Università degli Studi di Napoli Federico II

PhD program in

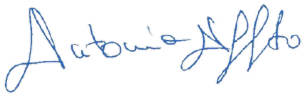
Information Technology and Electrical Engineering

PhD Student: Antonia Affinito

Cycle: XXXV

Training and Research Activities Report

Academic year: 2020-21 - PhD Year: Second



student signature

Tutor: prof. Alessio Botta

Co-Tutor:

Date: October 21, 2021

Training and Research Activities Report

PhD program in Information Technology and Electrical Engineering

PhD student: Antonia Affinito
Cycle: XXXV

1. Information:

- **PhD student: Antonia Affinito** **PhD Cycle: XXXV**
- **DR number: DR993885**
- **Date of birth: 10/08/1992**
- **Master Science degree: Computer Engineering**
- **University: University of Napoli "Federico II"**
- **Scholarship type: Unina**
- **Tutor: Prof. Alessio Botta**
- **Co-tutor:**

2. Study and training activities:

Activity	Type[1]	Hours	Credits	Dates	Organizer	Certificate[2]
"AI4NETS-AI/ML for data communication Networks-Tutorial"	Seminar	4	0.8	02/11/2020	Prof. Marco Mellia, Politecnico di Torino	No
"Robot Manipulation and Control"	Seminar	2.5	0.5	17/11/2020	Prof. Bruno Siciliano	Yes
"Telemedicina in Italia: casi di successo"	Seminar	1.5	0.3	17/11/2020	Prof. Giovanni D'Addio	Yes
Lecture 3 on Data Science-"Digital Project Management: practices, processes, techniques, tools and scientific approach"	Seminar	2	0.4	18/11/2020	Prof. Giuseppe Luongo	yes
"L'esperienza del progetto di tele-riabilitazione NEUROREAB"	Seminar	1.5	0.3	24/11/2020	Ing. Giovanni D'Addio	yes

Lecture 4 on Data Science- "#andràtuttobene: Images, Texts, Emojis and Geodata in a Sentiment Analysis Pipeline"	**Seminar**	**1.5**	**0.3**	**25/11/2020**	**Prof. Giuseppe Luongo**	**yes**
“Telemedicina, e-health e «mobile health» si può davvero usare il digitale nel percorso assistenziale?”	**Seminar**	**1.5**	**0.3**	**26/11/2020**	**Dott.ssa Simonetta Scalvini**	**yes**
“Patent Searching Best Practices with IEEE Xplore”	**Seminar**	**1**	**0.2**	**27/11/2020**	**Dott. Alessandra Scippa**	**yes**
Lecture 5 on Data Science: “At the Nexus of Big Data, Machine Intelligence, and Human Cognition"	**Seminar**	**1**	**0.2**	**2/12/2020**	**Prof. Giuseppe Luongo**	**yes**
(MSc Course): “Data Management”	**Course**		**6**	**28/09/2020-22/12/2020**	**Prof.ssa Flora Amato**	**Yes**
Lecture on Data Science: “From Photometric Redshifts to Improved Weather Forecasts: an interdisciplinary view on machine learning”	**Seminar**	**1**	**0.2**	**13/01/2021**	**Prof. Giuseppe Luongo**	**Yes**
Lecture on Data Science: "Cybercrime and e-evidence: the	**Seminar**	**2**	**0.4**	**20/01/2021**	**Prof. Giuseppe Luongo**	**yes**

criminal justice response"						
Lecture on Data Science: "The era of Industry 4.0: new frontiers in business model innovation"	**Seminar**	**1**	**0.2**	**27/01/2021**	**Prof. Giuseppe Luongo**	**Yes**
Lecture on Data Science: " Machine learning: Causality lost in translation"	**Seminar**	**1.5**	**0.3**	**10/02/2021**	**Prof. Giuseppe Luongo**	**yes**
Lecture on Data Science: "Approaches to Graph Machine Learning"	**Seminar**	**1**	**0.2**	**17/02/2021**	**Prof. Giuseppe Luongo**	**yes**
"Antonio Picariello Lectures: Visual Interaction and Communication in Data Science"	**Seminar**	**2**	**0.4**	**03/03/2021**	**Prof. Giuseppe Luongo**	**yes**
"Robo Ludens: A game design taxonomy for human-robot interaction"	**Seminar**	**1**	**0.2**	**05/03/2021**	**Prof. Giuseppe Luongo**	**yes**
"Ethics of qualification", Antonio Picariello Lectures series"	**Seminar**	**2**	**0.4**	**26/05/2021**	**Prof. Giuseppe Luongo**	**yes**
"End-to-end optimization of augmented experience services over cloud-integrated 5G networks"	**Seminar**	**4**	**0.8**	**15/06/2021;16/06/2021**	**Dr. Jaime Llorca, New York University**	**No**
SSIE 2021 – IEEE Italy	**Course**	**5 days**	**5**	**12-16/07/2021**	**University of Padova**	**yes**

PhD student: Antonia Affinito
Cycle: XXXV

Section Summer School. Machine Learning Theory.						

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**	**3.3**	**6**	**6**	**15.3**
Bimonth 2	**6**	**1.3**	**5**	**0**	**6.3**
Bimonth 3	**0**	**0.6**	**9**	**0**	**9.6**
Bimonth 4	**0**	**1.2**	**8**	**0**	**9.2**
Bimonth 5	**5**	**0**	**7.5**	**0**	**12.5**
Bimonth 6	**0**	**0.4**	**8**	**0**	**8.4**
Total	**11**	**6.8**	**43.5**	**6**	**61.3**
Expected	**30 - 70**	**10 - 30**	**80 - 140**	**0 – 4.8**	

3. Research activity:

The still growing number of connected devices generates a massive quantity of network traffic: the amount of data that has to be analysed is higher and higher, especially in current high-speed networks. At the same time, sophisticated network attacks are growing exponentially and getting them is more and more complicated.
In a first work, we worked at flow level for coping with the high-speed of current and future networks. However, even at the flow-level, the analysis of traffic for the detection of anomalies in high-speed networks requires huge computational power or data reduction techniques as flow records still represent a huge quantity of data. Therefore, we decided to analyse the network traffic at flow level applying Big Data techniques in order to solve the problems of huge quantity of data to analysed. In particular, we focused on the detection of the most spread network anomalies – port and net scan [3]. In the former case, an attacker probes a various TCP/UDP ports to find active and vulnerable services. In the latter case, the attacker scans a group of victim hosts on a single or a small number of ports. These two types of anomalies are the typical preliminary steps an attacker makes to find victims in a certain network.
Thereafter, we decided to filter the network traffic and to analyse only the Domain Name System (more simply DNS) packets. The DNS is a system able to convert human-readable names in their corresponding IP addresses. It is an indispensable component of the Internet world, distributed over a global network of DNS servers that are constantly in communication with each other to bring users to their websites or network resources.
The Domain Name System is also considered a valid tool to analyse a lower percentage of traffic and to extract interesting information about the network operation. In addition, new domain names are registered every day, but the 70% of them are “malicious”, “suspicious” or “not safe for work”. [1]
The first datasets we analysed were those provided by Cisco Umbrella and Alexa every day: they are lists that contain the Top One Million popular domain names and websites, respectively.

Looking at the trends of the most popular applications, divided by categories, we observed variations in the scores of the domain names during the COVID pandemic period [2]. The two providers of the lists adopt different methods for the ranking evaluation: the Umbrella list contains the most queried domains based on passive DNS; the Alexa's list contains the most popular sites visited by people that use Alexa's browser extensions. Our results show which application was most used and by which type of device during the quarantine period.
The second type of the datasets we analysed consists of a list of malicious hostnames collected by Cisco analysts. In particular, this dataset contains three types of domain names: C&C – domains associated with a Command-and-Control systems of botnets; Malware – domains associated with malware threats; Phishing – domains associated with phishing pages. In particular, the purpose of our work is to find out which type of DNS resolver, local or public, has the best security/response time ratio. The preliminary results related to the response code show that all analysed DNS resolvers protect us from most malicious hostnames with a different approach: some DNS resolvers return the "0" rcode but the IP address is related to a courtesy page; others return "NXDOMAIN" in the DNS response. The results of the response times show that local DNS resolvers are generally faster than public ones. In addition, the comparison between the two analysed public resolvers show that Google is faster than OpenDNS.
The third type of dataset that we are analysing is characterized by several zone files provided by the University of Twente. In particular, the goal of this project is to detect malicious domain names through their lifetime retrieved from the information in the zone files. Indeed, the lifetime of a domain is set to approximately 1-2 years for benign domains. The lifetime of a malicious domain name is shorter than that of a benign one.

[1]: Z. Chen, J. Javier Wang, K. Kwan; Newly Registered Domains: Malicious Abuse by Bad Actors. Palo Alto Company
[2] A. Affinito, A. Botta, and G. Ventre, "The impact of covid on network utilization: an analysis on domain popularity," in 2020 IEEE 25th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), 2020, pp. 1–6.
[3] A. Affinito, A. Botta, L. Gallo, M. Garofalo, and G. Ventre, "Spark-based Port and Net Scan Detection"; Proceedings of the 35th Annual ACM Symposium on Applied Computing

4. Research products

- *Antonia Affinito, Alessio Botta, Luigi Gallo, Mauro Garofalo, Giorgio Ventre; "Spark-based Port and Net Scan Detection"; The 35th ACM/SIGAPP Symposium on Applied ComputingACM SAC; published; 2020.*
- *Antonia Affinito, Alessio Botta, Giorgio Ventre; "The impact of Covid on network utilization: an analysis on domain popularity"; IEEE CAMAD 2020; online conference; published; 2020.*
- *Antonia Affinito, Alessio Botta, Giorgio Ventre; "Local and Public DNS Resolvers: do you trade off performance against security?"; submission at NOMS Conference 2022.*

5. Conferences and seminars attended

- *Network Traffic Measurement and Analysis Conference; TMA Conference 2021; online conference;*
- *3rd International Workshop on AI in Networks and Distributed Systems; Online Workshop.*

6. Periods abroad and/or in international research institutions

I started the period abroad the 14th of April 2021 and I will be here until the end of March 2022.
The hosting institution is the University of Twente (UT) in Enschede (Netherlands). The supervisor is Roland van Rijswijk-Deij, adjunct Professor.
We are currently working on the detection of the malicious domain names with two different approaches. The former one consists of studying the lifetime of the domain names in order to understand if there are different patterns in the lifetime of the malicious and benign hostnames. The latter one is the detection of the malicious domain names analysing some features provided by the data collected by Open Intel platform (from the UT University) and the Certificate Transparency Logs.
The number of months spent abroad during the second year is 5 months.

7. Tutorship

- *"Fondamenti di Informatica", where Prof. Alessio Botta is a lecturer; Number of hours=18.*

8. Plan for year three

For the next year, I plan to investigate the topic of the malicious domain names detection by looking at their lifetime. The goal is to analyze the domain names with a short lifetime in the zone file.
I also plan to work on the detection of malicious domain names analyzing some features provided by Open Intel platform and Certificate Transparency Logs. These two datasets are provided by the University of Twente (Netherlands) where I am spending my period abroad, under the supervision of two supervisors. The idea is to apply supervised and unsupervised machine learning algorithms in order to realize a classifier capable of detecting and distinguishing the different types of malicious domain names.
In addition, I also plan to investigate the behaviors of malicious traffic, with a focus on botnets for the IoT devices: network of devices controlled from a single command and control server. The DNS is typically used to establish links between IoT botnets and their C&C server. A possible idea is to detect the C&C activities using features provided by the DNS protocols as well as by the botnet traffic.