Jurnal Elektronika dan Telekomunikasi (JET), Vol. 19, No. 1, August 2019, pp. 26-31
Accredited by RISTEKDIKTI, Decree No: **32a/E/KPT/2017**
doi: 10.14203/jet.v19.26-31

CDF-based Flow Detection for Network Flow Sampling and Packet Capturing

Aris Cahyadi Risdianto [a, *], Nuryani [b]

[a] *School of Electrical Engineering and Computer Science*
Gwangju Institute of Science and Technology
123 Cheomdangwagi-ro, Buk-gu
Gwangju, Korea
[b] *Research Centre for Informatics*
Indonesian Institute of Sciences (LIPI)
Gedung 20 Lt.3, Jl. Cisitu (Komplek LIPI) No. 21/154D,
Bandung, Indonesia

Abstract

Providing an appropriate level of flow collection, relying on packet capturing or flow sampling method, is extremely hard due to various practical limitations and resources requirements. To address this challenge, this paper investigated a CDF (Cumulative Distribution Function)-based flow detection to decide between "known" and "unknown" flows. Therefore, a combined flow collection can be achieved to improve the collection's efficiency by sampling only the known flows and capturing the remaining unknown flows. As a preliminary experiment, detecting known and unknown flows was conducted over a long period by calculating the empirical CDF distance between each flow's rate and overall packet's rate distribution, called as FPR (Flow-to-Packet Ratio), with a threshold (FPR_{min}) based on a significant level of observed data. The result shows that unknown flow is detected for most of the recommended significant level values.

Keywords: flow detection, cumulative distribution function (CDF), flow sampling, and packet capturing.

I. INTRODUCTION

Providing an appropriate level of flow collection, relying on either distributed with fully captured or centralized with sampled, is extremely hard due to various practical limitations and requirements in packet capturing and flow sampling [1]. Simple flow collections can be collected through some available packet capturing tools. However, for monitoring wide-area networks, the packets need to be replicated and sent to a single location. Then replicated packets are sent to the packet capturing tools, with or without match-based packet filtering process. Another way to enable flow collections is by using an agent that combines packet samples into datagram and randomly sampled according to a pre-defined sampling interval/ratio before sending the datagram into the collector [2].

In this paper, flow is defined as a set of packets with a common property specified in the packet's headers, known as the flowkey, which observed within a period. The packet capturing reflects exact flows being processed but only valid at that particular point and also known as a resource-consuming process. In contrast, the flow sampling is limited due to sub-optimal flow sampling ratio/interval. However, it has a unique global view of flows in the whole network and requires a minimum resource. So, there is a need to utilize combined methods that providing an appropriate level of flow collection and improving efficiency while still keeping the specific level of accuracy. Several sampling methods for network flow monitoring are proposed to address these needs. Providing efficient building blocks for sampling and large flow detecting by using OpenFLow based methods in SDN switch can be used in various monitoring application [3]. A double-sampling and hold-based approach that includes two sample process, hold, and early removal process is proposed to maximize the flow information in the given limited resources [4]. Another approach is modular and self-adaptive measurement architecture that consists of management, sampling, and network plane to accommodate the selection and configuration of sampling technique [5]. R. Hofstede *et al.* [6] explained a novel traffic monitoring approaches as well as improving efficiency in processing and storing the traffic data. They used protocols such as Netflow and IPFIX, to perform flow monitoring, including packet observation, flow metering and export, data collection, and the final stage is data analysis. However, it needs improvement by combining packet analysis and flow monitoring. Another work tried to estimate the number of bytes and packets of the flow by using Maximum Likelihood Estimation (MLE). The expected relative error is defined as $\% \, error \leq 196. \sqrt{\frac{1}{s}}$ [7], means that avoiding flow sampling for achieving specific *error*, if the number of sampled flows is less than *s*. Performance can be improved by observing packet over a long period (not over short a time of sampling period) and counting the actual rate parameter (not only the packet count).

This paper describe a decision problem to select an appropriate flow collection that combines two methods,

* Corresponding Author.
Email: aris@nm.gist.ac.kr
Received: February 20, 2019 ; Revised: June 25, 2019
Accepted: August 01, 2019 ; Published: August 31, 2019

flow sampling and packet capturing, by monitoring a CDF (Cumulative Distribution Function) of each flow over a long period, then comparing with overall packet's characteristics. If the characteristic of flow is similar to the overall packet's characteristic, then it can be decided that the flow and packet are closely related, and considered as a *known flow*. On the other side, if it is significantly different, the flow is considered as an *unknown flow*. The particular problem is described in section 2. The proposed design of the system model to decide a set of *known* and *unknown flow* with a pre-defined level of accuracy is described in section 3. It is followed by experimental simulation and analysis in section 4 and 5, respectively. Finally, conclusion and several recommendations for future work are provided in section 6.

II. PROBLEM FORMULATION

As mentioned in the previous section, a flow that closely related with overall packets characteristic can be considered as *known flows* as depicted in Figure 1 (a) and (b). The flow can be collected by flow sampling for reducing the amount of collected data and resource requirement. The other flows with fixed-rate/packet size, suspicious flow, or wrong flow can be considered as *unknown flow* as depicted in Figure 1 (c) and (d).

This paper observed a set of packet input over a pre-defined long-time period, and classifies them based on the specific attributes, which can be considered as *flows*. Then, the statistical characteristic of the overall packet and flows were analyzed. For defining known or unknown flows, the statistical relation between flow and overall packets were compared by measuring the distance between distributions for specific statistics parameter (e.g., rates or counts). Under pre-defined threshold (i.e., distance value) or expected level of accuracy (i.e., a probability of error), the flow f_k can be decided into a set of known flows F_N or unknown flows F_U with (1).

$$\begin{cases} CDF\ distance\ between\ f_k\ and\ p \leq\ distance_{min}\ ,\ f_k \in F_N \\ CDF\ distance\ between\ f_k\ and\ p > distance_{min}\ ,\ f_k \in F_U \end{cases} \quad (1)$$

III. SYSTEM DESIGN

The proposed system design for calculating the CDF of the flow's rate value over a long period is depicted in Figure 2. This system helps us to distinguish known or unknown flows depending on the distribution distance between each flow's rate and overall packet's rate. *Flow Classifier* with pre-defined set of *flowkey* S (i.e., a combination of packet's header information) practically groups the packet input $p(S,t)$ which contains all the *flowkey* in $S\{1,2,3,\dots,K\}$, into K number of flows. So, each flow's rate over the time can be defined as in (2).

$$f_k(t) = p(S,t) \quad , k \in \{1,2,3,\dots,K\} \quad (2)$$

The essential point is a set of *flowkey* S with K number of *flowkey* elements, which generate exactly K number of flows.

Flow modeling tries to observe the flow rate's value over a specific time and then models the observed value into CDF. First, random variable X is defined with the value of $x_1, x_2, x_3, \dots, x_k$ to represent the observed rate values in bps (bit per second) for each flow $f_1, f_2, f_3, \dots, f_K$ in the duration of Δt as defined in (3) and (4). The total number of observed rate values is N, so each random variable $x_1, x_2, x_3, \dots, x_k$ have N values.

$$\mathrm{x}_k(t_n) = \int_{t_n}^{t_n+\Delta t} f_k(t_n)\, dt \quad (3)$$

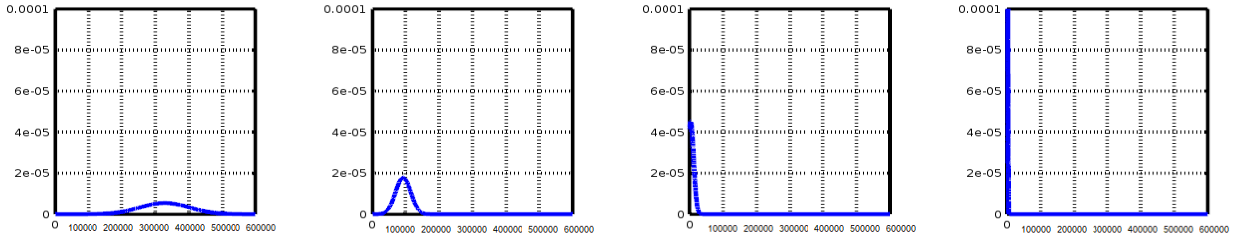
$$X_k = \{\mathrm{x}_k(t_1), \mathrm{x}_k(t_2), \dots, \mathrm{x}_k(t_n)\} \quad (4)$$

Due to this randomness characteristic, flow's or packet's rate values may not be possible to approximate with a single type of distributions (e.g., normal or uniform distribution function). CDF-based characteristic is selected to analyze the observed rate values distribution over a specific time. In this work, two types of CDF, nominal CDF, and empirical CDF were analyzed. Nominal CDF is used to analyze real-valued of packet/flow's rate random variable, as defined in X_k. It can denote as C_{X_k}, so flow $f_1, f_2, f_3, \dots, f_K$ have numerical distribution functions $C_{X_1}, C_{X_2}, C_{X_3}, \dots, C_{X_k}$. It is defined as in (5).

$$C_{X_k}(x_k) = P(X_k \leq x_k) \quad (5)$$

Empirical CDF tries to estimate the distribution function of the packet/flow's rate random variable based on real observed values from experiments. Rate random variable and empirical CDF can be denoted as $\dot{X}_k$ and $\hat{\mathrm{C}}_{X_k}$, so flow $f_1, f_2, f_3, \dots, f_K$ have empirical distribution functions $\hat{\mathrm{C}}_{X_1}, \hat{\mathrm{C}}_{X_2}, \hat{\mathrm{C}}_{X_3}, \dots,$. . If the number of observed rate values is N and a specific value of observed value is r, so $\hat{\mathrm{C}}_{X_k}$ can be described in (6).

$$\hat{\mathrm{C}}_{X_k} = \hat{\mathrm{C}}_{X_k}(N,r) = \frac{1}{N}\sum_{i=1}^{N} 1_{x_{k,i} \leq r} \quad (6)$$



(a) (b) (c) (d)

Figure 1. Probability Density Function (PDF) comparison for rate values distribution between packet and several classified flows; (a) overall packet, (b) known flows (similar distribution shape with packet), (c) very small rate flow (suspicious traffic), and (d) fixed & small rate flow (abnormal traffic).

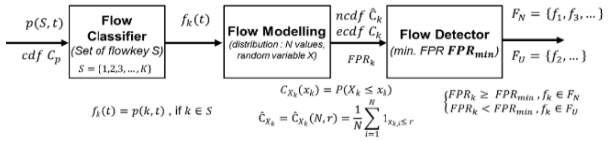


Figure 2. System model for detecting known and unknown flows.

where $1_{x_{k,i}}$ is defined as the indicator specific observed rate value (i.e., event) of flow f_k [8].

The ratio between each flow's rates with the overall packet's rates, called as FPR (Flow-to-Packet Ratio), is obtained using Kolmogorov-Smirnov Test (i.e., K-S Test) [9] for calculating a distance between empirical CDF of each flow and overall packet input. If flow's rate distribution $\hat{C}_{X_k}$ also, packet's rate distribution $\hat{C}_p$ for all observed values *N*, then the distance can be written in (7).

$$FPR_k = D_{k,p}(N) = \sup_r \left|\hat{C}_{X_k}(N,r) - \hat{C}_p(N,r)\right| \quad (7)$$

Flow Detector detects known flow or unknown flow by comparing the value of FPR_k with a minimum ratio value FPR_{min}, which may derive from a significant level that reflects a confidence level and the number of observed values of experiment results. Equation (8) is described the maximum distance between flow's rate distribution and packet's rate distribution with a number observed value N_k and N_p for specific significant value α.

$$FPR_{min} = D_{k,p}(min) = c(\alpha)\sqrt{\frac{N_k+N_p}{N_k.N_p}} \quad (8)$$

where $c(\alpha)$ is defined as in (9).

$$c(\alpha) = \sqrt{-\frac{1}{2}\ln\left(\frac{\alpha}{2}\right)} \quad (9)$$

By comparing FPR_k and FPR_{min}, the detector updates the set of normal flows F_N, and set of unknown flows F_U, as shown in (10).

$$\begin{cases} FPR_k < FPR_{min}\ ,\ f_k \in F_N \\ FPR_k \geq FPR_{min}\ ,\ f_k \in F_U \end{cases} \quad (10)$$

F_N can be easily collected by flow sampling because the flow's rate is well-distributed or correlated with overall packet's input rate. In contrast F_U need to be captured packet-by-packet, because the flow's rate is suspicious or too small for sampling that may cause a sampling error. Symbols and notations used in the system are described in Table 1.

IV. EXPERIMENTAL SIMULATIONS

Experimental simulation to collect observation data is conducted by generating a different type of packets with a different type of applications. CDF-based analysis requires data from packet-level monitoring as an initial observation data. The packets were captured and collected into the pcap-based file, and then were analyzed by using packet analyzing tools (i.e., Wireshark [10]) to get filtered packets based on specific flowkey (i.e., source IP, destination IP, and protocol) and generate a statistic report. This report is used to calculate the rate's distribution for each flow and overall packets. As a preliminary experiment, three network testing tools, known as D-ITG (Distributed Internet Traffic Generator) [11] were used, to generate packets from three pairs of IP source and destination, which are considered as known flows. One port scanning tool, known as nmap [12], is used to generate attack packets from a single pair of IP source and destination, which are considered as unknown flows (i.e., suspicious flows). The known flows' rates are expected to influence overall packet's rate, so the rate's distribution should be similar. The unknown flow's rate is very low because it sent only an initial packet of TCP communication (i.e., TCP sync message), so the rate's distribution is different.

TABLE 1
SYMBOLS AND NOTATIONS

Symbols	Notation	Description
Feature	S	Set of flowkey which contained in the packet header (ip address, protocol, tcp port, others)
Index	t	Indexes in the time domain
	k	Indexes for classified flows
	n	Indexes for observed rate values
Number	K	Number of classified flows based on feature S
	N	Number of observed rate values
	r	Specific rate value
Data	$p(S,t)$	Rate of packet input over time which containing all possible flowkeys (set of flowkey S)
	C_p	Nominal cumulative distribution function (ncdf) of the packet input
	$\hat{C}_p$	Empirical cumulative distribution function (ecdf) of the packet input
	F	Set of classified Flows
	$f_k(t)$	Rate of flow k over an observation period
	$x_k(t_n)$	Observed rate value of flow k for $t_n + \Delta t$
	x_k	The random variable of flow k for N observed values
	C_{X_k}	Nominal cumulative distribution function (ncdf) of flow k
	$\hat{C}_{X_k}$	Empirical cumulative distribution function (ecdf) of flow k
	FPR	Flow-to-packet ratio (Similar to SNR)
	FPR_k	Flow-to -packet ratio for flow k
	α	The confidence level of Observed Data
	FPR_{min}	Minimum flow-to-packet ratio to detect between Known and Unknown Flow
	F_N	Set of Known Flow (can be sampled)
	F_U	Set of Unknown Flow (need to be captured/inspected)

TABLE 2
TOTAL NUMBER OF CAPTURED PACKETS DURING OBSERVATION PERIOD

Flow	FlowKey	Number of Packets	Number of Bytes
Flow$_1$	172.16.1.13,172.16.1.10,tcp	26512	16836505
Flow$_2$	172.16.1.6,172.16.1.8,tcp	28617	18383512
Flow$_3$	172.16.1.12,172.16.1.14,udp	8906	5585816
Attack	172.16.1.102,172.16.1.16,tcp	18611	1079438
Ethernet	ff:ff:ff:ff:ff:ff,eth	193	15633

Table 2 shows the number of captured packets during an observation period (i.e., 10 minutes) for a different type of applications. Set of packets from three network testing tools are classified as flow$_1$, flow$_2$, flow$_3$, while a set of packets from port scanning tools is classified as an attack. Furthermore, another flow generated during the experiment was also observed. This flow is considered as Ethernet flow (use an Ethernet broadcast address as its destination), which is usually used by the host to send message to all hosts in the same Ethernet domain. Usually, Ethernet flow should be happened very rarely and incidentally, but in this experiment, it was sent regularly to generate another type of flow that can be considered as unknown flow (i.e., abnormal flow).

By leveraging packet-level monitoring for pcap-based raw observation data, each flow's rate graph and overall packet's rate also can be generated during an observation period, as depicted in Figure 3.

V. VERIFICATION AND ANALYSIS

A. CDF-based Analysis

As a preliminary analysis of the result depicted in Figure 4, a numerical CDF of all classified flow's rate and overall packet's rate can be analyzed, which is calculated by using (4). It is shown a comparison of numerical CDF for observed rate values from overall packets and classified flows. Network random generated flows flow$_1$, flow$_2$, flow$_3$ are most likely to have a similar shape with the overall packet, which is similar to our expectation as described in the previous section while the attack and ethernet flows have different shapes due to a statistical characteristic of both flows. However, intuitively, a CDF characteristic of each flow has a far distance from the overall packet's CDF characteristic. It may give inappropriate results in distance measurement between those two CDFs. So, the empirical CDF need to anayzed, which is derived by measuring frequency counts of specific rate's value from observed data, as formulated in (5). Figure 5 shows the empirical CDF comparison between all classified flow's rate and overall packet's rate. It intuitively describes the distribution distances between each flow and packet. Network random generated flows flow$_1$, flow$_2$, flow$_3$ rate's values have very close distance with the overall packet's rate value, while attack flow has far enough distance. Unfortunately, ethernet flow has a close distance with the overall packet's rate value and it against our expectation.

B. Distance Measurement

For further analysis, the distance between those rate distributions need to be calculated by applying the Kolmogorov-Smirnov test (K-S test) [9], as formulated in (6). However, the K-S test is only applied for empirical CDF because for nominal CDF, the distance between each flow's distribution and overall packet's rate distribution is intuitively very far. Thus, FPR value for each flow FPR_k is obtained based on this distribution distance, called as *K-S* (Kolmogorov-Smirnov) distance. The summary of K-S distance from all flows is shown in Table 3.

C. Hypothesis Testing

Table 3 also gives hypothesis testing [14] result that described on the P-values. By comparing P-value with a default significant level ($\alpha = 0.05$), it can be easily seen that K-S test results accept the null hypothesis (i.e., there is a similarity in the distribution) for all flows, except the *attack* flow. However, for further analysis or detecting each flow and categorizing it into known or unknown flows, another hypothesis need to be made by comparing each K-S distance with a minimum distance value, denoted as FPR_{min}. As a preliminary test, different FPR_{min} was adopted, valued from critical values derived from significant level and several observed data as formulated in (7). Table 4 shows the list of FPR_{min} for recommended significant level α as suggested by [13] for the number observed data N equal to 600.

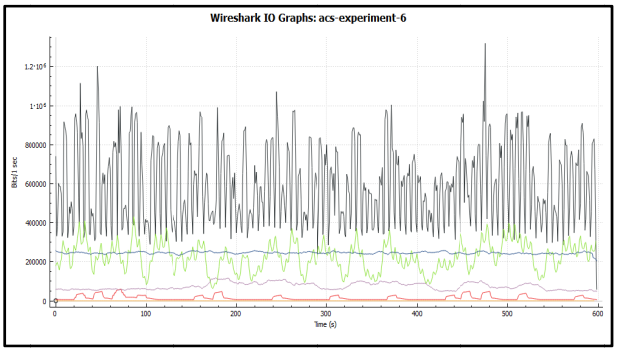


Figure 3. Flow rate graph from packet-level monitoring.

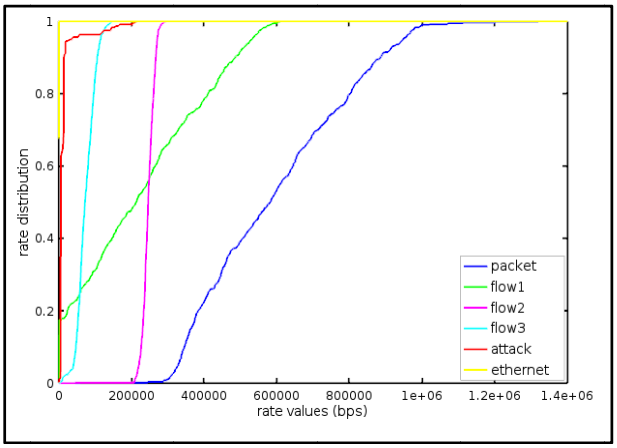


Figure 4. Numerical CDF for observed rate values from overall packets and classified flows.

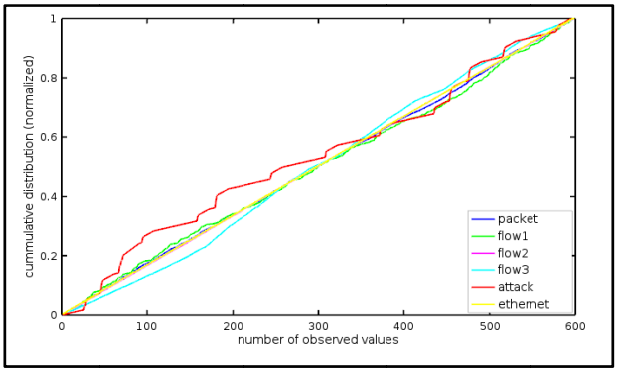


Figure 5. Empirical CDF for observed rate values from overall packets and classified flows.

TABLE 3
THE SUMMARY OF K-S DISTANCES

Flow	K-S distance = FPR_K	P-value
Flow$_1$	0.0284	0.967
Flow$_2$	0.0167	1.000
Flow$_3$	0.0584	0.251
Attack	0.1018	0.040
Ethernet	0.0184	1.000

TABLE 4
THE MINIMUM DISTANCES FOR DIFFERENT SIGNIFICANT LEVEL

A. Significant level (alpha)	B. Minimum distance = FPR_{MIN}
0.1	0.07066
0.05	0.07841
0.025	0.08546
0.01	0.093971
0.005	0.099929
0.001	0.112553

TABLE 5
K-S DISTANCE-BASED DETECTION RESULT

α	FPR_{MIN}	*Flow*$_1$	*Flow*$_2$	*Flow*$_3$	*Attack*	*Ethernet*
0.1	0.07066	F_N	F_N	F_N	F_U	F_N
0.05	0.07841	F_N	F_N	F_N	F_U	F_N
0.025	0.08546	F_N	F_N	F_N	F_U	F_N
0.01	0.093971	F_N	F_N	F_N	F_U	F_N
0.005	0.099929	F_N	F_N	F_N	F_U	F_N
0.001	0.112553	F_N	F_N	F_N	$\boldsymbol{F_N}$	F_N

By comparing each *FPR* in Table 3 with FPR_{min} in Table 4, another table to show detection results for all flows was generated. Table 5 show the summary of detection result for known flows F_N and unknown flows F_U with a different confidence level of observed data.

The detection result shows that our proposed method can consider *attack* flow as an "unknown" flow for almost all significant level α, except α = 0.001. This small value of significant level reflects a wide interval of confidence. Some references consider confidence level as 1 – α [14], so it can be said that our observed data is not achieved 99.9% of confidence level. It can be improved by increasing the number of observed data to decrease the value of FPR_{min}. However, the overall result shows that the unknown flow needs to be monitored by packet capturing for further inspection and analysis. Noted that, Ethernet flow is detected as "known" flow due to a few numbers of packet and consistency appearances during the observation period. It can be improved by reducing the granularity of the observation period, such as every 0.1 seconds (in this paper, the observation period is 1 second).

D. Comparison with Maximum Likelihood Estimator (MLE)

Relative sampling error in Maximum Likelihood Estimator (MLE) and confidence interval in the Kolmogorov-Smirnov test were also compared. It may give an initial idea about the value of relative error and confidence interval for the same observed packet's data. MLE's relative sampling error relies on the number of packet's samples (depends on the sampling rate and ratio), while the confidence interval relies on significant level (depends on the number of observation data). Finally, an improvement in CDF-based analysis over MLE can be achieved, and hopefully, a similar accuracy result with the MLE.

CONCLUSION

This paper describes an idea for CDF-based flow detection for network flow sampling and packet capturing. The result shows that the distance between flow's and packet's empirical CDF based rate distribution can be used to detect "known" and "unknown" flows with a most significant confidence level in distribution distance test. Known flows can be monitored using flow-level monitoring to reduce the overhead process, while unknown flows need to be captured packet-by-packet to increase accuracy. In the future, this approach can be implemented to monitor network traffic in real-time.

REFERENCES

[1] A. C. Risdianto, J. W. Kim, "A balanced collection of flow visibility for effective SDN-coordinated flow clustering and tagging," in *Proc. Korea Inst. Commun. Inform. Sci. Winter Conf. 2017*, Jeongseon, Korea, 2017.

[2] S. Panchen, P. Phaal, N. McKee (2001). InMon corporation's sFlow: A method for monitoring traffic in switched and routed networks.

[3] Y. Afek, A. B. Barr, S. L. Feibish, L. Schiff, "Sampling and large flow detection in SDN", in *Proc. 2015 ACM Special Interest Group Data Commun.*, London, UK, 2015, pp. 345-346.

[4] G. Cheng, Y. Tang, W. Ding, "A double-sampling and hold based approach for accurate and efficient network flow monitoring," in *Proc. Int. Conf. Computational Sci.,* China, 2007, pp. 857-864.

[5] J. M. C. Silva, P. Carvalho, S. R. Lima, "A modular architecture for deploying self-adaptive traffic sampling," in *Proc. Int. Federation Inform. Process. Int. Conf. Autonomous Infrastructure Manage. Security,* 2014, pp. 179-183.

[6] R. Hofstede, P. Čeleda, B. Trammell, I. Drago, R. Sadre, A. Sperotto, and A. Pras, "Flow monitoring explained: From packet capture to data analysis with netflow and ipfix, " *IEEE Commun. Surveys Tutorials*, vol. 16, no. 4, pp. 2037-2064, May, 2014.

[7] P. Phaal and S. Panchen. (2017, June). *Packet sampling basics*. [Online]. Available: http://www.sflow.org/packetSamplingBasics/index.htm.

[8] A. W. V. Vaart, *Asymptotic Statistics*. Cambridge: Cambridge University Press, 1998, p. 265.

[9] H. W. Lilliefors, "On the Kolmogorov-Smirnov test for normality with mean and variance unknown," *J. American Statistical Assoc.,* vol. 62, no. 318, pp. 399-402, Jun. 1967.

[10] Wireshark. (2017, June). *Wireshark* [Online]. Available: https://www.wireshark.org/.
[11] Dipartimento di Ingegneria Elettrica e delle Tecnologie dell'Informazione (2017, June). *D-ITG: Distributed Internet Traffic Generator* [Online]. Available: http://www.grid.unina.it/software/ITG/.
[12] G. Lyon (2017, June). *Nmap: the Network Mapper – Free Security Scanner* [Online]. Available: https://nmap.org/.
[13] University de Montreal. (2017, June). *Critical Values for two-sample Kolmogorov-Smirnov test (2-sided)* [Online]. Available: https://www.webdepot.umontreal.ca/Usagers/angers/MonDepotPublic/STT3500H10/Critical_KS.pdf.
[14] D. M. Lane. (2017, June). *Significance Testing and Confidence Intervals* [Online]. Available: http://onlinestatbook.com/2/logic_of_hypothesis_testing/sign_conf.html.