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

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

Abstract

Techniques for sampling Internet traffic are very important to understand the traffic characteristics of the Internet [14, 8]. In spite of all the research efforts on packet sampling, none has taken into account of self-similarity of Internet traffic in 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. 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. Our performance study shows that BSS gives a performance improvement of 40% and 20% (in terms of efficiency) as compared to static systematic and simple random sampling.*

1 Introduction

Techniques for sampling Internet traffic are very important to understand traffic characteristics of the Internet [14, 8]. If the sampled results faithfully represent Internet traffic, they can be utilized on a short-term basis for hot spot and DDoS attack detection [19], or on a long-term basis for traffic engineering [14] and accounting [9]. As such, 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 scalability. Logging/inspecting each individual packet for an origin-destination (OD) flow or sampling at a very high rate is obviously not scalable, due to the large volume of samples to be kept. On the other hand, if the sampling rate is inadequately low, the sampled results may not characterize faithfully actual traffic. What makes the problem even more difficult is the bursty nature of Internet traffic. As indicated in a number of recent empirical studies of traffic measurement from a variety of operational packet networks [20, 12, 22, 23], 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 activities (peaks) and low activities (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 sampling at a high rate increases the memory requirements for 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] focused on the issue of reducing the bandwidth needed for transmitting traffic measurements to a back-office system for later analysis, and devised a size-dependent flow sampling method. Choi *et al.* [2] proposed the notion of adjusting the sampling density upon detection of traffic changes in order to meet certain constraints on the estimation accuracy. 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 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 variance of the sampling results. 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 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 the sampled result in the ith sampling process

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

(i.e., the ith experiment). The average variance is then defined as $E(V) = E[E[(X_i - \bar{X})^2]]$, where the inner expectation is taken over all the samples in one sampling process, and the outer expectation is taken over all the sampling instances (For example, different starting sampling points in systematic sampling give different sampling instances). 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 processes 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 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 accurately estimate the mean for self-similar Internet traffic, especially when the sampling rate is small. We then propose, based on an important observation, a revised version of systematic sampling, 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 systematic sampling, it retains all the advantages of the latter.

Both the verification and validation in **T1**–**T3**, and the evaluation of BSS, are 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.

Note that the traffic process $f(t)$ considered in the paper is rather general, and can be either an individual OD-flow or the aggregate of several/all OD-flows that traverse a router. After $f(t)$ is specified, the proposed sampling technique can be used to, for example, estimate the mean of the aggregate traffic of several (selected) OD flows between the west and east coasts in the States. Note also that although it is feasible to log each and every individual packet and record the entire flow time series $f(t)$, the process of collecting such an enormous amount of samples can only be carried out at a small number of ISP routers that are equipped with DAG packet collection cards and large memory. The large amount of data is then analyzed off-line to better understand the traffic characteristics. Sampling remains as an effective and economical technique to on-line collect/estimate parameters that characterize the traffic.

The rest of the paper is organized as follows. After providing the background material in Section 2, we analyze in Section 3 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 retain the second order statistics (and hence Hurst parameter) of the original process. Then, we compare in Section 4 the average variance of the sampling results obtained by the three techniques. Following that, we demonstrate in Section 5 with both synthesized and real Internet traces that all three techniques fail to capture the real mean of Internet traffic, and elaborate on BSS. Finally we present in Section 6 a performance study (based on both synthesized and real traces). The paper concludes with Section 7.

2 Background

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

2.1 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 mentioned in Section 1, the traffic process can be an individual OD-flow or an aggregation of several/all OD-flows that traverse a router. To define self-similarity, we further define the aggregated series $f^{(m)}(\tau)$ as

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

$f^{(m)}(\tau)$ 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:

$$R(\tau) \sim \text{const} \cdot \tau^{-\beta}, \quad (2)$$
$$R^{(m)}(\tau) \sim R(\tau), \quad (3)$$

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 $H = 1 - \beta/2$.

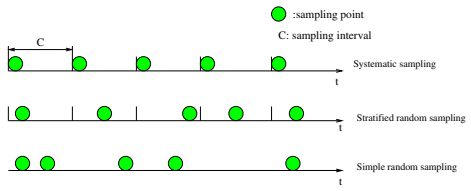


Figure 1. An illustration of the three sampling techniques.

2.2 Three Commonly Used Sampling Techniques

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.

3 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. Then we derive a *SNC* that a sampling technique has to satisfy in order to retain the autocorrelation structure of the original process.

3.1 Systematic Sampling

Let $f(t)$ and $g(t)$ denote, respectively, the original and sampled processes. Also, let $R_f(\tau)$ and $R_g(\tau)$ denote the autocorrelation function of $f(t)$ and $g(t)$, and $F(t)$ and $G(t)$ the CDF of $f(t)$ and $g(t)$, and H_f and H_g the Hurst parameter of $f(t)$ and $g(t)$, respectively. Without loss of generality, let t be discretized to be integers: $0, 1, 2, 3....$ For systematic sampling, let C be the sampling interval. Then we have[2]

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

Also,

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

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.

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

3.2 Stratified Random Sampling

Recall that in stratified random sampling, the time axis is divided into interval of length C, and one sample is randomly taken in each interval. Using the same notation as in Section 3.1, 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 interval 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 $\tau' = \frac{\tau_1 - \tau_2}{C}$.

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.

3.3 Simple Random Sampling

In simple random sampling, N samples are randomly selected from the entire population of M samples. That is, with probability $\rho = N/M$ 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

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

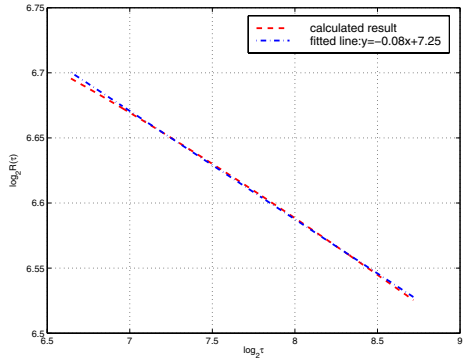


Figure 2. The calculated result of $R_g(\tau)$ is fit into the line $log_2 R_g(\tau) = -0.08 log_2 \tau + 7.25$; the real value of β is 0.1.

we have

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

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

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

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. As shown in Fig. 2, 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 find that the estimated value of β matches the real value very well, i.e., the sampled process can keep the autocorrelation structure of the original process.

3.4 Sufficient and Necessary Condition for Accurately Capturing the Hurst Parameter

In Section 3.1–3.3, 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 the 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 fully characterized 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 the function h for stratified random sampling method is

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

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

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

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

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

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*

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

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 $\tau \to \infty$, and hence the conclusion.*

4 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 used in each instance. Here by "instance," we mean an experiment carried out to take samples for a period of time. 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 in Section 1 as: 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,*

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

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 $R(\tau) \sim const \cdot \tau^{-\beta}$, it is easy to check that the condition in Eq. (16) holds.

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 tends to infinity as time goes to infinity. However, since we only consider the process in finite time periods in practice, we conjecture the above condition is still valid. To validate the conjecture, we carry out experiments on both synthetic and real Internet traffic and measure the average variance of sampling results (under the three techniques). Specifically, we generate in *ns-2* self-similar traffic with the 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 use 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. 3 gives the average variance of sampling results under the three sampling techniques. Note that Fig. 3 (b) gives the result for the real Internet trace set with 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. 3, 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 in Section 5 that it provides very biased estimates of the real mean for a self-similar process. To remedy this deficiency, we then devise a new variation of systematic sampling to improve the accuracy of sampling results, while retaining 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.

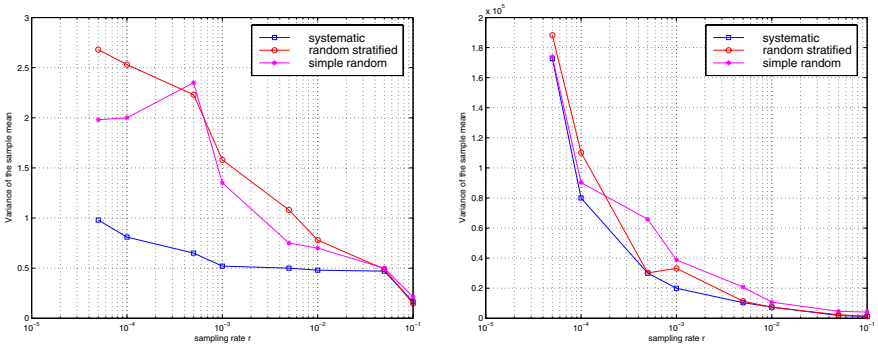
5 Biased Systematic Sampling for Heavy-Tailed Traffic

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.

5.1 Problem with Sampling a Self-Similar Process

It is well known that as the number of samples goes to infinity, both simple random sampling and systematic sampling provide an un-biased estimator of the real mean for stationary processes with finite mean and variance. (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 cannot be readily applied, and the sampled mean approaches the real mean slowly as the number of samples increases. This is because while a major portion of a self-similar process consists of "small values," a small portion of "extremely large values" contribute to the majority volume of the entire process. As these extremely large values do not occur very often, it is difficult to capture them (unless the process is sampled at an extremely high rate) and yet their effect on the estimate of the mean is enormously large. Similar observations have been made in the literatures [7, 21], but no effective solution has been proposed to counter this problem. In particular, 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 4. We change the sampling rate from 10^{-5} to 0.1 for synthetic traces, and from 10^{-5} to 10^{-3} for real Internet traces. (The reason why we used a smaller sampling rate for real Internet traces is due to the large volume of Internet traces. In fact, a sampling rate of 10^{-3} is considered high, given the fact that tera-bytes of traffic are generated per day.) As shown in Fig. 4, in the case of synthetic traffic traces, the discrepancy between the real mean and the sampled mean (obtained even with a sampling rate of 0.1) is quite notable. The discrepancy



(a) result for synthetic data (b) result for real Internet traffic

Figure 3. The average variance of sampling results (under systematic, stratified random, and simple random sampling) on both synthetic and real Internet traffic.

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.

5.2 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 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, the first step is 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 $f(t)$, X_r. 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 activities and low activities, 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 4. In the experiments, we set $a_{th} = X_r \times \epsilon$, where ϵ is called the normalized threshold and varies from 0.5 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. Fig. 5 gives the results for $\epsilon = 1.0$. The fitted Pareto distribution has the shape parameter $\alpha = 1.3$ for the case of synthetic traces, and $\alpha = 1.65$ for the case of real Internet traffic traces. For different values of ϵ, the value of α changes mildly from 1.2 to 1.8, but the heavy-tailed nature of B remains unchanged.

5.3 Detailed Description and Analysis of Biased Systematic Sampling

In this section, we propose, based on the observation made in Section 5.2 (i.e., B is heavy tailed), 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 taken with the value 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*).

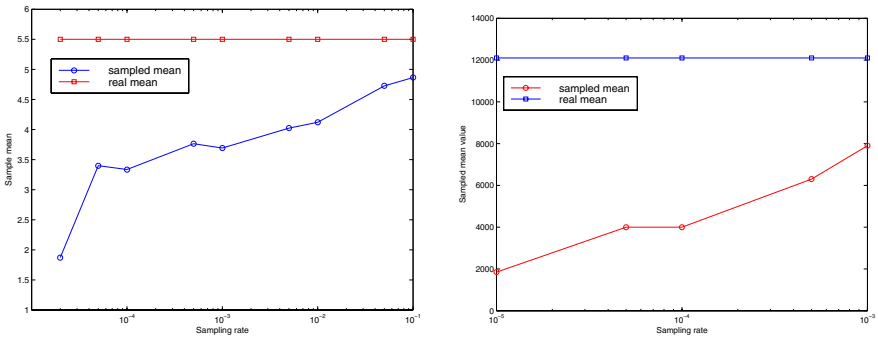
Analysis The rational behind BSS is as follows. A sample that is greater than a_{th} must fall in one of the 1-burst periods. Let the 1-burst period in which the sample falls be denoted as B. Suppose the sample is taken τ time units after the beginning of the 1-burst period B. We show that given B is 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

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

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

$$\wp(\tau) \sim 1 - \frac{c_1 e^{-c_2 \tau} - c_1 e^{-c_2(\tau+1)}}{c_1 e^{-c_2 \tau}} = e^{-c_2}. \tag{19}$$

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}. On the



(a) result for synthetic data (b) result for real Internet traffic

Figure 4. The sampled mean and the real mean of a self-similar process versus different sampling rates.

other hand, if 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

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

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 — very likely the extra samples taken contain extremely large values.

Parameters setting in BSS To complete the design, we have to determine the values of two important parameters used in BSS: the on-set threshold a_{th} and the number, L, of extra samples in each sampling interval C. For clarity of description, we assume that the original process $f(t)$ follows a Pareto distribution with shape parameter α and scale parameter ℓ. (This assumption has been corroborated by both the study in [1] and our own study [15]. For example, we verified that the CCDF of $f(t)$ can be fit into a Pareto distribution with shape parameter $\alpha = 1.5$ and $\alpha = 1.71$ for synthetic and real traces, respectively.)

Let X_r, X_s, and X_{bss} denote, respectively, the real mean, the sampled mean under systematic sampling, and the sampled mean under BSS. By the property of the heavy tail distribution, $X_r = \frac{\ell\alpha}{\alpha-1}$, where ℓ is the lowest value the original process can take. Also, let the difference, η, between X_r and X_s be defined as

$$\eta = 1 - \frac{X_s}{X_r}. \qquad (21)$$

Since the original process is self-similar, the sampled process is also self-similar with the same shape parameter α (Section 3). As a result, the probability that a sample is greater than a_{th} is $(\ell/a_{th})^{\alpha}$, where ℓ is the lowest value the original process can take. In other words, approximately $(\ell/a_{th})^{\alpha} \times N$ samples exceeds the on-set threshold, and trigger the operation of taking L extra samples. By a similar line of reasoning, approximately $(\ell/a_{th})^{\alpha} \times L$ samples (out of the L extra samples) exceed the threshold a_{th}, and are classified as (*qualified* among all the extra samples taken). The sampled mean of the set of *qualified samples* taken is *approximately* $\frac{a_{th}\alpha}{\alpha-1}$.

Now the sampled mean, X_{bss}, under BSS can be expressed as

$$X_{bss} = \frac{N \cdot X_r + (\frac{\ell}{a_{th}})^{2\alpha} \cdot N \cdot \frac{a_{th}\alpha}{\alpha-1} \cdot L}{N + L \cdot (\frac{\ell}{a_{th}})^{2\alpha} \cdot N}. \qquad (22)$$

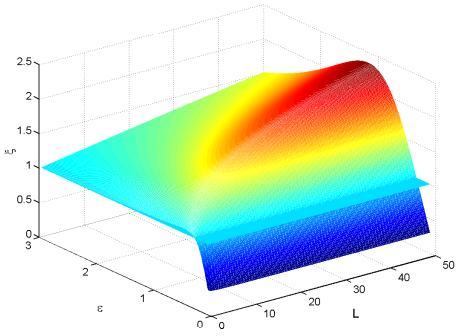


Figure 6. The relationship among L, ϵ and ξ in BSS.

Recall that $X_r = \frac{\ell\alpha}{\alpha-1}$ is the real mean, and hence Eq. (22) can be rewritten as

$$\begin{aligned} X_{bss} &= \frac{\ell\alpha}{\alpha-1} \cdot \frac{1 + L(\ell/a_{th})^{2\alpha-1}}{1 + L(\ell/a_{th})^{2\alpha}} \\ &\triangleq X_r \cdot \xi, \end{aligned}$$

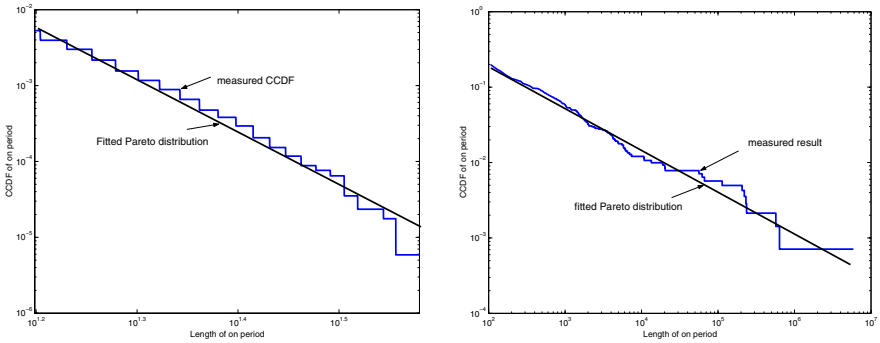
where

$$\xi = \frac{1 + L(\ell/a_{th})^{2\alpha-1}}{1 + L(\ell/a_{th})^{2\alpha}} \qquad (23)$$

is the *bias parameter*. If $\xi = 1$, then BSS is an unbiased sampling method. Given the values of ℓ and α, ξ is determined by L and a_{th}. In Fig. 6 we show the relationship between ξ L, and the normalized threshold $\epsilon = a_{th}/X_r$. The intersection of the plane of $\xi = 1$ and the surface of ξ gives the set of parameters that makes BSS unbiased. In particular, given any fixed value of L, there exists only one intersection point along the ϵ axis: $\epsilon = \frac{\alpha-1}{\alpha}$. This solution is, however, not feasible in practice, because $\epsilon = \frac{\alpha-1}{\alpha}$ for $1 \leq \alpha \leq 1.2$ is very small and suggests extra samples be taken in virtually every sampling interval. This translates to sampling at a very high rate.

A remedy to this problem is to allow BSS to be biased ($\xi > 1$). A key step along this direction is to determine the value of ξ. Note that $X_{bss} = X_r \cdot \xi$ (Eq. (23)) holds only when $N \to \infty$. In the case that N is finite, $X_{bss} \approx X_s \cdot \xi$. In order to have X_{bss} approach X_r in the finite-N case, we set $X_s \cdot \xi = X_r$ or $\xi = \frac{1}{1-\eta}$.

Tuning L and a_{th} in the case that η is known: If η is known, we can calculate $\xi = \frac{1}{1-\eta}$ and select appropriate values of L and a_{th} by



(a) Synthetic trace (b) Real Internet trace

Figure 5. The CCDF of the 1-burst period B for the case of $\epsilon = 1.0$, where ϵ determines the onset value, α, of the 1-burst period ($a_{th} = X_r \times \epsilon$).

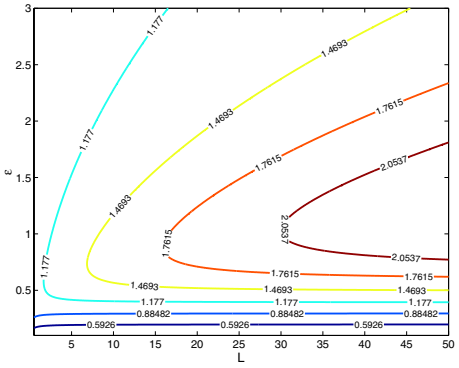


Figure 7. The contour of ξ.

intersecting the surface of ξ in Fig. 6 with the plane $\xi = \frac{1}{1-\eta}$. Fig. 7 gives the contour of ξ. The label on each contour curve indicates the value of ξ. Since all the points on the same contour curve render the same value of ξ, we can set one of the two parameters first and determine the other one accordingly.

Now to further determine the values of L and a_{th}, we take into account of the number of *qualified samples*, $N \cdot L \cdot (\frac{\ell}{a_{th}})^{2\alpha}$. This can be considered as the *overhead* in BSS. L and a_{th} should be so chosen that the number of qualified samples is as few as possible. That is, one should avoid the combination of a small value of ϵ and a large value of L. In our performance evaluation study, we set $\epsilon = 1$.

In summary, given the value of η, ξ can be calculated as $\frac{1}{1-\eta}$. Given the calculated ξ, the relation between a_{th} and L can be determined. Appropriate values of a_{th} and L can then be determined with the consideration of reducing the overhead of BSS as much as possible.

Tuning L and a_{th} without the knowledge of η In reality, as X_r is not known *a priori*, η cannot be readily obtained. In what follows, we discuss how to set the value of a_{th} given the value of ϵ, without the knowledge of X_r. Then we determine the value of L.

To determine the value of a_{th}, we propose an on-line tuning scheme. Before applying BSS, we first take N_{pre} samples (which we call *pre-samples*) from which we obtain an initial estimate of the mean and set the value of a_{th} accordingly. Then BSS commences, and we set the value of a_{th} as $a_{th} = E(X_{bss,i}) \times \epsilon$, where $X_{bss,i}$ is the sampled mean of the sample set that contains all the samples up to and including the ith *regular* sample (i.e., the set includes the *pre-samples*, the i samples and all the *qualified samples* taken so far). Note that during the course of taking extra qualified samples in a sampling interval, the value of a_{th} is not updated, 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 interval when the next *regular* sample is to be taken will the value of a_{th} be updated as $E(Y_i) \times \epsilon$.

Given the value of a_{th}, the value of η is needed to set an appropriate value of L. In the lack of the η value, we estimate it from the sampling rate r as follows. As shown in [21] (Chapter 3), if we define

$$V_n = N^{1-1/\alpha}(X_s - X_r), \tag{24}$$

then

$$V_n \to \varphi_\alpha, \quad \text{in distribution}, \tag{25}$$

where φ_α is an α-*stable* distribution. That is, V_n converges in distribution for large values of N, i.e., $|X_s - X_r| \sim N^{1/\alpha-1}$. Hence,

$$\eta = \frac{|X_r - X_s|}{X_r} \sim \frac{N^{1/\alpha-1}}{X_r}. \tag{26}$$

Let N_{total} be the total number of points in the original processes, and r the systematic sampling rate. Then $N = N_{total} \cdot r$, and

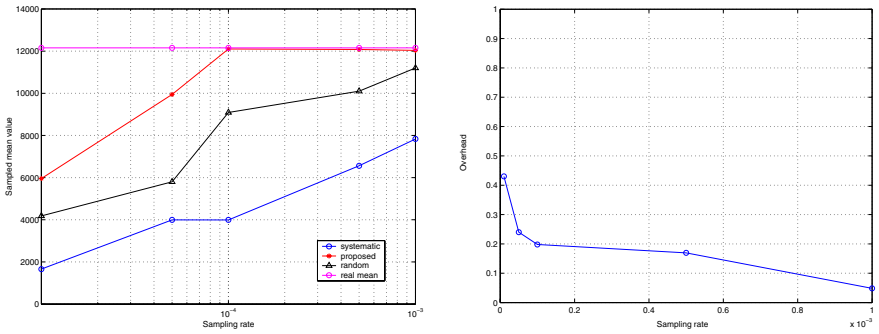
$$\eta \sim C_s \cdot r^{1/\alpha-1}, \tag{27}$$

where $C_s = \frac{N_{total}^{1/\alpha-1}}{X_r}$ is a constant less than 1 for $1 \le \alpha \le 2$. In our experimental study, we find that for synthetic traces ($\alpha = 1.5$), $C_s \in (0.08, 0.15)$ while for real traces ($\alpha = 1.66$), $C_s \in (0.05, 0.1)$.

In summary, Eq. (27) is used to estimate the value of η. With the value of η, one can obtain $\xi = \frac{1}{1-\eta}$. By plugging in both the values of ξ and a_{th} in Eq. (23), one can obtain the value of L.

6 Performance evaluation

To evaluate the performance of BSS, we have carried out several sets of experiments on both synthetic and real Internet traces. As BSS achieves its accuracy by sampling more "biased" samples of larger values, we use the following three metrics to evaluate BSS:



(a) Sampled mean (b) Sampling overhead

Figure 8. 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.

(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 $e = \frac{1-\eta}{log(N_{bss})}$ and N_{bss} is the total number of samples (including both the samples normally taken in systematic sampling and the *qualified samples* taken in BSS).

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.

Performance w.r.t. Sampled Mean, Overhead and Efficiency We use the same traces given in Section 4. For synthetic traces, we set the shape parameter of the on/off periods to be $\alpha \in (1.2, 1.6)$. Figures 9–8 give the sampled mean obtained by systematic sampling, simple random sampling, and BSS ((a)), and the sampling overhead incurred in BSS ((b)) for both synthetic and real Internet traces. Note that the result shown in Fig. 9 is for the synthetic trace with $\alpha = 1.3$ and mean value 5.68 kbytes/second, while that in Fig. 8 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. 9 (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. 9 (b), the overhead is below 0.2 for larger sampling rates ($\geq 10^{-4}$) and below 0.5 for smaller sampling rates, while $1 - \eta$ (Section 5.3) 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. 8, except that the sampling overhead is around 0.2.

Fig. 10 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 value of e for BSS is 0.36, while that for systematic and simple random sampling is 0.26 and 0.3, respectively, i.e., BSS achieves a performance gain of 40% and 20%, respectively, as compared to systematic and simple random sampling.

Performance w.r.t. Hurst Parameter and Average Variance 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. However, due to the space limit, we cannot include the experimental results, but refer the interested reader to [15] for a detailed account. As a synopsis of all the experimental results, the BSS can retain the Hurst parameter of the original process and achieve the same average variance as the systematic sampling method. 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.

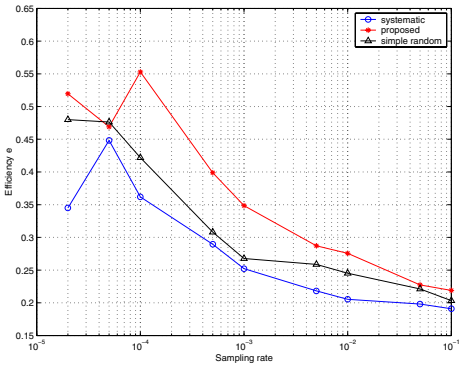
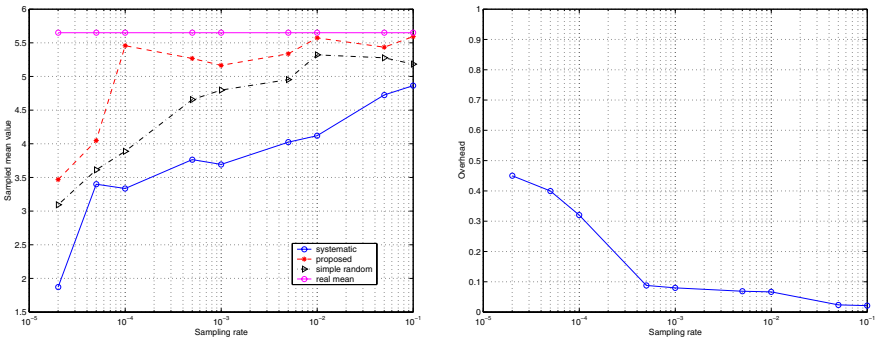


Figure 10. The efficiency of systematic sampling, simple random, and BSS for synthetic traffic.

7 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 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



(a) Sampled mean (b) Sampling overhead

Figure 9. The sampled mean obtained by systematic sampling, simple random, and BSS ((a)), and and the sampling overhead incurred in BSS ((b)) for synthetic traces.

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 40% and 20% (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 un-based 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.

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] *http://lion.cs.uiuc.edu/sampling2.pdf*

[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] 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.

[23] 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.