Department of Computer Science

CSE 590 Fundamentals of Data Science

Mini Project - 1

Anomaly Detection in Network Data

Author:
Arpit Singh
SBUID:110162005

Supervisor:
Dr. Leman Akoglu

September 17, 2015

1 Introduction

Huge amount of data is being exchanged over networks due to sudden increase in bandwidth and computational power over last decade. Often network data is huge and consist of packets from various protocols. Data Science provides handy tool to analyze such huge data to find trends and detect anomalies if any. In this project, objective is to analyze network traffic data[1] and find anomalies. The data looks like intranet data and recorded over period of few hours on machine. The data is available in two parts, one is ".csv" file and other is ".txt" file. On closer look, it is found that both of them maps to same data, but Acoutic.txt file contains much larger details that tetsnetworkdata.csv. Aim is to apply analytic tools observe trends and find out anomalies. Therefore extensive analysis is being done over the data provided and results are presented below.Data provided mainly consists of five fields:packet no,time,Source IP,Source port,Destination IP,Destination Port ,Protocol and info field.

2 Problem Statement and Approach taken

Objective is to find out anomaly in network data. Possible anomalies in the network could be surges in packet rate, unusual drop of packets, too many reset sent/received etc[2] Problem Statement is identified as below:

1. Determining the glitches in packet rate
2. Identifying the properties of anomalies
3. Determining the protocol found in anomalies
4. Determining the IP address or group of ip address involved in anomalies
5. Determining the port number targeted
6. Possible explanation of pattern or anomalies
7. Conclusion about issue. identifying what extra data required to come to exact cause of issue

2.1 Approach towards solving the issue:

First step taken to spot the anomaly was to parse the data and build a time series graph. Python[3] scripts were used to parse data and matlibplot[4] library is used to plot the graphs. Every step is deduced from the result of previous steps.Steps taken are enumerated below:

1. Time series graph of packet received per minute is drawn
2. Protocols dominating the traffic are identified by protocol analysis.
3. Per protocol time series is correlated with original packet time series
4. TCP data is analyzed to find out number of SYN,FIN and RST received
5. Since RST represent abnormal connection termination, RST data is parsed and stored in separate file.Time series of RST data is drawn and analyzed
6. Suitable bar graph are drawn to identify the dominating IP address and port no responsible for RST

3 Result and Discussion

First the time series of traffic which provides variations of packet received per minute is shown in Figure 1 As we can clearly see in that periodic spike is observed at regular time intervals. Thus, to identify which protocols are causing it, a bar graph is drawn in Figure 2 which identifies protocol forming the major part of traffic

Thus we observe that RST packets are comparable to SYN packets indicating a vast majority of connection are closed prematurely. This indicates an alarming situation where machine is targeted intentionally to start dead connection. This behavior can be shown for Dos attacks[5] as well as by port scanning software[6]

RST packet data from TCP data is further parsed to scrutinize the source IP ,destination IP ,source port and destination port. Figure 6 shows time series of RST packets Thus there is clear correlation between network spikes and RST packets From above data it is clear that machine is targeted for unwanted connections. However it could a DOS attack or it could be a port scanning software which is trying to find out which ports are open.

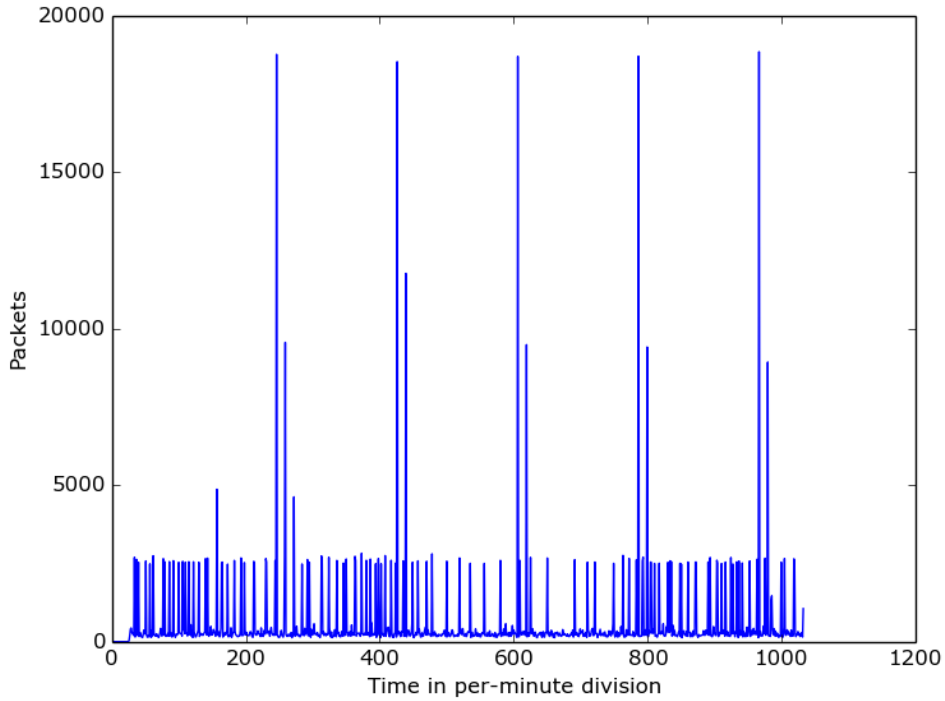

Figure 1: Time series graph of traffic where time is in per minute division

Figure 7 shows source IP vs packet data

Since most of data is from single IP, indicating that given IP is interface targeted. Figure 8 shows destination IP vs packet exchanged indicating multiple IP targeted host machine. However all destination IP are Port analysis is done to find out source port targeted and destination port sending RST packets Figure 9 and Figure 10 shows packets received per source port and destination port. However, figure 7,8,9,10 shows only targets receiving packets more than defined threshold and not all targets are shown All IP appear to be part of same intranet It is clear from figure that mysql port was targeted to connect with mysql on other machine

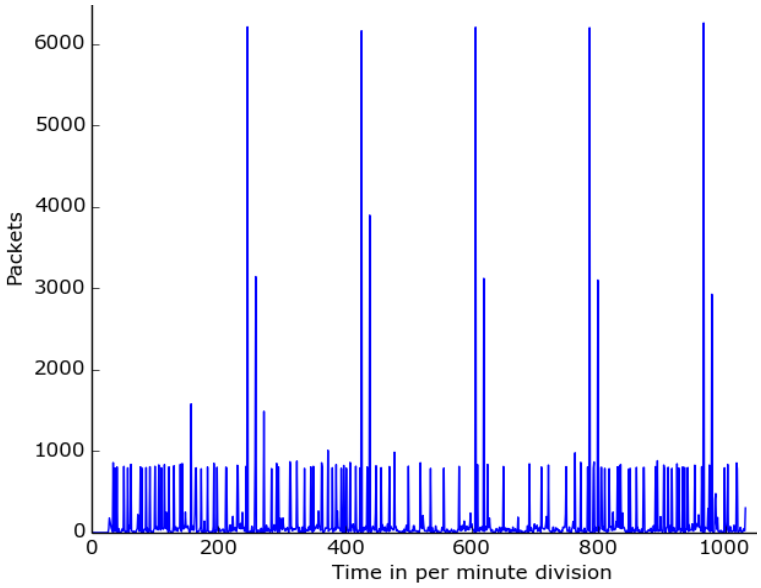

Figure 3 Time series of TCP packets

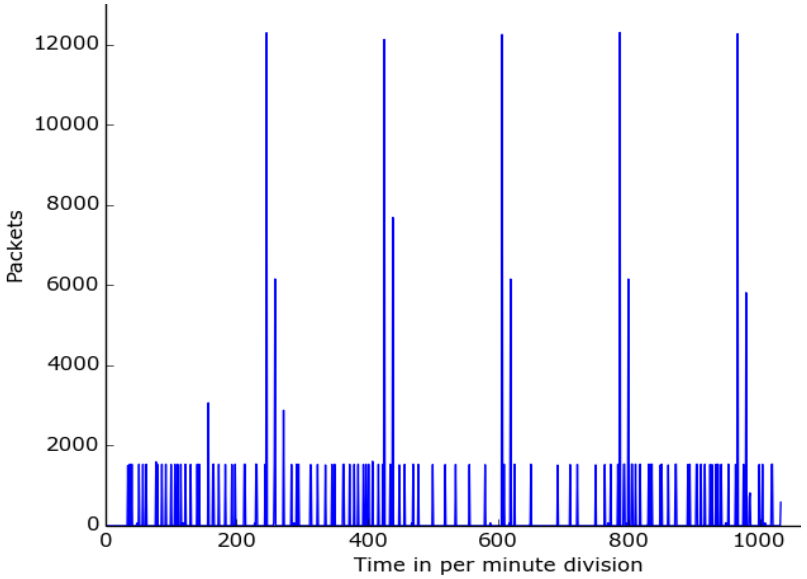

Figure 4 Time Series of FTP-DATA packets

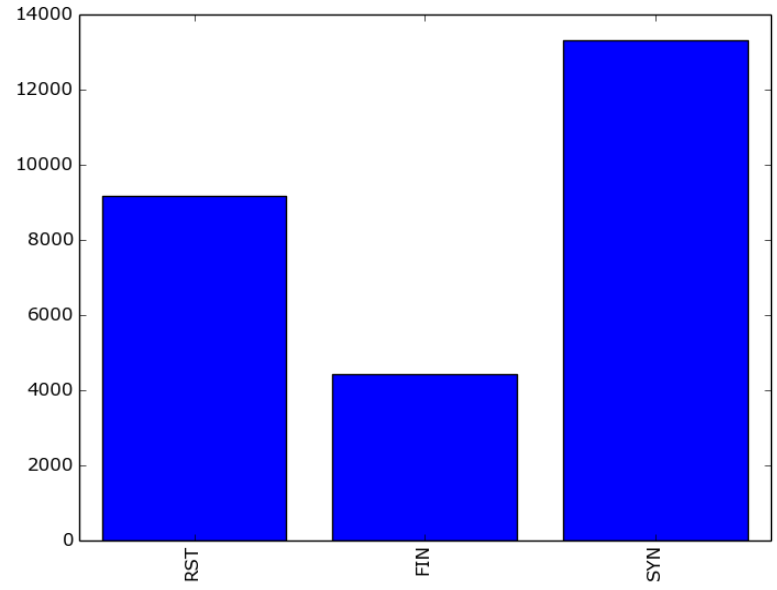

Figure 5 Count of types of TCP Packet

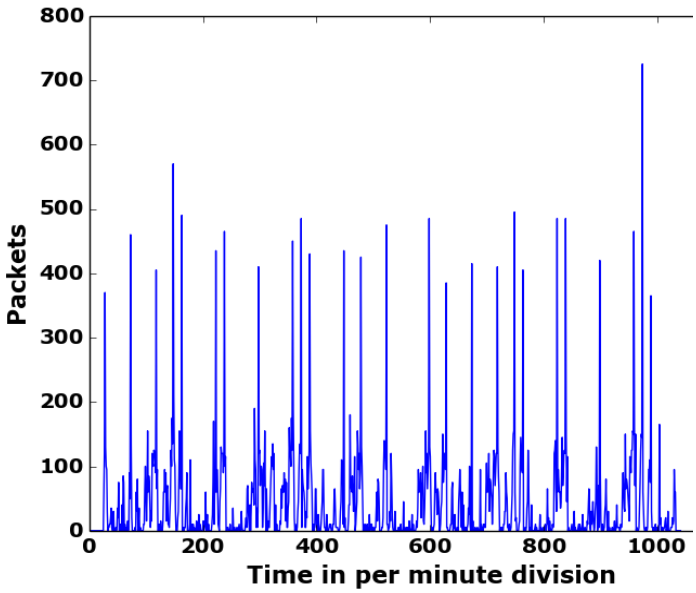

Figure 6 Time Series of RST Packets

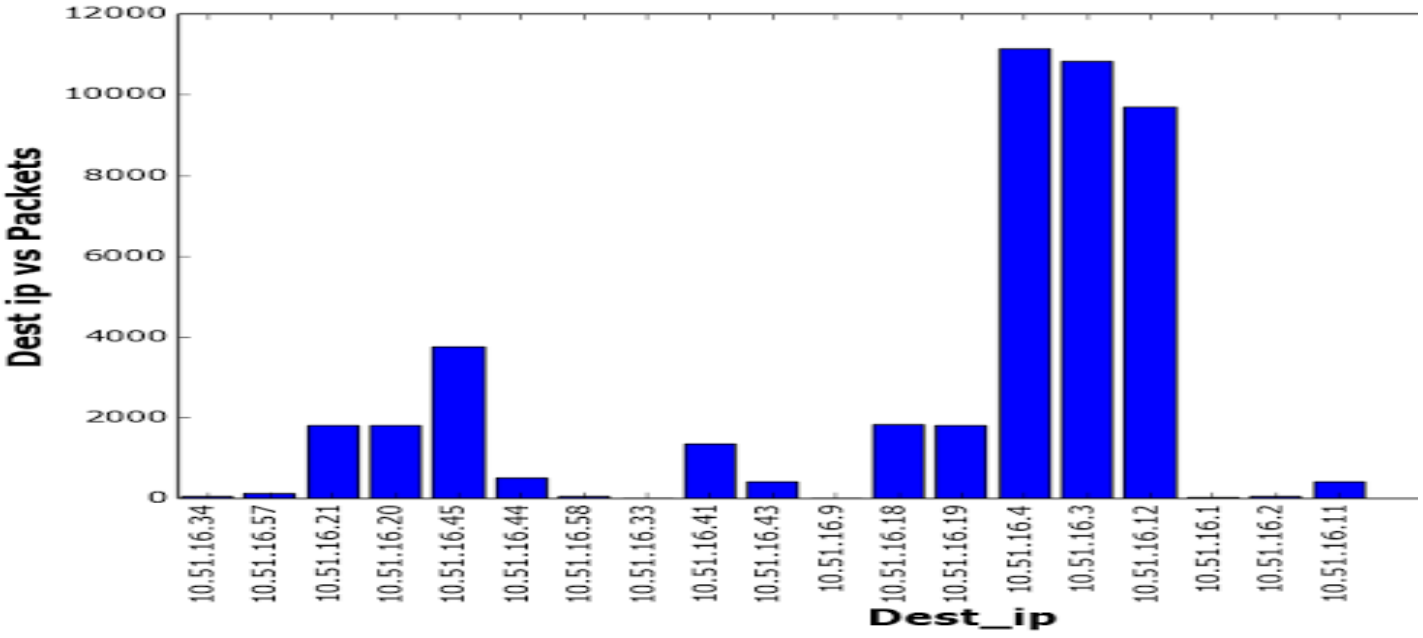

Figure 7 Destination IP vs packet received

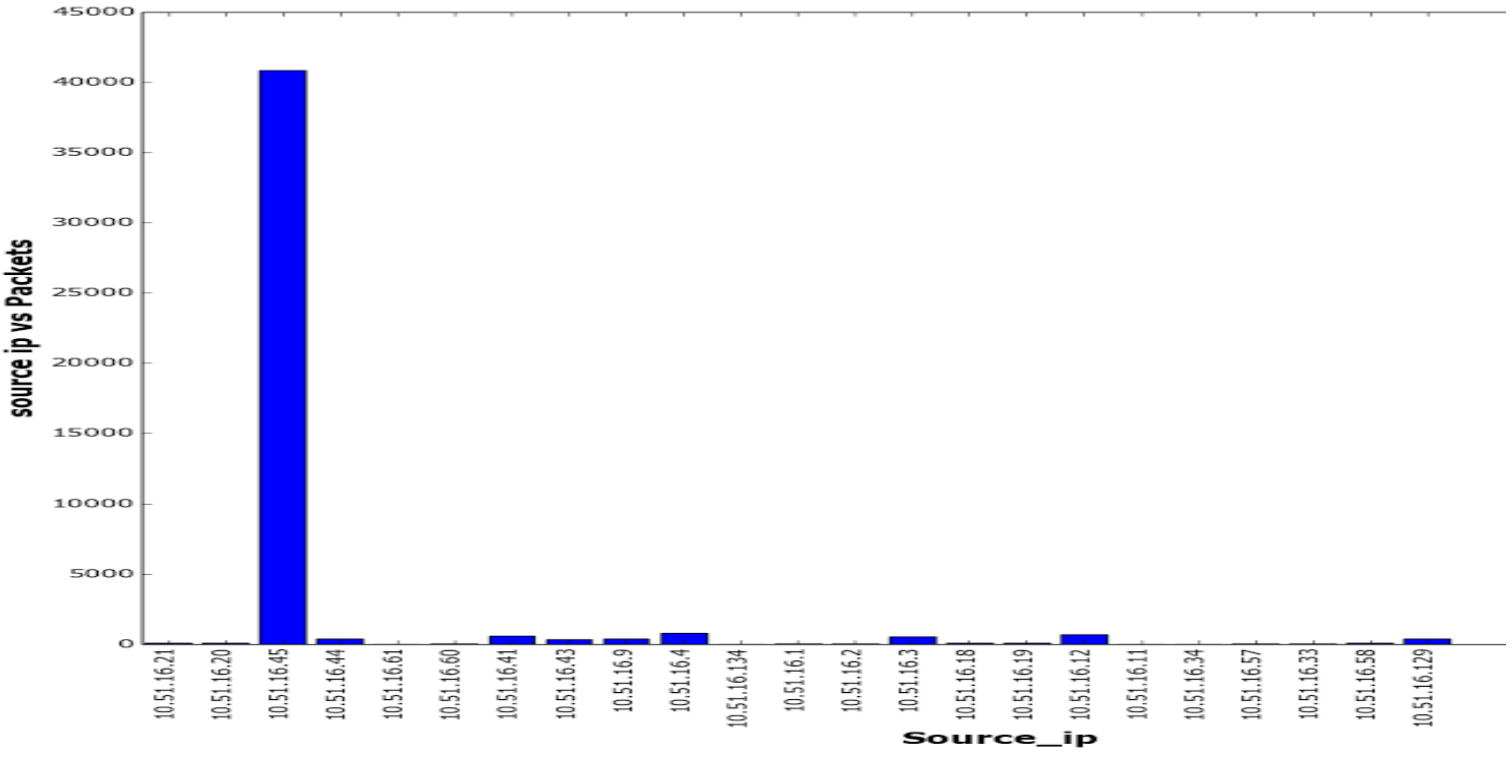

Figure 8 Packet received per source ip address

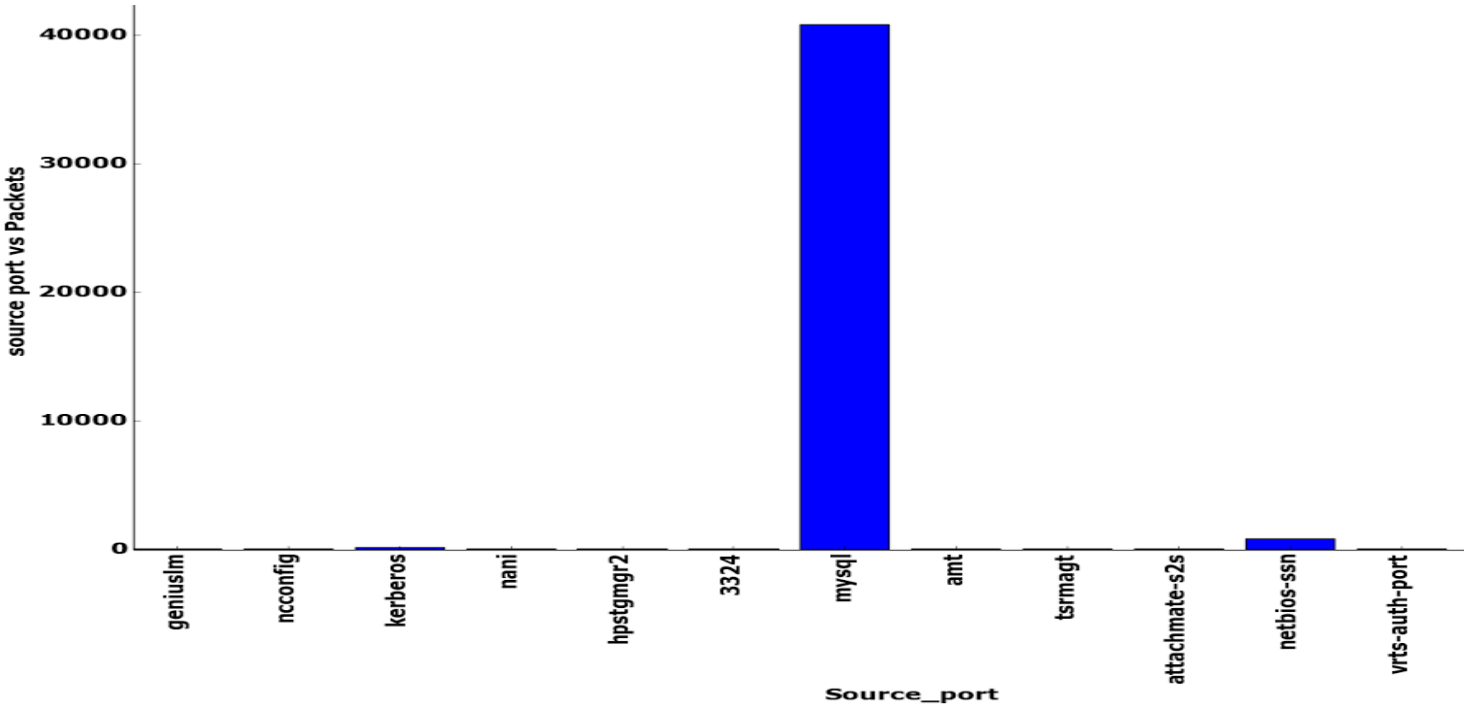

Figure 9 Packet received per source port

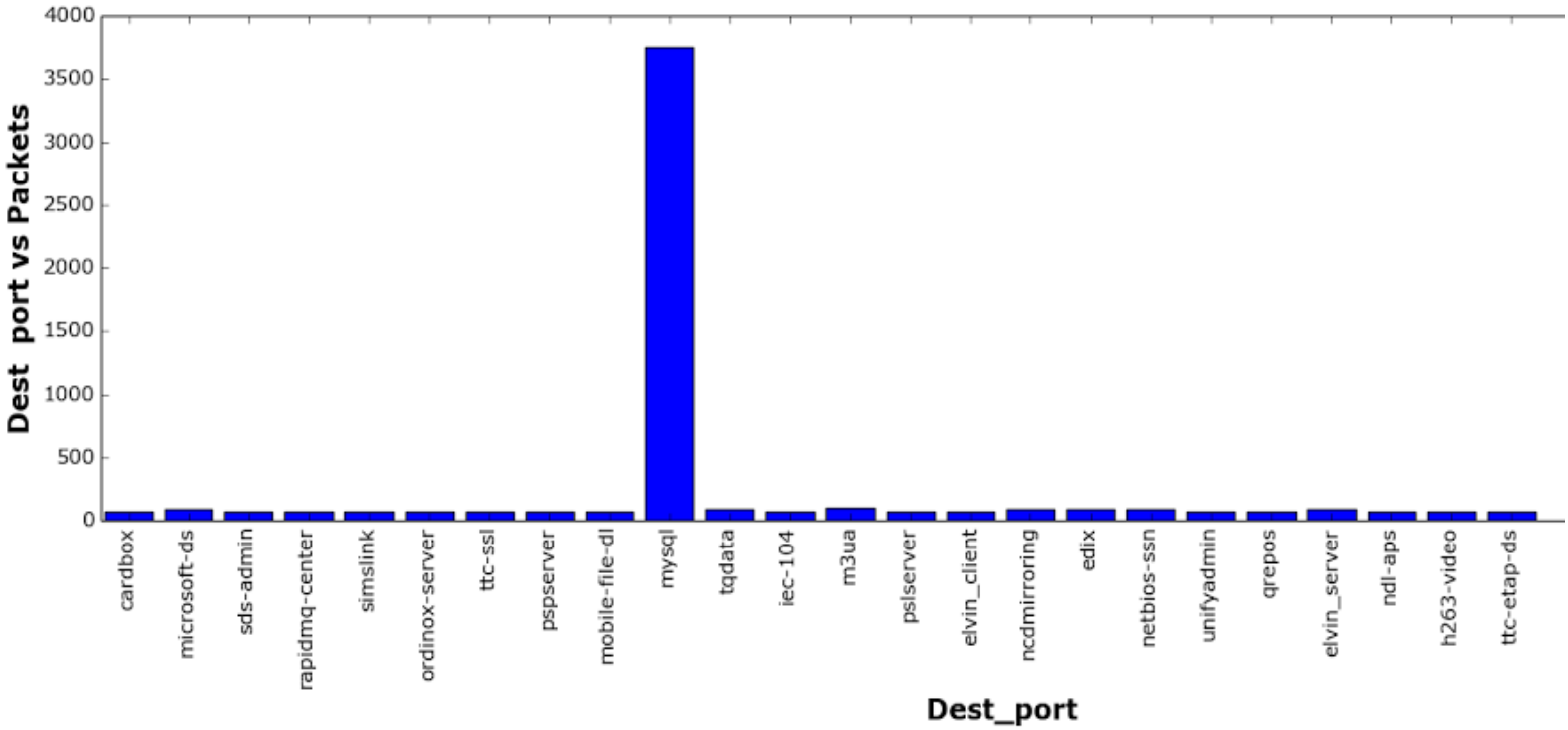

Figure 10 Packet received per destination port

4 Conclusion

Conclusion: we observe regular spikes which are tracked down to correlate with FTP-DATA as well as with TCP RST sent out RST Packets are sent at regular intervals indicating host machine is target of unwanted traffic.It is found that 1710 ports were targeted on host machine However st of the packets originated or targeted towards mysql port It is either a port scanning software running at regular intervals Someone has intruded into intranet and launched Dos attack on one particular machines Machines are requesting data from from SQL database using FTP-DATA bombarding it with connection requests regularly. It is possible that application requesting connections at fast rate might be faulty along with synchronization mechanism to ask data from single served might be required.

5 References

1. http://www3.cs.stonybrook.edu/~leman/courses/15CSE590/assignments.htm#mini1
2. https://en.wikipedia.org/wiki/Attack_(computing)
3. https://www.python.org/
4. http://matplotlib.org/
5. https://en.wikipedia.org/wiki/Denial-of-service_attack
6. https://en.wikipedia.org/wiki/Port_scanner