INTERNATIONAL JOURNAL OF COMMUNICATION SYSTEMS
Int. J. Commun. Syst. 2015; **00**:1–15
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/dac

Inside Packet Sampling Techniques: Exploring Modularity to Enhance Network Measurements

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

Centro Algoritmi, Universidade do Minho, 4750-057 Braga, Portugal

SUMMARY

Traffic Sampling is viewed as a prominent strategy contributing to lightweight and scalable network measurements. Although multiple sampling techniques have been proposed and used to assist network engineering tasks, these techniques tend to address a single measurement purpose, without detailing the network overhead and computational costs involved. The lack of a modular approach when defining the components of traffic sampling techniques also makes difficult their analysis. Providing a modular view of sampling techniques and classifying their characteristics is, therefore, an important step to enlarge the sampling scope, improve the efficiency of measurement systems, and sustain forthcoming research in the area. In this context, this paper defines a taxonomy of traffic sampling techniques based on a comprehensive analysis of the inner components of existing proposals. After identifying *granularity*, *selection scheme* and *selection trigger* as the main components differentiating sampling proposals, the study goes deeper on characterizing these components, including insights into their computational weight. Following this taxonomy, a general-purpose architecture is established to sustain the development of flexible sampling-based measurement systems. Traveling inside packet sampling techniques, this paper contributes to a clearer positioning and comparison of existing proposals, providing a road map to assist further research and deployments in the area.

Received ...

KEY WORDS: Packet sampling techniques; Traffic sampling classification; Traffic measurements; Traffic sampling taxonomy

1. INTRODUCTION

The growth in size and heterogeneity of today's communication networks has brought huge challenges to network planning and management activities. The need for efficient monitoring solutions, being crucial to assist service providers and network managers in these activities, is further stressed when considering recent paradigms such as Next Generation Networks and Cloud Computing, where aspects such as service convergence, mobility, virtualization and ubiquity are expected to coexist in a seamless network environment.

The most used and versatile strategy designed to assist network monitoring is traffic measurement. Traffic measurement techniques can be applied in real, emulated or simulated networks and the measurement points can be deployed directly in the network nodes, in dedicated equipment or in general-purpose devices connected to the network under analysis.

The huge traffic volume traversing high-capacity links, the distinct accuracy requirements associated with service types or monitoring activity, and the positioning of measurement points are main challenges when designing a measurement strategy. Assuming that monitoring activities should not interfere with the normal network operation, passive measurement methodologies adopt a non-intrusive approach, *i.e.*, they are based on real traffic in the network under analysis, in

*Correspondence to: pmc@di.uminho.pt; Tel. +351253604432

contrast to active measurement methodologies that resort to intrusive traffic, *i.e.*, probe packets are injected into the network for measurement purposes. A major difficulty associated with the usage of passive measurements is the volume of traffic involved, resulting in high-resource requirements for processing, storage and transmission of data [1]. Hence, the most common strategy used to mitigate this challenge is traffic sampling. Sampling consists of selecting a subset of packets that will allow to estimate parameters about all traffic, with compatible degrees of accuracy, avoiding processing it completely. In this way, packet sampling has become mandatory for effective passive network measurements, especially in the network core, reducing the amount of data to a manageable size [2].

Despite the substantial research work regarding packet sampling, the majority of existing proposals are focused on specific network measurement tasks, aiming at increasing the accuracy estimation of a single network metric or a small set of metrics. This scenario hampers the development of an encompassing measurement strategy based on traffic sampling able to support a large range of network management and planning activities, in a scalable way.

To sustain the development of configurable and efficient sampling techniques it is crucial to identify and understand the distinct features inherent to sampling. Analyzing sampling techniques through its constituent parts rather than a closed unit will allow to address issues such as accuracy estimation, sampling data overhead and computational weight within a narrower and simpler scope. A survey of the current literature on the topic reveals that a comprehensive classification of sampling techniques, including *adaptive* and *hybrid* techniques, and their inner components is still missing.

In this context, this work defines a taxonomy of sampling techniques with the aim to clarify sampling concepts and to provide a common ground for current and forthcoming research involving traffic sampling. By looking inside of traffic sampling techniques and identifying their main components, this taxonomy provides a modular view of sampling which can be explored both to adjust sampling configuration to specific measurement requirements and to enhance the performance of network measurement systems. In the classification criterion, the *granularity*, *selection scheme* and *selection trigger* are identified and proposed as the main components distinguishing current proposals. A comparative study demonstrates the taxonomy ability in guiding the modular design of classical and emerging sampling techniques. In addition, based on our previous work evaluating the computational requirements of different sampling techniques [3], some advices are given regarding the computational weight of each component and their suitability for specific network measurement tasks.

Taking the proposed taxonomy into consideration, we advocate a three-layer measurement architecture addressing key components to sustain a versatile and lightweight measurement strategy. This architecture has grounded the development of a traffic sampling framework able to be applied in both online and offline measurement scenarios.

The remaining of this paper is organized as follows: the related work is discussed in Section 2; the main characteristics of sampled-based measurement systems are introduced in Section 3; the proposed taxonomy of sampling techniques is presented in Section 4, including a comparison of current sampling proposals; and the conclusions are drawn in Section 5.

2. RELATED WORK

Traffic sampling techniques sustain a broad range of network tasks. As illustrated in Figure 1, examples of these tasks include:

(i) *network management* involving short, medium and long term planning and management of network operation, maintenance and provisioning of network services [4][5];

(ii) *traffic engineering* involving performance optimization, traffic characterization, traffic modeling and control [6][7][8];

(iii) *performance evaluation* of protocols and management tools, network reliability and fault tolerance [9][10][11];

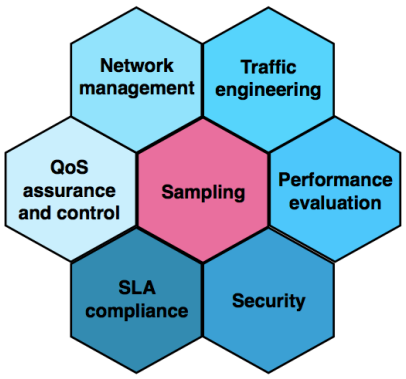


Figure 1. Usefulness of sampling

(iv) *network security*, including anomalies and intrusion detection, botnet and DDoS (Distributed Denial of Service) identification [12][13][14][15];

(v) *SLA (Service Level Agreement) compliance*, where auditing tools may resort to network sampling for measuring and reporting service levels [16];

(vi) *QoS control*, an area widely assisted by sampling, for measuring parameters such as delay, jitter and packet loss [17][18][19]. Most of these techniques claim better performance for each network task when compared with approaches widely deployed in measurement points, such as [1] [20].

Existing surveys on sampling usually assess the impact of packet sampling on various network monitoring activities [21] [22] [23] [24], identify the challenges of applying sampling and analyzing sampled network measurements [25] [26], or study the performance of traffic selection schemes [23] [27]. However, they usually do not address new sampling approaches such as adaptive techniques, restricting the analysis to the classical techniques referred in [1].

As regards the classification of traffic sampling techniques, a review of the literature demonstrates that none of the existing proposals define structurally all the components involved in sampling. Initial proposals for classifying traffic sampling techniques [28] were further developed and standardized within IETF [1]. These proposals classify the techniques concerning the packet selection scheme in use, e.g. systematic or random sampling, but exclude more elaborate sampling schemes. In [25], the relationship between traffic sampling and traffic aggregation aiming at data reduction on passive Internet measurements is introduced. This strategy involves grouping traffic streams according to flow keys observed within a period of time, instead of individual packets. The selection policy concepts are discussed in [2], however, the study is limited to count and event-driven approaches.

The present study, parsing classical and recent sampling proposals, brings an added value to the field, establishing a reference classification platform and a comparative overview for helping future research. Moreover, useful insights are given into the computational weight and the suitability of each approach for main network tasks. This knowledge allows designing more efficient and scalable sampling techniques.

3. SAMPLING-BASED MEASUREMENT SYSTEMS

Aiming at fostering the deployment of encompassing and flexible measurement strategies, this work proposes a three layer sampling architecture covering the main elements involved in traffic measurements. Each layer is modularly designed, which provides flexibility to accommodate mechanisms able to enhance the overall performance in current and forthcoming measurement scenarios.

3.1. Sampling concepts

Traffic sampling techniques share a set of concepts sometimes presented in an ambiguous way. To avoid inconsistencies, the most common terms are assumed in accordance with the following definitions:

- *Sample* - subset of network packets that are selected at the measurement point and then considered in the estimation of network parameters. This is also often referred as *sample event*, which consists in an individual action of selection and capture of packets from the stream under analysis;
- *Sample size* - number of packets or time interval in which all incoming packets at the measurement point are selected and captured to compose a sample. The sample size is controlled by triggers, that are responsible for starting and finishing each sample taking into account the packet position into the stream or its timestamp at the measurement point;
- *Interval between samples* - number of packets or time interval in which all incoming packets are ignored for measurement purposes. Likewise the sample size, the manipulation of interval between samples also resorts to triggers.

Figure 2 illustrates the above concepts.

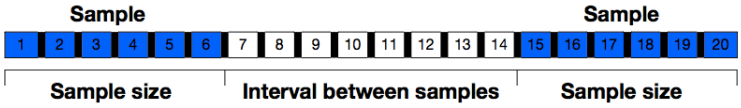


Figure 2. Basic sampling concepts

3.2. Measurement architecture

A sampling-based measurement architecture is foreseen as comprising three planes, as illustrated in Figure 3. The *management plane* includes tasks deployed directly in measurement points or in external management entities, namely: (*i*) map the measurement needs related to a specific network task into the more suitable sampling technique and its operational parameters; (*ii*) select and communicate with the measurement points which will perform packet sampling in order to set them up; and (*iii*) process the measurement results and provide a visualization component, when applicable, based on reports produced by the control plane. This also involves identifying an information model able to define managed objects in the network independently of specific implementations or protocols in use [29], as well as a standardized way for encoding information related to the sampling process, exporting and storage of the sampled data. In this way, the more comprehensive information model for packet sampling is the extended version of the IPFIX (IP Flow Information eXport) for PSAMP [30], which includes properties required by packet sampling reports that cannot be modeled using the basic IPFIX information model [31].

A modular design of the *control plane* allows a flexible sampling technique selection and configuration. Considering IETF PSAMP work and recent sampling proposals, a sampling taxonomy is here proposed to identify the inner characteristics distinguishing sampling techniques (see Section 4). The taxonomy also supports the definition of new sampling techniques which can be adjusted to each traffic/service measurement scenario.

In the control plane, the sampled packets received from the data plane are processed and the relevant field contents are extracted according to the network task measurement needs. These values are then aggregated (both in time and space) and exported following IETF guidelines (RFC6728, RFC6313), and using IPFIX specifications.

At the data plane, following the sampling rules defined in the control plane, packets are collected from the network link for subsequent use. Due to performance issues involved in reading packets from network links, mainly in high-capacity networks, the processes implemented in this plane must be kept simple, avoiding processing overhead. In this way, after collected, the unprocessed packets are reported to the control plane to be processed.

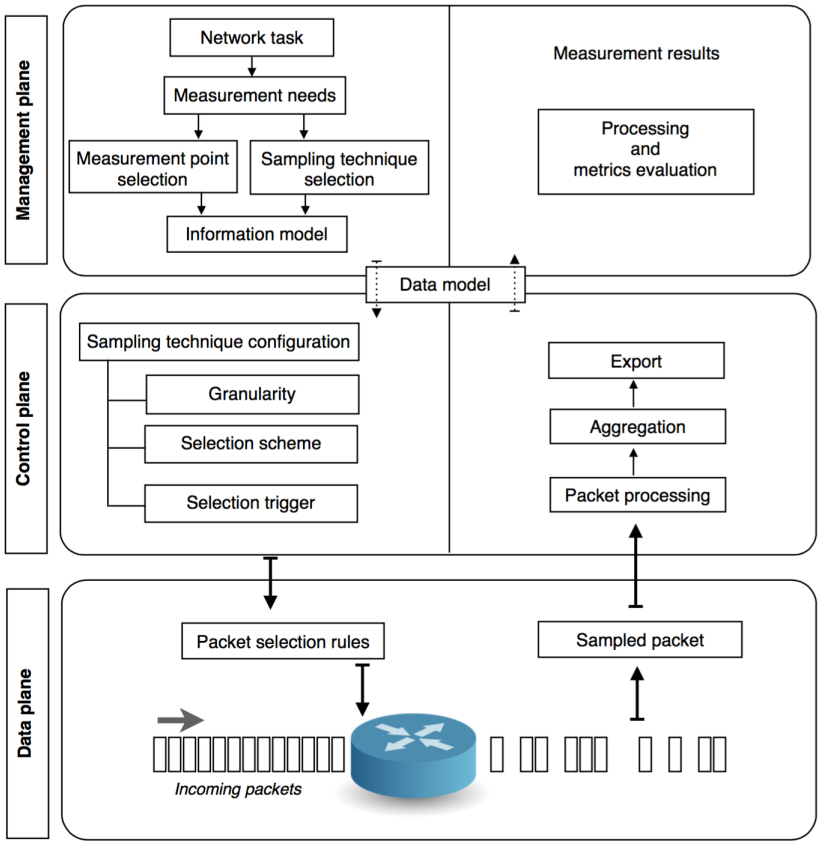


Figure 3. Architecture description

Regarding packet capturing, the measurement point implements an interface, also called capture device, which may be a NIC or any other device able to reading and collecting packets from the monitored link. In wired networks, where most traffic measurements are performed, the capture interface can be positioned *in-line* and in *mirroring* mode. While in in-line mode, the measurement point is directly connected to the monitored link between two hosts, usually resorting to a network tap that duplicates all observed traffic through passive splitting (on optical fiber links) or regeneration (in electrical copper networks), in mirroring mode, the network device forwarding packets can mirror packets from one or more ports to another port, in which the measurement point device is attached.

Considering that packets have to traverse several layers from the interface to the library (which is located at the top of the operating system's network stack), the overall capture performance depends on the efficiency in handing over packets from the capture device interface to the upper plane via the packet capture library. One of the various strategies proposed to improve this process is using a memory mapping technique in order to reduce the cost of copying packets from kernel-space to user-space through DMA (Direct Memory Access) [32].

Articulating the measurement scope, the required information model and the adequate sampling strategy is a major design issue for achieving an encompassing and efficient sampling solution. In this context, we have implemented a flexible sampling framework in order to select and configure each sampling technique according to the measurement purpose [33]. The framework allows to combine the sampling components defined above, and can be applied to both online and offline measurement scenarios.

4. TAXONOMY PROPOSAL

A wide coverage of the related literature shows that most of the sampling techniques, whether simple or complex, share a set of structural components, based on standard schemes, or as new strategies, arranged orthogonally with classical schemes or completely disjunct. In this way, describing these components in a modular and hierarchical structure able to foster a flexible and simple deployment of a comprehensive number of techniques represents a key role toward effective measurement systems based on sampling.

With the aim to provide a common ground for current and forthcoming sampling proposals, the proposed taxonomy fragments the sampling techniques 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 both classic and recently proposed sampling techniques. An overview of the taxonomy is illustrated in Figure 4, where:

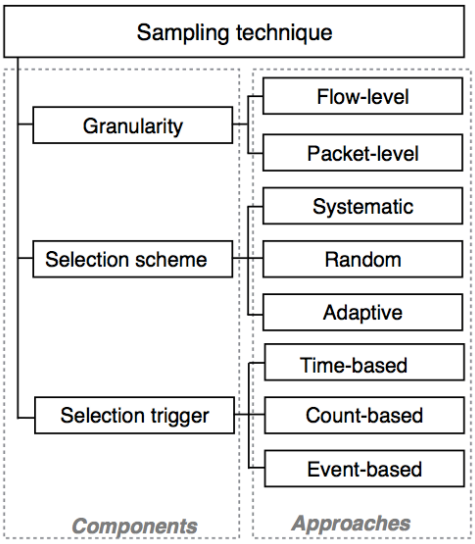


Figure 4. High-level view of sampling taxonomy

- *Granularity* - identifies the atomicity of the element under analysis in the sampling process: in a flow-level approach, the sampling process is only applied to packets belonging to a flow or to a set of flows of interest; in a packet-level approach, packets are eligible as single independent entities;
- *Selection scheme* - identifies the function defining which traffic packets will be selected and collected; this scheme may follow a deterministic, a random or an adaptive function;
- *Selection trigger* - determines the spatial and temporal sample boundaries; it may use a time-based approach, a count-based approach or an event-based approach.

4.1. Granularity

This component identifies which segment of traffic is considered in the sampling process and in the data reporting format. During the selection of packets, the sampler may consider all traffic traversing a measurement point or just part of it, targeting specific flows of interest. Generally, this decision depends on the network task (or measurement objective to fulfill), the network parameters being monitored, and the available communication and computational resources.

4.1.1. Flow-level Sampling According to RFC3697 [34], a flow is defined as a stream of packets sent by a particular source to a unicast, anycast or multicast destination, which exhibits specific properties or attributes in common. Traditionally, these properties (also called a flow key) are

identified based on five fields (5-tuple) of the packet header, namely source and destination IP addresses, source and destination ports, and type of protocol. In addition, RFC 2724 [35] and RFC 7011 [36] extend flow identification based on application layer information, MPLS (Multiprotocol Label Switching) labels or fields derived from packet treatment (e.g., next-hop IP address, etc.).

In terms of traffic sampling, the flow-level approach consists in applying the traffic capture policy only to packets belonging to a flow or a set of flows of interest. This involves classifying packets into flows before or during the sampling process [25]. Although considering a subset of flows may reduce the volume of data captured, stored and transmitted by the measurement point, this approach may increase the computational weight insofar all incoming packets must be processed to identify which flow they belong to. It also requires prior knowledge of which flows should be measured or some strategy to decide automatically which flows should be sampled.

Note that *flow-level sampling* is different from *flow sampling* (discussed in RFC 7014 [37]), that consists in capturing all packets that belong to a particular flow. Nevertheless, the strategies used in selecting flows, presented in [37] (*i.e.*, *systematic* and *random*), also may be applied to flow-level sampling context. In addition, different strategies can also target specific flows according to the measurement purpose. An example is the method introduced in [38], called *smart sampling*, that addresses the correct estimation of flow size distribution. These strategies are presented below already considering the necessary adaptations toward flow-level sampling:

- *Deterministic flow selection* - resorts to a deterministic function in which a set of flows are selected according to the number of flows arriving at the measurement point or during a time interval of observation. In the first case, the measurement point selects every Nth arriving flow to be considered for sampling, independently of the traffic type. Flow selection is then based on the first packet of a flow, upon which a counter is increased every time a packet belonging to a new flow arrives at the measurement point. If the counter is increased to a multiple of N, this flow will be considered for packet sampling. In the second approach, the packets from every flow observed at the measurement point between a time-based interval are elected to participate in the sampling process [37].
- *Random flow selection* - is based on a random process to select flows following a *n-out-of-N* or a *probabilistic* scheme. In *n-out-of-N*, n flows are selected out of the parent population, which consists of N incoming flows. It may involves generating n different random numbers in the range [1,N] and then selecting all flows with arrival position equal to one of the random numbers. In *probabilistic* flow selection, the decision of whether or not a flow is selected to participate in packet sampling follows a predefined probability that may be uniform (*i.e.*, with the same selection probability for all flows) or non-uniform (*i.e.*, where the selection probability can vary for different flows). The probabilistic scheme implies that the number of selected flows can vary [37].
- *Smart sampling* - aims at enhancing the ability to identify accurately the distribution of the traffic by controlling the flows that will participate in the sampling process according to a threshold previously defined. For every incoming flow, the measurement point maintains a counter with the size of the flow (in bytes or number of packets), then flows of size greater than the threshold are always selected, while smaller flows are selected with a probability proportional to their size [38].

Some works have demonstrated that conducting sampling only to a specific set of flows of interest can improve the estimation accuracy for tasks sensitive to flow sizes [39]. For instance, while traffic accounting targets mainly flows of large sizes [12], anomaly detection targets flows of small sizes [6]. In addition, as a group of packets share the same features (*i.e.*, flow key), it is possible to aggregate flows and thereby reduce the storage and transmission requirements to a manageable amount. In this way, flow-level sampling is expected to comply with IPFIX - *IP Flow Information Export* (RFC5470 [40], RFC6183 [41]).

4.1.2. Packet-level Sampling In this approach, incoming packets to a measurement point are considered single independent entities. Conversely to flow-level sampling, at packet-level the

packets do not need to be previously classified into flows, which may reduce drastically the computational requirements of processing every packet. Furthermore, collecting packets indistinctly turns packet-level sampling into a flexible and appropriate solution to be used in general purpose measurement tasks and aggregated estimations, in presence of diverse traffic types.

On the other hand, packet-level sampling may be difficult to deploy over large and high-speed networks, due to the challenges regarding the storage and transmission of measured data in environments with many flows.

There are many sampling techniques based on packet-level granularity and, following the increasing importance of traffic sampling, Cisco Systems has introduced a module in its tool for traffic monitoring - NetFlow. By default, NetFlow processes all incoming packets at the measurement point, keeping incremental statistical data about each flow in a cache memory. This requires high-processing resources, mainly in presence of high-speed links. In Sampled NetFlow a *systematic* selection scheme (discussed in Section 4.2.1) is used to process uniquely a subset of packets arriving at a defined interface, then flow records are assembled over the sampled packets. Furthermore, a *random* selection scheme (discussed in Section 4.2.2) was included in the tool aiming to increase the accuracy of the available statistics. Note that, although Netflow and Sampled Netflow are flow-based tools (state information is maintained per flow), in fact, the sampling selection scheme is packet-based, justifying the inclusion in this section.

As regards the exporting of sampled packets, a protocol based on IPFIX informational model (RFC5102 [42]), modified to report on single packets rather than on flows, is presented in RFC5474 [43]. This protocol defines mandatory contents for basic reports and an extended version able to include all fields required in RFC5102 [42].

4.2. Selection Scheme

The selection scheme identifies the selection function that determines the pattern under which packets will be selected and collected. This scheme can follow a *systematic*, a *random* or *adaptive* function.

4.2.1. Systematic Sampling In systematic sampling, the process of packet selection is ruled by a deterministic function which imposes a fixed sampling frequency, independently of the packet contents or treatment. In this scheme only equally spaced traffic portions are collected, i.e., sampling triggers are periodic (see Section 4.3) [1].

This approach is usually simple to develop and deploy, however, there is an inherent risk of obtaining biased samples if the packets being sampled exhibit a periodic structure which is rationally related to the deterministic function. This may lead to inaccurate results in parameter estimation due to the deterministic behavior of sampling and to the bursty nature of network traffic. Another potential drawback is that systematic sampling is to some extent predictable and, hence, open to deliberate manipulation [25].

Regarding the computational weight (e.g., CPU load and memory usage), the resources required by systematic techniques are directly related to the sampling frequency [3]. As discussed in [44] for flow accounting, the definition of the optimal number of samples depends on the expected accuracy, therefore, the accuracy level also impacts on the resources required.

4.2.2. Random Sampling The random selection scheme aims to avoid biasing samples by ruling the sampling frequency through a random process, usually resorting to a pseudorandom generator or to a probabilistic function.

The pseudorandom approach tries to avoid predictability choosing values exponentially distributed (for time-based trigger, discussed in Section 4.3) or geometrically distributed (for count-based trigger, discussed in Section 4.3.2) [25]. A common strategy following this principle is performed by inducing the random function generator to converge to a required sampling rate, ensuring that the sampling frequency distribution is limited by maximum and minimum values. The *n-out-of-N* technique presented in [1] and used in Cisco Sampled NetFlow follows this approach,

where n packets are randomly selected from a traffic population of N packets, generating numbers in the range $[1, N]$ and then selecting all packets that have the corresponding packet position.

As regards the probabilistic approach, the decision about the sampling frequency follows a predefined probability density function. The probabilistic function can be uniform, where all packets have an equal probability to be selected, or non-uniform, where the packets have different probability of selection. In [13], it is introduced a sampling technique based on a probabilistic scheme for anomaly detection, namely network scans, SYN flooding and worms. This technique divides time into strata and then selects an incoming packet with a probability, which is a decreasing function f of the predicted size of the flow the packet belongs to.

In [3], it is shown that for equivalent sampling frequencies, the random techniques require slightly more computational resources than systematic sampling. The computational burden may however vary according to the complexity of the probabilistic function adopted. As discussed in [2] [27], there is no clear advantage in the choice of random or systematic sampling for traffic classification and characterization.

Random approaches are also difficult to deploy for estimating multipoint metrics such as end-to-end delay, even in flow-level approaches, as the sampling processes running on the measurement points involved are not correlated, and there are no guarantees that samples will be composed by the same packets [45].

4.2.3. Adaptive Sampling In this approach, the sampling technique is endowed with the ability to change the selection of packets during the course of measurements. This flexibility aims at identifying the most important parts of a traffic stream according to the measurement needs or to save network resources during critical periods of its operation.

Adaptive packet sampling techniques can be based on linear prediction, fuzzy logic or other particular adaptive strategies and mechanisms that consider the traffic behavior, the packet content or the network status to rule sampling pattern changes. Although the use of an adaptive strategy may suggest higher consumption of resources, some adaptive techniques may introduce less computational weight, regarding CPU load and memory consumption, even when compared with classical techniques [3].

The multiadapative sampling technique presented in [46] uses linear prediction to identify changes in the network activity and therefore set a more suitable sampling pattern, by adjusting accordingly the sampling frequency (manipulating the interval between samples) and the sample size. This allows reducing the measurement overhead while keeping the accuracy in traffic characterization.

In adaptive sampling techniques based on fuzzy logic, a controller adjusts the sampling rate based on past experience of similar situations, determining the most suitable action for a particular traffic condition or measurement requirement. Some of these proposals discuss the characteristics and deployment of this type of controller, as in [47] and [48]. These approaches tend to require more resources as long-time databases are required to store the knowledge about past situations. They also tend to be less reactive and effective in accommodating new measurement needs.

4.3. Selection Trigger

In network measurements based on sampling, only a subset of all packets traversing the measurement point is selected and considered to estimate network metrics. To achieve this, a trigger is defined to determine the start and the end of a sample, and consequently the interval between samples. In this way, a selection trigger is classified as *time-based*, *count-based* or *event-based* as described below.

4.3.1. Time-based A time-based approach defines that the beginning and the end of a sample is determined based on packet arrival timestamping. Its deployment consists of using a first countdown timer within which all packets arriving at the measurement point are selected for the sample and a second countdown timer within which all packets incoming are ignored for measurement purposes.

As shown in Figure 5(a), when the trigger fires the beginning of a new sample, the measurement point waits for the first bit of the next incoming packet and starts the collection. When the trigger fires the end of sampling, the measurement point continues the collection until the last bit of the current packet and then interrupts the selection process.

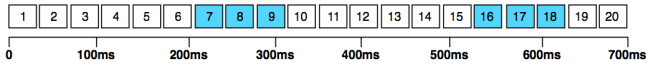
This may involve a simple deterministic function as discussed in [46], in which a *systematic* selection scheme is used [1] (presented in Section 4.2.1), where the sample size and the interval between samples are set at the beginning of the sampling process and remain invariant until the end. In the *adaptive* approach presented in [18], the interval between samples is decided dynamically based on the variance of an observed reference parameter while the sample size remains invariant.

In terms of performance, the work reported in [49] demonstrates that time-based triggers are less robust than count-based when applied in traffic characterization, being affected by the bursty nature of network traffic. However, they may be suitable for applications that require the analysis of consecutive packets, such as IDS [50].

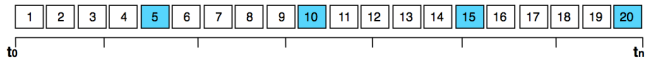
Traffic burstiness may also affect the volume of data involved in a sampling process, and hamper the adoption of strategies which define the optimal number of samples beforehand (as it is difficult to anticipate the number of packets arriving at the measurement point in each sampling interval). As discussed in [3], the amount of data collected in time-based techniques is often higher than in count-based ones. However, as there is no significant activity during the interval between samples, such as packet counter increments (introduced in Section 4.3.2), this approach may achieve a significant reduction in the ratio of CPU load and memory usage per MByte collected and stored.

4.3.2. Count-based The count-based approach defines that the beginning and the end of a sample are driven for the spatial position of the packet within the traffic stream, using counters which are independent of the packet arrival timestamp.

An example of using this approach involves a deterministic function in which the interval between samples corresponds to a predefined number of packets that must arrive at the measurement point before the beginning of a new sample. For this, the counter is decremented at every packet arriving at the measurement point; when the counter reaches zero, a new sample starts. This strategy is used in Sampled NetFlow and illustrated in Figure 5(b).



(a) Time-based



(b) Count-based

Figure 5. Example of time-based and count-based selection triggers

Having the possibility of anticipating which proportion of the traffic will be collected and stored, count-based techniques are suitable for environments with limited resources, or for applications where it is necessary to determine the optimal number of samples for a specific accuracy, as discussed in [44]. In this approach, every packet arriving at the measurement point must be processed to shift the packet counter, therefore, the computational weight involved is directly related to the total load of the traffic under analysis [3].

4.3.3. Event-based In this approach 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 content, the treatment of the packet at the measurement point or a more complex observation. The packet content corresponds to the union of the packet header (which includes link layer, network layer, and other encapsulation headers) and packet payload [1]. Sampling techniques based on this strategy may predefine some values of the packet header and then, all packets in

which these values match are selected for the sample. This approach is usually called *property match filtering* [1].

Hash-based techniques, such as presented in [51], are considered event-based, once the hash function is applied to the packet contents and then the packet is selected if the hash value falls in a selection range. This approach is sometimes used to emulate random sampling by selecting a proper range of hash values. Although event-based allows collecting a specific range of packets of interest, as it involves processing all incoming packets to identify the event. This may lead to workload overhead in the equipment.

4.4. Hybrid Techniques

There are several techniques that combine approaches of the same taxonomy component, usually from the selection trigger or selection scheme. These techniques aim at enhancing traffic sampling, although the overlap increases the computational cost for traffic measurement.

An example of a hybrid technique is the use of an event-based trigger that fires upon observation of a packet with specified contents, after which any incoming packets within the next t seconds are selected to compose the sample, using a *time-based* approach [25]. This solution may also be deployed in conjunction with a *count-based* approach by capturing the first n packets arriving at the measurement point after identifying the event. As exemplified in Figure 6, the event is the first packet of a new flow observed in the stream and the sample size is equal to 3. This strategy may be of interest for traffic classification or security tasks.

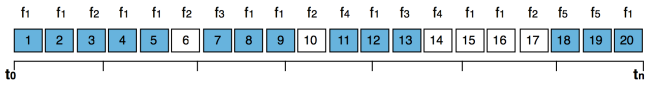


Figure 6. Hybrid technique - event-based and count-based

4.5. Comparative Summary of Sampling Techniques

After exemplifying how sampling components can be articulated to establish a particular sampling technique, this section provides a comparative summary on a comprehensive range of current sampling techniques in light of the taxonomy defined in Section 4.

The sampling taxonomy presented before allows to classify both classical and recent sampling techniques as well as to ground the definition of future proposals. Figure 7 exemplifies how the components defined in the taxonomy can be organized to deploy a sampling technique. The resulting sampling structure can be linear or hybrid (detailed in Section 4.4) depending on how sampling approaches are elected per component.

Figure 7(A) corresponds to a technique defined in [1] and available in most deployed sampling tools, *e.g.*, Cisco NetFlow and sFlow. Although the technique represented in Figure 7(B) is also defined in [1], it is scarcely deployed in the current network measurement panorama, even considering its importance for IDS [50]. The technique represented in Figure 7(C) illustrates the flexibility in deploying new sampling profiles. This technique might be used for monitoring a specific service and adapting the sampling frequency in response to some event observed into its own traffic. Some techniques able to be used in this context are presented in [12] [17].

Table I presents a summary of most used and referenced sampling techniques classified according to the proposed taxonomy, highlighting the network task to which the technique is oriented to. This comparative study, along with insights throughout this work, allows a clearer positioning of existing sampling proposals, being both a contribution for further research in the area and a road map for deciding on the most suitable sampling technique to use.

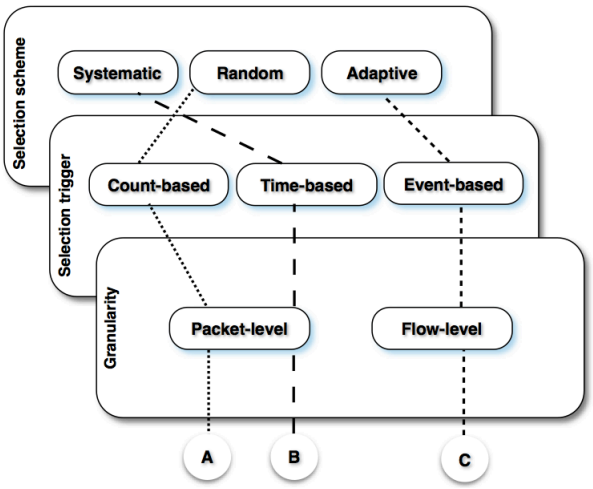


Figure 7. Example of sampling technique composition

Table I. Taxonomy of sampling techniques

Sampling Proposals	**Granularity**		**Selection scheme**			**Selection trigger**			**Network task**
	Pkt	Flow	Sys	Rnd	Adp	Cnt	Time	Event	
Adaptive linear prediction [47]	√				√		√		Traffic engineering
Adaptive non-linear sampling [6]		√			√	√			Traffic engineering
Adaptive random sampling [52]	√			√		√			Traffic engineering
Adaptive statistical sampling [18]	√				√		√		QoS assurance and control
Botnet-aware adaptive sampling[53]	√				√	√			Security
Distributed adaptive sampling [17]		√			√			√	QoS assurance and control
Flow statistics trivial sampling [19]	√		√	√		√			QoS assurance and control
Fuzzy regulator adapt. sampling [54]	√				√		√		Performance evaluation
Hash-based sampling [1]	√		√					√	Traffic engineering
Systematic count-based [1]	√		√			√			Traffic engineering
Systematic SYN sampling [2]	√		√			√		√	Traffic engineering
Systematic time-based [1]	√		√				√		Traffic engineering
Modified FLC sampling [48]	√				√		√		Performance evaluation
Multiadaptive sampling [46]	√				√		√		Traffic engineering
Opportunistic sampling [12]		√			√			√	Security
Random sampled NetFlow [55]	√			√		√			Network management
Resource conserving sampling [9]		√			√	√			Performance evaluation
Sample and hold [51]	√			√		√			SLA compliance
Sampled NetFlow [55]	√		√			√			Network management
sFlow [20]		√		√		√			Traffic engineering

5. CONCLUSIONS

Aware of the relevance and need of a common understanding in the traffic sampling arena, this work goes inside packet sampling techniques establishing a comprehensive taxonomy of their inner characteristics. After identifying *granularity*, *selection scheme* and *selection trigger* as the main differentiating components of sampling proposals, the study has further detailed each of these components, discussing also the involved computational weight. Following this taxonomy, a general-purpose sampling-based measurement architecture has been proposed to assist the deployment of flexible and lightweight measurement systems. Finally, classic and recent sampling techniques have been compared taking into account the proposed taxonomy, highlighting their applicability in the network management context.

Having demonstrated the ability to frame existing sampling techniques, future work will carry out an extensive and systematic comparative analysis of each measurement task supported by packet sampling. The aim is to provide reliable inputs to select the most suitable technique for specific network scenarios and measurement goals. The full specification of all components in the proposed sampling-based measurement architecture, covering aspects such as mapping the measurement requirements into a suitable sampling technique and selecting the packet fields of interest, the aggregation level and exporting format, is also a key aspect to be addressed in forthcoming work.

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. Zseby T, Molina M, Duffield N. Sampling and Filtering Techniques for IP Packet Selection RFC 5475. *Technical Report*, IETF 2009. URL http://datatracker.ietf.org/doc/rfc5475/.
2. Tammaro D, Valenti S, Rossi D, Pescapé A. Exploiting packet-sampling measurements for traffic characterization and classification. *International Journal of Network Management* Nov 2012; **22**(6):451–476, doi:10.1002/nem.1802. URL http://dx.doi.org/10.1002/nem.1802.
3. Silva JMC, Carvalho P, Lima SR. Computational weight of network traffic sampling techniques. *2014 IEEE Symposium on Computers and Communications (ISCC)*, IEEE: Madeira, Portugal, 2014; 1–6, doi:10.1109/ISCC.2014.6912467. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6912467.
4. Hu ZG, Zhang DL, Hou CP, Zhang JS. Adaptive sampling algorithm of network round-trip time. *Journal of Computer Applications*, vol. 30, Yingyong J (ed.), 2010; 319–322.
5. Duffield NG, Grossglauser M. Trajectory sampling for direct traffic observation. *ACM SIGCOMM Computer Communication Review* Oct 2000; **30**(4):271–282, doi:10.1145/347057.347555. URL http://dl.acm.org/citation.cfm?id=347057.347555.
6. Hu C, Wang S, Tian J, Liu B, Cheng Y, Chen Y. Accurate and Efficient Traffic Monitoring Using Adaptive Non-Linear Sampling Method. *IEEE INFOCOM 2008 - The 27th Conference on Computer Communications*, IEEE, 2008; 26–30, doi:10.1109/INFOCOM.2008.14. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4509609.
7. Yang L, Michailidis G. Sampled Based Estimation of Network Traffic Flow Characteristics. *IEEE INFOCOM 2007 - 26th IEEE International Conference on Computer Communications*, IEEE, 2007; 1775–1783, doi:10.1109/INFCOM.2007.207. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4215789.
8. Tune P, Veitch D. Towards optimal sampling for flow size estimation. *Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement*, IMC '08, ACM: New York, NY, USA, 2008; 243–256, doi:10.1145/1452520.1452550.
9. Mahmood AN, Hu J, Tari Z, Leckie C. Critical infrastructure protection: Resource efficient sampling to improve detection of less frequent patterns in network traffic. *Journal of Network and Computer Applications* 2010; **33**(4):491–502, doi:DOI: 10.1016/j.jnca.2010.01.003.
10. Lee M, Duffield N, Kompella R. Two Samples are Enough: Opportunistic Flow-level Latency Estimation using NetFlow. *2010 Proceedings IEEE INFOCOM*, IEEE, 2010; 1–9, doi:10.1109/INFCOM.2010.5462044. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5462044.
11. Kandula S, Mahajan R. Sampling biases in network path measurements and what to do about it. *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, IMC '09, ACM: New York, NY, USA, 2009; 156–169, doi:http://doi.acm.org/10.1145/1644893.1644912. URL http://doi.acm.org/10.1145/1644893.1644912.

12. Androulidakis G, Chatzigiannakis V, Papavassiliou S. Network anomaly detection and classification via opportunistic sampling. *IEEE Network* Jan 2009; **23**(1):6–12, doi:10.1109/MNET.2009.4804318. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4804318.
13. He X, Yang W, Wang Q. An Adaptive Traffic Sampling Method for Anomaly Detection. *2009 Fourth International Conference on Internet Computing for Science and Engineering*, IEEE, 2009; 142–146, doi:10.1109/ICICSE.2009.32. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5521614.
14. Wu Z, Hu R, Yue M. Flow-oriented detection of low-rate denial of service attacks. *International Journal of Communication Systems* 2014; :n/a–n/adoi:10.1002/dac.2805. URL http://dx.doi.org/10.1002/dac.2805.
15. Wu Z, Yue M, Li D, Xie K. Sedp-based detection of low-rate dos attacks. *International Journal of Communication Systems* 2015; **28**(11):1772–1788, doi:10.1002/dac.2783. URL http://dx.doi.org/10.1002/dac.2783.
16. Zseby T. Deployment of sampling methods for SLA validation with non-intrusive measurements. *Proceedings of Passive and Active Measurements Conference*, Fort Collins, 2002.
17. Serral-Gracia R, Cabellos-Aparicio A, Domingo-Pascual J. Packet Loss Estimation Using Distributed Adaptive Sampling. *Network Operations and Management Symposium Workshops, 2008. NOMS Workshops 2008. IEEE*, IEEE, 2008; 124–131, doi:10.1109/NOMSW.2007.22. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4509938.
18. Dogman A, Saatchi R, Al-Khayatt S. An adaptive statistical sampling technique for computer network traffic. *Communication Systems Networks and Digital Signal Processing (CSNDSP), 2010 7th International Symposium on*, 2010; 479–483, doi:http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5580380. URL http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5580380.
19. Gu Y, Breslau L, Duffield N, Sen S. On Passive One-Way Loss Measurements Using Sampled Flow Statistics. *IEEE INFOCOM 2009 - The 28th Conference on Computer Communications*, IEEE, 2009; 2946–2950, doi:10.1109/INFCOM.2009.5062264. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5062264.
20. Phaal P, Panchen S, McKee N. InMon Corporation's sFlow: A Method for Monitoring Traffic in Switched and Routed Networks. RFC 3176 (Informational) Sep 2001. URL http://www.ietf.org/rfc/rfc3176.txt.
21. Mai J, Chuah CN, Sridharan A, Ye T, Zang H. Is sampled data sufficient for anomaly detection? *Proceedings of the 6th ACM SIGCOMM on Internet measurement - IMC '06*, IMC '06, ACM Press: New York, New York, USA, 2006; 165, doi:10.1145/1177080.1177102. URL http://portal.acm.org/citation.cfm?doid=1177080.1177102.
22. Brauckhoff D, Tellenbach B, Wagner A, May M, Lakhina A. Impact of packet sampling on anomaly detection metrics. *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, IMC '06, ACM: New York, NY, USA, 2006; 159–164, doi:http://doi.acm.org/10.1145/1177080.1177101. URL http://doi.acm.org/10.1145/1177080.1177101.
23. Pescape A, Rossi D, Tammaro D, Valenti S. On the impact of sampling on traffic monitoring and analysis. *2010 22nd International Teletraffic Congress (lTC 22)*, IEEE, 2010; 1–8, doi:10.1109/ITC.2010.5608718. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5608718.
24. Carela-Español V, Barlet-Ros P, Cabellos-Aparicio A, Solé-Pareta J. Analysis of the impact of sampling on NetFlow traffic classification. *Computer Networks* Apr 2011; **55**(5):1083–1099, doi:10.1016/j.comnet.2010.11.002. URL http://dx.doi.org/10.1016/j.comnet.2010.11.002.
25. Duffield N. Sampling for Passive Internet Measurement: A Review. *Statistical Science* Aug 2004; **19**(3):472–498, doi:10.1214/088342304000000206. URL http://projecteuclid.org/Dienst/getRecord?id=euclid.ss/1110999311/.
26. Zseby T, Hirsch T, Claise B. Packet Sampling for Flow Accounting: Challenges and Limitations. *Passive and Active Network Measurement*, *Lecture Notes in Computer Science*, vol. 4979, Claypool M, Uhlig S (eds.). Springer Berlin / Heidelberg, 2008; 61–71. URL http://dx.doi.org/10.1007/978-3-540-79232-1_7.
27. Chabchoub Y, Fricker C, Guillemin F, Robert P. Deterministic Versus Probabilistic Packet Sampling in the Internet. *Managing Traffic Performance in Converged Networks*, *Lecture Notes in Computer Science*, vol. 4516, Mason L, Drwiega T, Yan J (eds.). Springer Berlin / Heidelberg, 2007; 678–689. URL http://link.springer.com/chapter/10.1007/978-3-540-72990-7_60.
28. Amer P, Cassel L. Management of sampled real-time network measurements. *[1989] Proceedings. 14th Conference on Local Computer Networks*, IEEE Comput. Soc. Press, 1989; 62–68, doi:10.1109/LCN.1989.65244. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=65244.
29. Dietz T, Claise B, Quittek J. Definitions of Managed Objects for Packet Sampling. RFC 6727 2012. URL http://datatracker.ietf.org/doc/rfc6727/.
30. Dietz, T and Claise, B and Aitken, P and Dressler, F and Carle G. Information Model for Packet Sampling Exports. *Technical Report*, IETF RFC 5477 2009. URL https://datatracker.ietf.org/doc/rfc5477/.
31. Claise, B and Trammel B. Information Model for IP Flow Information Export (IPFIX) - RFC 7012. *Technical Report*, IETF 2013. URL https://datatracker.ietf.org/doc/rfc7012/.
32. Hofstede R, Celeda P, Trammell B, Drago I, Sadre R, Sperotto A, Pras A. Flow Monitoring Explained: From Packet Capture to Data Analysis With NetFlow and IPFIX. *IEEE Communications Surveys & Tutorials* Jan 2014; **16**(4):2037–2064, doi:10.1109/COMST.2014.2321898. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6814316.
33. Silva J, Carvalho P, Rito Lima S. A modular sampling framework for flexible traffic analysis. *Software, Telecommunications and Computer Networks (SoftCOM), 2015 23rd International Conference on*, 2015; 200–204, doi:10.1109/SOFTCOM.2015.7314061.
34. Rajahalme J, Conta A, Carpenter B, Deering S. IPv6 Flow Label Specification. *Technical Report 3697*, Internet Engineering Task Force 2004. URL http://datatracker.ietf.org/doc/rfc3697/.
35. Handelman S, Stibler S, Brownlee N, Ruth G. RTFM: New Attributes for Traffic Flow Measurement. *Technical Report 2724*, Internet Engineering Task Force 1999. URL

http://datatracker.ietf.org/doc/rfc2724/.
36. Claise B, Trammell B. Specification of the IP Flow Information eXport (IPFIX) Protocol for the Exchange of Flow Information. RFC 7011 2013. URL http://datatracker.ietf.org/doc/draft-ietf-ipfix-protocol-rfc5101bis/.
37. DAntonio S, Zseby T, Hence C, Peluso L. Flow Selection Techniques. *Technical Report 7014*, Internet Engineering Task Force 2013. URL http://datatracker.ietf.org/doc/rfc7014/.
38. Duffield N, Lund C. Predicting resource usage and estimation accuracy in an IP flow measurement collection infrastructure. *Proceedings of the 2003 ACM SIGCOMM conference on Internet measurement - IMC '03*, ACM Press: New York, New York, USA, 2003; 179, doi:10.1145/948205.948228. URL http://dl.acm.org/citation.cfm?id=948205.948228.
39. Hohn N, Veitch D. Inverting sampled traffic. *Proceedings of the 2003 ACM SIGCOMM conference on Internet measurement - IMC '03*, ACM Press: New York, New York, USA, 2003; 222, doi:10.1145/948205.948235. URL http://dl.acm.org/citation.cfm?id=948205.948235.
40. Sadasivan G, Brownlee N, Claise B, Quittek J. Architecture for IP Flow Information Export. RFC 5470 2009. URL http://datatracker.ietf.org/doc/rfc5470/.
41. Kobayashi A, Claise B, Muenz G, Ishibashi K. IP Flow Information Export (IPFIX) Mediation: Framework. RFC 6183 2011. URL http://datatracker.ietf.org/doc/rfc6183/.
42. Quittek J, Bryant S, Claise B, Aitken P, Meyer J. Information Model for IP Flow Information Export. RFC 5102 2008. URL http://datatracker.ietf.org/doc/rfc5102/.
43. Duffield N, Chiou D, Claise B, Greenberg A, Grossglauser M, Rexford J. A Framework for Packet Selection and Reporting. *IETF RFC 5474* 2009; URL http://datatracker.ietf.org/doc/rfc5474/.
44. Choi BY, Bhattacharyya S. Observations on Cisco sampled NetFlow. *ACM SIGMETRICS Performance Evaluation Review* Dec 2005; **33**(3):18, doi:10.1145/1111572.1111579. URL http://portal.acm.org/citation.cfm?doid=1111572.1111579.
45. Henke C, Schmoll C, Zseby T. Empirical Evaluation of Hash Functions for PacketID Generation in Sampled Multipoint Measurements. *Passive and Active Network Measurement*, *Lecture Notes in Computer Science*, vol. 5448, Moon S, Teixeira R, Uhlig S (eds.). Springer Berlin / Heidelberg, 2009; 197–206. URL http://dx.doi.org/10.1007/978-3-642-00975-4_20.
46. Silva JMC, Carvalho P, Rito Lima S. A multiadaptive sampling technique for cost-effective network measurements. *Computer Networks* Dec 2013; **57**(17):3357–3369, doi:10.1016/j.comnet.2013.07.023. URL http://dx.doi.org/10.1016/j.comnet.2013.07.023 http://linkinghub.elsevier.com/retrieve/pii/S1389128613002491.
47. Hernandez EA, Chidester MC, George AD. Adaptive Sampling for Network Management. *Journal of Network and Systems Management* 2001; **9**(4):409–434. URL http://dx.doi.org/10.1023/A:1012980307500.
48. Xin Q, Hong L, Fang L. A Modified FLC Adaptive Sampling Method. *2009 WRI International Conference on Communications and Mobile Computing*, vol. 2, IEEE, 2009; 515–520, doi:10.1109/CMC.2009.56. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4797177.
49. Claffy KC, Polyzos GC, Braun HW. Application of sampling methodologies to network traffic characterization. *SIGCOMM Comput. Commun. Rev.* 1993; **23**(4):194–203, doi:http://doi.acm.org/10.1145/167954.166256. URL http://doi.acm.org/10.1145/167954.166256.
50. Shirali-Shahreza S, Ganjali Y. FleXam. *Proceedings of the second ACM SIGCOMM workshop on Hot topics in software defined networking - HotSDN '13*, ACM Press: New York, New York, USA, 2013; 167, doi:10.1145/2491185.2491215. URL http://dl.acm.org/citation.cfm?id=2491185.2491215.
51. Estan C, Varghese G. New directions in traffic measurement and accounting. *SIGCOMM Comput. Commun. Rev.* Oct 2002; **32**(4):323–336, doi:10.1145/964725.633056. URL http://dl.acm.org/citation.cfm?id=964725.633056 http://doi.acm.org/10.1145/964725.633056.
52. Choi BY, Park J, Zhang ZL. Adaptive random sampling for load change detection. *ACM SIGMETRICS Performance Evaluation Review* Jun 2002; **30**(1):272, doi:10.1145/511399.511376. URL http://dl.acm.org/citation.cfm?id=511399.511376.
53. Zhang J, Luo X, Perdisci R, Gu G, Lee W, Feamster N. Boosting the scalability of botnet detection using adaptive traffic sampling. *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*, ASIACCS '11, ACM: New York, NY, USA, 2011; 124–134, doi:http://doi.acm.org/10.1145/1966913.1966930. URL http://doi.acm.org/10.1145/1966913.1966930.
54. Giertl J, Baca J, Jakab F, Andoga R. Adaptive sampling in measuring traffic parameters in a computer network using a fuzzy regulator and a neural network. *Cybernetics and Systems Analysis* 2008; **44**(3):348–356. URL http://dx.doi.org/10.1007/s10559-008-9005-0.
55. Cisco IOS. NetFlow 2008.