Segurança em Redes de Comunicações
Security in Communications Networks
Second Project

Professor:
Paulo Salvador salvador@ua.pt;

Objective: Define SIEM rules to detect anomalous network behaviors and possibly compromised devices. Test the defined rules on a data log of IP traffic flows, identifying the compromised devices.

Description:
A corporate network has a SIEM system with the historic data of traffic flows on the network. Data was received from the main firewalls logging system. To implement a reliable Cybersecurity system it requires the implementation of alert rules, of possible attacks, based on anomalous behaviors.

In order to simplify data processing, data was exported to external files in parquet format. Rules should be implemented and tested using a python script. **You do not have to implement a formal SIEM.**

Consider the dataset (datasetX.zip file) with files dataX.parquet, testX.parquet and serversX.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 internal 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 the internal devices, from a full day and may contain anomalous behaviors resulting from illicit activities, such as internal botnet activities, data exfiltration, and remote C&C of devices.
The file serversX.parquet contains data from a full day of external accesses to the corporation servers (in network 200.0.0.0/24) from a small set of clients in the same external network, and may contain external users interacting with the corporation servers in an anomalous way (tip: it is not the amount of traffic or flows).
Each *.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 (for dataX and testX files identifies the internal device, for the serversX file identifies the external client);
- dst_ip: IPv4 destination address (identifies the external or internal server);
- proto: transport protocol used (tcp or udp);
- port: destination port;qq
- up_bytes: total of uploaded byes;
- down_bytes: total of downloaded bytes.

IMPORTANT NOTE: all public IPv4 addresses represent real networks, but (besides the flow statistics) only the owner and location are relevant. The real purpose/services of the same are not relevant! **DO NOT PERFORM SERVICE/VULNERABILITY SCANS ON THE IPv4 ADDRESSES!**

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 the external GeoLite2 databases (GeoLite2-ASN and GeoLite2-Country). Download both databases and place them on the same directory as the scrip/data files. Check also the provided python script with with basic examples on how to perform IP Geo-localization and DNS queries.

- Present a report with the proposed SIEM rules and rule tests (identifying the devices with anomalous behaviors). Submit via e-learning, in PDF format, until June 8th. Should be done by a group of 2 students. Exceptionally, can be done individually.

- Tasks:
  - Analysis of the non-anomalous behaviors: (i) identify the network private IPv4 network(s), (ii) identify internal server/services, (iii) describe and quantify traffic exchanges (upload/download statistics, ratios, and destination countries) from internal users with internal and external servers, and (iv) describe and quantify traffic exchanges (overall, ratios and interval between flows) from external users with the corporation public servers (4 points).
  - Definition of the SIEM rules and respective justification for detection of: (I) internal BotNet activities, (ii) data exfiltration using HTTPS and/or DNS, (iii) C&C activities using DNS, (iv) anomalous external destinations, and (v) external users using the corporate public services in an anomalous way. **Rules must be well defined based on measured values from data.** (6 points).
  - Test of the SIEM rules and identify the devices (IP addresses) with anomalous behaviors. **List the IP addresses of the devices identified as anomalous**. (6 points).
  - Written report; structure and content (4 points).