AD4IDS Sous-projet 1
Flow Classification

P-F. Marteau, ENSIBS

October 2023

1 Data loading

The ISCX files contain detailed flow information in XML format for each day. The flows have been generated (from the packet files (pcap)) using IBM QRadar appliance. The ”Tag” column indicates whether the flow is normal or part of an attack scenario.

The XML files contain the following: ”appName”, ”totalSourceBytes”, ”totalDestinationBytes”, ”totalDestinationPackets”, ”totalSourcePackets”, ”sourcePayloadAsBase64”, ”destinationPayloadAsBase64”, ”destinationPayloadAsUTF”,”direction”, ”sourceTCPFlagsDescription”, ”destinationTCPFlagsDescription”, ”source” ,”protocolName” ,”sourcePort”, ”destination” ,”destinationPort”, ”startDateTime”, ”stopDateTime”, ”Tag”

The ISCX data is available as a bunch of XML files :

- TestbedSatJun12Flows.xml
- TestbedSunJun13Flows.xml
- TestbedMonJun14Flows.xml
- TestbedTueJun15-1Flows.xml
- TestbedTueJun15-2Flows.xml
- TestbedWedJun16-1Flows.xml
- TestbedThuJun17-1Flows.xml
- TestbedWedJun16-2Flows.xml

- TestbedThuJun17-2Flows.xml

Write a Python program to load the data. The lxml library can be use to parse the XML and to convert each flow element into a Python dictionary structure. Thus, each file will be converted into a list (sequence) of dictionaries, each dictionary corresponding to a single flow.

2 Data indexing

As the data potentially will not fit into the RAM of your machine, we will process the XML files one by one, and use a hashtable on the Hard drive to store it. The shelve Python library can be use to implement this function, or more efficiently the Elasticsearch component.

Warning, the origin of each flow, namely the file from which it has been extracted, needs to be stored as well.

3 Data access

Write a set of access functions (API) to get elementary and aggregated information from the stored flows:

In particular write functions to:

- get the list of all the (distinct) protocols contained in the XML files
- get the list of flows for a given protocol
- get the number of flows for each protocols
- get the source and destination Payload size for each protocol
- get the total source/destination Bytes for each protocol
- get the total source/destination packets for each protocol
- get the list of all the (distinct) applications contained in the XML files
- get the list of flows for a given application
- get the number of flows for each application
- get the source and destination Payload size for each application

- get the total source/destination Bytes for each application
- get the total source/destination packets for each application
- etc.

Draw the "ranked" distribution #Flows v.s. #Packets, from the largest flow on the left to the smallest to the right. Use standard and log-log axis, and the matplotlib library.

4 Data preprocessing

To use common machine learning algorithms, in particular neural networks, we will need to convert all the so-called categorical fields (features) that describe the flow into numerical data.

For instance the application or protocol fields are string data. Propose some conversion procedure to encode all the categorical fields into numerical fields.

This question is not necessarily obvious to answer to, hence ask the Prof. to raise a discussion here!

5 Deliverable

Put your well commented code and a mini documentation (user install/manual + testing code) in a zip file and deposit it into the "Rendu" folder corresponding to "Sous-projet 1" on the ENT/Moodle dedicated to the IA&DA course.

Do not forget the list of licensed code you have reused if any!