An Analysis of Packet Sampling in the Frequency Domain

Luigi Alfredo Grieco [*]
DEE - Politecnico di Bari, Italy
a.grieco@poliba.it

Chadi Barakat
INRIA - Planet Group, Sophia Antipolis, France
chadi.barakat@sophia.inria.fr

ABSTRACT

Packet sampling techniques introduce measurement errors that should be carefully handled in order to correctly characterize the network behavior. In the literature several works have studied the statistical properties of packet sampling and the way it should be inverted to recover the original network measurements. Here we take the new direction of studying the spectral properties of packet sampled traffic. A novel technique to model the impact of packet sampling is proposed based on a theoretical analysis of network traffic in the frequency domain. Moreover, a real-time algorithm is also presented to detect the spectrum portion of the network traffic that can be restored once packet sampling has been applied. Preliminary experimental results are reported to validate the proposed approach.

Categories and Subject Descriptors

C.2.3 [**Computer-Communication Networks**]: Network Operations—*Network Monitoring*

General Terms

Measurement, Algorithms, Theory

Keywords

Packet sampling, Measurement, Aliasing, Variance

1. INTRODUCTION

Packet sampling techniques are very useful to reduce the complexity of network monitoring systems [5, 6]. They simply consist on capturing a subset of packets, which are then used to infer the original traffic properties. Packet sampling is known to introduce estimation errors that should be very carefully handled in order to correctly characterize the network behavior [13]. This problem has been faced by the scientific community in recent years and many novel analysis and sampling techniques have been proposed, e.g., [7, 8, 15, 10, 16, 4, 11].

These previous works, among others, have shed the light on many of the statistical properties of packet sampling. Several inversion methods [1] have followed combining stochastic analysis and statistical inference. In this paper, we look at packet sampling from another interesting perspective, that of the spectral density of the traffic bit rate averaged over some time intervals, called bins, and tracked over time. In fact, the traffic is not fixed, but varies over time forming a signal composed of several frequencies. We try to evaluate the parts of the spectrum that get altered because of sampling and identify efficient non-biased inversion methods. Our target is not only the volume of the traffic or its marginal distribution at some time instant, but rather how many frequencies we can still recover after sampling. This way we can make sure that the main frequencies in the original traffic are preserved, which is of major importance for applications like anomaly detection and network tomography [3, 14, 1]. By the help of Fourier Transforms, we develop an original theoretical framework able to explain the impact of packet sampling on the traffic spectral density. In particular, the error in the estimation of the traffic volume is modeled as an aliasing effect in the frequency domain [17]. Moreover, by leveraging the theoretical analysis, a real-time algorithm is also designed to detect the spectrum portion of the network traffic signal that can be restored once packet sampling has been applied. Preliminary experimental results are reported to validate the proposed approach.

The rest of the work is organized as follows: Sec. 2 overviews the related work; Sec. 3 formulates the problem of estimating the binned traffic rate in the frequency domain; in Sec. 4 the effects of packet sampling are modeled; in Sec. 5 a filterbank is proposed to process a stream of sampled packets and to estimate the portion of the spectrum of the original traffic that can be restored; Sec. 6 shows preliminary experimental results; finally the last section draws conclusions and future research.

2. RELATED WORK

We review the body of the literature relevant to our discussion. In [7] a method is proposed that allows the direct

[*] From March to June 2009 he has been working as visiting researcher at INRIA, Planete Research Group, Sophia Antipolis, France.

[1] In sampling terminology, *inversion* is the process of estimating original traffic properties from sampled measurements.

IMC'09, November 4–6, 2009, Chicago, Illinois, USA.
Copyright 2009 ACM 978-1-60558-770-7/09/11 ...$10.00.

inference of traffic flows by observing the trajectories of a subset of all packets traversing a domain. The key idea is to sample packets based on a hash function computed over the packet content. Using the same hash function yields the same sample set of packets in the entire domain, and enables the reconstruction of packet trajectories. The approach allows also to cope with unreliable report transport.

[8] focuses on the frequencies at which different numbers of packets per flow occur. In particular, the paper: (i) shows how to smooth the estimated flow size distribution in order to deal with short flows that disappear; (ii) uses maximum likelihood estimation to derive the full distribution of packet and byte flow lengths; (iii) exploits protocol level details to render the estimators more accurate.

In [9] a method for sampling flow records on some router interface is proposed. It is based on a threshold-based sampling strategy that sets the sampling probability according to the size of the flow record. The theoretical properties of the estimator have been derived. Moreover, it has been demonstrated that the algorithm has an accuracy slightly smaller than a modified version of the sample and hold algorithm proposed in [10]. Finally, several strategies to dynamically control the volume of the sampled traffic are proposed and compared.

In [15] the Sketch Guided Sampling (SGS) has been proposed. It sets the packet sampling probability according to an estimate of the size of the flow the packet belongs to. This translates into an increase in the packet sampling rate of the small and medium flows at slight expense of the large flows, resulting in much more accurate estimations of various network statistics. Other interesting proposals have been conceived to deal with large flows [2, 10, 16, 4, 11].

[13] demonstrates that it is impossible in practice to recover the spectral density of the packet arrival process and the distribution of the number of packets per flow using traditional packet based sampling. Thus, it proposes to sample flows rather than packets in order to achieve higher accuracy at the expense of an increased computational complexity.

3. PROBLEM FORMULATION IN THE FREQUENCY DOMAIN

Our objective is to estimate the amount of data sent from a sender node (S) to a receiver node (R) during consecutive time intervals of duration T, which will be referred to as bins. A node can be a net or a subnet with some IP address prefix, a domain, an edge router, etc. The estimation is carried out using packet sampling, i.e., each packet is captured with a uniform probability p. Packets captured during the same bin are summed together, then the resulting binned value is tracked over time to understand the traffic behavior.

To model the spectral density of the traffic signal, we divide the time axis into small time slots with size t_0. In each slot, no more than one packet can be transmitted. In practice, this t_0 corresponds to the transmission time of the smallest packet over the monitored link. We define $d(k)$ as the amount of data sent by S during the time interval $[(k) \cdot t_0, (k+1) \cdot t_0[$, where $k \in N$. To be more precise, if the transmission of an entire packet has been accomplished during the time interval $[(k) \cdot t_0, (k+1) \cdot t_0[$, $d(k)$ will be equal to the size of the sent packet, otherwise $d(k)$ will be equal to 0. Moreover, we take the bin size T to be an integer multiple of t_0, i.e., T is made by T/t_0 slots. The expected Fourier Trasform of $d(k)$ can be expressed as follows [17]:

$$E[D(f)] = \sum_{k=-\infty}^{+\infty} E[d_k] \cdot e^{-j2\pi k f t_0} = \sum_{n=-\infty}^{+\infty} D_0(f - \frac{n}{t_0}), \quad (1)$$

where $D(f)$ is the Fourier Transform of $d(k)$ and $D_0(f) = 0$, for $|f| > \frac{0.5}{t_0}$. This expression has a general validity because the spectrum of any discrete-time signal is periodic with period equal to $1/t_0$, if the time between two subsequent samples is equal to t_0. Basically, $D_0(f)$ is a function that we introduce and that includes all frequencies of the signal $d(k)$ in the interval $[-0.5/t_0, +0.5/t_0]$. Moreover, we define f_M, $f_M \leq \frac{0.5}{t_0}$, as the maximum frequency of the spectrum $D_0(f)$. To better clarify the meaning of our notation, Fig. 1 pictures a typical example for $E[D(f)]$.

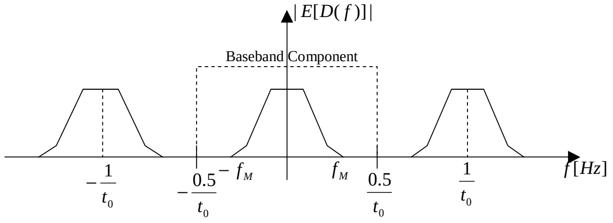


Figure 1: Expected spectrum of original packet stream $d(k)$.

As first step, we model the spectrum of the traffic signal under the ideal assumption of capturing all packets, i.e., $p = 1$. Given that the measurement bin lasts T/t_0 time slots, summing the data received in a bin time can be expressed as filtering $d(k)$ using a discrete-time filter with pulse response $h(k) = 1$ for $k = 0 \ldots T/t_0 - 1$, and $h(k) = 0$ for $k \geq T/t_0$. The corresponding transfer function is:

$$H(f) = e^{-j\pi f(\frac{T}{t_0}-1)t_0} \cdot sin(\pi f T)/sin(\pi f t_0). \quad (2)$$

$H(f)$ is a low-pass filter with cutoff bandwidth $B \approx \frac{0.445}{T}$ and static gain equal to T/t_0 [12]. Moreover, it is worth noting that the spectrum of $H(f)$ is periodic (with period $1/t_0$) because the corresponding pulse response is discrete. Thus, $H(f)$ acts as a low-pass filter in the frequency band $[-0.5/t_0, 0.5/t_0]$.[2] To provide a further insight into the filter $H(f)$, Fig. 2 plots the module of its transfer function obtained for $t_0 = 1s$ and $T = 10s$.

Being $H(f)$ a linear filter, it holds that the expected Fourier Transform of $\bar{d}(k)$, the filtered version of the traffic signal $d(k)$, is:

$$E[\bar{D}(f)] = H(f)E[D(f)] = \frac{T}{t_0} \sum_{n=-\infty}^{+\infty} \bar{D}_0(f - n/t_0), \quad (3)$$

where $\bar{D}_0(f) = t_0 \frac{H(f)D_0(f)}{T}$. The last equality in Eq. (3) holds because both $H(f)$ and $E[D(f)]$ are periodic functions with the same period $1/t_0$. Fig. 3 plots an approximated model of $E[\bar{D}(f)]$.

Now, we present our approach to move from a discrete-time signal representation to a continuous-time one. This is shown in Fig. 4: the signal $\bar{d}(k)$ is decimated by a factor

[2] Frequency components of $d(k)$ outside the interval $[-0.5/t_0, 0.5/t_0]$ can be filtered out only using an interpolator, i.e., a continuous time filter, that reconstructs a continuous version of the signal $d(k)$.

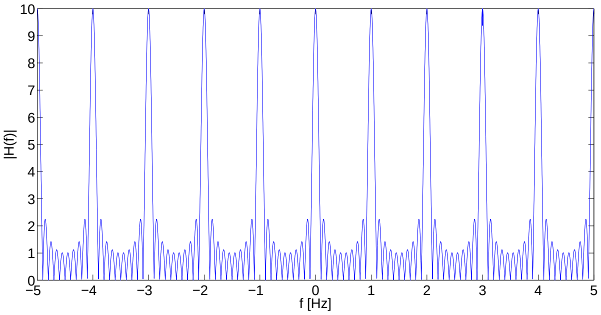


Figure 2: Module of $H(f)$ ($t_0 = 1s$ and $T = 10s$).

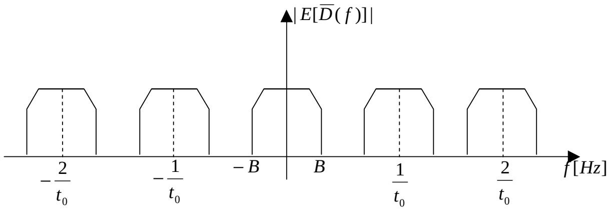


Figure 3: Approximated model of $E[\bar{D}(f)]$.

T/t_0, i.e., one sample of $\bar{d}(k)$ is taken every bin, then the resulting signal $\bar{d}_T(k)$ is processed with Zero Order Holder (ZOH), which is a device that keeps the output $\hat{d}(t)$ equal to the last received sample. Using the Poisson summation formula [17], the expected spectrum of $\bar{d}_T(k)$, i.e., the decimated version of $\bar{d}(k)$, is:

$$E[\bar{D}_T(f)] = \sum_{n=-\infty}^{+\infty} \bar{D}_0(f - n/T). \tag{4}$$

It is worth noting that the spectrum $E[\bar{D}_T(f)]$ is the sum of the functions $\bar{D}_0(f-\frac{n}{T})$, which are obtained by translating $\frac{T}{t_0}\bar{D}_0(f)$ by integer multiples of $\frac{1}{T}$ and by dividing the result by T/t_0. As a consequence, and given that the bandwidth of $\bar{D}_0(f)$ is $B \approx \frac{0.445}{T}$ [12], the decimation does not introduce aliasing. Moreover, the transfer function of the ZOH is:

$$G_{ZOH}(f) = e^{-j\pi fT} \cdot sin(\pi fT)/(\pi fT), \tag{5}$$

which is a low-pass filter with unitary static gain and bandwidth equal to that of $H(f)$. With respect to $H(f)$, the ZOH is also able to filter out all high frequency components of the input signal, so that, the expected spectrum of the continuous-time signal $\hat{d}(t)$ is no more periodic and can be expressed as follows:

$$E[\hat{D}(f)] = G_{ZOH}(f)E[\bar{D}_T(f)] \approx G_{ZOH}(f)\bar{D}_0(f). \tag{6}$$

This is no other than a low-pass filtered version of the baseband component of the expected spectrum of $d(k)$. The signal $\hat{d}(t)$ is the binned traffic rate that network operators track over time for management and monitoring purposes, thus our aim is to evaluate the impact of packet sampling on the spectrum of this signal and to propose conservative values for T and p to be used. Note that most of the difficulty comes from the fact that the spectrum of the original signal $d(k)$ is unknown from sampled traffic.

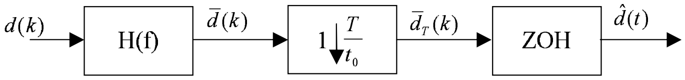


Figure 4: Continuous time reconstruction of original packet stream $d(k)$.

4. MODELING PACKET SAMPLING

Following the same methodology, we derive the spectrum of the traffic bit rate estimated from sampled packets. We show how this spectrum is related to the spectrum of the original traffic and we identify the part of the spectrum (i.e., the set of frequencies) that can be recovered without noise. We relate this finding to the values of the sampling rate p and the measurement bin T under consideration. One can use this result to set the values of p or T, or both together, such to avoid aliasing and to recover a traffic signal close, if not identical, to the original one. First, we state our main result then we follow with its derivation.

Main result: for a traffic rate signal with maximum frequency f_M in the baseband, an averaging interval T and a packet sampling rate p, estimation errors are fully avoided iff $\frac{0.445}{T} < \frac{p}{t_0} - f_M$. In the other cases, the estimation errors are due to frequency aliasing effects that cannot be filtered out.

Suppose that packets are sampled with some uniform probability $0 < p < 1$ and denote by $d_p(k)$ the volume of sampled data in the time slot $[(k)\cdot t_0, (k+1)\cdot t_0[$, $k \in N$. The signals $d(k)$ and $d_p(k)$ are related to each other, as for each k, $d_p(k)$ is equal to $d(k)$ with probability p and to 0 with probability $1-p$. Let us express the time-slot corresponding to the n-th captured sample of $d(k)$ as $t_n = (\frac{n}{p} + \Delta_n)t_0$, Δ_n being a random variable modeling the time between sampled packets. Under this hypothesis we can compute the spectrum of $d_p(k)$ as:

$$D_p(f) = \sum_{n=-\infty}^{+\infty} d(\frac{n}{p} + \Delta_n)e^{-j2\pi f(\frac{n}{p}+\Delta_n)t_0} \tag{7}$$

$$= \sum_{n=-\infty}^{+\infty} d(\frac{n}{p} + \Delta_n)e^{-j2\pi f\frac{n}{p}t_0}\left(1 + \sum_{i=1}^{+\infty}\frac{(-j2\pi f\Delta_n t_0)^i}{i!}\right). \tag{8}$$

Since we are interested in low-frequency components with $|f| < \frac{1}{T}$, we can safely assume that $f\Delta_n t_0 \ll 1$ or equivalently $\Delta_n \ll T/t_0$. This simply means that the bin size is very larger than the jitter of the inter-arrival time between sampled packets. Thus:

$$D_p(f) \approx \sum_{n=-\infty}^{+\infty} d(n/p + \Delta_n)e^{-j2\pi f\frac{n}{p}t_0}. \tag{9}$$

Assuming further that $E[d(\frac{n}{p} + \Delta_n)] \approx E[d(\frac{n}{p})]$, i.e., the stationarity interval of $d(k)$ is greater than $\Delta_n t_0$, we can compute the expectation of both members of Eq. (9) as follows:

$$E[D_p(f)] \approx \sum_{n=-\infty}^{+\infty} E[d(n/p)]e^{-j2\pi f\frac{n}{p}t_0}. \tag{10}$$

Thus, in the frequency band of interest, the spectrum of the sampled traffic $E[D_p(f)]$ can be viewed as the spectrum of the original traffic $E[d_k]$ sub-sampled with frequency $\frac{p}{t_0}$. Recalling the spectrum of the signal $E[d_k]$ reported in Eq. (1),

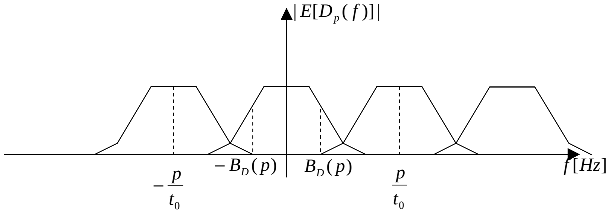


Figure 5: Expected spectrum of sampled packet stream $d_p(k)$.

it holds that [17]:

$$E[D_p(f)] \approx p \sum_{n=-\infty}^{+\infty} D_0(f - n \cdot p/t_0). \tag{11}$$

An example of this spectrum is plotted in Fig. 5 where we can see the aliasing introduced by packet sampling. In general, the entire baseband component $D_0(f)$ cannot be estimated from $D_p(f)$. We define $B_D(p)$ as the largest frequency component of $D_0(f)$ that can be restored from $D_p(f)$, i.e., only frequency components of $d(k)$ with $|f| \leq B_D(p)$ can be reconstructed from $d_p(k)$. In other words, if we filter $d_p(k)$ using a low-pass filter with bandwidth B, such as $H(f)$ whose cutoff bandwidth is approximately $\frac{0.445}{T}$, we have to impose $B \leq B_D(p) = \frac{p}{t_0} - f_M$ to achieve a correct estimate[3].

It is worth pointing out that in real cases, the maximum traffic baseband frequency f_M could be very close to $0.5/t_0$, thus it is not possible to fully avoid aliasing effects. As a consequence, $B_D(p)$ should be defined as the maximum frequency that can be estimated with a reasonably small error due to aliasing.

Once sampled, the reduced traffic $d_p(k)$ is filtered using $H(f)$ to obtain the signal $\bar{d}_p(k)$. Its average spectrum can be expressed as:

$$E[\bar{D}_p(f)] \approx pH(f) \sum_{n=-\infty}^{+\infty} D_0(f - n \cdot p/t_0). \tag{12}$$

By isolating the baseband component $\bar{D}_0(f)$, this can be rewritten as:

$$E[\bar{D}_p(f)] \approx p\frac{T}{t_0}\bar{D}_0(f) + pH(f) \sum_{n \neq 0} D_0(f - n \cdot p/t_0) \tag{13}$$

Finally, in order to move to a continuous time representation that models the averaging of the sampled traffic over bins of T/t_0 slots, the signal $\bar{d}_p(k)$ has to be decimated by a factor T/t_0 before being interpolated using a ZOH (see Fig. 6). Using the Poisson summation formula as done to derive Eq. (4), and provided that the filter $H(f)$ has removed the aliasing due to the sampling, the expected spectrum of $\bar{d}_{(p,T)}(k)$, i.e., the decimated version of $\bar{d}_p(k)$, can be written as:

$$E[\bar{D}_{(p,T)}(f)] \approx p \sum_{n=-\infty}^{+\infty} \bar{D}_0(f - n/T). \tag{14}$$

By applying the ZOH, one can extract a continuous time reconstruction of the sampled traffic whose spectrum is $\frac{pG_{ZOH}(f)}{D_0(f)}$, i.e., a low-pass filtered version of the base-band component of the average spectrum of $d(k)$ scaled by p. Compared to Eq (6), this implies that the signal $d_p(k)$ modeling the sampled traffic should be divided by p in order to compensate the scaling due to sampling and obtain the same spectrum as the time averaged reconstruction of the original traffic.

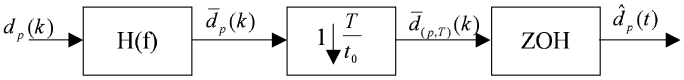


Figure 6: Continuous time reconstruction of sampled packet stream $d_p(k)$.

5. ESTIMATING $B_D(P)$ USING A FILTER-BANK

The distortion of the signal $d_p(k)$ is due to the tails of the spectrum of the signal $d(k)$ translated and folded together in the bandwidth of interest $[-B, +B]$. For a sampling probability p, we expect to have a number of replicas equal to $\frac{1-p}{p}$, see Fig. 5. If we assume a constant energy density n_0 for the tails of the spectrum of the original traffic $d(k)$, and we refer to $NE(B)$ as the energy of the noise in the bandwidth of interest $[-B, +B]$, we can write

$$NE(B) \leq 2 \cdot [(1-p)/p] \cdot B \cdot n_0. \tag{15}$$

The inequality holds because the replicas of $D_0(f)$ do not necessarily sum up with the same phase in the band of interest. Eq. (15) is very important because it tells us that as long as we increase p, not only the total energy of the noise decreases but also its first order derivative with respect to p decreases. As a consequence, by inspecting the behavior of the variance (i.e., proportional to the energy) of any low-pass filtered version of d_p/p, we can infer if the noise introduced in a bandwidth B is significant.

Herein, we leverage this theoretical finding to propose an algorithm that estimates $B_D(p)$ by properly processing the output of a bank of low-pass filters. The traffic is supposed to be sampled in the network at rate p, hence information on the original traffic, typically f_M and the shape of the baseband component, is not available. The only option left is either to down sample further the traffic at the monitor, or to play with the monitoring time bin. By calculating the variance of the traffic and tracking its behavior with the new sampling rate and the time bin, one can estimate the bandwidth $B_D(p)$ of the traffic signal that can be restored at p. Note that knowing $B_D(p)$ allows one operator or a router to properly select the minimum monitoring time resolution T at which sampled packets should be averaged over time without paying for aliasing. In fact, we remember that once $B_D(p)$ is estimated, the bin size T can be set as:

$$T = 0.445/B_D(p). \tag{16}$$

Before starting the description of the algorithm, it is important to underline two facts. If we filter the signal d_{p_1}/p_1, obtained using a sampling probability p_1, with two different low-pass filters, namely F_{B_1} and F_{B_2}, having respectively bandwidth B_1 and B_2, with $B_1 < B_2$, we expect to collect a larger amount of noise due to aliasing by using the second filter (see Eq. (15)). Moreover, if we filter two different signals d_{p_1}/p_1 and d_{p_2}/p_2, with $p_1 < p_2$, using the same low-pass

[3] It can be easily shown that the same result holds for any slot size which is an integer sub-multiple of t_0.

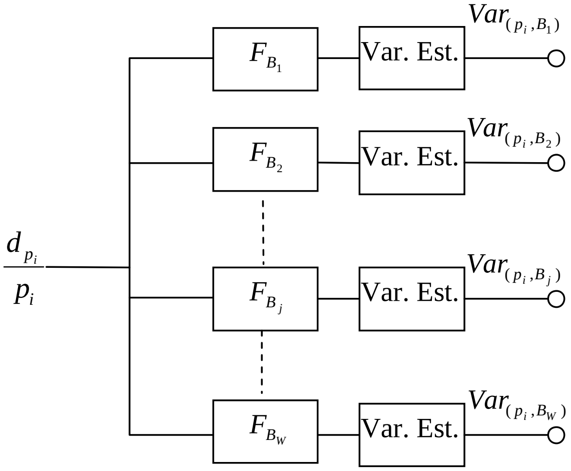


Figure 7: Filter-bank.

filter, we expect to collect a smaller amount of noise when the second signal is processed. In any case, when p is sufficiently large, the energy of the noise will be negligible (see Eq. (15)). *Thus we will estimate $B_D(p)$ as the largest bandwidth for which a perturbation of p does not introduce any significant change in the variance of the binned and sampled traffic.*

To accomplish this task, we propose a filter-bank made by W low-pass filters with bandwidths equal to B_1, B_2, $\dots$, B_W, $B_j < B_{j+1} \forall j$. The tool is described in Fig. 7. Each filter is fed with the inverted sampled traffic d_{p_i}/p_i. The output of each filter is analyzed over time bins inversely proportional to the filter bandwidth, according to Eq. (16). For each bin size, the variance of the filtered signal is evaluated. We define $Var_{(p_i,B_j)}$ to be the estimated variance of the output of the filter-bank i.

In parallel, L filter-banks, identical to the one described above, are applied to the signals d_{p_1}/p_1, $d_{p_2}/p_2, \dots, d_{p_L}/p_L$, obtained by further under-sampling d_p, with $p_1 < \cdots < p_L < p$, respectively. For a given bandwidth B_j, we compare the values of $Var_{(p_i,B_j)}$ obtained for several values of $p_i < p$. If we find that those variances are too dissimilar among them, we have to conclude that the frequencies in the interval $[-B_j, B_j]$ cannot be safely estimated using the current value of p. Thus a smaller B_j should be considered (see also Fig. 8). The bandwidth $B_D(p)$ we are looking for is the largest among those B_j not presenting an aliasing problem.

It could happen that the outcome of the analysis is that the current value of p is very large compared to the maximum allowed one, even for the largest B_j. This information could be fruitfully exploited to reduce the packet sampling probability inside the network.

Notice that, the proposed algorithm is based on linear filters. Thus, it can easily run in real time using cards equipped with Digital Signal Processors. Further details about the filter-bank will be provided in the next section.

6. EXPERIMENTAL RESULTS

We have processed a real packet trace collected in January 2009 over a trans-pacific 150 Mbps link [4]. Experiment parameters are listed in Tab. 1. Fig. 9 shows the module of the spectrum of the inverted sampled traffic $d_p(k)/p$, obtained for several values of p, when 10,000 packets of the aggregate trace are considered. By comparing the plot obtained for $p = 1$ with respect to the other ones, it is straightforward to note that: (i) only low frequencies of the original traffic can be recovered, even using a very high sampling probability as $p = 0.1$; (ii) the harmonic tones of the original traffic, i.e., those obtained for $p = 1$, appear translated in the frequency spectrum of the sampled traffic signals as expected by the Poisson summation formula; (iii) the noise across the continuous component of the traffic signal grows with $1/p$ as expected by inequality (15).

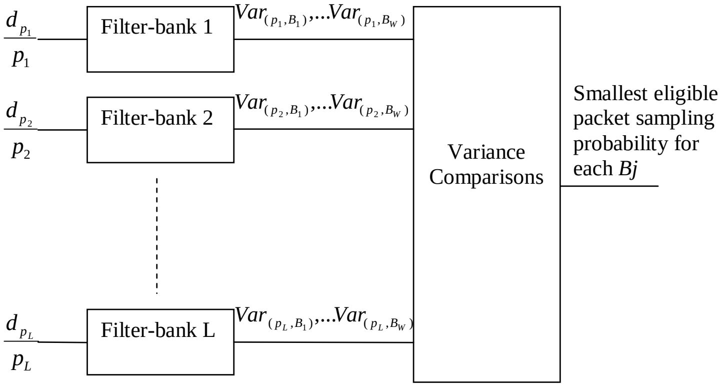

d_{p_1}
p_1
Filter-bank 1
$Var_{(p_1,B_1)}$,...$Var_{(p_1,B_W)}$
d_{p_2}
p_2
Filter-bank 2
$Var_{(p_2,B_1)}$,...$Var_{(p_2,B_W)}$
d_{p_L}
p_L
Filter-bank L
$Var_{(p_L,B_1)}$,...$Var_{(p_L,B_W)}$
Variance Comparisons
Smallest eligible packet sampling probability for each Bj

Figure 8: Filter-banks with variance comparisons.

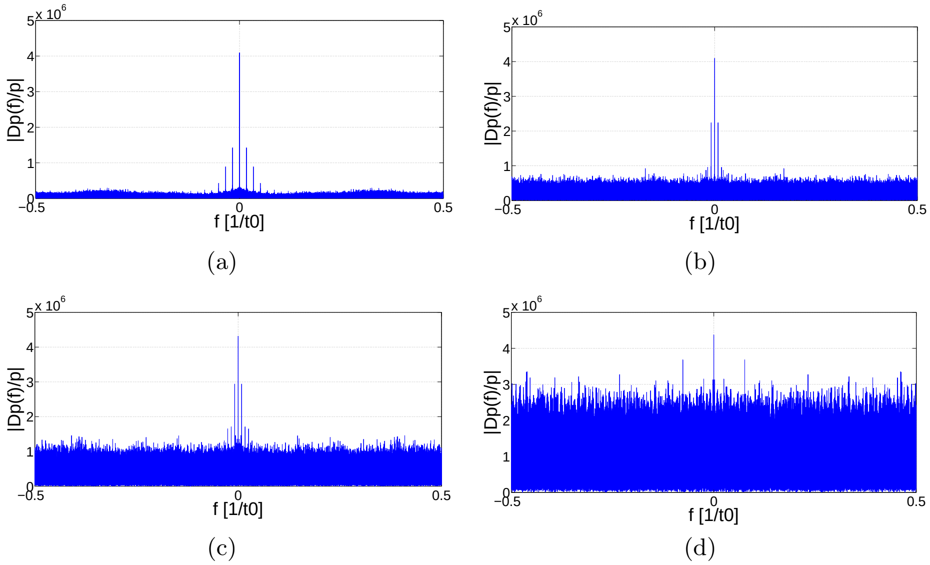
Param.	Description	Value
t_0	time-slot	$2.13 \cdot 10^{-6} s$
T	bin size	$[1s, 400s]$
p	sampling probability	$[10^{-4}, 1]$
L	no. of filter-banks	16
W	no. of filters in each filter-bank	11
th	threshold for Var. comparisons	[0.1, 0.2]
N	no. of Var. comparisons	2,3

Table 1: Experiment parameters.

Regarding the effectiveness of the filter-bank proposed in Sec. 5, we have processed the packet trace with values of p in the range $[10^{-4}, 1]$. Each decade of this range has been split into 3 octaves, so that we have considered $L = 16$ possible values for p. Moreover, we have considered $W = 11$ low pass filters, whose respective bandwidths, according to (16), allow the bin size T to range over the interval $[1s, 400s]$. Given B_j and $p = p_i$, the algorithm compares the ratios $\left|\frac{Var_{(p_{i-1},B_j)}}{Var_{(p_i,B_j)}}\right|, \left|\frac{Var_{(p_{i-2},B_j)}}{Var_{(p_i,B_j)}}\right| \dots \left|\frac{Var_{(p_{i-N},B_j)}}{Var_{(p_i,B_j)}}\right|$ with respect to a threshold $th_0 = 1 + th$. N is the number of consecutive sampling rates tested to decide on the appropriateness of B_j and p_i. The considered value of $p = p_i$ is admissible for the bandwidth B_j iff all the considered ratios are smaller than th_0. To further improve accuracy, we require that the ratio between $Var(p_i, B_j)$ and the square of the estimated traffic volume is smaller than th (small relative error).

In our experiments we have set N to 2 or 3, and we have varied th in the range $[0.1, 0.2]$. For each bandwidth B_j,

[4] The trace is available at `http://mawi.wide.ad.jp/mawi/samplepoint-F/2009/`



(a)

(b)

(c)

(d)

Figure 9: Baseband component of $D_p(f)/p$: (a) $p = 1$; (b) $p = 0.1$; (c) $p = 0.03$; (d) $p = 0.005$.

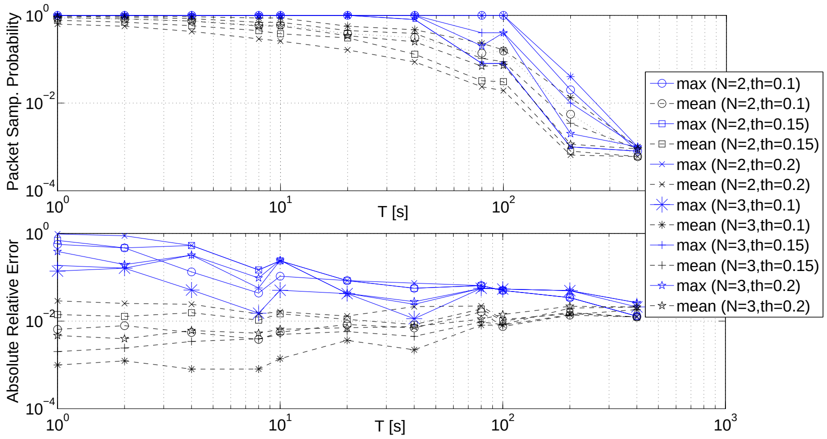


Figure 10: Minimum allowed packet sampling probabilities (up) and absolute relative errors (down).

we have recorded the smallest admissible value of p and the absolute relative estimation error of the traffic volume calculated over time windows of T seconds each.

Fig. 10 shows both maximum (solid lines) and average (dashed lines) values of the minimum allowed packet sampling probability and the absolute relative error, for each considered value of T. By looking at Fig. 10, it is clear that, regardless of N and th, our algorithm, for each bandwidth B_j (linked to T by Eq. (16)), provides admissible packet sampling rates that ensure negligible average estimation errors. It also clear how the required sampling rate increases when the traffic is monitored over smaller and smaller intervals. For example, seen from traffic spectrum viewpoint, the use of packet sampling rates smaller than 0.05 is only admitted for bin sizes T not smaller than $100s$. Finally, as expected, it is straightforward to note that the proposed algorithm becomes more conservative, i.e., provides smaller values of packet sampling probability, as N increases and th decreases.

7. CONCLUSIONS AND FURTHER RESEARCH

A novel technique to model the impact of noise caused by packet sampling is proposed in this paper, exploiting a theoretical analysis in the frequency domain. Moreover, a real-time algorithm is presented to detect the spectrum portion of the network traffic signal that can be restored once packet sampling has been applied. Preliminary experimental results have been reported to validate the proposed approach. Our future research will focus on the extension of the proposed approach to larger contexts as network-wide monitoring, application-level analysis and anomaly detection.

8. REFERENCES

[1] G. Androulidakis, V. Chatzigiannakis, and S. Papavassiliou. Network anomaly detection and classification via opportunistic sampling. *Network, IEEE*, 23(1):6–12, January-February 2009.

[2] C. Barakat, G. Iannaccone, and C. Diot. Ranking flows from sampled traffic. In *Proc. ACM CoNEXT 2005*.

[3] D. Brauckoff, B. Tellenbach, A. Wagner, M. May, and A. Lakhina. Impact of packet sampling on anomaly detection metrics. In *Proc. of ACM SIGCOMM IMC 2006*.

[4] B. Y. Choi, J. Park, and Z. L. Zhang. Adaptive packet sampling for accurate and scalable flow measurement. In *Proc. of IEEE Globecom 2004*.

[5] K. C. Claffy, G. C. Polyzos., and K. W. Braun. Application of sampling methodologies to network traffic characterization. *ACM SIGCOMM Comput. Commun. Rev.*, 23(4), 1993.

[6] N. Duffield. A framework for packet selection and reporting. In *IETF Draft (work in progress)*, Jun. 2008.

[7] N. Duffield and M. Grossglauser. Trajectory sampling with unreliable reporting. *IEEE/ACM Trans. on Networking*, 16(1):37–50, 2008.

[8] N. Duffield, C. Lund, and M. Thorup. Estimating flow distributions from sampled flow statistics. In *Proc. of ACM Sigcomm 2003*.

[9] N. Duffield, C. Lund, and M. Thorup. Learn more, sample less: Control of volume and variance in network measurement. *IEEE Trans. on Information Theory*, 51(5):68–80, 2005.

[10] C. Estan and G. Varghese. New directions in traffic measurement and accounting: Focusing on the elephants, ignoring the mice. *ACM Trans. Comput. Syst.*, 21(3), 2003.

[11] F. Hao, M. Kodialam, T. V. Lakshman, and S. Mohanty. Fast, memory efficient flow rate estimation using runs. *IEEE/ACM Trans. on Networking*, 15(6):1467–1477, 2007.

[12] F. Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of the IEEE*, 66(1), 1978.

[13] N. Hohn and D. Veitch. Inverting sampled traffic. *IEEE/ACM Trans. on Networking*, 14(1):68–80, 2006.

[14] P. Kanuparthy, C. Dovrolis, and M. Ammar. Spectral probing, crosstalk and frequency multiplexing in internet paths. In *Proc. of ACM SIGCOMM IMC 2008*.

[15] A. Kumar and J. Xu. Sketch guided sampling - using on-line estimates of flow size for adaptive data collection. In *Proc. of IEEE Infocom 2006*.

[16] T. Mori, M. Uchida, R. Kawahara, J. Pan, and S. Goto. Identifying elephant flows through periodically sampled packets. In *Proc. of ACM SIGCOMM IMC 2004*.

[17] J. Proakis and D. G. Manolakis. *Digital Signal Processing*. Prentice Hall, Int. Eds., 3 edition, 1996.