An In-Depth, Analytical Study of Sampling Techniques For Self-Similar Internet Traffic

Guanghui He and Jennifer C. Hou
Dept. of Computer Science
Univ. of Illinois at Urbana Champaign
Urbana, Illinois 61801
Email: {ghe,jhou}@cs.uiuc.edu

Abstract— **Internet traffic sampling techniques are very important to understand the traffic characteristics of the Internet [14], [8], and have received increasing attention. In spite of all the research efforts, none has taken into account the self-similarity of Internet traffic in analyzing and devising sampling strategies. In this paper, we perform an in-depth, analytical study of three sampling techniques for self-similar Internet traffic, namely static systematic sampling, stratified random sampling and simple random sampling. We show that while all three sampling techniques can accurately capture the Hurst parameter (second order statistics) of Internet traffic, they fail to capture the mean (first order statistics) faithfully, due to the bursty nature of Internet traffic. We also show that static systematic sampling renders the smallest variation of sampling results in different instances of sampling (i.e., it gives sampling results of high fidelity). Based on an important observation, we then devise a new variation of static systematic sampling, called *biased systematic sampling* (BSS), that gives much more accurate estimates of the mean, while keeping the sampling overhead low. Both the analysis on the three sampling techniques and the evaluation of BSS are performed on synthetic and real Internet traffic traces. The performance evaluation shows that BSS gives a performance improvement of 42% and 23% (in terms of efficiency) as compared to static systematic and simple random sampling.**

I. Introduction

Internet traffic sampling techniques are very important to understand the traffic characteristics of the Internet [14], [8]. If the sampled results faithfully represent Internet traffic, they can be utilized to monitor traffic on a short-term basis for hot spot and DoS detection [19] or on a long-term basis for traffic engineering [14] and accounting [9]. As such, the packet sampling approaches have been suggested by the IETF working groups IPFIX [16] and PSAMP [17]. Tools such as NetFlow [4] employ a naive 1-out-of-N sampling strategy in the router design.

The major challenge in employing sampling techniques is, however, scalability. Inspecting each individual packet for each flow or sampling at a very high rate is obviously not feasible, due to the large volume of traffic. On the other hand, if the sampling rate is not adequate, the sampled results may not reveal actual traffic characteristics. What makes the problem even more difficult is the bursty nature of the Internet traffic. As indicated in a number of recent empirical studies of traffic measurement from a variety of operational packet networks [20], [12], [23], [24], the Internet traffic is self-similar or long-range dependent (second order statistics, *LRD*) in nature. This implies the existence of concentrated periods of high activity (peaks) and low activity (valleys), i.e., burstiness, at a wide range of time scales. In the context of packet sampling, this implies that either the sampling rate must be high enough or the sampling strategy has to be judiciously devised so as to capture all the peaks and valleys in the traffic. As oversampling increases the memory requirements for the off-board measurement devices, and has the danger of making the sampling method unscalable, the latter approach (devising a sampling strategy that is able to capture the traffic characteristics) is preferred.

Several research efforts have been made to investigate the effectiveness of sampling techniques in measuring network traffic. Three commonly used sampling techniques, i.e., static systematic[1], stratified random and simple random, have been studied by Claffy *et al.* [3]. In particular, they explored various time-driven and event-driven sampling approaches with both random and deterministic selection patterns at a variety of time granularities. The results showed that event-driven techniques outperform time-driven ones, while the differences within each class are small. Cozzani and Giordano [6] used the simple random sampling technique to evaluate the ATM end-to-end delay. Estan and Varghese [13] proposed a random sampling algorithm to identify large flows, in which the sampling probability is determined according to the inspected packet size. Duffield *et al.* [9] focus on the issue of reducing the bandwidth needed for transmitting traffic measurements to a back-office system for later analysis and devise a size-dependent flow sampling method. The notion of adjusting the sampling density upon detection of traffic changes in order to meet certain constraints on the estimation accuracy was proposed in [2]. Finally, Duffield *et al.* [11], [10] investigated the issue of inferring stochastic properties of original flows (specifically the mean flow length, and the flow length distribution) from the sampled flow statistics.

In spite of all the research efforts, none has taken into account of the self-similarity of Internet traffic in devising sampling strategies. Three of the most important parameters for a self-similar process are the mean (first order statistics), the Hurst parameter (second order statistics), and the average

[1] In what follows, we omit "static" and simply name it systematic.

variance of the sampling results. In particular, the average variance of the sampling results is defined as follows: let $\bar{X}$ be the real mean of the parameter of interest in the original process, and X_i be the sampled result in the ith instance of sampling (i.e., the ith experiment). Then the average variance is defined as $E(V) = E[E[(X_i - \bar{X})^2]]$, where the inner expectation is taken over all the samples in one instance of sampling, and the outer expectation is taken over all the sampling instances (e.g., different starting sampling points in the systematic sampling technique give different sampling instances). The mean gives the most direct value of the traffic attribute to be measured. The Hurst parameter characterizes the second order statistics for a self-similar/LRD process, and is crucial for queuing analysis. The average variance is an index of the fidelity of the sampling results.

Although it has been reported in [21] that in sampling self-similar process with the three commonly used sampling techniques, the sampled mean is always smaller than the actual mean (i.e., the sampling techniques under-estimate the mean), no solution has been proposed to address this problem. The issues of whether the various sampling techniques accurately capture the Hurst parameter and/or render a small average variance have not been studied either. In this paper we close the gap and

T1. Investigate whether or not the three commonly used sampling techniques accurately capture the Hurst parameter. We also provide a sufficient and necessary condition (*SNC*) that a sampling strategy must satisfy in order to maintain the autocorrelation structure of the original process. Our derivation indicates that all the three methods satisfy the *SNC*.

T2. Verify whether or not the three commonly used sampling techniques render small average variances (and hence give high fidelity) by leveraging the results reported in [5]. Our research finding is that the systematic sampling method outperforms the other two.

T3. Demonstrate all three methods cannot provide accurate estimate of the mean for self-similar Internet traffic, especially when the sampling rate is small. We then propose, based on an important observation, a new variation of the systematic sampling technique, called *biased systematic sampling* (BSS), that gives much more accurate estimates of the mean, while keeping the sampling overhead low. As BSS is a variation of the systematic sampling technique, it retains all the advantages of the latter.

One thing worth mentioning is that, although it is not a problem for a router to count the incoming traffic and summarize the mean value of the total traffic going through the router, the obtained result instructs us little. In most cases, we are more interested in one or several original-destination flows (OD-flows). For example, we need to know the mean value of the aggregated traffic of 2 specified OD flows going between west coast and east coast in US. Under such similar case, the router counter fails to give the information we want and judiciously designed sampling techniques serve as an approach. Specifically, in this paper, we consider a generalized traffic process $f(t)$ without giving a rigid definition on it (Section II. It can be any individual OD-flow or the aggregation of several OD-flows going through a router. Our proposed method can be applied to any of these cases as long as a process $f(t)$ is specified.

Both the verification and validation in **T1–T3**, and the evaluation of BSS, have been performed on synthetic and real Internet traces. In particular, the real Internet traces were obtained from Lucent Technologies Bell Labs [18], contain millions of packets, and provide detailed packet level information for hundreds of pairs of end hosts.

The rest of the paper is organized as follows. After providing the background material in Section II, we investigate analytically in Section III whether or not the three sampling techniques accurately capture the Hurst parameter of the process to be measured and provide a *SNC* that a sampling strategy must satisfy in order to keep the second order statistics (and hence Hurst parameter). Then, we compare in Section IV the average variance of the sampling results obtained by the three techniques. Following that in Section V, we demonstrate with both synthesized and real Internet traces that all three techniques fail to capture the real mean of Internet traffic and present BSS in detail. Finally we present our performance study (again based on both synthesized and real traces) in Section VI. The paper concludes with Section VII.

II. Background

In this section, we introduce the self-similar processes and the three commonly used sampling techniques, and set the stage for subsequent derivation and discussion.

A. Self-Similar and Heavy-tailed Distribution

Let $\{f(t), t \in Z+\}$ be a time series which represents the traffic process measured at some fixed time granularity. As we have mentioned, the traffic process can be individual OD-flow or the aggregation of several OD-flows or any other flows the researchers are interested in. To make our approach a generic one, our definition on $f(t)$ is rather general. To define a self-similar process, we further define the aggregated series $f^{(m)}(\tau)$ as

$$f^{(m)}(\tau) = \frac{1}{m} \sum_{i=(\tau-1)m+1}^{\tau m} f(i). \tag{1}$$

$f^{(m)}(\tau)$ can be interpreted as follows: the time axis is divided into blocks of length m and the average value for each block is used to represent the aggregated process. The parameter τ is the index of the aggregated process, i.e., the τth block.

Let $R(\tau)$ and $R^{(m)}(\tau)$ denote the autocorrelation functions of $f(t)$ and $f^{(m)}(i)$, respectively. $f(t)$ is (asymptotically second-order) self-similar, if the following conditions hold:

$$\begin{aligned} R(\tau) &\sim \text{const} \cdot \tau^{-\beta}, && (2) \\ R^{(m)}(\tau) &\sim R(\tau), && (3) \end{aligned}$$

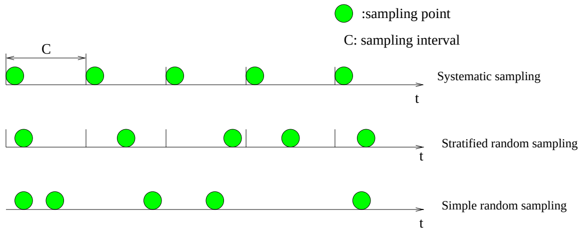


Fig. 1. An illustration of the three sampling techniques.

for large values of τ and m where $0 < \beta < 1$. That is, $f(t)$ is self-similar in the sense that the correlation structure is preserved with respect to time aggregation (Eq. (3)) and $R(\tau)$ behaves hyperbolically with $\sum_{\tau=0}^{\infty} R(\tau) = \infty$ (Eq. (2)). The latter property is also referred to as long range dependency (LRD).

Since self-similarity is closely related to heavy-tailed distributions, i.e., distributions whose tails decline via a power law with a small exponent (less than 2), we give a succinct summary of heavy-tailed distributions. The most commonly used heavy-tailed distribution is the Pareto distribution. A random variable X follows the Pareto distribution if its complementary cumulative distribution function (CCDF) follows:

$$Pr(X > x) \sim (k/x)^{\alpha}, x \geq k,$$

where α is the shape parameter and determines the decreasing rate of its tail distribution, and k is the scale parameter and is the smallest value X can take.

An important parameter that characterizes self-similarity/LRD is the Hurst parameter, defined as $H = 1 - \beta/2$. By the range of β, $1/2 < H < 1$. It can be seen from Eq. (2) that the larger H is, the more long-range dependent $f(t)$ is. A test for self-similarity/LRD can then be obtained by checking if H significantly deviates from $1/2$ or not.

B. Three Commonly Used Sampling Techniques

Generally speaking, the larger the sampling set, the more accurately the original process can be characterized. The price one has to pay is, however, the more CPU processing time and buffer space. Indeed there exists a trade-off between the sampling rate and the accuracy of sampling results. Three categories of sampling techniques have been commonly used in measuring Internet traffic: systematic sampling, stratified random sampling, and simple random sampling (Figure 1). In systematic sampling, every Cth element (e.g., packet) of the parent process is deterministically selected for sampling, starting from some starting sampling point. In stratified random sampling, the time axis is divided into intervals of length C, and one sample is randomly selected in each interval. In simple random sampling, N packets are randomly selected from the entire population.

III. Hurst Parameter of the Sampled Process

In this section, we first investigate whether or not the three sampling techniques accurately capture the Hurst parameter of Internet traffic. This is done by deriving the autocorrelation function of the sampled process obtained from the three sampling techniques. (Note that we do not intend to devise a procedure to estimate the Hurst parameter, but instead derive the Hurst parameter (through calculation of the autocorrelation function) of the sampled process and compare it with that of the original process.) Then we derive a *SNC* that a sampling technique has to satisfy in order to retain the autocorrelation structure of the original process.

A. Systematic Sampling

Let $f(t)$ and $g(t)$ denote the original and sampled process, and H_f and H_g the Hurst parameter of $f(t)$ and $g(t)$ respectively. Without loss of generality, t is discretized to be integer numbers: $0, 1, 2, 3....$ For systematic sampling, let C be the sampling interval. Then we have[2]

$$g(t) = f(Ct), t = 0, 1, 2, ... \tag{4}$$

Let $R_f(\tau)$ and $R_g(\tau)$ denote the autocorrelation function of $f(t)$ and $g(t)$, and $F(t)$ and $G(t)$ denote the CDF of $f(t)$ and $g(t)$ respectively. Then we have

$$\begin{aligned} R_g(\tau) &= E(g(t)g(t-\tau)) = E(f(Ct)f(Ct - C\tau)) \\ &= \int f(Ct)f(Ct - C\tau)dF(t). \end{aligned} \tag{5}$$

Let $Ct = u$. Then Eq. (5) can be re-written as

$$\begin{aligned} R_g(\tau) &= \int f(u)f(u-\tau)C^{-1}dF(t) \\ &= C^{-1} \cdot R_f(\tau). \end{aligned} \tag{6}$$

Hence $R_g(\tau) = C^{-1}R_f(\tau) \sim A\tau^{-\beta}$ as $\tau \to \infty$, where A is a constant. Also, we have $H_g = H_f = \frac{2-\beta}{2}$, where $0 < \beta < 1$. The above derivation implies that the sampled process obtained by the static systematic sampling technique has the same Hurst parameter as the original process.

B. Stratified Random Sampling

Recall that in stratified random sampling, the time axis is divided into interval of length C, and one sample is randomly selected in each interval. Using the same notation as in Section III-A, we have

$$\begin{aligned} R_g(\tau) &= E(g(t)g(t-\tau)) \\ &= E(f(Ct + \tau_1)f(Ct - C\tau + \tau_2)), \end{aligned}$$

where τ_1 and τ_2 are random variables that represent the time lags after the beginning of the tth and $(t-\tau)$th bucket respectively. $R_g(\tau)$ can be further written as

$$\begin{aligned} R_g(\tau) &= E(E(f(Ct+\tau_1)f(Ct - C\tau + \tau_2)|\tau_1, \tau_2)) \\ &= E(C^{-H-1}R_f(\tau + \frac{\tau_1 - \tau_2}{C})) \\ &= E(C^{-H-1}R_f(\tau + \tau')), \end{aligned}$$

where $\tau' = \frac{\tau_1 - \tau_2}{C}$.

[2] Without loss of generality, we denote the starting point of systematic sampling to be $t = 0$.

By Eq. (3), we have

$$\begin{aligned} R_g(\tau) &\sim E(D \cdot (\tau+\tau')^{-\beta}) \\ &= \int D \cdot (\tau+\tau')^{-\beta} f_{\tau'} d\tau', \end{aligned}$$

where D is a constant related to C, and $f_{\tau'}$ is the probability density function (pdf) of τ'. As both τ_1 and τ_2 are uniformly distributed in $[0, C]$, we have

$$f_{\tau'}(x) = \begin{cases} 1+x, & \text{if } -1 \le x \le 0, \\ 1-x, & \text{if } 0 \le x \le 1, \end{cases} \tag{7}$$

and hence

$$\begin{aligned} R_g(\tau) &\sim \int_{-1}^{1} D \cdot (\tau+\tau')^{-\beta} f_{\tau'} d\tau' \\ &\sim D\tau^{-\beta} \int_{-1}^{1} (1-\beta\frac{\tau'}{\tau}) f_{\tau'} d\tau' \\ &= D \cdot \tau^{-\beta} \text{as } \tau \to \infty. \end{aligned} \tag{8}$$

The last equality results from the fact that $E(\tau') = 0$. By Eq. (8), we conclude that the sampled process obtained by the stratified random sampling technique has the same Hurst parameter as the original process.

C. Simple Random Sampling

In simple random sampling, N samples are randomly selected from the entire population of M samples. That is, with probability $\eta = N/M$ a sample is selected. Let t_0 denote the sampling point in $f(t)$ corresponding to the tth sample $g(t)$. Then we have

$$\begin{aligned} R_g(\tau) &= E(g(t)g(t+\tau)) \\ &= E(f(t_0)f(t_0+a)) \ = \ R_f(a), \end{aligned}$$

where $a \ge \tau$ is a random variable. Since

$$\Pr(a = \tau+i) = \left(\begin{array}{c} \tau+i-1 \\ i \end{array} \right) \rho^{\tau}(1-\rho)^{i}, i = 0, 1, 2.., \tag{9}$$

we have

$$\begin{aligned} R_g(\tau) &= \sum_{a=\tau}^{\infty} R_f(a) \cdot \Pr(a) \\ &= \sum_{i=0}^{\infty} R_f(\tau+i) \left(\begin{array}{c} \tau+i-1 \\ i \end{array} \right) \rho^{\tau}(1-\rho)^{i} \\ &\sim \sum_{a=\tau}^{\infty} \Gamma a^{-\beta} \left(\begin{array}{c} a-1 \\ a-\tau \end{array} \right) \rho^{\tau}(1-\rho)^{a-\tau} \\ &= \sum_{a=\tau}^{\infty} \Gamma a^{-\beta} \frac{(a-1)!}{(a-\tau)!(\tau-1)!} \rho^{\tau}(1-\rho)^{a-\tau}, \end{aligned} \tag{10}$$

where Γ is a constant. Using the $Sterling$ equation, we can further approximate Eq. (10) as

$$\begin{aligned} R_g(\tau) &\approx \frac{\Gamma\rho^{\tau}}{\sqrt{2\pi(\tau-1)}(\tau-1)^{\tau-1}e^{-(\tau-1)}} \\ &\quad \sum_{a=\tau}^{\infty} \frac{a^{-\beta}(a-1)^{a-1/2}e^{-(a-1)}}{(a-\tau)^{a-\tau+1/2}e^{-(a-\tau)}} \cdot (1-\rho)^{a-\tau} \\ &= \frac{\Gamma\rho^{\tau}(1-\rho)^{-\tau}}{\sqrt{2\pi}(\tau-1)^{\tau-1/2}} \sum_{a=\tau}^{\infty} \frac{a^{-\beta}(a-1)^{a-1/2}(1-\rho)^{a}}{(a-\tau)^{a-\tau+1/2}} \\ &\triangleq \hat{\Gamma} \sum_{a=\tau}^{\infty} \frac{a^{-\beta}(a-1)^{a-1/2}(1-\rho)^{a}}{(a-\tau)^{a-\tau+1/2}}, \end{aligned} \tag{11}$$

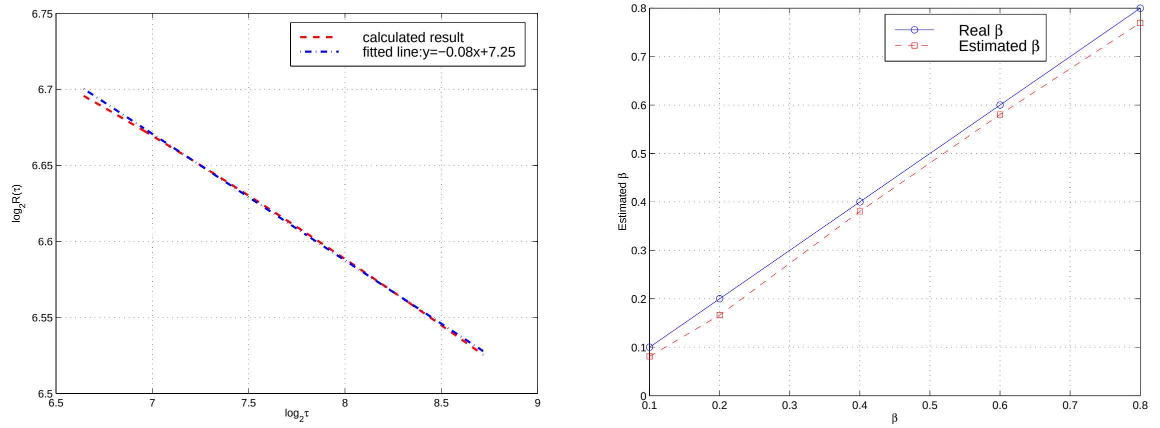
where $\hat{\Gamma} = \frac{\Gamma(\frac{\rho}{1-\rho})^{\tau}}{\sqrt{2\pi}(\tau-1)^{\tau-1/2}}$.

Since no closed form result can be obtained from Eq. (11), We use matlab to find the relation between $R_g(\tau)$ and τ. Specifically, We fit the value of $R_g(\tau)$ (calculated from Eq. (11)) to $const \cdot \tau^{\hat{\beta}}$ and depict the estimated value $\hat{\beta}$ and the real value of β in Fig. 2. In Fig. 2 (a) we fit the calculated result of $R_g(\tau)$ (after taking $\log_2$ on both τ and $R(\tau)$) to a line with slope $\hat{\beta} = -0.08$, where the real value is $\beta = 0.1$. By changing the real value of β from 0.1 to 0.8, we perform the same operation and report the estimated value of $\hat{\beta}$ in Fig. 2 (b). As shown in Fig. 2 (b), the values of $\hat{\beta}$ and β agree very well and hence $H_g \approx H_f$. Note the small gap between the values of $\hat{\beta}$ and β is due to the truncation error on the right hand side of Eq. (11), i.e., in calculating $R_g(\tau)$, we cannot sum up an infinite number of terms (from $a = \tau$ to ∞) and have to approximate the right hand side of Eq. (11) with a finite number of terms.

D. Sufficient and Necessary Condition for Accurately Capturing the Hurst Parameter

In Section III-A–III-C, we have shown that the sampled process generated by all three sampling techniques has the same Hurst parameter as the original process. A more general question is then: given a sampling technique, how do we check if the sampled process generated by this technique has the same Hurst parameter as the original process? To answer the question, we derive a sufficient and necessary condition (SNC) which a sampling technique has to satisfy in order to preserve the same second order statistics (and therefore Hurst parameter) in the thinned process.

We generalize the sampled process generated by a sampling method to be a point process $Z_n, n = 1, 2, 3...$, which represents the series of sampling points. The intervals between any two consecutive sampling points are defined as $T_i = Z_{i+1} - Z_i, i = 1, 2, ...$ T_i's are i.i.d random variables with the probability density function $h(x)$ for the continuous case and the probability mass function $H(x)$ for the discrete case. Note that Z_n is a renewal process with the renewal interval distribution h or H. A sampling method (and hence the sampled process generated by the sampling method) is generated by h or H. For example, the function H for systematic sampling is $H(C) = \Pr(T_i = C) = 1$ and $H(x) = 0$ for $x \neq C$, while



(a) The calculated result of $R_g(\tau)$ is fit into the line $log_2 R_g(\tau) = -0.08 log_2 \tau + 7.25$; the real value of β is 0.1.

(b) $\hat{\beta}$ versus β.

Fig. 2. Estimated and actual values of β.

the function h for stratified random sampling method is

$$h(x) = \begin{cases} \frac{1}{C^2}x, & \text{if } 0 \leq x \leq C, \\ \frac{2}{C} - \frac{1}{C^2}x, & \text{if } C \leq x \leq 2C, \end{cases} \tag{12}$$

where C is the length of each sampling bucket. For the simple random sampling technique with the sampling rate r, H can be expressed as

$$H(i) = Pr(T_i = i) = (1-r)^{i-1}r. \tag{13}$$

Under the assumption that the process $f(t)$ is wide sense stationary, we have

$$\begin{aligned} R_g(\tau) &= E\left(g(t)g(t-\tau)\right) \\ &= E\left(f(t+t_0)f(t+t_0-u)\right) \\ &= E\left(f(t)f(t-u)\right) \\ &= E\left(E\left(f(t)f(t-u)|u\right)\right) \\ &= \sum_{u=0}^{\infty} R_f(u)p(u), \end{aligned} \tag{14}$$

where $u = \sum_{i=1}^{\tau} T_i$ and $p(u)$ is the probability mass function of u. Note that $p(u)$ is the τth order convolution of $H(u)$, which we denote as $k(u,\tau)$ (as it is a function of both u and τ). Now we are in a position to derive the sufficient and necessary condition.

Theorem 1: Given any wide sense stationary (WSS) process $f(t)$, the sampled process $g(t)$ obtained from a sampling technique with h or H has the same second order statistics as the original process asymptotically if and only if the following condition holds

$$\sum_{u=0}^{\infty} R_f(u)k(u,\tau) \sim R_f(\tau), \tag{15}$$

where $k(u,\tau)$ is the τth order convolution of $H(u)$. *Proof:* By Eq. 14 we know that $g(t)$ retains the same second order statistics of $f(t)$ asymptotically, if and only if $R_g(\tau) \sim R_f(\tau)$, as $\tau \longrightarrow \infty$, and hence the conclusion.

Although Theorem 1 gives a sufficient and necessary condition for a sampling technique to retain the second order statistics of the original process, it cannot be readily applied, since $k(x,\tau)$ usually does not have a closed form, except for several extremely simple cases (e.g., for example the systematic sampling, in which $k(x,\tau) = \delta(x-\tau C)$, where $\delta()$ is the impulse *Dirac* function and C is the constant sampling interval).

In order to be able to apply Theorem 1, we propose a numerical method to calculate $k(x,\tau)$:

(S1) Calculate the Fourier transform of $H(x)$, $H(\omega)$. [3]

(S2) Let the Fourier transform of $k(x,\tau)$ (in terms of x) be $K(\omega,\tau)$. Then $K(\omega,\tau) = H(\omega)^{\tau}$.

(S3) Obtain $k(x,\tau)$ by deriving the inverse Fourier transform (IFT) of $K(\omega,\tau)$.

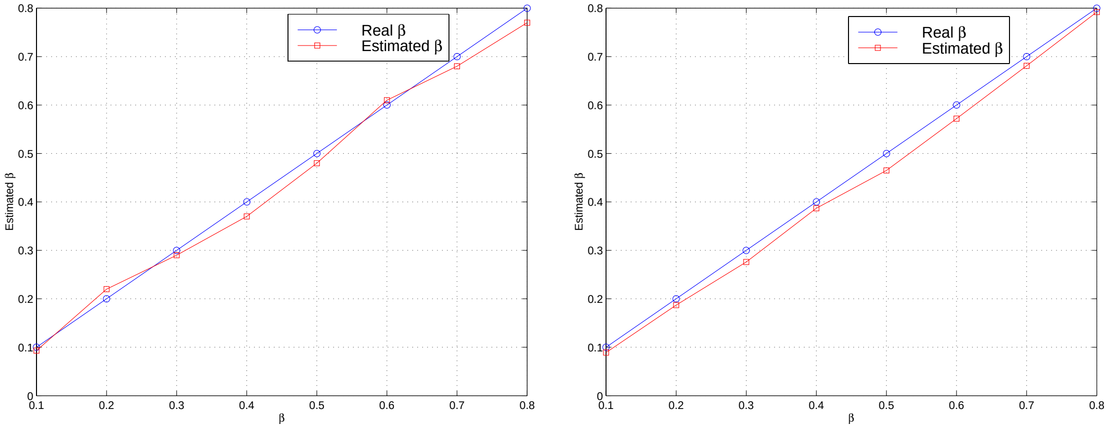
With $k(x,\tau)$, we can then calculate the left hand side of Eq. (15), and compare it against $R_f(\tau)$ as $\tau \longrightarrow \infty$. Since fast algorithms exist for both the Fourier and inverse Fourier transform, the above method provides a fast and reliable test in evaluating Eq. (15).

To validate the above proposed method for applying Theorem 1, we apply it to check the random stratified and simple random sampling techniques, and give the results in Fig. 3. As shown in Fig. 3, the estimated and real value of β agree extremely well, which is consistent with the derivation in Sections III-B–III-C.

IV. The Average Variance of Sampling Methods

Due to the randomness nature of stratified random sampling and simple random sampling, sampling results vary from one sampling instance to another, even if multiple instances of sampling are taken simultaneously and the same sampling rate is applied in each instance. Here by "instance," we mean each experiment made to take samples for a specific time interval.

[3] In the case that $H(x)$ cannot be expressed in a closed form, the proposed numerical method cannot be used to apply Theorem 1.



(a) stratified random sampling

(b) simple random sampling

Fig. 3. Estimated and real values of β under stratified random sampling and simple random sampling.

Even for systematic sampling, different starting sampling points may lead to different sampling results. If the variance of sampling results obtained from multiple instances is large, then one cannot rely on a single sampling instance to infer the entire process. To evaluate different sampling techniques in this aspect, we use the average variance of sampling results $E(V)$ as the index. Recall that $E(V)$ is defined as follows in Section I: let $\bar{X}$ be the real mean of the parameter of interest in the original process, and X_i be the sampled result in the ith instance of sampling (i.e., the ith experiment). Then the average variance is defined as $E(V) = E[E[(X_i - \bar{X})^2]]$.

Let V_{sy}, V_{rs} and V_{ran} denote, respectively, the variance of sampling results of systematic, stratified random and simple random sampling. To compare the three sampling techniques with respect to the average variance of sampling results, we leverage the results from [5] (Theorem 8.6):

Theorem 2: For a random process $f(t)$, with mean μ, variance σ^2, and autocorrelation function $R(\tau)$, if the following condition holds,

$$\delta_\tau = R(\tau+1) + R(\tau-1) - 2R(\tau) \geq 0, \qquad (16)$$

we have $E(V_{sy}) \leq E(V_{rs}) \leq E(V_{ran})$.
The result in Theorem 2 is actually quite intuitive. For systematic sampling, as the sampling interval remains unchanged among different sampling instances, the same second order statistic structure (e.g., the autocorrelation function) is retained. For the other two sampling techniques, different sampling instances have different second order statistic structures, although in the long run, they follow the same decreasing rule.

Theorem 2 gives a sufficient condition (Eq. (16)) in evaluating the three sampling techniques with respect to $E(V)$, given that the original process has finite mean and variance. To leverage Theorem 2, we first check whether the condition in Eq. (16) holds for a self-similar process. Using the fact that $R(\tau) \sim const \cdot \tau^{-\beta}$, we calculate δ_τ for different values of β and depict it in Fig. 4. As shown in Fig. 4, δ_τ is always positive regardless of the value of β, i.e., the condition in Eq. (16) holds.

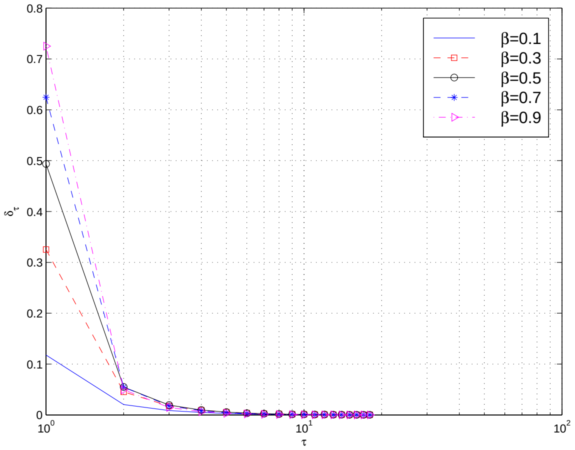


Fig. 4. δ_τ versus τ for different values of β.

In applying Theorem 2 we also need to verify if the process has finite mean and variance. A self-similar process (with $\alpha \in (1,2)$) has finite mean, but its variance goes to infinity as time goes to infinity. However, in practice we often consider finite time periods, and hence we conjecture the above condition is still valid. To verify the conjecture, we carry out experiments and measure the average variance of sampling results (under the three techniques) on both synthetic and real Internet traffic. In this experiments, we generate in *ns-2* self-similar traffic with Hurst parameter equal to 0.80 using the on-off model, where the on/off periods have heavy-tailed distributions with shape parameter $\alpha = \beta + 1$, $1 \leq \alpha \leq 2$. We also obtain real Internet traces from Lucent Technologies Bell Labs [18]. The set of traces was obtained on March 8, 2000, is in the *tcpdump* format, and contains detailed packet level information for hundreds of pairs of end hosts. The traces last for about 40 minutes and contains millions of packets. Fig. 5 shows the results. Note that Fig. 5 (b) gives the result for one of the trace sets with the Hurst parameter 0.62. Results for the other sets (that correspond to different servers) show similar trends and are not shown here. As shown in Fig. 5, systematic sampling does give the smallest average variance.

Although systematic sampling does capture the Hurst parameter and provide sampling results of small variance, we show in the next section that it provides very biased estimates of the real mean for a self-similar process. Due to this drawback, we then devise a new variation of systematic sampling to improve the accuracy of sampling results, while retaining all of its good properties. In the subsequent discussion, we will focus on systematic and simple random sampling, as stratified random sampling is a variation of systematic sampling.

V. Biased Systematic Sampling for Heavy-Tailed Traffic

In this section, we first show that both systematic sampling and simple random sampling fail to provide a good estimate of the actual mean for a self-similar process (e.g., Internet traffic). Then based on an important observation on self-similar processes (validated through experiments), we propose a new extension of systematic sampling to remedy the above deficiency. The dilemma here is that the major portion of a self-similar process consists of "small values," while a small portion of "extremely large values" contributes to the majority of the volume of the entire process (which in turn dramatically affects the mean of the process). Due to the massive amount of Internet traffic and the storage limitation, the sampling rate and hence the number of samples cannot be too large, but in order to capture the effect of these extremely large values (that occur not as often), one has to gather a large amount of samples. Similar observations have been made in the literature. For example, it has been reported in [7] that the steady-state behavior for self-similar workloads can be elusive, due to the fact that the average behavior depends on the presence of many small observations as well as a few large observations. The same observation has also been made in [21] on sampling Internet traffic, but no effective solution has been proposed to counter this problem.

A. Problem with Sampling a Self-Similar Process

By the central limit theorem (or the law of large numbers), the sampled mean can be used to approximate the real mean for any stationary process with finite mean and variance, as long as the sampling techniques are un-biased. It is well known that both simple random sampling and systematic sampling provide an un-biased estimator of the real mean for stationary processes with finite mean and variance as the number of samples goes to infinity. (In practice, a moderate number of samples suffice to provide a relatively good estimate of the real mean.) On the other hand, if the original process has infinite variance, e.g., a self-similar process, the law of large numbers does not hold, and the sampled mean approaches the real mean slowly, as the number of samples increases. As shown in [7], in order to achieve two-digit accuracy in the mean, the number of samples needed is up to 10^{22} for the case of $\alpha = 1.2$ (which corresponds to $H = 0.9$). Even for mild cases where $\alpha = 1.5$ ($H = 0.75$), still a million samples is required to achieve the desirable accuracy.

We carry out experiments to demonstrate the problem in the context of Internet traffic. In the experiments, we use the same set of synthetic and real Internet traffic traces used in Section IV. For synthetic trace, we change the sampling rate from 10^{-5} to 0.1, while for the real Internet trace, the sampling rate varies from 10^{-5} to 10^{-3}. (The reason why we used a smaller sampling rate is due to the large volume of Internet traces. In fact, a sampling rate of 10^{-3} is considered quite high, given the fact that tera-bytes of traffic is generated per day.) As shown in Fig. 6, in the case of synthetic traffic trace, the discrepancy between the real mean and the sampled mean (obtained even with a sampling rate of 0.1) is quite notable. The discrepancy becomes even more pronounced in the case of real Internet traces: the sampled mean obtained with a sampling rate of a 10^{-3} is approximately $\frac{2}{3}$ of the real mean, although in both cases the sampled mean increases steadily but slowly.

B. An Important Observation

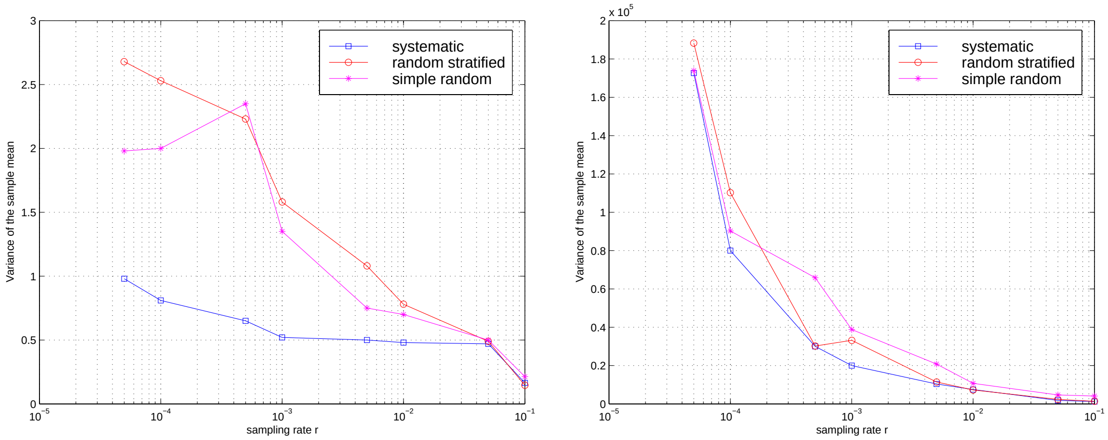
As mentioned above, the reason why the sampled mean is always far less than the real mean for a self-similar process is that the major portion of a self-similar process consists of "small values," while a small portion of "extremely large values," albeit occurring less often, contributes to the majority of the volume of the entire process. Without use of a sufficiently high sampling rate, the large values are less likely to be sampled and hence the sampled mean is always less than the real mean. If one could instrument the sampling method to capture these extremely large values, the discrepancy between the sampled mean and the real mean can be reduced.

To instrument a sampling method to capture extremely large values, we need to identify where they occur. For a self-similar process $f(t)$, we define another on-off process $q(t)$ as:

$$q(t) = \begin{cases} 1, & \text{if } f(t) > a_{th}, \\ 0, & \text{otherwise,} \end{cases} \tag{17}$$

where a_{th} is a constant approximately of the same order of magnitude as the mean of $f(t)$, $\bar{X}$. The process $q(t)$ consists of bursts of 1s and 0s. The length of the 1-burst period is a random variable (which we denote as B).

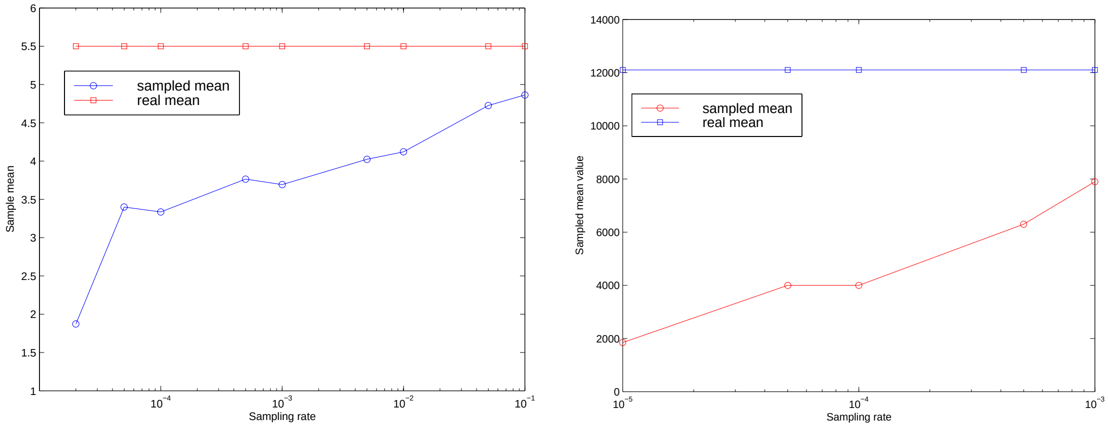
We conjecture that due to the self-similar properties of $f(t)$, B is heavy tailed. Intuitively this conjecture is made based on the fact that a self-similar process contains concentrated periods of high activity and low activity, and hence once the process goes beyond a_{th}, the time interval B during which it continuously remains above a_{th} is heavy-tailed. To validate the conjecture, we again carry out experiments on both the synthetic and real Internet traces introduced in Section IV. In the experiments, we set $a_{th} = \bar{X} \times \epsilon$, where ϵ varies from 0.3 to 1.5. For each fixed value of ϵ, we measure B and fit its CCDF to the most widely used heavy tailed distribution, the Pareto distribution. A line in a log-log plot indicates heavy-tailed behavior. Fig. 7 gives the results for $\epsilon = 0.5$. The fitted Pareto distribution has the shape parameter $\alpha = 1.3$ for the case of synthetic traces, while the shape parameter $\alpha = 1.65$ for the case of real Internet traffic traces. For different values



(a) result for synthetic data

(b) result for real Internet traffic

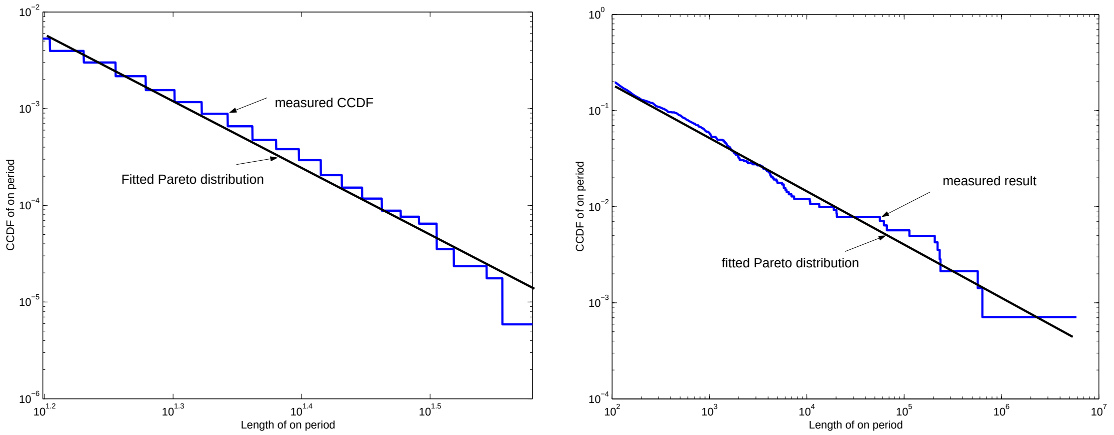
Fig. 5. The average variance of sampling results (under systematic, stratified random, and simple random sampling) on both synthetic and real Internet traffic.



(a) result for synthetic data

(b) result for real Internet traffic

Fig. 6. The sampled mean and the real mean of a self-similar process versus different sampling rates.



(a) Synthetic trace

(b) Real Internet trace

Fig. 7. The CCDF of the 1-burst period B for the case of $\epsilon = 0.5$, where ϵ determines the onset value, α, of the 1-burst period ($a_{th} = \bar{X} \times \epsilon$).

of ϵ, the value of α changes mildly from 1.2 to 1.8, but the heavy-tailed nature of B remains unchanged.

C. Detailed Description and Analysis of Biased Systematic Sampling

In this section, we propose, based on the observation made in Section V-B, a new variation of systematic sampling, called *biased systematic sampling* (BSS), that captures extremely large values more faithfully. Specifically, BSS is essentially systematic sampling with a sampling interval C, except that when a sample is greater than a threshold a_{th}, L extra samples are *evenly* taken in the current sampling interval C (i.e., the sampling interval for these extra samples is C/L). Among these extra samples, we only keep those that are greater than a_{th} (which we henceforth call *qualified samples*).

Analysis*:* The rational behind this design is as follows. A sample that is greater than a_{th} must fall in one of the 1-burst periods. Let the 1-burst period in which the sample falls be denoted as B. Suppose the sample is taken τ time units after the beginning of the 1-burst period B. We show given that B is heavily tailed, the probability that the next sample taken under BSS also exceeds a_{th} goes to 1 as τ goes to infinity. In other words, once a sample is taken with the value larger than a_{th}, it is highly possible that the values thereafter will still be larger than a_{th}. Specifically, such a probability can be expressed as

$$\begin{aligned} \wp(\tau) &= \Pr(q(\tau+1)=1|q(t)=1, 1\leq t\leq\tau) \\ &= 1-\frac{\Pr(B=\tau)}{\Pr(B\geq\tau)}. \end{aligned} \tag{18}$$

In the case that B is lightly tailed, e.g., the CCDF of B has an exponential tail, or $\Pr(B>x)\sim c_1e^{-c_2x}$, where c_1 and c_2 are two positive constants, Eq. 18 can be re-written as

$$\wp(\tau)\sim 1-\frac{c_1e^{-c_2\tau}-c_1e^{-c_2(\tau+1)}}{c_1e^{-c_2\tau}}=e^{-c_2}. \tag{19}$$

That is, in the case that B is lightly tailed, the probability that the samples taken exceed a_{th} does not become larger conditioning on the event that a sample has been identified to exceed a_{th}. In the case that B is heavily tailed, we have $\Pr(B>x)\sim cx^{-\alpha}$, where $1\leq\alpha\leq 2$ is the index of heavy-tailedness of the process, and hence

$$\wp(\tau)\sim 1-\frac{c\tau^{-\alpha}-c(\tau+1)^{-\alpha}}{c\tau^{-\alpha}}=(\frac{\tau}{\tau+1})^{\alpha}. \tag{20}$$

That is, $\wp(\tau)\to 1$, as $\tau\to\infty$. This implies given that B is heavily tailed, once a sample exceeds a_{th}, with a high probability the process will keep on large values. This lays the theoretical base for BSS, and ensures all the extra samples taken do increase the chance of capturing extremely large values.

A Rough Analysis of the Relationship between a_{th} and L*:* There are two important parameters used in BSS: the on-set threshold a_{th} and the number, L, of extra samples in each sampling interval C. In what follows, we perform an analysis on the relationship of these two parameters. In the analysis we assume that $f(t)$ follows a Pareto distribution with shape parameter α. This assumption is reasonable which has been demonstrated in[1]. Furthermore, we use the synthetic and real Internet traces as have been used in Section VI-B to show that $f(t)$ is heavy-tailed. The results are shown in Fig. 8. We show the CCDF of $f(t)$ and fit it to a Pareto distribution with shape parameter $\alpha=1.5$ and $\alpha=1.71$ for the synthetic and real traces respectively.

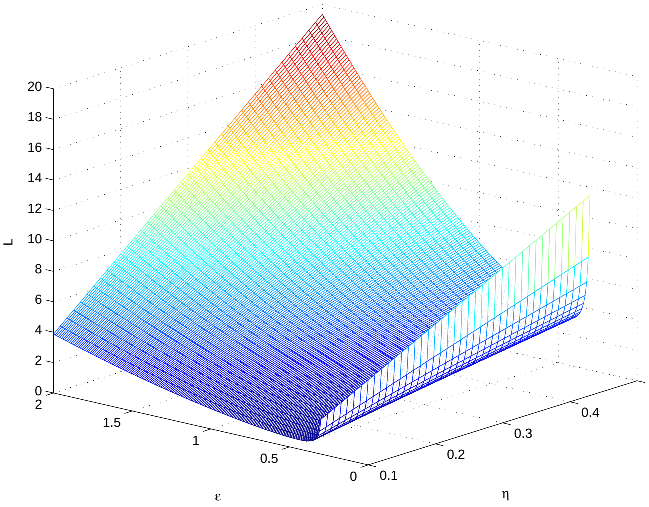


Fig. 9. The relationship among L, ϵ and η in BSS.

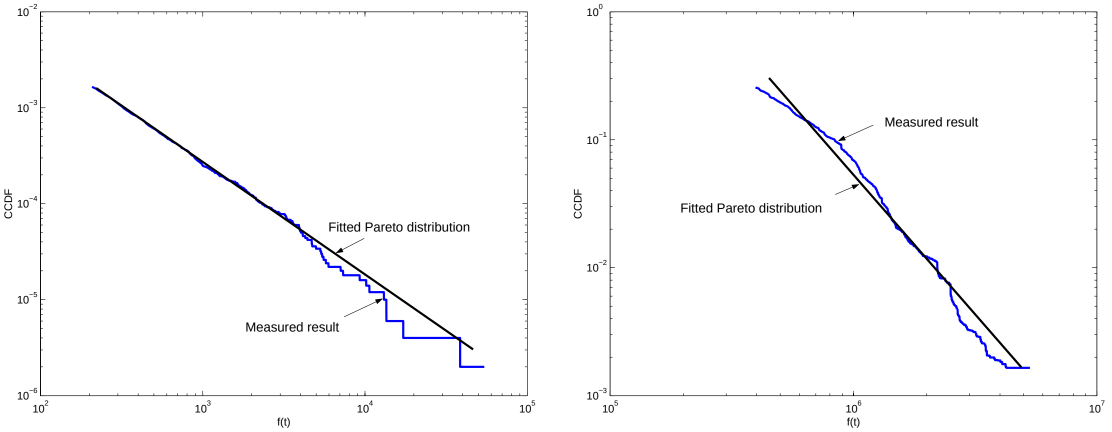
Let X_r and X_s denote the real mean and the sampled mean before the extra samples are taken, respectively, and let the difference, η, between X_r and X_s be defined as

$$\eta=1-\frac{X_s}{X_r}. \tag{21}$$

Suppose the traditional systematic sampling generates a total of N samples. Since the original process is self-similar, the sampled process is also self-similar with the same shape parameter α (Section **??**). Then each sample is greater than a_{th} with the probability $(\ell/a_{th})^{\alpha}$, where ℓ is the lowest value the original process can take. In other words, about $(\ell/a_{th})^{\alpha}\times N$ samples are above the threshold. After each of these samples is taken (and an on-set point is detected), L extra samples will be taken in each sampling interval of C in BSS. By a similar line of reasoning, we know that approximately $(\ell/a_{th})^{\alpha}\times L$ samples will be kept (*qualified samples* among all the extra samples taken), as they exceed the threshold a_{th}. The sampled mean of the set of *qualified samples* taken is *approximately* $\frac{a_{th}\alpha}{\alpha-1}$. As our objective is to make the mean of the entire set of $(N+L\cdot(\frac{\ell}{a_{th}})^{2\alpha}N)$ samples as close to the real mean $X_r=X_s(\frac{1}{1-\eta})$ as possible, we equate

$$\frac{N\cdot X_s+(\frac{\ell}{a_{th}})^{2\alpha}\cdot N\cdot\frac{a_{th}\alpha}{\alpha-1}\cdot L}{N+L\cdot(\frac{\ell}{a_{th}})^{2\alpha}\cdot N}=X_s(\frac{1}{1-\eta}), \tag{22}$$

The right hand side of Eq. 23 is the real mean X_r, while the left hand side is the new sample mean. After some algebraic



(a) Synthetic trace with $\alpha = 1.5$

(b) Real Internet trace with $\alpha = 1.71$.

Fig. 8. The CCDF of $f(t)$ and fitted Pareto distribution for synthetic and real Internet traces.

operations, we have

$$L = \frac{X_s \eta \ell^{-2\alpha} a_{th}^{2\alpha}}{(1-\eta)\frac{a_{th}\alpha}{\alpha-1} - X_s}. \tag{23}$$

The relationship among L, normalized a_{th} (ϵ)and η is shown in Fig. 9. As shown in the figure, L increases monotonically with both η and ϵ when ϵ is not small. L increases sharply when ϵ approaches 0. A large value of η implies the real mean has been extremely underestimated, and more samples of large values should be taken to amortize the bias. For a_{th}, although a larger value of ϵ (and hence larger a_{th}) implies that *qualified samples* taken are of large values, the probability in obtaining these large samples decrease faster ($\sim a_{th}^{-\alpha}, 1 \leq \alpha \leq 2$). As a result, the number of extra samples required increases accordingly. On the other hand, if ϵ is too small, the *qualified samples* are very likely to assume small values and help little in pushing the sampled mean toward the real mean value, therefore, more extra samples (larger L) are needed to get more larger *qualified samples*.

Next, we systematically study the setting of the parameters of L and a_{th} by carrying out an in-depth study of BSS

In-depth study of the BSS: Let us consider the expectation of the sampling results generated by BSS. To ease description, we define the following notations. We still use $f(t)$ to denote the traffic traces measured at some fixed time granularity, and it follows a Pareto distribution with shape parameter α. We define another two random variables:

$$Y = f(t), \quad \text{given } f(t) \geq a_{th}, \tag{24}$$

and

$$Z = f(t), \quad \text{given } f(t) \leq a_{th}. \tag{25}$$

Then it is straightforward to obtain the pdf of Y and Z to be:

$$p(y) = \frac{\alpha \ell^{\alpha} y^{-\alpha-1}}{(\ell/a_{th})^{\alpha}}, y \geq a_{th}, \tag{26}$$

and,

$$p(z) = \frac{\alpha \ell^{\alpha} z^{-\alpha-1}}{1-(\ell/a_{th})^{\alpha}}, \ell \leq z \leq a_{th}, \tag{27}$$

where ℓ is the smallest value $f(t)$ can take.

Let W denote the sampled process from BSS, then we have:

$$W = \begin{cases} Y, & \text{with probability } (\frac{\ell}{a_{th}})^{\alpha}, \\ Z, & \text{with probability } 1 - (\frac{\ell}{a_{th}})^{\alpha}, \\ Y', & \text{qualified samples.} \end{cases} \tag{28}$$

Given the number of systematic samples is N, the number of *qualified samples*, $L' = N \cdot L \cdot (\frac{\ell}{a_{th}})^{2*\alpha}$. The measured mean $\hat{W}$ is:

$$\hat{W} = \frac{1}{N+L'} \sum_{i=1}^{N+L'} w_i, \tag{29}$$

and,

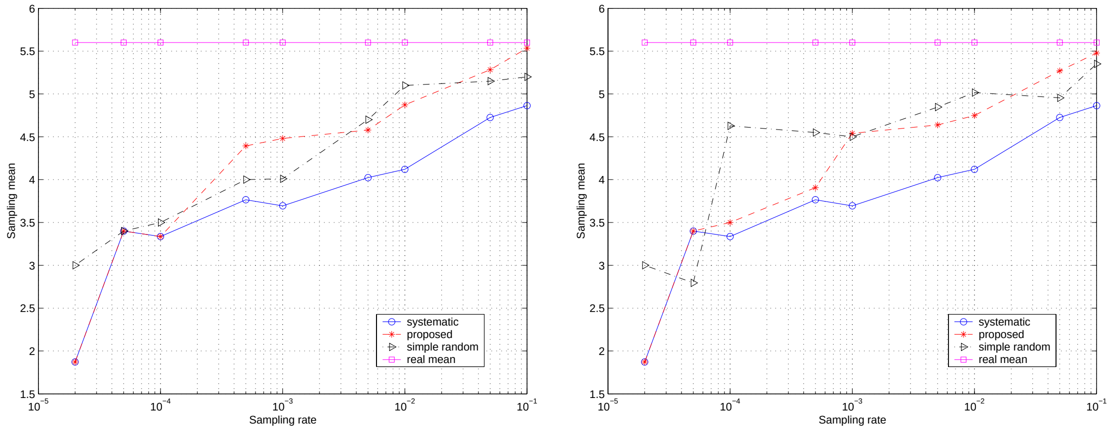
$$\begin{aligned} E(\hat{W}) &= \frac{1}{N+L'} \cdot \sum E(w_i) \\ &= \frac{\ell\alpha}{\alpha-1} \cdot \frac{1}{1+L(\ell/a_{th})^{2\alpha}} (\frac{\ell}{a_{th}})^{2*\alpha} L \frac{a_{th}\alpha}{\alpha-1} \\ &= X_r \cdot \xi, \end{aligned}$$

where $X_r = \frac{\ell\alpha}{\alpha-1}$ is the real mean, and

$$\xi = \frac{1}{1+L(\ell/a_{th})^{2\alpha}} (\frac{\ell}{a_{th}})^{2*\alpha} L \frac{a_{th}\alpha}{\alpha-1} \tag{30}$$

is called *bias parameter*. If $\xi = 1$, then BSS is an unbiased sampling method. ξ is determined by L and a_{th}, given ℓ and α. In Fig. 10 we show the relationship between ξ and L, normalized a_{th} (ϵ). We also draw the plane with $\xi = 1$ and the intersection of the two surfaces renders the set of parameters that make BSS unbiased.

One important observation is that, for each fixed L, there are two intersections along the ϵ axis. In other words, for a fixed L, there are two solutions to the equation $\xi(\epsilon) - 1 = 0$. To make it clearer, we show a slice of Fig. 10 when $L = 5$ in Fig. 11. We call the smaller one ϵ_1 and the larger one ϵ_2. One interesting finding is that ϵ_1 is almost the same for different L and $\epsilon_1 \approx \frac{\alpha-1}{\alpha}$ which is clearly shown in Fig. 10. But this solution is infeasible, since it causes $L < 0$ in Eq. (23). For the other solution ϵ_2, it increases with L.



(a) $L = 10,\epsilon = 2.55$ (b) $L = 8$, $\epsilon = 2.28$

Fig. 12. The sampled mean obtained by systematic sampling, simple random, and BSS for synthetic traces under two parameter settings. Both of them render $\xi = 1$.

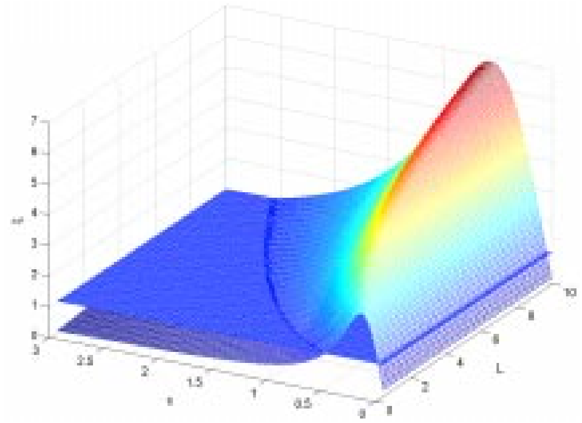


Fig. 10. The relationship among L, ϵ and ξ in BSS.

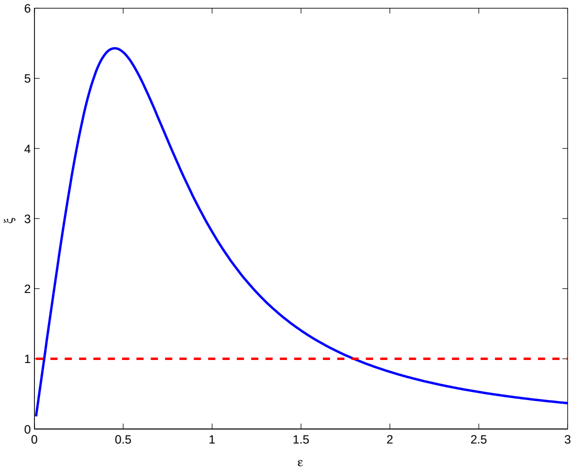


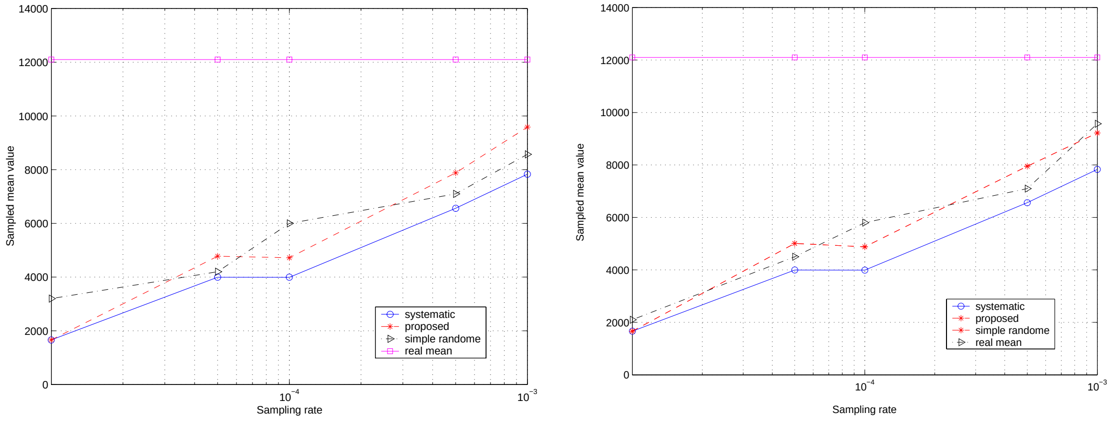
Fig. 11. A slice when $L = 5$, the relation between ξ and ϵ.

Next, we aim to answer the question: can BSS improve the performance of the systematic sampling method (providing more accurate estimate of the mean value) while keeping unbiased. To study this issue, we did experiments on both synthetic and real Internet traces. The results are shown in Fig. 12 and Fig. 13.

We can see that compared with systematic sampling, the unbiased BSS sampling produces exact the same results when the sampling rate is small. Under larger sampling rates, it improves a little, but still cannot reach the real mean quickly. The reason is that, in order to make BSS unbiased, for fixed L, a_{th} must be "large" enough so that the sampled results won't overshot. So, for small sampling rate, a_{th} is so high that few *qualified samples* can be obtained, and the sampled results resembles that of systematic sampling exactly. When the sampling rate increases, the chance in getting *qualified samples* increases. Due to the fact that the choosing of L and a_{th} makes $\xi = 1$, the sampled mean still cannot reach the real mean promptly.

Biased *BSS:* From above, we see that by carefully choosing L and a_{th} we can make BSS unbiased. Although this unbiased BSS improve the sampling performance when sampling rate is not too low, it suffers from the same problem as the systematic sampling. A remedy to this is to let BSS to be biased ($\xi > 1$) so that the real mean value can be reached quickly. The first hindrance we must hurdle is how to determine the value of ξ.

According to the definition of η $(= 1 - \frac{X_s}{X_r})$, in order to fill the gap between X_s and X_r, we can set $\xi = \frac{1}{1-\eta}$. If η is known, by setting $\xi = \frac{1}{1-\eta}$, we can choose appropriate L and a_{th} by intersecting the curve in Fig. 10 with a flat plane $\xi = \frac{1}{1-\eta}$. In Fig. 14 we show the contour of ξ. The labels on each contour line indicates the value of ξ. Therefore, a_{th} and L can be chosen according to the contour once the value of ξ is given. Since every point on the save contour line serves the goal, we can set one of the two parameters first and the other one can be herein determined.



(a) $L = 10,\epsilon = 1.809$ (b) $L = 8$, $\epsilon = 1.68$

Fig. 13. The sampled mean obtained by systematic sampling, simple random, and BSS for real traces under two parameter settings. Both of them render $\xi = 1$.

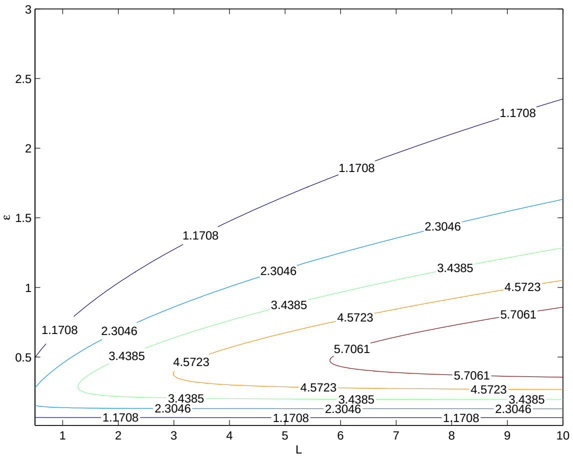


Fig. 14. The contour of ξ.

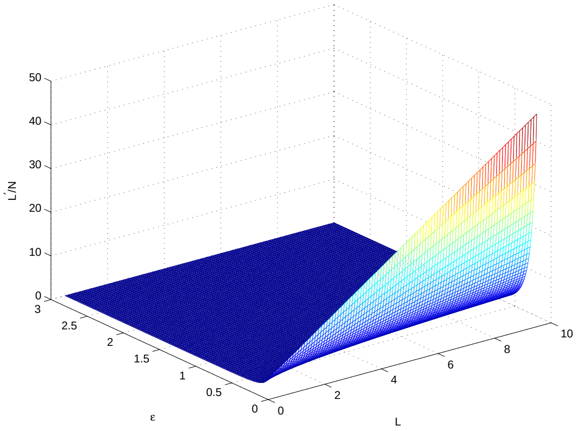


Fig. 15. The number of qualified samples.

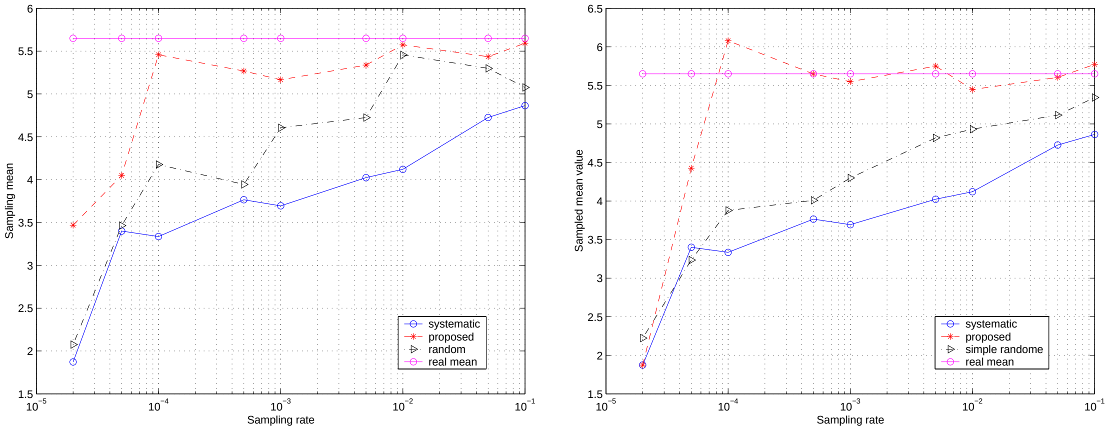
Before setting the parameters, we consider the number of *qualified samples*, L', which can be deemed as the cost of BSS (will be defined as the *overhead* in the next section). The optimal setting of L and a_{th} should make L' as small as possible. In Fig. 15 we show the relationship between $\frac{L'}{N}$ and L, ϵ according to the fact that $L' = N \cdot L \cdot (\frac{\ell}{a_{th}})^{2\alpha}$. One important observation is that to make L' small, we should avoid small ϵ and large L. We also notice that ϵ is more sensitive when it is small. Specifically, for $\epsilon < 0.5$, L' rockets.

Therefore, once η and X_r are known, given ϵ (hence a_{th}) or L, and $\xi = \frac{1}{1-\eta}$ we can faithfully determine the other one parameter to fulfill our objective. In Fig. 16 and 17, we show the result for both synthetic and real traces when one of ϵ and L is fixed and the other one is tuned [4].

Tuning L and a_{th} without knowledge of η: From above analysis, we know that the knot of selecting appropriate L and a_{th} lies in whether or not η and X_r can be obtained. In reality, sampling must be done online and η and X_r cannot be readily obtained. In what follows, we first discuss, given ϵ, the setting of a_{th} without the knowledge of X_r, then we set L by bridging the sampling rate r with η.

To tune a_{th}, we propose an on-line tuning scheme in determining the value of a_{th}. Before applying BSS, we first choose N_{pre} samples (called *pre-samples*) from which we obtain a rough estimate of the mean and assign a_{th} accordingly. After the operation of BSS commences, we set a_{th} as $a_{th} = E(Y_i) \times \epsilon$, where Y_i is the sampled mean of the sample set that contains all the samples till the ith sample (i.e., it includes the ith sample, the *pre-samples*, and all the *qualified samples* taken so far), and according to the above analysis, to save overhead (reduce the number of *qualified samples*), the normalized threshold ϵ should not be smaller than 0.5. Therefore, we set $\epsilon \in (1.0, 1.5)$ (in Section **??** we set $\epsilon = 1.0$). Note that when extra samples are chosen in a sampling interval C, we do not update a_{th} since whether or not to take extra samples in a sampling interval should be based on the same threshold. Only by the end of a sampling

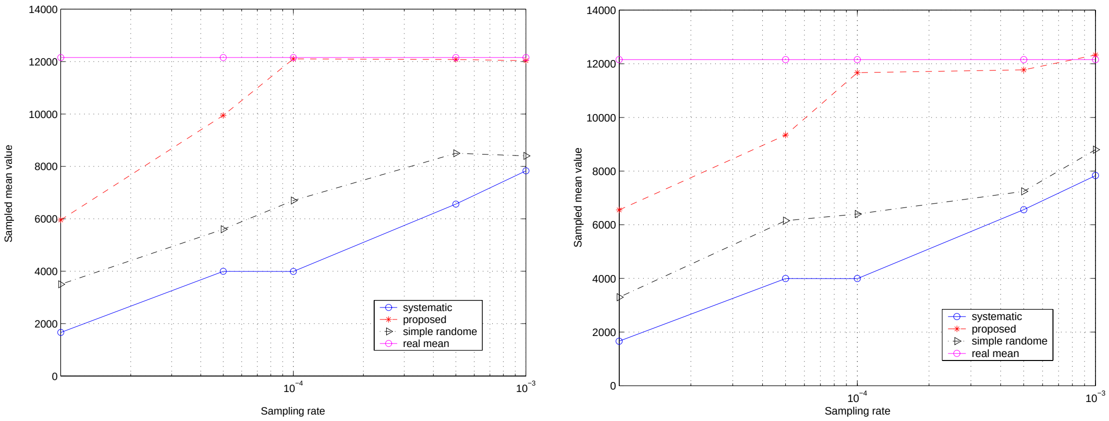
[4] The value of η and X_r are readily obtained since we have the entire traces.



(a) L is fixed to be 10

(b) ϵ is fixed to be 1.

Fig. 16. The sampled mean obtained by systematic sampling, simple random, and BSS for synthetic traces.



(a) L is fixed to be 30

(b) ϵ is fixed to be 1.

Fig. 17. The sampled mean obtained by systematic sampling, simple random, and BSS for real traces.

interval when the next *normal* sample is taken, we update a_{th} as $E(Y_i) \times \epsilon$.

Given a_{th}, in order to set appropriate L, η is indispensable. Although we cannot obtain the exact value of η, we can estimate it from the sampling rate r. The estimation procedure is shown as follows.

Let N be the total number of systematic sampling. We have:

$$X_s = \frac{1}{N}\sum_{i=1}^{N} f(t_i), \tag{31}$$

which is the sample mean. From [21] (Chapter 3), if we define:

$$V_n = N^{1-1/\alpha}(X_s - X_r), \tag{32}$$

where X_r is the real mean. Then

$$V_n \rightarrow \varphi_\alpha, \quad \text{in distribution}, \tag{33}$$

where φ_α is an α-*stable* distribution. In other words, V_n converges in distribution for large N. Thus we have another way to convey this: $|X_s - X_r| \sim N^{1/\alpha - 1}$, then,

$$\eta = \frac{|X_r - X_s|}{X_r} \sim \frac{N^{1/\alpha-1}}{X_r}. \tag{34}$$

Let r and N_t be the sampling rate and total number of points in the original process, then $N = N_t * r$. So we have:

$$\eta \sim C_s \cdot r^{1/\alpha - 1}, \tag{35}$$

where $C_s = \frac{N_t^{1/\alpha-1}}{X_r}$ is a constant less than 1 for $1 \leq \alpha \leq 2$. In reality, C_s may fluctuate mildly for different sampling rates. From our experimental study, we find that for the synthetic traces ($\alpha = 1.5$), $C_s \in (0.25, 0.35)$ while for the real traces ($\alpha = 1.66$), $C_s \in (0.2, 0.3)$.

Therefore, given a systematic sampling rate r, Eq. (35) can be applied to estimate η. With the knowledge of η, we can obtain $\xi = \frac{1}{\eta}$. At last, by plugging both ξ and a_{th} in Eq. 30, the value of L can be easily obtained. In Section **??**, ϵ is pre-set to be 1 and we apply this procedure in setting the parameters and the experiments generate relatively good results.

VI. Performance Evaluation

To evaluate the performance of BSS, we have carried out several sets of experiments on both synthetic and real Internet traces. As the increase in the accuracy of the sampled mean in BSS is obtained at the cost of sampling more "biased" samples of larger values, we use the following three metrics to evaluate BSS: (1) the sampled mean (accuracy); (2) the sampling overhead, defined as the ratio of the number of *qualified samples* to the number of samples taken by systematic sampling; and (3) the efficiency e, defined as $e = \frac{1-\eta}{log(N_t)}$ and N_t is the total number of samples (including both the samples normally taken in systematic sampling and the *qualified samples* taken in BSS). In addition to the above three metrics, we also verify whether the sampled process has the same Hurst parameter as the original process and calculate its average variance. The performance evaluation is made by comparing BSS against systematic and simple random sampling. As stratified random sampling is a variation of systematic sampling and yields similar performance as the latter, we do not include it in the comparison study.

A. Performance w.r.t. Sampled Mean, Overhead and Efficiency

We use the same traces as in Section IV. For synthetic traces, we set the shape parameter of the on/off periods to be $\alpha \in (1.2, 1.6)$. Figures 18–19 give the sampled mean obtained by systematic sampling, simple random, and BSS ((a)), and the sampling overhead incurred in BSS ((b)) for both the synthetic traces and real Internet traces. Note that the result shown in Fig. 18 is for the synthetic trace with $\alpha = 1.3$ and mean value 5.68 kbytes/second, while that in Fig. 19 is for the Internet trace with the real mean rate 1.21×10^4 bytes/second and the (measured) Hurst parameter 0.62. (Results for the other traces exhibit similar trends and hence are not shown here.) As shown in Fig. 18 (a), BSS generates much more accurate sampled means than the other two sampling techniques. The performance improvement is especially pronounced when the sampling rate is as small as 10^{-4}. As shown in Fig. 18 (b), the overhead is around 0.2, while $1 - \eta$ (Section V-C) is 0.922 for BSS and 0.66 and 0.81 for systematic sampling and simple random sampling, respectively. Similar conclusions can be made in Fig. 19, except that the sampling overhead is around 0.3.

Fig. 20 compares BSS against systematic sampling and simple random sampling with respect to the efficiency e for synthetic traces. BSS achieves higher efficiency than the other two sampling techniques. The average e for BSS is 0.37, while that for systematic and simple random sampling is 0.26 and 0.3, respectively, i.e., BSS achieves a performance gain of 42% and 23%, respectively, as compared to systematic and simple random sampling.

B. Performance w.r.t. Hurst Parameter and Average Variance

To verify whether or not BSS captures the Hurst parameter accurately, we use synthetic traffic with β ($H = \frac{2-\beta}{2}$) varying from 0.1 to 0.8 and give the result in Fig. 21. The Hurst parameter of the synthetic traces is calculated using a wavelet based tool provided by Abry *et al.* [22]. As shown in Fig. 21, the sampled process has the same value of β (and hence the same Hurst parameter) as the original process. This is not surprising, as BSS is a variation of static systematic sampling and the extra samples taken in each sampling interval are also taken in a systematic sampling fashion in each interval C. As a result, the sampled process generated by BSS keeps the same autocorrelation function as that generated by systematic sampling (which in turns is the same as that of the original process, Section III-A). Finally, Fig. 22 gives the the average variances of BSS and systematic sampling for both synthetic and real Internet traces. As shown in the figure, the average variances of these two methods almost overlap completely. This is not surprising due to the same reason stated above.

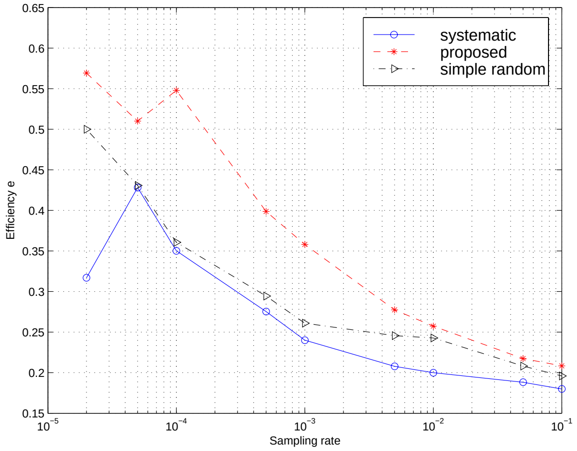


Fig. 20. The efficiency of systematic sampling, simple random, and BSS for synthetic traffic.

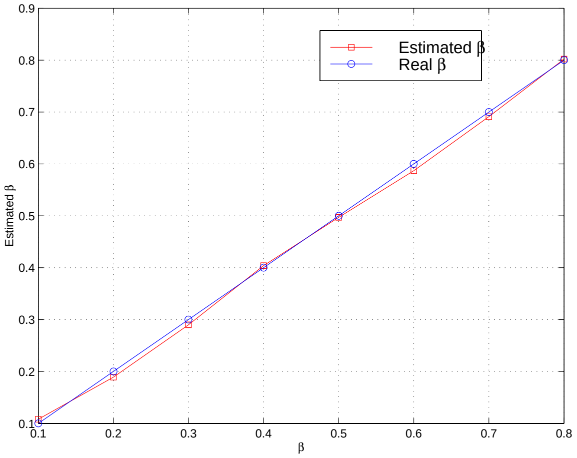
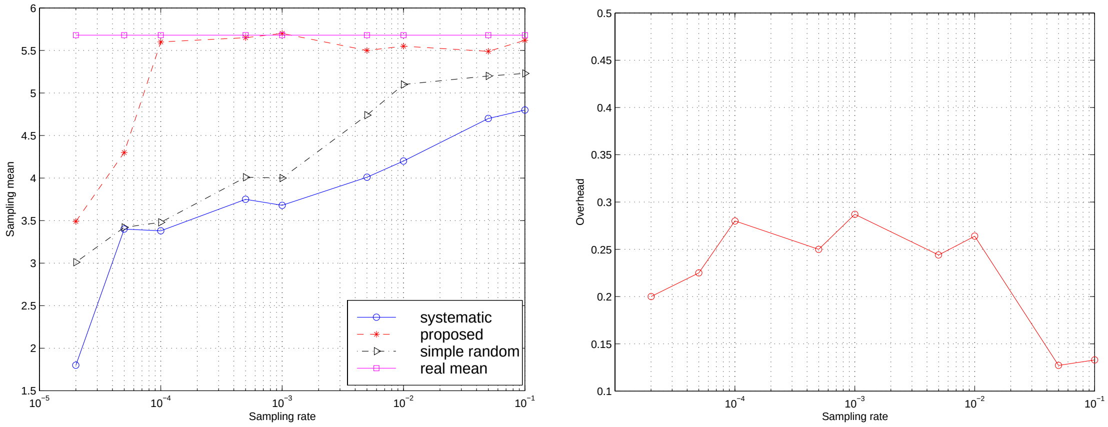


Fig. 21. The β values of the sampled process generated by BSS and and the real process.

VII. Conclusion

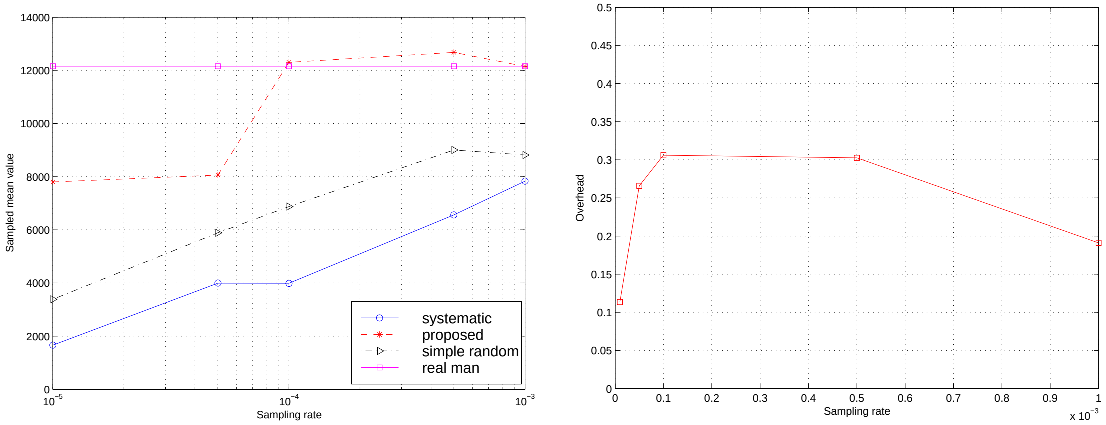
In this paper, we have investigated several important issues in employing sampling techniques for measuring Internet traffic. We show that while all three sampling techniques can accurately capture the Hurst parameter (second order statistics) of Internet traffic, they fail to capture the mean (first order statistics) faithfully, due to the bursty nature of Internet traffic. We also show that static systematic sampling renders the smallest variation of sampling results in different instances of sampling (i.e., it gives sampling results of high fidelity). Based



(a) Sampled mean

(b) Sampling overhead

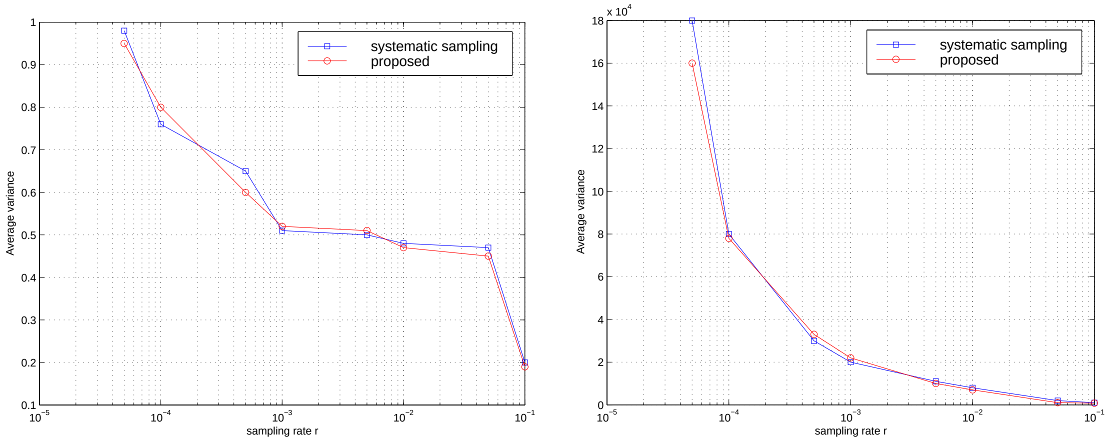
Fig. 18. The sampled mean obtained by systematic sampling, simple random, and BSS ((a)), and and the sampling overhead incurred in BSS ((b)) for synthetic traces.



(a) Sampled mean

(b) Sampling overhead

Fig. 19. The sampled mean obtained by systematic sampling, simple random, and BSS ((a)), and and the sampling overhead incurred in BSS ((b)) for real Internet traces.



(a) Average variance for synthetic traces

(b) Average variance for real Internet traces

Fig. 22. The average variances of BSS and systematic sampling.

on an important observation, we then devise a new variation of systematic sampling, called *biased systematic sampling (BSS)*, that gives much more accurate estimates of the mean, while keeping the sampling overhead low. Both the analysis on the three sampling techniques and the evaluation of BSS are performed on both synthetic and real Internet traffic traces. The performance evaluation shows that BSS gives a performance improvement of 42% and 23% (in terms of efficiency) as compared to static systematic and simple random sampling.

An important lesson learned from the work is that although un-biased sampling methods are usually preferred for processes with finite means and variances (where the law of large numbers guarantees that the sampled mean approaches the real mean exponentially fast as the number of samples increases), it may not be the case for a process with an infinite variance (e.g., self-similar Internet traffic with the Hurst parameter larger than 0.5). Due to the heavy-tailedness inherited in the self-similar process, the speed for the sampled mean to converge to the real mean is extremely slow, and therefore unbased sampling techniques often render un-satisfactory results. In this case, a biased sampling method is actually desirable. By biasing toward the *large* values of the process, one can reduce the discrepancy between the sampled mean and the real mean. In this paper we make a case where a biased sampling method outperforms un-biased ones.

We have also identified several new research directions. First, we will attempt to prove the conjecture made in Section V-B that the length of 1-burst periods in self-similar processes is heavy tailed. Second, although the methods presented in Section V-C for tuning the parameters, a_{th} and L, of BSS are devised based on analytical reasoning, they are engineering-oriented and may not render the optimal setting. We would like to study how to optimally set these parameters so as to strike a balance between the sampling overhead and the accuracy thus achieved.

REFERENCES

[1] J. Cao, W. S. Cleveland, D. Lin and D. X. Sun. The Effect of Statistical Multiplexing on Internet Packet Traffic: Theory and Empirical Study. *Bell Labs Tech. Report*, 2001.

[2] B. Y. Choi, J. Park and Z. L. Zhang. Adaptive Random Sampling for Load Change Detection. *ACM SIGMETRICS*, 2002, (Extended Abstract).

[3] K. C. Claffy, G. C. Polyzos and H. W. Braun. Application of sampling methodologies to network traffic characterization. In *Proc. ACM SIGCOMM'93*, September, 1993.

[4] Cisco netflow. *http://www.cisco.com/warp/public/732/Tech/netflow*.

[5] W. G. Cochran. Sampling Techniques. John Wiley & Sons, Inc., 1977

[6] I. Cozzani and S. Giordano. A Measurement based QoS evaluation. *IEEE SICON'98*, June, 1998.

[7] M. E. Crovella and L. Lipsky. Long-lasting Transient Conditions in Simulations with Heavy-tailed Workloads. *Proc. of the 1997 Winter Simulation Conference*.

[8] N. G. Duffield and M. Grossglauser. Trajectory sampling for direct traffic observation. in *Proc. ACM SIGCOMM'00*, pp. 271-282, August, 2000.

[9] N. Duffield, C. Lund and M. Thorup. Charging from sampled network usage. in *SIGCOMM Internet Measurement Workshop*, November, 2001.

[10] N. G. Duffield, C. Lund and M. Thorup. Properties and Prediction of Flow Statistics from Sampled Packet Streams. *ACM SIGCOMM Internet Measurement Workshop,* November, 2002.

[11] N. G. Duffield, C. Lund and M. Thorup. Estimating Flow Distributions from Sampled Flow Statistics. In *Proc. ACM SIGCOMM'03*, August, Germany, 2003.

[12] A. Erramilli, O. Narayan, and W. Willinger. Experimental queuing analysis with long-range dependent traffic. *IEEE/ACM Transactions on Networking*, April 1996.

[13] C. Estan and G. Varghese. New Directions in Traffic Measurement and Accounting. *ACM SIGCOM Internet Measurement Workshop*, 2001.

[14] A. Feldmann, A. Greenberg, C. Lund, N. Reingold, J. Rexford and F. True. Deriving traffic demands for operational ip networks: Methodology and experience. In *Proc. ACM SIGCOMM'00*, pp. 257-270, August, 2000.

[15] N. Hohn and D. Veitch. Inverting Sampled Traffic. *Proc. ACM IMC'03*, October 27-29, 2003, Florida, USA.

[16] Internet Protocol Flow Information eXport (IPFIX). IETF Working Group, *http://ipfix.doit.wisc.edu*.

[17] Packet Sampling (PASAMP). IETF working group, *http://ops.ietf.org/psamp/*.

[18] *http://cm.bell-labs.com/cm/ms/departments/sia/InternetTraffic/S-Net/*.

[19] R. Mahajan, S. M. Bellovin, S. Floyd, J. Ioannidis, V. Paxson and S. Shenker. Controlling high bandwidth aggregates in the network. http://www.aciri.org/pushback/, July, 2001.

[20] W. E. Leland, M. S. Taqqu, W. Willinger and D. V. Wilson. On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, February 1994.

[21] K. Park, W. Willinger. Self-similar network traffic and performance evaluation. Ch. 21, Wiley-Interscience.

[22] Roughan, Veitch and Abry. Real-time estimation of the parameters of long-range dependence (extended version). *IEEE/ACM Transactions on Networking*, vol.8, no.4, pp. 467-478, August 2000.

[23] 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 *Proc. ACM SIGCOMM'97*, pp. 149–157, 1997.

[24] W. Willinger, V. Paxson, and M. S. Taqqu. Self-similarity and heavy tails: structural modeling of network traffic. In R. Adler, R. Feldman, and M.S. Taqqu, editors, A Practical Guide to Heavy Tails: Statistical Techniques and Applications, Birkhauser, Boston, 1998.