Traffic model using a novel sniffer that ensures the user data privacy

ALBERT ESPINAL[1,*], REBECA ESTRADA[1], CARLOS MONSALVE[1]

[1]Escuela Superior Politecnica del Litoral, ESPOL, ECUADOR
aespinal@espol.edu.ec

Abstract:- Nowadays, the traffic over the networks is changing because of new protocols, devices and applications. Therefore, it is necessary to analyze the impact over services and resources. Traffic Classification of network is a very important prerequisite for tasks such as traffic engineering and provisioning quality of service. In this paper, we analyze the variable packet size of the traffic in an university campus network through the collected data using a novel sniffer that ensures the user data privacy. We separate the collected data by type of traffic, protocols and applications. Finally, we estimate the traffic model that represents this traffic by means of a Poisson process and compute its associated numerical parameters.

Keywords:-Traffic modeling, Data Privacy, Sniffers, Networks

Received: December 21, 2018. Revised: February 26, 2018. Accepted: March 26, 2019.
Published: April 30, 2019

1 Introduction

Understanding the behavior of the network traffic is crucial and an important prerequisite for planning the traffic engineering and apply quality of service; also, for traffic modeling and prediction.

Additionally, the network traffic is changing because of the convergence (voice, data and video). The applications are heterogeneous and complex; the number of mobile devices accessing the networks are increasing exponentially. New protocols like the IPv6 are present in the internet, and technologies such as Internet of Things (IoT) will allow the connection of millions of new devices. The study from Cisco Systems: forecast and trends [1], predicts that by 2022, the number of devices connected to IP networks will be more than three times the global population; the smartphone traffic will exceed PC traffic; and traffic from wireless and mobile device will account 71 percent of total IP traffic.

In packet-based networks, like the internet or the Local Area Networks (LANs), the transmission of information is performed in discrete packets [2]. When we need analyze and modelling the network traffic, we can to considerate two stochastically variables: the packet size and the inter-arrival time [3]. This study is focus on packet size (packet length).

Normally we can measure the traffic network by means of active polling and passive monitoring [4]. The active method generates new traffic, inject it into the network, while passive method consists on monitor, and capture the network traffic. In this case, we use the passive form for capture traffic, analyze the packet headers and produce statistics. One drawback of the method is the privacy of the data to be captured, because the traditional sniffers saves the entire packet: headers and payload. The passive measurement can be performed at various levels like byte, packet, flow, and session [5]. We use packet level because the most of the network's problems occur in this level; is independent of the protocols, and avoid the encrypted payload.

In this work, we propose to develop a sniffer that assures the user data privacy in order to analyze the traffic of a university campus network to estimate the model for such traffic.

The rest of the paper is organized as follows: section 2 provides information about related works; in section 3 we present the novel sniffer; in section 4 we show the data collection, classified by type of traffic, by protocols, and by application, according to the variable packet size. Section 5 presents the traffic model that characterize the realistic traffic analyzed. The paper ends with the conclusion in section 6.

2 Related works

Many works have analyzed the network traffic based on packet size, using methods such as statistical analysis, pattern recognition methods, length of the application messages, packet flows, user behavior, etc. Additionally, these studies had suggested models to simulate the realistic network traffic.

In [6], Sinha et al. observed that the internet traffic was bimodal at packet sizes of 40 and 1500 bytes, different to data in [7] that was tri-modal with packet sizes around 40, 765 and 1500 bytes. Wu et al. in [8] analyzed flow records and classified this by applications using machine learning. A study for identifying network

traffic based on message size analysis is present in [9], and a Gaussian model is proposed for characterize the application-level protocols. Lee et al. in [10] present a study about the self-similarity of traffic using bandwidth frequency distribution. In [11] a work that classify network traffic using three classification approaches based on transport layer ports, host behavior and flow features is present. In [12] Zhang et al. evaluate the amount of UDP and TCP traffic, in terms of flows, packets and bytes. A work over internet data traffic generated in a university campus and a model for predict internet data traffic is present in [13]. Cao et al. in [14] demonstrate that the number of active connections has an effect on traffic characteristics.

Regarding the traffic modelling, Vicari present in [15] a model for internet traffic from the user perspective, using distribution functions applied to data. In [16], Maheshwari et al. design a Hidden Markov model for network traffic and validate it for different packet sizes. A study for modeling TCP/IP traffic over a wireless network is present in [17]. Mueller in [18] specifies a traffic model based on object sizes at the application layer applied to wireless network.

3 Proposed sniffer

One of the critical issues in the process of capturing network traffic is the use of the sniffer. This is owing to the fact that they normally capture the entire packet, which includes headers and payload. Network administrators need some kind of confidentiality agreement to avoid problems because of the inappropriate use of the user information. This motivates our work. We propose to implement a sniffer that guarantees the privacy of the information avoiding the capture of the payload of the packages. In addition, with the deployment of IPv6, our sniffer would be able to differentiate a dual stack environment with IPv4. Finally, the sniffer should have low resource consumption, which allows a more efficient capture of the data.

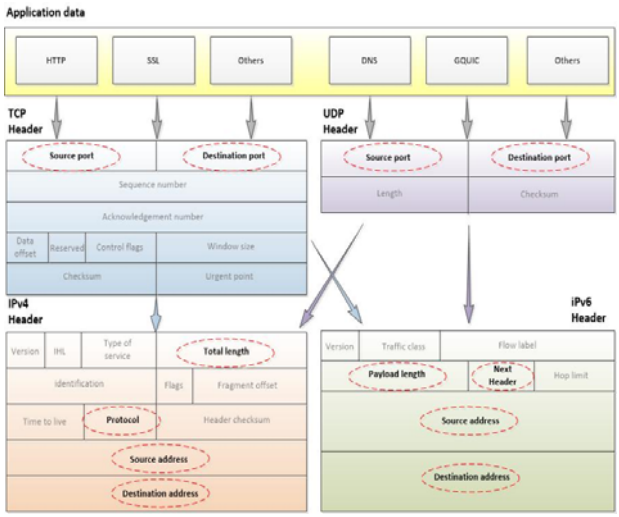

Fig. 1. Packet headers

The sniffer, called TinySniff, is written in C language and runs under Linux operating system. It is portable and lightweight software consumes small amount of resources (i.e. memory and CPU). Can capture traffic in LAN and WLAN scenarios, and store the headers captured in flat files, in text format.

TinySniff is design to capture the following fields in the header of a package for further analysis: total length (IPv4) o payload length (IPv6), source address, destination address, protocol (IPv4) or next header (IPv6), source port, and destination port, as shown in figure 1. An example of data in TXT format is showed in figure 2.

36156	94	TCP	45820	443
36157	94	TCP	45820	443
36158	94	TCP	45820	443
36159	94	TCP	45820	443
36160	94	TCP	45820	443
36161	1412	GQUIC	59096	443
36162	1412	GQUIC	59096	443
36163	1412	GQUIC	59096	443
36164	1412	GQUIC	59096	443
36165	1412	GQUIC	59096	443
36166	1412	GQUIC	59096	443
36167	86	ICMPv6		
36168	1412	GQUIC	37386	443
36169	1412	GQUIC	37386	443
36170	1412	GQUIC	37386	443

Fig. 2. Example of TinySniff output

4 Data collection and analysis

We implement a scenario for capture realistic traffic in an university campus network shown in figure 3. We install TinySniff on a desktop computer with Linux Ubuntu version 16.04 LTS. Its technical specifications are: AMD FX-8300 Eight-core processor, 24 GB of RAM, and two-network interface cards (NIC) Ethernet. One NIC is for PC management, and another for capture traffic. We connect the NIC for capture, in a gigabit port of access layer Cisco switch, and configure this port in trunking mode for access all VLAN traffic.

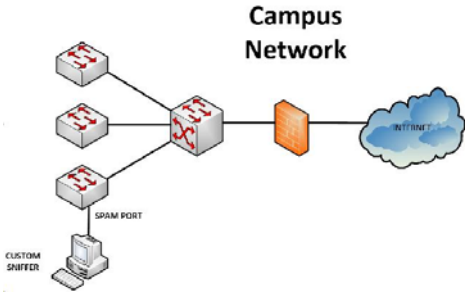

Fig. 3. Scenario of network traffic capture

The traffic capture was collected on October 25, 2018 during 5199 seconds between 08:54:33 and 10:21:12, peak traffic time. We collect near of 10 million of packets, with average 1899 packets per second and average packet size of 709 bytes. Then, this data was classified by type of traffic (e.g. IPv4, IPv6 and ARP), by protocols (e.g. TCP and UDP), and by applications (e.g. HTTP, DNS, GQUIC, etc.). Table 1, 2 and 3 present the traffic classification by type of traffic, by protocols, and by applications respectively.

From table 1, it can be observed that IPv4 traffic is still more considerable than IPv6 in this network. Without ARP packets (these are local traffic), IPv4 represents 97% of the total traffic compared to 3% of IPv6. Table 2 shows that TCP traffic is significantly higher with respect to UDP (91.42% versus 8.42%). Regarding IPv6, ICMPv6 traffic is considerable. Relating to applications, HTTP, SSL, TLS are the applications more relevant over TCP. GQUIC and MDNS over UDP.

Table 1. Data by traffic type.

Traffic Type	Frequency	Percent
IPv6	259.312	2,61%
IPv4	8.442.917	85,10%
ARP	1.001.556	10,09%
Others	217.789	2,20%
Total	9.921.574	100%

Table 2. Data by protocol.

Protocol	IPv4		IPv6	
	Frequency	Percent	Frequency	Percent
UDP	711.064	8.42%	78.553	30.29%
TCP	7.718.693	91.42%	30.341	11.70%
ICMP	13.160	0.16%	150.418	58.01%
Total	8.442.917	100%	259.312	100%

Table 3. Data by application

	Application	IPv4		IPv6	
		Frequency	Percent	Frequency	Percent
TCP	SSL	4.294.926	55,64%	26.592	87,64%
	HTTP	867.062	11,23%	0	0,00%
	Others	2.556.705	33,12%	3.749	12,36%
	Total	7.718.693	100 %	30.341	100 %
UDP	GQUIC	314.816	44,27%	9.494	12,09%
	MDNS	123.093	17,31%	37.293	47,47%
	SSDP	83.116	11,69%	4.154	5,29%
	BootStrap	47.422	6,67%	3.310	4,21%
	NETBIOS	36.124	5,08%	0	0,00%
	DNS	32.232	4,53%	1.594	2,03%
	Others	74.261	10,44%	22.708	28,91%
	Total	711.064	100 %	78.553	100 %

This work analyzes the variable packet size; the packet length usually is between 40 and 1500 bytes. To analyze the packet size, we take intervals of 10 bytes for discrimination (i.e. 0-10, 11-20, 21-30, etc.). Figure 4 shows the behavior of packet size according to traffic type (IPv4, IPv6, ARP). Figure 5 and 6 present the variable packet size for IPv4 protocol and for IPv6 respectively. The analysis of IPv4 applications (under TCP and UDP) and packet size are shown in figures 7 and 8.

From Fig. 5 to 8, we can see that there is a bimodal traffic distribution with 48.32% of packets around of 60 bytes size, and 38,42% around 1500 bytes. For the first size, all traffic types contribute to this trend, while for second size only IPv4 traffic contributes. If we analyze the IPv4 traffic, it can be observed that TCP is the main protocol over UDP and contributes over both bimodal trends.

This IPv4 traffic is bimodal too, with 40.27% of packets around 60 bytes and 45.13% around 1500 bytes. TCP packets are the main factor in this behavior with 41.66% around 60 bytes and 49.28% around 1500 bytes. HTTP, SSL and TLS are the main applications and represent more than 95% of total IPv4 TCP packets and contributes with 41.66% of packet around 60 bytes and 49.28% around 1500 bytes. UDP packets contributes mainly around 1400 bytes with 38.88%, and the main application for this behavior is GQUIC (around 1400 bytes). Other UDP applications contribute with packets between 60 and 300 bytes in a sparse form.

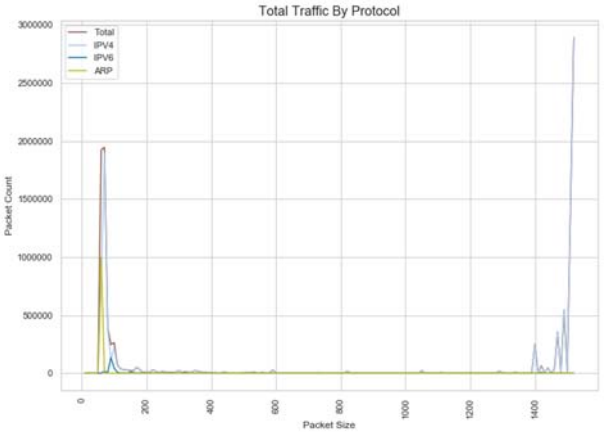

Fig. 4. Total traffic by packet size

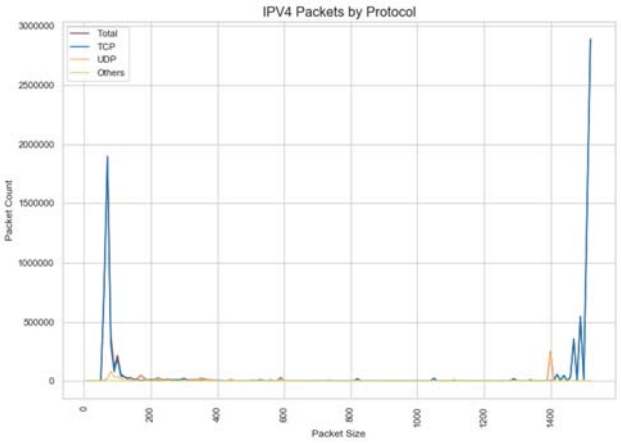

Fig. 5. IPv4 traffic by packet size

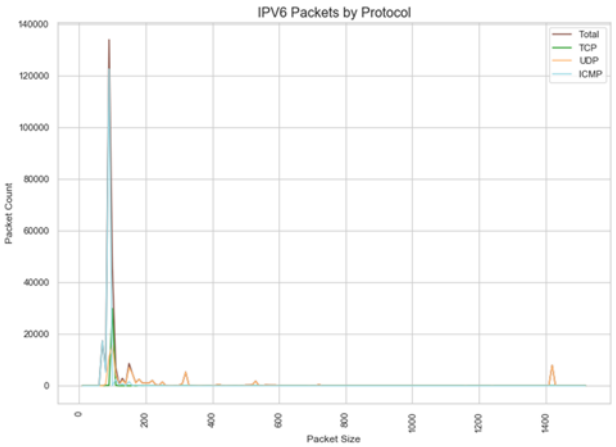

Fig. 6. IPv6 traffic by packet size

The analysis of IPv6 traffic show that contribute with small packets around 80 bytes with 85.88%, mainly ICMPv6 packets. TCP and UDP traffic over IPv6 are still limited in this university campus network. Applications as HTTP and SSL over TCP, and MDNS over UDP, are the most relevant.

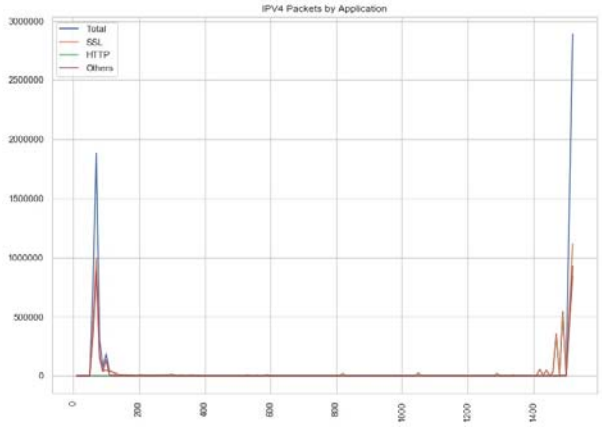

Fig. 7. IPv4 – TCP applications traffic by packet size

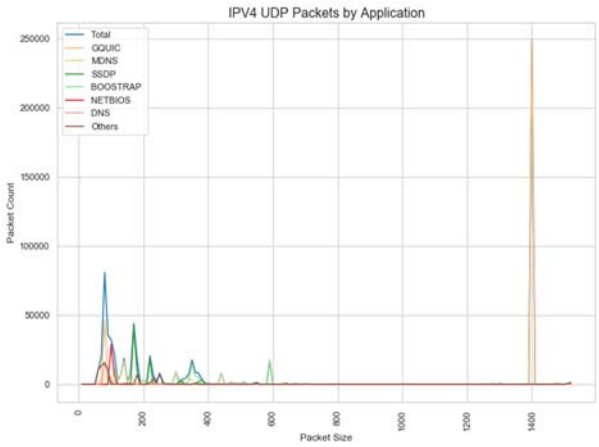

Fig. 8. IPv4 – UDP applications traffic by packet size

5 Traffic modelling

Taking into account the analysis of the network traffic analyzed in the previous section, we estimate some models using the Poisson probability distribution function, based on traffic type, protocols and applications.

For total traffic presented in fig. 4, results a fitted model as a mixture of two Poisson distributions with parameters $\lambda_1 = 84.38$, and $\lambda_2 = 1457.11$. The probability that the length of a packet belongs to the first distribution is 0.545, while for the second distribution the probability of a packet following that distribution is 0.455. Finally, the model is the result of the sum of two Poisson distributions as in (1):

$$P(X = x) = 0.545 * \frac{e^{-84.38}84.38^x}{x!} + 0.455 * \frac{e^{1457.11}1457.11^x}{x!} \quad (1)$$

Where x is the ocurrence of packet size variable. In fig. 9 we show the histogram of data and the simulate model for network traffic total.

For IPv4 network traffic the parameters are $\lambda_1 = 90.61$ and $\lambda_2 = 1458.72$. The probability that the length of a packet belongs to the first distribution is 0.469, while for the second distribution the probability of a packet following that distribution is 0.531. The model is showed in (2). For IPv6 network traffic, the model is as in (3), with parameters $\lambda_1 = 1083.92$, and $\lambda_2 = 103.86$. The probability that the length of a packet belongs to the first distribution is 0.0505, while for the second distribution the probability of a packet following that distribution is 0.9495. Fig 10 and 11 show these simulate models.

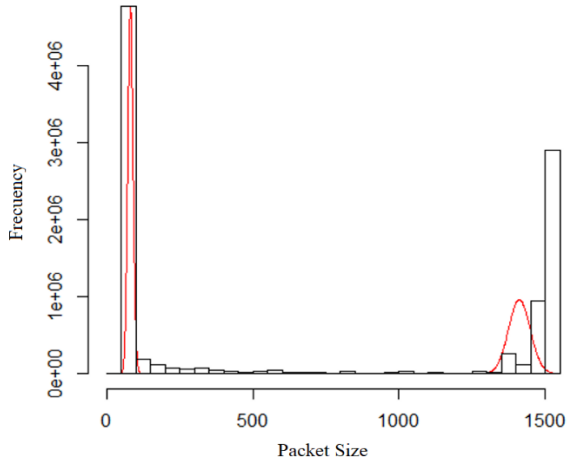

Fig. 9. Poisson model for Traffic Total

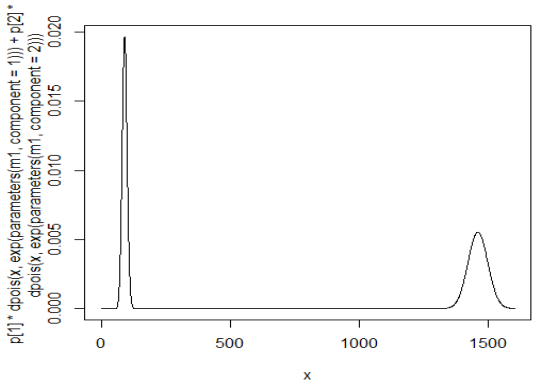

Fig. 10. Poisson model for IPv4 Traffic

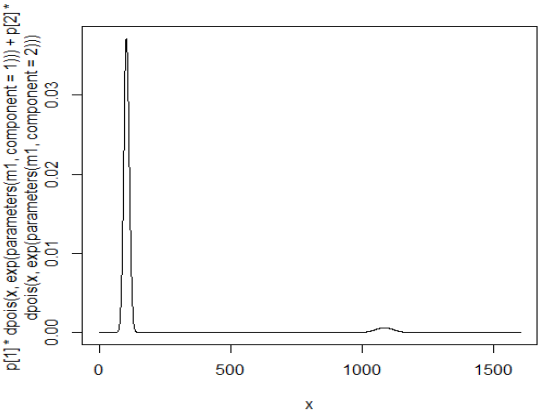

Fig. 11. Poisson model for IPv6 Traffic

$$P(X = x) = 0.469 * \frac{e^{-90.61}90.61^x}{x!} + 0.531 * \frac{e^{1458.72}1458.72^x}{x!} \quad (2)$$

$$P(X=x) = 0.0505 * \frac{e^{-1083.92}1083.92^{x}}{x!} + 0.9495 * \frac{e^{103.86}103.86^{x}}{x!} \quad (3)$$

Additionally, we present models for protocols TCP and UDP, over IPv4 and IPv6. Table 4 resume the parameters of the models, where λ_1 represent average occurrence in interval 1, λ_2 represent average occurrence in interval 2, P_1 is the probability for a packet following the first distribution, and P_2 is the probability of a packet following the second distribution. For IPv6 only one Poisson distribution is necessary for fit the data. Figures 12 to 15 show the simulation of these models; and the equations in (4) (5) (6) (7).

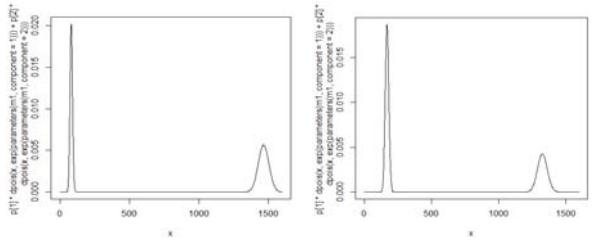

Fig. 12-13. Poisson models for IPv4 - TCP and UDP Traffic

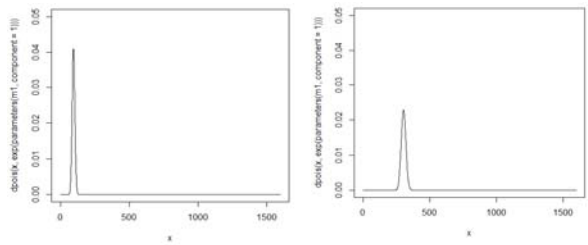

Fig. 14-15. Poisson models for IPv6 - TCP and UDP Traffic

$$P(X=x) = 0.544 * \frac{e^{-1467.92}1467.92^{x}}{x!} + 0.456 * \frac{e^{81.21}81.21^{x}}{x!} \quad (4)$$

$$P(X=x) = 0.611 * \frac{e^{-169.84}169.84^{x}}{x!} + 0.389 * \frac{e^{1327.37}1327.37^{x}}{x!} \quad (5)$$

$$P(X=x) = \frac{e^{-94.99}94.99^{x}}{x!} \quad (6)$$

$$P(X=x) = \frac{e^{-305.12}305.12^{x}}{x!} \quad (7)$$

Table 4. Data by application

Protocol		λ_1	λ_2	P_1	P_2
IP v4	TCP	1467.92	81.21	0.544	0.456
	UDP	169.84	1327.37	0.611	0.389
IP v6	TCP	94.99	-	-	-
	UDP	305.12	-	-	-

Finally, table 5 presents the parameters for the applications that mainly contribute to the total network traffic.

Table 5. Data by application

Protocol			λ_1	λ_2	P_1	P_2
IPv4	TCP	HTTP	1496.52	-	-	-
		SSL	87.93	1437.97	0.444	0.556
	UDP	GQUIC	1383.01	79.65	0.807	0.193
		MDNS	129.25	422.95	0.61	0.31
IPv6	TCP	SSL	94.36	-	-	-
	UDP	MDNS	212.75	-	-	-

6 Conclusions

This paper presents results for stochastic behavior of packet size variable using network traffic measurements in a university campus network. The results show that there is a bimodal traffic distribution with packets around 60 and 1500 bytes. IPv4 packets represents a big impact in this behavior, mainly TCP packets, and the applications that mark this trend are HTTP and SSL.

Network administrators can use these results to design better networks and optimize network traffic in order to give security policies, QoS provisioning, and ensure efficient utilization of resources.

We development models for characterize the network traffic based using mixture Poisson distribution and provide the best statistical fit to the packet size variable of the dataset considered in this paper. These models simulate the data by traffic type, protocols and applications. Research community can use these distribution parameters presented for built traffic models and apply in other studies in the areas of computer networking and traffic engineering.

The authors thank to technical staff from Electrical Engineering and Computer Science Faculty of Escuela Superior Politecnica del Litoral, ESPOL, by the facilities for capture of network traffic.

References

1. CISCO, “Cisco visual networking index: forecast and trends,” 2017. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html.
2. D. K. Arrowsmith, R. J. Mondrag, *Modelling Network Data Traffic*. (2005).
3. I. W. C. Lee, A. O. Fapojuwo, Stochastic processes for computer network traffic modeling. *Comput. Commun.* **29**, 1–23 (2005).
4. R. Pries, F. Warmer, D. Staehle, K. Heck, and P. Tran-Gia, Traffic measurement and analysis of a broadband wireless internet access. *IEEE Veh. Technol. Conf.* (2009).
5. S. Maheshwari, K. Vasu, C. Kumar, and S. Mahapatra, Measurement and Comparative Analysis of UDP Traffic over Wireless Networks. *Int. Conf. Wirel. Networks* (2011).
6. R. Sinha, C. Papadopoulos, and J. Heidemann, Internet Packet Size Distributions: Some Observations. *Network* 1–7 (2007).
7. W. John and S. Tafvelin, Analysis of internet backbone traffic and header anomalies observed. *dl.acm.org* (2007).

8. X. L. Wu, W. M. Li, F. Liu, and H. Yu, Packet size distribution of typical Internet applications. *2012 Int. Conf. Wavelet Act. Media Technol. Inf. Process. ICWAMTIP 2012* 276–281 (2012).
9. A. Hajjar, J. Khalife, and J. Díaz-Verdejo, Network traffic application identification based on message size analysis. *J. Netw. Comput. Appl.* **58**, 130–143 (2015).
10. S. Lee, Y. Won, and D. J. Shin, On the multi-scale behavior of packet size distribution in internet backbone network. *NOMS 2008 - IEEE/IFIP Netw. Oper. Manag. Symp. Pervasive Manag. Ubiquitous Networks Serv.* 799–802 (2008).
11. H. Kim, K. Claffy, M. Fomenkov, D. Barman, M. Faloutsos, and K. Lee, Internet traffic classification demystified: myths, caveats, and the best practices. *Proc. 2008 ACM Conex. Conf.* **50**, 1–12 (2008).
12. M. Zhang, M. Dusi, W. John, and C. Chen, Analysis of UDP traffic usage on internet backbone links. *Proc. - 2009 9th Annu. Int. Symp. Appl. Internet, SAINT 2009* 280–281 (2009).
13. O. J. Adeyemi, S. I. Popoola, A. A. Atayero, D. G. Afolayan, M. Ariyo, and E. Adetiba, Exploration of daily Internet data traffic generated in a smart university campus. *Data Br.* **20**, 30–52 (2018).
14. J. Cao, W. S. Cleveland, D. Lin, and D. X. Sun, Internet Traffic Tends Toward Poisson and Independent as the Load Increases. *Nonlinear Estim. Classif.* 83–109 (2013).
15. N. Vicari, Modeling of Internet Traffic : Internet Access Influence, User Interference, and TCP Behavior. Norbert Vicari Würzburger Beiträge zur Leistungsbewertung Verteilter Systeme. (2003).
16. S. Maheshwari, S. Mahapatra, and K. Cheruvu, Measurement and Forecasting of Next Generation Wireless Internet Traffic. (2018).
17. I. W. C. Lee and A. O. Fapojuwo, Analysis and modeling of a campus wireless network TCP/IP traffic. *Comput. Networks* **53**, 2674–2687 (2009).
18. Mueller, C. M. On the importance of realistic traffic models for wireless network evaluations. *COST 2100 12th MCM* 6–13 (2010).