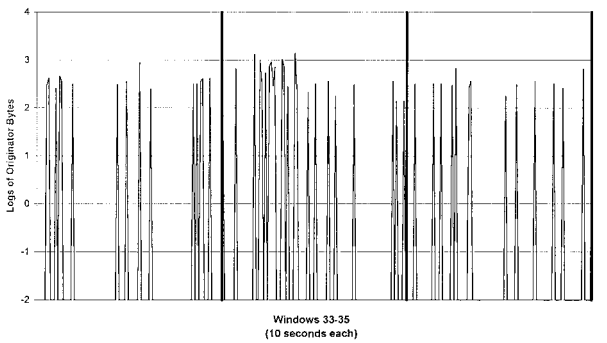


FIG. 6. Bursts in three consecutive windows

Figure 8 is the spectrum of residuals of the AR(5) fit. If there were long-range dependence, the spectra would have a much larger peak at the lowest frequency band.

There is no evidence for long-range dependence.

## 7. Conclusions

The data exhibit heavy tails and other characteristics reported in the network traffic literature. What differs here from standard network analysis is the statistical treatment and the focus on prediction. The logarithmic transformation is an effective treatment to use in analyzing this type of distribution. Analyzing the data using standard statistical techniques finds no evidence for long-range dependence in this set of data.

Evidence of nonstationarity exists in the mean and variances. In addition, there are episodic nonlinear events in these network traffic data as recorded in the Web log file of TCP traffic flows by Bro. If the results of the analysis of other traffic data using spectral analysis confirm the results reported here, episodic nonlinearity may be recognized as a traffic invariant, but further testing of other data will be necessary.

Among the questions that have to be examined are to what extent are the phenomena observed in TCP traffic a

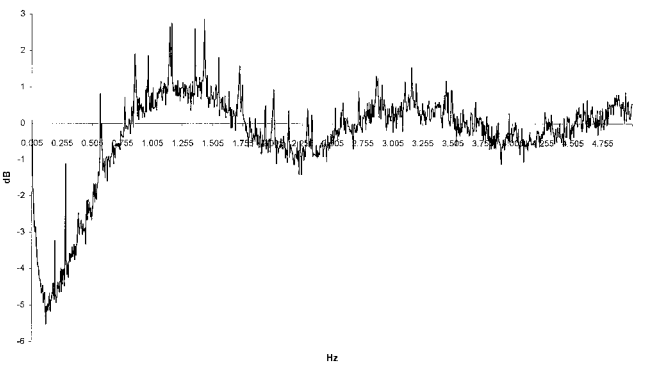


FIG. 8. Spectrum of residuals of an AR(5) to log of mean originator bytes, REJ = No, standard error = 0.21.

result of aggregation of behavior of other OSI Layers? Aggregation occurs in various devices on the Internet, for instance, to buffer traffic; that is, hold it while waiting for other traffic. Could such aggregation create the nonlinearities, leaving open the possibility that network traffic can be predicted by currently available techniques once the aggregation were corrected for?

In order to understand this question, analysis of traffic from other OSI Layers is necessary, particularly the Data Link Layer (Layer 2) and the Network Layer (Layer 3). This analysis may yield insights into the various patterns observed in TCP (roughly, Layers 4 and 5) traffic.

In addition to potential artifacts introduced by lower OSI Layers, Bro aggregates data. For instance, the 64 million-byte transaction is recorded as occurring at one specific time when it likely took place over a period of time. For the purpose for which Bro was written, this treatment of the data is appropriate but it might—might—not be appropriate if the question is predicting network traffic.

Therefore, in the quest for techniques to predict traffic on information networks, data from other OSI Layers and from other networks is necessary before the structure of such a conclusion could be drawn.

The evidence of nonlinearity if confirmed in other studies may mean that prediction of network traffic cannot be accomplished with current techniques. Then methods to approximate such predictions will need to be developed.

## Acknowledgments

The authors thank Vern Paxson who graciously supplied the data used in this analysis and patiently answered questions. We also thank one unnamed reviewer who made a number of helpful observations.

## References

Adler, R. (1997). "Discussion." Annals of Statistics, 25, p 1850.

Adler, R., Feldman, R., & Taqqu, M., Editors. (1998). A practical guide to heavy tails: statistical techniques and applications. Boston: Birkhauser.

Aitchison, J., & Brown, J.A.C. (1957). The lognormal distribution. Cambridge: Cambridge University Press.

Anderson, T.W. (1971). The statistical analysis of time series. New York: John Wiley & Sons, Inc.

Box, G.E.P., & Jenkins, G.M. (1970). Time series analysis: forecasting and control. San Francisco: Holden-Day.

Box, G.E.P., & Pierce, D.A. (1970). Distributions of residual autocorrelations in autoregressive integrated moving average models. Journal of the American Statistical Association, 65, 1509–1526.

Brock, W., Dechert, W., & Scheinkman, J. (1987). A test for independence based on the correlation dimension. Mimeograph. Department of Economics, University of Wisconsin at Madison.

Brock, W., Hscih, D., & LeBaron, B. (1991). Nonlinear dynamics, chaos, and instability: statistical theory and economic evidence. Reading, Massachusetts: M.I.T. Press.

Crovella, M., Taqqu, M., & Bestavros, A. 1998. Heavy-tailed probability distributions in the world wide web. In R. Adler, R. Feldman, & M. Taqqu (Editors). A practical guide to heavy tails: statistical techniques and applications (pp. 3–25). Boston: Birkhauser.

Edgeworth, F.Y. (1898). On the representation of statistics by mathematical formulae. Journal of the Royal Statistical Society, 61, 670–700.

Fuller, W.A. (1976). Introduction to statistical time series. New York: John Wiley & Sons, Inc.

Galton, F. (1879). The geometric mean in vital and social statistics. Proceedings of the Royal Society of London, 29, 365–367.

Gnedenko, B.V., & Kolmogorov, A. N. (1954). Limit distributions for sums of independent variables, translation by K.L. Chung. pp. 162–190. Cambridge, Massachusetts: Addison-Wesley.

Hamilton, J.D. (1994). Time series analysis. Princeton, New Jersey: Princeton University Press.

Hinich, M. (1996). Testing for dependence in the input to a linear time series model. Journal of Nonparametric Statistics, 6, 205–221.

Hinich, M. & Patterson, D. (1995). Detecting epochs of transient dependence in white noise. Mimeograph. University of Texas at Austin.

Hinich (2002). T23 is written in Fortran and is available at: http://www.molyneux.com/t23/ compiled for Windows95/98/NT and in source code. Documentation for the program is found in the source code.

Johnson, N.L. (1949). Systems of frequency curves generated by means of translation. Biometrika, 36, 149–176.

McAlister, D. (1879). The law of the geometric mean. Proceedings of the Royal Society of London, 29, 367–376.

Molyneux, R.E., & Williams, R.V. (2001). Internet measurement. Annual Review of Information Science and Statistics, 34, 287–339.

Mukherjee, A. (1994). On the dynamics and significance of low frequency components of internet load. Internetworking: Research and Experience, 5, 163–205.

Paxson, V. (1998). Bro: a System for detecting network intruders in real-time, Revised January 14, 1998. ftp://ftp.ee.lbl.gov/papers/brousenix98-revised.ps.Z

Priestley, M.B. (1981). Spectral analysis and time series. London: Academic Press.

Resnick, S. (1997). Heavy tail modeling and teletraffic data. Annals of Statistics, 25, 1805–1869.

SAS (2002). The statistical analysis system, SAS Institute, Cary, North Carolina. http://www.sas.com/.

Willinger, W., & Paxson, V. (1997). "Discussion." Annals of Statistics, 25, 1856–1866.

Willinger, W., Paxson, V., & Taqqu, M. (1998). Self-similarity and heavy tails: structural modeling of network traffic. In R. Adler, R. Feldman, & M. Taqqu, Editors. A practical guide to heavy tails: statistical techniques and applications. Boston: Birkhauser.

Woodruff, A., Aoki, P., Brewer, E., Gauthier, P., & Rowe, L. (1996). An investigation of documents from the world wide web. Fifth International World Wide Web Conference, May 6–10, 1996, Paris, France. http://www5conf.inria.fr/fich_html/papers/P7/Overview.html