

On the Self-Similarity of Web Query Traffic: Evidence, Cause and Performance Implications

Raju Balakrishnan
rajub@asu.edu

Subbarao Kambhampati
rao@asu.edu

School of Computing and Informatics
Arizona State University
Tempe AZ 85287

ABSTRACT

Temporal distribution of the web query traffic has a significant influence on the performance and scalability of large scale information systems. Most of the existing performance analysis assume a standard Poisson distribution. In this paper, we demonstrate that the temporal distribution of web query log traffic is statistically self-similar, and that the currently assumed Poisson query distribution does not capture the statistical properties of the distribution. Also, we propose a high variability aggregation generative model as the physical explanation for the observed self-similarity, and support the model by rigorous statistical analysis on queries of the individual users from the AOL query log. We also empirically demonstrate the implications of our findings on capacity planning and performance evaluation of information systems based on queuing analysis. Our results show that the assumed Poisson distribution over-estimates the usable capacity of web information systems significantly.

1. INTRODUCTION

Search engines (and other online databases) aim to provide sub-second response times for queries. To do this effectively, they not only need to improve query processing time (e.g. through optimized index structures etc.), but also need access to accurate models of query arrival. The latter is important as in order to respond to queries in bounded time, web systems need to be planned with sufficient capacity, so that the sum of queueing and service time for queries is acceptable for users. Temporal distribution of queries is thus an important parameter for capacity planning and performance evaluation of information systems like search engines and web databases.

Most theoretical frameworks and simulation studies for performance evaluation assume that temporal distribution of queries follows a Poisson distribution. For example, Chowdhury *et al.* assumed Poisson temporal distribution of query logs in their theoretical framework for analyzing search engines [9]. Similarly, simulation studies of search engine performance commonly assume Poisson arrival process of queries [7, 18, 19, 20].

In this paper, we start by attempting to analyze the validity of Poisson like process assumption. We use the publicly available AOL query logs [25] (containing 20 million queries by 650,000 users over a period of three months.) Our analysis demonstrates that the Poisson distribution fails to capture the statistical properties of the actual query distribution.

We also see that the Poisson distribution grossly over-estimates the percentage utilization before queue lengths and service quality suffer. At first glance, this might look like an obvious result, since after all as a memoryless process, Poisson distribution cannot be expected to model periodic trends like daily, weekly and monthly cycles in query arrival rates. We show however that the modeling inability transcends simple cycles. In particular, we will see that the burstiness in the arrival rate survives aggregation across time scales.

Results of our analysis, including burstiness at all scales, strongly suggest self-similar nature of query logs.¹ We confirm this by estimating the Hurst parameter for the observed distribution and showing that the estimates validate self-similar nature of the logs. Beyond invalidating Poisson like distributions, this proof of self similarity has the important implications as: (i) Modeling query traffic with any short range dependent distribution (Poisson has zero range dependency, a special case of short range dependent distribution) would require number of model parameters increasing with sample size (ii) Using a memoryless distribution like Poisson even for a short samples is not sound. We will discuss further on this in Section 4 below. Next we present (and validate) a generative model for queries, which provides a physical explanation for the observed self-similarity.

We demonstrate that the self-similarity and burstiness in all time scales have important implications on server utilization, throughput, and response time. Intuitively, if the input query arrival process is bursty (non-uniformly distributed) the server will be partially idle during low burst periods, since the traffic rate is less than the throughput of the system. Thus part of the server capacity is wasted, and the effective throughput of the server will be less than the maximum throughput. On the other hand, during high bursts the traffic rates will be much more than the mean service rate, and queue length grows to large values. Due to these effects the queue length becomes unbounded for much smaller mean query rates than the mean throughput for non-uniform traffic. Poisson distribution currently in use significantly over-estimates the usable capacity, since Poisson is much less burstier than actual self-similar query traffic. We verify this intuitively suggested difference in server utilization between Poisson distribution and actual query traffic by queueing experiments. This directly implies that the self-similarity is

¹The paper analyzes the temporal distribution of the query log. Time distribution of queries should not be confused with well-known power law distribution of query content.

a dominant characteristic of a number of web information system engineering issues such as capacity planning, and dimensioning of buffers.

Note that the implication of the self similarity is not limited to the search engines. Since the general web information system query generation follows the same process, the results will have impact on deep web in general — which is 500 times as large as web composed of HTML pages according to some estimates [1]. Over the last decade, the realization that the ethernet traffic is self-similar (c.f. [17]) has brought about significant changes in the engineering of network systems. We are sanguine that the current work will have similar impact on the design of search engines and other online database systems.

Rest of the the paper is organized as follows. Next section discusses the related work in the area. Section 3.1 explains the data set used. Section 3.2 gives the necessary statistical background required to understand self-similar processes and long range dependency. Section 4 demonstrates the visual difference between the query log distribution and Poisson-distribution, followed by formal proof by Hurst parameter estimation. In the next section we present a generative model to explain self-similarity and validate the model against the observed user-behavior. In section 6 we discuss how the observed self-similarity has crucial implications on performance of search engines and web databases with the help of queuing experiments.

2. RELATED WORK

Capacity planning and load balancing of web information systems has been extensively researched [9, 24]. Researchers frequently used simulation for performance analysis of web information systems [6, 9]. Many of these simulation studies assume that the arrival process is Poisson [9, 7, 18, 19, 20]. In their landmark paper Crovella and Bestavros [12] revealed self similarity of web server traffic. A generative model is suggested by Barford and Crovella [4]. Our work is specific to web search engine query logs rather than for generic web server requests.

Content distribution of query logs—frequency distribution of distinct queries—are analyzed in detail and commonly accepted to be following a power law [2]. Power law distributions are closely associated with self-similar distributions, and often considered to be a signature of self-similarity. But the content distribution is independent of temporal distribution, as any content distribution can be combined with any temporal distribution. For example, query frequencies may be distributed following a power law but queries may be arriving in a uniform distribution of one query per second, or a number according to a Poisson process. Hence the well-known power law distribution query log content does not imply the results presented in this paper.

The term “self-similarity” was introduced by Mandelbrot [22] and later found to have applications in many areas of natural sciences like hydrology, geophysics, and diverse fields of engineering [10]. A characteristic of self-similar (fractal) processes is long range dependence (LRD), which is often used to establish the existence of self similarity. LRD implies that the autocorrelation function for the distribution decays slowly with time shifts. In their classic paper Leland *et al.* [17] established the self similarity of ethernet traffic, which corrected the then pervasive belief in Poisson distribution of packet traffic and had important implication

in network performance and congestion control [17]. Following this land-mark paper, many other researchers verified the results in other networks and there has been extensive research in related areas [16, 28]. Beyond the statistical significance, fractal nature is found to have strong implications on queue lengths in network buffers and a dominant characteristics of a number of traffic engineering problems [13]. These two and a related set of high-impact papers on self-similarity of ethernet traffic and implications significantly changed traffic and network engineering. Our work can be viewed as a parallel effort on query traffic in information retrieval.

Considering query log analysis research, temporal distribution of query types are studied by Beitzel *et al.* [5]. Determining semantic similarity between the search engine queries based on temporal correlations is studied by Chien and Immorlica [8]. Badue *et al.* [3] makes a passing reference to existence of self-similarity of query logs, but the work is far from conclusive since they merely report value of H-Parameter without details of methodology etc. But in light of this paper, their work may be considered as evidence of self-similarity in a different workload.

3. PRELIMINARIES

3.1 Dataset Used

The data set used for our analysis is AOL query logs [25], containing approximately twenty million queries by 650,000 users over a period of three months-01 March 2006 to 31 May 2006. The logs include clicks on search results, queries, and time stamps of each query in granularity of seconds. Note a critical difference between the queries and clicks. Google and MSN clicks send requests directly to the third party URL and do not route through search provider’s server. But for AOL and Yahoo! the clicks are first sent to the their own server, and then redirected to the third party URLs. In typical web databases like Amazon and redirecting search engines like Yahoo!, the click on results will hit the server and will contribute to the workload. On the other hand, direct request from client side to third party URLs as in Google, does not contribute to workload of the server. It can be argued that only queries contribute to workload. But in our analysis, we do not distinguish between queries and clicks, since this approach is more applicable to web databases in general than just the special case of search engines. For the logs used, time stamps are recorded in granularity of seconds, hence temporal analysis the query logs in finer granularity is not possible. In the published logs, sixteen hours of queries on May 16 are inexplicably missing. Rather than using interpolation or patching. we limit our analysis to data before May 16. The query rate in logs have clear fixed period cycles based on time of the day.

3.2 Background on Self-Similar Processes

We now review some basic statistical definitions required to understand self-similar stochastic processes. A process is stationary if the mean and the variance do not vary over time. Autocorrelation function $\rho(\tau)$ of a stationary stochastic process X with mean μ and variance σ^2 is defined as,

$$\rho(\tau) = \frac{E[(X_t - \mu)(X_{t+\tau} - \mu)]}{\sigma^2}$$

Intuitively, autocorrelation function gives the similarity be-

tween the time series and a time shifted version of the time series.

Aggregation of the time series X of length T is given by averaging the time series in non-overlapping windows of size m .

$$X^m(i) = \sum_{t=(i-1)m+1}^{mi} X_t \quad \text{where } t \in T$$

Informally, statistical self-similarity means process may be roughly described as “*part represents the whole*”; or if a part of the distribution is projected to finer time scales, part will resemble the whole of the distribution in more coarse time scale. This is referred as scale free distributions, or fractal nature. More formal definitions are given below.

Second Order Self-Similar: Second order Self-Similarity means properties of the time series, at least the autocorrelation function, are preserved on aggregation. X is exactly self-similar if

$$\rho^{(m)}(\tau) = \rho(\tau) \text{ for } \tau \geq 0 \text{ and}$$

$$\text{Var}(X^{(m)}) = \sigma^2 m^{-\beta}$$

where $\rho^{(m)}(\tau)$ is the autocorrelation function for aggregation of aggregation level m and σ is standard deviation of original non-aggregate time series. The Hurst Parameter (Hurst Exponent) H of the self-similar process is defined as $H = 1 - \frac{\beta}{2}$

X is asymptotically self-similar if

$$\rho^{(m)}(\tau) = \rho(\tau) \text{ as } m \rightarrow \infty$$

Three implications of self-similarity are

1. No natural length of bursts.
2. Presence of bursts in all time scales.
3. Process does not smooth out on aggregation. The Self-Similar process has $0.5 < H \leq 1$ where as non-self similar processes have $0 \leq H \leq 0.5$. This means that the smoothing with aggregation is much slower for self-similar processes, the greater the degree of self-similarity, the slower will be smoothing with aggregation.

In the following, we discuss three mathematical manifestations of self similarity, (i) Long range dependence (ii) Slowly decaying variance (iii) Hurst effect.

Long Range Dependence: Long-Range dependence implies autocorrelation function does not decrease fast with the time shift. While autocorrelation decays exponentially with time shift for short range dependent processes, the decay is hyperbolic for LRD, leading to non-summable autocorrelation function. i.e

$$\sum_{\tau=-\infty}^{\tau=\infty} |\rho(\tau)| = \infty$$

Slowly Decaying Variance: While variance for an aggregate process varies inverse to the aggregation size for normal processes, for self-similar process variance decays more slowly with aggregation.

$$\text{Var}(X^{(m)}) = \sigma^2 m^{-\beta} \text{ Where } 0 < \beta < 1$$

Hurst Effect: Self-similarity gives elegant explanation to Hurst effect demonstrated by many natural phenomena. To explain Hurst Effect, let us define rescaled adjusted range statistic (R/S statistic) first. R/S statistic of a process X is given by

$$\frac{R(X)}{S(X)} = \frac{\max(0, W_1, W_2, ..W_n) - \min(0, W_1, W_2, ..W_n)}{S(X)}$$

$$\text{Where } W_k = \sum_{i=1}^k X_k - k\mu_n$$

The R/S statistic is measuring the ratio of the range $R(X)$ — the difference between the minimum and maximum observations — to the standard deviation $S(X)$ over different aggregation levels (scales). Generally, the R/S statistic increases with the value of aggregation, and the rate of increase is faster for LRD processes than short term dependent processes. For a process, the value of the R/S statistic can be expressed as a function of sample size used for calculation as,

$$E \left[\frac{R(X)}{S(X)} \right] \sim an^H$$

Where H is the Hurst parameter. Hurst parameter is also the slope of the curve R/S against number of samples used in log-log plot. $0 \leq H < 0.5$ for short range dependent processes and $0.5 < H \leq 1$ for LRD processes, and for a random walk $H = 0.5$. This faster increase of R/S statistic as sample size increases, when $H > 0.5$, is called the Hurst Effect. Many natural processes shows Hurst Effect with H value typically around 0.7, and fractal/self-similar distributions gave an explanation to observed long range dependence and Hurst parameter values.

The practical way to estimate degree of self-similarity is to measure the values of Hurst exponent. One way of estimating Hurst parameter is to plot the R/S line for different value of samples sizes, and find the slope of the fitted line by linear regression. There are several other methods in frequency and time domain to measure the Hurst parameter.

4. EVIDENCE OF SELF-SIMILARITY

4.1 Evidence of Burstiness

Before providing formal estimation of self-similarity, we provide a graphical evidence of bursty nature of the query log data at all time scales. We also show that this observed burstiness is not accounted by the Poisson distribution. In Figure 1, we show aggregation of the query logs in four different time scales-ranging from one second to thousand seconds time units. Plot 1(a)-(d) on the left are generated from real query logs and plots 1(a')-(d') on the right are generated using synthetic Poisson process of the same mean (3.45 queries/second) as the query traffic. Since the granularity of log data is one second time unit, finest time scale used is one second (Figure 1(a) and 1(a')). Each plot above is obtained by reducing the time resolution by a factor of ten-i.e. increasing time scale by ten.

The query traffic shows prominent daily cycles (time period 24 hours) of large amplitude of approximately 20000 in 1000s time scale, which hinders the visual comparison between the query plots and Poisson data plots. The cycles affect the larger time scales more since the shift due to

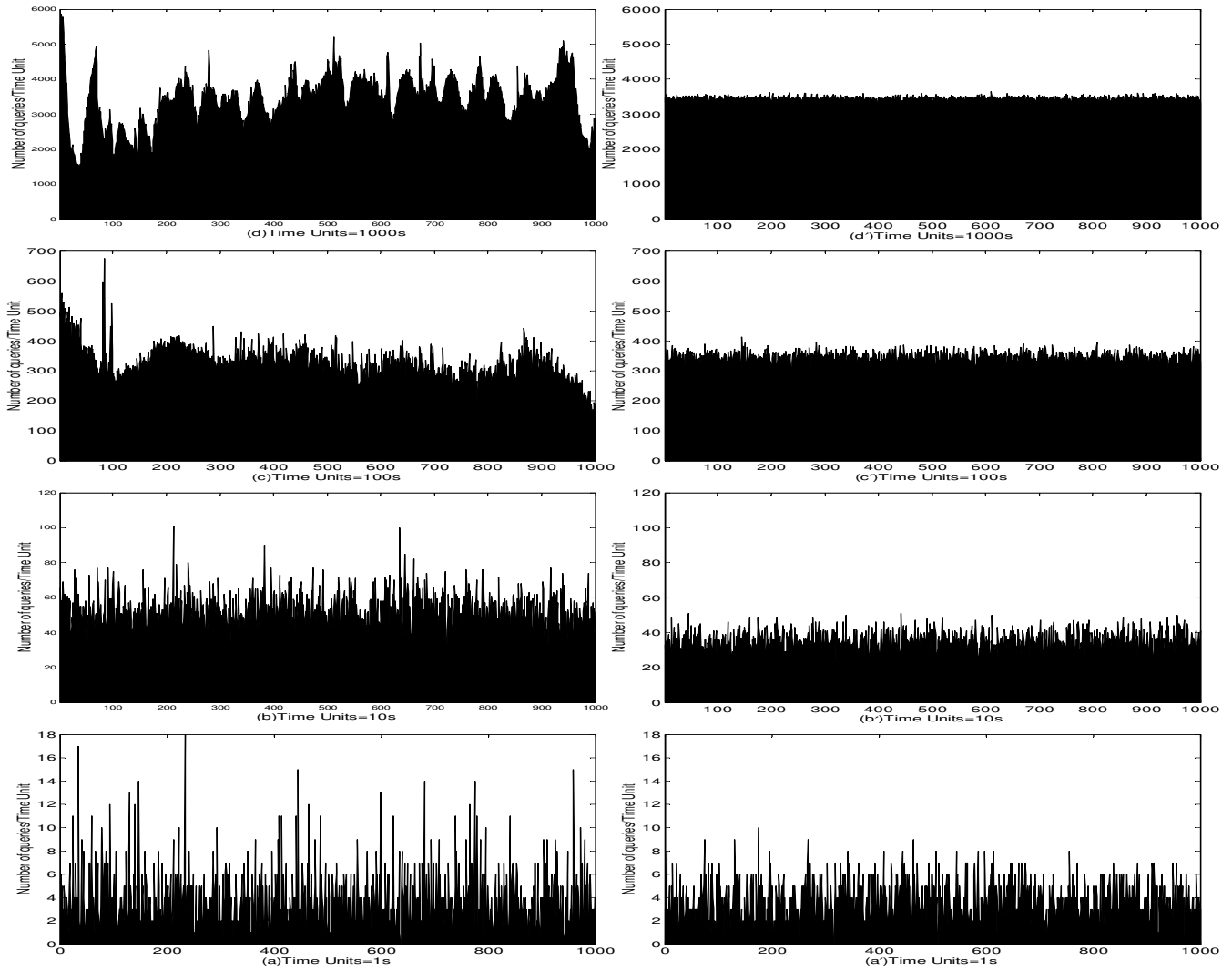


Figure 1: Aggregation of query logs *vs.* synthetic Poisson distribution. The plots (a) and (a') at the bottom are time distribution of query traffic and synthetic Poisson distribution respectively, with a sampling rate of one second. Each plot above is obtained by reducing the resolution of time scale of the plot below that by a factor of ten, or aggregating ten samples of the plot below. While plots (a) to (d) of real-traffic data preserve burstiness over aggregation, Poisson distribution plots (a') to (c') become smooth.

daily cycles is evident only during a time interval as lengthy as at least a few hours. To remove the daily cycles from 1000 and 100 second scale query data corresponding to plot 1(d) and (c), we removed the frequencies of period 24 hours and harmonics from the data. To do this, the time series is converted to frequency domain using 1000 point fast fourier transform, and frequencies corresponding to period 24 hours are removed with a few nearby frequencies. Also frequencies corresponding to second, third and fourth harmonics are removed, and signal is reconstructed in time domain. For 1(d) two neighboring lower and three neighboring higher frequency components of first harmonic are also removed. These frequencies were marked by clear peaks in frequency domain, and selection of frequencies to remove was based on visual inspection of power spectrum. Note that a visual examination of 1(d) and (c) reveals no cycles, suggesting effectiveness of our filtering method. Poisson distribution

for 1(c')-(d') is also filtered in the same way for the sake of uniformity, though removal has little effect as absence of a few components hardly affects almost uniform frequency distribution of size thousand. Data for plots 1(a) and 1(b) are not filtered and used as such.

In figure 1(a)-(d) of real traffic data three trends are evident; (i) Burstiness in all time scales: Even after thousand fold aggregation from figure 1(a)-(d) the burstiness of the traffic plot does not disappear. (ii) Lack of natural length of bursts: The figure shows burstiness ranging from seconds to hours. Full duration of the Figure 1(d) is 11.57 days, and some of the bursts have many hours of duration. (iii) Non-smoothing on aggregation: The thousand fold aggregation in Figure 1(d) again shows burstiness in traffic. These are exactly implications of self-similarity as described in section 3.2.

On the other hand, figure 1(a') to 1(d') show the behavior

| Method | Low | Confidence | Medium | Confidence | High | Confidence |
|-----------------------|-------|-------------|--------|-------------|-------|-------------|
| Aggregate Variance | 0.985 | 43.02 | 0.951 | 68.09 | 0.918 | 77.66 |
| R/S | 0.771 | 96.59 | 0.732 | 96.76 | 0.691 | 97.19 |
| Periodogram | 0.673 | | 0.592 | | 0.595 | |
| Absolute Moments | 0.930 | 51.06 | 0.896 | 60.15 | 0.855 | 65.04 |
| Variance of Residuals | 1.011 | 94.60 | 0.918 | 95.90 | 0.898 | 96.57 |
| Abry-Veitch | 0.545 | 0.537-0.553 | 0.546 | 0.538-0.554 | 0.565 | 0.557-0.573 |
| Whittle | 0.680 | 0.673-0.687 | 0.629 | 0.622-0.636 | 0.602 | 0.595-0.609 |

Table 1: The Hurst Parameter estimation for the three periods of traffic corresponding to low (2006-04-24 04:30:00), medium (2006-05-24 13:00:00) and high (2006-03-09 20:30:00) traffic periods. The time shown is mid point of the period. The confidence intervals or correlation coefficients are shown wherever applicable.

of synthetic Poisson process. The difference is obvious, the Poisson process smooths out with aggregation and resembles a uniformly distributed white noise at higher aggregations. The burstiness vanishes in coarse time scales, longer length bursts are absent, and bursts smooths out much faster with aggregation than actual query data. Thus, all three signatures of self-similarity present in the query traffic data discussed above are totally absent for Poisson process.

In our analysis, Figure 1(a) and (b) corresponding to short time periods is hard to discriminate from $1(a')$ and (d') visually. This naturally leads to questions of (i) Whether Poisson is a valid enough model for short intervals of query traffic. (ii) Whether basic Poisson model may be extended by number of Poisson distributions with different mean values to approximate the observed LRD. Let us answer question (ii) first: modeling an LRD process with any short term dependent process would require large number of parameters as the sample size to be modeled increases. On the other hand, a parsimonious single parameter (H-parameter) modeling is feasible for modeling using a LRD process [17]. Now going back to answer question (i), since Poisson process is fully memoryless, modeling even a very short length of LRD process with Poisson accurately would require infinite number of parameters. Otherwise, using memoryless Poisson process to model even small time durations of LRD process is not sound. Besides, different models for short intervals and long intervals is inelegant. The reader is referred to Leland *et al.* [17] for further discussion on problems in modeling LRD processes with short range dependent processes.

This analysis shows that Poisson modeling of query traffic is clearly inadequate in modeling the self-similar nature of real data, and is thus likely to give unrealistic results. We will elaborate this analysis in the next section, and discuss the consequences of self-similarity in the following sections.

4.2 H-Parameter Estimation

We estimate H-Parameter to demonstrate the long range dependency in query logs formally. Since there are number of manifestations for self-similarity as described in section 3.2, different methods in time and frequency domains are used in practice for the estimation. These methods work well in synthetic data, but in real-life data containing noise, cycles and trends different methods might estimate different values of H-Parameter. It is suggested to use multiple methods, report the correlation coefficients and confidence intervals by different methods, and visually inspect the data for trends and cycles [15, 11]. The chances of estimates agreeing on real data is small, but if most of the estimates

are above 0.5 the LRD is likely to exist. The confidence intervals provided by many estimators have only limited significance. Please check Karagiannis [15] and Clegg [11] for comprehensive understanding of imprecision in measuring H-Parameter.

Autocorrelation function and self-similarity is defined assuming stationary series as described in section 3.2. Our visual inspection of data plots (not shown here) showed that the data is stationary. Query traffic data clearly shows daily cycles and harmonics. Though the daily cycles are present in the data, the mean remains same over extended periods of time spanning days, apart from tell-tale signatures of self-similarity and LRD like burstiness and low frequency cycles.

Table 1 shows H-Parameter estimation for three different non-overlapping time periods corresponding to low, high, and medium traffic with mean number of queries 2.56, 5.10, and 6.69 per time unit respectively. The time periods are chosen at random, and 50000 samples corresponding to 13.89 hours of interval is used, with specified times in table 1 exactly at the middle of the interval. The data is used as such-without pre-processing. We used the estimators provided in SELFIS [14] toolkit for H-Parameter estimation. The estimation of H-Parameter is especially complicated in this data by daily cycles.

In Table 1, all estimators confirm LRD by estimating H-Parameter to be above 0.5. The imprecise nature of Hurst parameter estimation is evident in the Table 1 since the estimators are not agreeing on degree of self-similarity². The robust estimation of H-Parameter from the table is difficult, as the estimated values vary widely over this interval. But the fact that all the estimates of H-parameter is above 0.5 confirms existence of LRD in the data. The Hurst parameter might be in the range 0.65 to 0.75, since removing very low confidence estimates by Aggregate Variance and Absolute Moments and averaging gives the value of 0.697 for H-Parameter. Some estimators like R/S and Whittle show slight decrease in estimated values from low to high traffic periods, but this is not conclusive since it is not a general trend for all estimators and confidence intervals overlap even for these estimators.

5. GENERATIVE MODEL

Having established the existence of self-similarity, next

²As an attempt to increase the agreement between the estimators, we tried filtering and polynomial regression followed by subtraction of fitted curve to remove daily cycles, but confirming to prior observations [11], these methods found to be of little help.

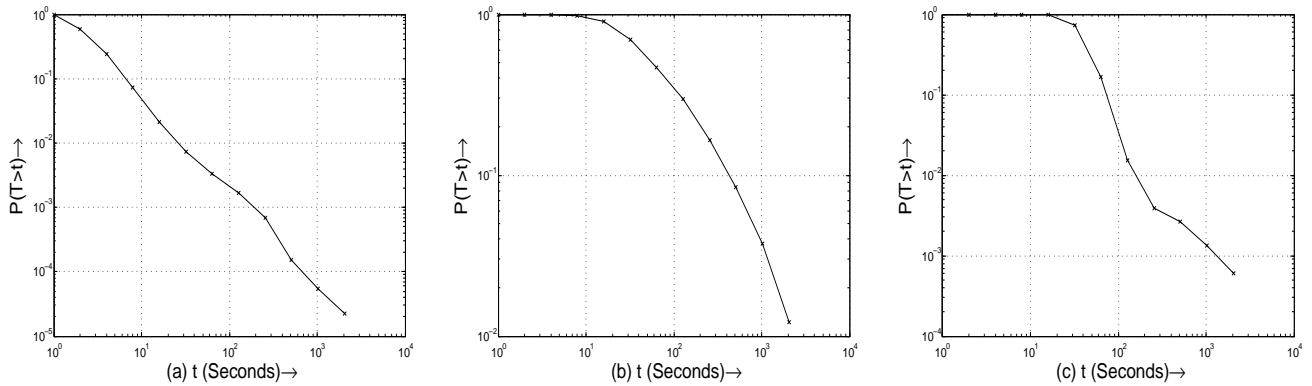


Figure 2: The power law distributions of user inter-query periods of top three user with 279430, 8695 and 8274 queries respectively. The X axis value t gives time in seconds, and Y axis is the fraction of inter-query periods greater than t for the user. All the plots show linear regions characterizing of slope between 1 and 2.

step is explaining the reason for observed self similarity in query logs. In this section, we provide a physical explanation for the existence of statistical-self similarity. We propose a generative process for self similar query traffic and establish the plausibility of the model by a rigorous statistical analysis on query logs.

Beyond the theoretical importance of providing a explanation for self similarity, a generative model has practical applications including,

(i) *Simulation studies for performance analysis:* Researchers and designers often use simulations to analyze the performance of information retrieval hardware and software [27, 21]. While the statistical distribution is sufficient for theoretical analysis and design of systems, generative model is required for simulations.

(ii) *Predicting query traffic behavior:* For example, consider the case of a closed enterprise database accessed by a few tens of thousands of employees, typically once in a week, will the query traffic be self-similar? Knowing the generative model and essential features of the model for self-similarity, the designer may predict the characteristics of aggregate query traffic, avoiding analysis of actual traffic, and customize the system design accordingly.

Query logs are an aggregation of queries by a large number of users. Hence an intuitive generative process providing physical explanation of self similarity of query logs must be an aggregation of queries from multiple sources. To propose and validate the generative process we follow a two step approach in the following two sections. In the next section, we describe a mathematical generative model for self-similar distributions proven by Taqqu *et al.* [26] and list out the required characteristics of query distributions of individual users. In Section 5.2 we analyze the individual user query distributions and demonstrate that the distributions indeed have the required statistical characteristics predicted by Taqqu *et al.*'s model.

5.1 High-Variability Aggregation Model

The generative model for self-similar distributions proved by Taqqu *et al.* [26] elucidates the characteristics required of individual sources for the aggregation of data from sources to be self similar. Taqqu's theorem says,

Aggregation of a number of ON/OFF sources with ON or OFF periods exhibiting infinite variance will lead to self-

similarity in the resulting aggregate process.

Here, the *ON* periods correspond to periods in which a source produces data. For example, the period in which a user sends query in context of query logs, or like the period in which a network node transmits data in context of network traffic. *OFF* periods correspond to idle periods of the source, or periods in which source produces no data, like period in which user does not issue any query. The infinite variance described in the theorem is same as the infinite variance known to be demonstrated by power-law distributions. The *ON* and *OFF* periods need not be following the same distribution for the aggregate distribution to be self-similar.

Taqqu *et al.* further proved that the required condition is that the inter-arrival time of sources, T should be a power law distribution as,

$$P(T > t) = ct^{-\alpha} \quad 1 < \alpha < 2 \quad (1)$$

and the hurst parameter H of the aggregate traffic related to α by,

$$H = \frac{(3 - \alpha)}{2} \quad (2)$$

where α is the slope of the constituting power law distributions. If the value of alpha is between the 1 and 2 the value of H-Parameter will be between 1 and 0.5, the characteristic range for self-similar distributions. This model was used by Willinger *et al.* [28] to explain observed self-similarity in local area network traffic.

Taqqu's theorem provides explanation of self-similarity in web query logs, and provides a plausible generative process. In context of query logs, the theorem states that the high variability (infinite variance) of query periods (*ON* periods) or time between consecutive queries (*OFF* periods) in individual users results in self-similarity of the aggregate query traffic. Intuitively, for query logs, inter-query periods (time between consecutive queries) is likely to be following power law distribution rather than query periods. That is, the model predicts a power law distribution of inter-query time periods of the individual users as the reason for observed self similarity, with exponent value between one and two.

5.2 User Level Query Log Analysis

In this section we analyze query distributions of individual users to demonstrate that the source level requirements

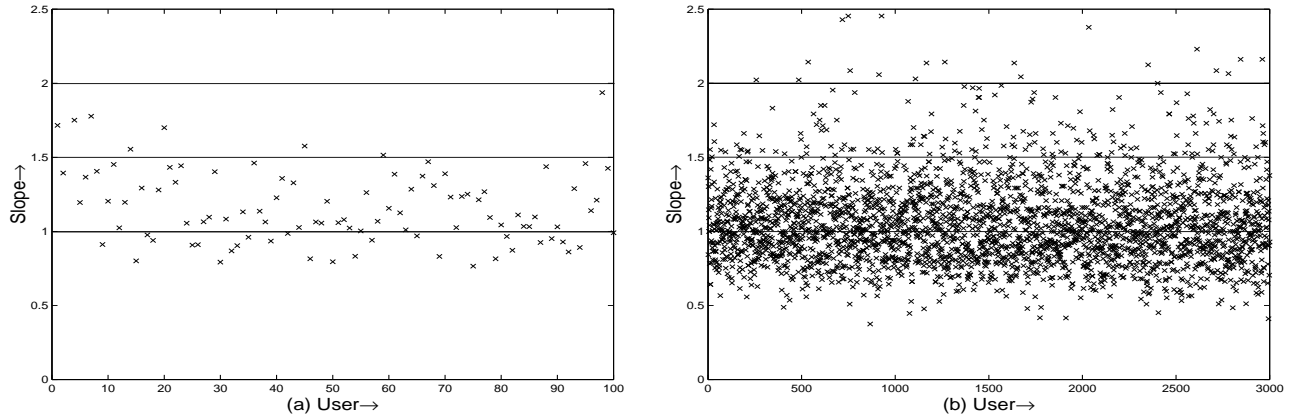


Figure 3: The scatter plots of slopes of power law distributions of users (a) Top 100 users, the majority of slopes are between 1 and 1.5 the mean slope is 1.17 and the weighted mean of the slopes with weights equal to the fraction of queries by the users is 1.43 (b) 3000 random users, a good fraction of slopes is between one and two, the visual examination clearly says at-least half of the slopes are between the one and two. Formal analysis shows that mean slope is 1.0541 and weighted mean is 1.0546

for self-similarity enumerated in previous section are satisfied. We extracted inter query periods of individual users for analysis, i.e. inter arrival times of queries with the same user id. Inter arrival times greater than 4096 seconds (1.14 hours) are not considered to remove the effect of daily cycles and trends based on time of the day. We focused on most active users, since majority of the traffic is likely to be dominated by active users. The sampling time for *ON* and *OFF* periods is one second, i.e. any time duration of one second with a query is counted as an *ON* period and any one second duration without a query is counted as an *OFF* period. Note that since logs used are sampled in one second duration, any higher resolution experiments is infeasible. Since the clicks on query results are logged with same time stamps as queries in log, clicks are removed from plots. The inter query periods are distributed into buckets, with bucket sizes exponentially increasing in powers of two. The exponentially increasing bucket sizes assures roughly the same number of samples correspond to each data point in log-log plot. We performed a three level analysis as, (i) a visual examination of inter arrival time distribution of top three users (ii) formal analysis of inter arrival time distributions of top 100 users (ii) formal analysis of distributions of 3000 random users.

5.2.1 Top Three Users

We illustrate the inter-arrival time plots for three top users (ranked by number of queries issued by the user), as shown in figure 2(a)-(c). The number of queries issued by first, second and third users are 279430, 8695 and 8274 respectively, and plotted in 2(a), 2(b), and 2(c). This gives a visual evidence of power law distribution and slope ranges in individual user query distribution

The observations are:

1. All the three plots in Figure 2(a) to 2(c) show clear straight line regions of slope between 1 and 2 (slope is negative, strictly speaking).
2. Flat regions are observed roughly from 1 to 16 seconds- at the beginning of the plots (see (b) and (c)). This is

intuitive since users are not likely to issue two queries in an interval less than a 15 seconds.

3. For the user corresponding to 2(a) flat region is absent, indicating anomalous behavior of large number of queries in a small interval, which is a typical characteristic of a robot or a ISP. Part of the log for this user in fact contains such “fast” queries. For this paper, we need not distinguish between human users and robots, unlike log analysis for personalization and user preference elicitation studies, since they contribute to the load irrespective of the fact that they are robots or humans.

5.2.2 Top Hundred Users

To provide a more formal evidence of power law distributions with slopes between one and two, we checked queries by top hundred users. Though the “eyeballing” method is good enough for a few users, it is not practical to analyze large number of users. To analyze this set of hundred users, we created buckets same way as we described for the top three users above. But instead of plotting, we examined the maximal slope between region corresponding to 16 seconds and 4096 seconds by linear regression. The whole region is divided into four sub-regions of size three buckets each. We fitted a line against each of this sub-region by linear regression. Slope of these lines are used as the slope of the region. The maximum of these slopes is used as the slope for that particular user. Note that the slopes we found by this method are lower bounds, as a manual analysis of slopes by visual examination may find slopes of better region than analysis method we followed. The distribution of slopes for these hundred users are shown in Figure 3(a). The mean of the slopes is found to be 1.17. Since the fraction of query logs contributed by each user is proportional to the number of queries issued by the user, we calculated the weighed mean of the users where the weights are the fraction of total queries contributed by the user. Value of the weighted mean was found to be 1.43. The higher value of weighed mean suggests that active users tend to have higher slopes. These results confirm that the top 100 users exhibits the

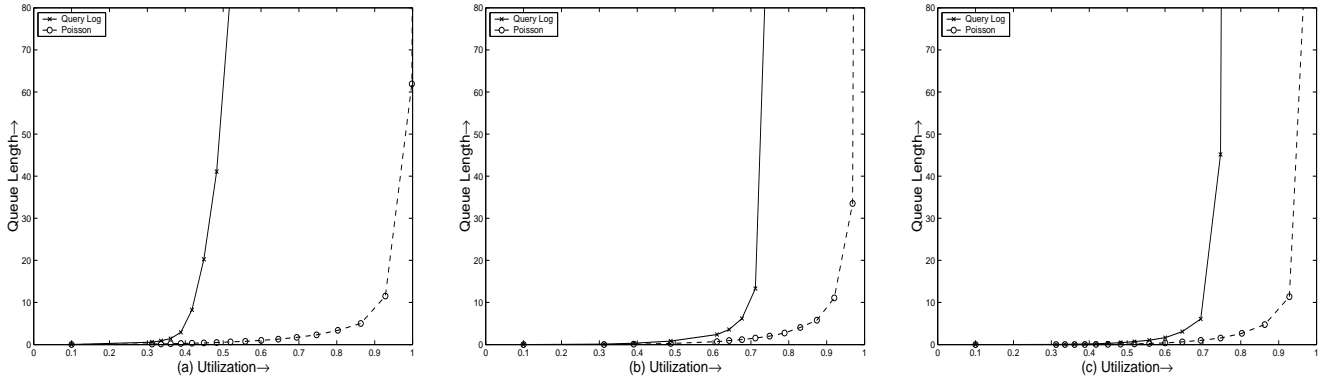


Figure 4: The plots from left to right show utilization *vs.* mean queue length for query traffic and synthetic Poisson distribution for (a) Low (from 2006-04-24 08:40:00) (b) Medium (from 2006-05-24 13:00:00) (c) High (from 2006-03-09 24:40:00) traffic periods with mean traffic 1.381, 3.965 and 7.46 respectively. The Maximum utilization—utilization corresponding to the “knee-point” of the curve—predicted by the Poisson far exceeds the actual possible utilization given by the curve corresponding to query logs. Though the difference is most prominent for low traffic periods, for all the three the predicted utilization exceeds actual utilization at least by 35%.

infinite variability with slopes between 1 and 2. We also did a correlation analysis of resulting lines from linear regression to the original log-log plot to check how good the regions resembles the straight lines. The average magnitude of correlation-coefficient is found to be 0.987 indicating good fit to power law distribution.

5.2.3 Random User

We also analyzed inter-query time for 3000 random users to check whether the slope is between one and two. We used the same method used for top 100 users described above. Only users having greater than 400 queries are considered, since a reasonable number of queries are required for the distributions to be meaningful. The distribution of the slopes are shown in Figure 3(b). From the distribution it is clear that a large number of users have slopes between one and two. As a formal estimate, the mean slope is 1.0541 and weighed mean is 1.0546. The correlation coefficient of reverse fitted lines is found to be 0.982. Also, note that there is no noticeable difference between weighted and non-weighted mean, indicating similar slopes for active and less active users.

5.2.4 Discussion on User Level Analysis

The empirical evidence provided in these three experiments along with Taqqu’s theorem stated above proves that the aggregation of infinite variance model is the generative process for observed self-similarity. Though the observed value of α (slopes) and H (Hurst Parameter) are not exactly same as the relation predicted by Equation 2, this is not surprising given the approximate nature of both the estimates.

As a note on implementability of this model for generation of self-similar query logs, the generative processes for power laws are well known [23]. Hence self-similar query log for simulations studies can be generated by aggregating number of power law distributions. Also H-Parameter of aggregate query log can be decided by changing power law exponent of individual sources, as exponents of individual sources and H-Parameter of aggregate are related by Equation 2.

In addition to providing a generative model for query traffic, the infinite variance in source level provides an additional proof for the self-similarity of query traffic. This model also implies that other search engines and web-databases query logs are also likely to be self-similar, since the statistical properties of user behavior are likely to be the same.

6. PRACTICAL IMPLICATIONS OF SELF-SIMILARITY

In this section, we discuss the impact of self-similarity of query logs on the performance parameters and engineering of web information systems. The burstiness of traffic demonstrated in Section 4.1 combined with intuitive implication of lower performance for busy traffic suggested in Introduction implies lower server utilization for actual query logs, compared to smoother Poisson process. To validate this implication empirically, we perform our experiments on a generic single server queuing system to make sure that our analysis is agnostic of search engine architecture, and results are valid across architectures. Specifically, we will illustrate through queuing experiments that the self-similar distribution, especially one of the manifestations of it — Long Range Dependency described in Section 3.2 — has a significant impact on the performance parameters of the web-information systems. We also discuss implications of LRD in query traffic on load balancing strategies for distributed servers.

For our queuing experiments we assume a single server, single queue and unbounded waiting room. The throughput of the server is assumed to be Poisson, following current literature [9]. Utilization is the ratio of mean arrival rate to mean service rate. To simulate different utilization levels of the server, the mean throughput is changed, keeping the arrival rate fixed. We perform two sets of experiments with a single server queuing system to show that (i) The memoryless Poisson process considerably overestimate the maximum utilization and performance of servers. (ii) The long range correlation accounts for the reduced utilization of servers in actual query logs.

6.1 Sever Utilization

In our first queuing experiment, we compare the queue length growth of the Poisson process with the actual self-similar query logs. For this, we used two arrival-service queues with the same mean service rates. For one queue, the arrival process is the actual arrivals from the query logs, and for the second queue the arrival process is the synthetic Poisson process generated with the same mean as the query arrival process. Thirty thousand consecutive samples of query arrival corresponding to an interval of 8.33 hours (30000 seconds) of the query logs are used. The results of three queuing experiments shown in Figure 4(a)-(c) correspond to query logs of low (mean traffic 1.381), medium (mean traffic 3.965) and high (mean traffic 7.46) traffic time periods.

Results in Figure 4(a)-(c) show that the intuitive suggestion of higher queue length implied by the burstiness described above in this section is indeed true. In figure the “knee-point” of the queue-length utilization curve is the point of interest, since beyond this point the queue length grows fast. The knee of the curve corresponding to the actual traffic is at a significantly lower utilization level than that of the Poisson traffic. While the Poisson predicts around 90% usable capacity, the actual usable capacity is only around 40-70%, as observed in curves corresponding to query traffic. This means that Poisson over estimates usable capacity by 35 to 100%. Also note that the queue length is much larger for actual query traffic in regions before knee point, compared to Poisson traffic. This implies that the required buffer size in this range is more than what is predicted by the Poisson process. Overall, Poisson significantly underestimates the system capacity required to guarantee bounded delay time for a mean query traffic rate.

6.2 Effect of LRD

Since the Poisson distribution is memoryless—not having short term and long term correlations—we explore the contribution of long term correlations against the contribution of short term correlations in reduced utilization using actual query logs. We used the high traffic period query logs from above queuing experiments, since performance studies are most interesting at high traffic periods. To examine the effect of long term correlations, we repeated queuing experiments removed the long range correlations from the query logs by an “external-shuffle”. External shuffle means dividing the time series into buckets and shuffling the order of buckets, while keeping the sequence of samples inside the bucket unchanged [13]. This preserves the short-term correlations up to a lag of bucket size, but removes all long range correlations with ranges above the bucket size. The plot of shuffled query arrival time series using a bucket size 20 in Figure 5 shows that the resulting curve moves towards the queue length characteristics of Poisson process. Otherwise, the usable capacity of the system increases for the shuffled processes due to the removal of long-term correlations. This shows that the long term correlations accounts for decrease in the usable capacity of the server.

In a second shuffling experiment, we performed an “internal-shuffle” of query logs, with the same bucket size 20. Internal shuffle means dividing the time series into buckets and shuffling the order of samples inside the buckets, while keeping the sequence of buckets unchanged [13]. This will remove all short term correlations less than the buckets size but

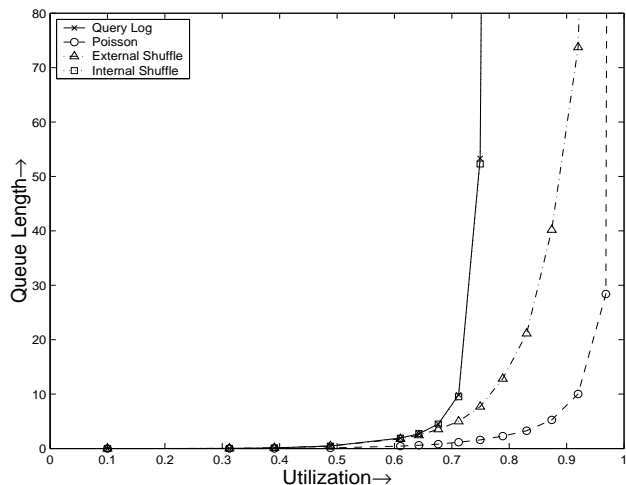


Figure 5: The plots from left to right show utilization *vs* mean queue length for query traffic, internally shuffled query traffic (query traffic and internally shuffled traffic exactly overlaps and not distinguishable in graph), externally shuffled query traffic, and Poisson distribution. Note that while the internal shuffle does not make any difference in predicted queue length, the external shuffled query logs moves towards the Poisson traffic since long range correlations are removed.

will keep all the long range correlations greater than bucket size. In Figure 5, the queue of internally shuffled query logs exactly overlaps with the actual query logs *vs* increased utilization for externally shuffled query logs. This clearly indicates that the removal of short term correlations has no effect on the effective utilization, and the difference in queue lengths between memoryless Poisson distribution and actual query logs is entirely due to LRD. These shuffling experiments gives clear evidence for the negative impact of LRD, a by-product of self-similar traffic, on the performance of web information systems.

Load balancing strategies of distributed retrieval systems [7, 24] might be optimized considering the effect of shuffling. Systems in which the capabilities and the document collections are duplicated, like meta-search systems and mirrored systems, the queries from a single queue is distributed to one of the multiple nodes for processing, with more than one node capable of doing the same task. If the queries are allocated in a maximally shuffled manner to minimize LRD of queries allocated to each node, utilization of nodes may be improved. We are leaving this as a future research topic.

7. CONCLUSION AND FUTURE WORK

Current simulation studies for performance evaluation and capacity planning most commonly assume that the temporal distribution of queries follows Poisson distribution [9, 7, 18, 19, 20]. In this paper, we questioned this assumption of Poisson distribution. Our analysis of the large-scale AOL query logs provides significant evidence that the real queries follow a self-similar distribution. In particular, our analysis showed burstiness at all time-scales, confirming scale-invariance of distribution. We also estimated and showed that Hurst parameter for the query logs is above 0.5, proving the self-similarity and Long Range Dependence formally.

We then turned our attention to (i) explaining the observed self-similarity and (ii) understanding its consequences on the performance. For the former, we were able to provide an explanation establishing infinite variance of inter-query periods of the individual user by rigorous analysis. For the latter, we showed that Poisson distribution grossly overestimates the possible utilization, when compared to the actual data. We also showed that the removal of long-range dependencies from the data does bring the estimates back in line with estimates based on memoryless Poisson assumption — further confirming the importance of taking the self-similarity into account while designing systems. We believe that the realization of self-similarity of query logs will likely have far-reaching impact in engineering of search engines and online databases (similar to the way the self-similarity of ethernet traffic (c.f. [17]) has had a profound effect on the engineering of network systems).

8. REFERENCES

- [1] White paper on deep web. <http://grids.ucs.indiana.edu/courses/xinformatics/searchindik/deepwebwhitepaper.pdf>.
- [2] E. Adar, D. S. Weld, B. N. Bershad, and S. D. Gribble. Why we search? visualizing and predicting user behavior. In *Proceedings of WWW*, pages 161–170. ACM, May 2007.
- [3] C. Badue, R. Baeza-Yates, and B. Ribeiro-Neto. Modeling performance-driven workload characterization of web search systems. In *Proceedings of CIKM*. ACM, 2006.
- [4] P. Barford and M. Crovella. Generating representative Web workloads for network and server performance evaluation. In *Proceedings of the 1998 ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, pages 151–160. ACM New York, NY, USA, 1998.
- [5] S. M. Beitzel, E. C. Jensen, A. Chowdhury, D. Grossman, and O. Frieder. Hourly analysis of a very large topically categorized web query log. In *Proceedings of SIGIR*. ACM, 2004.
- [6] F. CACHED, V. Plachouras, and I. Ounis. Performance analysis of distributed architectures to index one terabyte of text. In *Advances In Information Retrieval*, volume 2997, pages 394–408. Springer, March 2004.
- [7] B. Cahoon, K. S. McKinley, and Z. Lu. Evaluating the performance of distributed architectures for information retrieval using a variety of workloads. In *Transactions on Information Systems*, volume 18, pages 1 – 43. ACM, January 2000.
- [8] S. Chien and N. Immerlica. Semantic Similarity Between Search Engine Queries Using Temporal Correlation.
- [9] A. Chowdhury and G. Pass. Operational requirements for scalable search systems. In *Proceedings of CIKM*. ACM, November 2003.
- [10] K. Chuang, J. Huang, and M. Chen. Power-law relationship and self-similarity in the itemset support distribution: analysis and applications. *The VLDB Journal The International Journal on Very Large Data Bases*, pages 1–21.
- [11] R. G. Clegg. A practical guide to measuring hurst-parameter. In *International Journal of Simulation: Systems, Science and Technology*, pages 3–14, October 2006.
- [12] M. Crovella and A. Bestavros. Self-similarity in World Wide Web traffic: evidence and possible causes. *Networking, IEEE/ACM Transactions on*, 5(6):835–846, 1997.
- [13] A. Erramilli, O. Narayan, and W. Willinger. Experimental queueing analysis with long-range dependent packet traffic. In *IEEE/ACM Transactions on Networking*, volume 4. ACM/IEEE, April 1996.
- [14] T. Karagiannis, M. Faloutsos, and M. Molle. A user-friendly self-similarity analysis tool. In *SIGCOMM Computer Communication Review*, volume 33, pages 81–93. ACM, 2003.
- [15] T. Karagiannis, M. Faloutsos, and R. Riedi. Long-range dependence: Now you see it, now you don’t! In *In Proceedings of GLOBECOM*, volume 3, pages 2165–2169. IEEE, November 2002.
- [16] T. Karagiannis, M. Molle, and M. Faloutsos. Long-range dependence—ten years of internet traffic modeling. In *IEEE Internet Computing*. IEEE, September 2004.
- [17] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of ethernet traffic (extended version). In *IEEE/ACM Transactions on Networking*, volume 2, pages 1–15, February 1994.
- [18] Z. Lu and K. McKinley. Partial collection replication versus caching for information retrieval systems. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 248–255. ACM New York, NY, USA, 2000.
- [19] Z. Lu and K. McKinley. Partial Collection Replication for Information Retrieval. *Information Retrieval*, 6(2):159–198, 2003.
- [20] Z. Lu, K. McKinly, and B. Cahoon. The hardware/software balancing act for information retrieval on symmetric multiprocessor. In *EurpPar*, pages 521–527, 1998.
- [21] Z. Lu, K. McKinly, and B. Cahoon. Partial collection replication versus caching for information retrieval systems. In *Research and Development in Information Retrieval*, pages 248–255, 2000.
- [22] B. Mandelbrot. How long is the coast of britain? statistical self-similarity and fractional dimension. In *Science*, volume 156, pages 636–638, May 1967.
- [23] M. Mitzenmacher. A brief history of generative models for power law and lognormal distributions. In *Internet Mathematics*, pages 226–251. ACM, May 2003.
- [24] A. Moffat, W. Webber, and J. Zobel. Load balancing for term-distributed parallel retrieval. In *SIGIR*, pages 348 – 355. ACM, 2006.
- [25] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In *Proceedings of First International Conference on Scalable Information Systems*, June 2006.
- [26] M. S. Taqqu, W. Willinger, and R. Sherman. Proof of fundamental results in self-similar traffic modeling. In *Computer Communication Review*, volume 27, pages 5–23. ACM, April 1997.
- [27] A. Tomasic and H. Garcia-Molina. Performance of

inverted indices in distributed text document retrieval system. In *Parallel and Distributed Information Systems*, pages 8–17, 1993.

- [28] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson. Self-similarity through high-variability: Statistical analysis of ethernet lan traffic at the source level. In *IEEE/ACM Transactions on Networking*, volume 5, pages 394–408. Springer, February 1997.