Analysis of the Hurst Exponent for M2M Applications

[1,]Stephen P. Emmons, [2,]Farhad Kamangar

The University of Texas at Arlington

ABSTRACT:

Machine-to-Machine (M2M) applications represent a class of systems that make significant demands on the global Internet infrastructure. It is important to understand and characterize the behavior of these systems so that they may be effectively engineered for scalability and reliability. This paper analyzes the "Hurst exponent" (H) for several M2M applications of varying sizes with data collected for up to six months. The Hurst exponent is a measure of the long-term dependence and self-similarity typical of Internet-based communications systems. The analysis focuses specifically on the use of the "Rescaled-Range Statistic" (R/S), and also includes several other methods for comparison, including "Absolute Moment," "Aggregate Variance," "Higuchi," and "Residuals of Regression." The results show that such applications indeed exhibit strong long-term dependence, consistent with results found for other Internet communications studies.

KEYWORDS: *GSM radio communications, Hurst exponent, long-range dependence, M2M applications, R/S statistic, self similarity, vehicle tracking systems*

I.   INTRODUCTION

The Internet is commonly thought to be the medium for people communicating with websites, sharing files, streaming audio and video, and other uniquely human activities. However, there are other users of the Internet, making demands of its resources. Kevin Ashton, co-founder of the Auto-ID Center at MIT, is credited with first using the phrase "The Internet of Things." He described "things" such as embedded devices using the same shared infrastructure as humans, noting that computer processors are found in home appliances, all types of vehicles, and many other places. The number of these devices is larger than the number of people using the Internet. Systems with such devices communicating on the Internet are often called "Machine-to-Machine" (M2M) applications.It is critical to study M2M applications in order to understand their behavior, and in particular, their impact on the infrastructure of the Internet. One such class of systems is vehicle-tracking applications. These applications require both "wired" Internet resources, as well as the wide-area wireless communications infrastructure operated by cellular carriers.Today,"smartphones" are common in the popular consciousness, but less visible are vehicles and other mobile assets (e.g., pallets, rail cars, etc.) that use basically the same technology for capturing local information and reporting it to aggregation points for analysis. These systems must be engineered for scalability and reliability, taking into consideration their future growth and demand on communications infrastructure.Key to understanding the communications requirements of M2M applications is an analysis of the message traffic generated by embedded devices. Typically, their behavior may be defined in probabilistic terms using one of many models describing the frequency of message arrival, potential for collisions, and duration of "bursts" of messages [1]. Such models have various parameters that map their theoretical characteristics to specific operational conditions. It is important to choose the correct model, as well as the right values for the parameters that govern it. In doing so, one attribute of the message traffic with important implications for the system's behavior is the "Hurst exponent" [2].This paper analyzes the long-term dependence and self-similarity characteristics of four commercial vehicle tracking applications by using the Hurst exponent. Each application has from several hundred to several thousand vehicles, and the dataset for each of them includes message arrival counts collected over up to six months. Five different methods are used and compared in this analysis.The organization of the remaining sections of this paper is as follows:
Section 2 provides background on the Hurst exponent, briefly describing its development and use.

Then Section 3 reviews a cross-section of related works that elaborate on or apply the Hurst exponent in contexts similar to M2M applications.A description of the vehicle-tracking applications studied is presented in Section 4, followed by the results of the Hurst analysis of these systems in Section 5 using several estimation methods.And finally, Section 6 summarizes the conclusions of the analysis and outlines future work.

II.   BACKGROUND

The Hurst exponent represents an estimated measure of "long-range dependence" (LRD) or "self-similarity" of time-series data. It is named for Harold Edwin Hurst who first introduced the concept in the field of hydrology [2]. In the 1950s, he introduced the "Rescaled Range Statistic" (R/S) as a measure of the variability of a given time-series by finding the rescaled range (R) of values and dividing it by the standard deviation (S).The rescaled range for a time-series $X = \{x_1 \ldots x_n\}$ is defined as the difference between maximum and minimum of the mean-adjusted range series for a given sub-sequence of X given a length t = 1...n and an

$$\bar{X}_t^i = \frac{1}{t} \sum_{k=i}^{i+t} x_k \qquad (1)$$

$$Y_t^i = \sum_{k=i}^{i+t} (x_k - \bar{X}_t^i) \qquad (2)$$

$$R_t^i = max(Y_t^i) - min(Y_t^i) \qquad (3)$$

offset i = 0 . . . (n − t):

Thus $(R/S)_t^i$ for a sub-sequence of X is defined as follows:

$$S_t^i = \sqrt{\sum_{k=i}^{i+t} (x_k - \bar{X}_t^i)^2} \qquad (4)$$

$$(R/S)_t^i = \frac{R_t^i}{S_t^i} \qquad (5)$$

Note that when t = n, then I = {0} and $(R/S)_t^0 = (R/S)_n$. Furthermore, it is convenient to consider only the expectation of $(R/S)_t^i$ as follows:

$$(R/S)_t = E[(R/S)_t^i] = \frac{1}{t} \sum_{k=i}^{i+t} (R/S)_t^i \qquad (6)$$

Hurst found that the expectation of (R/S) as the time scale increases approaches the product of some constant C and n to the power of another constant H as follows:

$$E[(R/S)_n] = Cn^H \qquad (7)$$

In this form, the exponent H can range between 0 and 1. Values of H such that 0.5 < H < 1 result from time series values having a tendency to continue in the same direction as other recent values; thus exhibiting "long-term memory" in its behavior. Values of H such that 0 < H < 0.5 are found for time series that are strongly "mean regressive." In between, a value of 0.5 suggests uncorrelated, random values with no tendency to increase of decrease based on prior history.H can be estimated as the slope of the least-squares fit of the expectation of (R/S) for exponentially larger values of t up to n plotted on the log/log scale. Figure 1 illustrates the boundaries of this range with several reference examples. For each example having 2^{16} samples, the graph shows values for the $(R/S)_n$ for values of n = 2^i where i = 1...16, plotting the results on a log/log scale:

- A simple monotonically-increasing number sequence from 1 to 2^{16} is fully self-similar at all scale levels, and as expected, shows a slope of 1.
- 2^{16} random samples shows a slope 0.5.
- A set of 2^{16} samples toggling between 1 and -1, thus always regressing toward a mean of 0, likewise shows a slope of 0.
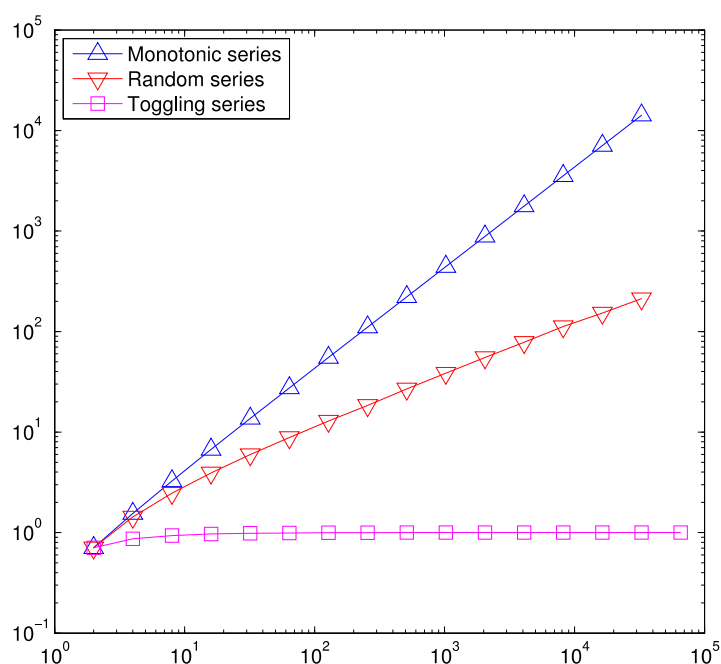
Figure 1: Three reference examples of (R/S) values on a log/log scale
whose slopes correspond to H values of 1, 0.5, and 0

Note that as an estimate of H, the slope of the log/log plot of (R/S) values can be highly influenced by the choices of t used. Figure 1 provides a hint of this fact, since the first few (R/S) values in the examples for slopes 0.5 and 0 trend slightly differently than the rest. Section 5 will elaborate on this. Starting in the late 1960s, Mandelbrot related the Hurst exponent to the "roughness" of a time series [3][4]. In his work with fractals, which are known to exhibit high self-similarity at many (or all) scales, he found that higher H values occur with "smoother" shapes, whereas lower H values correspond to shapes that are more "wild" or "rough."In the early days of the Internet, Leland et. al used an analysis of the Hurst parameter applied to Ethernet data communications [5]. They found that the "burstiness" of such data traffic appeared to occur equally at small and large time scales, similar to Mandelbrot's fractals. This discovery was significant because the LRD indicative of such behavior is not consistent with a "memoryless" Poisson process that had been used by the telecommunications industry in the past to model the occurrence of telephone calls for capacity planning [5]. In the early 1990s, the Poisson model had been used by many in industry and government around the world when forecasting long-term Internet growth planning. In a special joint publication by both the IEEE and ACM, they successfully brought attention to the greater capacity requirements that such self-similar communications would demand [5].

III. RELATED WORK

Many methods have been studied for calculating the Hurst exponent using alternatives to (R/S).

As a contributor to the aforementioned study led by Leland, Taqqu, along with Levy, had previously investigated several Hurst estimators, including aggregated variance, periodigrams, various Whittle methods, and even simple visual observation [6].Later, Taqqu et. al analyzed additional methods [7] such as the eponymous Higuchi estimator [8] and "Residuals of Regression" by Peng [9].More recently, Jones and Shen developed a method of estimating H by looking at level crossings for faster calculation [10].Overall, the various available methods exhibit strengths and weaknesses based on the type of data being considered. While (R/S) is more computationally expensive than many others with a time-complexity of $O(n^2)$, it is generally recognized as the standard.Since Leland et. al, there have been many studies of the LRD qualities of data communications involving the Hurst exponent.Nash and Ragsdale used an analysis of the Hurst exponent for critical systems as part of their characterization of systems communications as a precursor to establishing criteria for an automated intrusion detection method [11].Idris et. al used a Hurst exponent estimate as a reference against which to detect anomalies in a system [12]. Their method focused on determining an effective window size of recent traffic activity, calculating a current H and comparing it to an expected value.Park et. al explored the LRD qualities of

several TCP/IP data sets representing significantly higher data rates than those considered earlier by Leland et. al [13]. They used a variety of methods to estimate the Hurst exponent and confirmed high values (i.e., between 0.8 and 0.9) for such data.Clegg used the Hurst exponent to help establish the parameters of a Markov-modulated process for simulating self-similar network traffic, overcoming the kinds of limitations found with past models as identified by Leland et. Al [1].Dobrescu et. al examined how the self-similarity of various networks, such as LANs and WANs, are affected by their topology, exploring multiple methods for measuring H, as well as related statistical measures [14].

IV. TARGET SYSTEMS

The datasets analyzed in this paper come from four commercial vehicle-tracking applications used by

Table 1: Dataset attributes

Dataset	Vehicles	Weeks
A	175	26
B	285	26
C	650	13
D	1000-3000	26

both business and consumer users.

Table 1 identifies each application by a letter "A" through "D," along with the number of vehicles it had, and the total time span of the message count data collected. Each dataset contains a progressively larger number of device-attached vehicles with data collected for at least over three months, and in most cases over six months. The largest dataset, "D," saw an increase in the number of devices by a factor of three over the six month period.For the purpose of this study, the primary focus concerns two key elements of these applications: First, there is an embedded processor and wireless GSM radio, or "device." Messages are sent as single UDP/IP packets over GPRS to report both periodic and exception-based information about the condition of an attached vehicle.The devices in the study come from several manufacturers, including Enfora (now Novatel), Xirgo, and Calamp. The duty cycle of the devices can vary across applications, but the typical configuration involves an hourly report of location when the attached vehicle is powered off and a report of location every minute when it is on.In addition, the devices may also report their status when the vehicle is powered on or off, crosses a predefined speed threshold, enters or leaves a predefined "geofence," or detects a state change for an attached digital and/or analog sensor.Second, there is an aggregation service known to the devices to which they send their various messages. This service receives, and in some cases acknowledges, messages from the different device types, each having its own data format, and normalizes/conditions the information contained in the message for subsequent analytical processes.Each application has a similar service configuration, but typically, is comprised of one or more virtual servers hosted in a "private cloud" in sufficient number and with sufficient allocation of resources (e.g., CPU, RAM, SAN) to handle the operational demands of the overall system.
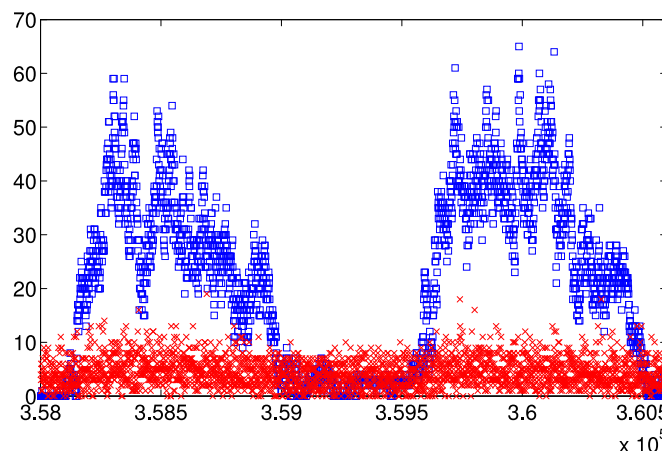
Figure 2: Message counts per minute categorized by ignition on vs. off
for a 2-day period from dataset "A"

Figure 2 shows a subset of data from dataset "A" having 175 devices representing a fleet of vehicles for a construction company. The horizontal axis represents a two-day period with one-minute resolution, and the vertical axis shows the aggregated number of messages received in one minute. The red 'x's represent the counts per minute from devices with their ignition off, whereas the blue squares represent message counts from devices with their ignition on. The larger magnitude of "ignition on" counts reflects the higher reporting rate for this condition. A clear diurnal cycle is evident given that the vehicles are active with their ignition on during the course of their business day, with little or no activity during the night-time hours.For dataset "A" the ratio of messages with ignition on to those with ignition off was approximately 3 to 1. Histograms of the distribution of these types of messages in Figure 3 depict a strong half-normal frequency distribution, with the "ignition on" messages exhibiting a Gaussian mixture model. The high values on the left side of the graph for devices with ignition on indicates that a small percentage of the population of devices are on at the same time.
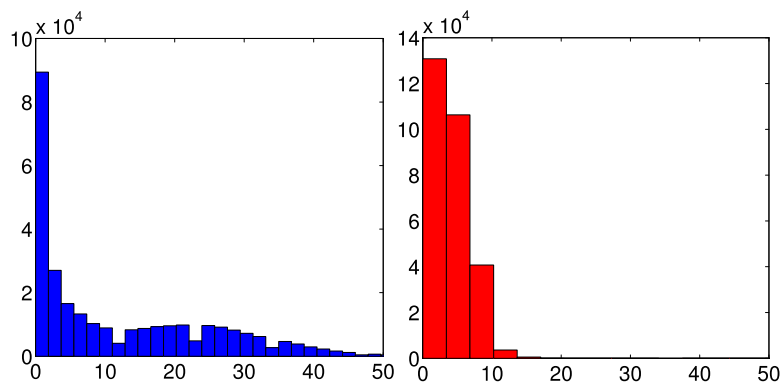

Figure 3: Distribution of message counts by minute for 175 vehicles with ignition on and off

V.   RESULTS

For the purpose of this study, values of (R/S) were calculated for the per-minute message counts of each application up to six months at intervals for powers of 2 from 2^1 to 2^{18} minutes in duration.Since Figure 2 showed visually different message arrival characteristics for "ignition on" vs. "ignition off," the possibility existed that the estimate of H might be different for each type of count history. One might imagine a possible M2M application that only reports values for vehicles with the ignition on, or another that only reports hourly, along with some exceptions representing locally-detected changes in state.
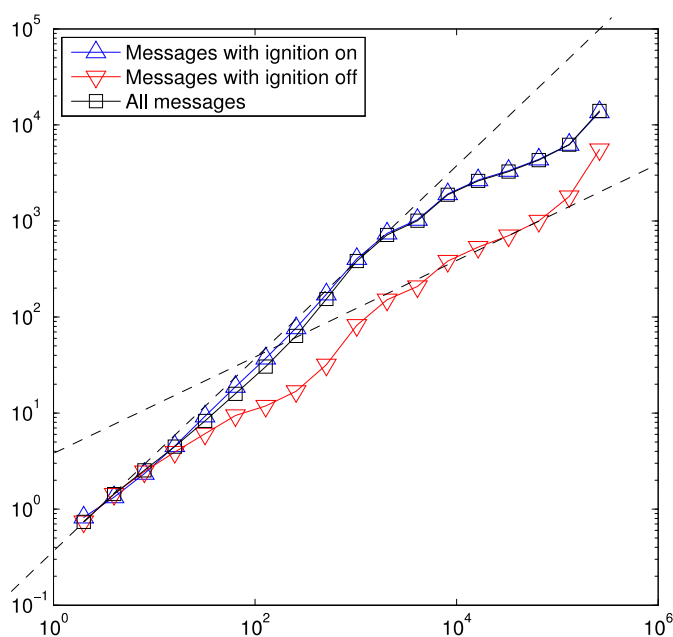

Figure 4: Changing slope for system "A" over differing scales for message counts
with ignition on, off, and combined

Figure 4 shows three plots for (R/S) values representing only "ignition on," "ignition off," and all combined message counts for dataset "A." Bracketing dotted reference lines with slopes 0.5 and 1 show that, at different time scales, the (R/S) values trend back and forth between these boundaries of the LRD/self-similar range for H.
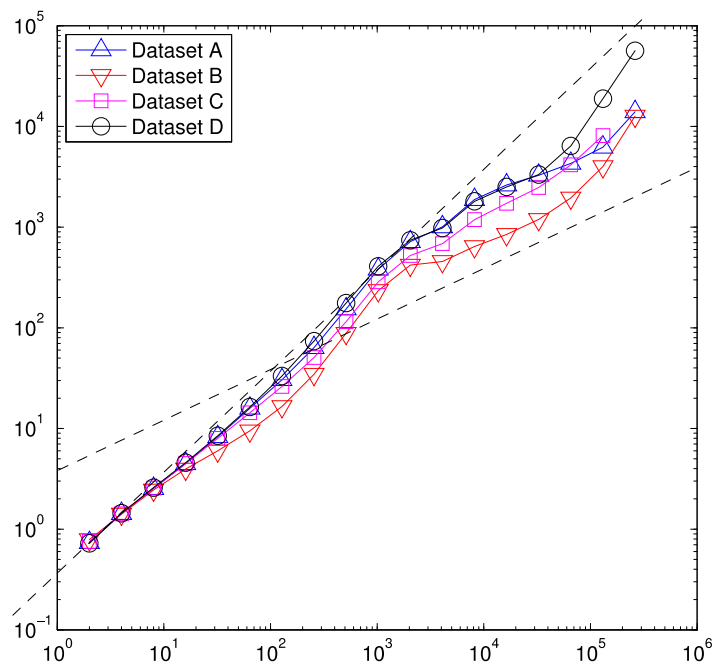

Figure 5: Log/log plot of averaged (R/S)$_n$ values at many time scales for each target system

The same characteristic change in slope is also visible for each of the other datasets "B" through "D" as shown in Figure 5. Here the data shows strong LRD with slope close to 1 for time scales 2^1 to 2^{11}, at which point the slope drops closer to 0.5. Then at time scale 2^{15}, the slope again increases back closer to 1. The time scales of lower slope correspond to the daily and weekly periodicity of the data that are necessarily "mean regressive," and as such would be expected show a decrease in the value of H accordingly.

Table 2: Dataset Hurst estimates

Dataset	$H_{(R/S)}$	H_{AM}	H_{AV}	H_{Hig}	H_{RR}
A	0.8535	0.9098	0.9114	1.2101	1.0005
B	0.7928	0.8784	0.8717	1.0698	1.0002
C	0.8502	0.9237	0.9109	1.1131	1.0003
D	0.9180	0.9531	0.9501	1.3412	0.9994

Table 2 summarizes the values of H using (R/S) for each dataset, emphasizing that the overall LRD quality of each time series remains high, despite the fact that, at some time scales, they appear less so.

Interestingly, these periods of lower H occur at aggregation levels typically being examined by system operators for such applications. As a result, the casual observer might incorrectly assume the LRD for these applications is low.While estimating H using (R/S) is well-established as the standard, it is nevertheless computationally expensive as mentioned previously, and many other methods have been devised to estimate H more quickly [7][8][9][10].As examples of these methods, Table 2 includes results using the "Absolute Moment" (AM), "Aggregate Variance" (AV), "Higuchi" (Hig), and Peng's "Residuals of Regression" (RR) methods from Taqqu and Levy [7] using MatLab. While these methods produced results in a few seconds or less

each versus up to 20 minutes using (R/S), all resulted in values of H higher than expected. Furthermore, all values for Higuchi and most values for Peng were found to be invalid by being greater than the maximum of 1.

VI.   CONCLUSIONS AND FUTURE WORK

The systems in this study vary in number of vehicles, and while similar, are not identical in their purpose and behavior. However, all exhibit consistently high self-similar message frequency patterns at all scales measured. As a result, it should be reasonable to assume that similar system behavior would continue to persist at even larger scales as the systems grow.This, along with many other analyses of Internet-based communications systems of all types, consistently shows that they exhibit LRD characteristics. While Park et. al [13] found high Hurst exponent values between 0.8 and 0.9 for TCP/IP communications with its stream-oriented protocol overhead, the systems in this study using more simplistic UDP/IP communications, often without message acknowledgement, likewise have H values in the same range.Differences in aggregate shape of the data at time scales corresponding to days and weeks appeared to exhibit less LRD due to their "mean regressive" periodicity, but high LRD character remained true at both lower and higher scales.The well-established standard for estimating H based on (R/S) is computationally expensive, motivating the desire for alternative methods, but such methods, while confirming the high LRD of the systems in this study, appear to produce results too different from (R/S) to be equivalent.Since these systems were studied at the level of UDP packet arrival at "host"-level aggregation services, in cellular wireless communications, two additional layers of communications exist below this. For GSM networks, there is a layer of GPRS messaging and its underlying SS7 signaling protocol for which a similar Hurst analysis may be of value in the future.

REFERENCES

[1]   R. G. Clegg, "Simulating internet traffic with Markov-modulated processes," in UK Performance Engineering Workshop, 2007.
[2]   H. E. Hurst and A. S. of Civil Engineers. Hydraulics Division, Long-term Storage Capacity of Reservoirs. American Society of Civil Engineers, 1950.
[3]   B. B. Mandelbrot, "Long-run linearity, locally gaussian processes, H-spectra and infinite variances," 1969.
[4]   B. B. Mandelbrot and J. R. Wallis, "Computer Experiments with Fractional Gaussian Noises: Part 2, Rescaled Ranges and Spectra," Water Resources Research, vol. 5, no. 1, pp. 242–259, 1969. [Online]. Available: http://dx.doi.org/10.1029/WR005i001p00242
[5]   W. E. Leland, M. Taqqu, W. Willinger, and D. V. Wilson, "On the Self-Similar Nature of Ethernet Trafc (Extended Version)," IEEE/ACM Transactions on Networking, vol. 2, no. 1, pp. 1–15, 1994.
[6]   M. Taqqu and J. Levy, "Using renewal processes to generate LRD and high variability," in DEPENDENCE IN PROBABILITY AND STATISTICS: A Survey of Recent Results,
E. Eberlein and M. Taqqu, Eds. Birkhaeuser,
Boston, 1986, pp. 73–89.
[7]   M. Taqqu, V. Teverovsky, and W. Willinger,
"Estimators for Long Range Dependence: An Empirical Study," Fractals, vol. 3, pp. 785–798, 1995.
[8]   T. Higuchi, "Approach to an irregular time series on the basis of the fractal theory," Phys. D, vol. 31, no. 2, pp. 277–283, Jun. 1988.
[9]   C. Peng, S. Buldyrev, S. Havlin, M. Simons,
H. Stanley, and A. Goldberger, "Mosaic organization of DNA nucleotides," Physical review E, vol. 49, no. 2, pp. 1685–1689, 1994.
[10]   O. Jones and Y. Shen, "Estimating the Hurst Index of a Self-Similar Process via the Crossing Tree," IEEE Signal Processing Letters, vol. 11, no. 4, pp. 416–419, Apr. 2004.
[11]   D. Nash and D. Ragsdale, "Simulation of self-similarity in network utilization patterns as a precursor to automated testing of intrusion detection systems," IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 31, no. 4, pp. 327–331, Jul. 2001.
[12]   M. Y. Idris, A. H. Abdullah, and M. A. Maarof, "Iterative Window Size Estimation on Self-Similarity Measurement for Network Traffic Anomaly Detection," International Journal of Computing & Information Sciences, vol. 2, no. 2, pp. 84–91, 2004.
[13]   C. Park, F. Hernandez-Campos, L. Le, J. S. Marron, J. Park, V. Pipiras, F. D. Smith, R. L. Smith, M. Trovero, and Z. Zhu, "Long Range Dependence Analysis of Internet Traffic," Computer Networks: The International Journal of Computer and Telecommunications Networking - Special issue: Long range dependent traffic,
vol. 48, no. 3, pp. 401–422, 2005.
[14]   R. Dobrescu, D. Hossu, and R. Ulrich,
"Self-similarity Tests for Internet Traffic," Control Engineering and Applied Informatics, vol. 11,
no. 4, pp. 11–17, 2009.