The International Journal Of Engineering And Science (IJES)
|| Volume || 3 || Issue || 4 || Pages || 41-52 || 2014 ||
ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805

Analysis of Pattern Recognition Techniques for Detecting Traffic Anomalies

Maksha D.D. & Zirra P.B
[1]Adamawa State University, Mubi, PMB 25 Mubi Adamawa state-Nigeria
dmaksha@yahoo.com

--ABSTRACT--

Network traffic anomalies stand for a large number of the internet traffic and highly affect the performance of the network resources. Detecting these threats is a time consuming task and is laborious that network operators face daily. In spite of the advantages of these methods, researchers have reported several common drawbacks that affect their use in practice. The generation of packet header, IP addresses as well as the communication between client and the server were achieved through simulation. A server and a client were designed in Java programming language using server and socket classes. Hough transform is then applied to the picture in order to obtain a Hough space. From the Hough space, points that constitute a line in the picture, based on the threshold that are identified. The lines drawn in the Hough space representing the points that intersects. These straight lines indicate abnormal behavior in the simulated network communication. They are the identified anomalies in the traffic. Statistical method has common drawback as lack of ground truth data and approximate evaluation of the method. This method take advantage compared to other method which stated that graphical representation reduce the dimension of network traffic and provide intuitive output as it is not possible with current statistical method. Therefore network operators should adopt the current method of detecting anomaly which in turn increases performance and quest for development.

Keyword: *Network Traffic, Data, Anomaly, Hough Space*

Date of Submission: 25 March 2014 Date of Publication: 20 April 2014

I. INTRODUCTION

The internet has become a common medium for communication and information exchange, providing many attractive services for ordinary users. A victim of its own success, internet traffic is still growing at a fast rate and contains an increasing amount of anomalies such as configuration failures and attacks. These improper uses of network resource consume bandwidth and adversely affect the performance of networks. Thus, these anomalies penalize legitimate applications from using an optimal amount of network resources. Since the core of the internet is particularly deteriorated by anomalous traffic, quick and accurate detection of anomalies in the backbone traffic has been a hot topic[6].Also, the success of internet services result in a constant network traffic growth along with an increasing number of anomalies such as remote attacks (examples; Denial Of Service (DOS) attack, port scan, worm spreading, phishing, hackers, virus,) bad configuration, and interception of data on network. These anomalies represent a large fraction of the internet traffic that is unwanted and penalizes legitimate users from accessing 'optimal network resources. Therefore, detecting and diagnosing these threats are crucial tasks for network operators that are trying to maintain the internetresources made available. Due to the important traffic volume involved, quick and accurate identification of anomalies in the internet traffic requires automation. Intensive studies have been carried out in this field but the proposed anomaly detection methods still have important drawbacks that affect their practical usage in real environment[6].

Detecting Anomalies

The pattern-recognition based detector takes advantage of image processing techniques to provide intuitive output and parameter set.The main idea of the detection method is to monitor the traffic in Two Dimensional (2-D) pictures where anomalies appear as "line" which is easily identifiable using a pattern recognition technique called Hough transform. This method overcomes several shortcomings of current statistical-based anomaly detectors; similarly to the statistical base detector it requires constant attention from network operators to be optimally used[5].

Packed Header Data

Packet is a small group of digits having a well-defined format whichcan form part of complete digital message [2]. A whole setof such packets constitute the message. Forms of signal encoding may thenbe used which facilitate automatic error detection. Packet switching is amethod of transmitting a digital message as a series of short numbered frame across communication system. Also packet may be described as a buffer of certain size, full of data that is transmitted from one machine to another. This data buffer is transmitted differently depending on which "data link" network in use, such as Ethernet, Fibre Distributed Data Interface (FDDI), toke-ring, Wide Area Network (WAN), point to point links and so forth. The end result is the same entire bundle called a buffer, is transmitted from one machine to another and arrives at the destination intact (hopefully).However, header enables long message to be broken down into a large number of packages before transmission and correctly re-assemble it at its destination. From network point of view, a protocol header contains information that is specific for that protocol.

II. STATEMENT OF THE PROBLEM

Statistical analysis is an appealing approach to solve the anomaly detection problem. However, resulting anomaly detections suffer from high error rate as investigating the output of statistical tools and tuning their parameter set in accordance to the analyzed network traffic is challenging. Hence pattern-recognition based detector is reliable detectors that can be use conveniently and successfully, with good output. Also by applying a pattern-recognition techniques using Two Dimensional (2-D), it will be able to detect new and unknown anomalies. It report traffic that is featuring anomalous behaviour.

III. AIM AND OBJECTIVES OF THE STUDY

The aim of this research is to develop an anomaly detector through packet data header using image processing techniques of 2-D transformation.The objectives include:

[1] To produce a means of detecting anomalies in Internet network traffic.

[2] To detect some issues (e.g. TCP SYN flood attack) that occurs in TCP protocol before it affects the network system.

IV. RELATED WORK

The methods of detecting anomalies in internet traffic data are numerous. Different features of the internet can be used to accomplish this task. The internet traffic feature chosen for this research is the packet header data. Using that information, it is easy to plot this information into the graph and detect any unusual pattern referred to as an anomaly.A set of flows altering the distribution of at least one of the four following traffic features: the sources IP address, destination IP address, source port and destination port can be investigated. This has sprung an areaof research that has recently received a lot of attention [7].Internet traffic anomaly detection aims at identifying anomalous traffic that is transmitting in the core of the internet, where the monitored traffic is asymmetric due to routing policies, thus flows are incomplete[4].There are two approaches to anomaly detection: Network intrusion detection and Internet traffic anomaly detection[5]. The goal m intrusion detection is toprotect a network from remote threats. Thus the detection method is monitoring the traffic at the edge of the protected network where complete flows and packet pay load are usually accessible.

In contrast, internet traffic anomaly detection aim at identifying anomalous traffic that is transmitted in the core of the internet where the monitored traffic is asymmetric due to routing policies, thus flows are incomplete [5]. The work is dedicated exclusively to internet traffic anomalydetection, thus in this dissertation anomaly detection refer only to this specific domain.

Volume based approaches are monitoring the number of bytes, packets or flows transmitted over time and aims at detecting abnormal variance that represent abusive usage of network resources or resource failure. Several methods have been proposed to effectively identify local and global traffic volume variances that stand for respectively short and long lasting anomalies.[1] proposed a method base on wavelet that inspects the traffic volume at different frequencies. Their approach makes use of the wavelet analysis to dissect the traffic into three distinct signals representing local, normal and global variance of the traffic. The decomposed signal are analyzed by a detection procedure that finds the irregularities and reports the period of time they occur. Since the three signals represent the traffic at different time scales, this approach is able to report short and long lasting anomalies. Nevertheless, as the whole traffic is aggregated into a single signal diagnosing the detected anomalies is challenging and anomalous are left unknown.[6]proposed a detection method that detect and diagnoses anomalies in large scale works. First, their approach monitors the traffic using a matrix in which each cell represents the traffic volume of alink of the network at a certain time interval. Second, the main behaviour of the traffic is extracted from his matrix with the Principal Component Analysis (PCA) and anomalies are detected in residual traffic. Finally, the origin and destination nodes of the network that are affected by the anomalous traffic are identified and reported.Anomaly detectors themselves have to be evaluated. This is done so as to compare their efficiencies and determine the most suitable anomaly detector for any given condition. Providing ground truth data evaluate anomaly a detector is a challenge that has been addressed several times in the past. We distinguish two different approaches to evaluate an anomaly detector, namely using simulated or real traffic [3].According to [4] providing real internet traffic for the evaluation of anomaly detectors is challenging for two main reasons:

[1] Labeling anomalous traffic in real internet traffic is difficult because of the lack of truth worthy method and the traffic volume that makes manual labeling unpractical.
[2] Proving internet traffic is inherently problematic because of the privacy issues.

recently proposed a data set containing real backbone traffic where anomalies are precisely located. In this work the traffic is captured at different points in the used network, which is supposed to be anomaly free, and the researchers generate two kinds of anomalies; flash crowd and Distributed Denial of Service (DDOS) attacks. Their experiment consisted of different scenarios where the intensity of the anomalies varies. Thus, the sensitivity of the detector to DDOS and flash crowd is easily identified. However, there are only a few kinds of anomaliesin their data and they are not a realistic representation of the diverse anomalies found on the internet.

According to [4] adeeper understanding of the Measurement and Analysis on the WIDE Internet (MAWI) traffics is achieved by analyzing the traffic with three anomaly detectors based on different theoretical background, These are:

[1] Principal component analysis (PCA)
[2] Gamma modeling
[3] Kullback-leiblerdivergence

Statistical sequential change-point detection has been applied successfully to network anomaly detection;[11] characterized network anomalies with management Information Size Base (MIB) variable undergoing abrupt changes in a correlated fashion.Given a set of MIB variables sampled at a fixed time-interval, the authors compute a network health function by combining the abnormality indicators of each individual MIB variable. This network health function can be used to determine whether there is an anomaly in the network. [12]detect SYN flooding attacks based on the dynamics of the differences between the number of SYN and FIN packets, which is model as a stationary periodic random process. The non - parametric cumulative sum (CUSUM) method is then used to detect the abrupt changes in the observed time series and thus detect the SYN flooding attacks.[10] develop a traffic anomaly detection scheme based on Kalman Filter. Unlike [6] process the link data using a Kalman Filter rather than PCA analysis to predict the traffic matrix one step 4 into the future. After the prediction is made, the actual traffic matrix is estimated based on new link data. Than the difference between theprediction and the actual traffic volume anomaly based on different threshold’s methods. Kalman filter has been applied successful to a wide variety of problems involving the estimation of dynamics of linear system from incomplete data. Thus, it is a promising tool for network anomaly detection together with other more complicated models of non-linear dynamics.

V. METHODOLOGY

The two traffic features to be considered are:

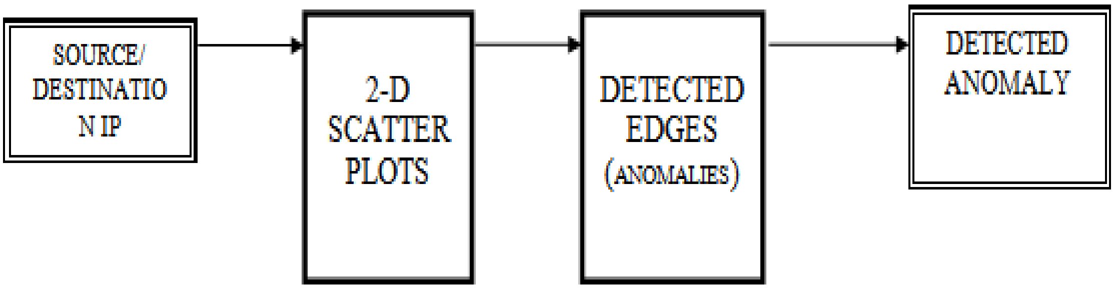
[1] Source IP address
[2] Destination IP address

The reason for the use of only two parameters for our proposed anomaly detection is because the two are sufficient to implement our detector. Some other parameters like source and destination ports cannot be used, because time of recording the data stream can also be used especially when evaluating the efficiencies of various anomaly detectors. However, for this study the destination and source IP addresses are deemed sufficient in order to achieve our aim and objectives.

Translating the collection of these source and destination IP addresses over a period of time will require performing the following.

[1] We make the horizontal axis to represent time.
[2] The vertical axis to represent either the source IP or destination IP addresses.
[3] The shade of the plot (whether dark or light) indicates the amount of packets.
[4] The apparent “lines” represent excessive use of traffic features.

Operation of the system Anomaly Detection



Hough Transform Algorithm for 2-D Scatter Plots

To detect any anomaly in form of line or edges, Hough Transform Algorithm is employed as follows:

[1] Take a point (x^1, y^1) in the image, all lines which passes through that pixel to produce an equation in the form $y^1 = mx^1 + c$ (1)

[2] For varying value of m and c equation (1) can be re-write as $c = x^1 m + y^1$ (2) where x^1 and y^1 are constants and m and c are variables.

[3] Get a straight line on a graph. Each different line through the (x^1, y^1 `) corresponds to one of the points on the line in (m,c) space (termed as Hough Space).

[4] Take all pixels which lie on the same line with (x , y^1) space to represent the lines which passed through a single point in (m,c) space.

[5] Collect consecutive detected lines to derive an algorithm for pattern recognition.

[6]

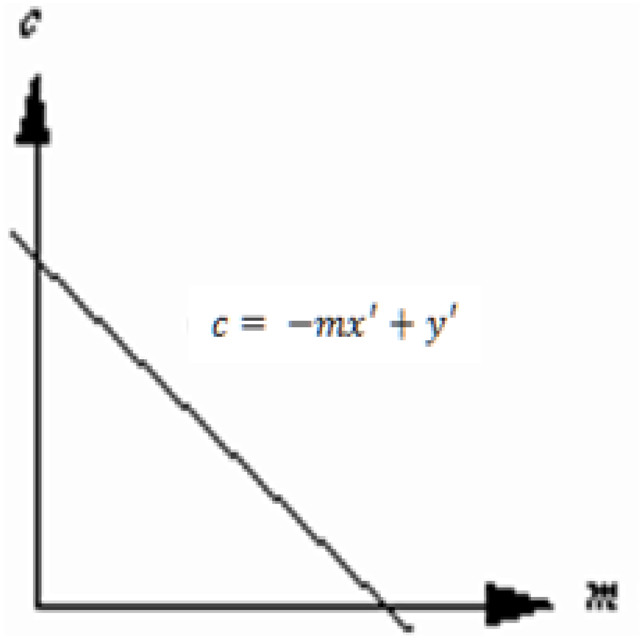


Fig. 3.4(A): Principle Of Hough Transform

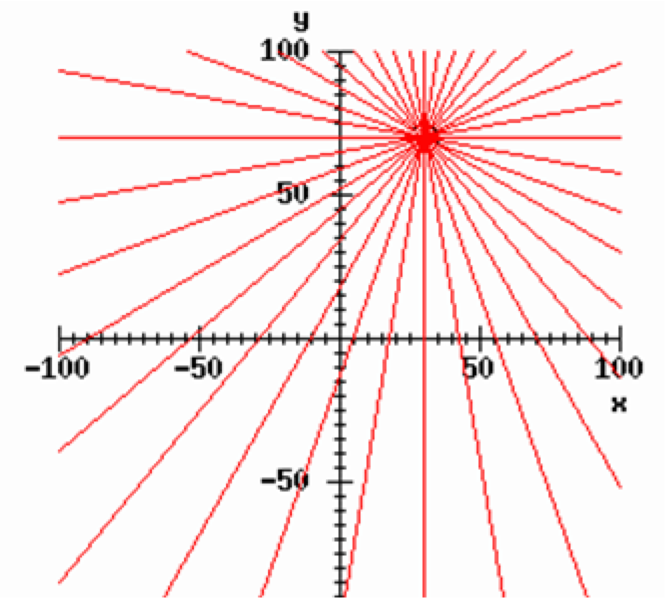


Fig. 3.4(b): Principle of Hough Transform

lgorithm 1: Anomaly Detection and Classification

```
define F as number of traffic features used (F = 2)
set the sliding window at the beginning of the data
while. (window: =eof) do
    for (all packets in the window)
        plot packet in P pictures and store header information
    end for
    for (all pictures) do
        compute the Hough space for the picture
        extract lines from the Hough space
        for (all lines found) do
            implement a new event (using event handler e)
            retrieve all packet header from line
            e←summarizes traffic features from packet headers
            if (anomaly with main feature (a) = main features of e) then
                add etoa
            else
                create a new anomaly
            end if
        end for
    end for
    close Slide the window
end while.
```

IV. RESULTS AND DISCUSSION

The results of different cases were tested. For the detected Traffic anomalies through packet header data, the user will start the server. The server will wait for a communication from the client. When the client has started the system, it will attempt to established connection with the server. Once the connection is established immediately, communication will start flowing between the client and the server. The generation of packet header, IP addresses as well as the communication between client and the server were achieved through simulation. A server and a client were designed in Java programming language using server and socket classes. First the server is started and it waits for connection from a client. Then the client is started, which connects to the server and the communication is established. The period of a complete communication is known as a session.

Case I

Figure 4.1a has the IP address at the y-axis and the time on the horizontal axis. The points seen are the scatter plots. Figure 4.1b contain the same axis as in Figure 4.1a, but is a Hough space. It contains three intersections. The number of intersection in the Hough space produced. The detected anomalies in Figure 4.1c. The straight lines in Figure 4.1c are the anomalies. The number of intersection correspond to the number of points on the scatter plot that constitute a line base on the threshold, hence the anomaly.

Case II

In case 2, Figure 4.2a contains points greater than Figure 4.2a, but the points are two scatrer since the points are two scatrered, they will not be able to make an intersection in the Hough space in Figure 4.2b. Since there were no intersection in Figure 4.2b, the points will remain as it is without producing any anomaly. No enough points are found to constistute a line. Sonsequently no intersection appear on Figure 4.2b which is the Hough space.

These straight lines as seen in Figure 4.1c and 4.2c respectively indicate abnormal behavior in the simulated network communication. They are the identified anomalies in the traffic. Comparing this output with [11] asserted that statistical method that has common drawback as lack of ground truth data and approximate evaluation of the method.This method take advantage compared to [3] which stated that graphical representation reduce the dimension of network traffic and provide intuitive output as it is not possible with current statistical method.

IV. CONCLUSION

The recommendations that have been put forward will no doubt in improving performance, efficiency and effectiveness. If these recommendations are implemented and the existing anomaly detection is reviewed, then the objective of the anomaly detection will no doubt be attainable. This study has help in contributing to the knowledge as it increased the accuracy in detecting anomalies in network. This is because anomalies are being identified as picture and picture can easily be identify and read easily compared to other techniques that will involves sampling without ground truth data as used in statistical techniques. Further, this method takes advantage of graphical representation to reduce the dimension of network traffic.

REFERENCES

[1] P. Barford, J. Kline andD. Ploika (2002).*A signal analysis of network traffic Anomalies.*New Jersey: Prentice Hall.

[2] K.G. Beauchamp (1990).*Computer communication*. 2nded. Chapman& Hall: London

[3] X. Braukhoff, A.Dimitropoulos J. Wagner. AndK. Salarnatian (2009).*Anomaly extraction in back bone networks using association rule.*Rio de Janerrio: Brazil.

[4] R. Fontugne , K. Fukuda & J. Himura (2010).*Estimated Speed and Scanning Activities with Hough-transform.*Sokendai: Japan.

[5] R. Funtugneand K.Fukuda (2011). *An analysis of longitudinal tcppassive measurements*.Retrieved November, 20, 2012. from http:// www.fukuda-lab.org/mawilab/.

[6] A. Lakina, M. Crovella& C. Diot (2004).*Diagnosing networking -wide traffic anomalies.* New York: Willy.

[7] A. Lakina , M.Crovella M. & C.Diot(2005).*Mining anomalies using traffic feature distribution.* New York: McGraw- Hill Companies.

[8] A. Lakina,K. Papaginnaki, M. Crovella, C.Diot, EN. Ko1aczk andN. Taft (2004).*Structure analysis ofNetwork traffic flows*. New York: Spring-Verlag.

[9] P.Owezarski (2010).*A database of anomalies traffic for accessing profile based.* Retrieved December 7, 2012 from http://www.abolene.iu.edu/routage.

[10] A. Soule,K. Samatian and N. Taft (2005).*Combining filtering and statistical methodfor anomaly detection.* New York: Wiley.

[11] M.Ji. Thottan, Y. Yan, and G.Liu (2003).Anomaly detection in IP networks.*International Journal of Computer Science and Communication Network*.51(8), 2191-2204.

[12] W.Wang,X. Zhang, W. Shin, and S. Jin (2002).*Detecting SYN flooding attack.*Retrieved December 20, 2012 from http://www.tcpdump.org/.

APPENDIX

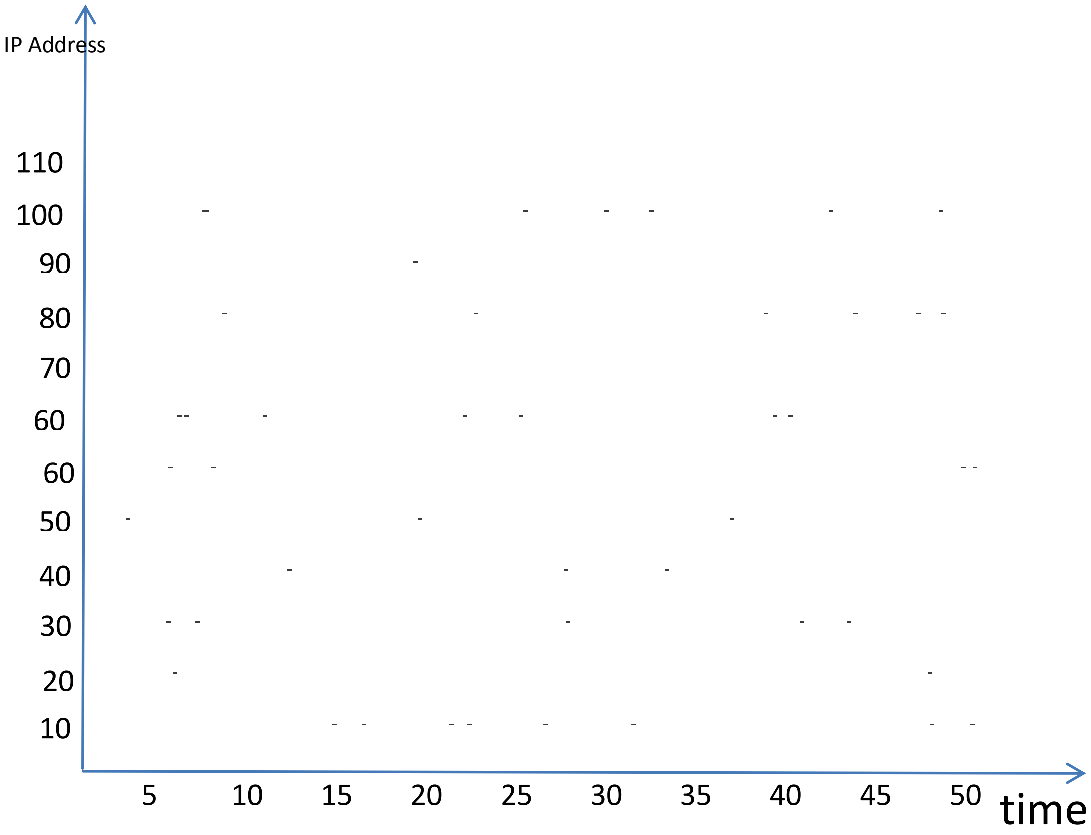
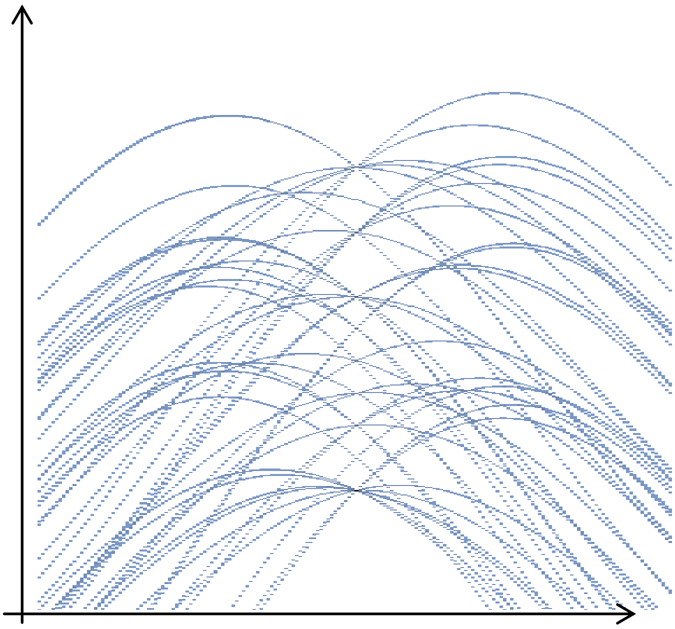


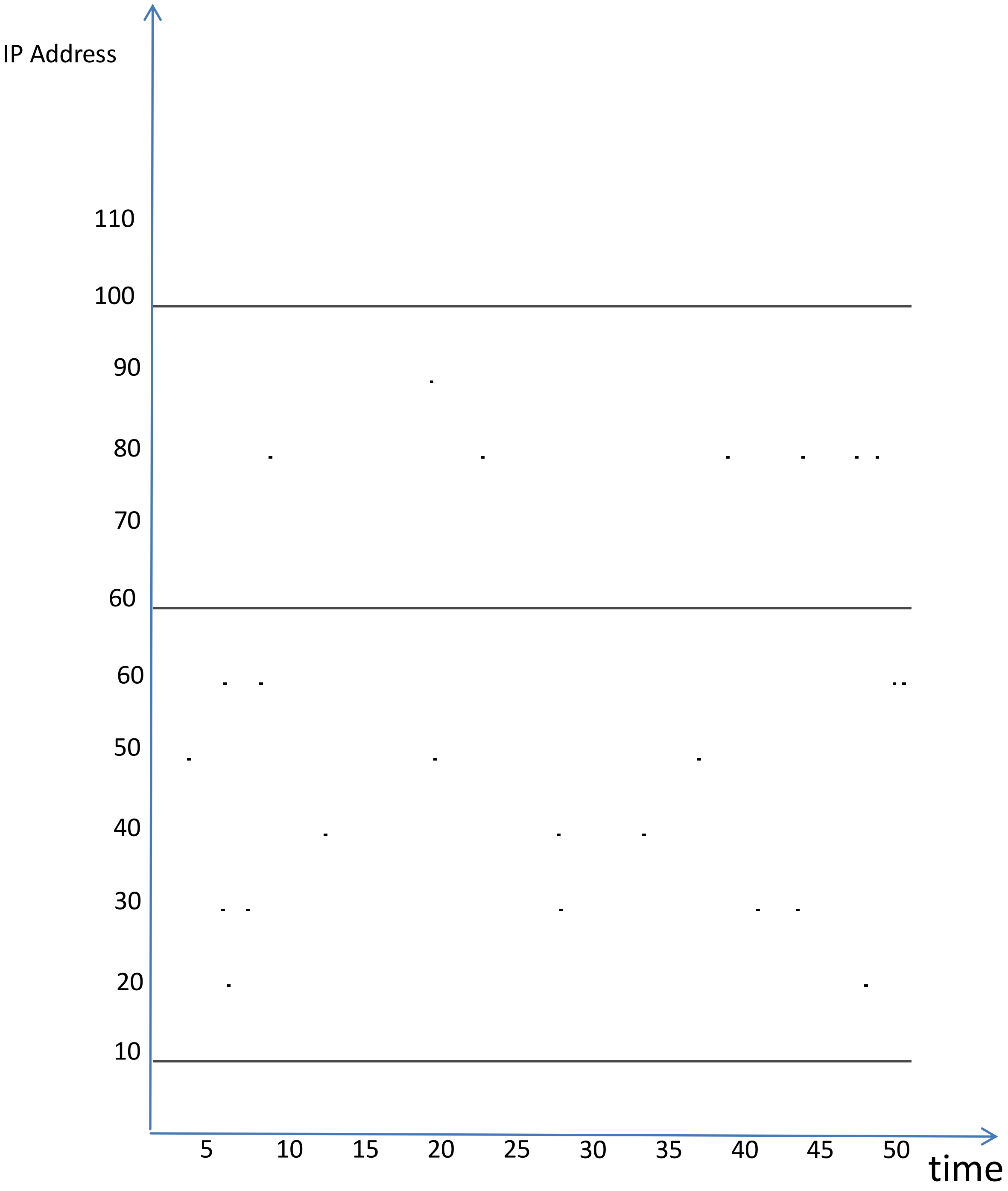
Fig. 4.1b: Hough Space

IP Address



Time

Fig. 4.1c: 2D Scatter Plot with detected line


IP Address
110
100
90
80
70
60
60
50
40
30
20
10
5
10
15
20
25
30
35
40
45
50
time

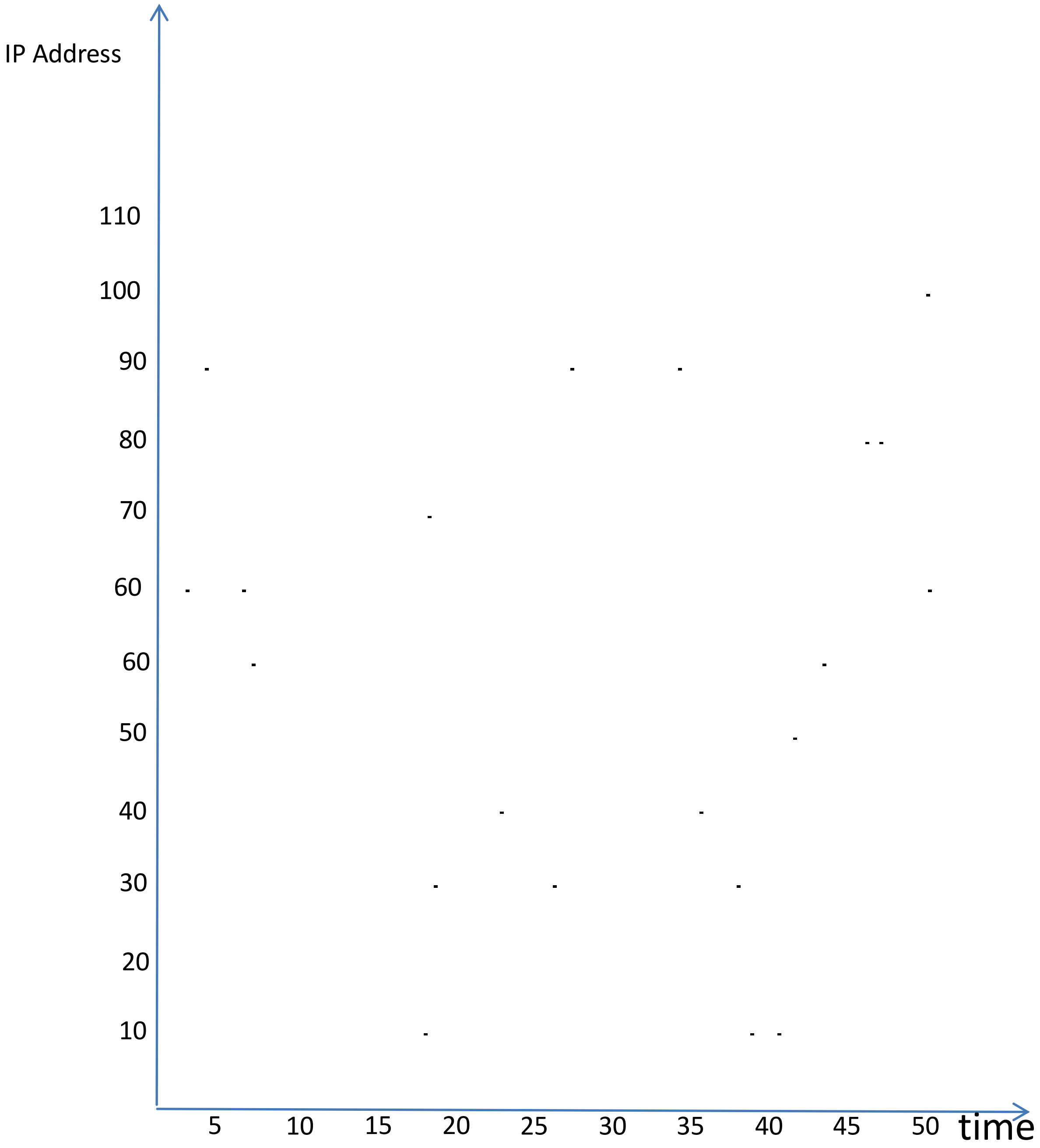
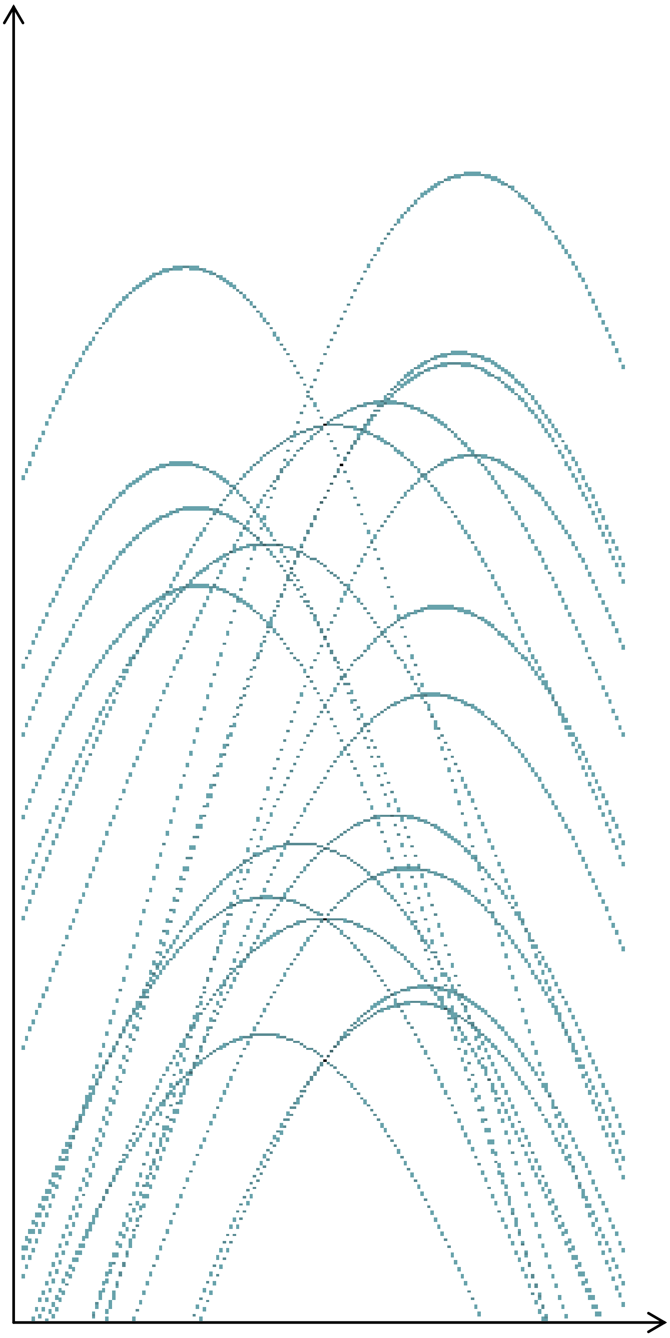
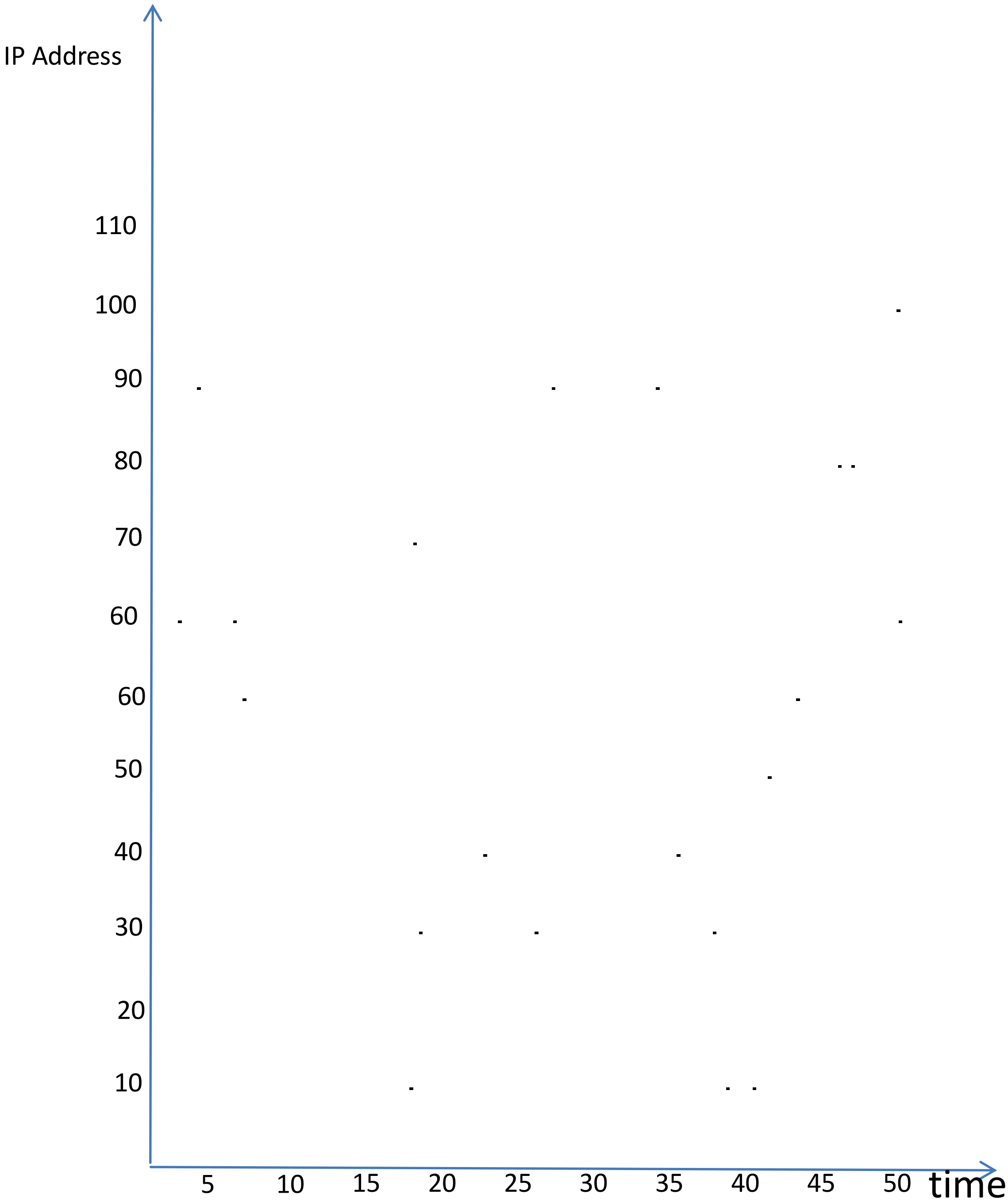


Fig. 4.2c: 2D Scatter Plot with no line detected


Time


IP Address
110
100
90
80
70
60
60
50
40
30
20
10
5
10
15
20
25
30
35
40
45
50
time

SHORT BIOGRAPHY

Maksha Daniel Didfee obtained his BSc degree in computer at Adamawa State University Mubi, Nigeria in 2010. Currently running MSc computer science same Adamawa State University, Mubi. His area of interest is computer network.

Doctor Peter Buba Zirra is Lecturer with Adamawa State University, Mubi Nigeria. He obtained his Doctorate degree in Computer Science from Modibbo Adama University of Technology, Yola in 2012, MSc in Computer Science from Abubakar Tafawa Balewa University, Bauchi in 2006, MBA (Finance) from University of Maiduguri, Borno state in 2000 and had his B.Tech in computer science, 1994 same AbubakarTafawaBalewa University, Bauchi. His area of interest includes Computer Network and Security. He is happily married with two children.