Preliminary Results on Exploring Data Exhaust of Consumer Internet of Things Devices

Alexander Loginov
Faculty of Computer Science
Dalhousie University
Halifax, Canada
loginov@cs.dal.ca

Jeffrey Adjei
Faculty of Computer Science
Dalhousie University
Halifax, Canada
jeffrey.adjei@dal.ca

Nur Zincir-Heywood
Faculty of Computer Science
Dalhousie University
Halifax, Canada
zincir@cs.dal.ca

Srinivas Sampalli
Faculty of Computer Science
Dalhousie Univeristy
Halifax, Canada
srini@cs.dal.ca

Kevin de Snayer
Calian Group Ltd.
Canada
kevin.desnayer@calian.com

Terri Dougall
Calian Group Ltd.
Canada
t.dougall@calian.com

***Abstract*—In this paper, we apply a machine learning classifier to the publicly available consumer Internet of Things (IoT) traffic traces to explore the nature and extent of any potential data exhaust. To this end, we propose two feature sets and compare them against the baseline flow feature set and the results from the previous works. Evaluations show the improvement in performance obtained using the proposed feature sets and the variety of information that can be extracted from the captured IoT traffic regardless of encryption.**

***Index Terms*—IoT security, data exhaust, network traffic.**

I. Introduction

Internet of Things (IoT) technology makes our daily lives more convenient and intelligent with automatically functioning devices controlled through physical or cyber interactions. In general, IoT devices can be grouped into two categories, namely, cyber-physical system IoTs and consumer IoTs. Cyber-physical system IoT devices offer services such as power utilities, manufacturing plants, or factory automation. Consumer IoT devices offer services such as personal digital assistants, home security or climate control. Despite the positive prospects for the spread of IoT technologies, a major problem is the nature of the technology that gives rise to new concerns about privacy and security. These devices have the potential to gather information about their users and their surrounding environments by combining sensor information from cameras, microphones, motion sensors and their Internet connectivity. Much of this information could have major security as well as privacy implications. For example, when devices surreptitiously record audio and share this information over the Internet with device manufacturers and/or unknown third parties, this can result in not only privacy concerns but also in security problems. This type of system/device behaviour is referred as data exhaust, i.e. data generated as trails or information by-products resulting from digital/online activities. Most of the Consumer IoT devices lack any functionality/property that could indicate data exhaust for the users to protect themselves. In this work, we apply a Machine Learning (ML) based approach to analyze data exhaust using the metadata of the publicly available consumer IoT device traffic datasets [1]–[4]. This enables us to explore the nature and extent of potential data exhaust of these IoT devices. To this end, we explore the following research questions:

- **Q1:** To what extent consumer IoT devices and non-IoT devices can be separated in the captured traffic?
- **Q2:** Can we infer the category of a device based on the captured traffic?
- **Q3:** Can we infer the type of a device based on the captured traffic?
- **Q4:** Can we detect different voices based on the captured traffic if the device is voice activated?
- **Q5.** Can we detect different activities (interactions) based on the captured traffic if the device is voice activated?

The rest of the paper is organized as follows. Section II summarizes the related work. Section III introduces the proposed feature set, datasets used and the methodology followed. Section IV details the evaluations, results and comparisons to the related work. Finally, conclusions and the future work are discussed in Section V.

II. Related Work

There are several key challenges that limit our understanding of data exhaust from consumer IoT devices and their security and privacy implications. In general, ground truth about data trails and by-products in the IoT device ecosystem are not readily available. For the vast majority of IoT devices, it is not feasible to modify the device firmware, or employ proxy techniques to identify data exhaust that might be leaking. Previous works characterize IoT device traffic from different perspectives. Some works focus on whether encryption is used, and if so, whether it is misused [1]. Others analyze traffic from many different IoT devices to identify encrypted and/or unencrypted traffic and their vulnerability to different types of attacks [2]. Several works investigate the communication channels between IoT devices and their cloud services to

characterize their traffic [4]. Others study encryption and authentication protocol weaknesses of these devices [REF]. These studies cover data exhaust for different categories of IoT devices such as medical devices, office and home automation solutions. Additionally, several works in the literature focus on intrusion detection systems using machine learning [5] while others focus on a policy enforcement approaches for detecting malicious behaviours [6]. Some research addresses the problem of automatically verifying and enforcing the compliance of a given IoT device according to its specification [7]. Other approaches use statistical techniques on the device traffic to profile users and their activities [3]. In this work, our goal is to explore consumer IoT devices, and their interactions over the Internet. In doing so, we aim to study the nature and extent of any potential data exhaust.

III. Methodology

In this work, the methodology follows the design and implementation of a network traffic analysis pipeline including: (i) selecting and extracting features from traffic traces; (ii) training and testing the classifier using those features; and (iii) analyzing the ML model solutions and their performance. To achieve this, we employ the Random Forest classifier since it has been reported as the best performing classifier in [1], [2]. And, we employ four publicly available datasets, namely NEU-SNS [1], CIC-IoT [2], UPC-IoT [3] and ISOT-CID [4], since these were the most recent publicly available datasets reported in the literature in this field. These not only ensure the replicability of our research but also enable us to explore the research questions introduced in Section-I. Out of these four datasets, the first three include IoT device network traffic, whereas ISOT-CID dataset does not include IoT traffic. This allows us to investigate the similarities and differences between IoT and non-IoT network traffic in our evaluations.

A. Datasets

The NEU-SNS dataset consists of traffic from 81 IoT devices (54 unique types of devices) located in the UK and USA IoT labs (testbed). The dataset includes IoT devices from the following six categories [1]: Cameras (14 devices), Smart hubs (7 devices), Home Automation (10 devices), TV (5 devices), Audio (7 devices), and Appliances (11 devices).

The CIC-IoT dataset consists of malware and benign traffic of IoT devices [2]. This dataset includes some different (such as Door Lock brands) and some similar (such as camera brands) IoT devices compared to the NEU-SNS dataset. The UPC-IoT dataset employed in this work only includes Amazon Alexa traffic traces. It contains 300'000 raw PCAP traces with all the communications between the Amazon Echo device and Amazon Alexa servers with 100 different voice commands repeated 500 times in two different languages [3].

On the other hand, the ISOT-CID dataset is collected from a cloud environment and includes more than 2.5 terabytes of traffic traces, such as normal activities and a variety of attacks. These include web traffic generated by more than 160 legitimate users, traffic generated by 100 robots, performing tasks such as account registration, reading/posting and commenting on blogs as well as browsing various pages [4].

B. Proposed feature selection and extraction

To study the research questions stated in Section-I, we used Python 3.9.5 programming language and the Random Forest (RF) classifier with the newly developed feature sets. All features used to train the RF classifier are generated from the flow statistics gathered from packet sizes and timestamps. Our preliminary research has shown that the flow statistics gathered from packet sizes in bytes and timestamps have different fingerprints for the types and interaction of IoT devices. Our observations indicate that while 'data.len' related information is more helpful for research questions Q1 to Q3, the distributions of bytes send/received over the flow duration seems to be preferable for Q4 and Q5. Based on the above, we propose two feature sets, namely feature set 1 (FS1) for Q1, Q2 & Q3, and feature set 2 (FS2) for Q4 & Q5. For this purpose, the features extracted from the network traffic traces are shown in Table I. These fields are used to generate flows.

TABLE I: IP Fields extracted

Name	Description
frame.time_epoch	Epoch Time
frame.time_delta	Time delta from previous captured frame (TD)
frame.protocols	Protocols in the frame
frame.len	Frame length on the wire (FL), seconds
ip.src	IP Source Address
ip.dst	IP Destination Address
tcp.srcport	TCP Source Port
tcp.dstport	TCP Destination Port
udp.srcport	UDP Source Port
udp.dstport	UDP Destination Port
data.len	Data Length (DL), bytes

The flows are then aggregated by the following fields (flow keys) [10]: 'frame.protocols', 'ip.src', 'ip.dst', 'tcp.srcport', 'tcp.dstport', 'udp.srcport' and 'udp.dstport'. The flow aggregation algorithm supports bi-directional flows (client-server and server-client sub-flows) [8] and generates two sets of features (time and size frames statistics within the flow) - FS1 and FS2 shown in Tables II and III respectively. We also used the 'Tranalyzer2' flow generator and packet analyzer [9] to generate the third feature set, shown in Table IV. Hereafter, this will be referred as the baseline to evaluate the effect of proposed feature sets based on aggregated flows.

IV. Evaluations and Results

In the following evaluations, the RF classifiers are trained with the default parameters on the training partitions and the trained models are tested on the test partitions of the respective datasets (Section III-A). The following details the proposed FS1 and FS2 in combination with the datasets used:

- FS1 and Tranalyzer2 feature sets are generated from the NEU-SNS and ISOT-CID datasets. These datasets are labelled as IoT and non-IoT, respectively. They are used to study Q1
- FS1 and Tranalyzer2 feature sets are generated from NEU-SNS dataset. In this case, the dataset is labelled

TABLE II: Proposed Feature Set - FS1

#	Description
1	Different protocols in the flow[a]
2	Number of different protocols (depth)
3	Flow duration
4	Min TD value in the flow
5	Mean TD value in the flow
6	Median TD value in the flow
7	Max TD value in the flow
8	TD variance in the flow
9	# of frames in the flow
10	Min FL value in the flow
11	Mean FL value in the flow
12	Median FL value in the flow
13	Max FL value in the flow
14	TD variance in the flow
15–18	FL percentiles (20, 40, 60, 80)
19	Mean DL value in the flow
20	Median DL value in the flow
21	# of frames in the client-server sub-flow
22	# of frames in the server-client sub-flow
23	Total DL in the client-server sub-flow
24	Total DL in the server-client sub-flow

[a]Categorical feature encoded with 'OdrinalEncoder'.

TABLE III: Proposed Feature Set - FS2

#	Description
1	Protocols in the flow[a]
2	Flow duration[b]
3	Min TD value in the flow[b]
4	Mean TD value in the flow[b]
5	Median TD value in the flow[b]
6	Max TD value in the flow[b]
7	TD variance in the flow[b]
8	# of frames in the flow[b]
9	Min DL value in the flow[b]
10	Mean DL value in the flow[b]
11	Median DL value in the flow[b]
12	Max DL value in the flow[b]
13	DL variance in the flow[b]
14–23	Bytes send/received distribution over the flow duration[b]

[a]Categorical feature encoded with 'OdrinalEncoder'.
[b]Features 2-23 are calculated for the flow and client-server and server-client sub-flows resulting in the total number of features equal to 67.

according to IoT device types to study Q2.

- FS1 feature set is generated by the Sompy Door Lock from the CIC-IoT dataset which is labelled as 'IoT.'

- FS2 feature sets are generated from the UPC-IoT English and Spanish datasets (respectively) and labelled as 'voice' for studying three voices and 'interaction' for studying 100 interactions available in the datasets.

It should be noted here that we used stratified sampling to randomly split each of the above datasets into the training (70%) and test (30%) partitions. Moreover, we used the following metrics to evaluate the performance of the models: precision (1), recall (2), and F1-score (3).

TABLE IV: Feature Set - Tranalyzer2

#	Name	Description
1	dir	Flow direction[a]
2	duration	Flow duration
3	numHdrDesc	Number of different headers descriptions
4	numHdrs	Number of headers (depth) in hdrDesc
5	hdrDesc	Headers description[a]
6	l4Proto	Layer 4 protocol
7	numPktsSnt	Number of transmitted packets
8	numPktsRcvd	Number of received packets
9	numBytesSnt	Number of transmitted bytes
10	numBytesRcvd	Number of received bytes
11	minPktSz	Minimum layer 3 packet size
12	maxPktSz	Maximum layer 3 packet size
13	avePktSize	Average layer 3 packet size
14	stdPktSize	Standard deviation layer 3 packet size
15	minIAT	Minimum IAT
16	maxIAT	Maximum IAT
17	aveIAT	Average IAT
19	stdIAT	Standard deviation IAT
20	pktps	Sent packets per second
21	bytps	Sent bytes per second
22	pktAsm	Packet stream asymmetry
23	bytAsm	Byte stream asymmetry

[a]Categorical feature encoded with 'OdrinalEncoder'.

$$precision = \frac{TP}{TP + FP} \tag{1}$$

$$recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{3}$$

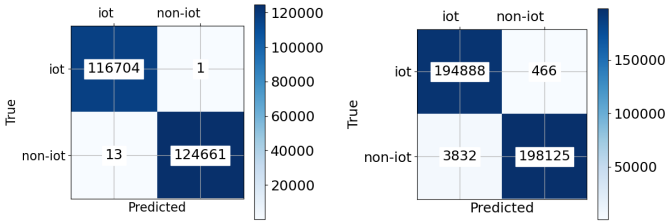
where: **TP** (true positives) are positive instances, labelled as positive. **TN** (true negative) are negative instances, labelled as negative. **FP** (false positive) are negative instances, labelled as positive. **FN** (false negative) are positive instances, labelled as negative. **D3** - FS1 feature set generated by the Sompy Door Lock from the IoT-23 dataset and labelled 'IoT.'

A. Exploring - Q1

The following evaluation is intended to show to what extent IoT and non-IoT traffic can be separated (binary classification). In this evaluation FS1 feature set is compared to the baseline feature set to measure the impact of aggregating flows. Figure 1 and Table 5 show the results. The results show that proposed FS1 features sets based on aggregated flow performs better than the baseline.

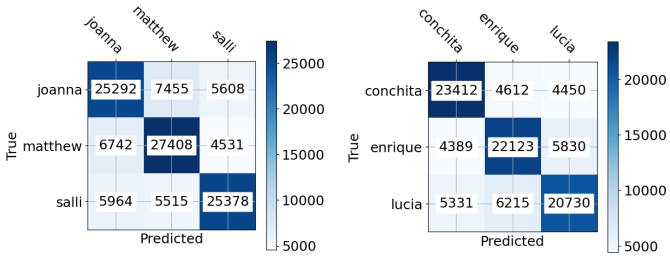
TABLE V: Precision, Recall and F1 scores. Experiment E1

Metric	FS1		Tranalyzer2	
	IoT	non-IoT	IoT	non-IoT
precision, %	99.99	100	98.07	99.77
recall, %	100.00	99.99	99.76	98.1
F1 score, %	99.99	99.99	98.91	98.93



(a) FS1 feature set. (b) Tranalyzer2 feature set.

Fig. 1: Exploring Q1 - Confusion matrices.



(a) English voices. (b) Spanish voices.

Fig. 2: Confusion matrices. Experiment E4.

B. Exploring - Q2 and Q3

In this case, the goal is to study how well a device's category could be inferred from the captured traffic. In this evaluation, FS1 feature set is compared with the baseline feature set. Moreover, to compare the number of inferable IoT devices with the results in [1], we assumed the same conditions: the device is inferable when its F1 score is greater than 75%. All IoT devices are grouped in the same six categories [1]. The results are shown in Tables VI to VIII.

TABLE VI: #Devices (per category) using FS1, F1>75%.

Category					**VPN**			
	US	**UK**	**US∩**	**UK∩**	**US**	**UK**	**US∩**	**UK∩**
Appliances	8 (9)	4 (4)	1 (2)	2 (2)	7 (9)	4 (4)	2 (2)	2 (2)
Audio	4 (5)	5 (6)	3 (4)	3 (4)	5 (5)	5 (6)	4 (4)	3 (4)
Cameras	11 (11)	8 (8)	5 (5)	5 (5)	11 (11)	8 (8)	5 (5)	5 (5)
Home Auto	9 (9)	6 (6)	5 (5)	5 (5)	9 (9)	6 (6)	5 (5)	5 (5)
Smart Hubs	7 (7)	6 (6)	6 (6)	6 (6)	7 (7)	6 (6)	6 (6)	6 (6)
TV	5 (5)	4 (4)	4 (4)	4 (4)	5 (5)	3 (4)	4 (4)	3 (4)

The total number of devices is shown in parentheses.

TABLE VII: #Devices (per category) Baseline, F1>75%.

Category					**VPN**			
	US	**UK**	**US∩**	**UK∩**	**US**	**UK**	**US∩**	**UK∩**
Appliances	2 (9)	1 (4)	0 (2)	0 (2)	2 (9)	2 (4)	0 (2)	1 (2)
Audio	2 (5)	2 (6)	1 (4)	1 (4)	3 (5)	1 (6)	2 (4)	0 (4)
Cameras	10 (11)	7 (8)	4 (5)	4 (5)	11 (11)	7 (6)	5 (5)	5 (5)
Home Auto	8 (9)	5 (6)	5 (5)	5 (5)	8 (9)	4 (6)	5 (5)	4 (5)
Smart Hubs	5 (7)	5 (6)	4 (6)	5 (6)	4 (7)	5 (6)	4 (6)	5 (6)
TV	4 (5)	3 (4)	3 (4)	3 (4)	4 (5)	3 (4)	3 (4)	3 (4)

The total number of devices is shown in parentheses.

TABLE VIII: Over all: #Devices, F1>75%.

Feature Set		**VPN**
FS1	126 (132) or 95.4%	126 (132) or 95.4%
Tranalyzer2	89 (132) or 67.4%	91 (132) or 68.9%

The total number of devices is shown in parentheses.

The results obtained with the FS1 feature set outperforms the results obtained with the baseline feature set as well as the results given in [1]. Moreover, we used FS1 to study Q3 (device classification per type). Results show that 46 out of 54 (85%) unique types of devices could be detected correctly (F1 score >75%). The full list of all available IoT devices can be found in Table 1 in [1]. The error analysis indicates that most errors are caused by mislabeling the devices within the three groups:

- Google Home and Home Mini.
- Amazon Echo Dot, Echo Plus and Echo Spot
- Samsung Washer and Dryer.

We assume that the devices from the above groups use similar firmware, which affects the classification results. After re-labelling devices within the groups as 'Google Home,' 'Amazon Echo' and 'Samsung Washer/Drier', respectively, we were able to correctly detect (F1 score >75%) 48 out of 50 in other words 96% of the unique types of IoT devices. These results are better than the results given in [1] as well as [5], where an average of 87.3% is given for inferring device types using RF classifier over 20 IoT devices.

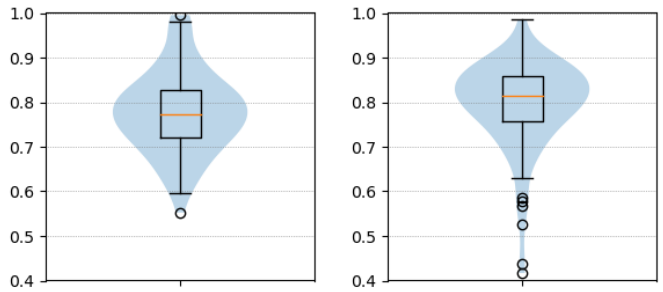
Furthermore, we utilized the RF model trained for Q1 and tested it on 130 FS1 flows generated from the 'Sompy Door Lock' traffic using the CIC-IoT dataset. This device was not available during the training of the model used here. In this case, all 130 flows are labelled correctly reaching 100% F1 score. This seems to show that the trained model with the proposed feature set could detect new IoT devices. Further research is necessary to evaluate other new devices.

C. Exploring - Q4

The goal of this evaluation is to study how well different voices used to interact (issue a voice command or ask a question) with a Smart Voice Assistant (Amazon Alexa) could be detected in the captured traffic. The following experiments are conducted using the UPC-IoT datasets (Section III-A). Table IX summarizes the results, and Figure 2 shows the confusion matrices for English and Spanish Voices.

TABLE IX: Voice Classification Results.

Voice	**precision, %**	**recall, %**	**F1 score, %**
Salli	71.45	68.86	70.13
Matthew	67.88	70.86	69.34
Joanna	66.56	65.94	66.25
Conchita	70.66	72.09	71.37
Enrique	67.14	68.4	67.77
Lucia	66.85	64.23	66.51



(a) English voices. (b) Spanish voices.

Fig. 3: Q5 - F1 scores distributions.

In these evaluations, it is possible to achieve F1 scores of 66% and 71%. On one hand these scores are better than random guessing. On the other hand, they are high performances in the case of ternary classification. However, they demonstrate that potential adversaries could extract information about the number of voices, i.e. persons, in a household or office environment using such consumer IoT devices. These results indicate that this is possible using only the available metadata present in the captured network traffic trace even if the traffic is encrypted as in the datasets used in this study.

D. Exploring - Q5

In Q5, the goal is to study how well English and Spanish voice interactions with Amazon Alexa can be detected from the captured traffic. We also analyze whether the importance of the features changes depending on the language of interaction. The complete lists of available interactions (commands) can be found in [3]. Table X summarizes the results and Figure 3 shows the F1 score distributions for English and Spanish voice interactions. These results show that different voice interactions with Amazon Alexa could be reliably detected in the captured encrypted Alexa traffic traces regardless of the language of interaction used.

TABLE X: Q5 - Voice interaction classification Results.

Metric	English voices			Spanish voices		
	precision, %	recall, %	F1 score, %	precision, %	recall, %	F1 score, %
Mean	78.0	77.7	77.7	80.0	79.6	79.5
Std dev.	8.0	11.0	9.1	7.2	12.6	9.8
Min	62.8	45.3	55.3	57.1	30.8	41.7
25%	72.4	71.7	72.1	75.4	75.9	75.7
50%	77.0	78.5	77.3	80.8	83.0	81.3
75%	82.3	64.1	82.8	85.0	87.6	85.8
Max	99.5	99.6	99.5	99.2	97.8	98.5
F1 >75%	63 (100)[a]			77 (100)[a]		

[a]The total number of unique interactions is shown in parentheses.

V. Conclusion and Future Work

We applied a RF classifier to explore data exhaust of consumer IoT devices. We used the metadata of the publicly available traffic traces of IoT devices. We proposed two different feature sets: (i) FS1 based on packet time and size statistics, and (ii) FS2 based on the distributions of time and size information exchanged over aggregated flows. We studied five research questions, Q1 - Q5, and compared the results to the baseline flow feature set and to the related works using the same traces. The evaluations for Q1 to Q3 show that IoT and non-IoT traffic can be separated using RF with a 99.99% F1 score. Moreover, the trained model can also infer the category and type of an IoT device with a high accuracy. It can also reliably detect the traffic of IoT devices not seen during the training. Further evaluations for Q4 and Q5 reveal that the number of users (voices used to interact) of Amazon Alexa Smart Home Assistant in the household/office can be classified. Also, the different user-device voice interactions can be reliably detected with an F1 score reaching to 99%. Future research will explore other non-IoT and IoT traffic traces as well as analyze other types of new IoT devices.

Acknowledgement

This research is supported by the Mitacs and Calian Group funding program. The research is conducted as part of the Dalhousie NIMS Lab[1].

References

[1] J. Ren, D. Dubois, D. Choffnes, A.M. Mandalari, R. Kolcun, H. Haddadi, "Information Exposure for Consumer IoT Devices: A Multidimensional, Network-Informed Measurement Approach," Proc. of the Internet Measurement Conference (IMC), 2019.

[2] S. Garcia, A. Parmisano, M.J. Erquiaga, "IoT-23: A labeled dataset with malicious and benign IoT network traffic (Version 1.0.0) [Data set]," Zenodo, 2020 http://doi.org/10.5281/zenodo.4743746.

[3] R. Barceló-Armada, I. Castell-Uroz, P. Barlet-Ros, "Amazon Alexa traffic traces," Computer Networks, vol. 202, 2022.

[4] A. Aldribi, I. Traore, P.G. Quinan, O. Nwamuo, "Documentation for the ISOT Cloud Intrusion Detection Dataset," Technical Report # ECE-2020-10-10, University of Victoria, ECE Department.

[5] S. Dadkhah, H. Mahdikhani, P. K. Danso, A. Zohourian, K. A. Truong and A. A. Ghorbani, "Towards the Development of a Realistic Multidimensional IoT Profiling Dataset," 2022 19th Annual International Conference on Privacy, Security & Trust (PST), Fredericton, NB, Canada, 2022, pp. 1-11, doi: 10.1109/PST55820.2022.9851966.

[6] I. Hafeez, M. Antikainen, A. Y. Ding and S. Tarkoma, "IoT-KEEPER: Detecting Malicious IoT Network Activity Using Online Traffic Analysis at the Edge," in IEEE Transactions on Network and Service Management, vol. 17, no. 1, pp. 45-59, March 2020, doi: 10.1109/TNSM.2020.2966951.

[7] A. Hamza, D. Ranathunga, H. H. Gharakheili, M. Roughan, and V. Sivaraman, "Clear as MUD: Generating, Validating and Applying IoT Behavioral Profiles," 2018 Workshop on IoT Security and Privacy. ACM, New York, NY, USA, 8–14. https://doi.org/10.1145/3229565.3229566.

[8] S. Burschka, B. Dupasquier, "Tranalyzer: Versatile high performance network traffic analyser," 2016 IEEE Symposium Series on Computational Intelligence (SSCI), 2016.

[9] Tranalyzer2 Tarantula version 0.8.14lmw1, Retrieved October, 2022 from https://tranalyzer.com.

[10] B. Trammell, A. Wagner, B. Claise, "RFC7015, Flow Aggregation for the IP Flow Information Export (IPFIX) Protocol," Proposed standard proposal, https://www.rfc-editor.org/rfc/rfc7015, 2013.

[1]https://projects.cs.dal.ca/projectx/