Weekly Learning Log

In order to demonstrate how much, I have learnt over each week, I will write an entry in the learning log each week.

The Challenge

The official challenge specification is as follows:

“Create **a raspberry pi, npcap and wireshark** based **packet sniffing network monitor** that can be attached to **wired and wireless networks** and **passively monitor network communications**. Time permitting explore how anomalous transmissions can be detected using machine learning based modelling of normal network behaviour. This project also requires an analysis of the data protection and privacy implications of the technology.”

Re written in a more understandable way for myself:

“Enable a raspberry pi to sniff packets on the network (wired/wireless). Then write code to analyse the packets in order to provide insights into the network. Time permitting enhance this analysis to include machine learning techniques. This project also includes looking into the legality of the device and the implications it has on peoples privacy.”

Week Of 16th-23rd January

**Unknown Terminology**

QEMU

“Flash” a Pen Drive/Micro USB

BIOS

GNU-GRUB bootloader

Sudo

NPCap

NMap

KVM

**Meaning**

QEMU: QEMU is a generic and open source machine emulator and virtualizer. I will use QEMU to emulate a raspberry pi.

“Flash” a Pen Drive/Micro USB: Flashing a Pen Drive/Micro USB is essentially formatting the memory device so that it can store an operating system.

BIOS: Stands for Basic input/output system. Used by the microprocessor to boot the operating system on start-up. It also manages data flow between the computer's operating system and attached devices such as the hard disk, video adapter, keyboard, mouse and printer.

Sudo: On Unix-like operating systems, the sudo command (sudo stands for "superuser do") allows a user with proper permissions to execute a command as a superuser(administrator). When typed into the console, the console will prompt you for the root password.

NPCap: NPCap is the Nmap Project's packet sniffing (and sending) library for Windows. Is a part of a wider application called Nmap.

NMap: Nmap ("Network Mapper") is a free and open source utility for network discovery and security auditing. Uses NPCap.

KVM: Kernel-based Virtual Machine (KVM) is an open source virtualization technology built into Linux. It allows for a computer running Linux to create and run virtual machines. KVM virtualisation is said to be the fastest, with only a roughly 4.5% drop in efficiency between the emulated version and the stand alone device being emulated.

**Useful Links**

USB Flashing Software Used: <https://rufus.ie/>

NPCap Download Link: <https://nmap.org/npcap/>

Ubuntu Dual Boot Tutorial: <https://www.youtube.com/watch?v=u5QyjHIYwTQ>

**Comments**

I have downloaded and used a lot of different operating systems for this project. I enjoy working with them after getting over my initial worries.

Week Of 24th-31st January

**Unknown Terminology**

Network Adapter

Structured Variable

Code Tag

**Meaning**

Network Adapter: A network adapter is the component of a computer's internal hardware that is used for communicating over a network with another computer. NPCap “listens” on one or more of these devices in order to sniff packets on the network.

Structured Variable: In C programming, a struct (or structure) is a collection of variables (can be of different types) under a single name. They are similar to dictionaries in other languages. They can be used a pseudo objects however C is not designed for OOP, so this can lead to the program being too complex or inefficient.

Code Tag: The HTML <code> tag defines a fragment of computer code in the HTML document.

**Useful Links**

W3 Schools Code Tag: <https://www.w3schools.com/tags/tag_code.asp>

C Structs:

<https://www.tutorialspoint.com/cprogramming/c_structures.htm>

HTML Character Entities Sheet: https://dev.w3.org/html5/html-author/charref

**Comments**

Not too many new terms in this entry as the majority of the time was spent on my tutorial web page. The character entities of HTML came in useful whenever I wrote code examples.

Week Of 31th- 7th February

**Unknown Terminology**

WinSCP

TShark

TermShark

SFTP

DoS attack

SYN DoS attack

Scapy

**Meaning**

WinSCP: WinSCP is an open source free SFTP client, FTP client, WebDAV client, S3 client and SCP client for Windows. Its main function is file transfer between a local and a remote computer. In my project I have used it to create an SSH connection in order to freely send and receive files from it.

TShark: A terminal version of Wireshark. I use this on the raspberry pi as it is less taxing on the computer’s resources than Wireshark GUI version.

TermShark: A terminal version of Wireshark, however unlike TShark, TermShark provides a GUI in the terminal window itself.

SFTP: Stands for Secure File Transfer Protocol. It is a separate protocol packaged with SSH that works in a similar way over a secure connection.

DoS: a cyber-attack in which the perpetrator seeks to make a machine or network resource unavailable to its intended users by temporarily or indefinitely disrupting services of a host connected to the Internet.

SYN DoS attack: A SYN flood is a form of denial-of-service attack in which an attacker sends a succession of SYN requests to a target's system in an attempt to consume enough server resources to make the system unresponsive to legitimate traffic.

Scapy: Scapy is a packet manipulation tool for computer networks, originally written in Python. It can forge or decode packets, send them on the wire, capture them, and match requests and replies.

**Comments**

I learnt a lot about networking this week. I feel as though I have made a lot of progress, as I can now capture packets using the raspberry pi. I found using WinSCP very difficult and have yet to get the SFTP server running. Learning about DoS attacks was very interesting to me, as these attacks manipulate basic networking principles .I found Scapy a very easy to use python library which allows me to generate any traffic I want on my network. This will be invaluable for testing.

Week Of 8th- 14th February

**Unknown Terminology**

Flask

Port Number

Mu(IDE)

Thonny(IDE)

Random Forest Tree Algorithm (Research)

HTTP DoS Attack

DHCP DoS Attack

**Meaning**

Flask: Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. It has no database abstraction layer, form validation, or any other components where pre-existing third-party libraries provide common functions.

Port Number: A communication endpoint, used to identify a specific process to which packets need to be forwarded or received.

Mu(IDE): Mu is a Python editor for beginning programmers, designed to make the learning experience more pleasant.

Thonny(IDE): Thonny is an integrated development environment for Python that is designed for beginners.

Random Forest Tree Algorithm: Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees.

I will need to research more into this algorithm, however I believe it will be the first algorithm I will try and use to implement the packet analysis.

HTTP DoS Attack: A flood of HTTP requests sent to the same webserver in order to overload the server and deny other users from accessing the webserver.

DHCP DoS Attack: Sending multiple IP requests to the DHCP server, in order to starve the server of IP addresses. This prevents normal user’s from obtaining an IP address, and thus denied access to the internet.

**Comments**

I have a lot to learn in regard to machine learning. It will be very difficult to understand and implement. This is compounded by the fact that I cannot follow along with the tutorial series on fast.ai.

I will have to figure out a way of continuously downloading my packet files from the flask server.

Week Of 14th- 21th February

I am continually learning about Machine Learning.

I am continually learning about pandas. Pandas is a high level data manipulation library which is going to be used to store the data that the machine learning algorithm will use.

A problem I had when cleaning the data was obtaining the number of the “same” packets. This is crucial part of the cleaning process as with this number, you can then deduce the packets per second. A high packets per second would then constitute to a DDOS attack.

After a lot of googling I managed to find the answer:

df['PerSec'] = df.groupby(cols)['SourceIP'].transform('count')

The function groupby allows for the grouping of rows based on a particular collumn. If the value in the collumn is the same between two rows, they are “grouped” together. I then count the number of rows in each row’s group to obtain the occurrence. Then I transform the collumn PerSec to store the occurrence values.

**Unknown Terminology**

Honeypot

**Meaning**

Honeypot: A honeypot is a computer or computer system intended to mimic likely targets of cyberattacks. It can be used to detect attacks or deflect them from a legitimate target. It can also be used to gain information about how cybercriminals operate.

**Comments**

A lot of time spent this week was dedicated to data cleaning. As a result , little else has been learnt, apart from pandas.

Week Of 21th- 27th February

This week was spent trying to find appropriate datasets for my model in order to carry out an experiment which validates that the model works.

**Unknown Terminology**

Google Scholar

Labelled Dataset

Botnet

Survey Paper

**Meaning**

Google Scholar: Google Scholar is a freely accessible web search engine that indexes the full text or metadata of scholarly literature.

Labelled Dataset: Labelled data typically takes a set of unlabelled data and augments each piece of that unlabelled data with some sort of meaningful "tag," "label," or "class" that is somehow informative or desirable to know. After obtaining a labelled dataset, machine learning models can be applied to the data so that new unlabelled data can be presented to the model and a likely label can be guessed or predicted for that piece of unlabelled data.

Botnet: A botnet is several Internet-connected devices, each of which is running one or more bots. Botnets can be used to perform distributed denial-of-service attack, steal data, send spam, and allows the attacker to access the device and its connection.

Survey Paper: A Survey paper comprises the interpretation, the author draws out after reviewing and analysing several research papers(which are already published) on a specific topic.

**Comments**

Finding appropriate datasets is hard, manipulating them will be harder as some reach the size of 6GB. Wireshark can’t handle files that size so I’m assuming I will be using tshark.

Week Of 27th- 4th March

This week was spent primarily writing the blog post, so not many “new” things were learnt. I am still continually learning about pandas and NumPy as I edge closer to finishing the model.

**Unknown Terminology**

os-release file

IPFire

**Meaning**

os-release file: An os-release file is a newline-separated list of environment-like shell-compatible variable assignments. It contains data that is defined by the operating system vendor and should generally not be changed by the administrator.

Example of an os-release file:

Ip Fire IPFire is a hardened open source Linux distribution that primarily performs as a router and a firewall.

**Comments**

Had to learn a little more linux in order to obtain the necessary details for the blog post.

Week Of 4th- 11th March

This week was spent improving and refining the code for the machine learning model. Used a new python library called PyPDF which is used to create and manipulate PDF files, which the model visualiser produces.

**Unknown Terminology**

PyPDF

Decision Tree Classifier

**Meaning**

PyPDF: The package is a pure-Python PDF library that you can use for creating, splitting, merging, cropping and transforming pages in your PDFs.

Decision Tree Classifier: Decision tree learning is one of the predictive modelling approaches used in statistics, data mining and machine learning. It uses a decision tree to go from observations about an item to conclusions about the item's classification.

**Comments**

Though this entry in the learning log looks sparse, I can assure that I’m learning more in libraries such as sklearn and pandas.

Week Of 12th- 19th March

**Unknown Terminology**

PyPi

Kali Linux

LOIC

**Meaning**

PyPi: The Python Package Index, abbreviated as PyPI is the official third-party software repository for Python. PyPI primarily hosts Python packages in the form of archives called sdists (source distributions)

Kali Linux: Kali Linux is a Debian-derived Linux distribution designed for digital forensics and penetration testing. It is maintained and funded by Offensive Security.

LOIC: Low Orbit Ion Cannon (LOIC) is network stress testing tool, meaning it is designed to test how much traffic a target can handle for planning future resource expectations.

**Comments**

In order to fully test my machine learning model, it would have to face the worlds toughest attacks. Kali Linux facilitates these sorts of attacks, so I did a bit of research into it. It is amazing how such powerful tools are freely available to everyone. I suppose if everyone has access to them, the easier they are to counter.

Week Of 20th- 27th March

**Unknown Terminology**

tail

GPIO pins

Sh file

Bash Script

.rule file

**Meaning**

tail: tail is a program available on Unix, Unix-like systems and FreeDOS used to display the tail end of a text file or piped data.

GPIO pins: A general-purpose input/output is an uncommitted digital signal pin on an integrated circuit or electronic circuit board whose behavior—including whether it acts as input or output—is controllable by the user at run time.

Sh file - An SH file is a script programmed for bash, a type of Unix shell (Bourne-Again SHell). It contains instructions written in the Bash language and can be executed by typing text commands within the shell's command-line interface.

Bash script – A bash script is a plain text file which contains a series of commands.

.rule file – Used to store a set of rules for an application to use or filter content by. Used in applications such as Snort or Suricata to filter packet input.

**Comments**

This week I was looking into linux a bit more whilst also doing research for the next hand in.

Week Of 28th- 5th April

**Unknown Terminology**

Feature

dir

**Meaning**

Feature - a feature is an individual measurable property or characteristic of a phenomenon being observed by the machine learning algorithm

Dir – An alternative to the more popular ls linux command. It exists for backwards compatibility reasons.

**Comments**

Summary

Here is a list of some (not all) of what I have learnt over the course of this module:

* C programming
* Flask Webservers
* Dual Booting
* Linux Commands
* Packet Analysis
* Machine Learning
* Pandas
* Problem Solving
* Research Skills