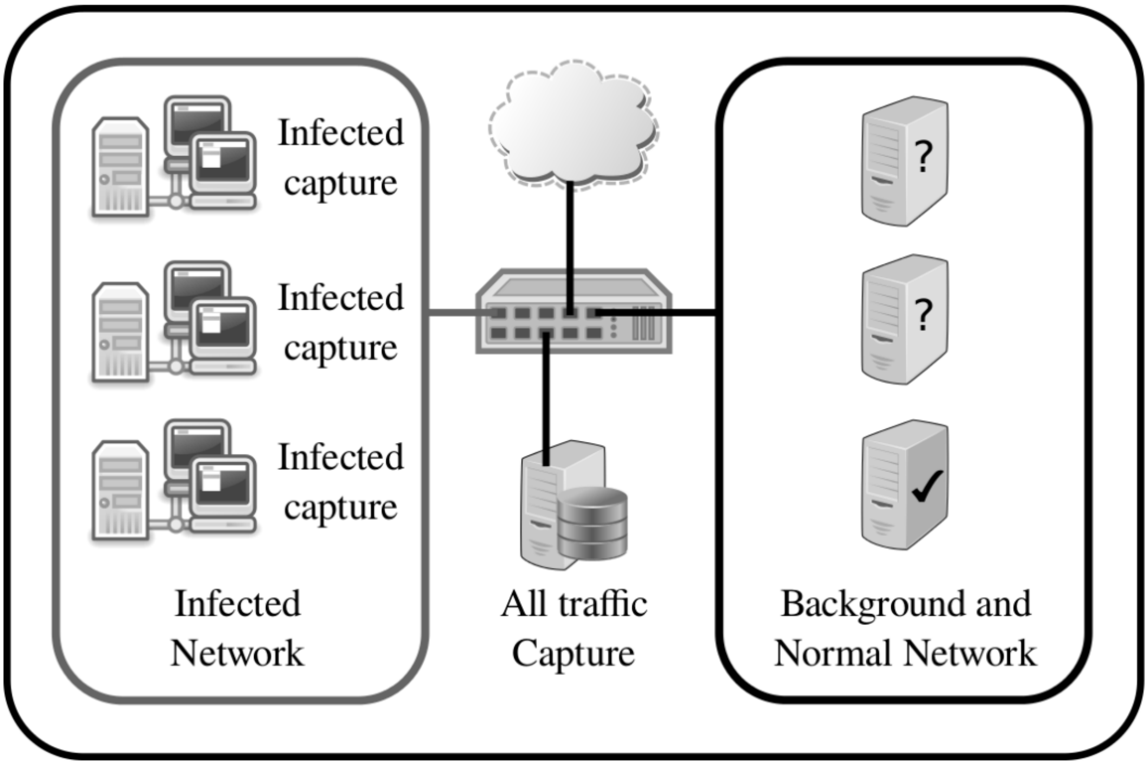
Thanawat Tejapijaya

Weekly report Week 3 (5/6/2023 - 9/6/2023)

As I said last week that I will wait for the svm to finish, as a consequence it hasn't finished running, so I won’t use it. This week, I did data analysis such as finding the history of the dataset like the purpose of making it, how it was made etc.
This is the note that I made:

- botnet is a network of computer that get infected by malware and under control of a single attack party without owner’s knowledge
- Why did they make this dataset?
    - lack of good dataset, it is hard to find third party dataset for botnet that is large, no Background, botnet and normal data that is labeled and from real botnet
    - to make a general public dataset for botnet that can be use to compare each other to find the best botnet detection method
    - creator want to compare the botnet anomaly detection methods but lack of botnet dataset with the characteristics that is needed
    - goals of the dataset
        - must contain real botnet
        - must contain unknown large network traffic
        - must have labels for training and evaluation methods
        - must include different types of botnets
        - must have several bots infected at the same time
        - must have NetFlow files to protect privacy of users (CSV format may not suitable for some algorithm unlike NetFlow and pcap)
- How did this dataset created



  -
  - Traffic was captured on both one of the University router and Linux host
    - Use traffic from linux host for labeling purpose
    - Data from university router used to create final dataset
  - use tcpdump to captured traffics
  - How were they labeled?
    - Assign all to Background
    - Assign Normal if they match certain filter
    - Assign Botnet to all traffic that come from or to any known infected ip address
- Why divide it into 13 files?
  - Each scenario (file) contain of different behavior of malware as state in table below

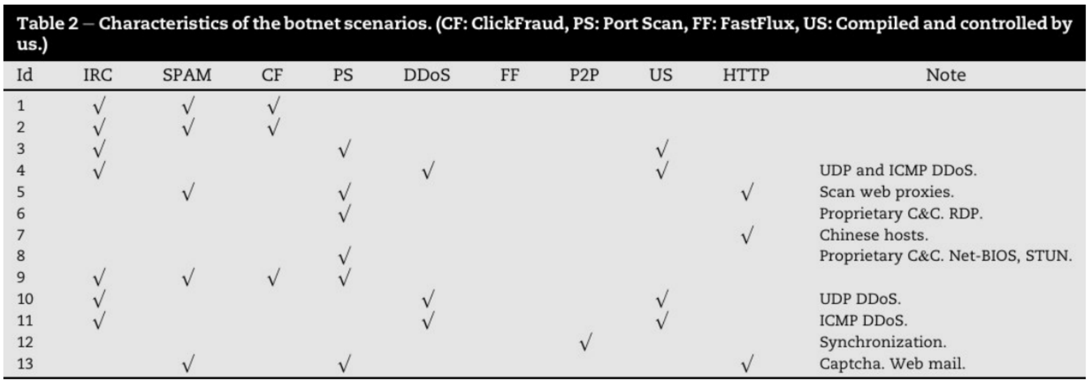


Table 2 – Characteristics of the botnet scenarios. (CF: ClickFraud, PS: Port Scan, FF: FastFlux, US: Compiled and controlled by us.)

Id	IRC	SPAM	CF	PS	DDoS	FF	P2P	US	HTTP	Note
1	√	√	√							
2	√	√	√							
3	√			√				√		
4	√				√			√		UDP and ICMP DDoS.
5		√		√					√	Scan web proxies.
6				√						Proprietary C&C. RDP.
7									√	Chinese hosts.
8				√						Proprietary C&C. Net-BIOS, STUN.
9	√	√	√	√						
10	√				√			√		UDP DDoS.
11	√				√			√		ICMP DDoS.
12							√			Synchronization.
13		√		√					√	Captcha. Web mail.

  -
  - Each of them represent of different malware

Thanawat Tejapijaya

- Output of this dataset/ evaluation
    - binary classification (0 or 1)
    - use error metrics score to evaluate such as accuracy, f1score etc.

And I have prepared for the 1st presentation in Prof Kotani seminar.

Next week I will move on to implementing a paper on the feature extract, feature selection part.