Analysing Traffic Flows Through Sampling: A Comparative Study

João Marco C. Silva, Paulo Carvalho and Solange Rito Lima
Centro Algoritmi, Universidade do Minho, Braga, Portugal
Email: joaomarco@di.uminho.pt, pmc@di.uminho.pt, solange@di.uminho.pt

***Abstract*—Understanding network workload through the characterization of network flows, being essential for assisting network management tasks, can benefit largely from traffic sampling as long as an accurate snapshot of network behavior is captured. This paper is devoted to evaluate the real applicability of using sampling to support flow analysis. Considering both classical and emerging sampling techniques, a comparative performance study is carried out to assess the accuracy of estimating flow parameters through sampling. After identifying the main building blocks of sampled-based measurements, a sampling framework has been implemented to provide a versatile and fair platform for carrying out the testing and comparison process. Through an encompassing coverage of representative sampling techniques, the present study aims to provide useful insights regarding the use of sampling in traffic flow analysis.**

I. Introduction

The importance of traffic characterization for planning and managing effectively todays networks is undeniable. Associating network traffic with the corresponding applications and studying flows characteristics allows gathering valuable information about network usage and, hence, devising solutions able to accommodate applications' requirements. However, the massive volume of data to process impair applying traffic classification and characterization algorithms efficiently. In this way, packet sampling is viewed as a promising strategy to cope with large data amounts by resorting to a subset of the entire traffic to classify and characterize network usage.

Most of the current research efforts in developing sampling-based measurement tools are usually focused on classical sampling approaches (*i.e.,* systematic and random packet capture), not covering the use of recent sampling techniques. This includes adaptive sampling approaches deployed to bring flexibility and scalability to network management tasks. In addition, many of the current classification approaches do not consider sampling, assuming that the input data is based on the entire traffic.

The ability to generate accurate inputs about traffic profiles is a crucial requirement imposed to a sampling technique in order to supply the traffic classifier properly. To reduce the computational burden involved, traffic data is usually handled as flow statistics, including flow identification, flow size (number of packets), flow load and duration.

In this context, the main objective of the present work is to assess the applicability and performance of sampling techniques for network flow analysis. This involves analyzing the accuracy of flow statistics produced by current and recently proposed sampling techniques in capturing the characteristics of traffic flows crossing the networks. The test methodology resorts to a sampling framework developed with the purpose of implementing different sampling techniques in a flexible way, allowing the combination of their inner characteristics in forthcoming operational scenarios. The performance study is carried out using recent traffic traces gathered at Portuguese National Statistics Institute network.

Facing the above considerations, the contributions of this work are threefold: (i) identification of functional layers and tasks involved in a sampling-based measurement architecture; (ii) adoption of a unified sampling taxonomy identifying the inner characteristics distinguishing sampling techniques (providing the basis for the sampling framework); and (iii) evaluation of the impact of using distinct sampling techniques for network flow characterization.

This paper is organized as follows: the related work is discussed in Section II; the sampling-based measurement architecture and corresponding traffic sampling taxonomy are introduced in Section III; the methodology of tests used in the flow analysis is presented in Section IV; the performance evaluation results are discussed in Section V; and the conclusions are included in Section VI.

II. Related work

The usefulness of traffic sampling has been explored in multiple network tasks, namely: network security - for anomaly and intrusion detection, botnet and DDoS identification [1]; SLA compliance and QoS control - for estimating parameters such as packet delay, jitter and loss [2]; traffic engineering - to assist traffic classification and characterization [3].

The importance of accuracy in traffic classification and characterization based on sampled packets is increasing at the same pace as traffic sampling is becoming mandatory to reduce the burden of traffic analysis. Current research on this topic is typically focused on identifying the complexity and limitations introduced by the missing data during the analysis of sampled traffic [4] [5], the statistical performance of different classification approaches, such as machine learning-based [3] [6]. However, these works only consider the widely deployed systematic count-based or random count-based sampling approaches, in which packets are selected according to their position in the stream under analysis by resorting to a deterministic or probabilistic function [7], respectively. Recent approaches, such as adaptive sampling, or even already

978-1-4673-7194-0/15/$31.00 ©2015 IEEE

standardized proposals, such as time-based sampling [7], are usually not covered.

Facing this gap within sampling research, this paper aims to provide useful insights regarding the use of sampling for flow analysis, resorting to an encompassing coverage of representative sampling techniques applied to real traffic.

III. A SAMPLING-BASED MEASUREMENT ARCHITECTURE

A sampling-based measurement architecture can be viewed as comprising three planes, as illustrated in Figure 1. The *management plane* includes tasks deployed directly in measurement points or in external management entities. It is responsible for selecting and configuring the measurement points to participate in the sampling process, and estimating the relevant metrics using data reports sent by the control plane. The metrics estimation resorts to flow statistics which are received from the lower plane following IP Flow Information eXport (IPFIX) specifications.

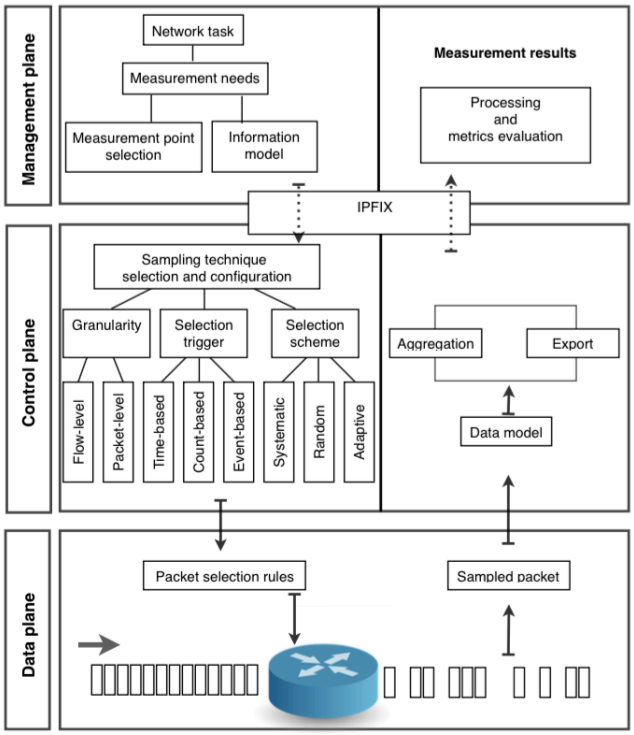


Fig. 1. Architecture description

A modular design of the *control plane* allows a flexible sampling technique selection and configuration. Considering IETF PSAMP work [7] and recent sampling proposals, a sampling taxonomy is used to identify the inner characteristics distinguishing sampling techniques and also supporting the definition of new sampling techniques which can be adjusted to each traffic/service measurement scenario. The proposed taxonomy defines that sampling techniques can be classified into three well-defined components according to the *granularity*, *selection scheme* and *selection trigger* in use. Then each component is further divided into a set of approaches commonly followed in existing sampling techniques.

- *Granularity* - identifies the atomicity of the element under analysis in the sampling process: in *flow-level* approach, the sampling process is only applied to packets belonging to a specific set of flows of interest; in *packet-level* approach, packets are eligible as independent entities.
- *Selection scheme* - identifies the function defining which traffic packets will be selected and collected; this scheme may follow a *systematic* approach, in which the process of packet selection is ruled by a deterministic function that imposes a fixed sampling frequency, independently of the packet contents or treatment; a *random* approach, that rules the sampling frequency through a random process, usually resorting to a pseudo-random generator or to a probabilistic function; or an *adaptive* approach, in which the sampling technique is endowed with the ability to change the selection of packets during the course of measurements aiming to identify the most important parts of a traffic stream according to the measurement needs or to save network resources during critical periods.
- *Selection trigger* - determines the spatial and temporal sample boundaries; it may use a *time-based* approach, in which the sampling beginning and end are driven based on the packets arrival time at the measurement point; a *count-based* approach, in which the sampling boundaries are defined based on the packet position in the incoming stream; or an *event-based* approach, in which the decision on when a sample starts and ends takes into account some particular event observed in the traffic being monitored. This event may be some value in the packet contents, the treatment of the packet at the measurement point or a more complex observation.

As presented in Figure 1, at the control plane, the sampled packets received from the data plane are processed and the relevant field contents are extracted according to the classification algorithm requirements. These values are then aggregated (both in time and space) and exported following IETF guidelines [8] [9], and using IPFIX specifications.

At data plane, traffic is collected from network interfaces by applying the sample rules defined in the control plane. The unprocessed packets are then reported to the control plane to be processed, simplifying the data plane.

Following this architecture, firstly introduced in [10], a framework implemented in Java using *libpcap*, connects the sampling components in order to enable a versatile deployment of sampling techniques. This framework can be applied to both online and offline measurement scenarios and currently supports a large number of classical and recently proposed sampling techniques. These techniques are summarized in Table I, as well as their components under the proposed taxonomy. The most relevant techniques for this work and corresponding notation are briefly explained in Section IV-A. As an example, the notation Syst identifies the systematic technique, the notation C, T and Evt identifies the selection trigger and _F a flow-level granularity.

TABLE I
SAMPLING TECHNIQUES AVAILABLE IN THE FRAMEWORK

Technique	Granularity		Selection trigger			Selection scheme		
	Packet-level	Flow-level	Count-based	Time-based	Event-based	Systematic	Random	Adaptive
SystC	✓		✓			✓		
SystC_F		✓	✓			✓		
SystT	✓			✓		✓		
SystT_F		✓		✓		✓		
SystEvt	✓				✓	✓		
SystEvt_F		✓			✓	✓		
RandC	✓		✓				✓	
RandC_F		✓	✓				✓	
LP	✓			✓				✓
LP_F		✓		✓				✓
MuST	✓			✓				✓
MuST_F		✓		✓				✓

IV. FLOW ANALYSIS: METHODOLOGY OF TESTS

In order to assess the traffic sampling impact on flow analysis, the methodology of tests consists in applying different sampling techniques to real traffic scenarios, evaluating the estimation accuracy of common flow parameters. The study also extends the analysis to flow classification for different sampling frequencies of the same technique. The following sections detail the sampling techniques under evaluation, the flow parameters used in the comparative study, and the traffic scenario in use.

A. Sampling techniques

The sampling techniques evaluated correspond to: (i) the main techniques currently used in network measurement tools, i.e., *SystC - systematic count-based* and *RandC - random count-based* [7], following a deterministic or uniform random function for packet selection, respectively; (ii) *SystT - systematic time-based*, a technique also defined in [7] and scarcely deployed in current measurement points due to the computational complexity in manipulating timestamps with fine grain precision; and (iii) two adaptive techniques, *i.e.*, *LP - adaptive linear prediction* [11] and *MuST - multiadaptive sampling* [12]. In the present study, due to the nature of the traffic classification process where traffic is addressed as an aggregate, all these techniques have packet-level granularity, being flow identification and statistics then derived from the sampled packets.

B. Comparative parameters

For comparing the ability of distinct sampling techniques in assisting network flow analysis correctly, several flow parameters are considered, namely: (i) the amount of flows identified; (ii) the percentage of heavy-hitter (HH) flows identified, where the notion of heavy hitter refers to 20% of the largest flows in terms of size (number of packets) [13]; (iii) the utilization share at transport level; (iv) the utilization share at application level; and, (v) the accuracy of load estimations for the identified flows.

Considering that when flow characterization is based on sampling only a subset of the packets is available, estimating the underlying metrics involves the usage of statistical estimators to overcome missing data. In particular, the load estimation of each flow is an additional challenge as it needs to be often inferred from a small number of collected packets. Following the discussion in [4] and the notation in Table II, the specific estimators in this comparative work are as follows:

- *Flow Mean Packet Size ($\bar{X}_f$):* the average number of sampled packet sizes of flow f.

$$\bar{X}_f = \frac{\sum_{i=1}^{n_f} X_i}{n_f} \tag{1}$$

- *Estimated Flow Size (S_f):* the estimated number of packets in flow f.

$$S_f = N * \frac{n_f}{n_s} \tag{2}$$

- *Estimated Flow Load (L_f):* the byte count of an individual flow f.

$$L_f = S_f * \bar{X}_f \tag{3}$$

TABLE II
NOTATION

X_i	the size of the ith sampled packet of flow f
n_f	number of sampled packets of flow f
n_s	total number of sampled packets
N	estimated total number of packets ($n_s/sampling\_frequency$)

Regarding the estimated flow load, this work applies an innovative way to assess accuracy by resorting to a nonparametric method to estimate the density distribution of load estimation (i.e., KDE - Kernel Density Estimation method) and thereby fostering the discussion on the estimation bias when applying each sampling technique. Each distribution corresponds to a nonparametric probability density function estimated using the Kernel method and a Gaussian smoothing scale. This method consists in drawing a continuous and smooth density distribution, weighted by the distance from a central value (the Kernel), where the population is inferred from a finite number of observations. In this context, as defined in [14], let $(L_{f1}, ..., L_{fn})$ be the estimated load of all identified flows (n) for which the density p is under evaluation. The shape of this function using the kernel estimator is given by:

$$\hat{p}_{bw}(L_f) = \frac{1}{n}\sum_{i=1}^{n} K(\frac{L_f - L_{fi}}{bw}), \tag{4}$$

where $K()$ is the kernel scaled by a Gaussian function, and bw is a smoothing parameter called bandwidth which defines the variance of the kernel in order to concentrate the density distribution within a specific interval. This interval is defined using the standard deviation of the smooth kernel when considering both unsampled traffic and traffic resulting from all sampling techniques.

When useful, the present study includes the mean absolute error (MAE) and the mean square error (MSE) of the estimated

values, which are commonly used to evaluate the accuracy of estimators [4]. Note that, the evaluation of flow classification methodologies and tools is beyond the scope of this work, which resorts to a port-based classification technique for distinguishing flows.

C. Traffic scenario

The data used to evaluate sampling accuracy consists of four traffic traces collected from a high-utilization production network, providing multiple services such as videoconference, VoIP, distributed databases, private cloud, ftp, etc. Each trace, gathered at Portuguese National Statistics Institute network, corresponds to a twenty minutes capture in different workload periods aiming to include load scenarios in which flow analysis is commonly used. The traffic traces are then handled as an aggregate, reflecting a continuous and heterogeneous network activity period of 252,087 individual unidirectional flows, comprising nearly 3 million packets.

V. Evaluation Results

This section includes the main test results evaluating the performance of the sampling techniques described in Section IV. After performing an initial tuning of the systematic count-based technique (SystC) to assess the impact of sampling frequency, the discussion is focused on evaluating the accuracy of all sampling techniques under study when estimating the flow parameters defined in Section IV-B.

A. Identifying existing flows and heavy hitters

Intuitively, the sampling frequency is directly proportional to the estimation accuracy, as increasing the collected traffic results in a larger dataset for statistical analysis. While this may be true for analysis carried out using the same sampling technique, when the sampling selection scheme changes, the trade-off between overhead and accuracy may differ significantly. Even within the same technique, studying this trade-off may bring useful information for reducing the amount of traffic collected and processed as, for several flow parameters, this reduction does not have a significant impact on accuracy.

Considering the evaluation of SystC sampling technique, the results in Table III confirm that decreasing the number of sampled packets leads to a decrease on the number of flows identified. As illustrated in Figure 2(a), while SystC 1/8[1] identifies 40% of existing flows, SystC 1/256 only detects 2,3% of flows. However, reducing the sampling frequency does not impact significantly on the accuracy in the identification of the heavy-hitter flows. For instance, despite SystC 1/256 manipulates only 6% of the traffic processed by the technique SystC 1/16, the accuracy of heavy hitters identification is almost equivalent. This performance analysis is useful to guide activities in which accounting for heavy flows is relevant.

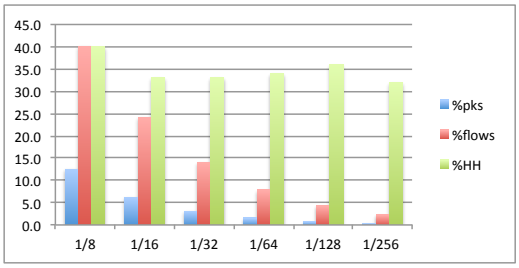
Regarding the different sampling techniques[2], a larger number of sampled packets also implies a larger number of flows identified, as illustrated in Table IV. Nevertheless, count-based techniques are more efficient as, for the same proportion of sampled packets, the percentage of identified flows is significantly higher. This is visible when comparing the results of SystC 1/8 and SystT in Figures 2(a) and (b), respectively. This is due to the distinct packet selection policies in use, as the process of packet selection in count-based techniques increases the probability of capturing distinct flows, contrarily to time-based techniques in which packets are selected sequentially.

TABLE III
SystC flow identification

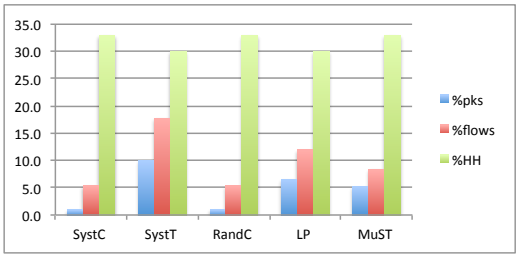
	# Sampled packets	# Distinct flows	% Heavy hitters
SystC 1/8	373124	101168	40%
SystC 1/16	186562	61022	33%
SystC 1/32	93280	35523	33%
SystC 1/64	46639	19854	34%
SystC 1/128	23319	10891	36%
SystC 1/256	11665	5889	32%

TABLE IV
Sampling techniques flow identification

	# Sampled packets	# Distinct flows	% Heavy hitters
SystC	29849	13652	33%
SystT	296579	44590	30%
RandC	29849	13577	33%
LP	196007	30186	30%
MuST	156382	21009	33%



(a) SystC



(b) Sampling Techniques

Fig. 2. Identifying flows - comparative analysis

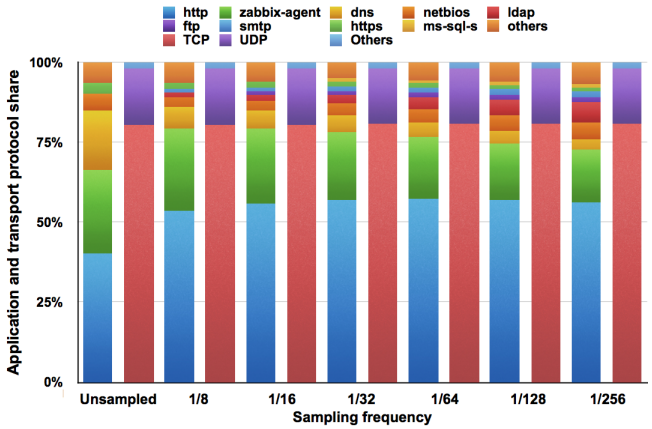
B. Utilization share at transport and application level

Regarding the analysis at transport level, the reduction on the number of sampled packets promoted by a lower sampling frequency of SystC does not affect accuracy, as presented in

[1]The notation SystC 1/8 denotes that one packet is collected from each eight incoming packets at the measurement point.

[2]Despite of the values presented in Table III, the following comparative evaluation tests use the frequency 1/100 for SystC and RandC techniques, as suggested in [6]. For SystT technique, the sampling frequency in use is 100/1000 as it led to the best results for the analysis performed.

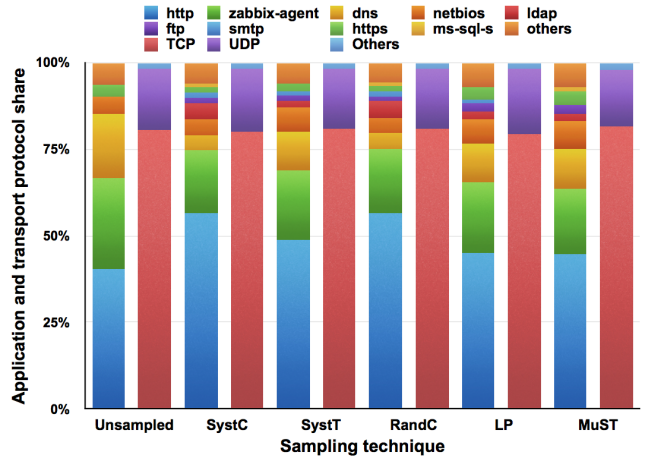
Figure 3 and confirmed by the low MAE and MSE throughout all frequencies. However, considering the application level, the estimated distribution is significantly affected, resulting in an overestimation of http traffic.



	1/8	1/16	1/32	1/64	1/28	1/56
MAE	0.0099	0.0464	0.1738	0.1337	0.1232	0.3045
MSE	0.0017	0.0026	0.0383	0.0262	0.0189	0.1383

Fig. 3. Analysis at application and transport level - systematic count-based

Although changing the sampling technique also maintains the accuracy in classifying the transport protocol (see Figure 4), the classification of applications shows more variability for the different techniques. As presented in Figure 4, time-based techniques lead to a more realistic distribution of the application share, with MuST providing the more accurate result. Globally, the results evince that an adequate yet small fraction of network traffic is able to provide a useful panoramic view of the protocolar mix of network flows.



	SystC	SystT	RandC	LP	MuST
MAE	0.2464	0.0999	0.0462	0.7521	0.7634
MSE	0.0817	0.0112	0.0330	0.7992	0.7671

Fig. 4. Analysis at application and transport level - comparative techniques

C. Load estimation

Attending to the formulation in Section IV-B, the results in Figure 5 show the distribution of the estimated flow load L_f (in logarithmic scale) when applying the different sampling approaches. The resulting graphics demonstrate the ability to represent the load distribution of all flows identified in the traffic trace, instead of only the more significant ones. This analysis plays a key role for traffic characterization and resource management activities.

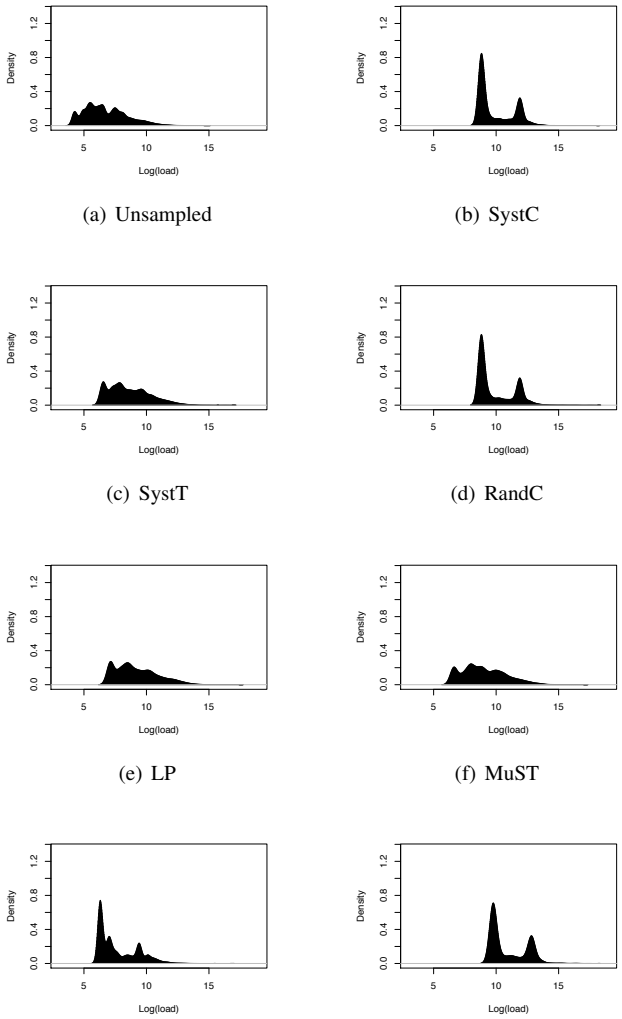
The results show that time-based techniques achieve a distribution closer to the real flow behavior (unsampled case in Figure 5 (a)) when compared with the count-based approaches (Figures 5 (b) and (d)). In addition, time-based approaches also lead to more accurate estimations of individual flows; this is observed through a better adjustment on the x-axis, meaning that the load estimations are closer to the real values. This suggests that a positive aspect (sparse packet selection) in flow identification becomes a drawback of the count-based techniques in flow dimensioning, since the current heuristics for flow load estimation consist in linear extrapolation proportional to the sampling frequency. This implies that a lower sampling frequency leads to overestimation of the flow load, deforming the density distribution to the right, and also concentrating the trend of load estimations (as detailed in Figure 5 (g) and (h), for two sampling frequencies of SystC technique), which evinces statistical loss of accuracy. This may interfere with network tasks in which the classification of small flows are of particular interest, such as intrusion detection and DDoS attacks.

Conversely, once time-based techniques select successive packets, the bursty behavior of larger flows tends to be better identified and dimensioned, resulting in more accurate flow load distributions, as presented in Figures 5(c), (e) and (f).

Table V includes statistics (mean, standard deviation and mode) to complement the above flow load analysis for the sampling techniques under study. As shown, the average flow load, the dispersion and the mode of the load estimations corroborate the behavior depicted in Figure 5. For the systematic count-based technique, as the sampling frequency decreases, the overestimation and concentration tendency is stressed.

TABLE V
FLOW LOAD STATISTICAL DESCRIPTION

	Mean	Standard deviation	Mode
Unsampled	6.08	1.48	4.18
SystC	9.78	1.36	8.79
SystT	8.04	1.47	4.49
RandC	9.78	1.37	8.79
LP	8.40	1.50	6.82
MuST	8.74	1.54	6.58
SystC 1/8	7.26	1.40	6.26
SystC 1/16	7.91	1.39	6.89
SystC 1/32	8.61	1.38	7.62
SystC 1/64	9.34	1.37	8.33
SystC 1/128	10.06	1.37	9.03
SystC 1/256	10.79	1.38	9.73



(a) Unsampled (b) SystC

(c) SystT (d) RandC

(e) LP (f) MuST

(g) SystC 1/8 (h) SystC 1/256

Fig. 5. Density of flow load estimation

VI. Conclusions

The present research work was focused on evaluating the performance of distinct traffic sampling techniques to support network flow analysis. Attending to the potential relevance of combining these tasks to allow cost-effective network management, evaluating the real applicability of current and emerging sampling techniques for flow characterization is an important step to avoid misleading estimation of network usage.

Globally, the obtained results evince that: (i) distinct sampling techniques led to variable yet reduced sampling data volumes attending the overall traffic; this gain in traffic handling has a cost regarding the small to moderate ratio of flows identified. Nevertheless, count-based techniques revealed to be more efficient as they allow to identify more flows with the same amount of sampled traffic. Heavy-hitter flows are generically detected in the same proportion for all sampling techniques, independently of the sampling frequency; (ii) despite the small number of detected flows, the protocolar analysis both at transport and application levels provided a comprehensive snapshot of the protocolar scene, being MuST the more accurate technique at application level. The sampling frequencies adopted for SystC technique did not affect the protocolar analysis significantly; (iii) flows load estimation is hard to achieve accurately as it needs to be inferred from a reduced number of packets of each flow. Using specific estimators to overcome missing data, time-based techniques are more effective in estimating flows load, leading to a probability density function close to the unsampled flow case.

In summary, a higher sampling frequency does not necessarily lead to more accuracy and, depending of the network tasks assisted by sampling, one can take advantage of either a low-frequency sampling or a specific sampling technique in order to improve the trade-off between overhead and accuracy when studying network flows behavior.

Acknowledgements

This work has been supported by FCT - *Fundação para a Ciência e Tecnologia* in the scope of the project UID/CEC/00319/2013.

References

[1] G. Androulidakis, V. Chatzigiannakis, and S. Papavassiliou, "Network anomaly detection and classification via opportunistic sampling," *Network, IEEE*, vol. 23, no. 1, pp. 6–12, 2009.

[2] Y. Gu, L. Breslau, N. Duffield, and S. Sen, "On Passive One-Way Loss Measurements Using Sampled Flow Statistics," in *INFOCOM 2009, IEEE*, 2009, pp. 2946–2950.

[3] D. Tammaro, S. Valenti, D. Rossi, and A. Pescapè, "Exploiting packet-sampling measurements for traffic characterization and classification," *International Journal of Network Management*, pp. 451–476, 2012.

[4] B.-Y. Choi and S. Bhattacharyya, "Observations on Cisco sampled NetFlow," *SIGMETRICS Perform. Eval. Rev.*, vol. 33, no. 3, pp. 18–23, 2005.

[5] T. Zseby, T. Hirsch, and B. Claise, "Packet Sampling for Flow Accounting: Challenges and Limitations," in *Passive and Active Network Measurement*, ser. LNCS. Springer, 2008, vol. 4979, pp. 61–71.

[6] V. Carela-Español, P. Barlet-Ros, A. Cabellos-Aparicio, and J. Solé-Pareta, "Analysis of the impact of sampling on NetFlow traffic classification," vol. 55, no. 5, pp. 1083–1099, Apr. 2011.

[7] T. Zseby, M. Molina, N. Duffield, S. Niccolini, and F. Raspall, "Sampling and Filtering Techniques for IP Packet Selection," RFC 5475, Internet Engineering Task Force, Mar. 2009.

[8] G. Muenz, B. Claise, and P. Aitken, "Configuration Data Model for the IP Flow Information Export (IPFIX) and Packet Sampling (PSAMP) Protocols," RFC 6728, Internet Engineering Task Force, 2012.

[9] B. Claise, G. Dhandapani, P. Aitken, and S. Yates, "Export of Structured Data in IP Flow Information Export (IPFIX)," RFC 6313, IETF, 2011.

[10] J. M. C. Silva, P. Carvalho, and S. R. Lima, "Enhancing Traffic Sampling Scope and Efficiency," in *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, Apr. 2013.

[11] E. A. Hernandez, M. C. Chidester, and A. D. George, "Adaptive Sampling for Network Management," *Journal of Network and Systems Management*, vol. 9, no. 4, pp. 409–434, 2001.

[12] J. M. C. Silva, P. Carvalho, and S. R. Lima, "A multiadaptive sampling technique for cost-effective network measurements," *Computer Networks*, vol. 57, no. 17, pp. 3357 – 3369, 2013.

[13] R. Krishnan, L. Yong, A. Ghanwani, N. So, and B. Khasnabish, "Mechanisms for Optimizing Link Aggregation Group (LAG) and Equal-Cost Multipath (ECMP) Component Link Utilization in Networks," RFC 7424, IETF, Jan. 2015.

[14] B. W. Silverman, *Density estimation for statistics and data analysis*. CRC press, 1986.