PhD in Information Technology and Electrical Engineering
Università degli Studi di Napoli Federico II

PhD Student: Idio Guarino

Cycle: XXXVI

Training and Research Activities Report

Year: First

Tutor: prof. Antonio Pescapè

Date: October 28, 2021

1. Information:

- **PhD student: Idio Guarino**
- **DR number: 995139**
- **Date of birth: 05/03/1990**
- **Master Science degree: Computer Engineering University: University of Naples "Federico II"**
- **Doctoral Cycle: XXXVI**
- **Scholarship type: no scholarship. Funded by Consortium GARR through the awarded "O.Carlini" research grant.**
- **Tutor: Antonio Pescapè**
- **Co-tutor:**

2. Study and training activities:

Activity	Type[1]	Hours	Credits	Dates	Organizer	Certificate[2]
#andràtuttobene: Images, Texts, Emojis and Geodata in a Sentiment Analysis Pipeline	Seminar	1.5	0.3	25/11/2020	Prof. F.Amato, Prof. G.Luongo	Y
Patent searching best practices with IEEE Xplore	Seminar	1.0	0.2	27/11/2020	Eszter Lukacs	Y
How to get published with the IEEE	Seminar	1.5	0.3	02/12/2020	Paul Henriques	Y
Exploiting Deep Learning and Probabilistic Modeling for Behavior Analytics	Seminar	1.0	0.2	09/12/2020	Prof. F.Amato, Prof. G.Luongo	Y
Data Driven Transformation in WINDTRE through Managers voice	Seminar	2.0	0.4	16/12/2020	Prof. F.Amato, Prof. G.Luongo	Y
From Photometric Redshifts to Improved Weather Forecasts: an interdisciplinary view on machine learning	Seminar	1.0	0.2	13/01/2021	Prof. F.Amato, Prof. G.Luongo	Y
Cybercrime and e-evidence: the criminal justice response	Seminar	1.0	0.2	20/01/2021	Prof. F.Amato, Prof. G.Luongo	Y
Advances in Machine Learning for Modelling	Seminar	1.5	0.3	27/01/2021	Prof. Gustau	Y

and Understanding in Earth Sciences					Camps-Valls	
AI LEGAL: Artificial Intelligence for notary's sector - a case study	Seminar	1.0	0.2	27/01/2021	Prof. F.Amato, Prof. G.Luongo	Y
The era of Industry 4.0: new frontiers in business model innovation	Seminar	1.0	0.2	03/02/2021	Prof. F.Amato, Prof. G.Luongo	
Machine learning: Causality lost in translation	Seminar	1.5	0.3	10/02/2021	Prof. F.Amato, Prof. G.Luongo	Y
Robo Ludens: A game design taxonomy for human-robot interaction	Seminar	1.0	0.2	05/03/2021	Dr. John Edison Muñoz Cardona	Y
Scientific Programming and Visualization with Python	Course	20.0	2.0	08/03/2021	DIST-Prof. A.Botta	Y
Big data and Computational Linguistics	Seminar	2.0	0.4	10/03/2021	Prof. F.Amato, Prof. G.Luongo	Y
Machine Learning and Security	Seminar	2	0.4	17/03/2021	Dott. Fabio de Gaspari	Y
Data Science for Patient Records Analysis	Course	10.0	2.5	10/02/2021-18/03/2021	Prof. M.Cinque and Prof. C. Bavacchio	Y
Digital Forensics' methods, practices and tools	Course	10.0	3.0	03-10/11/2021	Dr. G. Cozzolino	Y
Statistical data analysis for science and engineering research	Course	12.0	4.0	17/02/2021-04/03/2021	Prof. R. Pietrantuono	Y
Distributional Semantics Methods: How Linguistic features can improve the semantic representation	Seminar	1.5	0.3	28/04/2021	Prof. F.Amato, Prof. G.Luongo	Y
Big Data and Business Intelligence	Course	-	6.0	II-Semester	Prof. G. Sperlì	Y
Ethics of quantification	Seminar	2.0	0.4	26/05/2021	Prof. F.Amato, Prof. G.Luongo	Y

End-to-End Optimization of Augmented Experience Services over Cloud-Integrated 5G Networks	Seminar	4.0	0.8	15-16/06/2021	Prof. A.M. Tulino	N
5G: Esposizione ai Campi Elettromagnetici e Metodologie di Misura	Seminar	4.0	0.8	16/07/2021	Prof. N. Pasquino	Y
Machine Learning - Neural networks and deep learning	Course	-	6.0	II-Semester	Prof. R. Prevete	Y

1) Courses, Seminar, Doctoral School, Research, Tutorship
2) Choose: Y or N

2.1. Study and training activities - credits earned

	Courses	Seminars	Research	Tutorship	Total
Bimonth 1	**0**	**1.4**	**8.6**	**0**	**10**
Bimonth 2	**0**	**1.4**	**8.4**	**0**	**9.8**
Bimonth 3	**11.5**	**1.3**	**2.2**	**0**	**15**
Bimonth 4	**6**	**0.4**	**3.6**	**0**	**10**
Bimonth 5	**0**	**1.6**	**3.4**	**0**	**5**
Bimonth 6	**6**	**0**	**4.2**	**0**	**10.2**
Total	**23.5**	**6.1**	**30.4**	**0**	**60**
Expected	**30 - 70**	**10 - 30**	**80 - 140**	**0 – 4.8**	

3. Research activity:

During my first year, I deepened into the motivations, applications, and issues of network traffic analysis with a specific focus on mobile and encrypted traffic. Indeed, traffic analysis represents a key prerequisite for security and QoS enforcement, and additional appeal is arising for mobile traffic analysis due to its potential for valuable profiling information (e.g., to advertisers and security agencies), while also implying privacy downsides (e.g., recognition of health or dating apps, or in bring-your-own-device scenarios). Specifically, my study activity was focused on the analysis of mobile network traffic generated by applications for Collaboration and Communication (C&C in the following) whose usage has significantly increased due to Covid-19 pandemic. Such applications include instant messaging, video conferencing, distance learning, and video streaming (with shared chat).
To this end, I firstly analyzed recent studies on the composition of network traffic and the impact of C&C apps during the lockdown periods imposed by many governments in order to contain Covid-19 pandemic which have forced millions of citizens to spend more time at home and adopt remote working modes (so-called "smart working") wherever possible. In detail, these studies have shown that the implementation of these restrictions has generated increased demand in terms of Internet traffic from residential users (+15-20% by volume) [1] for remote work, entertainment, commerce, and education. In addition, the sudden change

imposed on the lifestyles of millions of people has generated abrupt and obvious changes to the nature of traffic traversing the Internet. Recent studies show how these events have had a non-negligible impact also on the performance of networks, which—finding themselves having to manage a suddenly higher demand—have guaranteed lower performance levels compared to the period immediately preceding (for example, showing increases in the variability of delays and loss rates) [2].
As a consequence, since the networks have to respond to a higher and qualitatively different demand, the need arises to design new tools for an effective and efficient management of the traffic generated by the users accessing remote services. This situation, which is still in progress, introduces new and interesting research challenges, aimed at identifying innovative solutions that support monitoring, management and engineering activities of the networks themselves, in order to guarantee their correct functioning in compliance with the expected security levels.

However, the evolution of the nature of traffic traversing network devices have progressively rendered less effective techniques for traffic analysis proposed in the past. This introduces the need for rapid adaptation of the strategies and information elements on which these techniques are based. To a large extent, the traffic generated by modern applications is encrypted, leading to the impossibility of interpreting and exploiting the information exchanged between parties. Moreover, established communications often refer to dynamically assigned port numbers or are subject to NAT protocol interference. This makes techniques based on payload inspection and port analysis inapplicable. In addition, the prevalence of applications with servers hosted on (a few) cloud infrastructures further confounds IP address-based methods. Finally, the use of application protocols as transport sublayers (TLS and HTTP) common to even very different applications further complicates the problem. The difficulties related to a correct classification and prediction of traffic are exacerbated in the case of mobile apps, characterized by high numbers and continuous updates, possibly generating similar traffic patterns and with complex fingerprints. The latter is due to the scarce number of training samples per app and device/OS/version diversity. Consequently, such moving-target nature of mobile-app traffic can significantly impair the ability to collect and effectively analyze sufficient amounts of traffic.

In accordance with the promising trend of methodologies based on AI techniques and recent research results obtained through the application of such methodologies in the context of computer networks and in particular traffic analysis [3, 4], my research work aims to provide innovative tools based on Artificial Intelligence (AI) methodologies, with particular reference to Deep Learning (DL) architectures. The use of such DL techniques is considered the natural facilitator for the design of traffic analysis tools capable of bringing out the distinctive "fingerprint" of each application from the observed "raw" traffic. Specifically, my research activities are aimed at supporting two fundamental tasks in network monitoring and management: traffic classification and prediction. Both are among the traffic analysis activities that can be defined as horizontal, i.e. preparatory to the effective and efficient management of network infrastructures and to study and optimize their evolution. The former allows traffic to be categorized according to heterogeneous and complementary views, such as: the typology, the generating application or the particular service used (enabling the optimized management of the network according to the conveyed contents). The latter allow one to obtain information on how the traffic characteristics will evolve in time (allowing one to predict the network needs, both in view of an optimized provisioning of resources and of the application of new techniques, architectures, and services).

In line with recent studies, given the growing interest in the traffic generated by C&C apps, and given the lack of sufficient data to study its characteristics as well as to experimentally validate the investigated approaches, a part of my activity has been dedicated to the collection of a dataset of real traffic traces generated by these C&C apps by exploiting and optimizing the MIRAGE architecture [5], which proved to be an effective tool for the realization of "human-generated" datasets. To this end, the collection of traffic traces has been performed on a voluntary basis, involving the students of the University of Napoli Federico II and proceeding to the collection of data in a controlled environment to overcome the constraints related to data management (mainly arising from issues in terms of privacy) and to capture heterogeneous data adherent to a habitual use of the considered applications. It is worth noticing that traffic data, in addition to the collection performed in a controlled environment, have been properly anonymized before carrying out the traffic analysis tasks.

Motivated by the fact that knowing the characteristics and the peculiarity of the traffic crossing the network is critical for a number of network-related task such as identification/classification and prediction, I addressed the characterization and modeling of the network traffic generated by some of the top popular C&C apps when they are used by the user to perform some activity relevant to remote work [C1].
Specifically, this study has provided a characterization at trace level of the traffic, which allowed one to identify peculiar characteristics in terms of downstream and upstream bit rate and packet rate as well as downstream volume and packet fraction. Additionally, the inspection of the traffic at flow level, highlighted how the presence of concurrent biflows depends upon both app and activity, and also follows interesting trends with the time. Finally, the modeling analysis by means of Markov chains, has highlighted some similar transition patterns for different apps performing the same activity, while different patterns have arisen for certain apps.
Moreover, I have addressed the analysis of the traffic from the viewpoint of the classification employing state-of-art deep learning approaches to assess to which degree the apps, their different use cases (activities), and the pairs app-activity can be told apart from each other [C2]. To this end, I also investigated the early behavior of the biflows composing the traffic and the effect of tuning the dimension of the input, via a sensitivity analysis. The experimental analysis highlights the figures of the different architectures, in terms of both traffic-classification performance and complexity w.r.t. different classification tasks, and the related trade-off.
Finally, from the viewpoint of network prediction, I explored different training strategies to reduce the number of models, adopting the Markov Chains to model mobile video apps traffic at packet-level [C3]. I discussed and experimentally evaluated the prediction effectiveness of the proposed approaches by comparing the performance of app models with models trained on a specific category of video apps and a model trained on the mix of all video traffic. Experimental results showed that, by training models on traffic characteristics that group information with a coarser granularity than per-app one, it is possible to reduce the number of models with very little performance loss (or even a slight improvement) compared to those obtained with models trained at the finest granularity (i.e. one model for each app), which ideally should provide the best performance.

4. Research products:

[C1] *"Characterizing and Modeling Traffic of Communication and Collaboration Apps Bloomed With COVID-19 Outbreak",* Idio Guarino, Giuseppe Aceto, Domenico Ciuonzo, Antonio Montieri, Valerio

Persico, Antonio Pescapè, 2021 IEEE 6th International Forum on Research and Technology for Society and Industry (RTSI), published, 2021

[C2] *"Classification of Communication and Collaboration Apps via Advanced Deep-Learning Approaches",* Idio Guarino, Giuseppe Aceto, Domenico Ciuonzo, Antonio Montieri, Valerio Persico and Antonio Pescapè, 2021 IEEE International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), accepted, 2021

[C3] *"Mobile Network Traffic Prediction Using High Order Markov Chains Trained at Multiple Granularity",* Idio Guarino, Alfredo Nascita, Giuseppe Aceto, Antonio Pescapè, 2021 IEEE 6th International Forum on Research and Technology for Society and Industry (RTSI), published, 2021

5. Conferences and seminars attended

2021 IEEE 6th International Forum on Research and Technology for Society and Industry (RTSI), Virtual mode, 08/09/2021. Paper presentation
2021 IEEE International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), virtual mode, To be defined. Paper presentation

6. Activity abroad:

7. Tutorship

References

[1] A. Feldmann, O. Gasser, F. Lichtblau, E. Pujol, I. Poese, C. Dietzel, D. Wagner, M. Wichtlhuber, J. Tapiador, N. Vallina-Rodriguez, O. Hohlfeld, and G. Smaragdakis, "The Lockdown Effect: Implications of the COVID-19 Pandemic on Internet Traffic", ACM Internet Measurement Conference (IMC '20).

[2] M. Candela, V. Luconi, and A. Vecchio, "Impact of the COVID-19 pandemic on the Internet latency: A large-scale study", Elsevier Computer Networks, Volume 182, 2020, 107495, ISSN 1389-1286.

[3] G. Aceto, D. Ciuonzo, A. Montieri and A. Pescapé, "DISTILLER: Encrypted Traffic Classification via Multimodal Multitask Deep Learning", Elsevier Journal of Network and Computer Applications, 2021 (in press).

[4] G. Aceto, G. Bovenzi, D. Ciuonzo, A. Montieri, V. Persico and A. Pescapé, "Characterization and Prediction of Mobile-App Traffic using Markov Modeling", IEEE Transactions on Network and Service Management, 2021 (in press).

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