LUCA GIOACCHINI

Consortium GARR | THE ITALIAN EDUCATION & RESEARCH NETWORK

Automatic Detection of Coordinated Events in Darknet Traffic Through Unsupervised Data Mining Techniques

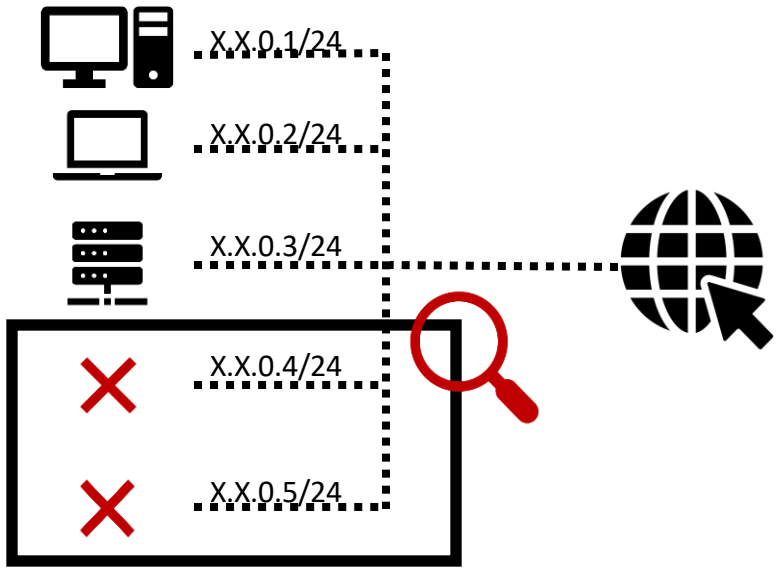
GIORNATA DI INCONTRO
BORSE DI STUDIO GARR
"ORIO CARLINI"
ROMA
21/04/2021

POLITECNICO DI TORINO
SmartData@PoliTO

SmartData@PoliTO

Problem Definition

- Darknets: sets of **passive** IP addresses not hosting any services and saving the received traffic.
- Incoming traffic is unsolicited, thus **anomalous** by definition.
- **Coordinated** source IPs can be a threat (e.g., distributed attacks from botnet).
- Huge amount of received traffic makes impossible manual analysis.



Research Questions

1. How to **automatically detect coordination** among source IPs through darknet traffic analysis?

   a) **Traffic intensity and type**. Hypothesis: Coordinated IPs generate similar traffic volume to similar services.

   b) **Temporal relationships**. Hypothesis: Coordinated IPs follow similar temporal pattern within darknets.

2. If the developed model is able to identify coordination, is it **maintained over time**?

Research Questions

Darknet Data Collection

Research Questions

Darknet Data Collection

Ground Truth Construction

Research Questions

- **Darknet Data Collection**
  - **Ground Truth Construction**
    - **Features Engineering**

Research Questions

Darknet Data Collection

Ground Truth Construction

Features Engineering

IPs Distance Metric Definition

Research Questions

Darknet Data Collection

Ground Truth Construction

Features Engineering

IPs Distance Metric Definition

Ground Truth Separation in Features Space

Research Questions

- Darknet Data Collection
  - Ground Truth Construction
    - Features Engineering
      - IPs Distance Metric Definition
        - Ground Truth Separation in Features Space
          - Clustering Algorithms

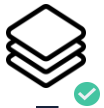
Research Questions

Darknet Data Collection

Ground Truth Construction

Features Engineering

IPs Distance Metric Definition



Ground Truth Separation in Features Space

Clustering Algorithms

Temporal Evolution of Clusters Membership

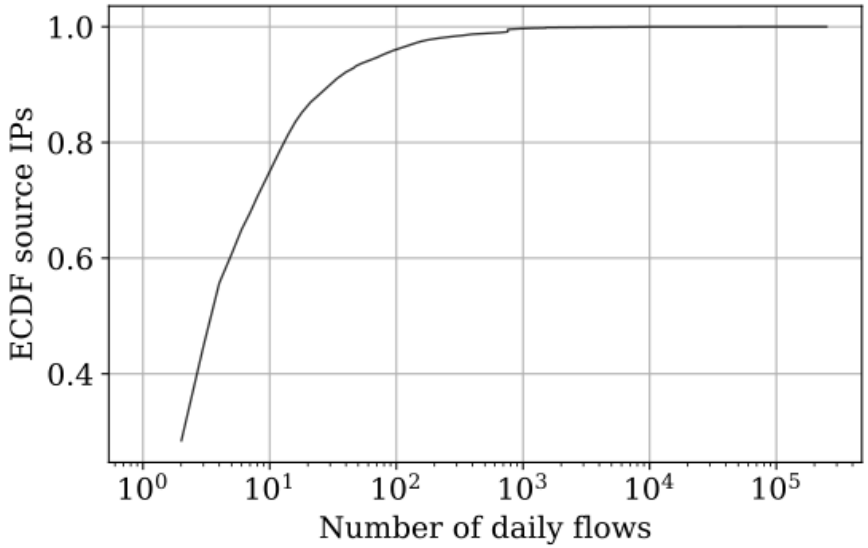
Dataset Overview

Case study: Daily darknet traffic.

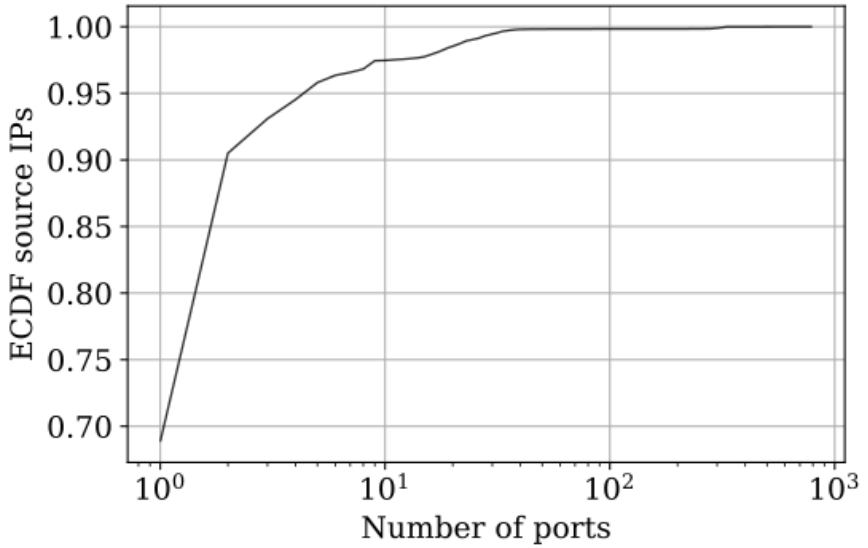
101 008 distinct source IPs.

1 920 058 total flows.

Flows per IP



Reached ports per IP



The most of the traffic is generated by few source IPs.
Only 10% of them reaches more than 1 port.

Ground Truth

Group of Source IPs whose coordination is known **a-priori**.
Classes discovered through Reverse DNS lookup (rDNS).

Label	# Flows	# IPs	# Dst Ports	Top-3 Ports
GT1	16 661	359	23	10443, 4567, 9000
GT2	15 309	49	42	-, 53, 8000
GT3	9 366	12	337	222, 80, 6666
GT4	8 651	5	1	-
GT5	760	10	1	53
GT6	693	1	2	19, 1434
GT7	36	4	1	80
GT8	11	4	2	23, 443

Strong unbalancing among Ground Truth classes.
Some unique behaviors emerge from a first evaluation.

Features Engineering

- Raw features generation:
  General information, protocols, services, etc.

- Features extraction:
  - Principal Components Analysis (PCA);
  - Self-Organizing Maps (SOM);

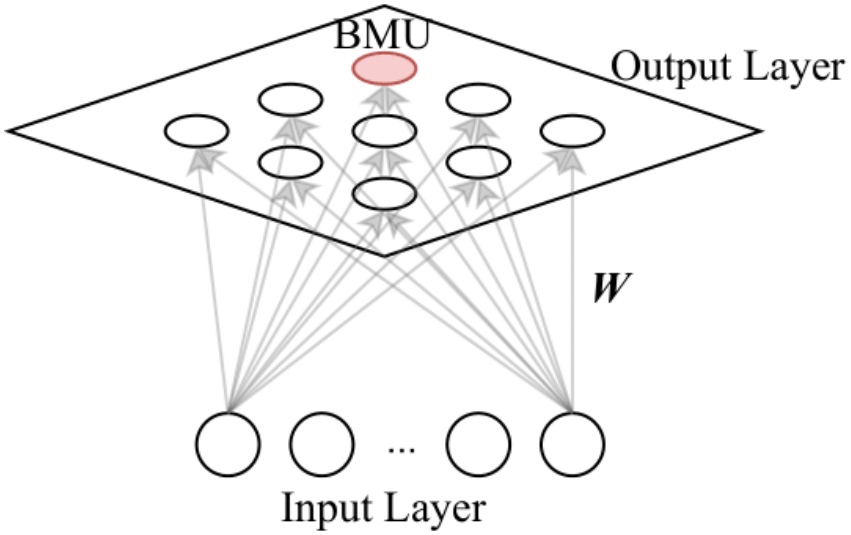
38 Generated Features.

Selected features:
Amount of daily flow per TCP flag,
SOM Best Matching Unit coordinates.

Features Engineering

Self-Organizing Maps

- Two-layers neural network.
- Maps multi-dimensional data onto a **2D neurons grid.**
- For each input sample, its output **winning neuron** is activated in its 2D space.



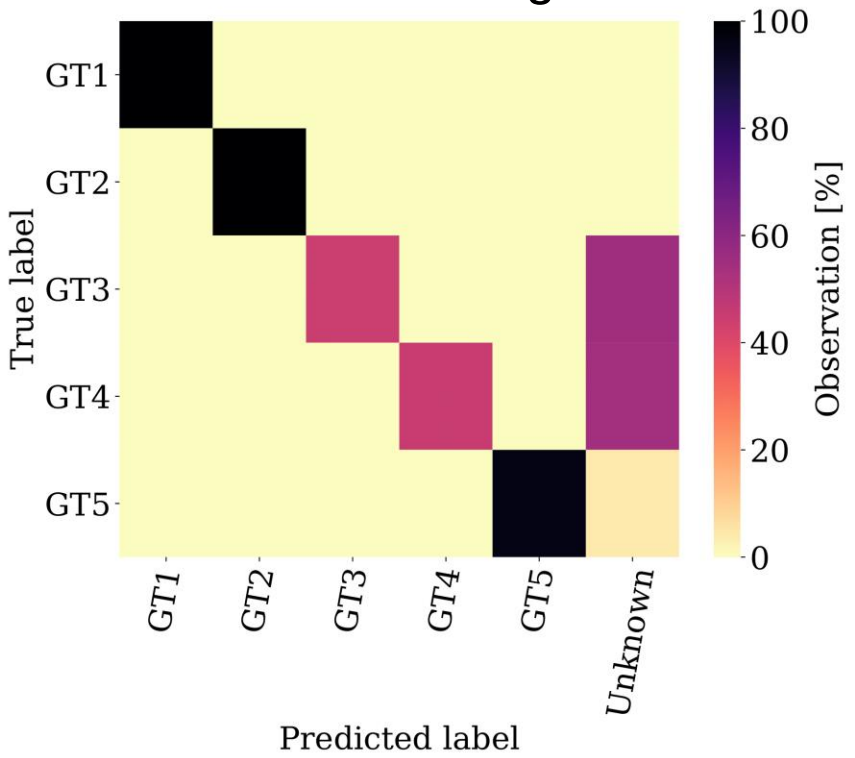
Ground Truth Separation in Features Space

Evaluate if nodes belonging to the same Ground Truth class belong also to the same neighborhood in the features space.

- Leave-One-Out 1-Nearest-Neighbor classifier.
- Tested both single day and over 10 days of traffic.

Ground Truth Separation in Features Space

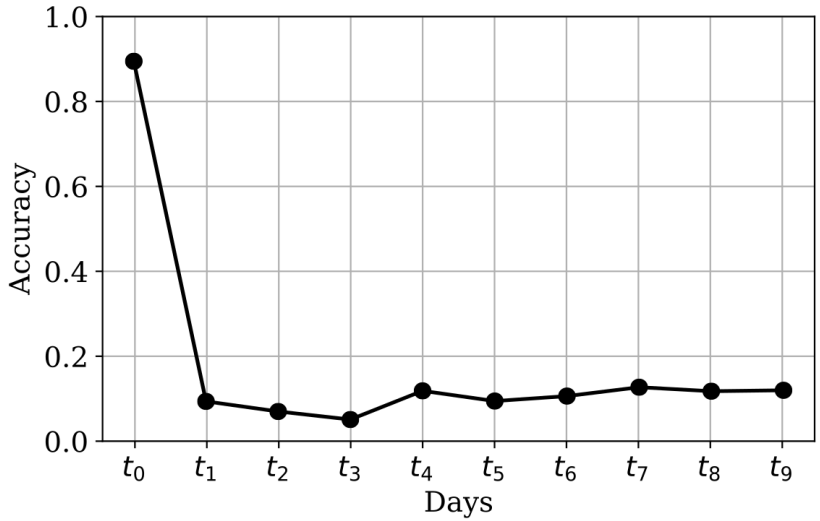
Confusion matrix resulting from the Leave-One-Out 1-Nearest-Neighbor classifier



High accuracy for a single day of traffic: 98.8%

Ground Truth Separation in Features Space

Analyze 9 subsequent days to evaluate consistency of neighborhood in features space over time.



Classifier accuracy evolution over time.
Traffic intensity features

The selected features highlight coordination only for short time ranges. The model fails in 'tracking' rapid traffic changes.

Clustering Algorithms

Algorithms comparison. Each algorithm has its heuristic for the hyperparameters.

Clustering algorithms:

- DBSCAN;
- Hierarchical Agglomerative Clustering;
- k-Means.

Quality metrics: **Silhouette** and number of found clusters.

kMeans best case.
However, Silhouette < 0.5 for all the cases.

Clustering Algorithms

Resulting silhouette for a single day: 0.47

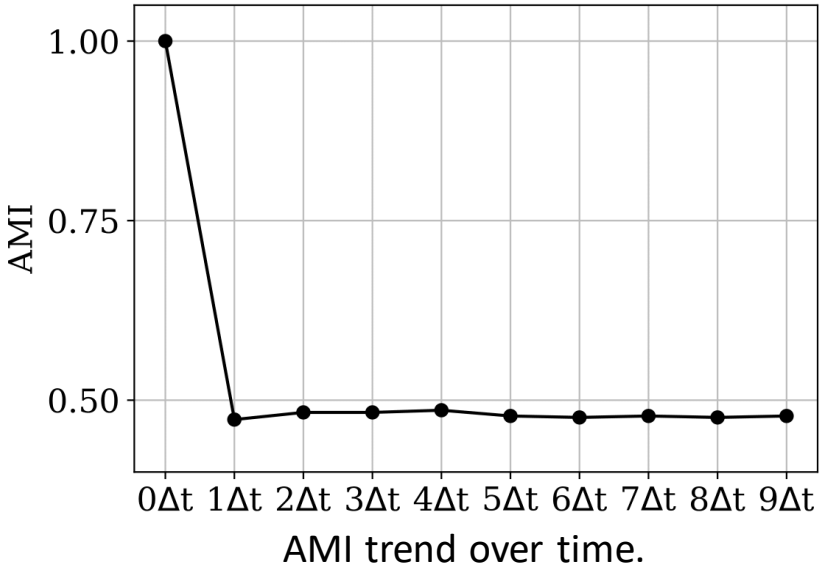
Main outcomes:

- Cluster 1: 100% ICMP traffic;
- Cluster 2: 100% GRE traffic;
- Cluster 3: 68% Telnet traffic;
  24% HTTPS traffic;
  all IPs sent <20 flows.

Some similar behaviors successfully detected w.r.t. protocols and level of traffic volume.

Temporal Evolution of Clusters Membership

Run kMeans for **10 subsequent days**.
Quality metric: **Adjusted Mutual Information** between each day and the first one.



AMI trend over time.

The algorithm is not able to spot the nodes similarities for the considered case.

Conclusions

- **Traffic types and intensity** information highlight coordination among ground truth classes only for **short-time** ranges;
- kMeans detects similar behaviors on the basis of protocols and level of traffic volume;
- kMeans **fails** in detecting coordination for long-time ranges.

Future Works

- Ground Truth expansion;
- Investigating if **temporal relationships** among source IPs can highlight coordination;
- Improving performance achieving long-time ranges robustness;
- Testing **other** clustering/community detection algorithms.

Thank you for your attention.

Questions?

Luca Gioacchini