Machine learning cybersecurity

**pcap file feature extraction**

## LAB 1: Writing a script to Extract features from pcap file

**Lab Description:** This lab is to extract features from pcap files in order to represent the raw data in the vector space model.

Example of pcap file:



* No.: The number of the packet in the capture file

Time: The timestamp of the packet.

Source: The address where this packet is coming from.

Destination: The address where this packet is going to.

Protoco:l The protocol name in a short (perhaps abbreviated) version.

Length: The length of each packet.

* You are required to write a python script to extract features from pcap files and represent them in the vector space model.
* The labels will be assigned based on the type of devices that send the packets.
* Since we have 5 devices (assistant, camera, miscellaneous, mobile and outlet), we will create 5 labels or classes.
* We can judge the device type by the source ip address of the package:
	+ Each ip address of the device is 192.186.1.#, and # is replaced by:

Assistant: 111,30,42,59,70

Camera: 128,145,78

Miscellaneous: 216,46,84,91

Mobile: 45

Outlet: 222,67

* We will extract 18 features from each of the package from the 5 devices. The 18 features are:



**Lab Environment:**

* The students should have access to a machine with Linux
* The environment for python is required as well as some libs such as numpy, tensorflow, pandas and sklearn.
* We will also need tshark.

Installation command: sudo apt install tshark

**Lab Files that are Needed:**

* For this lab you will need several pcap files, which are:

tcpdump-20181001-0029.pcap

tcpdump-20181001-0129.pcap

…

### Lab exercise 1

* Import the required libraries.



* Define a dictionary to store the string for each device you want to analyze.



* + ip\_filter is the name of the dictinoary
	+ 'TCP\_Outlet' is the key of the filtering string for a small power outlet device
	+ ' tcp && (ip.src==192.168.1.222) || (ip.src==192.168.1.67)' means to filter the tcp package from either 192.168.1.222 or 192.168.1.67, which are the ip addresses of outlet.
	+ You are required to define filtering strings for the other 3 devices (Assistant, Camera, Miscellaneous)
	+ Please note the single quotes and double quotes in the string
* Open a csv file to store the labels and features. The header should be the name of features and the first column should be the label.



* label\_feature\_small.csv is the name of the file
* 'a' means append data to an existing file
* You need to write string 'label' and 18 features names into the file
* Filter out all the packets from the 5 different devices in the original pcap files and save the result into 5 new pcap files.



* glob.glob will process all the pcap files in the original\_pcap folder
* -r means to read the local pcap file
* oriPcapFile is one of the pcap files in the original folder
* -w- and -Y means to write the packets matching the filter to the specified file
* ip\_filter[k] is the filtering command for the device k
* >> will append the filtering results to the pcap files in the filtered\_pcap folder
* The pcap files are named with k
* Create the labels and extract the features for them with the newly generated pcap files.



* glob.glob will get all the pcap file names in the filtered\_pcap folder
* The names of the pcap files will be used to create the label
* Use tshark command to extract features from the currently processed file
* -r means to read the local pcap file
* ftdPcapFile is the name of the currently processed file
* Use several -e options for each feature (field) that you want to extract from the pcap file, such as -e ip.len
* -T option should be selected if you want to use -e option
* os.popen(tsharkCommand).read() will excute the command and save the result to the allFeatures parameter
* Before writing the label and features to the csv file, you need to convert the tab-separated results to the comma-separated results and breaking the results at line boundaries.

## What to Submit

You should submit a lab report file which includes:

* + The steps for how you processed data
	+ The necessary code snippet of your feature extraction script.
	+ The screenshot of the results
	+ You can name your report "Lab\_pcap\_feature\_extraction\_yourname.doc".