PhD in Information Technology and Electrical Engineering

Università degli Studi di Napoli Federico II

PhD Student: Giampaolo Bovenzi

XXXIV Cycle

Training and Research Activities Report – Second Year

Tutor: Antonio Pescapé

Information

I'm **Giampaolo Bovenzi** and I achieved MS degree (summa cum laude) at the same University in October 2018. Currently, I am a PhD Student attending the XXXIV Cycle of the Information Technology and Electrical Engineering (ITEE) at the Department of Electrical Engineering and Information Technology (DIETI) of the University of Napoli Federico II. My tutor is the prof. **Antonio Pescapé** and I am a member of Traffic research group whose activities are carried out in the field of Computer Networks.

Study and Training Activities

During the second year of PhD program, I attended the seminars reported in the following.

Seminars

1. *Virtualization technologies and their applications (lesson 1 and 2)*, Dott. Luigi De Simone, 06-07/04/2020, Credits 0.4.
2. *How to Publish Open Access with IEEE to Increase the Exposure and Impact of Your Research*, IEEE Xplore Webinar, 23/09/2020, Credits 0.2.

Credits Summary

	Credits year 1	Credits year 2								Credits year 3	
	Summary	Estimated	1 bimonth	2 bimonth	3 bimonth	4 bimonth	5 bimonth	6 bimonth	Summary	Estimated	Total
Modules	**23.6**	**10**	0	0	0	0	0	0	**0**	**7**	**30.6**
Seminars	**5.1**	**5**	0	0	0.4	0	0	0.2	**0.6**	**5**	**10.7**
Research	**31.3**	**45**	10	10	9.6	10	10	9.8	**59.4**	**48**	**138.7**
Total	**60**	**60**	10	10	10	10	10	10	**60**	**60**	**180**

Research activity

Techniques for (encrypted) Traffic Characterization, Classification, and Prediction

In my PhD, I am working in the context of monitoring and management of computer networks, with specific focus on (encrypted) traffic characterization, classification, and prediction. This research activities want to enhance fine-grain knowledge of Internet traffic facing nowadays complexity, e.g., *growth of peripheral networks* and related *traffic volume*, *heterogeneity* of Internet connected devices, *increment of security flaws*.

During my second year of PhD, I've worked on stochastic, machine learning and deep learning-based traffic characterization, classification, and prediction techniques, that are effective also in presence of encrypted traffic and preserve privacy of users, not requiring payload content inspection, but leveraging statistical features extracted from the Internet traffic.

I am applying these techniques to the detection of network threats (i.e. botnets), which broke privacy of users and affect both critical Internet infrastructure (e.g. cloud) and home networks (e.g. IoT), and to mobile application internet traffic at packet-level granularity to overcome recent contributions which are limited to traffic prediction at an aggregate scale [1, 2].

Research Description

Internet Traffic Characterization is of the utmost importance to understand traffic peculiarities, predict its characteristics, enforce traffic engineering, perform network planning and provisioning, manage the QoS, profile user activities, identify anomalies, emulate real traffic for testing purposes, etc. However, this process is challenged by the nature of the traffic traversing today's networks that is impacted by the way users behave, interact, and access the network. In fact, operators have experienced in the last years tremendous growth of the traffic to be managed in their networks, mostly generated by mobile devices [3], and understanding the nature of the communications flowing through the Internet is essential for operators to properly manage the network [4]. Traffic Classification (TC) (i.e. the association of traffic aggregates with specific applications or groups of applications) and Traffic Prediction (TP) (i.e. the forecast of future behavior of traffic aggregates generated by specific applications or groups of applications) are enabled by the effective characterization of the network traffic under analysis and can be seen as two interacting components.

According to the latest Ericsson mobility report [5], between Q3 2018 and Q3 2019, mobile data traffic has grown 68%, being fueled by both the rising number of smartphone subscriptions and the increasing average data volume per subscription. Overall, it is forecasted that mobile subscriptions will reach 8.9 billion by 2025, corresponding to a mobile data traffic of 160 exabytes per month against the 38 exabytes of 2019. Moreover, capillary diffusion of Internet-enabled devices (e.g., IoT) and the growth of network link capacity and coverage introduce new challenges in implementing effective and efficient TC, calling for the design and the deployment of highly-scalable architectures, to permit a feasible (time-constrained) fine-grained analysis of huge amounts of heterogeneous network traffic. Just think that the number of Internet of Things (IoT) devices has been estimated at about 7B in 2018, with a 3x expected growth by 2025, considering both consumer and industrial applications [6]. Unluckily, beyond the scalability issues, IoT devices are often characterized by a low-cost production process, including hardware and software design choices (e.g., insecure network services and interfaces, unsafe update mechanism, outdated components), resulting in huge security concerns and high exposure to several vulnerabilities [7].

These phenomenon underline the even growing complexity of nowadays Internet, and they exacerbate the need for accurate characterization, modeling, classification, and predictability of network traffic at fine grain, but the implementation of effective approaches for fine-grained traffic inspection must overcome several challenges. The broad adoption of encrypted protocols, e.g. Transport Layer Security (TLS), blocks the road to modeling approaches based on packet inspection, as encrypted network traffic represents the majority of mobile traffic (80% of all Android apps, and 90% of apps targeting Android 9or higher). Thus, modeling approaches cannot rely on clear-text patterns to identify application fingerprints or a specific execution state [8], or even reconstruct application message boundaries. Also, mobile traffic is an extremely complex and dynamic phenomenon [9]. Indeed, generated traffic can show wildly different and complex fingerprints, due to the multifold nature of both tasks carried on by means of mobile terminals and user activities within the same app, besides potential device/OS/app-version diversity [9, 10].

Because of the nature of nowadays encrypted traffic, Machine Learning (ML) models represent the natural enabler, able to provide decisions based on the sole traffic flow features [11] and to overcome shortcomings due to the application of common solutions (e.g., those based on payload inspection or earlier port-based ones [12]). Leveraging Machine Learning approaches to accomplish this task also fulfils the requirement of preserving privacy by performing the classification solely based on the statistical features of the encrypted traffic, without decrypting its content.

The applied solutions to face the scalability and performance problem (e.g., retrain) rely on divide-et-impera approach to TC. The adoption of hierarchical approaches allows for performance improvement, by splitting the TC task in sub-problems. These results in a possible training complexity increase which can be managed by model parallelism, due to its scalability and modularity. Accordingly, the appeal of hierarchical ML-based TC has recently attracted the interest of the scientific community [13-15].

In particular, a hierarchical solution is proposed in **[C1]** to cope with malicious traffic generated by IoT botnets. The solution is a Network intrusion Detection System based on ML and DL one-/multi-class classifiers. The NIDS is hybrid, characterized by a first stage that performs anomaly detection, and by a second stage that conducts open-set attack classification. Both stages end in a threshold-controlled censoring system, that enables a fast path for benign traffic at stage 1 and an unknown attack detection at stage 2. Moreover, I design a lightweight solution for the stage 1 to enable the in-device implementation. This anomaly detection solution is an extension to classical Deep Autoencoders (DAE) to multimodality, named Multimodal DAE (M2-DAE). The proposal is evaluated against an open-set misuse detector, which leverage the best multi-class classifier we tested (i.e. Random Forest) to perform benign/class of attack classification with threshold-based unknown attack discovery. Proposal beats counterpart in terms of complexity and overall accuracy over the recently released dataset BotIoT [16].

On the same extent of classification performance enhancement, I conduct a study to face training time reduction, which is capitalized in **[J1]**. In this study I fused advantages carried out by Big Data (BD) analytics data parallelism and by hierarchical learning model parallelism to evaluate the trade-off between performance in terms of classification accuracy, speedup of training task, and deployment on a public cloud infrastructure cost. In particular, nodes of a hierarchical classifiers could be simultaneously trained because each one leverage its own dataset. Moreover, each training dataset could be further split into several portions by the means of BD infrastructure, by obtaining a distributed training over several workers machines. Evaluation is performed on the Anon17 dataset [17], which is a collection of Anonymity Tools (ATs) related traffic grouped into three diverse granularity labels, i.e. from AT to traffic category down to applications, and the results confirm the effectiveness of the proposal by reducing training time up to a 3.5 factor.

Finally, in **[J2]** (currently under review) I focused on mobile applications traffic characterization, modeling, and prediction. In this work I applied Markov-based characterization techniques to define a statistical metric (based on the Hellinger distance) to visually discriminate the traffic generated by mobile applications. The outcome of this phase is also useful to properly group applications into several homogeneous sets. Moreover, after the characterization phase, I compared the predictive power of Markov and ML-based internet traffic modeling techniques. The idea was to obtain easy explainable predictions by validating High Order Markov Chain (HOMC) effectiveness with respect to state of the art ML predictors. Results, carried out MIRAGE-2019 [18] dataset, show effectiveness of HOMC against counterpart and margins of improvement.

Products

Journal Papers

1. [J1] **Bovenzi, Giampaolo**, Giuseppe Aceto, Domenico Ciuonzo, Valerio Persico, and Antonio Pescapé. *A Big Data-enabled Hierarchical Framework for Traffic Classification*. IEEE Transactions on Network Science and Engineering (2020).

Conference Papers

1. [C1] **Bovenzi, Giampaolo**, Giuseppe Aceto, Domenico Ciuonzo, Valerio Persico, and Antonio Pescapé. *A Hierarchical Hybrid Intrusion Detection Approach in IoT Scenarios*. GLOBECOM 2020 IEEE Global Communications Conference.

Under review

1. [J2] Aceto, Giuseppe, **Giampaolo Bovenzi**, Domenico Ciuonzo, Antonio Montieri, Valerio Persico, and Antonio Pescapé. *Characterization and Prediction of Mobile-App Traffic using Markov Modeling.* IEEE Transactions on Network and Service Management (2020).

Conferences and Seminars

I have not attended any conferences during my second PhD year.

Activity Abroad

I have not carried out any activity abroad during my second PhD year.

Tutorship

- Co-supervision of M.Sc. student (Idio Guarino) thesis on “Study, implementation, and experimental evaluation of techniques for modeling and prediction of mobile application traffic using Markovian approaches”.
- Co-supervision of M.Sc. student (Nicola Esposito) thesis on “Traffic Modeling and Prediction of Mobile Applications by Composition of Deep Hybrid Networks”.
- Co-supervision of B.Sc. student (Diego D’Anna) thesis on “Data Traffic Analysis of Mobile Video Applications”.
- Co-supervision of B.Sc. student (Alessandro Testa) thesis on “Anomaly-based Intrusion Detection System for IoT: The Kitsune Model”.
- Co-supervision of B.Sc. student (Salvatore Santella) thesis on “Impact of the Discrete Wavelet Transform on Mobile Traffic Prediction”.

References

[1] Maier, G., Schneider, F., & Feldmann, A. (2010, April). A first look at mobile hand-held device traffic. In International Conference on Passive and Active Network Measurement (pp. 161-170). Springer, Berlin, Heidelberg.

[2] Falaki, H., Lymberopoulos, D., Mahajan, R., Kandula, S., & Estrin, D. (2010, November). A first look at traffic on smartphones. In Proceedings of the 10th ACM SIGCOMM conference on Internet measurement (pp. 281-287).

[3] Aceto, G., Ciuonzo, D., Montieri, A., & Pescapé, A. (2018). Multi-classification approaches for classifying mobile app traffic. Journal of Network and Computer Applications, 103, 131-145.

[4] A. Dainotti, A. Pescapé, and K. C. Claffy, *Issues and future directions in traffic classification*, IEEE Network, vol. 26, no. 1, 2012.

[5] Jonsson, P., Carson, S., Blennerud, G., Kyohun Shim, J., Arendse, B., & Husseini, A. (2019). Ericsson mobility Report. 2019. Ericsson: Stockholm, Sweden.

[6] State of the IoT 2018: Number of IoT devices now at 7B – Market accelerating, last access 18/11/2020.

[7] OWASP Internet of Things Project, last access 18/11/2020.

[8] Dai, S., Tongaonkar, A., Wang, X., Nucci, A., & Song, D. (2013, April). Networkprofiler: Towards automatic fingerprinting of android apps. In 2013 Proceedings IEEE INFOCOM (pp. 809-817). IEEE.

[9] van Ede, T., Bortolameotti, R., Continella, A., Ren, J., Dubois, D. J., Lindorfer, M., ... & Peter, A. (2020, February). FLOWPRINT: Semi-Supervised Mobile-App Fingerprinting on Encrypted Network Traffic. In Network and Distributed System Security Symposium, NDSS 2020. Internet Society.

[10] Taylor, V. F., Spolaor, R., Conti, M., & Martinovic, I. (2017). Robust smartphone app identification via encrypted network traffic analysis. IEEE Transactions on Information Forensics and Security, 13(1), 63-78.

[11] A. Montieri, D. Ciuonzo, G. Aceto, and A. Pescape, *Anonymity services Tor, I2P, JonDonym: Classifying in the dark (web)*, IEEE Trans. Depend. Sec. Comput., pp. 1–1, 2018.

[12] N. Cascarano, A. Este, F. Gringoli, F. Risso, and L. Salgarelli, *An experimental evaluation of the computational cost of a DPI traffic classifier*, in IEEE Global Communications Conference, 2009, pp. 1–8.

[13] J. Yu, H. Lee, Y. Im, M.-S. Kim, and D. Park, *Real-time classification of Internet application traffic using a hierarchical multi-class SVM*, KSII Transactions on Internet & Information Systems, vol. 4, no. 5, 2010.

[14] Y. n. Dong, J. j. Zhao, and J. Jin, *Novel feature selection and classification of Internet video traffic based on a hierarchical scheme*, Computer Networks, vol. 119, pp. 102–111, 2017.

[15] L. Grimaudo, M. Mellia, and E. Baralis, *Hierarchical learning for fine grained internet traffic classification*, in IEEE 8th International Wireless Communications and Mobile Computing Conference (IWCMC), 2012, pp. 463–468.

[16] Koroniotis, N., Moustafa, N., Sitnikova, E., & Turnbull, B. (2019). Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset. Future Generation Computer Systems, 100, 779-796.

[17] Shahbar, K., & Zincir-Heywood, A. N. (2017). Anon17: Network traffic dataset of anonymity services. Faculty of Computer Science Dalhousie University, Tech. Rep.

[18] Aceto, G., Ciuonzo, D., Montieri, A., Persico, V., & Pescapé, A. (2019, October). MIRAGE: Mobile-app Traffic Capture and Ground-truth Creation. In 2019 4th International Conference on Computing, Communications and Security (ICCCS) (pp. 1-8). IEEE.