Progress Report

Thanawat Tejapijaya

Content

- Background
- Dataset Analysis
- Data Visualization

Background

What is Cybersecurity? How it’s important?

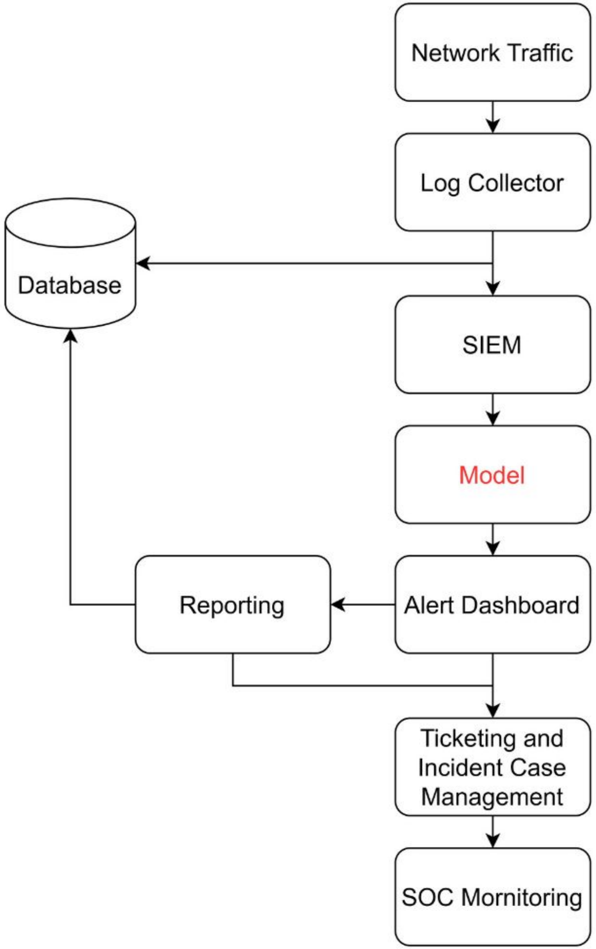
- Cybersecurity is an protection of an hardware, network and program from digital attack
- Cybersecurity divide into 2 main categories
  - Red team - Attacker such as hacker, pentester etc.
  - Blue team - Protector such as security operation center, incident response etc.
- Cybersecurity is one of the top priorities in most of the organization
  - It will help protect those organization data and secret
  - without it your sensitive information may be leak anytime

Security Operation Center (SOC)

- is an security team that help monitor an entire organization IT infrastructure
- Purpose of soc
  - Prepare for the cyber threat
  - Planning on how to handle each threat (Incident response)
  - Prevent Cyber threat
  - Protect sensitive data in that organization
  - Reduced Cybersecurity cost

Where do we implement the model?

- Security Information and Event Management(SIEM) is system that help soc team organize, detect and responds to security threat
- Model is Machine Learning model (ML)
- Purpose of using ML in soc
    - SIEM contain too much noise
    - data from SIEM is HUGE
    - shortage on expert



Dataset Analysis

pictures are from "An empirical comparison of botnet detection methods" Sebastian Garcia, Martin Grill, Jan Stiborek and Alejandro Zunino. Computers and Security Journal, Elsevier. 2014. Vol 45, pp 100-123. http://dx.doi.org/10.1016/j.cose.2014.05.011

CTU 13 Dataset

- Dataset that capture the network traffic of Botnet*
- have 13 files in total
  - Each files contain different attacks

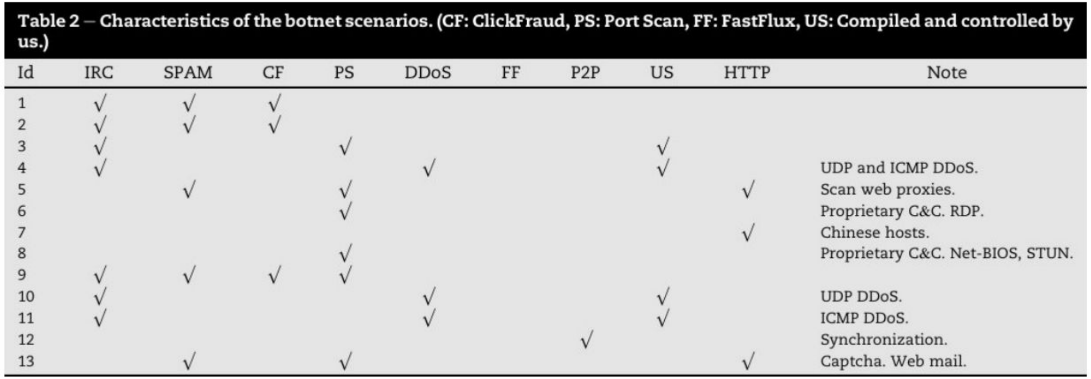
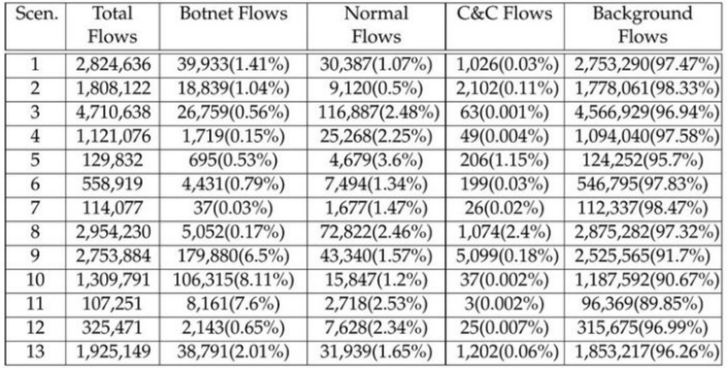


Table 2 – Characteristics of the botnet scenarios. (CF: ClickFraud, PS: Port Scan, FF: FastFlux, US: Compiled and controlled by us.)

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.



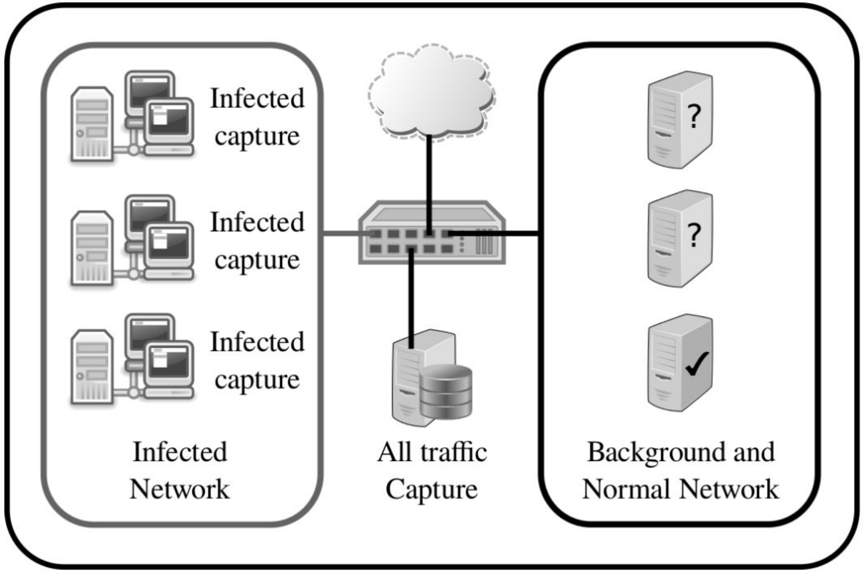
Scen.	Total Flows	Botnet Flows	Normal Flows	C&C Flows	Background Flows
1	2,824,636	39,933(1.41%)	30,387(1.07%)	1,026(0.03%)	2,753,290(97.47%)
2	1,808,122	18,839(1.04%)	9,120(0.5%)	2,102(0.11%)	1,778,061(98.33%)
3	4,710,638	26,759(0.56%)	116,887(2.48%)	63(0.001%)	4,566,929(96.94%)
4	1,121,076	1,719(0.15%)	25,268(2.25%)	49(0.004%)	1,094,040(97.58%)
5	129,832	695(0.53%)	4,679(3.6%)	206(1.15%)	124,252(95.7%)
6	558,919	4,431(0.79%)	7,494(1.34%)	199(0.03%)	546,795(97.83%)
7	114,077	37(0.03%)	1,677(1.47%)	26(0.02%)	112,337(98.47%)
8	2,954,230	5,052(0.17%)	72,822(2.46%)	1,074(2.4%)	2,875,282(97.32%)
9	2,753,884	179,880(6.5%)	43,340(1.57%)	5,099(0.18%)	2,525,565(91.7%)
10	1,309,791	106,315(8.11%)	15,847(1.2%)	37(0.002%)	1,187,592(90.67%)
11	107,251	8,161(7.6%)	2,718(2.53%)	3(0.002%)	96,369(89.85%)
12	325,471	2,143(0.65%)	7,628(2.34%)	25(0.007%)	315,675(96.99%)
13	1,925,149	38,791(2.01%)	31,939(1.65%)	1,202(0.06%)	1,853,217(96.26%)

*botnet is a network of computer that get infected by malware and under control of a single attack party without owner's knowledge

Why this dataset was made?

- lack of good dataset of botnet
- lack of general public dataset that can be use as an standard for comparing each botnet anomaly detection
- lack of dataset that contain real botnet action and behavior
- lack of Background, Normal and botnet labeled dataset
- lack of dataset that was captured in the real world scenarios

How was the dataset created?



picture are from "An empirical comparison of botnet detection methods" Sebastian Garcia, Martin Grill, Jan Stiborek and Alejandro Zunino. Computers and Security Journal, Elsevier. 2014. Vol 45, pp 100-123. http://dx.doi.org/10.1016/j.cose.2014.05.011

- Infected Network consist of Virtual Machine(VM) that running Window XP SP2 operation System
- Capture Network Traffic from both Infected Host and Router

How it’s labeled?

- Final dataset is from Router
- Captured by tcpdump tool
- Use data that was captured from Infected Hosts for labeling purpose
- Step of Labeling
    - Assign all traffic to be Background
    - Assign it Normal if they match certain filter
    - Assign Botnet if they are known or from infected host by looking at the ip address

Output and Evaluation

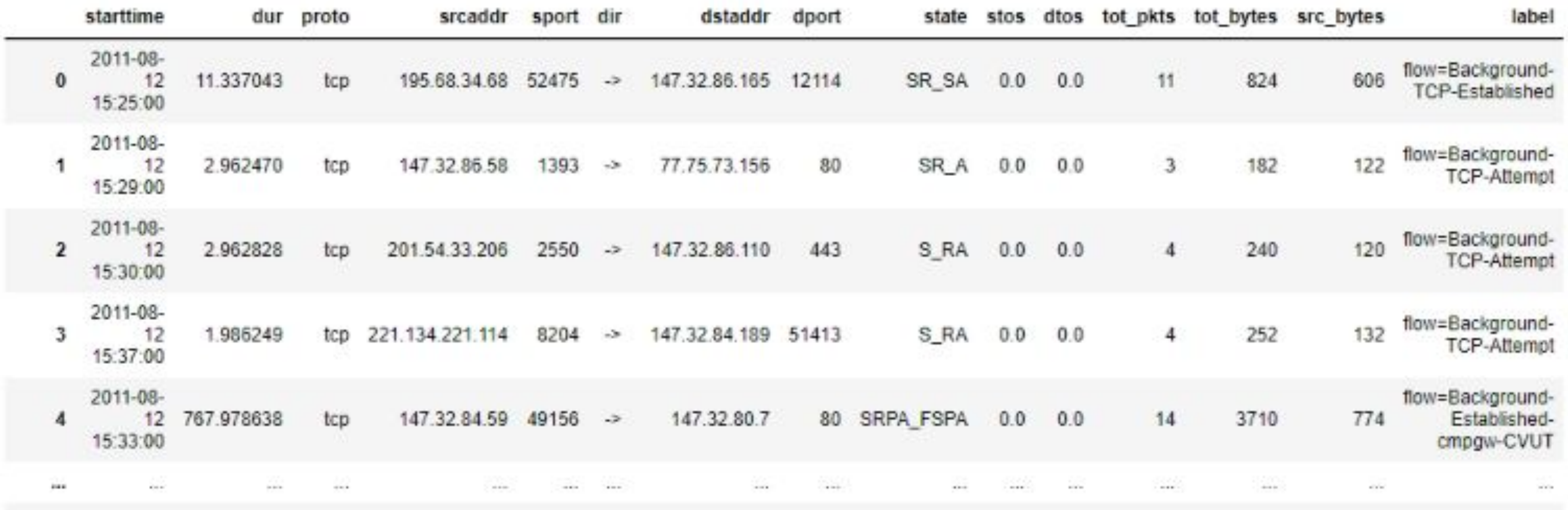
- To Find best anomaly detection method for botnet
- Result will be either 1 or 0 (1 if it is botnet else 0)
- Evaluate by using error metrics score
    - Accuracy
    - Precision
    - F1 Score
    - Error rate
    - False positive rate

* True Positive is when it is an botnet and get detect as botnet

* True Negative is when it is normal and get detect as non-botnet

Which files do I use to train

- Using capture20110812.binetflow to do data analyze (Longest Captured Duration)
  - Contain 4710637 rows -> 4165814 rows (after drop null and duplicates row)
  - How dataset looklike



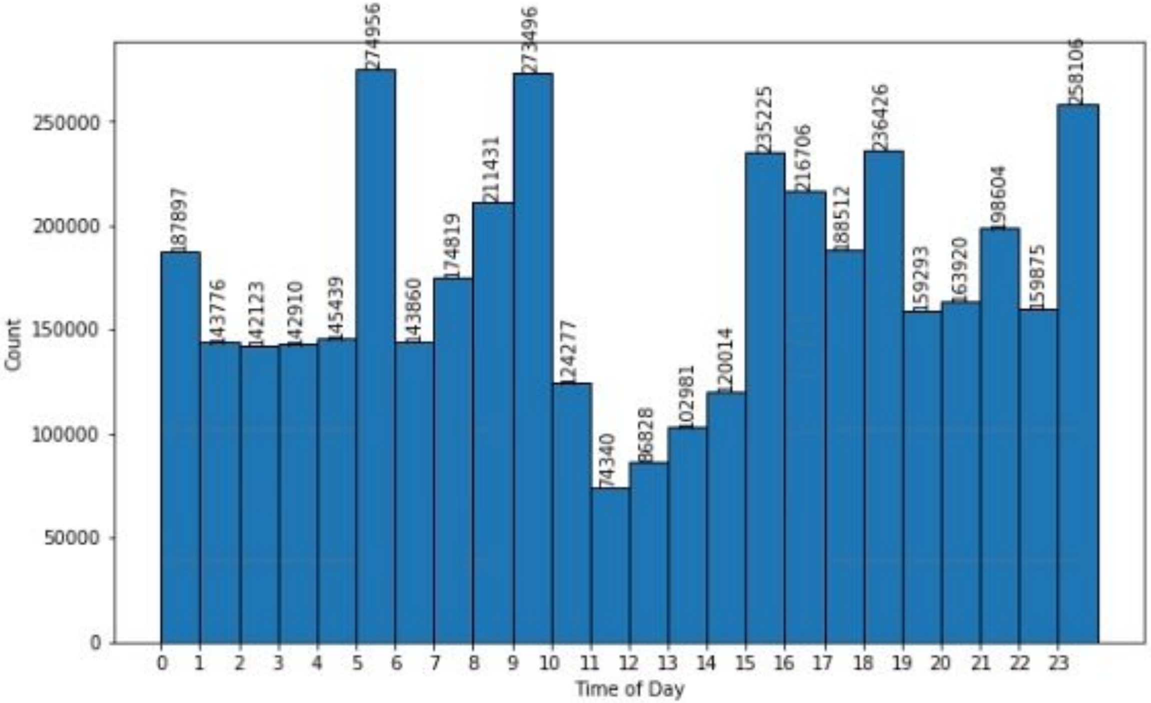
	starttime	dur	proto	srcaddr	sport	dir	dstaddr	dport	state	stos	dtos	tot_pkts	tot_bytes	src_bytes	label
0	2011-08-12 15:25:00	11.337043	tcp	195.68.34.68	52475	->	147.32.86.165	12114	SR_SA	0.0	0.0	11	824	606	flow=Background-TCP-Established
1	2011-08-12 15:29:00	2.962470	tcp	147.32.86.58	1393	->	77.75.73.156	80	SR_A	0.0	0.0	3	182	122	flow=Background-TCP-Attempt
2	2011-08-12 15:30:00	2.962828	tcp	201.54.33.206	2550	->	147.32.86.110	443	S_RA	0.0	0.0	4	240	120	flow=Background-TCP-Attempt
3	2011-08-12 15:37:00	1.986249	tcp	221.134.221.114	8204	->	147.32.84.189	51413	S_RA	0.0	0.0	4	252	132	flow=Background-TCP-Attempt
4	2011-08-12 15:33:00	767.978638	tcp	147.32.84.59	49156	->	147.32.80.7	80	SRPA_FSPA	0.0	0.0	14	3710	774	flow=Background-Established-cmpgw-CVUT
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Data Visualization

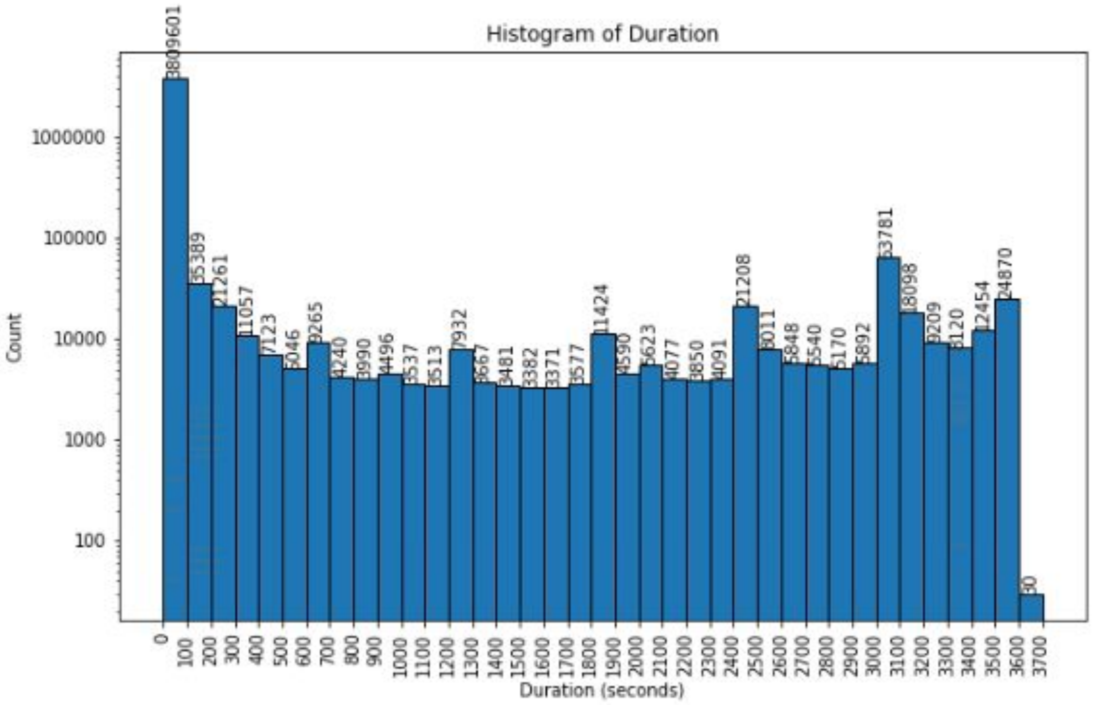
15 Features

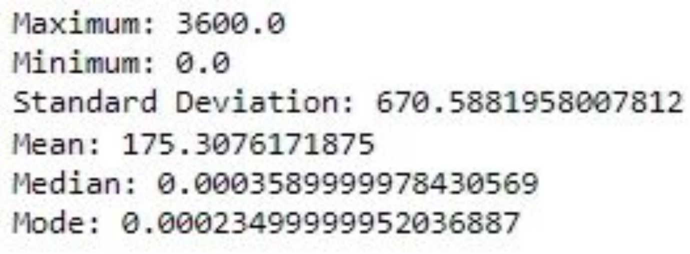
Start Time

Summation of count



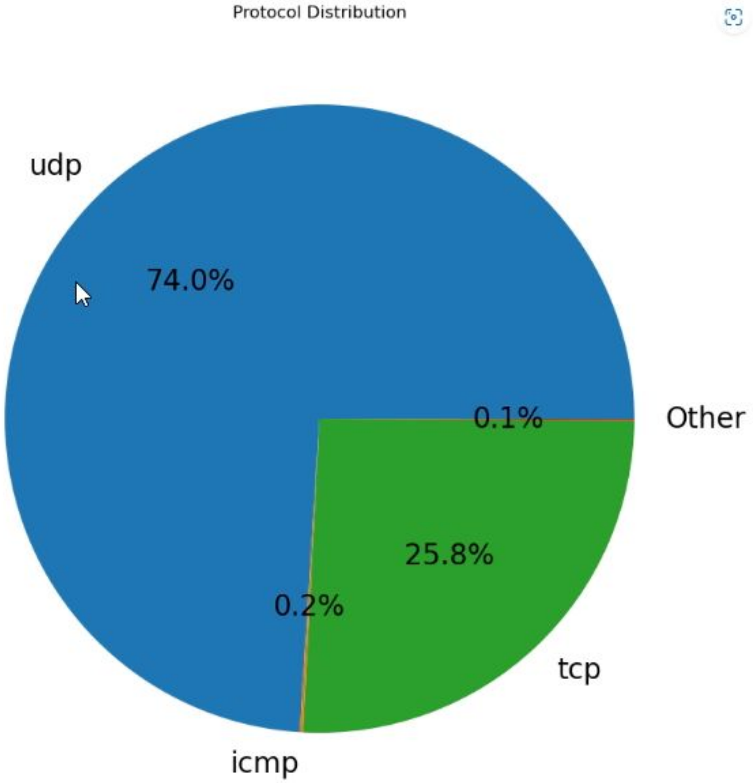
Dur (Duration) - connection duration of each packet





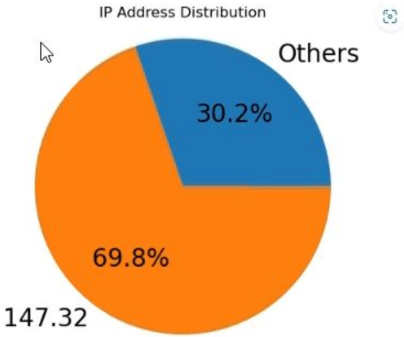
```
Maximum: 3600.0
Minimum: 0.0
Standard Deviation: 670.5881958007812
Mean: 175.3076171875
Median: 0.0003589999978430569
Mode: 0.00023499999952036887
```

Protocol

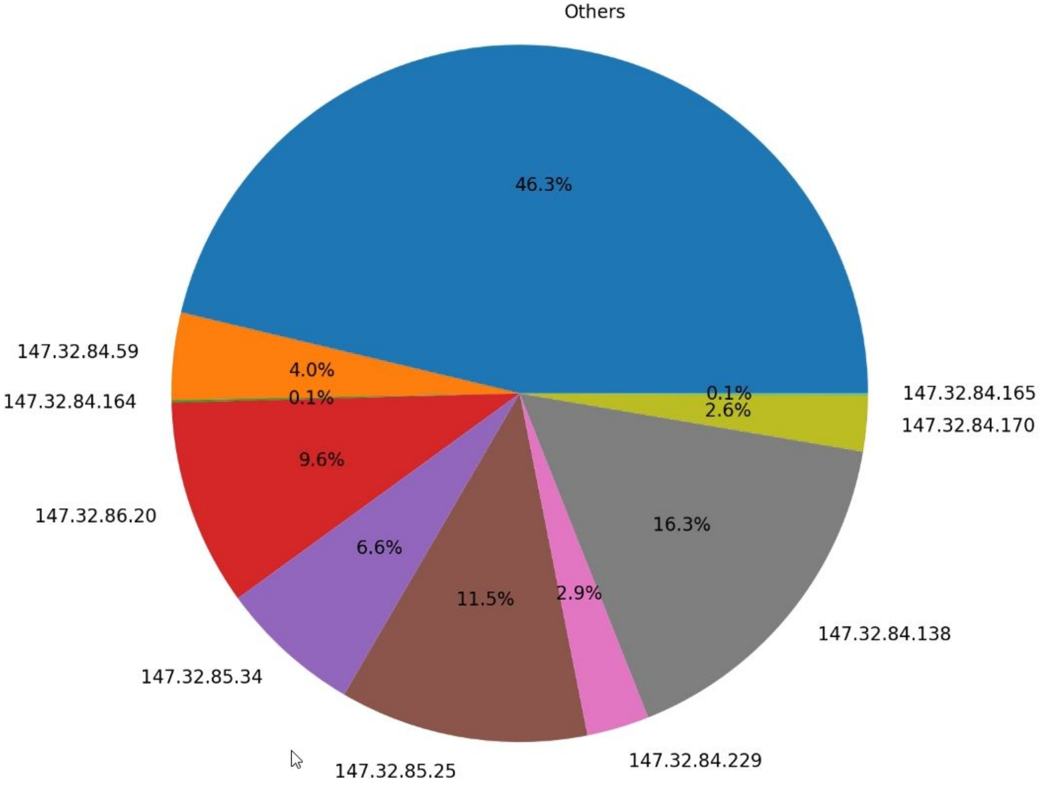


- Protocol of Each packet use
- 17 Protocols in total
- Piechart shows only top 3 protocol were use on this dataset

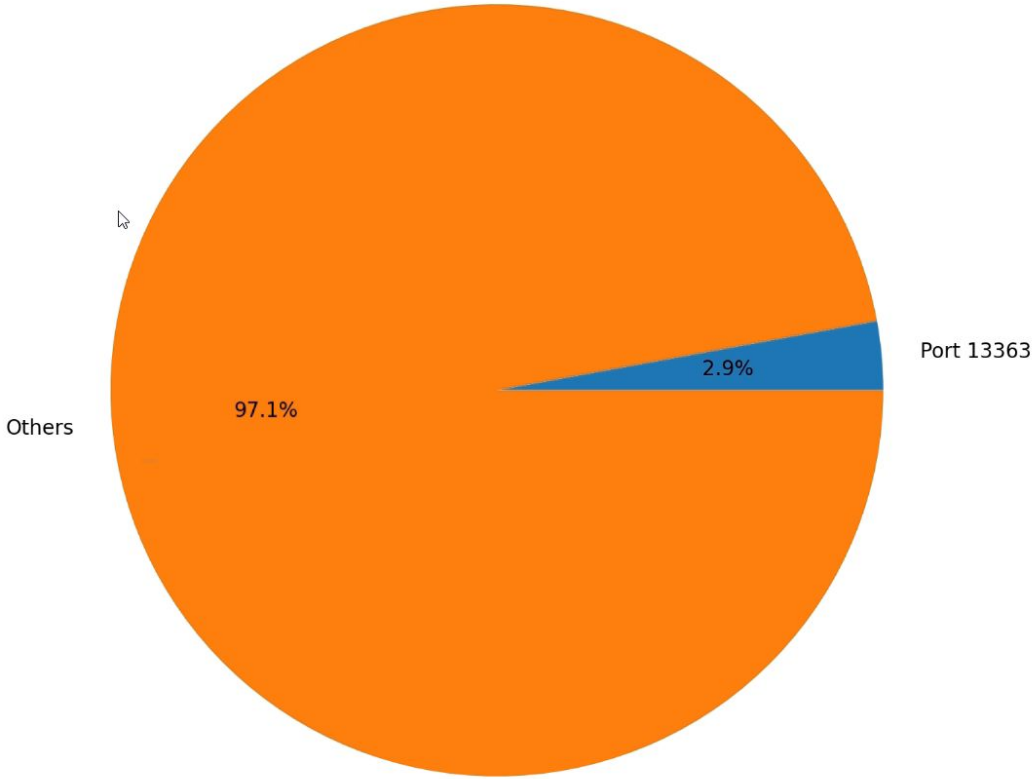
SrcAddr



- Source IP address
- 359,987 of different IP address
- 69.8% has ip address in form of 147.32.X.X

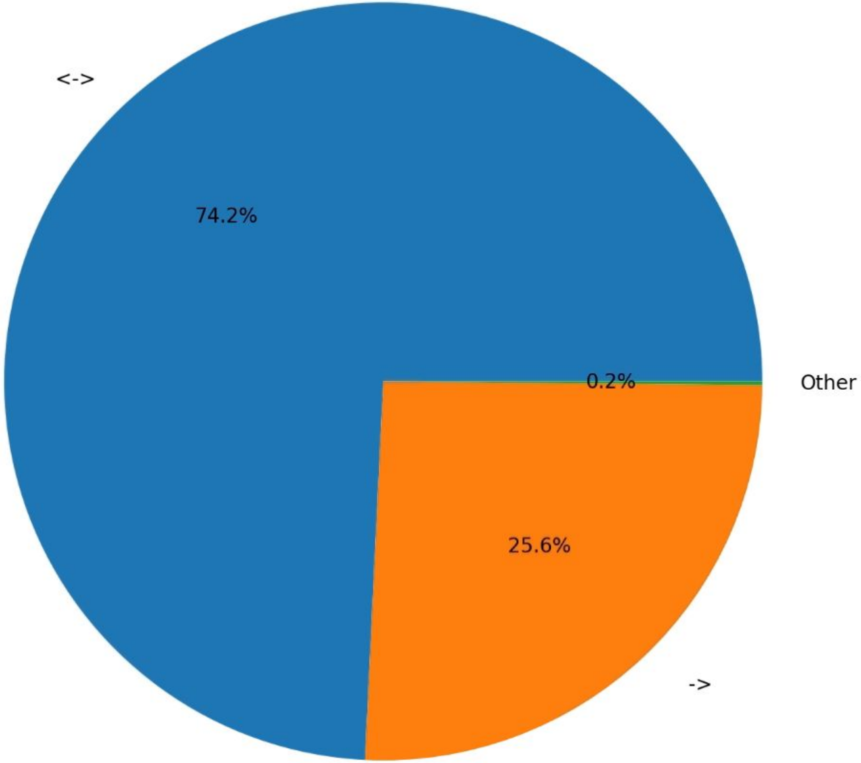


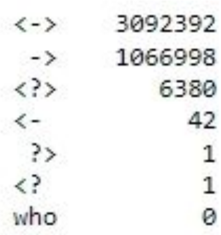
sport



- Source ports
- 64666 of different port
- Port 13,363 is the port that computer inside network use to communicate with each other

Dir - Direction of the network flow



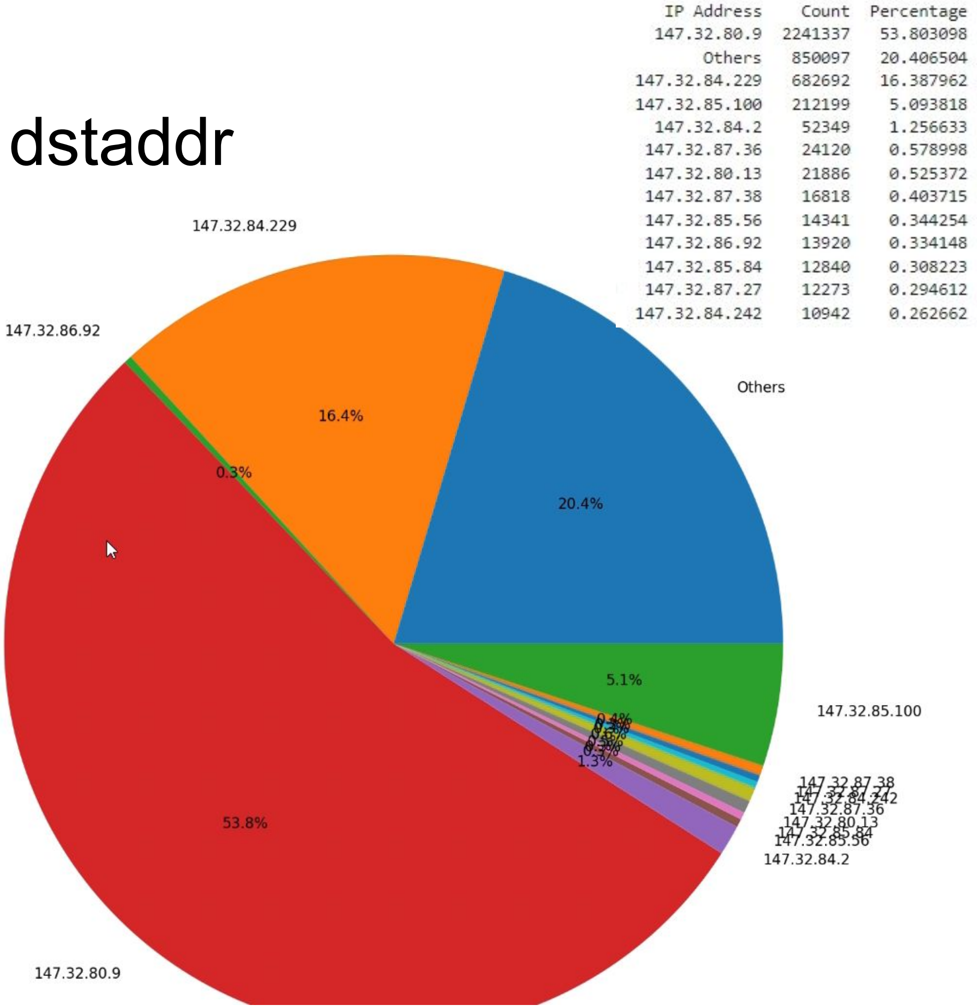


```
<->   3092392
 ->   1066998
<?>      6380
<-         42
 ?>         1
<?          1
who         0
```

74.2% is bidirectional

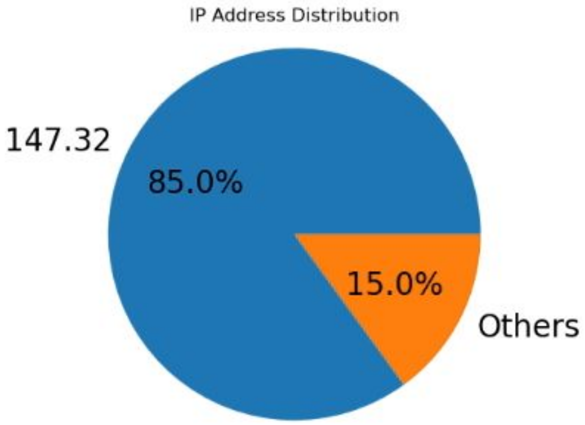
25.6% is from source to destination

dstaddr

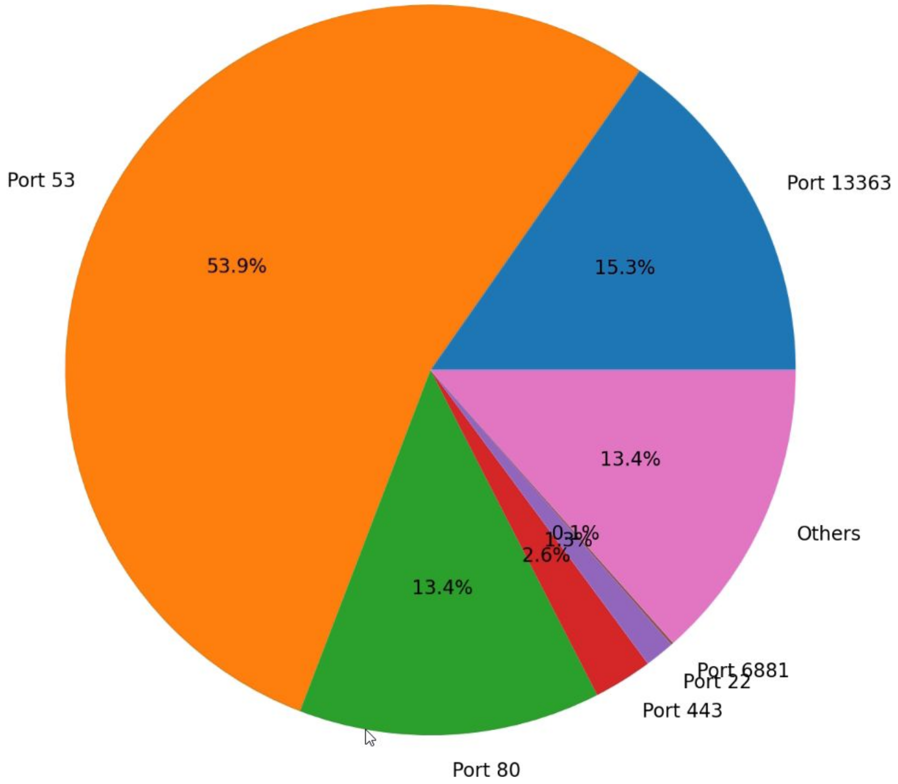


```
  IP Address    Count  Percentage
 147.32.80.9  2241337   53.803098
      Others   850097   20.406504
147.32.84.229  682692   16.387962
147.32.85.100  212199    5.093818
  147.32.84.2   52349    1.256633
 147.32.87.36   24120    0.578998
 147.32.80.13   21886    0.525372
 147.32.87.38   16818    0.403715
 147.32.85.56   14341    0.344254
 147.32.86.92   13920    0.334148
 147.32.85.84   12840    0.308223
 147.32.87.27   12273    0.294612
147.32.84.242   10942    0.262662
```

- destination address
- 125482 of different IP
- 85% has ip address in form of 147.32.X.X

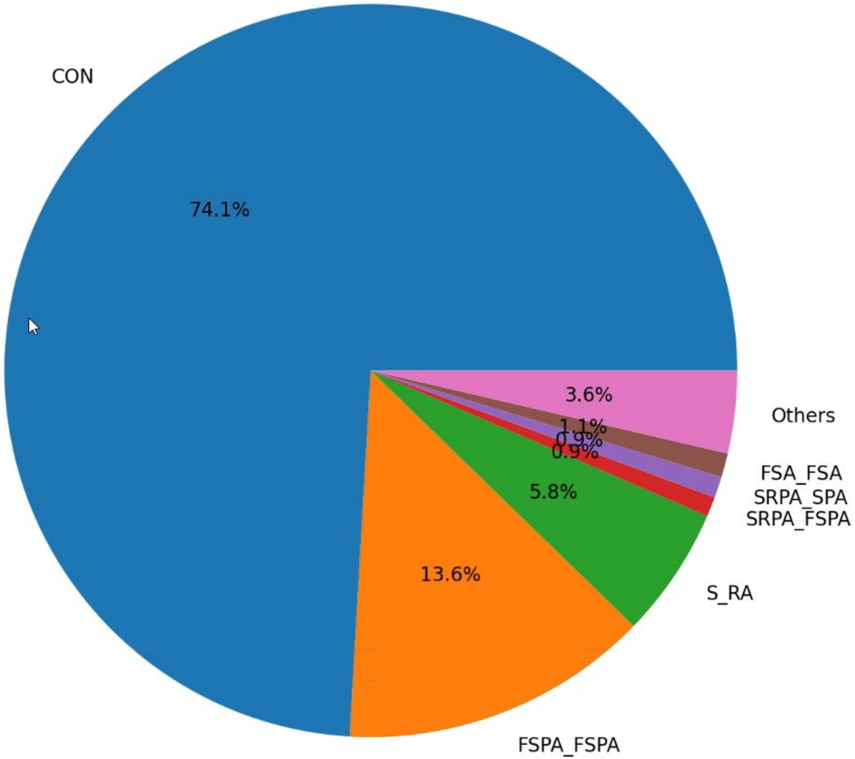


dport



- destination port
- 84591 of different port
- port 13363 is same as sport
- port 53 is Domain Name System (DNS)
- port 80 is Hyper transfer Text Protocol (HTTP)
- port 443 is Hyper transfer Text Protocol over TSL/SSL (HTTPS)
- port 6881 is Bittorrent (unofficial)

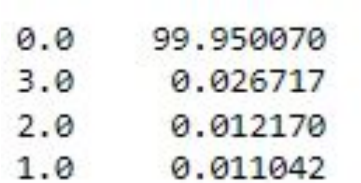
State



- The state is protocol dependent and _ is a separator for one end of the connection
  - CON mean connected (UDP)
  - S mean Synchronized (TCP)
  - F mean FIN (TCP)
  - A mean Acknowledge (TCP)
  - P mean push (TCP)
  - R mean reset (TCP)

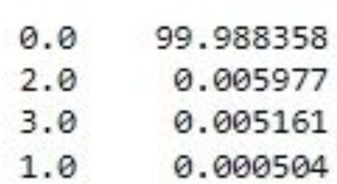
stos & dtos

stos



```
0.0    99.950070
3.0     0.026717
2.0     0.012170
1.0     0.011042
```

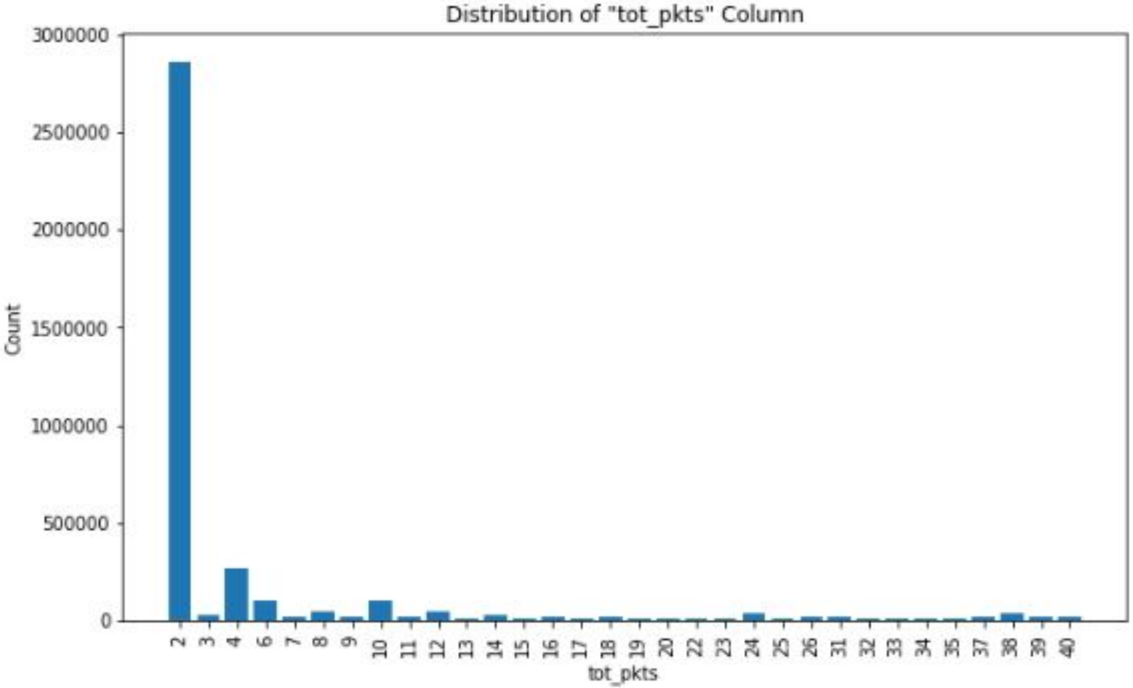
dtos



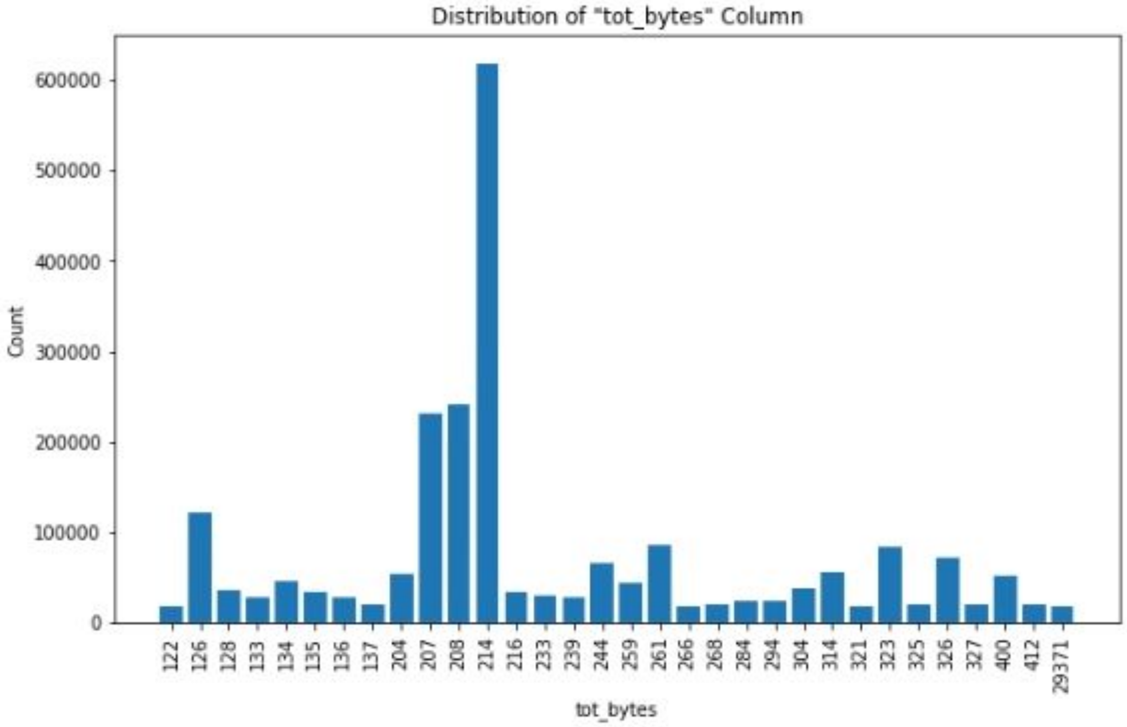
```
0.0    99.988358
2.0     0.005977
3.0     0.005161
1.0     0.000504
```

- Number that tell priority of the packet
  - 0 mean routine
  - 1 mean priority
  - 2 mean immediate
  - 3 mean flash
- 99.9% is 0

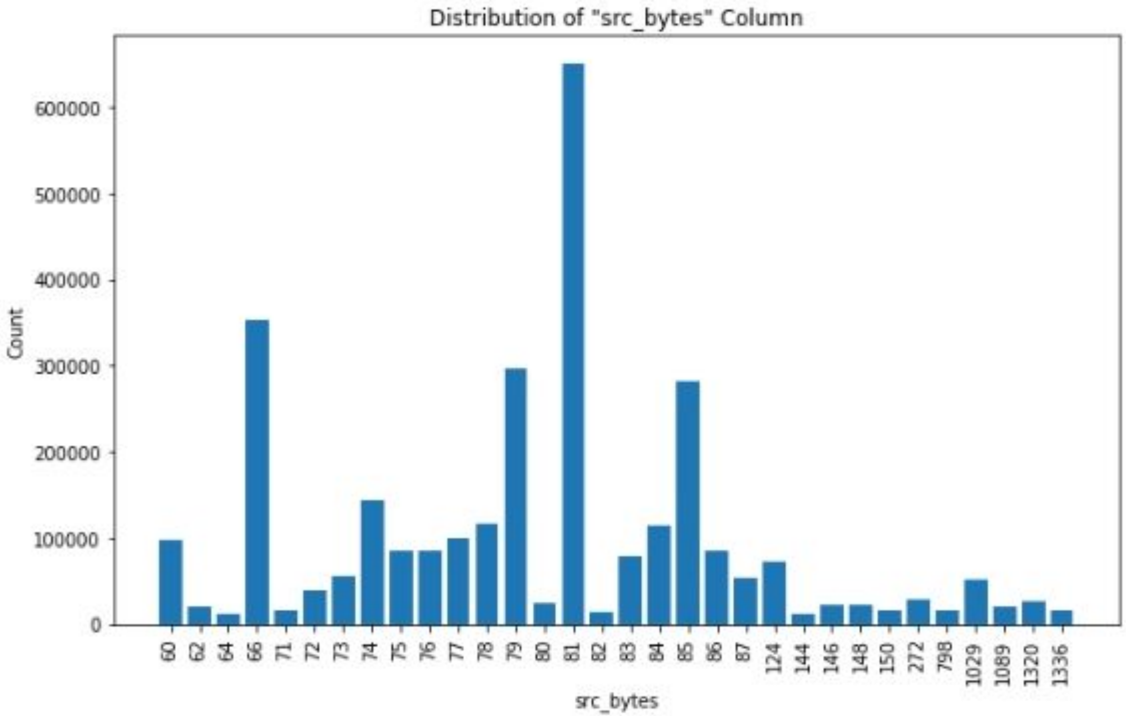
tot_pkts - Total numbers of transaction of each Packet



tot_bytes - total numbers of transaction Bytes



src_bytes - total numbers of transaction Bytes from Source



Label

Case 1

“Flow=From-Botnet” -> 1 contain 26822

Others -> 0 contain 4683816

“->” Refer to “change to”

Case 2

flow=Background -> 0 contain 2340042

flow=To-Backgro -> 1 contain 2225846

flow=From-Norma -> 2 contain 116303

flow=From-Botne -> 3 contain 26822

flow=From-Backg -> 4 contain 1041

flow=To-Normal- -> 5 contain 562

flow=Normal-V44 -> 6 contain 22