ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE,TAKES YOU FAR

XXXIV Cycle

Machine learning for anomaly detection in network traffic

Francesca Soro

Supervisors: Prof. Marco Mellia, Prof. Antonio Lioy

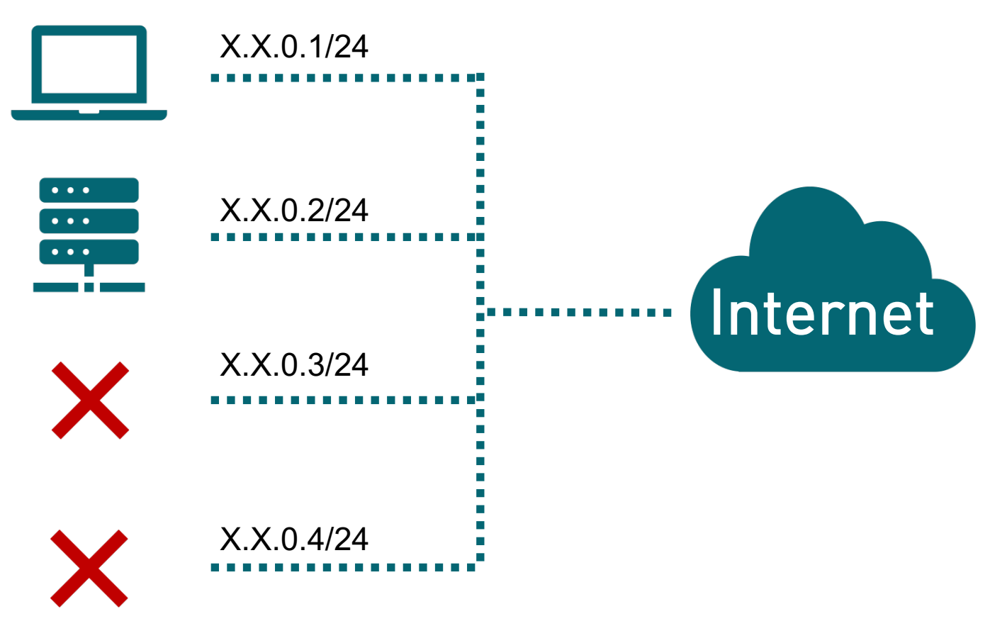
Research context and motivation

- The Internet traffic scenario is nowadays experiencing a continuous growth in both **volume** and **complexity,** given the wide variety and increasing amount of connected devices
- New threats and anomalies showing unpredictable and unseen fingerprints are generated everyday, making the design of an **efficient automatic cybersecurity system** a problematic task
- Automatic detection of **zero-days attacks** is almost impossible, given the lack of already labelled data, and so far only **knowledge-based systems** adopting **signature-based** or **novelty detection** are applied
- The current state of things and the large amount of available raw data, call for a **big-data approach** to help extract meaningful information

Addressed research questions/problems

The first step to recognize anomalous traffic automatically, is the characterization of anomalous traffic behavior. We start from one of the most relevant sources of this kind of traffic: **darknets.** A darknet is defined in literature as *a set of IP addresses that is advertised without answering any traffic*. Therefore, all the traffic hitting the darknet is by definition **unsolicited**, and in most cases of malicious nature. Considering this assumptions, we are able to use such traffic for several tasks:

- **Misconfiguration** detection
- **Botnets** monitoring
- **DDoS attacks** identification
- IPv4 address space **utilization estimation**
- **Internet censorship** analysis



To understand the spread of anomalous traffic around the world, and how dependent it is from the geographical location of the destination, we compare three darknets:

- /15 located in the Netherlands **(131,072 IPs, 30GB/day)**
- /19 located in Brazil **(8,192 IPs, 2.5GB/day)**
- 3*/24 located in Italy **(768 IPs, 420 MB/day)**

The comparison of different traffic destinations across the world allows us to better understand similarities and differences across darknets, and the impact of the size and of the allocated address space on the received traffic.

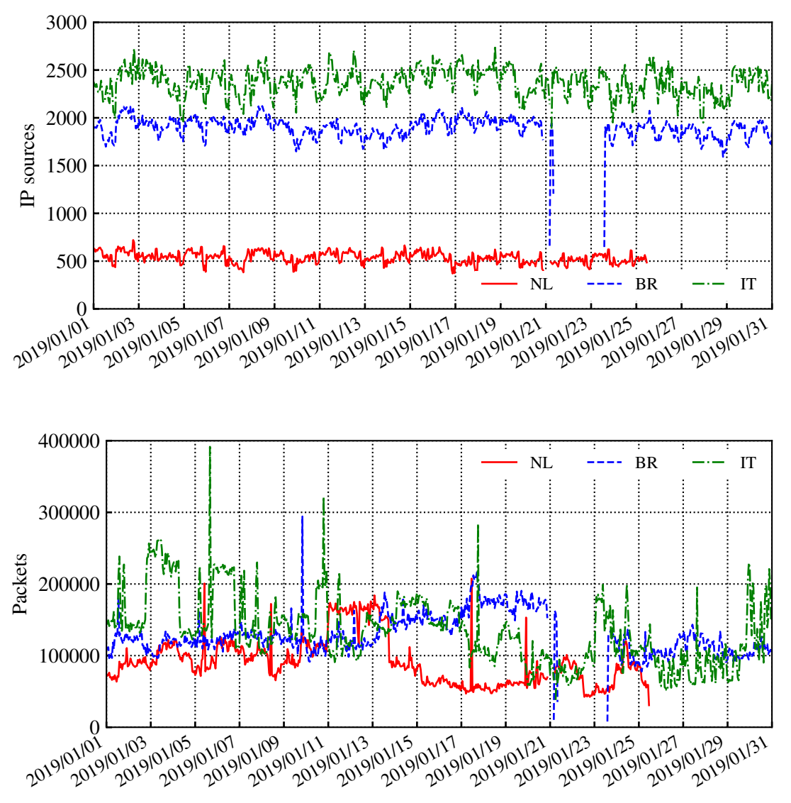
Submitted and published works

- P Casas, F Soro, J Vanerio, G Settanni, A D'Alconzo , *Network security and anomaly detection with Big-DAMA, a big data analytics framework*, IEEE 6th International Conference on Cloud Networking (CloudNet), Prague, 2017, pp. 1-7
- A Faroughi, R Javidan, M Mellia, A Morichetta, F Soro, M Trevisan, *Achieving Horizontal Scalability in Density-based Clustering for URLs*, IEEE International Conference on Big Data (Big Data), Seattle, 2018, pp. 3841-3846
- A Safari Khatouni, F Soro, and D Giordano*, A Machine Learning Application for Latency Prediction in Operational 4G Networks*., 2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM). IEEE, 2019.
- F Soro, I Drago, M Trevisan, M Mellia, J Ceron, J Santanna, *Are darknet all the same? On darknet visibility for security monitoring*,IEEE LANMAN 2019,Paris, 2019 (still to be published)

Adopted methodologies

Characterisation of traffic in terms of:

- Traffic **types** (scan, backscattering, UDP, ICMP, …)
- **Temporal patterns**
- Traffic **origins** (AS and Country of sources)
- **Per-port breakdown**



Evaluation of the effects of **parameter tuning**:

- Impact of the **length of the observation period**
- Impact of the **darknet size**

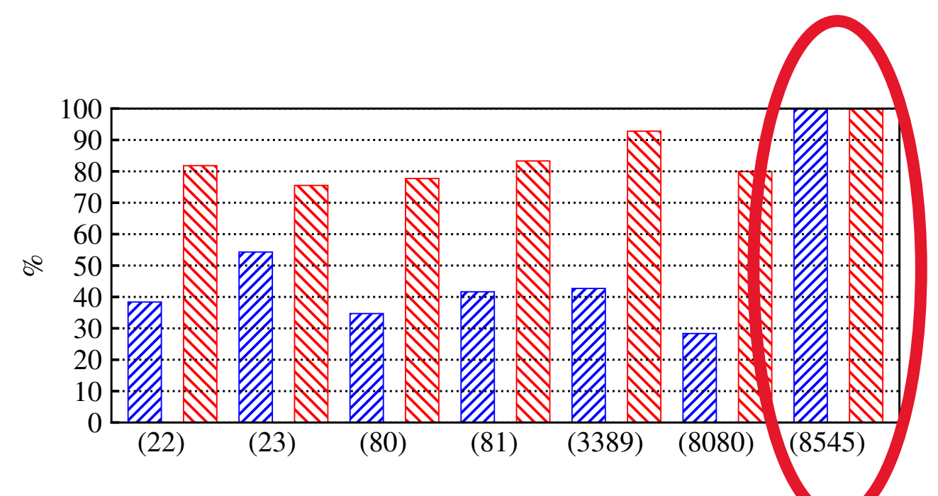
Both evaluated by means of the **Jaccard Index**:

$$\frac{set(ASes_{d\_1}) \cap set(ASes_{d\_2})}{set(ASes_{d\_1}) \cup set(ASes_{d\_2})}$$

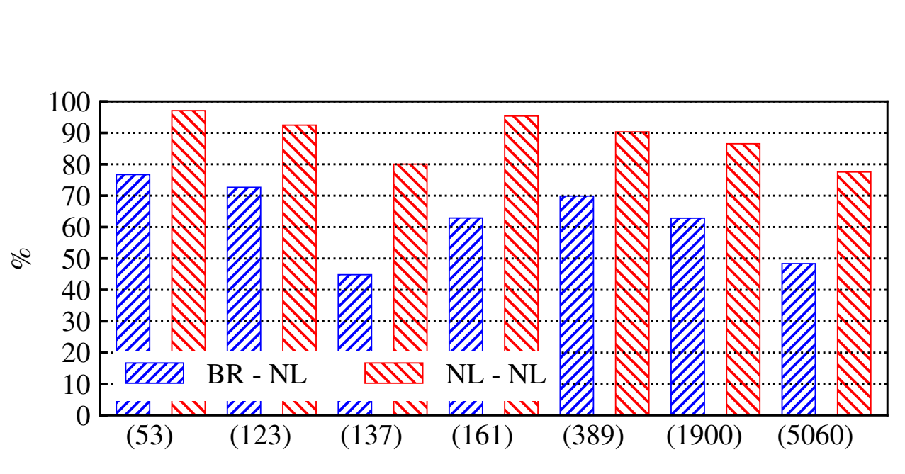
Novel contributions

From the similarity analysis, we are able to derive several useful findings on the source of anomalous traffic.

1. How spread are the sources targeting each port?

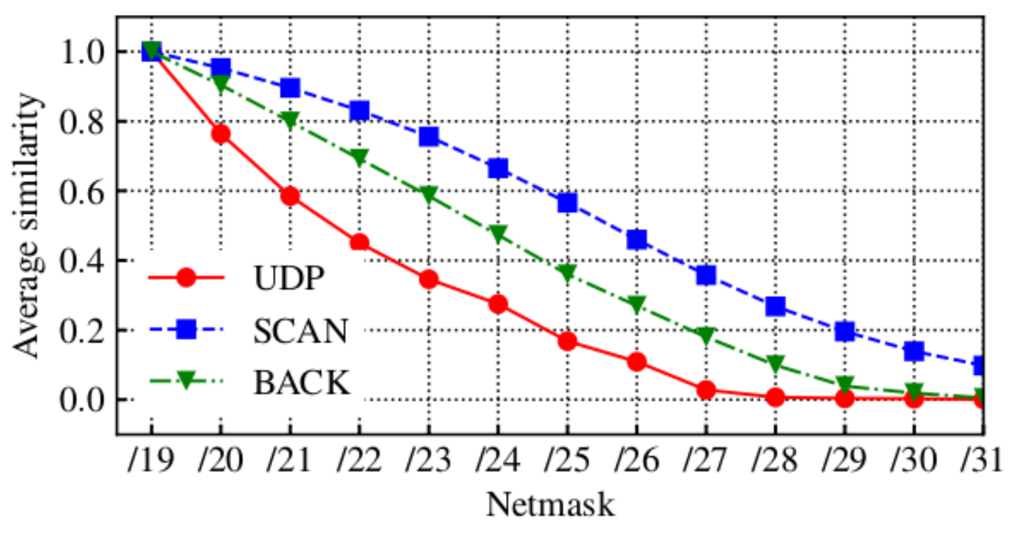


TCP

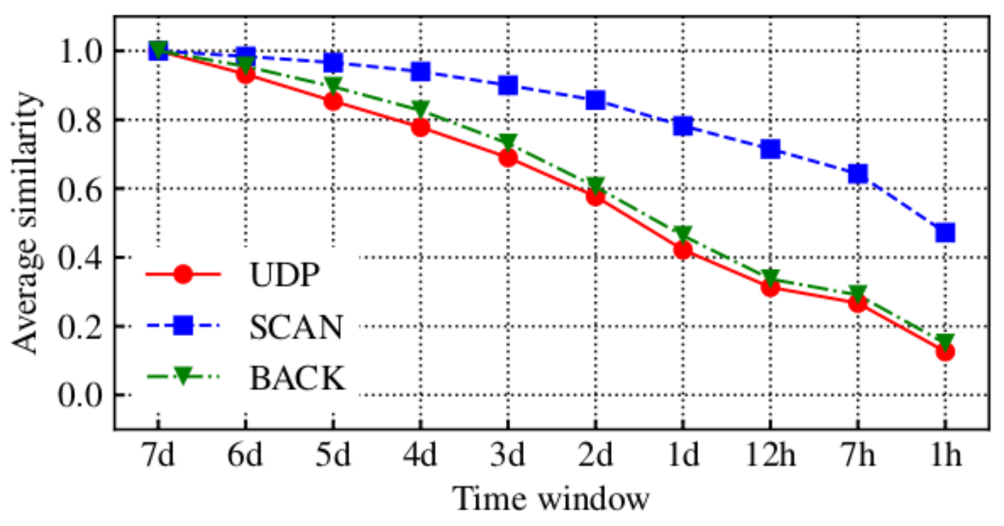


UDP

2. Which darknet size shall we choose for losing as less phenomena as possible, while saving some address space? Which is a reasonable observation time for having a full understanding of what is targeting the darknet?



Impact of the changes in the network size (NL)



Impact of the changes in the observation period (NL)

Future work

- Enrichment of the observed scenario with a ***responder*** honeypot device that answers all the incoming traffic. The responder establishes a connection with the unknown source to have further insights on its behavior, and to allow the collection of potential **attack fingerprints**
- Development of a **machine learning framework** that allows the detection of anomalies in the darknet traffic baseline, analyzed port by port. Such framework takes advantage of common off-the-shelf **supervised and unsupervised Machine Learning algorithms** to have a complete view on the raised alarms, and select only the most relevant ones.

List of attended classes

- 02LWHRV – Communication (16/11/2018, 6.67)
- 01QTEIU – Data science for Networks (15/2/2019, 50)
- 01TYNIY – Latent Variables-based Multivariate Data Analysis for Knowledge Discovery (29/3/2019, 33.33)
- 08IXTRV – Project management (10/4/2019, 6.67)
- 01RISRV – Public speaking (14/1/2019, 6.67)
- 01SWPRV – Time management (13/2/2019, 2.67)
- 01TGRRV – Uso degli strumenti e delle strategie per un efficace uso del tempo (22/3/2019, 5.33)

POLITECNICO
DI TORINO

PhD program in
Electrical, Electronics and Communications Engineering