# A Multiadaptive Sampling Technique for Cost-effective Network Measurements

João Marco C. Silva, Paulo Carvalho, Solange Rito Lima*

*Departamento de Informática, Centro Algoritmi*
*Universidade do Minho, Braga, Portugal*

**Abstract**

The deployment of efficient measurement solutions to assist network management tasks without interfering with normal network operation assumes a prominent role in today's high-speed networks attending to the huge amounts of traffic involved. From a myriad of proposals for traffic measurement, sampling techniques are particularly relevant contributing effectively for this purpose as only a subset of the overall traffic volume is handled for processing, preserving ideally the correct estimation of network statistical behavior.

In this context, this paper proposes MuST - a multiadaptive sampling technique based on linear prediction, aiming at reducing significantly the measurement overhead and still assuring that traffic samples reflect the statistical characteristics of the global network traffic under analysis. Conversely to current sampling techniques, MuST is a multi and self-adaptive technique as both the sample size and interval between samples are self-adjustable parameters according to the ongoing network activity and the accuracy of prediction achieved.

The tests carried out demonstrate that the proposed sampling technique is able to achieve accurate network estimations with reduced overhead, using

☆This work is funded by FEDER through the Competitiveness Factors Operational Programme - COMPETE and Portuguese National Funds through FCT - Foundation for Science and Technology under the Project: FCOMP-01-FEDER-0124-022674. This paper extends the work previously presented at the International Conference on Computer Communication and Networks, ICCCN 2012.

*Corresponding author, Tel.+351253604436

*Email addresses:* joaomarco@di.uminho.pt (João Marco C. Silva), pmc@di.uminho.pt (Paulo Carvalho), solange@di.uminho.pt (Solange Rito Lima)

throughput as reference parameter. The evaluation results, obtained resorting to real traffic traces representing wired and wireless aggregated traffic scenarios and actual network services, prove that the simplicity, flexibility and self-adaptability of the proposed technique can be successfully explored to improve network measurements efficiency over distinct traffic conditions. For optimization purposes, this paper also includes a study of the impact of varying the order of prediction, i.e., of considering different degrees of past memory in the self-adaptive estimation mechanism. The significance of the obtained results is demonstrated through statistical benchmarking.

*Keywords:*
Sampling techniques, Traffic measurements, Linear prediction, Adaptive sampling

---

## 1. Introduction

In today's Internet, network measurement techniques deal with massive traffic volumes which, in many cases, have to be processed online to provide feedback to real-time management and traffic engineering tasks. To operate properly, these tasks have to rely on an accurate view of the status of the network and of provided services. Therefore, key aspects to balance when designing an efficient network measurement solution are the estimation accuracy and the measurement overhead.

Traffic sampling techniques have been extensively used to reduce the impact of performing traffic measurements on operational networks. In these techniques, a subset of packets is selected and then used to estimate network parameters, avoiding processing all network traffic [1].

To face the drawbacks of common sampling techniques (deterministic or random), adaptive sampling techniques have been proposed (see discussion in Section 2). In *Adaptive Sampling* techniques the packet selection process considers the value of a reference parameter (e.g., throughput) observed during a measurement period. In this way, the sampling process becomes more flexible and self-adaptive, i.e., the packet selection criterion may change dynamically over the measurement period.

Despite the evolution of adaptive techniques in estimating network performance parameters correctly, their main target has not been on reducing the overhead associated with data volume involved in the sampling process. This aspect directly impacts on monitoring costs and efficiency, being particularly limitative in high-speed networks handling massive traffic volumes. The efficiency of a sampling technique may be assessed through a proper balancing between estimation accuracy and measurement overhead.

In this context, this paper presents a new multiadaptive sampling technique (MuST) based on linear prediction, which aims to reduce the amount of data involved in the network measurements without compromising the estimation accuracy. Therefore, the main objective is to reduce the measurement overhead and still assure that sampled traffic reflects the statistical characteristics of the global traffic under analysis.

For this purpose, the traffic selection process considers the levels of network activity, being configured to reduce the measurement impact when the network activity increases or the measurement process tends to overload the measurement points (MPs). The multiadaptive behavior of the proposed technique is achieved considering both the interval between samples and the sample size as adaptive parameters, bounded by proper thresholds to guarantee the representativeness of samples in capturing the network behavior. A proof-of-concept is provided using real traffic traces representing distinct traffic scenarios. The results demonstrate the effectiveness and versatility of the present proposal, outperforming conventional sampling techniques.

The remaining of this document is organized as follows: the debate on representative sampling approaches and the motivation for the present proposal is included in Section 2; the multiadaptive sampling technique, its design goals, definition and operation are described in Section 3; the proof-of-concept objectives and methodology are presented in Section 4; the evaluation results are discussed in Section 5; and finally, the conclusions are drawn in Section 6.

## 2. Related Work

Existing sampling techniques can be distinguished according to the methodology adopted to select the packets that will integrate a traffic sample.

In *Systematic Sampling*, a deterministic packet selection function is used, based on (i) the packet position (count-based); (ii) the packet arrival time to the measurement point (time-base); or (iii) the packet contents (content-based). Despite its simplicity, the traffic pattern resulting from deterministic sampling may still overload measurement points and produce biased samples [2].

*Random Sampling* techniques try to avoid biasing the samples by selecting packets according to a random function [2]. Nevertheless, these techniques cannot be deployed to estimate multipoint metrics, such as end-to-end delay, since the sampling processes on the two measurement points involved are not correlated and there are no guarantees that samples will be constituted by the same packets [3].

Adaptive sampling techniques are generally developed for a specific estimation parameter, such as packet loss [4] [5] or delay and jitter [6] [7]. A larger group of proposals are aimed at traffic characterization and SLA monitoring [8] [9] [10] [7]. Common adaptive techniques are usually based on Fuzzy Logic or Linear Prediction. In adaptive sampling based on fuzzy logic [11] [12], a controller adjusts the sampling rate based on past similar experiences, determining the most appropriate action for the current traffic conditions [13]. This approach requires a long-term database to store the knowledge and the possible action for each situation.

Linear prediction based techniques [14] [15] try to forecast network behavior based on an observed parameter in past samples. In these techniques, when the prediction is correct, the sampling rate can be reduced, while inaccurate predictions indicate a change in network activity and, therefore, an increase in sampling rate is required to determine the new pattern behavior [13]. In this sampling approach just a fixed number of samples is stored by the measurement point which are then used in the prediction process. However, if the sampling frequency increases more resources will be required from the measurement point, precisely in a critical moment of its operation.

As discussed, current sampling techniques have as main objective estimating accurately parameters of interest regarding network status, but not necessarily the efficiency. The efficiency involves, beyond high accuracy, the ability to perform measurements with minor interference with the normal network operation. Therefore, the deployment of sampling techniques able to identify critical periods in network activity adjusting their dynamics accordingly, is crucial to reduce resource requirements on network nodes and to reach efficiency.

The sampling technique proposed in this paper attempts to decrease resource consumption related to the processing, storage and transmission of captured packets during high network activity periods, while maintaining the accuracy of network statistical behavior estimation. The following section details the multiadaptive sampling technique design goals, definition and operation.

## 3. MuST - Multiadaptive Sampling Technique

### *3.1. Concepts*

Traffic sampling techniques share a set of concepts sometimes presented in an ambiguous way. To avoid misunderstanding, the most common terms were adopted in accordance with the following definitions (see illustration in Figure 1):

- *Sample:* subset of network packets that are selected at the measurement point and are considered in the estimation of network parameters. These packets are also used by the adaptive measurement algorithms as input for the estimation of the reference parameter;
- *Sample size:* time interval in which all incoming packets at the measurement point are selected and captured to compose a sample unit;
- *Interval between samples:* time interval in which all incoming packets are ignored for measurement purposes. During this time, the behavior of the network is not considered for the estimation of parameters and for the configuration of the adaptive sampling rate;
- *Reference parameter:* observed value in each sample, e.g., throughput, used as measurement parameter and input for configuring the sampling adaptation parameters.


Figure 1: Sampling concepts

### *3.2. Design Goals*

The multiadaptive sampling technique proposed in this work aims to improve the sampling reactivity and efficiency, considering both the interval between samples and the sample size as adjustable parameters [16]. In this way, the present proposal pursues the following design goals:

(i) The adaptive nature of the technique should be driven by simplicity of implementation and low consumption of resources;

(ii) The adaptive sampling process should be defined in order to minimize the impact of sampling on the normal network operation while keeping high-accuracy levels. Therefore, the technique should gauge the past and current network activity in order to estimate adequate parameters to guide the traffic selection process.

The first goal motivates the adoption of a sampling approach based on linear prediction, as proposed in [13]. In fact, as discussed in the previous section, the use of linear prediction leads to lighter solutions when compared to fuzzy logic adaptive approaches. However, in the adaptive process described in [13], the underlying processing overhead is still significant as the whole network traffic is considered for the definition of the reference parameter, regardless of packets belonging or not to a sample. In MuST, the technique described in [13] is modified so that only packets belonging to previously collected samples are taken into account to drive future sampling decisions, clearly reducing the processing overhead.

To pursue the second goal, an adaptive sampling process is defined to react autonomously and self-adapt to distinct network loads and traffic characteristics. In this process, the sampling frequency is increased or decreased whenever there is a noticeable increase or decrease in the network activity in order to allow detecting new traffic patterns as required. As increasing the sampling frequency implies higher consumption of resources (processing and storage), the sample size should be consequently reduced to mitigate the overhead increase. To guarantee that a representative amount of data is obtained when capturing the network behavior, the adaptive parameters need to be properly bounded by thresholds.

The sampling frequency can be adjusted varying the interval between samples, resorting to a linear prediction function for adapting the sampling frequency based on [13]. The new predictive function is presented in Equation 1, whereas the set of rules determining the change factor in the interval between samples are presented in Equation 2 and Table 1. This formulation is detailed in the following section.

As none of the techniques available so far present features for adjusting dynamically the sample size, this paper proposes a set of rules for varying the sample size according to the level of network activity. These rules increase and decrease the sample size considering the observed network behavior. The dynamic variation in sample size complies with the criteria presented in Table 2. This will also be detailed below.

*3.3. Multiadaptive Sampling Technique Description*

MuST takes into account the last $N$ samples to estimate the future value of the reference parameter, which is then used to determine the next interval between samples and the size of the next sample. Thus, for a sampler of order $N$, the expected value $X_p$ of the reference parameter for the next collected sample is defined as

$$X_p = X_N + \frac{\Delta T_N}{N-1} \sum_{i=1}^{N-1} \left( \left| \frac{X_{i+1} - X_i]}{\Delta T_i} \right| \right). \quad (1)$$

In Equation 1, the variable $X$ represents the values of the reference parameter of the last $N$ samples, being $X_N$ the value of the most recent sample. A second variable $T$ represents the intervals between samples, where each $\Delta T_i$ is the time elapsed between the end of the sample $X_i$ and the beginning of the sample $X_{i+1}$, i.e., $\Delta T_i = T_{i+1} - T_i$ for all $1 \leq i \leq N$ and $N > 1$.

*3.3.1. Defining the interval between samples*

When a new sample is collected, the corresponding value of the reference parameter $S$ is compared with the expected value $X_p$ in order to determine a factor of change $m$. Depending on the value of $m$, a set of rules is applied to define the sampling interval $\Delta T_{N+1}$, which will determine the start of the next sample. The factor $m$, obtained comparing $X_p$ and $S$, is given by

$$m = \begin{cases} \frac{X_p}{S} & if S \neq 0; \\ 1 & otherwise. \end{cases} \quad (2)$$

The fraction in Equation 2 returns a value close to 1 when the expected value $X_p$ is close to the current value of $S$, corresponding to a correct estimate. In this case, the range of values for $m$ is defined as varying between $m_{min} = 1 - \sigma$ and $m_{max} = 1 + \sigma$, i.e.

$$1 - \sigma < m < 1 + \sigma$$

where $\sigma$ allows to adjust the degree of adaptiveness (or reactivity) in the estimation process. As in [13], considering a 10% variation in the reference parameter (representing a variation in the network activity) leads to an adequate regulation in presence of multiple traffic types. Therefore, these values are set as $m_{min} = 0.9$ and $m_{max} = 1.1$ .

If $m < m_{min}$ the predicted value of the reference parameter was underestimated, indicating more network activity than expected. Thus, the interval between samples is decreased according to $m$ variation to achieve more accurate values in the following predictions.

On the other hand, if $m > m_{max}$ the value of the reference parameter was overestimated in the prediction, and the network activity is slowing down. In this case, the interval between samples is exponentially increased in order to converge faster to its maximum value, reducing the measurement overhead.

If the value of $S$ is null, representing that no traffic has been captured, e.g., due to a reduced network load or a temporary link failure, $m$ assumes a unitary value, which allows to keep the adaptive sampling parameters stable[1]. In this case, the current $\Delta T_N$ is assumed as the next interval between samples.

Table 1 lists the rules used to generate the next sampling interval $\Delta T_{N+1}$.

Table 1: Rules to define the next interval between samples

| current $m$ | next $\Delta T$ |
|---|---|
| $m < m_{min}$ | $\Delta T_{N+1} = m\ \Delta T_N$ |
| $m_{min} \leq m \leq m_{max}$ | $\Delta T_{N+1} = \Delta T_N$ |
| $m > m_{max}$ | $\Delta T_{N+1} = 2\ \Delta T_N$ |

An additional threshold is defined to prevent $\Delta T$ from increasing indefinitely, thus guaranteeing a minimum number of samples to obtain representative data for new predictions. Similarly, the maximum frequency of sampling is also limited so that the sampling interval does not tend to zero, which would result in capturing all traffic. These limits should weight and be adjusted according to the existing link capacity. In the present study, similarly to [13], a minimum and a maximum interval between samples of 0.1s and 8s, respectively, showed adequate for the traffic scenarios under analysis.

Figure 2 illustrates the evolution of the interval between samples as a function of a linear variation of the $m$ factor. As shown, the reactivity is smoother when the network activity increases, i.e., the next $\Delta T$ decreases proportionally to $m$ when $m < m_{min}$. Conversely, the reactivity is higher

[1]Note that, $X_p$ keeps incorporating $S$ becoming null after $N$ samples with $S = 0$. This behavior is adequate to resume properly the adaptive process when new traffic is detected. In practice, network links tend to exhibit load as, at least, control traffic is crossing the links. Therefore, successive iterations with $S = 0$ are likely due to link failure, which would be detected and handled at a higher network management layer.

in presence of low network activity, i.e., the time interval between samples varies exponentially when $m > m_{max}$.

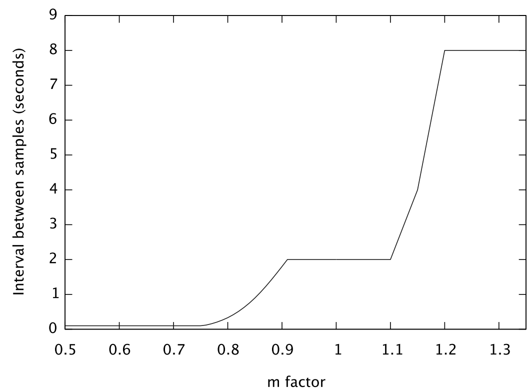



Figure 2: Evolution of interval between samples according to $m$ variation

*3.3.2. Defining the sample size*

For adapting the sample size, the factor $m$ is also considered as an indicator of network activity. Table 2 presents the rules used to define the next sample size, where $\Delta S_N$ represents the current sample size and $\Delta S_{N+1}$ the size of the next sample to be collected.

According to Table 2, in moments of increased activity, the sample size is decreased proportionally to $m$. This reduction in sample size, associated with the higher frequency in the sampling process, aims at reducing the overhead at measurement points. In presence of less network activity, the sample size is adjusted by a factor $k^2$, with $k = 0.15$ (see rational in Section 5.3). This allows to collect more data about the network in less critical periods of its operation, in sparse sampling events.

Similarly to the definition of the time interval between samples, the variation of sample sizes is also bounded. The imposed thresholds avoid small samples, which make difficult estimating parameters statistically, as well as samples excessively large, closely matching a total traffic capture. These

[2]The parameter $k$ was firstly introduced and experimentally tested in [7] with the aim of defining the variation in the time interval between samples based on a set of comparative statistics between adjacent samples. In the present proposal, the parameter $k$ is considered as the changing factor in the sample size, being defined to force an overhead reduction.

Table 2: Rules to define the next sample size

| current $m$ | next $\Delta S$ |
|---|---|
| $m < m_{min}$ | $\Delta S_{N+1} = m\ \Delta S_N$ |
| $m_{min} \leq m \leq m_{max}$ | $\Delta S_{N+1} = \Delta S_N$ |
| $m > m_{max}$ | $\Delta S_{N+1} = \Delta S_N + (k\ \Delta S_N)$ |

limits also depend on the existing links capacity, being here considered a minimum and maximum sample size of 0.1s and 2s, respectively. Figure 3 shows the evolution of the sample size for a linear variation of the $m$ factor.

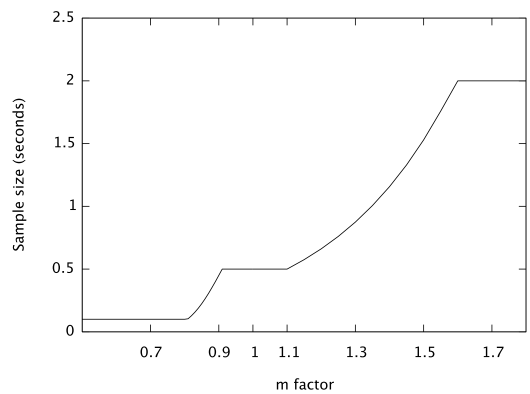



Figure 3: Evolution of sample size according to $m$ variation

For completeness, the pseudocode of MuST algorithm representing the overall sampling operation is included in Appendix A.

## 4. Proof-of-concept

The proof-of-concept aims: (i) to assess MuST ability to capture the network behavior correctly with reduced overhead; (ii) to compare the performance gain of the strategy facing currently used techniques; and (iii) to demonstrate the versatility of the proposed strategy, evaluating its effectiveness in distinct traffic scenarios. The impact of: (i) the factor $k$, (ii) the order of prediction and (iii) the network activity on the measurement efficiency are also evaluated.

The efficiency of sampling techniques is here assessed as a trade-off between accuracy and overhead. Accuracy is measured by comparing the statistical properties of the time series resulting from the total traffic trace with the corresponding outcome of applying each specific sampling technique. In more detail, as presented in Table 4, the statistical analysis studies the correctness of throughput evaluation by measuring both the variability of each time series and the relative mean error resulting from the comparative process. In [17], accuracy of estimation is further discussed, evaluating the optimal number of samples necessary to achieve a given accuracy level in the context of the total load and flow size. The overhead of each sampling technique is evaluated measuring and comparing the volume of sample data (packets and bytes) and the number of samples[3] involved.

Thus, for testing purposes, a statistical and visual analysis of the reference parameter evaluated using several sampling techniques is compared to the values obtained using the total traffic volume. Although the statistical parameters in use are common in representative research on adaptive sampling, this work extends previous works by cross-checking the sampling results against the total traffic. Common evaluation approaches only compare the performance of the proposed techniques to the Systematic techniques [2].

As mentioned, the reference parameter adopted in the tests is network throughput, which is commonly used for a graphical and statistical representation of the network activity in real time.

### *4.1. Evaluation Scenarios*

To assess the sampling optimization levels, MuST is compared to the Systematic Time-based technique [2] and Adaptive Linear Prediction technique [13]. *Systematic Time-based Sampling (ST)* is one of the most common sampling technique currently used. In the sampling process the packet selection follows a deterministic function based on the arrival time at the measurement point, i.e., the sample size and the time between samples are set at the beginning and remain unchanged along the sampling process. In this work, as suggested in [13], the operational parameters time between samples and sample size were set to 0.5s and 0.1s, respectively. In *Adaptive Linear Prediction Sampling (LP)* the time interval between samples is adjusted based on the level of network activity using linear prediction, while the sample size is fixed along the sampling process. Here, the initial interval between samples

[3]This parameter is relevant because, for each new sample, resources are required to evaluate the sampling adaptation parameters, namely the new sample size and the interval after which it should start. Thus, the higher the number of samples is, the more calculations at the measurement point are required.

Table 3: Traffic Scenarios

| Traffic Label | Characteristics | Duration | Available |
|---|---|---|---|
| **SIGCOMM08** | Captured during the SIGCOMM 2008 conference, including all the participants communications through IEEE 802.11a access points. | 8 hours | CRAWDAD[18] |
| **OC-48** | Captured passively in a backbone link OC48 in a large ISP from the US West Coast. | 5 minutes | CAIDA [19] |
| **SIP VoIP** | Capture of VoIP traffic, using SIP and G.711 encoding, in a Brazilian university campus with Fast Ethernet links. | 20 hours | University of Sergipe, Brazil |
| **Video streaming** | Capture of a live stream transmission of the STS-135 launching mission conducted by NASA on July 8, 2011. The stream was broadcast in High Definition 720p, through NASA TV channel and encoded in MPEG-4 using RTP. | 15 minutes | University of Minho, Portugal |

and sample size were set to 0.2s and 0.1s, respectively. Another configurable parameter of this technique is the order of prediction, which was set as in the original specification [13], i.e., $N = 2$. Although the technique takes into account the last two samples to decide the next interval between samples, the measurement point still needs to maintain information about all traffic since the beginning of its execution. This is needed because the technique analyzes the evolution of the reference parameter based on the accumulated amount of data until the end of the sample $N$, relative to the sample $N - 1$.

Table 3 illustrates the characteristics of the real traffic traces used in the tests. These traces were selected to provide a representative range of traffic scenarios to assess the versatility of MuST. Although the evaluation tests are based on traffic traces previously captured, the adaptive parameters of the sampling techniques are configured dynamically during the sampling process. Similarly to a real-time operational environment, there is no previous knowledge of the characteristics of the traffic passing through.

### *4.2. Statistical analysis*

The main goal of the statistical analysis is to determine the equivalence between the total traffic behavior and the behavior estimated through the sampling process, evaluating simultaneously the overhead associated with the data volume selected by the measurement technique. Additionally, the efficiency of the proposed strategy facing other proposals is also quantified.

In the analysis, the techniques are compared through the metrics - average throughput, coefficient of variation (CV), peak-to-average ratio, coefficient of correlation, relative mean error (RME) - defined in Table 4.

Table 4: Statistical Parameters

| | | | |
|---|---|---|---|
| **Overhead** | **Measurement Goal:** Evaluate the overhead associated with the number of packets processed, stored and transmitted by the measurement point where the sampling technique is deployed. | | |
| | **Metric** | **Description** | **Units** |
| | N. of Packets | Total number of packets captured during the sampling process for each sampling technique. | #packets |
| | Data Volume | Sum of all packets collected with each sampling technique. For this metric, the total length field within IP header is used. | Mbyte |
| | N. of Samples | Total number of samples captured during the sampling process. | #samples |
| **Throughput Estimation** | **Measurement Goal:** Evaluate the sampling technique accuracy when estimating throughput. | | |
| | **Metric** | **Description** | **Definition / Units** |
| | Throughput | Quantifies the volume of traffic transferred per unit of time (data rate). | Ratio between the total amount of traffic transmitted and the corresponding time interval (kbps) |
| | Coefficient of Variation | Measures the variability of packet time series. In network measurements, helps to identify and characterize the traffic burstiness. | Ratio between sample standard deviation and throughput. |
| | Peak-to-average ratio | Complementary descriptive statistics to measure the variability of the packet time series. | Ratio between the peak and average throughput in a measurement interval. |
| | Correlation | Expresses the correlation between the sampled traffic and total traffic, studying statistically the relationship between the corresponding series. | Coefficient of correlation $\rho$ (Pearson method [20]): $0.7 < \rho \leq 1$, strong correlation; $0.3 < \rho \leq 0.7$, moderate correlation; $0 \leq \rho \leq 0.3$, weak correlation |
| | Relative Mean Error | Assesses the discrepancy between the mean of the total traffic and its sampled version. | $$RME = \frac{\lvert M_{total} - M_{estimated} \rvert}{M_{total}} \quad (3)$$ $M_{total}$ is the average throughput of total traffic; $M_{estimated}$ is the average throughput of the sampled traffic [7]. |

## 5. Evaluation Results

### 5.1. Overhead reduction

The benchmarking carried out show that MuST leads to a significant overhead reduction for all traffic types under consideration. Figure 4 illustrates the number of packets, the volume of data and the number of samples[4] for all traffic scenarios when applying each sampling technique against the total traffic. Note that, on average, the sampled traffic resulting from MuST corresponds to 5.8% of total data in the original trace. Table 5 details the results, quantifying the overhead for each traffic type and sampling technique under analysis.

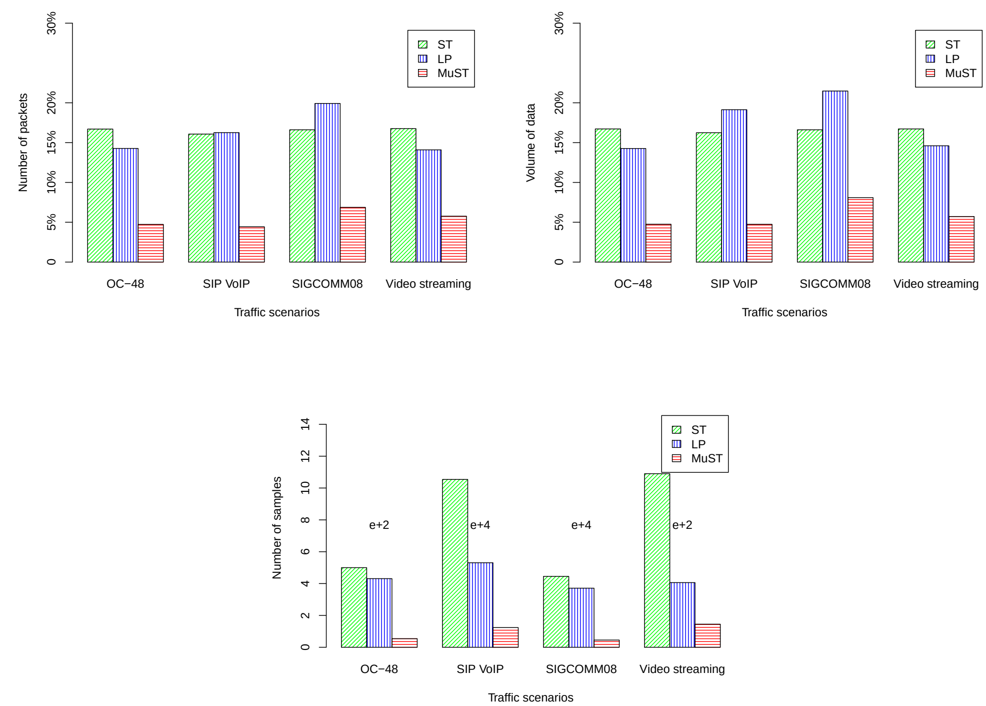



Figure 4: Overhead reduction using ST, LP and MuST

[4]In this case, the use of scientific notation aims to accommodate the different scales for the traffic traces considered.

Table 5: Overhead reduction for all traffic types

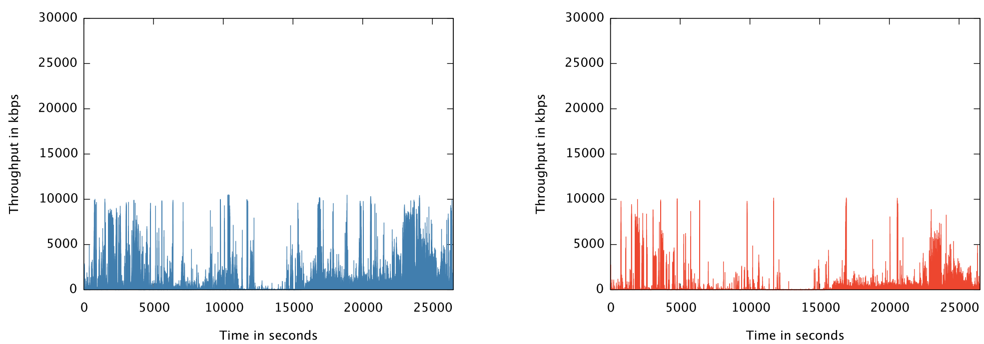
| **Traffic / Parameter** | **Total** | **ST** | **LP** | **MuST** |
|---|---|---|---|---|
| **OC-48** | | | | |
| **Number of packets** | 6550395 | 1094002 | 935066 | 310427 |
| **Data volume (MBytes)** | 3189.67 | 533.39 | 454.92 | 151.70 |
| **Number of samples** | | 500 | 431 | 54 |
| **SIP VoIP** | | | | |
| **Number of packets** | 1172378 | 188471 | 190572 | 51939 |
| **Data volume (MBytes)** | 168.68 | 27.41 | 32.27 | 8.0 |
| **Number of samples** | | 105385 | 53081 | 12380 |
| **SIGCOMM08** | | | | |
| **Number of packets** | 4513615 | 749977 | 899145 | 323467 |
| **Data volume (MBytes)** | 2382.92 | 395.97 | 511.93 | 192.82 |
| **Number of samples** | | 44511 | 37104 | 4534 |
| **Video streaming** | | | | |
| **Number of packets** | 97248 | 16303 | 13705 | 5606 |
| **Data volume (MBytes)** | 101.76 | 17.02 | 14.86 | 5.83 |
| **Number of samples** | | 1090 | 406 | 145 |

Comparing to the systematic technique, the multiadaptive technique reduces in 66.2% the number of collected packets, in 63.2% the total volume of sampled data, and in 88.7% the number of samples used to characterize all traffic scenarios. Considering all traffic scenarios and taking the adaptive LP technique as reference, MuST achieves decreases of 66% in the number of collected packets, of 65.6% in the total data volume, and of 81.1% in the number of processed samples.

These results clearly attest the performance improvement of applying MuST technique in measurement points. However, to prove measurement efficiency the statistical representativeness of the sampled traffic has to be evaluated, verifying its ability to capture the real traffic behavior. This evaluation is provided below.

### *5.2. Throughput estimation*

To describe traffic behavior, a visual representation of network throughput is commonly available in monitoring tools, illustrating the network workload during each measurement interval. Figure 5 presents the instantaneous throughput measured in 1s time intervals for SIGCOMM08 traffic. As depicted, the multiadaptive sampling technique represents closely the total traffic behavior when visually compared.




(a) Total traffic (b) MuST

Figure 5: Throughput for SIGCOMM08 traffic

As regards throughput estimation, Table 6 presents the statistical comparison resulting from applying each sampling technique for all traffic traces. The results show that ST and MuST techniques achieve high-quality representations of traffic behavior when comparing average throughput, peak-to-average ratio, CV and correlation with the corresponding total traffic statistics. The adaptive LP technique tends to overestimate throughput when using low-rate traffic traces (e.g. VoIP traffic). The best match of this technique was achieved for OC-48 trace, which represents traffic in a high-speed link of an ISP.

The correlation analysis of the instantaneous throughput time series (considering measurement intervals of 1s), obtained comparing each sampling strategy to the original trace, presents high coefficients of correlation (above 0.81), corroborating statistically the visual comparison illustrated in Figure 5. The accuracy achieved using MuST is ratified when considering the relative mean error of the estimated average throughput (see Table 6).

For a more detailed analysis, Figure 6 presents the relative error spread for each sampling technique when applied to the SIGCOMM08 trace. As shown, for all tested techniques the error distribution is similar, concentrated near to zero and with a well-bounded tail. The figure also shows that the higher relative errors result from overestimation, expressed by the positive tail in the histogram.

Table 6: Overall estimated behavior

| Traffic / Parameter | Total | ST | LP | MuST |
|---|---|---|---|---|
| **SIGCOMM08** | | | | |
| **Throughput** | 730.95 | 728.76 | 1130.27 | 765.45 |
| **RME** | | 0.002 | 0.54 | 0.04 |
| **Peak-to-average ratio** | 14.39 | 15.16 | 10.05 | 13.304 |
| **CV** | 1.90 | 2.07 | 1.56 | 2.07 |
| **Correlation** | | 0.91 | 0.91 | 0.90 |
| **SIP VoIP** | | | | |
| **Throughput** | 21.85 | 21.30 | 50.49 | 21.49 |
| **RME** | | 0.02 | 1.31 | 0.01 |
| **Peak-to-average ratio** | 158.40 | 132.14 | 58.00 | 118.57 |
| **CV** | 5.28 | 5.41 | 2.90 | 4.64 |
| **Correlation** | | 0.71 | 0.88 | 0.80 |
| **OC-48** | | | | |
| **Throughput** | 87103.35 | 87390.26 | 86467.86 | 87409.52 |
| **RME** | | 0.003 | 0.007 | 0.003 |
| **Peak-to-average ratio** | 1.40 | 1.35 | 1.34 | 1.19 |
| **CV** | 0.10 | 0.09 | 0.10 | 0.16 |
| **Correlation** | | 0.84 | 0.81 | 0.82 |
| **Video streaming** | | | | |
| **Throughput** | 1275.84 | 1279.17 | 2998.53 | 1268.58 |
| **RME** | | 0.002 | 1.35 | 0.005 |
| **Peak-to-average ratio** | 19.13 | 19.98 | 8.41 | 11.02 |
| **CV** | 3.93 | 3.96 | 2.50 | 3.01 |
| **Correlation** | | 0.99 | 0.98 | 0.98 |

Regarding the multiadaptive technique, when analyzing the trade-off between overhead and accuracy reduction in Tables 5 and 6, it is clear that this technique promotes a significant improvement compared to the other sampling techniques considered. These results show that, despite the significant reduction on the traffic volume considered, the MuST ability to capture the real traffic behavior correctly is not compromised. For several statistical parameters, the proposed technique outperforms the conventional techniques, which is even more relevant attending to the significant decrease on measurement overhead.

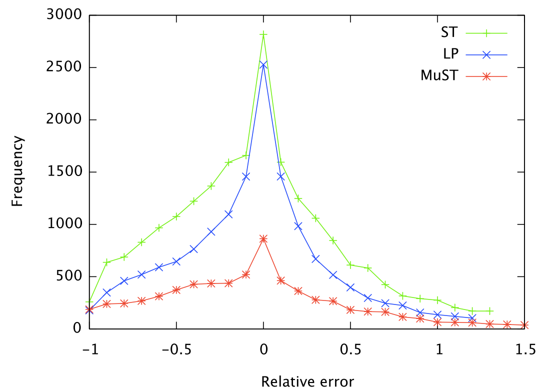


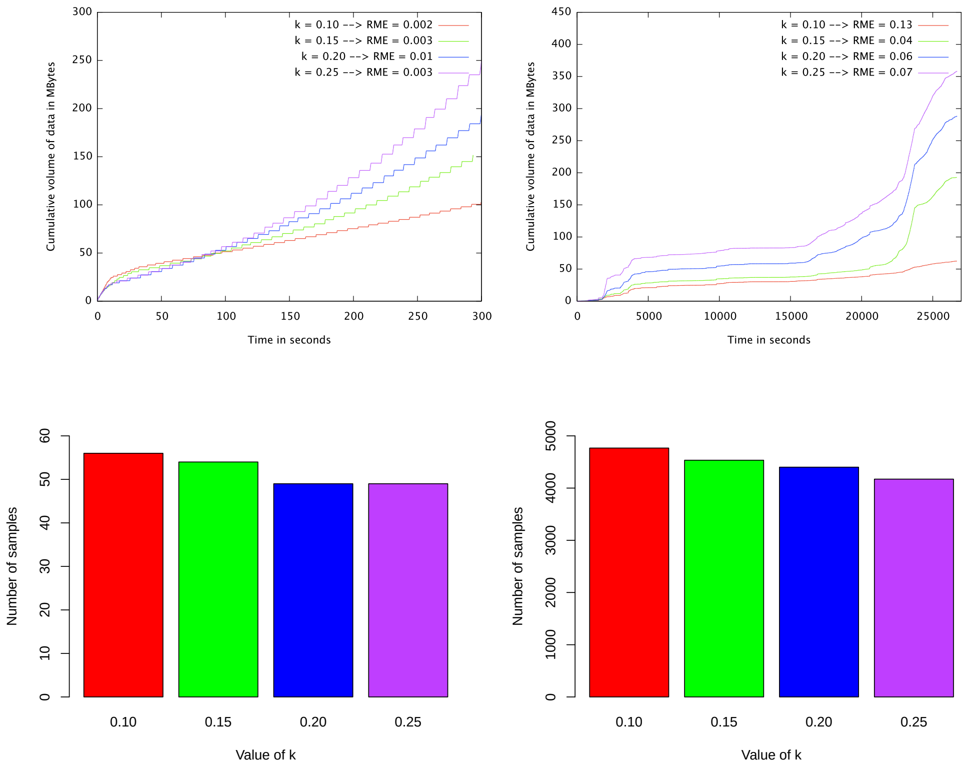

Figure 6: Histogram of relative error for SIGCOMM08 traffic

### *5.3. Studying the impact of k*

The following tests aim to assess the impact of $k$ when adapting the sample size (see rules in Table 2) allowing to evaluate the performance of MuST for two distinct traffic scenarios: one typically more regular (OC-48) and other with higher variability (SIGCOMM08). For each traffic type, the analysis is focused on the cumulative volume of sampled data (in Mbytes), RME and number of samples, for different values of $k$. As the results illustrate (see Figure 7), a $k = 0.10$ leads to the lowest volume in sampled data, however, it also causes an increase in RME (one order of magnitude) in the presence of bursty traffic. As shown in Figure 7 (b), with $k = 0.10$, the adaptive behavior of MuST is less sensitive in capturing traffic bursts (see behavior around 22500s). The results also show that a value of $k = 0.15$ presents a good compromise between the variables under study (data volume and RME) for both traffic types, although it does not lead to the smallest number of samples for the values of $k$ considered. The obtained differences for the number of samples are, however, not significant, meaning that MuST is well-behaved for $k = 0.15$.

### *5.4. Studying the impact of the order of prediction*

Similarly to the study of $k$ presented above, the following experiments aim at evaluating the impact of varying the order of prediction $N$ on MuST performance. Figure 8 presents the results comparing the overhead and relative mean error for OC-48 and SIGCOMM08 traces and distinct values of

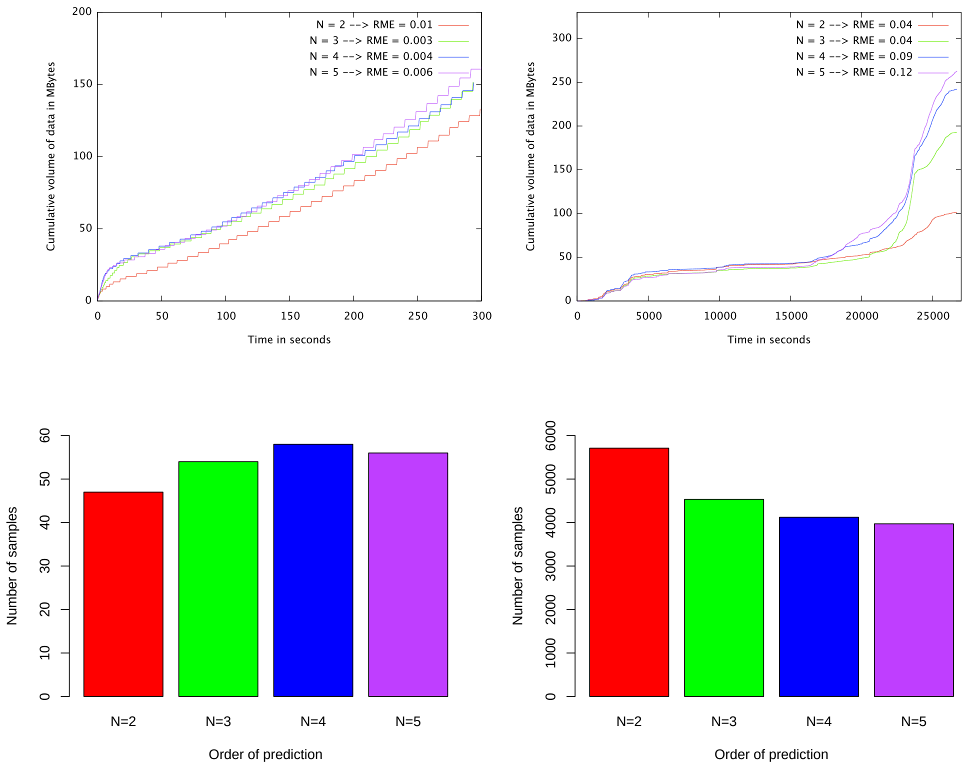



(a) OC-48 (b) SIGCOMM08

Figure 7: Impact of $k$ on MuST performance: Cumulative data volume (up); Number of samples (down)

$N$. Notice that higher values of $N$ correspond to configure the adaptive traffic selection process with more information of past samples, i.e., including more past memory. Generally, a larger past memory leads to less reactive mechanisms, and shorter past memory improves the reactivity to short term traffic fluctuations. The degree of this reactivity may affect the stability of sampling mechanisms.

According to the obtained results, an increase in $N$ conducts to an overhead increase regarding the volume of data collected, for all traffic types. However, for $N = 5$, it is clear that collecting more data does not lead to higher accuracy in the throughput estimation (RME); this is particularly visible in Figure 8 (b). This means that lower reactivity also reduces the ability to correctly measure bursty traffic. Nevertheless, when taking $N = 2$, cor-




(a) OC-48 (b) SIGCOMM08

Figure 8: Impact of $N$ on MuST performance: Cumulative data volume (up); Number of samples (down)

responding to the smallest overhead regarding the amount of sampled data, it does not imply a gain in accuracy, as visible for OC-48 traffic. Considering the number of samples, varying $N$ does not provoke a linear distribution for all traffic types. This is related to each particular traffic characteristics, which is inline with the adaptive behavior of MuST when facing traffic fluctuations, described in Section 3.2.

The results in the previous sections were obtained with an order of prediction $N = 3$, which represents a good compromise between overhead and accuracy. For the heaviest traffic trace (SIGCOMM08) the results could be further improved considering $N = 2$ as the evident overhead reduction does not penalize RME.

Globally, the values obtained for RME are very low for all traffic scenarios.

Therefore, depending on the objective or usefulness of throughput estimation, the value of $N$ can be tuned to achieve an optimal compromise between overhead and efficiency.

### *5.5. Studying the impact of the activity period*

According to [9], adaptive sampling techniques tend to be less accurate when the network activity is low. The tests included in this section intent to assess MuST ability in estimating network throughput correctly irrespectively of the network load period.

When observing Figure 5 it is clear that the time period between 12500s and 14000s corresponds to a low activity period; conversely, the period between 23000s and 25000s corresponds to a peak period in network activity. Therefore, the performance of each sampling technique was evaluated for these two time intervals in order to verify its versatility in estimating network behavior despite the ongoing network activity.

Table 7: Impact of network activity - SIGCOMM08

| **Low network load** | | | | |
|---|---|---|---|---|
| **Parameter** | **Total traffic** | **ST** | **LP** | **MuST** |
| **Number of packets** | 5291 | 592 | 257 | 210 |
| **Data volume (MBytes)** | 1.19 | 0.15 | 0.05 | 0.05 |
| **Number of samples** | | 2467 | 904 | 305 |
| **Throughput (kbps)** | 236.03 | 266.33 | 4.96 | 298.31 |
| **RME** | | 0.12 | 0.97 | 0.26 |
| **High network load** | | | | |
| **Parameter** | **Total traffic** | **ST** | **LP** | **MuST** |
| **Number of packets** | 901589 | 149634 | 226401 | 105985 |
| **Data volume (MBytes)** | 661.90 | 110.09 | 166.02 | 78.44 |
| **Number os samples** | | 3333 | 5183 | 245 |
| **Throughput (kbps)** | 769.81 | 771.46 | 1624.14 | 771.26 |
| **RME** | | 0.002 | 1.1 | 0.001 |

The results presented in Table 7 corroborate the results obtained in [9] as the LP technique clearly underestimates network throughput for low activity periods. Moreover, this technique is also inaccurate for high activity periods, overestimating throughput. Irrespectively of the activity period, ST and MuST techniques achieve positive results considering the estimation accuracy. However, the efficiency of MuST is higher as the estimation process involves less overhead, when considering all the related metrics (number of samples, number of packets and data volume). For high activity periods the accuracy-overhead trade-off was even more encouraging.

The above results evince that the proposed MuST technique is a step forward regarding classic adaptive sampling techniques, being simultaneously more accurate and versatile than its competitors.

## 6. Conclusions

Although current traffic measurement sampling techniques aim at estimating network parameters correctly, they do not address efficiency as a major concern. Beyond high accuracy, sampling techniques need to be sensitive enough to identify critical periods of network activity, adjusting their dynamics in order to reduce measurement overhead.

In this context, this paper has presented MuST, a multiadaptive traffic sampling technique able to improve the trade-off between network parameters estimation overhead and accuracy. By changing both the interval between samples and the sample size according to the observed network activity, this technique was able to capture correctly network throughput with very low overhead, particularly in periods of high activity, in which the network operation is critical.

The performance of this technique was evaluated using the Systematic Time-based and Adaptive Linear Prediction as comparative techniques. Using real traffic traces representing distinct profiles, the study provides a comparative statistical analysis of measurement overhead and estimation accuracy, under distinct traffic load scenarios. This analysis has evinced the effectiveness and flexibility of the present proposal, outperforming the conventional techniques.

To improve sampling efficiency, the impact of the order of prediction on the results was evaluated. A similar analysis was also carried out for the configuration parameters of the algorithm, attesting the values adopted as suitable for the evaluated traffic types.

The obtained results demonstrated that the self-adaptability and simplicity of MuST are important issues for achieving a versatile sampling solution which can be explored to improve network measurements efficiency over distinct traffic conditions.

## References


[1] S. K. Thompson, Sampling, 2nd Edition, Wiley-Interscience, 2002.

[2] T. Zseby, M. Molina, N. Duffield, Sampling and Filtering Techniques for IP Packet Selection, RFC 5475 (Mar. 2009).

[3] C. Henke, C. Schmoll, T. Zseby, Empirical evaluation of hash functions for packetid generation in sampled multipoint measurements, in: S. Moon, R. Teixeira, S. Uhlig (Eds.), Passive and Active Network Measurement, Vol. 5448 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2009, pp. 197–206.

[4] R. Serral-Gracia, A. Cabellos-Aparicio, J. Domingo-Pascual, Packet loss estimation using distributed adaptive sampling, in: Network Operations and Management Symposium Workshops, 2008. NOMS Workshops 2008. IEEE, 2008, pp. 124 –131.

[5] J. Sommers, P. Barford, N. Duffield, A. Ron, Improving accuracy in end-to-end packet loss measurement, in: Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications, SIGCOMM '05, ACM, New York, NY, USA, 2005, pp. 157–168.

[6] Z.-G. Hu, D.-L. Zhang, C.-P. Hou, J.-S. Zhang, Adaptive sampling algorithm of network round-trip time, Journal of Computer Applications 30 (2) (2010) 319–322.

[7] A. Dogman, R. Saatchi, S. Al-Khayatt, An adaptive statistical sampling technique for computer network traffic, in: Communication Systems Networks and Digital Signal Processing (CSNDSP), 2010 7th International Symposium on, 2010, pp. 479 –483.

[8] T. Zseby, T. Hirsch, B. Claise, Packet sampling for flow accounting: Challenges and limitations, in: M. Claypool, S. Uhlig (Eds.), Passive and Active Network Measurement, Vol. 4979 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2008, pp. 61–71.

[9] C. Hu, S. Wang, J. Tian, B. Liu, Y. Cheng, Y. Chen, Accurate and efficient traffic monitoring using adaptive non-linear sampling method, in:

INFOCOM 2008. The 27th Conference on Computer Communications. IEEE, 2008, pp. 26 –30.

[10] A. N. Mahmood, J. Hu, Z. Tari, C. Leckie, Critical infrastructure protection: Resource efficient sampling to improve detection of less frequent patterns in network traffic, Journal of Network and Computer Applications 33 (4) (2010) 491 – 502.

[11] J. Giertl, J. Baca, F. Jakab, R. Andoga, Adaptive sampling in measuring traffic parameters in a computer network using a fuzzy regulator and a neural network, Cybernetics and Systems Analysis 44 (2008) 348–356.

[12] Q. Xin, L. Hong, L. Fang, A modified flc adaptive sampling method, in: Communications and Mobile Computing, 2009. CMC '09. WRI International Conference on, Vol. 2, 2009, pp. 515 –520.

[13] E. A. Hernandez, M. C. Chidester, A. D. George, Adaptive sampling for network management, Journal of Network and Systems Management 9 (2001) 409–434.

[14] Y. Lu, C. He, Resource allocation using adaptive linear prediction in wdm/tdm epons, AEU - International Journal of Electronics and Communications 64 (2) (2010) 173 – 176.

[15] Y. Wei, J. Wang, C. Wang, A traffic prediction based bandwidth management algorithm of a future internet architecture, Intelligent Networks and Intelligent Systems, International Workshop on 0 (2010) 560–563.

[16] J. Silva, S. Lima, Multiadaptive sampling for lightweight network measurements, in: Computer Communications and Networks (ICCCN), 2012 21st International Conference on, 2012, pp. 1 –7.

[17] B.-Y. Choi, S. Bhattacharyya, On the accuracy and overhead of Cisco sampled Netflow, SIGMETRICS Perform. Eval. Rev. 33 (3) (2005) 18–23.

[18] A. Schulman, D. Levin, N. Spring, CRAWDAD trace umd/sigcomm2008/pcap/ethernet (v. 2009-03-02), Downloaded from http://crawdad.cs.dartmouth.edu/umd/sigcomm2008/pcap/Ethernet (Mar. 2009).

[19] C. Shannon, E. Aben, kc claffy, D. Andersen, N. Brownlee, The CAIDA OC48 traces dataset - 20030424-005000-0-anon, Downloaded from http://www.caida.org/data/passive/passive_oc48_dataset.xm (Apr. 2003).

[20] K. Pearson, On a New Method of Determining Correlation Between a Measured Character A, and a Character B, of which Only the Percentage of Cases Wherein B Exceeds (or Falls Short of) a Given Intensity is Recorded for Each Grade of A, 1st Edition, Vol. 7, Biometrika, 1909, pp. 96–105.

## Appendix A. MuST algorithm

This appendix includes the pseudocode for Algorithm 1 which allows to predict the reference parameter, and for Algorithm 2 which represents the overall sampler operation, in which all decisions regarding the new sample size and interval between samples are made attending to the predicted reference parameter.

**input** : $X$ - Reference parameter vector
$\Delta T$ - Current time interval between samples
$T$ - Times vector
**output**: forecast - Reference parameter prediction

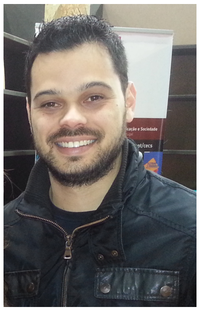
**for** $i \leftarrow 1$ **to** $order - 1$ **do**
  $sum \leftarrow sum + abs((X[i+1] - X[i])/T[i])$;
**end**
$forecast \leftarrow X[order] + ((\Delta T/(order - 1)) * sum$;
**return** $(forecast)$;

**Algorithm 1:** Pseudocode for the reference parameter predictor

**input** : $\Delta T_{current}$ - Initial interval between samples
$\Delta S_{current}$ - Initial sample size
**output**: Sampled traffic

**begin**
  **for** $i \leftarrow 1$ **to** *order* **do**
    newSample($\Delta T_{current}$, $\Delta S_{current}$); */* new sample capture */*;
    $X[i] \leftarrow$ referenceParameter(); */* stores ref. par. of last sample */*;
    $T[i] \leftarrow \Delta T_{current}$; */* kept unchanged */*
  **end**
  **repeat**
    */* Predictor() corresponds to Algorithm 1 */*;
    $X_p \leftarrow$ Predictor($X$, $\Delta T$, $T$);
    newSample ($\Delta T_{current}, \Delta S_{current}$);
    $S \leftarrow$ referenceParameter();
    $m \leftarrow 1$;
    **if** $S$ **then**
      $m \leftarrow X_p/S$
    **end**
    **case** $[m < m_{min}]$ */* underestimation */*
      $\Delta T_{next} \leftarrow m * \Delta T_{current}$;
      $\Delta S_{next} \leftarrow m * \Delta S_{current}$;
    **case** $[m_{min} \leq m \leq m_{max}]$ */* correct estimation*/*
      $\Delta T_{next} \leftarrow \Delta T_{current}$;
      $\Delta S_{next} \leftarrow \Delta S_{current}$;
    **case** $[m > m_{max}]$ */* overestimation */*
      $\Delta T_{next} \leftarrow 2 * \Delta T_{current}$; $k$=0.15;
      $\Delta S_{next} \leftarrow (1 + k) * \Delta S_{current}$;
    */* Interval Between Samples thresholds (sec) */*
    **if** $\Delta T_{next} < MinIBS$ **then**
      $\Delta T_{next} \leftarrow MinIBS$ ;
    **end**
    **if** $\Delta T_{next} > MaxIBS$ **then**
      $\Delta T_{next} \leftarrow MaxIBS$;
    **end**
    */* Sampling Size thresholds (sec) */*
    **if** $\Delta S_{next} < MinSS$ **then**
      $\Delta S_{next} \leftarrow MinSS$;
    **end**
    **if** $\Delta S_{next} > MaxSS$ **then**
      $\Delta S_{next} \leftarrow MaxSS$;
    **end**
    $\Delta T_{current} \leftarrow \Delta T_{next}$; $\Delta S_{current} \leftarrow \Delta S_{next}$;
    updateVectors($\Delta T_{current}, S$); */* update vectors T and X */*
  **until** *endOfSampling*;
**end**

**Algorithm 2:** Multiadaptive technique pseudocode



**João Marco Silva** graduated in 2008 from University Federal of Sergipe, Brazil, received his MSc in Networks and Communication Services from the University of Minho, Braga, Portugal, in 2011. He is currently a PhD student in Informatics at the same University. His research interests include traffic sampling, network management, network monitoring, traffic modeling, quality of service, traffic classification and characterization.

**Paulo Carvalho** graduated in 1991 and received his PhD degree in Computer Science from the University of Kent at Canterbury, United Kingdom, in 1997. He is currently Associate Professor in Computer Communications, Department of Informatics, University of Minho, Portugal. His main research interests include broadband technologies, multiservice networks, mobile networks, and teletraffic analysis.

**Solange Rito Lima** graduated in 1992, received her MSc and PhD degrees in Computer Science from the University of Minho, Portugal, in 1997 and 2006, respectively. She is currently Assistant Professor in Computer Communications at University of Minho. Her research interests include multiservice networks and protocols, traffic control, monitoring issues in IP networks and emerging network technologies.