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 – First 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 first year of PhD program, I attended the courses and seminars reported in the following. In September 2019, I also attended the International Summer Schools on Cyber at the University of Padua that has been included in the list of courses.

Courses

1. *Author Seminar - How to publish a scientific paper*, ad hoc module, Dr. Aliaksandr Birvkov and Elisa Magistrelli, 26/11/2018, 0.4 Credits.
2. *Ciberconflitti*, ad hoc module, Gian Piero Siroli, Francesco Vestito, Prof. Simon Pietro Romano, and Daniele Amoroso, 28/11/2018, 0.8 Credits.
3. *Data Science and Optimization*, ad hoc module, Prof. Manlio Gaudioso, Prof.ssa Laura Palagi, and Prof.ssa Enza Messina, 05/02/2019-07/02/2019, 1.2 Credits.
4. *Advanced techniques for software robustness and security testing*, ad hoc module, Prof. Roberto Natella, 16/01/2019, 30/01/2019, 13/02/2019, 20/02/2019, 07/03/2019, 19/03/2019, 02/04/2019, 3 Credits.
5. *Big Data*, ad hoc module, Prof. Antonio Picariello and Ing. Giancarlo Sperlì, 26/02/2019, 27/02/2019, 03/03/2019-05/03/2019, 3 Credits.
6. *Strategic Orientation for STEM Research & Writing*, ad hoc module, Chie Shin Fraser, 15/03/2019, 18/03/2019, 22/03/2019, 25/03/2019, 01/04/2019, 08/04/2019 6 Credits.
7. *Machine Learning*, ad hoc module, Anna Corazza, Francesco Isgrò, Stefano Olivieri, Roberto Prevete, and Carlo Sansone, 06/05/2019-10/05/2019, 13/05/2019-15/05/2019, 17/05/2019, 20/05/2019, 4.2 Credits.
8. 2019 International Summer School on “Machine Learning and Security”, doctoral school, 09/09/2019-13/09/2019 (40 h), 5 Credits.

Seminars

1. *Parallel and Distributed computing with MATLAB*, Ing. Stefano Marrone, 21/11/2019, Credits 0.4.
2. *An Introduction to Blockchains*, Ida Rejeki Siahaan, 03/12/2018, Credits 0.4.
3. *Bitcoin e Blockchain oltre l'hype*, Gabriele Sabbatini, Lorenzo Giustozzi, Marco Monaco, and Felice Balsamo, 18/01/2019, Credits 0.6.
4. *MATLAB and Embedded Systems*, Ing. Stefano Marrone, 28/03/2019, Credits 0.4.
5. *IEEEXploreTraining and Authorship Workshop*, Dr.ssa Eszter Lukacs, 04/04/2019, Credits 0.5.
6. *Internet censorship: enforcement, detection, and circumvention*, Credits 2.
7. *In-network Machine Learning for Networks*, Dr. Roberto Bifulco, 14/06/2019, Credits 0.4.
8. *Applying Semi-Supervised Learning to App Store Analysis*, Daniel Rodriguez, 12/07/2019, Credits 0.2.
9. *On Reinforcement Learning for computing channel capacity with feedback*, Prof. Haim Permuter, 21/10/2019, Credits 0.2.

Credits Summary

	Credits year 1							
	Estimated	1 bimonth	2 bimonth	3 bimonth	4 bimonth	5 bimonth	6 bimonth	Summary
Modules	20	1.2	4.2	9	4.2	0	5	**23.6**
Seminars	5	0.8	0.6	0.9	2.4	0.2	0.2	**5.1**
Research	35	8	5.2	0.1	3.4	9.8	4.8	**31.3**
	60	10	10	10	10	10	10	**60**

Research activity

Techniques for encrypted and anonymous Traffic Classification

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

Notably, I am working on machine learning (and deep learning) -based traffic classification techniques, that are effective 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 several security-concerns traffic categories, from (i) the classification of Anonymity Tools (ATs) (i.e. Tor, I2P, and JohnDonym) generated traffic, which ATs provide anonymization of end-users and encryption of exchanged traffic, to (ii) the detection of network threats (i.e. botnets), which broken privacy of users and affect both critical Internet infrastructure (e.g. cloud) and home networks (e.g. IoT).

Moreover, thanks to the collaboration with the CINI consortium and University of Campania Luigi Vanvitelli, I had the opportunity to apply Deep Learning techniques to biomedical images.

Research Description

Traffic Classification (TC) is the association of traffic aggregates with specific applications or groups of applications and is of utmost importance in Internet quality-of-service enforcement, traffic engineering, network security, etc. Related methodologies & tools support activities such as network monitoring, security assessment, application identification, anomaly detection, accounting, advertising, and service differentiation. In particular, understanding the nature of the communications flowing through the Internet is essential for operators to properly manage the network [1].

The classical TC process (i.e. Deep Packet Inspection) is more and more hampered by **encryption** (and **anonymization**) of Internet Traffic. These obfuscation techniques are required by the growing criticality of the activities users perform online and related privacy concern. In particular, preserving the anonymity of users is an aspect that has gathered the attention of the research community, that over last years put the effort in designing and developing tools able to achieve privacy at varying degrees. As a result, at present several Anonymity Tools (ATs) are freely available. ATs act as facilitators for Internet users, allowing them to obfuscate the communication as well as the nature of the exchanged contents to any eavesdropping entity. On one hand, ATs challenge Internet authorities in the discovery of cybercrimes, e.g., selling copyrighted or malicious software, drugs, guns, child porn, and stolen digital identities or hiding online frauds, extremism, hacking, and abuses. On the other hand, they are essential in sharing crucial information through the Internet, e.g. when censorship is enforced by non-democratic actors, or for the sole right to privacy [2]. The latter aspect confirms their original significance to keep the Internet as a fully-available common good. As a result, ATs like Tor are currently widespread, with a turnout up to two millions of users. Net of this, we put ATs among critical networks of which is important to understand the behaviour for **ethical reasons**.

Nowadays, the Internet is facing a growth in complexity about traffic volume and heterogeneity. The former due as to the growing of peripheral networks (i.e. IoT) as to the band increasing. Whereas, the latter refers to both Internet-enabled devices and related traffic. However, some efforts are required to enhance fine-grain classification of Internet traffic. 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 7$B in 2018, with a 3x expected growth by 2025, considering both consumer and industrial applications [3].

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 [4]. Once compromised, IoT devices may be leveraged collectively in the form of a botnet for malicious purposes, with their nature constantly evolving according to these newly exploitable vulnerabilities. As a specific application of TC, Network Intrusion Detection Systems (NIDSs) are meant to monitor network traffic to determine when a system is being targeted by (or is a source of) a network attack, thus providing the tools to implement security countermeasures.

The IoT environment is characterized by high dynamism from multiple viewpoints: the number and variety of devices, their spatial distribution, and the evolution of attacks. A NIDS aimed IoT scenarios must cope with these additional challenges, and its design is expected to aim at detecting yet unknown attacks, while retaining high efficiency in performing its tasks. Accordingly, NIDSs constitute a fundamental building block in today's networks and implement two main approaches, anomaly detection and misuse detection, aiming at capturing any deviation from the profiles of normal activities, and identifying patterns of known attacks, respectively [5].

Exploring Solution

Because of the specific nature of nowadays encrypted traffic, Machine Learning (ML) classifiers represent the natural enabler, able to provide decisions based on the sole traffic flow features [6] and to overcome shortcomings due to the application of common solutions (e.g., those based on payload inspection or earlier port-based ones [7]). Leveraging Machine Learning (ML) 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 explored 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 [8-10].

For the sake of example, Hierarchical Classification (HC) represents a **perfect match for TC of ATs**, as (i) it allows fine-grained tuning and design, potentially leading to classification performance gains; (ii) it also brings a number of “practical” benefits by design, at cost of moderate complexity increase. For example, re-training does not involve all the nodes in the hierarchy when new applications leveraging anonymity networks are released. Also, distributed deployment of TC tasks, thanks to the modularity of the framework, is enabled in the network (thus hierarchical classification could be achieved through chaining of

virtualized network functions, each associated to a classifier). The same advantages could be exploited to build more robust, scalable, and performing NIDS, benefiting from HC intrinsic operational and (re-)training efficiency.

Finally, **Big Data (BD) analytics** could manage and capitalize the huge amount of Internet traffic. In this direction, initial effort has been put forward by scientific community and industry to apply BD technologies, e.g. Apache Spark or Apache Hadoop, to ML-based TC, exploiting data parallelism [11-13]. The integration between HC and BD implementing double parallelism, i.e. that integrates the advantages of model parallelism given by hierarchical TC, with those originated by data parallelism, provided by BD technologies.

Products

Journal Papers

1. [J1] Montieri, Antonio, Domenico Ciuonzo, **Giampaolo Bovenzi**, Valerio Persico, and Antonio Pescapé. *A Dive into the Dark Web: Hierarchical Traffic Classification of Anonymity Tools.* IEEE Transactions on Network Science and Engineering (2019).

Conference Papers

1. [C1] Piantadosi, Gabriele, **Giampaolo Bovenzi**, Giuseppe Argenziano, Elvira Moscarella, Domenico Parmeggiani, Ludovico Docimo, and Carlo Sansone. *Skin Lesions Classification: A Radiomics Approach with Deep CNN.* In International Conference on Image Analysis and Processing, pp. 252-259. Springer, Cham, 2019.

Under review

1. **Bovenzi, Giampaolo**, Giuseppe Aceto, Domenico Ciuonzo, Valerio Persico, and Antonio Pescapé. *Double Parallelism in Traffic Classification: Big Data-enabled Hierarchical (BDeH) Framework.* IEEE Network (2019).
2. **Bovenzi, Giampaolo**, Giuseppe Aceto, Domenico Ciuonzo, Valerio Persico, and Antonio Pescapé. *H2ID: Hierarchical Hybrid Intrusion Detection for Security of IoT Devices.* IEEE International Conference on Communications (2020).

Conferences and Seminars

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

I made the following presentations:

1. Ital-IA, primo Convegno Nazionale sull'Intelligenza Artificiale del CINI – *Classificazione Gerarchica del Traffico di Reti Anonime con Machine Learning*, 18/03/2019.
2. 2019 International Summer School on "Machine Learning and Security" – *Final presentation*, 13/09/2019.

Activity Abroad

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

Tutorship

I have not carried out any tutorship during my first PhD year.

References

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

[2] G. Aceto and A. Pescapé, "Internet censorship detection: A survey," Computer Networks, vol. 83, pp. 381–421, 2015.

[3] State of the IoT 2018: Number of IoT devices now at 7B – Market accelerating, last access 28/10/2019.

[4] OWASP Internet of Things Project, last access 28/10/2019.

[5] P. Mishra, V. Varadharajan, U. Tupakula, and E. S. Pilli, *A detailed investigation and analysis of using machine learning techniques for intrusion detection*, IEEE Communications Surveys & Tutorials, 2018.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] V. D'Alessandro, B. Park, L. Romano, C. Fetzer et al., *Scalable network traffic classification using distributed support vector machines*, in IEEE 8th International Conference on Cloud Computing (ICCC), 2015, pp. 1008–1012.

[12] Z. Yuan and C. Wang, *An improved network traffic classification algorithm based on Hadoop decision tree*, in IEEE International Conference of Online Analysis and Computing Science (ICOACS), 2016, pp. 53–56.

[13] L.-V. Le, B.-S. Lin, and S. Do, *Applying big data, machine learning, and SDN/NFV for 5G earlystage traffic classification and network QoS control*, Transactions on Networks and Communications, vol. 6, no. 2, p. 36, 2018.