SPADA: SPark Anomaly Detection Ace

Antonia Affinito*, Alessio Botta*+, Luigi Gallo*, Mauro Garofalo*, and Giorgio Ventre*+

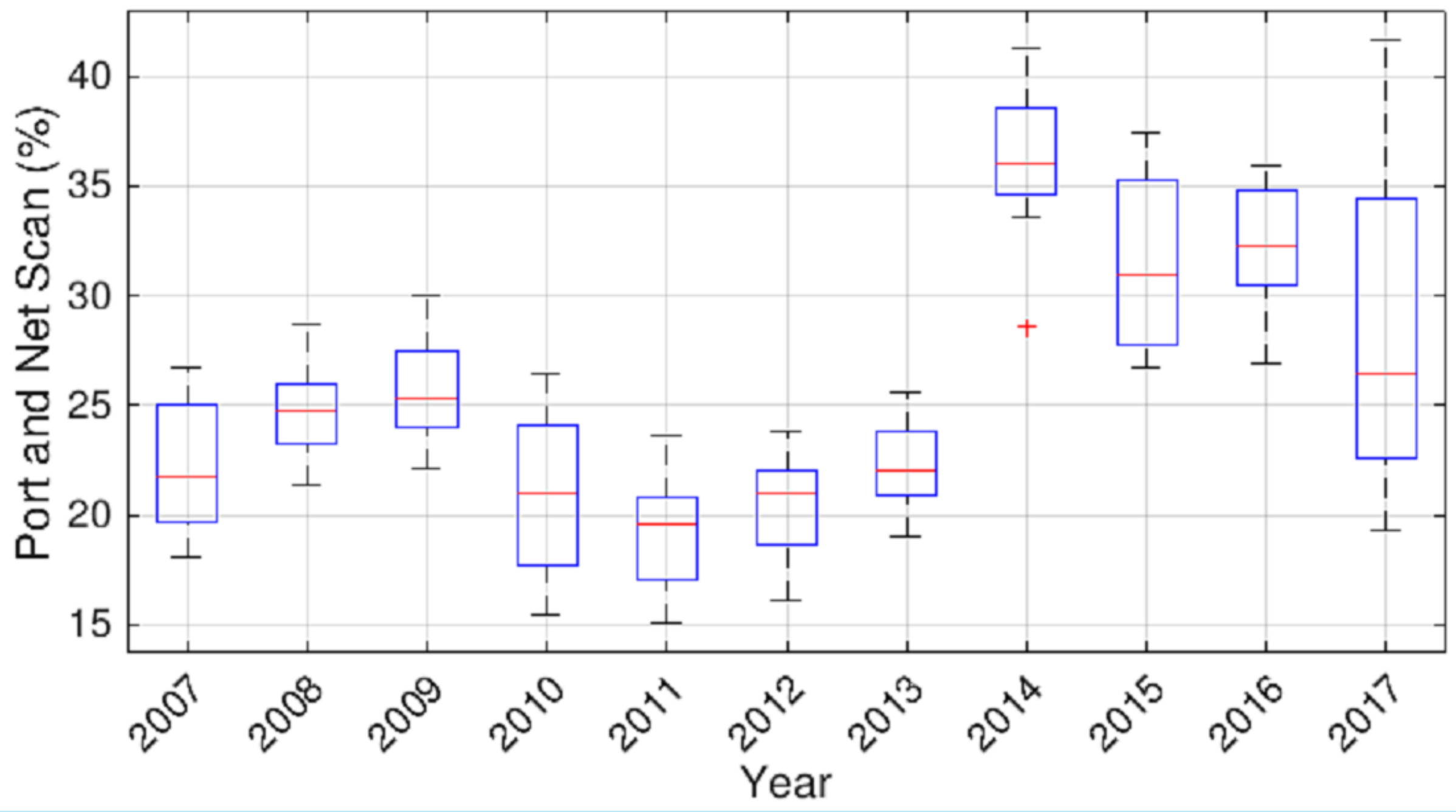
*Department of Electrical Engineering and Information Technologies, University of Napoli “Federico II” – Italy
+ NM2 SRL, Italy

{antonia.affinito, luigi.gallo10}@studenti.unina.it {a.botta, mauro.garofalo, giorgio}@unina.it

INTRODUCTION AND MOTIVATION

- An important category of network anomaly detection are port and net scan
  - The percentage of attacks and anomalous event in network traffic is constantly growing (see Figure 1)
- We present an approach to detect anomalies in high-speed networks working at flow-level
  - We use Apache Spark to cope with the problem of the large amount of data to be analyzed
  - We implement a simple threshold-based detection algorithm in Spark and test it by using several real traces

Figure 1: Malicious activities during the last 11 years



ALGORITHM

- We consider the ratio between the number of flows generated and received by the same IP address together with other important features

$$\frac{FlowS}{FlowD}+\frac{\alpha}{PKTF}-\frac{BPP}{\gamma}-\beta\frac{FlowS}{IPC}>TH$$

- FlowS: number of generated flows;
- FlowD: number of received flows;
- PKFT: average number of packets per flow;
- BPP: average number of bytes per packet;
- TH: threshold.

TOOLS

- A platform for distributed processing of Big Data [2]
- Very fast both in storage and data processing because of *in-memory* processing

MAWI (Measurement and Analysis of the Wide Internet)

- An archive of traces of real traffic provided by the MAWI Working Group [4]
- Traffic captured every day from 14:00 to 14:15 on a transoceanic link

MAWILab

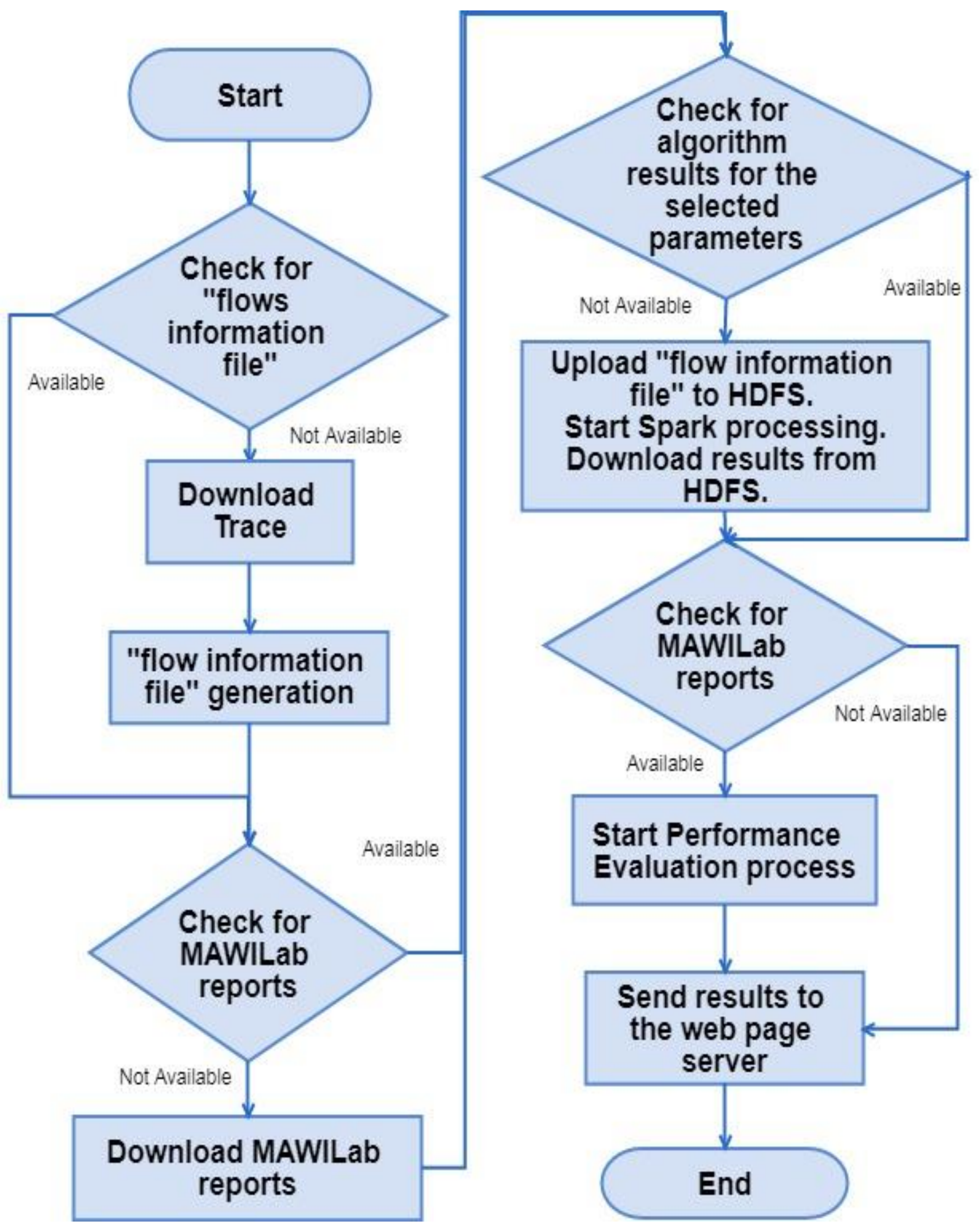
- An approach for the identification of network anomalies in MAWI [3]
- Uses four detectors: Principal Component Analysis (PCA), Gamma distribution, Kullback Leibler (KL) divergence, and Hough transformation

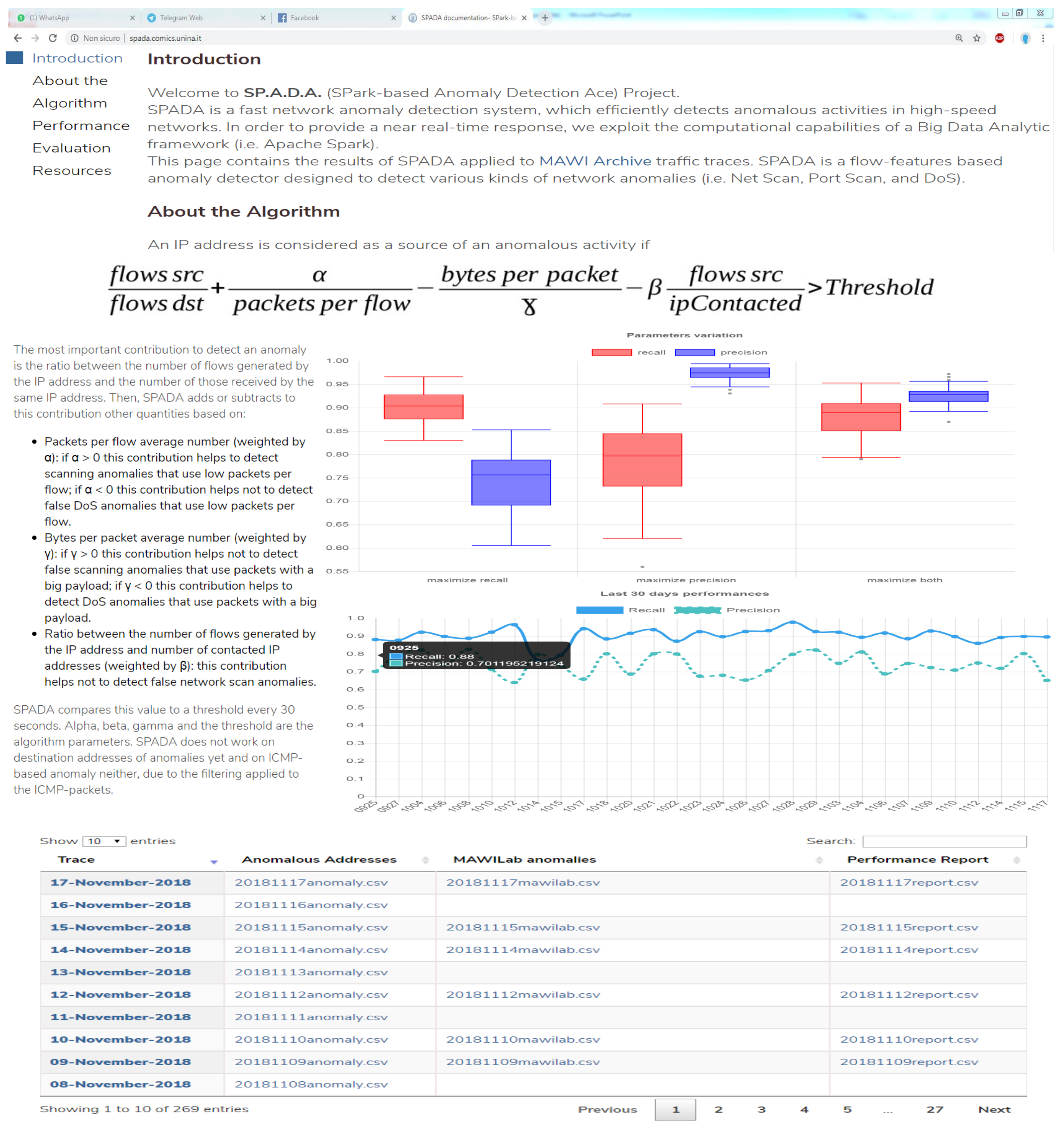
SPADA

SPADA is a system able to run automatically all operations needed for the analysis of a traffic trace. Every day

- It downloads the trace to be analyzed
- It generates a flow file using the TIE tool
- It run the algorithm on SPARK
- It publishes the results of the comparison on web portal

Visit us at:
spada.comics.unina.it


Start
Check for "flows information file"
Available
Not Available
Download Trace
"flow information file" generation
Available
Check for MAWILab reports
Not Available
Download MAWILab reports
Check for algorithm results for the selected parameters
Not Available
Available
Upload "flow information file" to HDFS. Start Spark processing. Download results from HDFS.
Check for MAWILab reports
Not Available
Available
Start Performance Evaluation process
Send results to the web page server
End



Introduction

Welcome to **SP.A.D.A.** (SPark-based Anomaly Detection Ace) Project.
SPADA is a fast network anomaly detection system, which efficiently detects anomalous activities in high-speed networks. In order to provide a near real-time response, we exploit the computational capabilities of a Big Data Analytic framework (i.e. Apache Spark).
This page contains the results of SPADA applied to MAWI Archive traffic traces. SPADA is a flow-features based anomaly detector designed to detect various kinds of network anomalies (i.e. Net Scan, Port Scan, and DoS).

About the Algorithm

An IP address is considered as a source of an anomalous activity if

$$\frac{flows\ src}{flows\ dst}+\frac{\alpha}{packets\ per\ flow}-\frac{bytes\ per\ packet}{\gamma}-\beta\frac{flows\ src}{ipContacted}>Threshold$$

The most important contribution to detect an anomaly is the ratio between the number of flows generated by the IP address and the number of those received by the same IP address. Then, SPADA adds or subtracts to this contribution other quantities based on:

- Packets per flow average number (weighted by α): if α > 0 this contribution helps to detect scanning anomalies that use low packets per flow; if α < 0 this contribution helps not to detect false DoS anomalies that use low packets per flow.
- Bytes per packet average number (weighted by γ): if γ > 0 this contribution helps not to detect false scanning anomalies that use packets with a big payload; if γ < 0 this contribution helps to detect DoS anomalies that use packets with a big payload.
- Ratio between the number of flows generated by the IP address and number of contacted IP addresses (weighted by β): this contribution helps not to detect false network scan anomalies.

SPADA compares this value to a threshold every 30 seconds. Alpha, beta, gamma and the threshold are the algorithm parameters. SPADA does not work on destination addresses of anomalies yet and on ICMP-based anomaly neither, due to the filtering applied to the ICMP-packets.

Trace	Anomalous Addresses	MAWILab anomalies	Performance Report
17-November-2018	20181117anomaly.csv	20181117mawilab.csv	20181117report.csv
16-November-2018	20181116anomaly.csv		
15-November-2018	20181115anomaly.csv	20181115mawilab.csv	20181115report.csv
14-November-2018	20181114anomaly.csv	20181114mawilab.csv	20181114report.csv
13-November-2018	20181113anomaly.csv		
12-November-2018	20181112anomaly.csv	20181112mawilab.csv	20181112report.csv
11-November-2018	20181111anomaly.csv		
10-November-2018	20181110anomaly.csv	20181110mawilab.csv	20181110report.csv
09-November-2018	20181109anomaly.csv	20181109mawilab.csv	20181109report.csv
08-November-2018	20181108anomaly.csv		

Showing 1 to 10 of 269 entries

The system allows to

- Detect anomalies on high-speed networks;
- Carry out longitudinal analysis on the anomalies in one of the most used trace reports
- Provide to the scientific community an always-updated reference for comparison

PRELIMINARY RESULTS

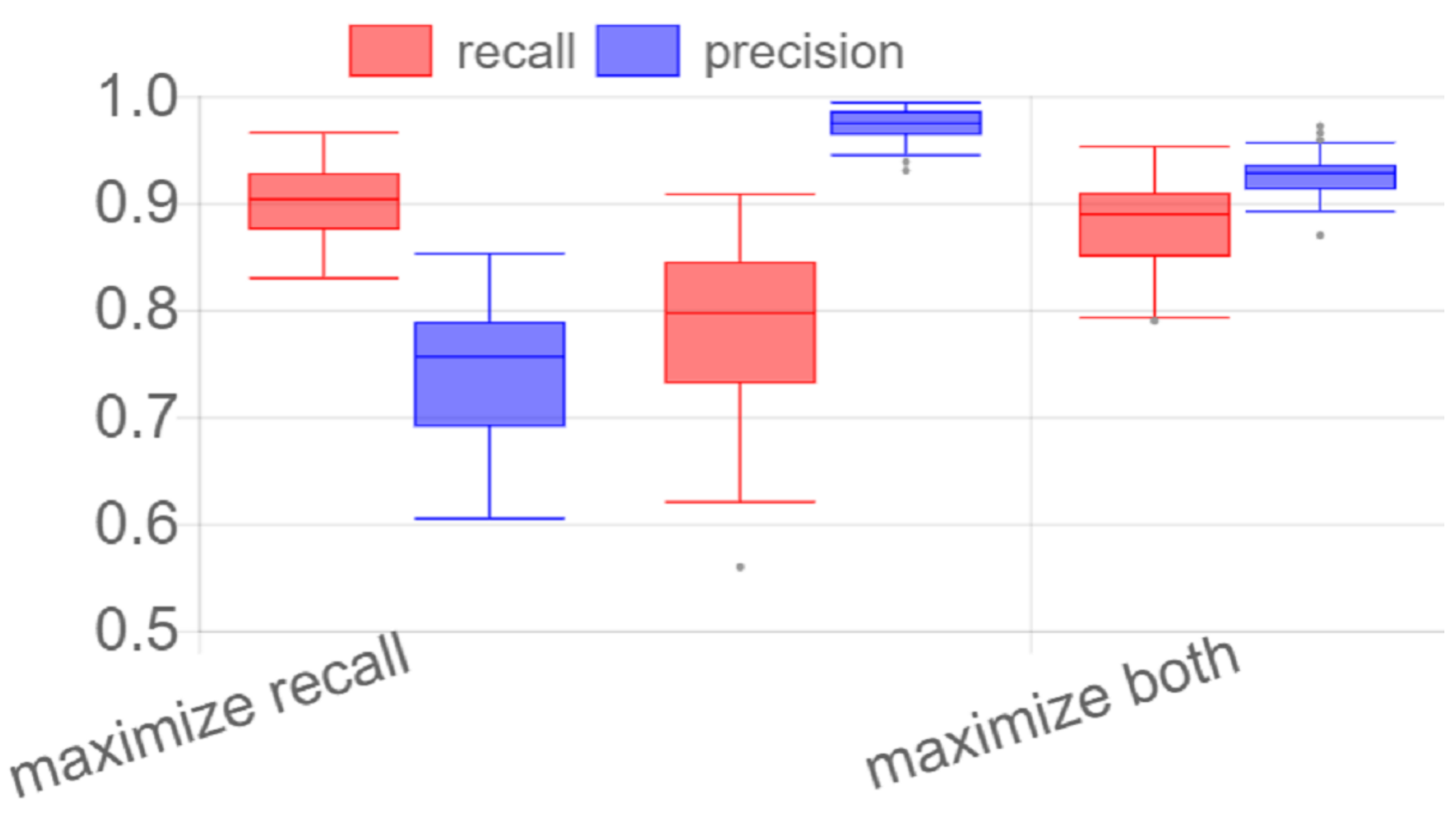


Figure 2: maximizing recall and/or precision

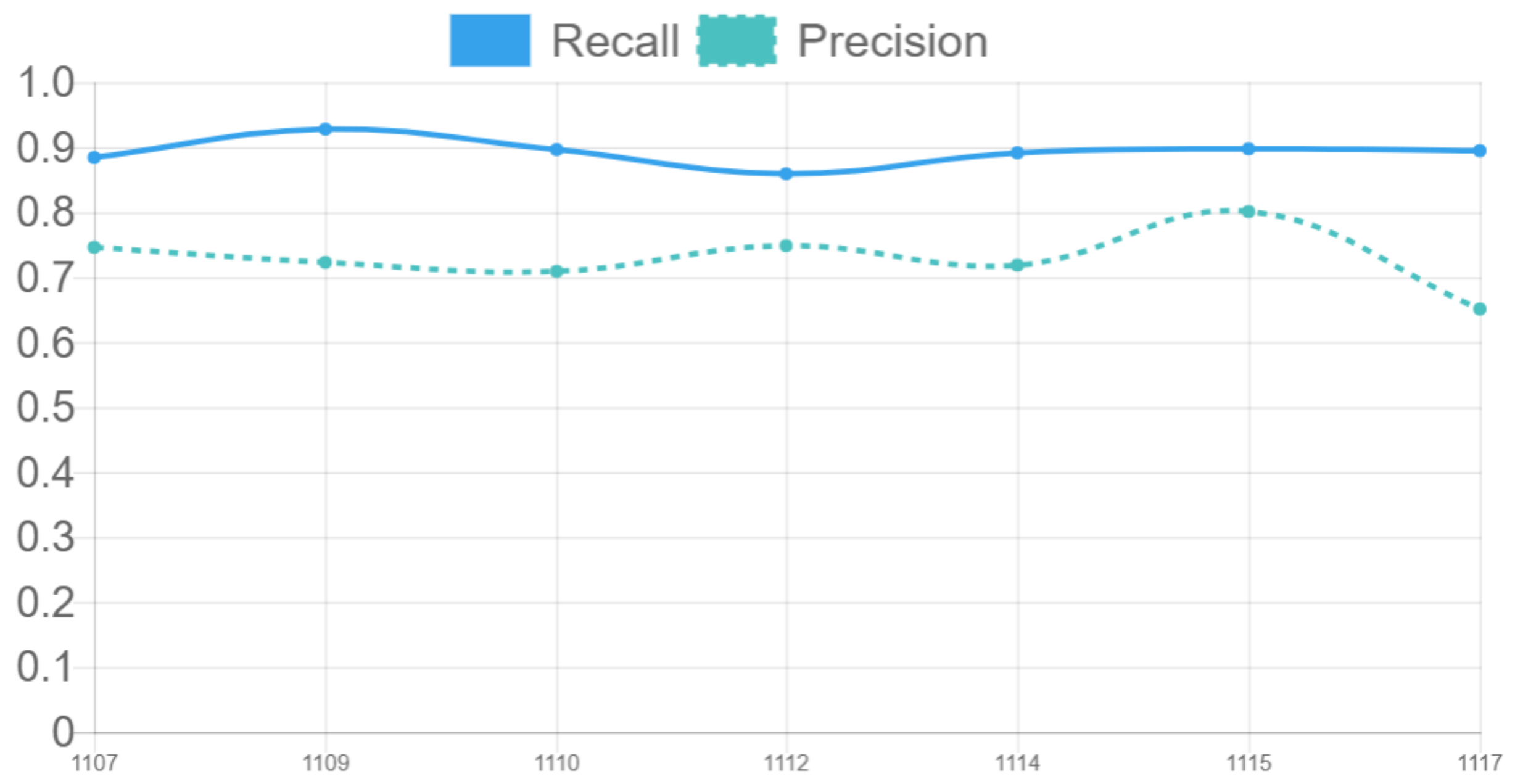


Figure 3: Last 7 days of SPADA

- Figure 2 and 3 show SPADA performance measured on 100 traffic traces with different values of the parameters
- The algorithm is more effective in detecting malicious scanning activity than MAWILab
- It allows to obtain a new ground truth starting from MAWILab

Acknowledgement

This work is partially funded by Cisco for University of Napoli and by MIUR art. 11 DM 593/2000 for NM2 SRL.

References

[1] SPADA documentation- SPark-based Anomaly Detection Ace. http://spad7a.comics.unina.it/
[2] Big data analytics on apache spark- https://link.springer.com/content/pdf/10.1007%2Fs41060-016-0027-9.pdf
[3] MAWILab – Home- http://www.fukuda-lab.org/mawilab/
[4] P. Borgnat, G. Dewaele, K. Fukuda, P. Abry, and K. Cho. Seven years and one day: Sketching the evolution of internet traffic. In IEEE INFOCOM April 2009
[5] Casas