Collecting and Characterizing a Real Broadband Access Network Traffic Dataset

Martin Andreoni Lopez*‡, Renato Souza Silva†, Igor D. Alvarenga*, Gabriel A. F. Rebello*
Igor J. Sanz*, Antonio G. P. Lobato*, Diogo M. F. Mattos*‡, Otto C. M. B. Duarte*, and Guy Pujolle‡

*Grupo de Teleinformática e Automação - Universidade Federal do Rio de Janeiro (COPPE/UFRJ)
Rio de Janeiro, Brazil - Email:{martin, alvarenga, gabriel, sanz, antonio, diogo, otto}@gta.ufrj.br

†Laboratório de Redes de Alta Velocidade (RAVEL) - Universidade Federal do Rio de Janeiro (COPPE/UFRJ)
Rio de Janeiro, Brazil - Email:{renato@ravel.ufrj.br}

‡Laboratoire d'Informatique de Paris 6 - Sorbonne Universities, UPMC Univ Paris 06
Paris, France - Email: Guy.Pujolle@lip6.fr

***Abstract*—Broadband Internet access security relies in the implementation of perimeter policies and in the adoption of access control lists. These measures are precarious because they are based on common and not frequently updated profiles that lack residential users threat information. In this paper, we analyze and profile residential users traffic from fixed broadband Internet access networks of a large telecommunication operator for a period of one week, and we obtain the profile of security alarms generated by an intrusion detection system. The results show that the proposed characterization allows the classification of alerts with a sensitivity of 93% in the differentiation of legitimate and anomalous flows and allows a 73% reduction of the traffic directed to the traffic analyzer, thus validating the collected dataset and enabling more dynamic and efficient access network security.**

I. Introduction

Internet access is present in 45.5% of Brazilian households [1] and more than 75% in the US [2]. However, Internet access infrastructure providers and regulatory agencies face the security of the Internet access service in a precarious way and without preventive actions to mitigate possible losses. Internet access security is implemented collectively by security perimeters [3]. Generally, there are no individual measures on the provider side to guarantee home network security [4]. The installed security restrictions become obsolete as network usage evolves, and adopts new technologies, such as cloud computing and the Internet of Things which bring new threats to home networks [5]. Thus, infrastructure providers systematically need to analyze security alerts and perimeter restrictions.

Defining new security policies depends on knowing the security threats that are present on the network to ensure greater accuracy in defense and mitigation of attacks [6], [7]. Accessing real network traffic is one of the major research challenges related to Internet traffic analysis and knowledge extraction from threat situations. Although there are some datasets available for research [8], [9], [10], [11], the data are often not actual or have been created with artificial attack patterns and threats. Other limitations on accessing real data fall on regulatory issues and user privacy, as the data may contain sensitive or confidential information.

In this paper, we analyze and characterize the traffic of a broadband Internet access network by differentiating normal from anomalous traffic, that is marked as threat by an IDS. We collected real and anonymized data from a major telecommunications operator[1]. The dataset is created by capturing 5 TB of access data of 373 residential broadband users in the city of Rio de Janeiro, Brazil. The dataset contains legitimate traffic, attacks and other security threats. An Intrusion Detection System (IDS) inspects the traffic and then summarizes a set of flow features associated with either an IDS alert or a legitimate traffic class. Finally, we evaluate the performance of a decision tree classifier to identify suspicious flows in real time.

In a previous work, we study the adequacy of classifiers and flow computation in the identification of network anomalies [12]. Other proposals create datasets based on synthetic and out-of-date attacks for the study of network security [10]. There are also proposals that collect data from honeypots [13] predict attackers behavior based on stochastic processes [14]. In contrast with other works, we analyze and characterize a real dataset consisting of captured data packets from residential fixed broadband users of a major telecommunications operator within a week. The results show that we achieve 93% accuracy in the identification of flows that generates security alerts, based only on flow statistic. Adopting the proposed classification has the potential to reduce up to 73% of the traffic passed to packet analysis tools, including those that scan higher layers.

The remainder of the paper is organized as follows. Section II covers related work. The problem of processing, characterization of flows, characterization of alerts and the procedure of data collection are presented in Section III. Section IV explains the analysis procedure and presents its results. Finally, Section V concludes the paper.

II. Related Work

Building a dataset that allows the analysis of the actual usage profile of the network is a challenge because the networks

[1] Anonymized data can be consulted through email contact with authors.

are in constant changes and the data can be captured in different locations of the network. Heidemann and Papadopoulos point out the network ideal locations for data capture, address the difficulties in anonymizing and extracting knowledge from anonymized data, and discuss the main datasets available [6]. Among the datasets for security research, the most commonly used is KDD [10], whose characterization was performed by Tavallaee *et al.* [15].

In addition to the location of the data collection, another important factor to avoid the biased contamination of the dataset is the sample size. Shiravi *et al.* discuss the creation and analysis of a real dataset to detect intrusions [16]. The authors also generated three categories of real-traffic datasets that are available to the research community. The main disadvantage of those datasets lies on the fact that all anomalous traffic was collected from simulated attacks in controlled environments. Our work uses an IDS to identify malicious and anomalous traffic. While there is no guarantee of completeness in this approach, we avoid the problems arising from the insertion of artificial attacks such as the creation of a biased dataset.

New types of attacks have been emerging due to the creation of new services and different types of traffic. Therefore, old datasets like in [10] can no longer be used to benchmark modern intrusion detection systems. The proposed approach in [16] analyzes different samples of TCP services (HTTP, SSH, FTP, SMTP, IMAP and POP3) in relation to the number of requests in time and compares the curves obtained with distributions of known probabilities. Similar distributions are then used to assemble attack profiles that can be reproduced synthetically. The dataset generated in this paper includes, among others, the same services and can generate up-to-date attack profiles.

Some research papers explore the use of honeypots for the assembly of real datasets. Chen *et al.* analyze a dataset created from 491 TCP honeypots to study the stochastic characteristics of attacks [14]. Song *et al.* propose to modify the honeypots to emulate proactive systems that even visit malicious pages and join botnets [13]. In addition, to increase the realism of the collected data, both proposals seek to overcome the difficulty of differentiating legitimate and malicious traffic by considering that all the traffic collected in the honeypots comes from attacks. However, it is not possible to guarantee, or verify, this premise. In this paper, we use an IDS to identify malicious traffic to address this challenge. The main difference between the methods lies in the probability of occurrence of false positives and false negatives in the dataset, respectively prevalent in their work [13], [14], and in this work.

Drapper-Gil *et al.* present the characterization of virtual private network traffic based on temporal characteristics of the flows [17]. Thus, the authors propose the use of 8 features to classify flows into 14 different types, including VPN and non-VPN flows. The results show that the decision trees machine learning algorithm presents a slightly better performance than the nearest k-neighbors algorithm. However, the authors do not assess whether the characteristics used for classification are those that best describe the dataset, nor do they assess the profile of network usage. The reduction of dimensionality of the dataset can be done through feature selection methods, by selecting the ones that best describe the data, or through techniques that take the data to another vector space [18]. This paper uses the techniques described in these works to carry out an evaluation of the relevance of attributes regarding to the identification of anomalies, also resulting in a set of eight main characteristics.

Kato *et al.* propose the use of deep learning to perform the characterization of network traffic, since they advocate that deep learning is capable of extracting more complex patterns than other techniques [19]. Nie *et al.* use a Bayesian network to detect anomalies and model the network traffic matrix [20]. The results show that traffic forecasting is accurate, but they focus on the cloud computing environment and do not aim to predict end-user traffic. This paper offers a complementary dataset to the one used by Kato *et al.*, since it is only composed of end-user traffic.

In this paper, we create a real dataset of the captured Internet access packets of 373 fixed broadband users of a major telecommunications operator in Rio de Janeiro, during a week between February 24 and March 4, 2017. The recommendations on the location are observed [6], as well as the methodology used to characterize [15]. Security alarms are identified by analyzing the data through an intrusion detection tool, in contrast with the synthetic [16] and honeypot-based [13], [14] approaches. In this way, it is intended to enable the study of classification methods in updated data to complement and update previous studies [16], [19].

III. Traffic Analysis and the Residential User Access Dataset to the Internet

The characterization of alert profile in an access network requires the monitoring, processing and management of large volumes of data generated in real time. This large data volume is processed by an Intrusion Detection System (IDS) and is also correlated with network flow information [21], [12]. Security Information and Event Management (SIEM) tools perform this type of task at a high financial cost and can still lead to delays in threat detection. On average, the reaction to security threats is taken after 123 hours of occurrence and, in case of a data leak, the delay in identifying this security flaw reaches 206 days [22]. By knowing the most common profile of security alerts, it fastens the detection of the leaked information and exploited vulnerabilities.

Figure 1 shows a typical access topology for the broadband service composed of a Customer Premises Equipment (CPE) connected to a Digital Subscriber Line Asymmetric Multiplexer (DSLAM), a transport network, such as Multiprotocol Label Switching (MPLS) network, and a section aggregator Broadband Remote Access Server (BRAS) that authenticates the session of the users through a RADIUS server, also responsible for auditing the network usage. Thus, in an access network for fixed broadband users, the inspection is performed only after the aggregation of the traffic, since there are no

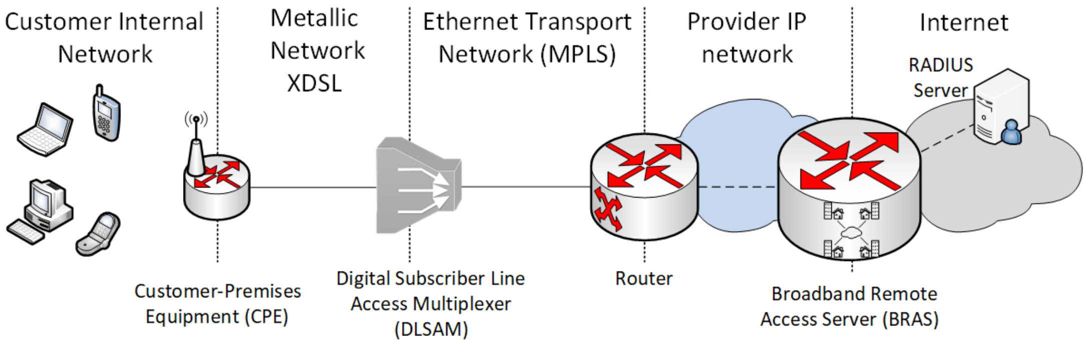

Customer Internal Network
Metallic Network XDSL
Ethernet Transport Network (MPLS)
Provider IP network
Internet
RADIUS Server
Customer-Premises Equipment (CPE)
Digital Subscriber Line Access Multiplexer (DLSAM)
Router
Broadband Remote Access Server (BRAS)

Figure 1. Typical topology of broadband access network. The connection between the Home Gateway and the Internet is authenticated, and registered by the Radius server. The traffic is encapsulated in Point-to-Point Protocol over Ethernet (PPPoE) sessions between the user's home and the Broadband Remote Access Server (BRAS). Traffic inspection and collection occurs after BRAS.

nodes that allow the inspection of the data in the users premises or in the perimeter closest to the users.

The analyzed traffic is composed of the aggregated traffic coming from the high capillarity, last mile, of different users with a wide variety of service profiles accessed by each user and generating a large data volume. Therefore, the problem of alert profile characterization is a complex Big Data problem [23], which requires appropriate processing tools. The main idea of this paper is to generate, analyze and characterize a dataset that represents as accurately as possible the residential fixed broadband user profile to train classifiers.

The analyzed dataset was created from the capture of raw packets containing real Internet Protocol (IP) traffic information of the residential users. Traffic was collected and recorded uninterruptedly for one week through the `tcpdump` [2] software. The process of collecting and writing the files did not use any packet filters and, therefore, all packets on the network were raw and recorded directly in the dataset. The physical collection structure has been configured by mirroring the aggregate traffic of one DSLAM to another port of the transport network metro Ethernet switch. The mirroring of the DSLAM port on the switch allows all traffic originated from or destined to the DSLAM to be cloned to a computer running an Ubuntu Linux operating system. To ensure high-speed storage and to allow easy data transport, the dataset was written to an external hard drive with a USB 3.0 interface. Figure 2 shows the basic topology and the assembled structure for data collection. It is worth to mention that analyzing all traffic from operator is out of scope, thus data consumption samples satisfy the needs for the proposed characterization.

The data capture procedure ensures no loss in port mirroring at 1 Gb/s. Thus, 100% of the traffic generated by the 373 customers was collected and recorded in the dataset, totaling 5 TB of information. Although the average available speed at each port of the DSLAM is approximately 12 Mb/s, generating a hypothetical aggregate traffic of more than 4 Gb/s, it was verified that during the entire capture process, aggregate real traffic did not exceed 800 Mb/s. Aggregate traffic is composed by round-trip, uplink and downlink traffic. It is worth noting that all the captured traffic comes from fixed broadband sessions.

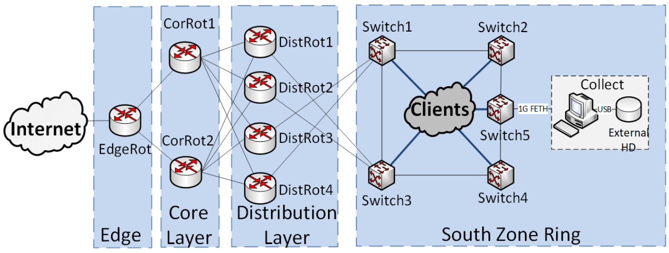


Figure 2. Topology of the data collection structure of the main DSLAM port with 373 broadband clients.

IV. The Data Analysis

The analysis of captured data from the telecommunications operator's network was divided into three stages. The first stage handles the raw data capture files through a network intrusion detection system (IDS) and then generates a summary of the data in the form of flows. We use the flow definition based on RFC 7011 [24]. A flow is defined as the set of packets collect during a window time in a monitoring point sharing common features. These features, include characteristics or information of the packets or functions employed in the packet or in transport or application header. In this paper, we abstract the flow in 43 features.

The second stage analyzes the distribution of the main characteristics of the dataset, highlighting differences between the normal traffic and the traffic that generates security alerts. Finally, the third step is to compare the use of classifiers to verify the accuracy in classifying normal and anomalous traffic. The classification process enables isolating potentially malicious traffic from all the traffic of the telecommunications operator and, then, forwarding only the portion of the traffic that can generate alerts to the IDS, which reduces the analyzed traffic load.

The first stage of data analysis was based on the extraction of the characteristics of the flows represented by the captured packets, as well as the verification of possible alerts through an IDS. Since the packets comes from residential clients

[2] Available at http://www.tcpdump.org.

with Asymmetric Digital Subscriber Line (ADSL) access, the captured traffic is encapsulated in Point-to-Point Protocol over Ethernet (PPPoE) sessions which make the analysis of packets harder for some IDS that do not perform the inspection of this type of protocol, such as Snort [25]. Therefore, in order to perform traffic classification on different types of alerts, the Suricata IDS[3], Version 3.2, was used with its most recent signature database.

The classification between normal traffic and alert was performed based on Suricata signatures since there was no previous knowledge about threat information. Because the data is real and hence untagged, it is not possible to ensure that all flows are legitimate or, even after IDS classification, that all alert flows are indeed malicious. However, the IDS classification is used as the reference for the proposed classification of this paper.

Parallel to the classification of the packets by the IDS, the captured packets were decapsulated from the PPPoE session using the tool `stripe`[4] and were summarized in flows through the `flowtbag`[5] tool. In addition, a Python application was developed to process the output of the Suricata IDS, and the summarized flow features in order to correlate which flow was reported as an alert by the IDS. Thus, it was possible to obtain a flow dataset with the respective class labels. As we remove the source and destination IP addresses from the flow feature set to ensure the data anonymization, the dataset presents 43 characteristics of each flow plus the class to which each flow belongs. The output class, characteristic 44, is given by the type of alert generated by the IDS or 0 in the case of a normal flow.

The second stage is to extract knowledge from the data. The characterization of the traffic is important for retrieving the parameters of systems models to evaluate the overall performance of the network [26]. For this purpose, the free and open source data analysis platform KNIME[6], version 3.3.1, was used. At first, data analysis focuses on Principal Component Analysis (PCA) in order to find out which characteristics carry most information. Thus, Figure 3 shows the six most important components of the dataset. We selected the six main components since they represent 99% of the dataset variance. The components were ordered based on its associated eingenvalue. The most relevant features for traffic characterization are: source port, destination port, total flow volume, quantity of packets in the stream, volume of the outbound and inbound sub-flows and volume of data in headers in the outbound and inbound flows. Using these characteristics, the behavior of normal traffic and alerts was analyzed.

The first analyzed features are the source and destination ports. Figure 4 shows the source and destination ports of the flows. The figure focuses on the 1024 first doors (from 0 to 1023), as they are the operating system restricted ports. Usually, these ports are used by daemons that execute services with system administrator privileges. The flow definition used considers the source port to be the port that initiates the TCP connection. Because the dataset portrays home users, it is expected that most connections will be destined to restricted and dynamics ports. Thus, it is remarked that the number of alerts coming from connections that the destination port is in the range of restricted ports is relatively low to the total number of connections on these ports, Figure 4(b). When considering the flows, in which the source port is in the range of restricted ports, almost all flows are labeled as alert by the IDS, shown in Figure 4(a). Another important fact is that most of the analyzed flows reflect the use of the DNS service (UDP 53) and HTTPS and HTTP services (TCP 443 and 80). The prevalence of HTTPS services over HTTP reflects the shift that major Internet content providers, such as Google and Facebook, have done to use encrypted service by default to ensure users privacy and security.

The relation between the most accessed services and flow duration is shown in Figures 5(a) and 5(b). The duration of analyzed flows is mostly less than 30 ms, characterizing the use of DNS, HTTP and HTTPS services. A good approximation for average flows duration is the Erlang distribution, with $k = 1$ and $\lambda = 3.7$. The approximation was calculated by matching the distribution to the data and minimizing the mean square error between the distribution and the obtained data. However, by the Kolmogorov-Smirnov test[7] the hypothesis that the data follow such a distribution should be rejected.

Regarding the protocols used, the prevalence of UDP flows is evident and refers to DNS queries. It is worth mentioning that the number of alerts generated by UDP flows is more than 10 times greater than the number of alerts generated by TCP flows. Another important point is that the number of flows that generate alerts is approximately 26% of total flows.

Figure 6 shows the characterization of the number of packets per flow, in both traffic directions. In both directions, communication occurs with up to 32 packets in 95% of cases and with up to 100 packets in 98.5%. The result shows that the connections in the residential scenario are mostly connections with few packets. Alert-generating flows generally have fewer packets than normal traffic flows, since 95% of the alert streams present up to 22 packets.

Considering the amount of data transferred in each flow, Figure 7 compares the round-trip flows in relation to volume in bytes. The disparity of the traffic volume in both directions of the communication is evident. While in one way 95% of traffic has a maximum volume of 6.4 kB, in the other way, the same traffic share presents up to 18 kB. The probability distribution that best fits the data volume, verifying the one that minimizes the mean square error, is the log-normal distribution with $\mu = 5.67$ and $\sigma = 31.68$. Validation of adequacy to log-normal distribution was performed using the Kolmogorov-Smirnov test on a random subsampling of the data to a statistical significance of 95%. This result demonstrates that the resi-

[3] Available at https://suricata-ids.org.

[4] Available at https://github.com/theclam/stripe.

[5] Available at https://github.com/DanielArndt/flowtbag.

[6] Available at https://www.knime.org/

[7] The Kolmogorov-Smirnov test checks whether one of the probability distributions differs from the hypothesized distribution based on a finite number of samples.

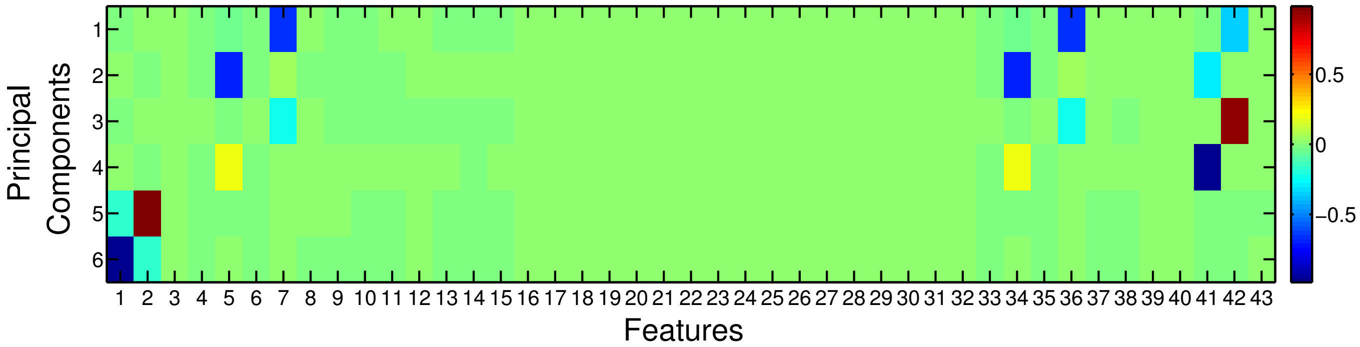
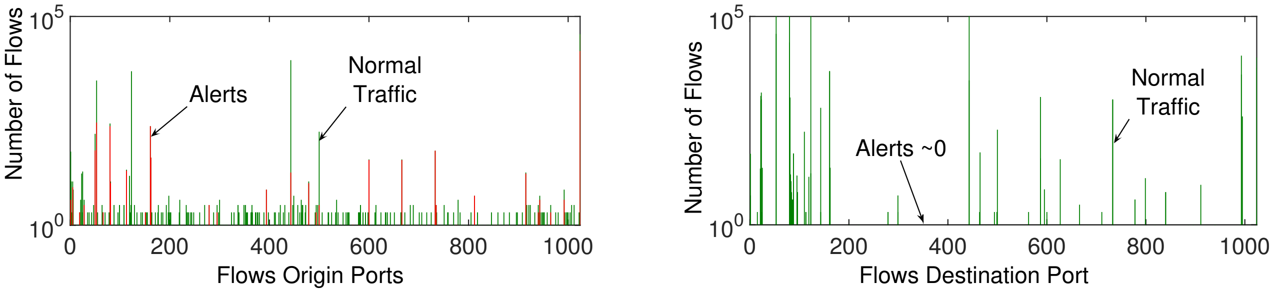


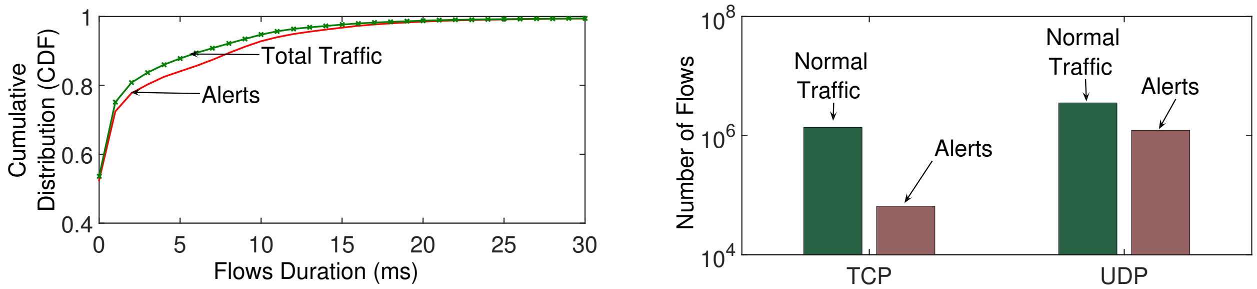
Figure 3. Visualization of the six components obtained with the Principal Component Analysis (PCA) on the dataset, in descending order of eigenvalue. The eight most relevant features are: source port (1), destination port (2), total flow volume (5), number of packets in the stream (7), volumes of the outgoing (34) and back flow (36), data volume in headers in the outbound flows (41) and back (42).



(a) Source Ports Distribution.

(b) Destination Ports Distribution.

Figure 4. Ports used in flows. Comparison of the use of the lowest 1024 ports (restricted ports) in the evaluated flows. Because they are home users, the largest number of flows originating from these ports are flows that generate alerts.



(a) Flow Duration.

(b) Transport Protocols Used.

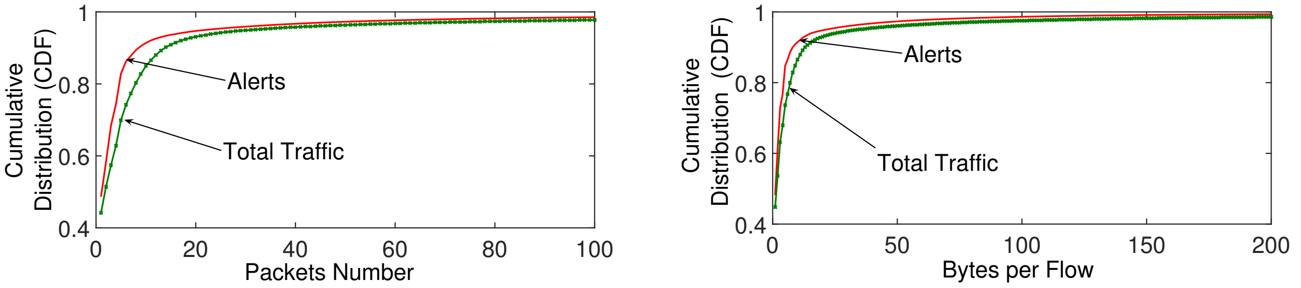
Figure 5. Cumulative Probability Density Function (CDF) for the distribution of the duration of flows in milliseconds and number of flows per transport protocols. A) The flows that generate alerts are shorter in duration than the average flow. B) The legitimate flows with UDP are numerous due to DNS (port 53 UDP). The number of alerts in UDP is more than 10 times greater than in TCP flows.

dential broadband user profile is a content consumer. Another interesting point is that the flows that generate alerts have a similar traffic volume profile in both directions. Asymmetric traffic is more typical of the legitimate users.

Figure 8 shows the behavior of the sub-flows generated in each connection. A sub-flow is considered a flow in one direction. This characteristic was pointed out by the PCA method as one of the most important characteristics to describe the dataset. However, the statistical behavior of the data volume of the sub-flows is the same as the total flow. This is because the flows are mostly of short duration, evidenced in Figure 5(a), and thus do not generate sub-flows. Data analysis showed that the flows do not pass to the idle state.

Another important feature is the total amount of data in the packet headers. Figure 9 shows that in both directions of the flows, both alert and total traffic have the same behavior. In particular, there is symmetry in the round-trip traffic in terms of the volume of data in the headers. It highlights that malicious traffic do not rely on the usage of header options;

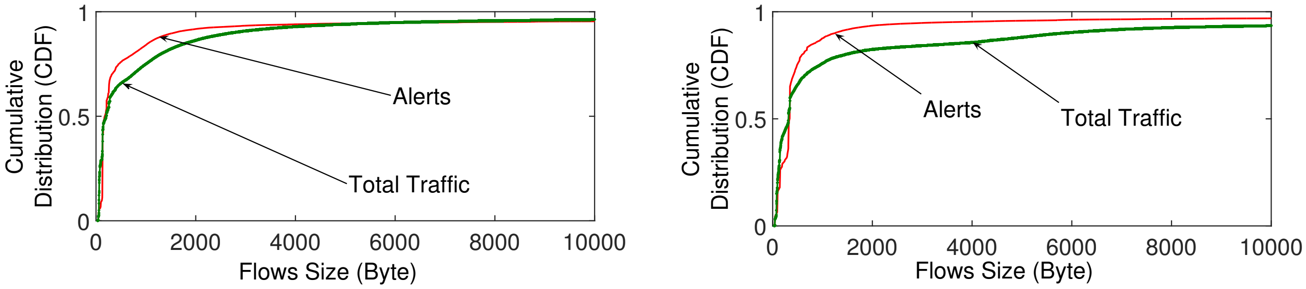
At the end of the knowledge extraction phase, the profile of the alerts generated by IDS was analyzed. Figure 10 shows which are the main classes of alerts triggered by the IDS.



(a) Flows in the Uplink Direction.

(b) Flows in the Downlink Direction.

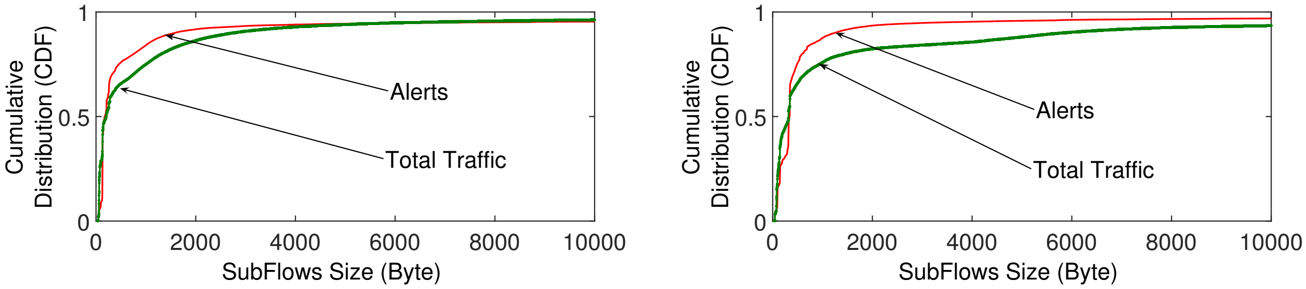
Figure 6. Cumulative Probability Density (CDF) function for the distribution of the number of packets per flow. Flows that generate alerts tend to have fewer packets.



(a) Flows in the Uplink Direction.

(b) Flows in the Downlink Direction.

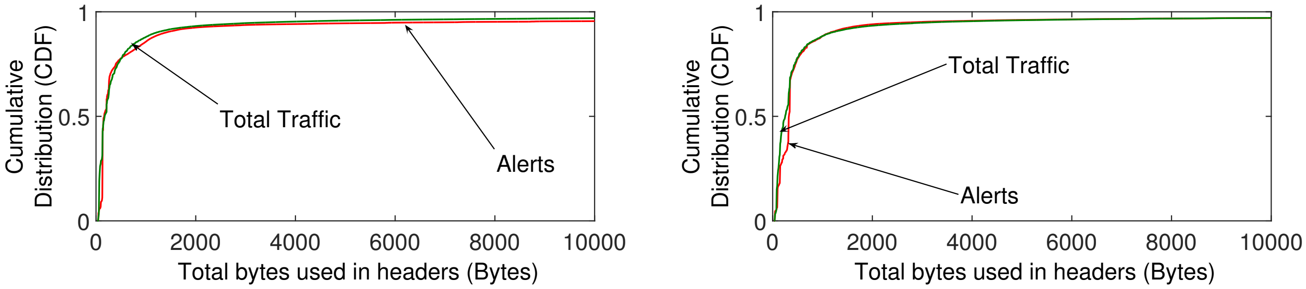
Figure 7. Cumulative Probability Density (CDF) function for volume distribution in bytes by flow. Flows that generated alerts tend to have smaller volumes in transferred bytes.



(a) Flows in the Uplink Direction.

(b) Flows in the Downlink Direction

Figure 8. Cumulative Probability Density (CDF) function for volume distribution in bytes by sub-flow in each flow. Flows that generate alerts tend to have smaller volumes in bytes that are transferred in sub-flows.



(a) Flows in the Uplink Direction.

(b) Flows in the Downlink Direction.

Figure 9. Cumulative Probability Density Function (CDF) for volume distribution in bytes of the data trafficked in headers. The behavior of traffic that generates alerts is very similar to total traffic.

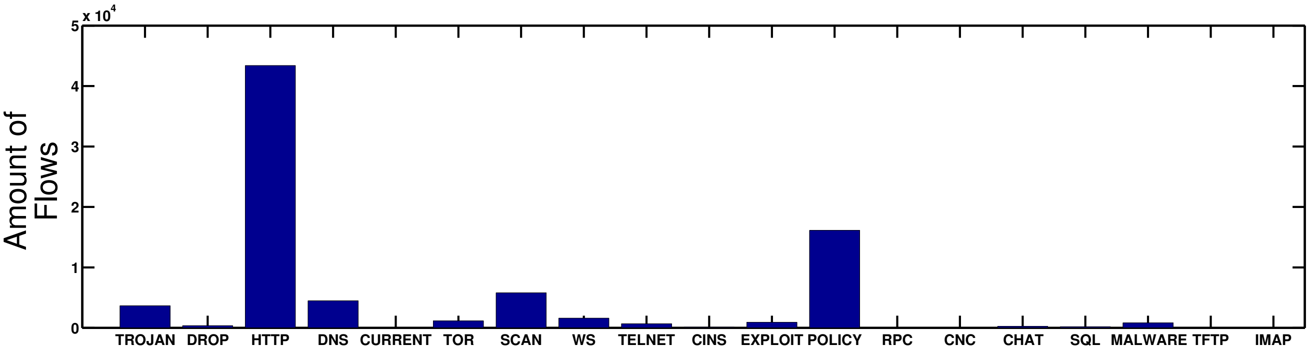


Figure 10. Distribution of the main types of alerts in the analyzed traffic.

Alerts for attacks against HTTP are the most frequent. This class of alerts includes SQL injection attacks through HTTP calls and XSS attacks (cross-site scripting). Home users can execute these attacks, as they use the parameters of HTTP calls to insert some malicious code into the servers and, therefore, are not filtered by access rules. Other important alerts are port scanning, vulnerability scanning, and execution of malicious applications (trojan and malware). The scans are generally intended to identify open ports and vulnerabilities in user premises (home gateway). Alerts for trojan and malware identify activities typical of known malicious applications that aim to create and exploit vulnerabilities in the devices of the home users. The other alerts refer to mechanisms of information theft and to Byzantine-attack signatures on common protocols, such as IMAP and Telnet [8].

The third stage of the data analysis is to design a classifier to differentiate flows into normal or anomalous classes. As the classification overhead is lower than the IDS processing time, the main goal is to carry out the classification of the flows with a considerable precision to optimize the traffic analysis performed by the telecommunications operator, since only the traffic classified as suspicious should be sent to an IDS service.

The first classifier approach considers a neural network of two hidden layers with 20 neurons in each layer, with six neurons in the input layer and one neuron in the output layer. The learning algorithm used was the MultiLayer Perceptron (MLP) with the weight adjustment through backward propagation algorithm. Modeling the neural network takes into account that the input is composed by the six principal components calculated by the PCA algorithm. The output of the network is the probability that the flow is marked as suspicious. A flow is considered suspicious when the probability is greater than 0.5 and normal otherwise. Table I shows the performance of the neural network in a cross-validation scenario of 10 rounds[9]. It is verified that the accuracy of the neural network was 0.847, with sensitivity of 0.625 in the alert class.

In a second approach, we evaluate the use of a Decision Tree-based classifier. The input of the classifier is also the six principal components extracted from the PCA. The output of the classifier is the tag in one of the two possible classes: alert or normal. Table II shows the classification results using the Decision Tree in a 10 folds cross-validation. The accuracy achieved by the classifier was 0.956. The sensitivity in the alert class was 0.934 and 0.976 in the normal class. This result shows that the use of this classifier as a pretreatment of the flows reduces the load in the telecommunications operator's traffic analyzer by up to 73%, with a sensitivity of 0.934 in the suspicious traffic.

Table I
Two Hidden Layers Neural Network Classification.

	Precision	Sens.	Spec.
Alert	0.671	0.625	0.891
Normal	0.870	0.891	0.625

Table II
Decision Tree Classification.

	Precision	Sens.	Spec.
Alert	0.933	0.934	0.976
Normal	0.976	0.976	0.934

V. Conclusion

The knowledge about the network usage profile is important to better scale the network and to identify its main services. The identification of the main security alerts in the network allows the design of possible countermeasures. In this paper, we presented the creation of a security dataset from a real telecommunication operator network located in the city of Rio de Janeiro, Brazil. The dataset represents the use of the fixed-line access service of 373 home users. The analysis of the data allows identifying that the main services accessed are those of DNS and web services. The flow profile is characterized by fast connections, up to 30 ms, with a data transfer of up to 18 kB in 95% of cases.

The obtained results through the application of a neural network and a decision tree demonstrate that the class-labeling applied to the dataset is consistent with the observable patterns.

[8] Mainly used for remote configuration of network equipment.

[9] TP:True Positive; FP: False Positive; TN: True Negative; FN: False Negative; Sens.: Sensitivity; Spec: Specificity.

Based on the characterization of normal usage and alerts, we propose the use of a decision tree based classifier to identify suspect flows and, then, to reduce the load on the traffic analyzer. The results show that using a simple classifier, we are able to reduce traffic sent to the traffic analyzer by up to 73%, with the ability to identify up to 93% of the flows that generate alerts on the network.

As future work, we intend to use this dataset in the validation of a flow-processing architecture for classification by decision trees, verifying the accuracy of real-time traffic analysis. In addition, the performance of other classification and training algorithms in real time will be evaluated.

Acknowledgment

This research is supported by INCT of the Future Internet, CNPq, CAPES, and FAPERJ.

References

[1] IBGE, *Síntese de indicadores sociais : uma análise das condições de vida da população brasileira*. Rio de Janeiro: IBGE, 2016, vol. 36.

[2] T. File and C. Ryan, "Computer and internet use in the united states: 2013," *United States Census Bureau*, no. P20-569, pp. 1–14, May 2013.

[3] S. Bhatt, P. K. Manadhata, and L. Zomlot, "The operational role of security information and event management systems," *IEEE Security Privacy*, vol. 12, no. 5, pp. 35–41, Sep. 2014.

[4] D. M. F. Mattos and O. C. M. B. Duarte, "AuthFlow: authentication and access control mechanism for software defined networking," *Annals of Telecommunications*, pp. 1–9, 2016.

[5] D. Puthal, S. Nepal, R. Ranjan, and J. Chen, "Threats to networking cloud and edge datacenters in the internet of things," *IEEE Cloud Computing*, vol. 3, no. 3, pp. 64–71, May 2016.

[6] J. Heidemann and C. Papadopoulos, "Uses and challenges for network datasets," in *Proceedings of the IEEE Cybersecurity Applications and Technologies Conference for Homeland Security (CATCH)*. Washington, DC, USA: IEEE, Mar. 2009, pp. 73–82.

[7] D. M. F. Mattos, O. C. M. B. Duarte, and G. Pujolle, "Reverse update: A consistent policy update scheme for software-defined networking," *IEEE Communications Letters*, vol. 20, no. 5, pp. 886–889, May 2016.

[8] CAIDA, "Supporting research and development of security technologies through network and security data collection," http://www.caida.org/funding/predict/, 2007, accessed 25-07-2017.

[9] T. H. D. Kotz and I. Abyzov, "CRAWDAD: A community resource for archiving wireless data at dartmouth," http://crawdad.cs.dartmouth.edu/, 2004, accessed 25-07-2017.

[10] NSL-KDD, "NSL-KDD data set for network-based intrusion detection systems," http://iscx.cs.unb.ca/NSL-KDD/, 2009, accessed 25-07-2017.

[11] A. Shiravi, H. Shiravi, M. Tavallaee, and A. A. Ghorbani, "Toward developing a systematic approach to generate benchmark datasets for intrusion detection," *Computers & Security*, vol. 31, no. 3, pp. 357 – 374, 2012.

[12] A. Lobato, M. Andreoni Lopez, and O. C. M. B. Duarte, "An accurate threat detection system through real-time stream processing," Grupo de Teleinformática e Automação (GTA), Universidade Federal do Rio de Janeiro (UFRJ), Tech. Rep. GTA-16-08, 2016.

[13] J. Song, H. Takakura, Y. Okabe, and K. Nakao, "Toward a more practical unsupervised anomaly detection system," *Inf. Sci.*, vol. 231, pp. 4–14, May 2013.

[14] Y.-Z. Chen, Z.-G. Huang, S. Xu, and Y.-C. Lai, "Spatiotemporal patterns and predictability of cyberattacks," *PLOS ONE*, vol. 10, no. 5, pp. 1–19, 05 2015.

[15] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the kdd cup 99 data set," in *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, July 2009, pp. 1–6.

[16] A. Shiravi, H. Shiravi, M. Tavallaee, and A. A. Ghorbani, "Toward developing a systematic approach to generate benchmark datasets for intrusion detection," *Comput. Secur.*, vol. 31, no. 3, pp. 357–374, May 2012.

[17] G. Draper-Gil, A. H. Lashkari, M. S. I. Mamun, and A. A. Ghorbani, "Characterization of encrypted and vpn traffic using time-related features," in *Proceedings of the 2nd International Conference on Information Systems Security and Privacy*, 2016, pp. 407–414.

[18] C. Pascoal, M. R. de Oliveira, R. Valadas, P. Filzmoser, P. Salvador, and A. Pacheco, "Robust feature selection and robust PCA for Internet traffic anomaly detection," in *2012 Proceedings IEEE INFOCOM*, Mar. 2012, pp. 1755–1763.

[19] N. Kato, Z. M. Fadlullah, B. Mao, F. Tang, O. Akashi, T. Inoue, and K. Mizutani, "The deep learning vision for heterogeneous network traffic control: Proposal, challenges, and future perspective," *IEEE Wireless Communications*, vol. PP, no. 99, pp. 2–9, 2017.

[20] L. Nie, D. Jiang, and Z. Lv, "Modeling network traffic for traffic matrix estimation and anomaly detection based on bayesian network in cloud computing networks," *Annals of Telecommunications*, pp. 1–9, 2016.

[21] K. Wu, K. Zhang, W. Fan, A. Edwards, and P. S. Yu, "RS-Forest: A rapid density estimator for streaming anomaly detection," in *2014 IEEE International Conference on Data Mining*, Dec. 2014, pp. 600–609.

[22] P. Clay, "A modern threat response framework," *Network Security*, vol. 2015, no. 4, pp. 5–10, Apr. 2015.

[23] M. D. Assunção, R. N. Calheiros, S. Bianchi, M. A. Netto, and R. Buyya, "Big data computing and clouds: Trends and future directions," *Journal of Parallel and Distributed Computing*, vol. 79, no. 1, pp. 3 – 15, 2015, special Issue on Scalable Systems for Big Data Management and Analytics.

[24] E. B. Claise, E. B. Trammell, and P. Aitken, "Specification of the IP Flow Information Export (IPFIX) Protocol for the Exchange of Flow Information," RFC 7011 (Informational), Internet Engineering Task Force, 2013 2013.

[25] M. Roesch *et al.*, "SNORT: Lightweight intrusion detection for networks." in *Lisa*, vol. 99, no. 1, 1999, pp. 229–238.

[26] D. M. F. Mattos, L. H. G. Ferraz, L. H. M. K. Costa, and O. C. M. B. Duarte, "Virtual network performance evaluation for future internet architectures," *Journal of Emerging Technologies in Web Intelligence*, vol. 4, no. 4, pp. 304–314, 2012.