DATI, AI E ROBOTICA @POLITO

RICERCA, TRASFERIMENTO TECNOLOGICO E SUPPORTO ALLE AZIENDE SUI TEMI FONDAMENTALI DEI BIG DATA, INTELLIGENZA ARTIFICIALE, ROBOTICA E RIVOLUZIONE DIGITALE

AI for Traffic Anomaly Detection

Author: Francesca Soro

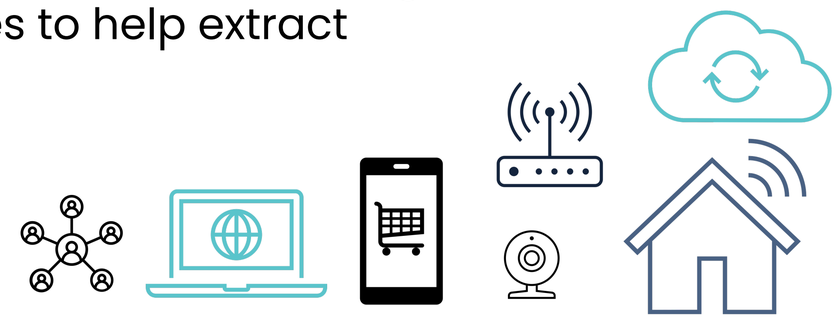
Politecnico di Torino
SmartData@PoliTO

Motivation and background

The **Internet traffic** is continuously growing in both **volume** and **complexity**, and a wider and wider variety of **connected devices** is rising on the market

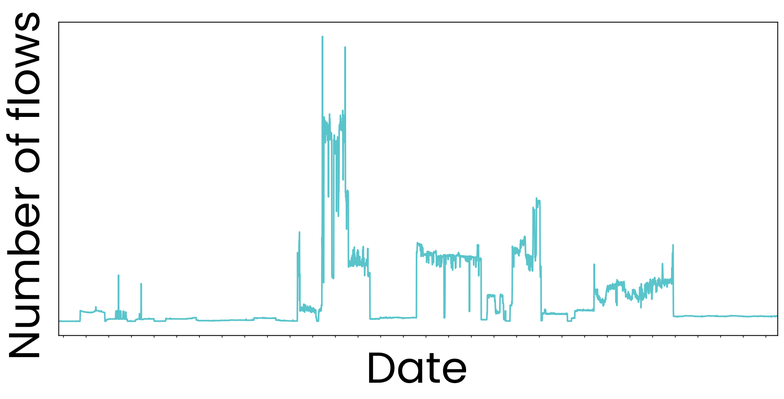
New cyber-attacks showing unseen fingerprints are generated everyday, making the design of an **efficient automatic cybersecurity system** problematic

The large amount of available **raw data** calls for **big-data** and **artificial intelligence** approaches to help extract knowledge



Data collection infrastructure

Unsolicited traffic is a fundamental resource to analyse for **understanding cybersecurity threats**.



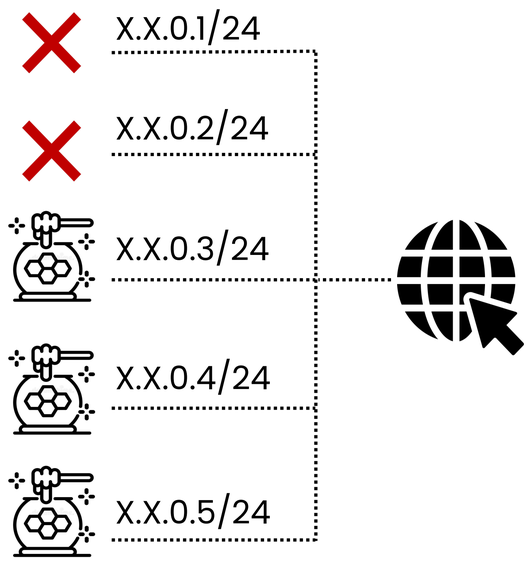
I use two types of infrastructures:

Darknets: IP addresses advertised without hosting any service. They act as **passive sensors** in monitoring activities to highlight phenomena such as **network scans** (both malicious and legitimate), and traffic due to **bugs or misconfigured machines.**

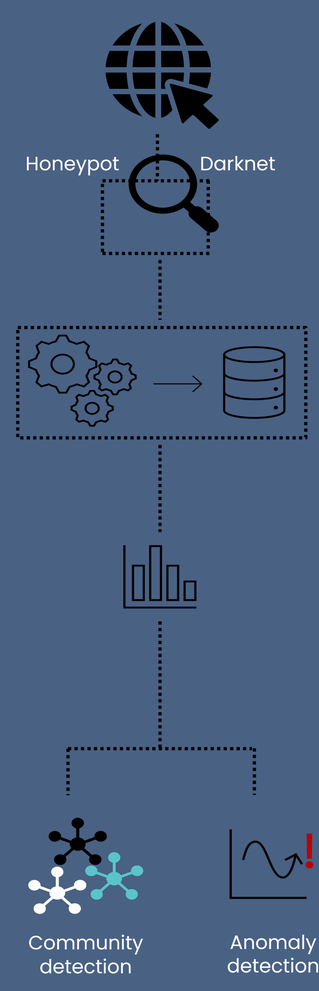
They anyway allow an **intrinsically limited visibility**, as they do not answer traffic requests.

Honeypots: **active sensors** which answer unsolicited traffic to obtain a deeper insight on the attackers' behaviour.

A large number of existing honeypots only answer to **selected vertical services** (HTTP, SSH, RDP, etc.).



Knowledge extraction pipeline



Step 1: Passive traces collection

Capture traffic hitting the infrastructure is captured as **raw .pcap files**.

Step 2: Traces processing and log storage

Process raw data are to obtain a **human-readable log format**, then store such files in a high-end **Hadoop cluster**.

Step 3: Events characterization

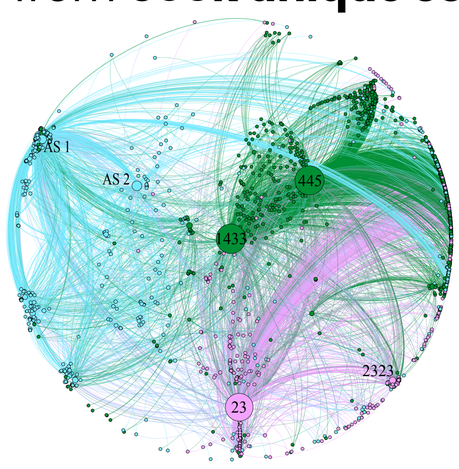
Aggregate logs in a meaningful way to generate **useful features** from the recorded events.

Step 4: Data analytics

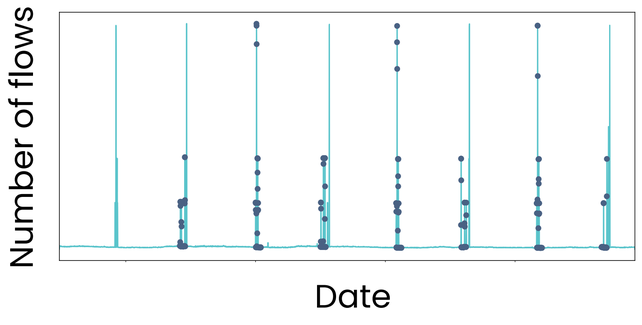
Run **graph mining** algorithms to detect **communities** among remote sources; perform **changepoint detection** in traffic flows.

Results

In a single month (15th April – 15th May 2021) the infrastructure received around **1.3 billion flows** from **600k unique sources**.



Depicting the most active sources as a **graph** based on their activity allows the detection of **coordinated actions** and **possibly malicious botnets**.



New **anomaly detection techniques** allow to mark changepoints and burst events on traffic profiles, raising **alerts on new attacks**

Conclusions and future work

Extracting significant information from unwanted network traffic is a particularly challenging task, due to the high volume of requests reaching the monitoring infrastrucutures everyday. Creating a solid data pipeline – enhanced with AI - is fundamental for recognizing and preventing new threats.