Segurança em Redes de Comunicações
Security in Communications Networks
Second Report

Professors:
Paulo Salvador salvador@ua.pt;
António Nogueira nogueira@ua.pt;

Objective: Based on data log of IP traffic flows, define SIEM rules to detect anomalous network behaviors and possibly compromised devices.

Description:

A corporate network has a SIEM system with the historic data of traffic flows on the network. To implement a reliable Cybersecurity system it requires the implementation of alert rules, of possible attacks, based on anomalous behaviors.
Consider the dataset (datasetX.zip file) with files dataX.parquet and testX.parquet, where X is the remainder of the division of the sum of the student numbers by 10:

in python: X=(num_mec1+num_mec2) % 10

Using data from one full day (file dataX.parquet) define the typical behavior of the network devices. This data was already fully analyzed and no illicit behavior was detected. You may assume that the IPv4 private address of each device does not change over time and is assigned to the same end-user.
The file testX.parquet contains data from a fully day and may contain anomalous behaviors resulting from illicit activities within the network, such as internal botnet activities, data exfiltration, and remote C&C of devices.
The *.parquet data files contain the list of all observed IPv4 data flows with the following information about each flow (columns):

- timestamp: time of observation of the first packet of the flow, in 1/100 of seconds from 0h of the day;
- src_ip: IPv4 source address;
- dst_ip: IPv4 destination address (internal or external);
- proto: transport protocol used (tcp or udp);
- port: destination port;
- up_bytes: total of uploaded byes;
- down_bytes: total of downloaded bytes.

Data is structured using pandas, and stored in parquet format. See: https://pandas.pydata.org/ and https://parquet.apache.org/ . Check the provided python script (sampleScript.py) with basic examples on how to read and process the data file. Geo-localization based on the IPv4 address must rely on external databases (GeoIP_DBs.zip). Check also the provided python script with with basic examples on how to perform IP Geo-localization and DNS queries.

- Present a report of the data analysis, proposed SIEM rules and rule tests (anomalous device detection).
- Submit via e-learning, in format PDF, until June 8th.
- Should be done by a group of 2 students. Exceptionally, can be done individually.
- Tasks:
  - Non-anomalous behavior analysis and description (4 points).
  - Anomalous behavior detection, description, and possible causes (4 points).
  - SIEM rules (4 points).
  - SIEM rules test and identification of the devices with anomalous behaviors (4 points).
  - Written report; structure and content (4 points).