Machine Learning Approaches to Cyber Security

Shantanu Prasad Burnwal

A Thesis Submitted to

Indian Institute of Technology Hyderabad

In Partial Fulfillment of the Requirements for

The Degree of Master of Technology

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Department of Electrical Engineering

June 2016

Declaration

I declare that this written submission represents my ideas in my own words, and where ideas or words of others have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the Institute and can also evoke penal action from the sources that have thus not been properly cited, or from whom proper permission has not been taken when needed.

S P Burnwal.

(Signature)

Shantanu Prasad Burnwal.

(Shantanu Prasad Burnwal)

ee14mtech11037

(Roll No.)

Approval Sheet

This Thesis entitled Machine Learning Approaches to Cyber Security by Shantanu Prasad Burnwal is approved for the degree of Master of Technology from IIT Hyderabad

(Dr. Phanindra Jampana) Examiner
Dept. of Chemical Eng
IITH

(Dr. Sumohana Channappayya) Examiner
Dept. of Electrical Eng
IITH

(Dr. Ketan P Detroja) Adviser
Dept. of Electrical Eng
IITH

(Dr. M. Vidyasagar) Co-Adviser
Dept. of Electrical Eng
IITH

(Dr. K. Sri Rama Murty) Chairman
Dept. of Electrical Eng
IITH

Acknowledgements

.

Dedication

.

Abstract

Cyber-security is used to identify cyber-attacks while they are acting on a computer or network system to compromise security of the system. We discuss the concept of Hidden Markov Model with the Large Deviation Theory approaches because now a days statistical anomaly detection with Large Deviation theory approach have been used to find attack signatures in network traffic. We present two different approaches to characterize traffic: a model-free approach and a model-based approach. Model free approach is method of types based approach using Sanov's theorem whereas model based approach is HMM based approach uses Large deviation theory. We discuss how these theories can be applied for anomaly detection from network traffic. We study their effectiveness in anomaly detection. We will discuss how much these statistical methods affective on spatio-temporal traffic data. We also discuss about how length of traffic data affect our Markov model. How our estimated model is related with true but unknown model.

Contents

Chapter 1

Introduction

A cyber-infrastructure or cloud is a media for exchanging and utilizing information. Cloud contains digital data and its supporting infrastructure consist of softwares and hardwares, which are being utilized for traffic flow, data processing, privacy protection, monitoring, supervision and control etc. This makes it more important to keep it safe. Because many private , national and international essential and emergency services needs uninterrupted Internet supply. A cyber attack can be launched by some anti-social activists, which can compromise the security of any one of the following systems. Hence it may cause chaos, threat to our economy or national security etc. A normal or nominal user or a group of users are one who don't intend to intrude on the cyberspace of other user. To secure cyberspace of a user or cyber-infrastructure against these anti-social activists or threats cyber-security professionals and researchers has been engaged in to design a variety of defense systems. Researchers and Professionals are maintaining confidentiality, availability and integrity of data from various hackers. The confidentiality refers to the ability to save sensitive data from third parties, availability refers to do normal tasks like accessing free data or uploading data in cyberspace. Whereas integrity refers to the a complete cyber-infrastructure without any loopholes.

Generally operating systems like Windows have there own firewall and security system for protection against malicious cyber attacks or viruses. They have there own cryptography which protects user information. People also uses anti-virus softwares for their system or infrastructure for protection against threats. These approaches are used to create a shield for users. However these methods appears to be not fully protective because of flaws in design or flaws in hardware or software infrastructure. Researchers always tries to patch up those flaws but attackers always finds a way through these security systems. For all these reasons we need better methodology for a reliable cyber protection. Blocks of a better cyber defense system can be shown as:

As shown in figure 1.1 feature extraction, their analysis and decision making are most important steps of attack detection. Feature extraction contains all information like IP addresses of source and destination, protocol, ports of source and destination, time and duration of data, number of packets etc. Analysis part has different methods to detect anomalous behavior which system haven't seen before. The decision statement is made when analysis method catches some anomalous behavior.

Traditionally human analysts watch over these sequence of alerts given by cyber attack identification system and signals attack accordingly. But this is a difficult and time consuming task for an analyst when number of alerts generated are very high. This task becomes more difficult when en-

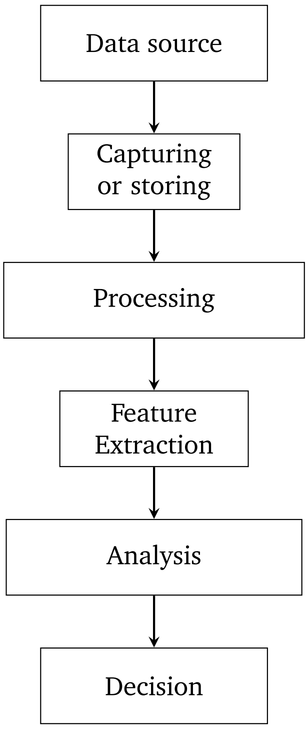


Figure 1.1: Steps of Cyber security

vironment changes rapidly. Machine learning is a good approach for cyber attack detection because it could deal with above problems effectively. First it gathers knowledge from the training data and then makes predictions on new data based on knowledge gained from previous data. This makes machine learning approach more efficient.

Cyber attack detection system (CADS) monitors the activities that occurs in a computing resource to detect violations of security policies of an organization. The intention of CADS can be summarized as follows:

- Increase attack detection rate also called True Positive (TP) i.e. Detect most of the attacks (Malicious or non-malicious, external opportunistic or deliberate attacks).
- Reduce false alarm rates also called False Positive (FP) i.e. accuracy of attack detection should be very high.
- Detect attacks in shortest time, thus to reduce damage caused by the attacks.

The above requirements have involved researchers to develop different machine learning algorithms that fulfill above goals to prevent systems from cyber attack.

Before proceeding lets talk what is machine learning. Machine learning is about automatic learning of models from sample data. It is a computational process where learning models uses rules or mathematical functions or logics for finding relationship between input and output. That is suppose we have an observed dataset X, parameters θ and a model $f(\theta)$. Then in machine learning we try to minimize errors $E(f(\theta), X)$ between truth and the learning model $f(\theta)$. In general we have a predicted output from the model $f(\theta)$ and observed sample data. We calculate the errors and use optimization algorithms to find accurate parameters θ for the accuracy of our learning model $f(\theta)$.

A CADS generally has to deal with problems like large network traffic volumes which is continuously varying with time because of human activity, highly uneven data and the difficulty to realize decision boundaries between normal and abnormal data and a requirement for continuous adaptation to a constantly changing environment. Generally the challenge is to efficiently classify behaviors of network systems. There are two general approaches for classifying behaviors of a network system for cyber attacks detection:

- Pattern recognition
- Anomaly detection

These two method work in complementation of each other.

1.1 Patttern Recognition

Pattern recognition is a supervised machine learning approach. Pattern-recognition techniques identify and store signature patterns of known attacks. Then they match the subject's observed behavior with patterns of known attack signatures and signal attacks when there is a match. Pattern-recognition techniques have been used in many places. Attack signatures can be characterized as strings, event sequences, activity graphs, or attack scenarios. Different types of rules, state transition diagrams or decision trees are used to match or identify pattern of attack signatures. Although pattern recognition is useful for known attacks but it can't stop novel or unusual attacks Therefore attack signature patterns needs to be updated timely manner to remain useful in this dynamic world where attack scenarios, protocol or configuration always keep on changing.

Pattern recognition approach uses different machine learning algorithms such as artificial neural networks(ANN), genetic programing (GP), support vector machines(SVM), Fuzzy rules based methods. Such methods uses Bayesian network (BN), classification and regression tree or decision tree based approach for feature extraction. These methods are being used for attack detection. But these all methods as told are ineffective towards novel attacks. These methods also have irregular performance for different attack types. So we need a different and little more concrete approach for attack detection which can overcome these characteristics.

1.2 Anomaly Detection

For a subject like a user or host machine or network of interest an anomaly-detection technique is consisting of establishing a norm profile, observe the activities of the subject and signal attacks when the subject's observed activities differ appreciably from its norm profile. So anomaly method can stop unusual and novel attack but it also have a drawback that it gives false alarm when it comes across an irregular behavior, which need not to be an anomaly. Pattern recognition and anomaly detection techniques can improve detection capability when used together. One of the technique is statistically anomaly detection technique is discussed in this report. It construct a statistical profile of a subject's normal activity from historic data. Many cyber-attacks require a series of related events to accomplish, it is possible to improve attack detection performance by incorporating the ordering of events. But statistical profile many times do not consider the in which event occurs to the subject. In a paper I have studied that Markov model are computationally intensive due to their use of the Bayes parameter estimation for learning the norm profile and a likelihood ratio test for inferring anomalies. Considering large amounts and high frequency of events in a computer and network system, those techniques are not applicable to cyber-attack detection in real time. Previously they had used Markov Chain (MC) model in [1], [2], for attack detection. But now a days as our computer's computational efficiency have been increased they have started researching with s-step Markov process. Markov chain process with s-step Markov process have been discussed in chapter 2 of this report.

1.3 Note

In this report anomaly detection method is described with some procedures because those papers contains details regarding anomaly detection method only. And also earlier work has showed that systems based on pattern matching had detection rates below 70% [3], [4] Furthermore, such systems need constant (and expensive) updating to keep up with new attack signatures. As a result, more attention has to be drawn to methods for traffic anomaly detection since they can identify even novel (unseen) types of anomalies.

Chapter 2

Literature Survey

A lot of research is done on Network anomaly detection. And still a lot of research is being carried out to make network anomaly detection perfect. In this section we will discuss approximately all anomaly detection methods that are in research and has been carried out till now. We will go through hows their modeling is done, how they work and their efficiency towards network anomaly detection. Let us start with the statistical approach then we will move towards the deterministic approach. As per statistical approach [2], [5], [6] and [7], we have found that statistical approach gives satisfactory result. Though research is being carried out to check whether these satisfactory results could provide perfect results. We have two types of statistical approaches.

- Model free approach
- Model based approach

In the upcoming sections we will see these approaches in details. Before moving forward let us see the network topology we need to monitor. We monitor a computer or a router for anomaly detection. So instead of calling them router, computer or traffic, lets say we are observing a system. Because these approaches we going to see, can be easily implemented on any of them. There are four different representations of data to be observed.

- Bytes
- Packets
- Flows
- Windows

A collection of Bytes forms a Packet, a collection of packets forms a flow and a collection of flow forms a window. We use flow and window based methods for attack detection. The traffic has variations in data flow due to human activity. Lets take an example of flow representation in traffic. Each flow consist of IP address (x), transmission time (t), size of the data in flow (d) and duration of flow (d_f). As we know each user is known by its IP address. When number of users becomes very high then it becomes nearly impossible to characterize the behavior of each user. So as to simplify this situation we form user groups to manage behavior. We use clustering algorithm to form user groups. Each user group has a cluster center and each user has spectral distance from the user. These clusters can be represented as:

$$f = (k(x), g(x), d, d_f) \tag{2.1}$$

Where $k(x)$ represents the cluster label and $g(x)$ represents the distance of cluster center from the user x. Suppose $x^i = (x_1^i, x_2^i, x_3^i, x_4^i) \in \{0, ..., 255\}^4$ and $x^j = (x_1^j, x_2^j, x_3^j, x_4^j) \in \{0, ..., 255\}^4$ are two i.p. addresses. Then their distance is given by $g(x^i, x^j) = \sum_{k=1}^{4} 256^{4-k}|x_k^i - x_k^j|$. $g(x)$ can be given as $g(x, \bar{x}^k)$. Where $\bar{x}^k$ represent kth cluster center. In this representation it is clear that we are using cluster label, distance between i.p. addresses, download rate or data transfer rate i.e. d/d_f, data size and duration of flow as features.

Whereas there is another representation of traffic. As we have seen that IP addresses was one way to form cluster center. But other then IP address clusters can also be formed based on number of flows. So we can represent network as:

$$f = (n_f(x), h(x), d, d_f) \tag{2.2}$$

Where $n_f(x)$ represents the number of flows user x have with the server or cluster center. $h(x)$ represents the distance of user from the server or cluster center. So here we can see that this method can be applied to those systems which work with a The first way of representation is used for techniques such as statistical and support vector machine anomaly detection. Whereas the second way of traffic representation is used for some other techniques such as Adaptive resonance theory based anomaly detection method.

Upcoming sections are arranged as follows. First we go through some important terms of HMM and LDT in section 2.1 and 2.2. Because these explanation will help us to understand statistical anomaly detection method. Section 2.3 and 2.4 describes those statistical approach i.e. model free approach and Markov model based approach for anomaly detection respectively. Section 2.5 describes how the above statistical approaches can be extended from temporal anomaly detection to spatio-temporal anomaly detection. In section 2.6 we will study about support vector machine approach for anomaly detection.

2.1 Hidden Markov Model

A stochastic process is defined as a sequence of random variables $X = x_1, ..., x_n$ for $n \to \infty$. As we know x takes values in from a finite set. Let's say the set A is a finite set of cardinality n. Set A can be written as $A = a_1, ..., a_n$. Here we will define a Markov chain model and their associated terms and then s-step Markov model.

2.1.1 Markov Chain

A process $\{X_t\}_{t=0}^{\infty}$ is said to be a Markov chain if for every $t \geq 0$ and every sequence $a_0, ..., a_t \in N^{t+1}$ we can say

$$Pr\{X_t = a_t | X_0 = a_0, ...X_{t-1} = a_{t-1}\} = Pr\{X_t = a_t | X_{t-1} = a_{t-1}\} \tag{2.3}$$

In a Markov process the length of the tail on which the conditioning is carried out is always one. In Markov process if the current state $i \in A$, the next state $j \in A$, then we can define the quantity $a_{ij}(t)$ as

$$a_{ij}(t) = Pr\{X_t = j|X_{t-1} = i\}, i, j \in A. \tag{2.4}$$

Where $a_{ij}(t)$ is the probability of making transition from the current state i at time t to the next state j at time $t+1$. A $n \times n$ matrix $A(t) = a_{ij}(t)$ is said to be state transition matrix of Markov process at time t. The matrix is said to be homogeneous if it is a constant matrix and independent of time and non-homogeneous otherwise. $A(t)$ is a stochastic matrix for all t. That is:

$$a_{ij}(t) \in [0, 1] \quad \text{and} \quad \sum_{j=1}^{\infty} a_{ij} = 1 \; \forall i, j \in A \tag{2.5}$$

Suppose $\{X_t\}$ is a Markov process in a state space N and we don't know state transition matrix. Let's say A is the unknown STM. We have observed Markov process sequence as $x_1^l = \{x_1, ..., x_l\}$. Now by this observation we would like to evaluate the terms of state transition matrix that makes this observed sequence most likely. So we evaluate likelihood of the observed sequence x_1^l if A is the state transition matrix of the underlying Markov processes. The likelihood can be written as:

$$L(x_1^l|A) = Pr\{x_1\} \prod_{t=2}^{l} Pr\{x_t|x_{t-1}\} = Pr\{x_1\} \prod_{t=2}^{l} a_{x_{t-1}x_t} \tag{2.6}$$

Taking log both sides and after further simplification it can be written as:

$$LL(x_1^l|A) = logPr\{x_1\} + \sum_{t=2}^{l} loga_{x_{t-1}x_t} = logPr\{x_1\} + \sum_{i=1}^{n}\sum_{j=1}^{n} v_{ij} loga_{ij}$$

Here v_{ij} denote the number of times state sequence ij appears in the sample path x_1^l. And we are writing precisely $a_{ij} = a_{x_{t-1}x_t}$ times. Now we have the constraint equation as well: $\sum_{j=1}^{n} a_{ij} = 1 \forall i \in N$. Then the updated objective equation can written with Lagrangian as:

$$LL(x_1^l|A) = logPr\{x_1\} + \sum_{i=1}^{n}\sum_{j=1}^{n} v_{ij} loga_{ij} + \sum_{t=1}^{n} \lambda_t (1 - \sum_{j=1}^{n} a_{ij})$$

Now we want to minimize this objective function. Then we get the following conclusions:

$$\frac{\partial J}{\partial a_{ij}} = \frac{v_{ij}}{a_{ij}} - \lambda_i = 0$$
$$\Rightarrow a_{ij} = \frac{v_{ij}}{\lambda_i}.$$

Now after further solving this equation we can get a closed form solution. Because it is readily available to us from the constrained equation:

$$\sum_{j=1}^{n} a_{ij} = 1 \Rightarrow \sum_{j=1}^{n} \frac{v_{ij}}{\lambda_i} = 1 \Rightarrow \lambda_i = \sum_{j=1}^{n} v_{ij} = \overline{v_i}$$

Therefore maximum likelihood estimate of the given matrix A is given by:

$$a_{ij} = \frac{v_{ij}}{\overline{v_i}} \tag{2.7}$$

Eq. (2.7) gives the state transition matrix values. But it contains a lot of zeros. To avoid zero entries in state transition matrix we do some adjustments in the matrix. Because the sample path with zero entries gives a likelihood zero or log-likelihood of minus infinity. To avoid this anomaly we make some adjustments. These adjustments are called Laplacian Correction.

2.1.2 s-step Markov process

s-step Markov process is also same as Markov chain with one difference. In s-step Markov process the length of the tail on which the condition is done is s instead of one as in Markov chain process. A process $\{X_t\}_{t=0}^{\infty}$ is said to be a s-step Markov if for every $t \geq 1$ and every sequence $a_0, ..., a_t \in N^{t+1}$ we can say:

$$Pr\{X_t = a_t | X_0 = a_0, ...X_{t-1} = a_{t-1}\} = Pr\{X_t = a_t | X_{t-s}^{t-1} = a_{t-s}^{t-1}\} \tag{2.8}$$

We can convert s-step Markov process to Markov chain process by changing the state space to N^s. We can define Z_t as $Z_t = X_{t-s+1}^{t}$. Then Z_t is defined over a block of s-states. Therefore Z_t is a Markov Process over state space N^s. Now equation (2.8) can be written as:

$$Pr\{Z_t = u | Z_{t-1} = v\} = Pr\{X_t = a_t | Zt - 1 = v\} \tag{2.9}$$

Here $u = (a_{t-s+1}, ..., a_t)$ and $v = (a_{t-s}, ..., a_{t-1})$. As v is previously known so first $(s-1)$ terms of u are already known. Therefore 2.9 is valid from that point of view. As $u, v \in N^s$. So this time state transition matrix can be written as $n^s \times n^s$ matrix. Here only n components in a row will be non-zero and other entries are zero because components corresponding to v are only non-zero.

2.1.3 Recurrent and Transient states

A state is said to be recurrent state in a Markov process if during transition from one state to another state, the process attain a state where it can't further attain any other state and form a loop for all $t \geq k$. Here k represents the time instant at which the state attains recurrent state. Set of all states excluding recurrent states are called transient states. Now as it was discussed earlier that $\{X_t\}$ is a Markov process assuming values in a finite set N of cardinality n and $A(t)$ denote the state transition matrix at time t. If c_0 is its initial distribution. Then state $A(t)$ is distributed according to

$$A(t) = c_0 A(1) A(2), ..., A(t-1) \tag{2.10}$$

Now if the system attains a recurrent state at time T. If π is that stationary state or vector then for all $t \geq T$ we can write the above equation as:

$$\pi \times A(T) = A(t)$$

Here π is called the stationary distribution of $A(t)$. Now we want to know whether a given stochastic matrix A have a stationary distribution and if it have then how to determine the set of all stationary distributions. For that we need to follow a theorem given in [[8], Theorem 4.7]. Some part of it is described here. For full details please go through the reference.

2.1.4 spectral radius

Given a matrix A, the spectrum of A consists of all eigenvalues of A and is denoted by $spec(A)$. If $spec(A) = \lambda_1, ..., \lambda_n$, then

$$\rho(A) = max\{|\lambda_i| : \lambda_i \in spec(A)\}$$

is called the spectral radius of A. $Spec(A)$ can't contain negative and complex numbers. $\rho(A)$ is always real and non-negative.

2.1.5 Canonical Form of a matrix

Suppose $A \in \Re_+^{n\times n}$ is a non-negative matrix. Then to capture the location of positive elements in the matrix we form an incidence matrix T corresponding to A:

$$\begin{aligned} T \in \{0,1\}^{n\times n}, t_{ij} &= 1 \; \forall a_{ij} > 0 \\ &= 0 \quad \forall a_{ij} = 0 \end{aligned}$$

Since A has n rows and columns, we can think of $N := 1, ..., n$ as the set of nodes of a directed graph and place a directed edge from node i to node j if and only if $t_{ij} = 1$. A path length l from node i to node j is a sequence of length $l+1$ of the form $\{i_0, ..., i_l\}$, where $i_0 = i, i_l = j$, and in addition $a_{i_s i_{s+l}} > 0$ for all $s = 0, 1, ...l$. Now if a^l_{ij} denote the $ij_t h$ element of A^l. Therefore from matrix multiplication formula:

$$a^l_{ij} = \sum_{s_1=1}^{n} ...... \sum_{s_{l-1}=1}^{n} a_{is_1} a_{s_1 s_2} ....a_{s_{l-1}j}$$

Now it is obvious that $a^l_{ij} > 0$ if and only if there is a path from node i to node j. A node i is said to be inessential if there exists a node j (of necessity, not equal to i) such that $i \to j$, but $j \not\to i$. Otherwise, i is said to be essential. With the above definition, we can divide the set of nodes $N = \{1, ..., n\}$ into two disjoint sets: $\boldsymbol{\iota}$ denoting the set of inessential nodes and $\boldsymbol{\varepsilon}$ denoting the set of essential nodes. Now after similarity transformation A can be converted in such a way that all nodes in $\boldsymbol{\varepsilon}$ comes first, followed by all nodes in $\boldsymbol{\iota}$. Let Π denote the permutation matrix then the matrix A can be written as:

$$\Pi^{-1} A \Pi = \begin{array}{ccc} & \boldsymbol{\varepsilon} & \boldsymbol{\iota} \\ \boldsymbol{\varepsilon} & P & 0 \\ \boldsymbol{\iota} & R & Q \end{array} \tag{2.11}$$

Then from [[8], Theorem 4.7, statement 8] it can be written as: If $c \in S_n$ be an arbitrary initial distribution. If $\boldsymbol{\iota}$ is non-empty, permute the component of c to be compatible with eq. (2.10), and write $c = [c_\varepsilon c_\iota]$. Partition $c_t = cA^t$ as $c_t = [c_\varepsilon^{(t)} c_\iota^{(t)}]$. Then $c_\iota^{(t)} \to \infty$ as $t \to \infty$, irrespective of c.

To proof this as its written above if $\boldsymbol{\iota}$ is non-empty, and $m = |\boldsymbol{\iota}|$ denote the cardinality of $\boldsymbol{\iota}$. Then partition A^m as:

$$A^m = \begin{array}{ccc} & \boldsymbol{\varepsilon} & \boldsymbol{\iota} \\ \boldsymbol{\varepsilon} & P^m & 0 \\ \boldsymbol{\iota} & R^{(m)} & Q^m \end{array}$$

Then we can say that each row of $R^{(m)}$ contains a nonzero element. Then we can say each row sum of Q^m is strictly less than one. Now as we have seen earlier that $\rho(A) \leq r(A) = \mu(A) \leq 1$ [8]. Where A is a stochastic matrix with each row sum equal to one. In this case $\mu(Q^m) \leq 1$. So we can say that $\rho(A) \leq 1$. Since $\rho(Q^m) = \rho(Q)^m$, it follows that $\rho(Q) \leq 1$. If A is in the form eq. (2.10) and $\boldsymbol{\iota}$ is non-empty, then A^t has the form:

$$A^t = \begin{array}{ccc} & \boldsymbol{\varepsilon} & \boldsymbol{\iota} \\ \boldsymbol{\varepsilon} & P^t & 0 \\ \boldsymbol{\iota} & R^{(t)} & Q^t \end{array}$$

Now we know that $\rho(Q) \leq 1$ from the above explanation. And also $Q^t \to \infty$ as $t \to \infty$. Hence if we write $c^t = cA^t$, then it implies that:

$$c^t_{\boldsymbol{\iota}} = c_{\boldsymbol{\iota}} Q^t \to 0 \quad \forall t \to \infty$$

Irrespective of the value $c_{\boldsymbol{\iota}}$have initially. Therefore we can conclude that the states in $\boldsymbol{\iota}$ are referred as transient states and those in $\boldsymbol{\varepsilon}$ are referred as recurrent states. Therefore we can conclude that irrespective of the initial distribution of the Markov chain, we have:

$$Pr\{X_t \in \boldsymbol{\iota}\} \to 0 \quad \forall t \to \infty$$
$$Pr\{X_t \in \boldsymbol{\varepsilon}\} \to 1 \quad \forall t \to \infty$$

So we can say that set $\boldsymbol{\varepsilon}$ are recurrent states, and these disjoint equivalence classes $\boldsymbol{\varepsilon}_1, ..., \boldsymbol{\varepsilon}_s$ are referred as communicating classes.

2.2 Large Deviation Theory

In this section we will discuss about Sanov's theorem [8] for i.i.d. sequence as well as Markov chain sequence. For that earlier we need to define a few things, which are as follows. Suppose $A = \{a_1 ....a_n\}$ is a finite set. $M(A)$ denote the set of all probability distributions on the set A. $\mu \in M(A)$ is a fixed but possibly unknown probability distribution and X is a random variable assuming values in A with distribution μ. In order to estimate μ we generate independent samples $x_1.....x_l....$, where each sample x_i belongs to set A, and distributed according to μ and independent of x_j for $j \neq i$. Let we have generated samples from first l experiments. Then we used to construct empirical distribution as follows:

$$(\hat{\mu}^l_1)_i = \frac{1}{l} \sum_{j=1}^{l} \mathbf{I}_{\{} x_j = a_i\} \tag{2.12}$$

Where **I** denote the indicator function. Now $(\hat{\mu})_{x=1}^{l}$ is also a probability distribution on A and a random element of $M(A)$. Thus we think as $l \to \infty$ the process converges to the true measure μ that is generating the samples. Let us define $\gamma \in M(A)$ is some set of probability distribution and $(\hat{\mu})_{x=1}^{l} \in \gamma$ is a sequence of real numbers. Now suppose that $\mu \ni \bar{\gamma}$, where set $\bar{\gamma}$ denotes the closure of set γ in the total variation metric. Thus the true measure μ that is generating the random samples does not belong to the closure of set γ. If $\mu \ni \hat{\mu}$ then we can say that the sequence of real numbers $Pr\{\hat{\mu}\}_{x=1}^{l}$ converges to zero. Large deviation theory tells us the rate at which the sequence converges to zero and how the rate depends on γ and μ. Suppose this sequence converges to zero at an exponential rate, that is

$$Pr\{\hat{\mu}(x_1^l) \in \gamma\} = c_1 exp(-lc_2)$$

Here c_2 is called rate of convergence. Now taking log both side we can write:

$$\frac{1}{l} logPr\{\hat{\mu}(x_1^l) \in \gamma\} = \frac{logc_1}{l} - c_2$$

As we can see that as $l \to \infty$ the negative of this quantity approaches to c_2. Now let us define the rate function.

2.2.1 Rate function

Let $\hat{\mu}(x_1^l)$ be defined as in (2.12). Then the function $I : M(A) \to \Re_+$ is said to be a rate function of the stochastic process $\hat{\mu}(x_1^l)$ if

- I is lower semi-continuous.
- For every set $\gamma \subseteq M(A)$, the relationship holds:

$$-\inf_{v\in\gamma^0} I(v) \leq \lim_{l\to\infty} \inf \frac{1}{l} logPr\{\widehat{\mu(x_1^l)} \in \gamma\} \leq \lim_{l\to\infty} sup \frac{1}{l} logPr\{\hat{\mu}(x_1^l) \in \gamma\} \leq -\inf_{v\in\bar{\gamma}} I(v) \tag{2.13}$$

Where γ^0 denote interior of the set γ. Again a function $f : M(A) \to \Re$ is said to be lower semi-continuous if

$$v_i \to v^* \to f(v^*) \leq \lim_i \inf f(v_i)$$

2.2.2 Sanov's theorem

Let us first understand method of types. Let us fix the integer l denote the length of the multi-sample. So now empirical distribution can take finite values as shown in equation (2.12), and every element of $\hat{\mu}(x)_1^l$ is a rational number with l in denominator. Let $\varepsilon(l, n)$ denote the set of all possible empirical distribution that can result from the multi-sample of length l and set $A = a_1......a_n$ has cardinality n. And we define two multi-sample to be equivalent if their empirical estimates are equal. $T(v, l)$ represent type class v of the multi-sample. Then we can say that [8]:

- Cardinality of $|\varepsilon(l,n)|$ is given by:

$$|\varepsilon(l,n)| = \frac{(l+n-1)!}{l!(n-1)!} \tag{2.14}$$

- The log likelihood of each multi-sample in $T(v,l)$ is given by:

$$logP_{\mu}^{l}\{x\} = -lJ(v,\mu) \tag{2.15}$$

Where $J(v,\mu)$ known as loss function. (Defined below.)

- The cardinality of $T(v,l)$ is given by:

$$|T(v,l)| = \frac{l!}{\prod_{i=1}^{n} l_i} \tag{2.16}$$

- Lower and upper bound of $|T(v,l)|$ is given by:

$$|\varepsilon(l,n)|^{-1}\exp[lH(v)] \leq |T(v,l)| \leq \exp[lH(v)], \quad \forall v \in \varepsilon(l,n) \tag{2.17}$$

Where $H(v)$ denote entropy of distribution v.

2.2.3 Loss function

Loss function between two probability measure μ and v is defined as:

$$J(v|\mu) = \sum_{i=1}^{n} v_i \log \frac{1}{\mu_i}$$

Where n represents cardinality of set $A = a_1, ...a_n$. This function have minimum value J_{min} when $v = \mu$. Otherwise the difference between this $Jmin$ and J (when $v \neq \mu$) is known as Kullback Laibler divergence i.e.

$$D(v||\mu) = J(v|\mu) - J(v|v)[when v = \mu, J_{min}]$$

2.2.4 Sanov's Theorem for i.i.d. processes

The stochastic process $\hat{\mu}(x)_1^l$ satisfies Large deviation property with the rate function $I(v) = D(v||\mu)$. Where $D(v||\mu)$ is called Kullback Laibler Divergence. Proof: As first condition for rate function is it should be lower semi-continuous function. But here $I(v)$ is not only lower semi-continuous but in fact its continuous function v. So now we need to proof the second criteria i.e. inequality criteria as told in section 2.1 point 2. Suppose $v \in \varepsilon(l,n)$. Now we need to evaluate the probability of empirical distribution $\hat{\mu}(x_1^l)$ to be equal to v.

$$Pr\{\hat{\mu}(x_1^l) = v\} = Pr_{\mu}^{l}(T(v,l)) = |T(v,l)|Pr_{\mu}^{l}(\{x_1^l\})$$

Here the sequence $x_1......x_l$ is generated by some unknown frequency μ we are trying to estimate and v is the estimated frequency, i.e. $\{x_1^l\} \in T(v,l)$. Now from equation (2.14) and from right inequality of equation (2.17) we can write:

$$Pr\{\hat{\mu}(x_1^l) = v\} = |T(v,l)| Pr_\mu^l(\{x_1^l\}) \leq \exp|lH(v)| \exp -lJ(v,\mu) = \exp[-lD(v||\mu)] \qquad (2.18)$$

Similarly using (2.15) and the left inequality (2.17) we can write:

$$Pr\{\hat{\mu}(x_1^l) = v\} = |T(v,l)| Pr_\mu^l(\{x_1^l\}) \geq |\varepsilon(l,n)|^{-1} \exp|lH(v)| \exp -lJ(v,\mu) = |\varepsilon(l,n)|^{-1} \exp[-lD(v||\mu)] \qquad (2.19)$$

Combination of equation (2.18) and (2.19) we can write it as:

$$|\varepsilon(l,n)|^{-1} \exp[-lD(v||\mu)] \leq Pr\{\hat{\mu}(x_1^l) = v\} \leq \exp[-lD(v||\mu)]$$

This above equation gives the lower limit and upper limit for $\hat{\mu}(x_1^l) \in T(v,l)$. Now since $\gamma \in M(A)$ is any probability distribution on A. Then using (2.19)

$$\begin{aligned} Pr\{\hat{\mu}(x_1^l) \in \gamma\} &= \sum_{v \in \gamma \cap \varepsilon(l,n)} Pr\{\hat{\mu}(x_1^l) = v\} \\ &\leq |\gamma \cap \varepsilon(l,n)| \sup_{v \in \gamma \cap \varepsilon(l,n)} Pr\{\hat{\mu}(x_1^l) = v\} \\ &\leq |\varepsilon(l,n)| \sup_{v \in \gamma} \exp[-lD(v||\mu)] \end{aligned}$$

Hence it can be written by taking log both sides and divide by l throughout the equation:

$$\frac{1}{l} \log Pr\{\hat{\mu}(x_1^l) \in \gamma\} \leq \frac{1}{l} \log|\varepsilon(l,n)| + \sup_{v \in \gamma} -D(v||\mu) \qquad (2.20)$$

As we know the number of experiments should be a large number. So we can write that $l >> n$. So we can write that $|l+n-1| \leq |2l|$. But if we think that l must be at least equal to n. Then (2.14) can be modified as:

$$|\varepsilon(l,n)| = \frac{(l+n-1)!}{l!(n-1)!} \leq \frac{(2l)!}{(n-1)!l!} = \frac{(2l)^{(n-1)}}{(n-1)!} \quad \forall l \geq n$$

So we can say from above equation that as $l \to \infty$ first term of equation 2.20 approaches zero. Since $\gamma \subseteq \hat{\gamma}$ so the rest of the equation can be written as:

$$\lim_{l \to \infty} \sup \frac{1}{l} \log Pr\{\hat{\mu}(x_1^l) \in \gamma\} \leq \sup_{v \in \gamma} -D(v||\mu) \leq -\inf_{v \in \hat{\gamma}} D(v||\mu)$$

This constitutes the right hand part of the (2.13) as given in the definition of the rate function. Now to prove left hand part of the definition as given in (2.13) we assume ϕ is a probability distribution in $M(A)$. We have got this distribution after performing l experiments. So we are assuming $\phi \in M(A)$ is same as $v \in M(A)$, but two outcome of same experiment. So we can say that every $\phi_i \in \phi$ and every $v_i \in v$ is a rational number with denominator l. Now $l\phi_i$ and lv_i both are integer values. Now let us define $q_i = \lceil l\phi_i \rceil$ and $r_i = \lfloor l\phi_i \rfloor$. Now if we assume c_i equal to either q_i or r_i for each index i. Then we can say that $\sum_{i=1}^n c_i = l$. Then we can say that type class v can be written as $v_i = c_i/l$.

Now from the above equation it can be written as radius of total variation metric ϕ as:

$$\rho(\phi, v) = \frac{1}{2}\sum_{i=1}^{n} |\phi_i - c_i| \leq \frac{n}{2l}$$

Now suppose v is an interior point in γ and $B(v)$ represents open ball in γ that contains v. Let there exists a sequence $\{v_l\} \in \varepsilon(l, n)$ for all l. And $v_l \to v$ as $l \to \infty$. Also $v_l \in (v)$ for sufficiently large l. Then from (2.19) we can write:

$$Pr\{\hat{\mu}(x_1^l) \in \gamma\} \geq Pr\{\hat{\mu}(x_1^l) = v_l\} \geq |\varepsilon(l, n)|^{-1} \exp\left[-lD(v_l||\mu)\right]$$
$$\frac{1}{l}\log Pr\{\hat{\mu}(x_1^l) \in \gamma\} \geq -\frac{1}{l}\log|\varepsilon(l, n)| - D(v_l||\mu)$$

As seen earlier that the first term on the right hand side approaches zero as $l \to \infty$. Now suppose $\gamma^0 \subseteq \gamma$.Then we can write:

$$-\inf_{v_l \in \gamma} D(v||\mu) \geq -\inf_{v_l \in \gamma^0} D(v||\mu)$$

So as $l \to \infty$ we have seen earlier that $v_l \to v$. Then above equation can ve written as:

$$\lim_{l \to \infty} \frac{1}{l}\log Pr\{\hat{\mu}(x_1^l) \in \gamma\} \geq -\inf_{v \in \gamma^0} D(v||\mu)$$

This proves the left inequality of the definition of the rate function. Therefore it can be concluded that Kullback-Laibler is the rate function for finite alphabets.

2.2.5 Large deviation theory for Markov chain

Previously we studied Sanov's theorem for an i.i.d. sequence. Here we will discuss Sanov's theorem for Markov chain sequence. Because Sanov's theorem is key in cyber security project. As we have seen earlier that if we have s-step Markov process of cardinality N. Then we can transform that s-step Markov process into Markov chain process of cardinality N^s. In this new Markov chain only transition matrix gets converted from $n \times n$ to $n^s \times n^s$.

2.2.6 Entropy rate

Suppose $\{X_t\}_{t \geq 0}$ is a stationary stochastic process assuming values in finite alphabet $A = \{a_0, ..., a_n\}$. Then entropy rate of the process is given by $H_r(X_t)$, and defined as [8]:

$$H_r(\{X_t\}) = \lim_{t \to \infty} H\left(\frac{X_t}{X_0^{t-1}}\right) \tag{2.21}$$

Where $H(.)$ is the conditional entropy. Suppose X and Y are random variables assuming values in finite set $A = a_1......a_n$ and $B = b_1......b_m$ respectively. $\phi \in S_{nm}$ denote their joint probability distribution. $\phi_x \in S_n$ and $\phi_y \in S_m$ denote the marginal probabilities of variables X and Y respectively.

Then conditional entropy of Y with respect to X is given by:

$$H\left(\frac{Y}{X}\right) = \sum_{i=1}^{n}(\phi_x)_i H(\phi_{Y|x=a_i})$$

It can also be written as:

$$H(Y/X) = H(X,Y) - H(Y)$$

Therefore as per given definition of Entropy rate and property of conditional probability it can be written as:

$$H_r(\{X_t\}) = \lim_{t\to\infty} H(X_t/X_0^{t-1}) = \lim_{t\to\infty} H(X_0^t) - H(X_0^{t-1})$$

Here we need to know that if $\{X_t\}_{t\geq 0}$ is a stationary stochastic process assuming values in a finite alphabet A. Time average of Entropy rate approaches a constant value $c \geq 0$ as $t \to \infty$. That is [8]:

$$H\left(\frac{X_t}{X_0^{t-1}}\right) \to c \qquad \forall t \to \infty \tag{2.22}$$

Entropy rate for a Markov process is given by [8]:

$$H_r(v) = \sum_{i=1}^{n} \bar{v}_i H(a_i) \tag{2.23}$$

2.2.7 Relative Entropy rate

Suppose $\{X_t\}$ and $\{Y_t\}$ are two stationary stochastic processes assuming values in a common finite set A. Then the relative entropy rate is defined as:

$$D_r(\{X_t\}||\{Y_t\}) = \lim_{t\to\infty} \frac{D(X_0^t||Y_0^t)}{t} \tag{2.24}$$

If limit exists then only relative entropy rate will be defined otherwise not defined. Like entropy rate, relative entropy rate can also be written as [8]:

$$D_r(v||\mu) = D(v||\mu) - D(\bar{v}||\bar{\mu})$$

Here $v, \mu \in M(A^k)$ probability distribution on A^k. Whereas $\bar{v}, \bar{\mu} \in M^{k-1}$ are probability distribution on A^{k-1}. There is a difference of one dimension in between them. Mathematically it can be represented as:

$$\bar{v}_i = \sum_{j\in A} v_{ij} = \sum_{j\in A} v_{ji}$$

Like entropy rate, relative entropy rate of Markov process is given by [8]:

$$D_r(v||\mu) = \sum_{i=1}^{n} \bar{v}_i D(a_i||b_i) \tag{2.25}$$

Where a_i and b_i are ith row of transition matrix A and B.

2.2.8 Rate function for Markov Process

Here we discuss about the doublet frequency because every frequency can be further transformed into doublet frequency. $\{X_t\}$ is a Markov process from a finite set called $A \in S_n$. After observing l experiments we have an output $x_1......x_l$. Let ϕ_1^l denote the empirical estimate of l observations. Then we can write:

$$\phi_i(x_1^l) = \frac{1}{l}\sum_{j=1}^{l} I_{x_j=i} \quad \forall i \in A \tag{2.26}$$

Here ϕ is an approximation of stationary distribution of the Markov chain. Since we are working on doublet frequency. So we can define:

$$\theta_{ij} = \frac{1}{l-1}\sum_{k=1}^{l-1} I_{x_k X_{k+1}=ij} \tag{2.27}$$

Here we can say that $\theta \in M(A^2)$ is an estimate of the true estimate $\phi \in M(A^2)$ which are generating doublet frequencies. As defined earlier that as $\theta \in S_n^2$, therefore $\bar{\theta} \in S_n$. But $\bar{\theta}$ is not a stationary distribution. Because as per definition of stationary distribution:

$$\bar{\theta}_i = \sum_{j=1}^{n} \theta_{ij} \neq \sum_{j=1}^{n} \theta_{ji}$$

As we can see that its a non-stationary function. We make this a stationary function by making an assumption that the outcome of $(l+1)th$ is the first observation and make it a cyclic path. Therefore the sample path-length will be $l+1$. So the new estimate is given by:

$$v_{ij} = \frac{1}{l}\sum_{k=1}^{l} I_{x_k x_{k+1}=ij} \tag{2.28}$$

Here v is a always stationary unlike θ. Here again we start with method of types to evaluate rate function for Markov process. Suppose $\hat{\mu}(x_1^l)$ is the empirical measure as defined in (2.28). If two observations have the same empirical estimates then we can say that they are of same type. Let us define $\varepsilon(l,n,2) \subseteq \varepsilon(l,n^2)$. Since every distribution in $\varepsilon(l,n^2)$ are not stationary. So we have selected $\varepsilon(l,n,2)$ of sample length l and of finite alphabet of length l. As earlier defined $T(v,l)$ define the type class $v = \hat{\mu}(x_1^l) \in \varepsilon(l,n,2)$. Therefore we can say [8]:

- The cardinality of $\varepsilon(l,n,2)$ is given by:

$$|\varepsilon(l,n,2)| \leq (l+1)^{n^2} \tag{2.29}$$

- The cardinality of type class $T(v,l)$ is given by:

$$(2l)^{2n^2} \exp lH_r(v) \leq |T(v,l)| \leq l \exp lH_r(v) \tag{2.30}$$

Where $H_r(.)$ is the entropy rate function.

- The log likelihood of $T(v,l)$ is given by:

$$\log PrX_1^l = x_1^l = l[J(v|\mu) - J(\bar{v}|\bar{\mu})] \tag{2.31}$$

Here v represents $v(x_1^l)$ and $\bar{v}$ represents $\bar{v}_1^l$ as defined in 2.28.

From (2.30) and (2.31) we can tell that:

$$\begin{aligned} \log|T(v)| &\leq lH_r(v) \\ \log|T(v,l)| &\leq l(J(v|\mu) - J(\bar{v}|\bar{\mu})) \end{aligned}$$

So from the above equations if we write that:

$$\delta(v,l) = \frac{1}{l}\log Pr\{\hat{\mu}(x_1^l) = v\}$$

Then it can be written as:

$$\begin{aligned} \delta(v,l) &\leq H_r(v) - J(v|\mu) + J(\bar{v}|\bar{\mu}) + o(1/l) \\ &= H(v) - H(\bar{v}) - J(v|\mu) + J(\bar{v}|\bar{\mu}) + o(1/l) \\ &= -D_r(v||\mu) + o(1/l) \end{aligned} \tag{2.32}$$

similarly it can be written using the left inequality of (2.30) and (2.31) as:

$$\delta(v,l) \geq -D_r(v||\mu) + o(1/l) \tag{2.33}$$

Therefore we can say that if $\gamma \subseteq M(A^2)$ be any set of the probability distribution in A^2. Then we can write:

$$\begin{aligned} Pr\{\hat{\mu}(x_1^l) \in \gamma\} &= \sum_{v\in\varepsilon(l,n,2)\cap\gamma} Pr\{\hat{\mu}(x_1^l) = v\} \\ &\leq |\varepsilon(l,n,2)\cap\gamma| \sup_{v\in\varepsilon(l,n,2)\cap\gamma} Pr\{\hat{\mu}(x_1^l) = v\} \\ \frac{1}{l}\log Pr\{\hat{\mu}(x_1^l) \in \gamma\} &\leq \frac{1}{l}\log|\varepsilon(l,n,2)| + \sup_{v\in\gamma}\delta(l,v) \\ \limsup_{l\to\infty}\frac{1}{l}\log Pr\{\hat{\mu}(x_1^l) \in \gamma\} &\leq \sup_{v\in\gamma} -D_r(v||\mu) = -\inf_{v\in\gamma} D_r(v||\mu) \end{aligned}$$

This is the right inequality of the rate function as given in (2.13). Now suppose $v \in \gamma$. $B(v)$ is an open ball in $M(A^2)$ that contains v. Now there exists a sequence of elements $\{v_l\} \in \gamma \cap \varepsilon(l,n,2)$. Now as we know as $l \to \infty$,$v_l \to v$. Therefore we can write that:

$$\begin{aligned} Pr\{\hat{\mu}(x_1^l) \in \gamma\} &\geq Pr\{\hat{\mu}(x_1^l) = v_l\} \\ \frac{1}{l}\log Pr\{\hat{\mu}(x_1^l) \in \gamma\} &\geq \delta(l,v_l) \\ &\geq -D_r(v_l||\mu) + o(1/l) \end{aligned}$$

Therefore it can be written that:

$$\lim_{l\to\infty}\inf \frac{1}{l}\log Pr\{\hat{\mu}(x_1^l)\in\gamma\} \geq -D_r(v||\mu) \quad \forall v\in\gamma^0$$
$$\lim_{l\to\infty}\inf \frac{1}{l}\log Pr\{\hat{\mu}(x_1^l)\in\gamma\} \geq \sup_{v\in\gamma^0} -D_r(v||\mu) = -\inf_{v\in\gamma^0} -D_r(v||\mu)$$

Where $\gamma^0 \subseteq \gamma$. So this proves the left inequality of the rate function as given in (2.13). Therefore we can conclude that relative entropy rate $(D_r(v||\mu))$ act as a rate function in Markov processes.

2.3 Model Free Approach

Model free approach is a statistical approach. Here we treat the sequence of data flow through the system is an i.i.d. sequence. As per this approach sliding window is applied on the sequence of data. Let $f^i = (k(x^i), g(x^i), d^i, d_f^i)$ represents a flow attribute in the sequence of data flow as discussed earlier in network traffic data representation part. Lets say $F_j = \{f^1, f^2, ..., f^{F_j}\}$ represents sequence of flows in j^{th} window. Similar to F_j we assume that we have used F_{ref} for training or as reference. Now first we quantize each reference flow of reference windows to to a closest symbol in discrete alphabet say $\sum$. Lets say these discrete symbols are represented as $\sum(\omega) = \{\sigma(f^1), \sigma(f^2), ..., \sigma(f^{F_j})\}$.

Now empirical measure of current flow sequence $F = \{f^1, f^2, ..., f^F\}$ is evaluated as frequency distribution vector:

$$\boldsymbol{\mathcal{E}}^F(\rho) = \frac{1}{|F|}\sum_{i=1}^{|F|} 1\{\sigma(f^i) = \rho\} \tag{2.34}$$

Where $1\{.\}$ denotes the indicator function and $\sigma(f^i)$ denotes the flow state in $\sum$ that f^i maps to. Now as we know that we have probability vector from reference flows F_{ref}. And just now we have calculated probability vector from empirical measure using (2.34). Lets say that $\mu(\sigma)$ is the reference probability measure of flow state σ. Then we use Sanov's theorem [8] to compare these probability vectors.

Suppose ν is the estimated probability vector after quantization of flows. Then with the help of Sanov's theorem we can check normality with the equation as:

$$H(\nu|\mu) = \sum_{\sigma\in\sum} \nu(\sigma) log(\nu(\sigma)/\mu(\sigma)) \tag{2.35}$$

The above equation is also known as relative entropy of ν with respect to μ. Now if we say that ϵ is the tolerable false alarm rate. Then the model free anomaly detector is given by:

$$I(F) = 1\{I_1(\boldsymbol{\mathcal{E}}^F) \geq \eta\} \tag{2.36}$$

Here η is given by $\eta = -\frac{1}{n} log\epsilon$. $I_1(\boldsymbol{\mathcal{E}}^F)$ is given by relative entropy of estimated probability vector ν with respect to μ. It has been shown in [5] that (2.36) is asymptotically Neyman-Pearson optimal.

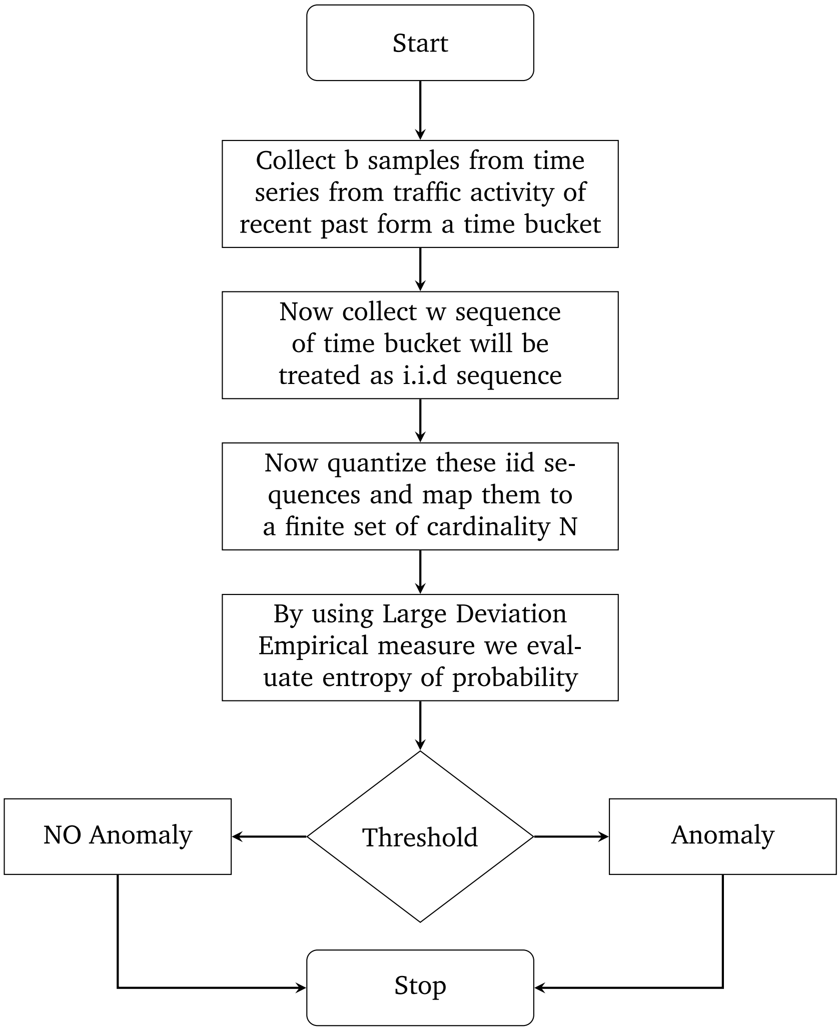

Start
Collect b samples from time series from traffic activity of recent past form a time bucket
Now collect w sequence of time bucket will be treated as i.i.d sequence
Now quantize these iid sequences and map them to a finite set of cardinality N
By using Large Deviation Empirical measure we evaluate entropy of probability
Threshold
NO Anomaly
Anomaly
Stop

Figure 2.1: Model free approach

2.4 Model based approach

In model based approach we will discuss about another statistical approach for anomaly detection i.e. Markov based approach. In Markov model we are using Markov chain rule to form a model of the given flow data.

Markov chain process is a stochastic process to be a sequence of random variables $\{X_0, X_1, X_2, ...\}$ or $\{X_t\}_{t=0}^{\infty}$ assuming values in finite alphabet $A = \{a_1, ..., a_n\}$. We will discuss here about finite state Markov process. A process $\{X_t\}_{t=0}^{\infty}$ is said to be a Markov chain process if for $t \geq 1$ and every sequence $a_1, ..., a_t \in N^t$, we can say that

$$Pr\{X_t = a_t | X_1 = a_1, ..., X_{t-1} = a_{t-1}\} = Pr\{X_t = a_t | X_{t-1} = a_{t-1}\} \tag{2.37}$$

So in Markov Chain model state transition occurs in one step. And again if state transition from time t to time $t + 1$ is independent of time then Markov Chain becomes stationary Markov chain. Here in this method model is formed assuming state transition as stationary one. A stationary state transition MC probability is given as:

$$Pr\{X_t = i_t | X_{t-1} = i_{t-1}\} = Pr\{X_t = j | X_{t-1} = i\} = p_{i,j} \tag{2.38}$$

Therefore $p_{i,j}$ represents probability of the system in state j at time t such that it was at state i at time $t - 1$. Therefore state transition matrix for this n number of state MC can be written as:

$$P = \begin{matrix} p_{1,1} & p_{1,2} & \cdots & p_{1,n} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n,1} & p_{n,2} & \cdots & p_{n,n} \end{matrix}$$

Whereas we have some initial probability distribution represented as:

$$Q = [q_1, q_2, ..., q_n]$$

q_i is defined as the probability that the system is in state i at time zero.

This is an alternate method for anomaly detection. In model free approach we assumed that the sequence of flows are i.i.d. i.e. all states are independent of each other. Whereas in this approach we have an assumption states are dependent on it predecessor state. As we have seen earlier in section 2.3 the states are same. But in quantized flow state $\sum(F) = \{\sigma(g^1), \sigma(g^2), ..., \sigma(g^F)\}$ each state $\sigma(g^i)$ is dependent on state $\sigma(g^{i-1})$. Hence in this model as per the above matrix equations flow state transitions can be formed under Markovian assumption as:

$$\boldsymbol{\mathcal{E}}_B^F(\sigma^i, \sigma^j) = \frac{1}{|F|} \sum_{l=2}^{|F|} 1\{\sigma(f^{l-1}) = \sigma^i, \sigma(f^l) = \sigma^j\} \tag{2.39}$$

In eq.(2.39) $1\{.\}$ represents indicator function. Where $\sigma(f^l)$ denotes a flow state in $\sum$ which f^l gets mapped to. Here we use F_{ref} to form a finite state transition matrix. This reference finite state transition matrix can be formed using eq. (2.39). The matrix can be written as:

$$\boldsymbol{\pi} = \{\pi(\sigma^i, \sigma^j)\}_{1,j=1,....,|\sum|}$$

As we can see that state transition matrix is a $\sum \times \sum$. And again empirical measurement is evaluated using same formula (2.39) with support $\sum \times \sum$. Lets say P is the estimated state transition matrix given by:

$$\mathbf{P} = \{p(\sigma^i, \sigma^j)\}_{1,j=1,....,|\sum|}$$

Each state transition matrix under Markovian assumption is associated with probability matrix of the form $\{p(\sigma^j|\sigma^i)\}_{1,j=1,....,|\sum|}$ where $p(\sigma^j|\sigma^i) = p(\sigma^i, \sigma^j)/p(\sigma^i)$. Where $p(\sigma^i) = \sum_{i=1}^{|\sum|} p(\sigma^i, \sigma^j)$ denotes the marginal probability of flow state in $\mathbf{P}$. Now following a similar procedure as in i.i.d.'s model free approach we apply Sanov's theorem under Markovian assumption is given in [6] and [8], we have:

$$H(\mathbf{P}|\boldsymbol{\pi}) = \sum_{i,j=1}^{|\sum|} p(\sigma^i, \sigma^j) log \frac{p(\sigma^j|\sigma^i)}{\pi(\sigma^j|\sigma^i)} \tag{2.40}$$

The above equation (2.40) is also known as relative entropy of $\mathbf{P}$ with respect to $\boldsymbol{\pi}$. Therefore in model based anomaly detection method the false alarm indication rate for F is given by:

$$I_B(F) = 1\{I_2(\boldsymbol{\mathcal{E}}_B^F) \geq \eta\} \tag{2.41}$$

As said earlier η depends on the tolerable false alarm rate. η is given as $\eta = \frac{1}{n} log\epsilon$. Where ϵ is the tolerable false alarm rate. $I_2(\boldsymbol{\mathcal{E}}_B^F)$ is the relative entropy rate of the estimated probability matrix $\mathbf{P}$ with respect to true probability matrix $\boldsymbol{\pi}$. Again as said earlier it has been proved in [5] that model based detector is asymptotically Neyman-Pearson optimal.

2.5 Incorporating spacial information

Till now we have studied only about temporal information. Yet, activity traces of interest can be collected in many locations and attacks at one place may be precursors or aftershocks of attacks elsewhere. So we introduce spacial information as well in the previous models. Consider a traffic activity as $X_1, ...X_n$ where $X_i \in \Re_d$. Where X_i represents the network feature of interest in slot i at all d locations we would like to monitor. Previous method can be extended from a scalar X_i to a vector X_i. For large d one would need to estimate for longer time to estimate its parameters and the anomaly detection algorithm would need longer samples to identify an anomaly. The only additional requirements in order to incorporate spatial information in the manner suggested is that the network elements must be synchronized and being able to exchange the time series of the network features they monitor.

2.5.1 spatio-temporal anomaly detection by model free approach

We will first decide that we want to monitor d number of network elements. Now $Y_t^{b^*,j} = Y_{(t-w+1)}^{b^*,j}, ..., Y_t^{b^*,j}$ represents w most recent partial sums of packet data of size b^* with $j = 1, ..., d$. Now for each network element we apply the temporal approach as discussed in section 2.3 to evaluate underlying alphabet and empirical measure.

- Now create the multi-dimensional alphabet $\sum^d = \{\alpha_1^{(1)}, ...\alpha_{N_1}^{(1)}\}......\{\alpha_1^{(d)}, ...\alpha_{N_1}^{(d)}\}$. Then compute the associated empirical measure (law) μ^d from past (anomaly-free) observations.

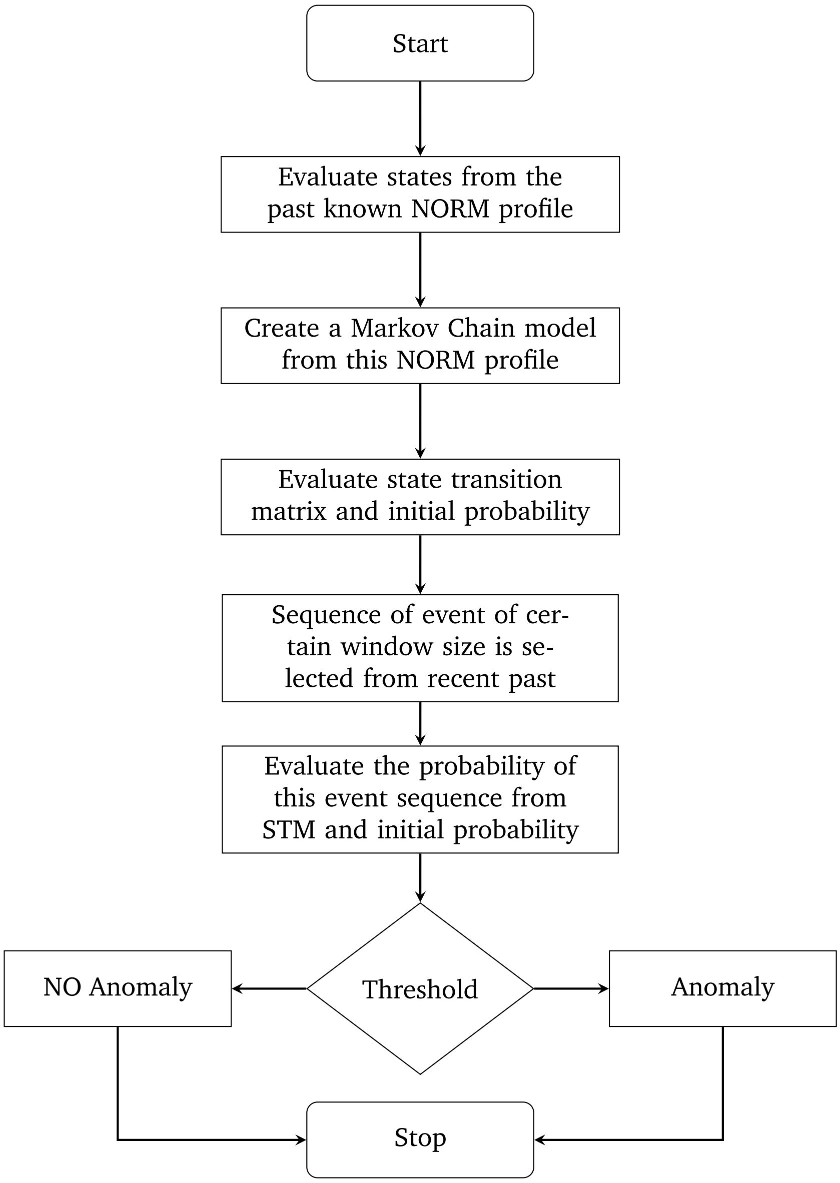

Start
Evaluate states from the past known NORM profile
Create a Markov Chain model from this NORM profile
Evaluate state transition matrix and initial probability
Sequence of event of certain window size is selected from recent past
Evaluate the probability of this event sequence from STM and initial probability
Threshold
NO Anomaly
Anomaly
Stop

Figure 2.2: Model based approach

- For every t we compute $Y_t^{b^*}$ of w most recent partial sums and compute its measure as $\varepsilon_{w,b^*}^{Y_t^{b^*}} = v_{t,w}^d$ where d represents the network element. And this represents the fraction of occurrence of $\sum^d$ letters in the trace $Y_t^{b^*}$.
- Same as scalar case we evaluate $\rho_{t,w}$ and approximates the probability that the trace $Y_t^{b^*}$ is drawn from the probability law μ^d. We can use the same test as (2.36) for anomaly.

The parameters w, b^* can be selected to improve the performance of the algorithm. After working with the statistics as fiven in [8] it is found that $b^* = 3$ and $w = 20$ are good values to work with.

2.5.2 spatio-temporal anomaly detection by model based approach

As before we assume d represents network elements or features we want to monitor. Suppose time series of network activity is represented as $X_n^j = \{X_1^j, ..., X_n^j\}$, where $j = 1, ..., d$. For each time series we split it into M_j subintervals or states for every j as discussed in section 4.2. Every M_j is selected using AIC criteria [?]. By using this M_j we form a multi-dimensional Markov chain. For example if $s_1^j, ...s_M^j$ represents the states of every network element then by putting it together we can form $s^1, ..., s^d$. Where $S^j \in \{s_1^j, s_2^j, ...\}$, i.e we are cascading every network elements state in a series. In this way we form a trace $X_n = \{X_n^1, ..., X_n^d\}$. Y_t denote the output of the multidimensional network at time t. Then:

- We will use past anomaly free data to form transition probability vector P^d.
- Now for each consecutive time slot t we observe the output vector $Y_{t,n} = (Y_{t-n+1}, ..., Y_n)$ and their empirical measures as $\varepsilon_{n,2}^{Y_{t,n}} = d_{t,n}^d$. Where d represents dimension of the Markov chain vector.
- Then $\rho_{t,n}$approximates the probability that the trace $Y_{t,n}$ is drawn from the Markov model with transition probability matrix p^d. and then compare using (2.40) for anomaly testing.

2.6 SVM base approach

This is also a model based approach. We can say it as decision boundary based approach. Because in this approach SVM forms a decision quadratic boundary between normal or regular activity and unusual or irregular activities. This technique is also popular by the name of 1st class SVM. Lets represent our dataset as $\mathcal{X} = \{x^1, ..., x^{\mathcal{X}}\}$. $\mathcal{X}$ can have values either $+1$ or -1. We have to find a hyperplane

$$f(x) = w^T \phi(x) - b \tag{2.42}$$

such that we can correctly classify a point x by its sign. Where w and b are constant vector and a constant. $\phi(x)$ is a function that maps the test point to a hyperplane to get more accurate classification. Suppose $\mathcal{X}$ data is not linearly separable then $\phi(\cdot)$ takes it to some higher dimension where it becomes linearly separable. Lets take an ideal case where every training data in $\mathcal{X}$ gets mapped by $\phi(x^{\mathcal{X}})$ to a linearly separable hyperplane where

$$w^T \phi(x) - b = 1 \quad \text{and} \quad w^T \phi(x) - b = -1 \tag{2.43}$$

separates data in $\mathcal{X}$ such that

$$y_i(w^T\phi(x^i) - b) \geq 1 \quad \forall i = 1, ..., |\mathcal{X}| \tag{2.44}$$

Where $y_i \in \{+1, -1\}$ for all $x^i \in \mathcal{X}$. So per given in (2.43) and (2.44) the hyperplane can be found by minimizing $\frac{1}{2}w^T w$. But in practice we hardly get a hyperplane linearly separable. So we try to find a hyperplane with lowest possible misclassification error. To complete this task we update our (2.44) as

$$y_i(w^T\phi(x^i) - b) \geq 1 - \xi_i \; \forall i = 1, ..., |\mathcal{X}| \tag{2.45}$$

Where $\xi_i > 0$. This allows a few misclassification error. Therefore we can find a hyperplane by [9]

$$\min \frac{1}{2}w^T w + c\sum_{i=1}^{|\mathcal{X}|}\xi_i \tag{2.46}$$

such that (2.45) satisfied and $\xi_i \geq 0$. Here c is a positive constant. Therefore it can be written as

$$L(w, b, \xi, \alpha, \beta) = \frac{1}{2}w^T w + c\sum_{i=1}^{|\mathcal{X}|}\xi_i - \sum_{i=1}^{|\xi|}\alpha_i[y_i(w^T\phi(x^i) - b) + \xi_i - 1] - \sum_{i=1}^{|\xi|}\beta_i\xi_i \tag{2.47}$$

Where α and β are Lagrange multiplier vectors and are non-negative. Therefore by taking derivatives we get

$$w = \sum_{i=1}^{|\mathcal{X}|}\alpha_i \mathbf{y}_i x^i \tag{2.48}$$

Therefore from (2.46) and (2.48) we can say from [9]

$$\sum_{i=1}^{|\mathcal{X}|}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{|\mathcal{X}|}\alpha_i\alpha_j y_i y_j \phi^T(x^i)\phi(x^j) \quad \forall i = 1, ..., |\mathcal{X}| \tag{2.49}$$

such that $\sum_{i=1}^{|\mathcal{X}|}\alpha_i y_i = 0$ and $c \geq \alpha_i \geq 0$. Therefore we can find our optimizer named α^*. Then we can find w^* by finding vector normal to the hyperplane. Since the inner product $\phi^T(x^i)\phi(x^j)$ is there we can use kernel function to replace that product. Using the kernel function [10] the above (2.42) can be written as

$$f(x) = \sum_{i=1}^{|\mathcal{X}|} y_i\alpha_i^* K(x, x^i) - b* \tag{2.50}$$

Where $k(u, v) = exp(-\gamma(u - v)^T(u - v)$ and $\alpha^* > 0$. Those elements of $\mathcal{X}$ for which $\alpha > 0$ are called support vectors. These points define the decision boundary. Lets say $\mathcal{X}_{sv} \in \mathcal{X}$ represents sets of all points which supports hyperplane. Therefore b^* is recovered by subset of those support vectors as

$$b^* = \frac{1}{|\hat{\mathcal{X}_{sv}}|} = \sum_{i=1}^{|\mathcal{X}|} 1_{\{x^i \in \hat{\mathcal{X}_{sv}}\}}(\sum_{j=1}^{\mathcal{X}}\alpha_j^* K(x^j, x^i) - y_i)$$

Now suppose training set $\mathcal{X}$ have only data of one type i.e. whose labels are all $+1$. Then the challenge comes how to separate outliers by using one-class SVM. So to do that we search a furtherest point from support vector points and consider it in outliers. Mathematically we are searching for a hyperplane for a set of non-negative margins $\{\xi_1, ..., \xi_{|\mathcal{X}|}\}$, represented by w and b i.e.

$$w^T \phi(x^i) \geq -\xi_i \quad \forall i = 1, ..., |\mathcal{X}|$$

We can find this hyperplane by QP [11]

$$\min \frac{1}{2} w^T w + \frac{1}{v|\mathcal{X}|} \sum_{i=1}^{|\mathcal{X}|} \xi_i - b \quad \forall i = 1, ..., |\mathcal{X}| \tag{2.51}$$

Where $w^T \phi(x^i) - b \geq -\xi$ and $\xi_i > 0$. Where v is a tunable parameter. It is used for separating anomaly. It is used for effectively tuning false alarm rate. As we know SVM uses binary classification. Therefore the binary classification can be written as

$$\min \frac{1}{2} \sum_{i,j=1}^{|\mathcal{X}|} \alpha_i \alpha_j K(x^i, x^j) \tag{2.52}$$

Such that $\sum_{i=1}^{|\mathcal{X}|} \alpha_i = 1$ and $0 \leq \alpha_i \leq \frac{1}{v|\mathcal{X}|}$. Here optimal α^* represents those support vectors in $\mathcal{X}$ which are used for anomaly detection.

2.6.1 First approach

For network anomaly detection flow based method is used for SVM. Consider a sequence of norm flows for training purpose $F = \{f^1, ..., f^{|F|}\}$. Where each flow is represented as $f^i = \{k^i, d(w^i), b^i, t^i, d_t^i\}$. We don't use time element in flow because in flow base SVM importance of time of transmission is very low. Time is important when we use time specific SVM where we see which server has how much at that specific time. But in this case time has low importance. So we ignore time for the time being. And we also omit cluster center from flow sequences for SVM operation because each cluster center represents a normal user. And distance from a normal user is more important than name of cluster center. Now the new data set can be written as $Z = z^1, ..., z^{|Z|}$. Where $z^i = d(w^i), b^i, d_t^i$. Therefore the redial basis vector compares test data against normal dataset. The anomaly detector can be represented as

$$sgn(f(z)) = sgn(\sum_{i=1}^{|Z|} \alpha_i^* K(z^i, z) - b^*) \tag{2.53}$$

Where $K(\cdot, \cdot)$ can be written as:

$$\min \frac{1}{2} \sum_{i,j=1}^{|Z|} \alpha_i \alpha_j K(z^i, z^j) \tag{2.54}$$

such that $\sum_{i=1}^{|Z|} \alpha_i = 1$ and $0 \leq \alpha_i \leq \frac{1}{v|Z|}$. Here v is the allowable false alarm rate. And b^* is given by

$$b^* = \frac{1}{|\hat{Z_{sv}}|} \sum_{i=1}^{|Z|} 1_{\{z^i \in \hat{Z_{sv}}\}} \sum_{j=1}^{|Z|} \alpha_j^* K(z^i, z^j) \tag{2.55}$$

Where $\hat{Z_{sv}} = \{z^i \in Z : 0 < \alpha_i^* < \frac{1}{v|Z|}\}$

2.6.2 Second approach

Here in second approach we use combination of model free approach and model based approach along with SVM to find anomalies from traffic. It is a window based anomaly detection approach. For each window i with flows F_i, we evaluate model free empirical measures as discussed in section 3. So for each F_i we get an empirical measure $\boldsymbol{\mathcal{E}}^{F_i}$. Same sequence of flows are used in Markov model based approach as discussed in section 4 and we evaluate empirical estimates $\boldsymbol{\mathcal{E}}_B^{F_i}$. Then we create a sequence of feature vector for window i as

$$Y^j = \{\boldsymbol{\mathcal{E}}^{G_i}, \boldsymbol{\mathcal{E}}_B^{G_i}, |G_i|\}$$

Then we use a series of windows say $\boldsymbol{\mathcal{Y}} = \{Y^1, ..., Y^{|\boldsymbol{\mathcal{Y}}|}\}$. Now we apply SVM approach on these windows to find anomalies. Since the dimensionality of the Y^j is to high so we use PCA first to reduce the dimensionality.

Chapter 3

Contribution

In statistical approach with Markov models we always assume that samples or events generated by Markov process. We use sample path from traffic for estimating State transition matrix(STM) and stationary distribution. The sample path for training models always have some length l. Because when we start using our model for a system we have a time constraint with us for training those models. Sample path of length l does not give true STM which generating those samples. So when we have finite sample length l we always get estimated parameters of the model. This true values and estimated values has some difference. Lets call these differences an error. This error can be positive or negative. Because suppose a_{ij} is true value of STM element and $\hat{a}_{ij}$ is the estimated value of estimated STM. Then the error can be given by $a_{ij} - \hat{a}_{ij}$. So we want to know how this length of sample path effect the modulus of the error. In 1963 W. Hoeffding had given a theorem or inequality called Hoeffding's Inequality [12] and [13] for i.i.d. sequence. But the theorem was given only for i.i.d. sequences. Here we have a markov process and the sample data in these sequences are not i.i.d. So we want to extend this inequality theorem to find a relationship between maximum error and length of sample path.

3.1 Extention of Hoeffding's Inequality for Markov Process

Suppose $\{X_t\}$ is a Markov process assuming values in $[1, N]$. Denote these by $\{1, ..., N\}$, where these are not integers but just labels. Suppose we look at a path of length l. Denote the sample path by $x_1, ..., x_l$. Where $x_i \in X_t$. To create a close path or cycle we create a ghost transition $x_{l+1} = x_1$. Based on this sample path, we form two estimates named as stationary distribution and STM. For each index $i \in [1, N]$, we count how many times the state i occurs in the sample path for stationary distribution and for each pair (i, j), we count how many times the pair ij occurs in the sample path for STM. In symbols,

$$\hat{\pi} = \frac{1}{l} \sum_{t=1}^{l} I_{x_t = i} \quad \hat{a}_{ij} = \frac{1}{l} \sum_{t=1}^{l} I_{(x_t x_{t+1}) = (ij)}$$

Here $\hat{a}_{ij}$ represents elements of estimated STM $\hat{A}$. where we take $x_{l+1} = x_1$, to make these estimates consistent; that is

$$\hat{\pi} = \hat{\pi}\hat{A}$$

We would like to estimate how quickly these estimates of STM $\hat{A}$ and stationary distribution $\hat{\pi}$ converge to their true values, namely true transition matrix A and the true stationary distribution π. For this purpose we would like to use Hoeffding's inequality [12]. Lets say $l_i = l\hat{\pi}_i$, that is the number of times that state i occurs in the sample path of length l. We know from the Hoeffding's inequality that

$$Pr\{l_i > l(\pi_i + \gamma)\} = Pr\{\hat{\pi}_i > \pi_i + \gamma\} \leq exp(-2l\gamma^2)$$

On the other side we can also write

$$Pr\{l_i < l(\pi_i - \gamma)\} = Pr\{\hat{\pi}_i < \pi_i - \gamma\} \leq exp(-2l\gamma^2)$$

Where $\gamma \geq 0$ and $\gamma \leq \pi_i$. So now with probability $1 - p$ we can say that

$$l_i \geq l(\pi_i - \gamma), \quad \forall i \in [N]$$

Where $q = Nexp(-2l\gamma^2)$. Now we can bound how close each estimate $\hat{\pi}_i$ is to the true value π_i. Now we know that each transition is independent of every other transition. So we can treat all of these transitions as independent. Therefore, again by using Hoeffding's inequality, we can say that

$$Pr\{|\hat{\pi}_i - \pi_i| > \gamma\} \leq 2Nexp(-2l\gamma^2) \tag{3.1}$$

Since we can say with confidence of at least $1 - q$, that $l_i \geq l(\pi_i - \gamma)$ for every index $i \in [N]$. Moreover, if $l_i \geq l(\pi_i - \gamma)$, then

$$exp(-2l_i\epsilon^2) \leq exp[-2l(\pi_i - \gamma)\epsilon^2] \quad \forall i \in [N] \tag{3.2}$$

Since we know for a Markov process that

$$\sum_{j=1}^{N} a_{ij} = 1 \quad \forall (i, j) \in [N]$$

The above equation is also valid for the estimated value $\hat{a}_{ij}$. Now as we know state i occurs l_i times in the sample path of length l. Therefore as we each transition is independent of every other transition we can write

$$Pr\{\hat{a}_{ij} > a_{ij} + \epsilon\} \leq exp(-2l_i\epsilon^2)$$

Where $\epsilon \geq 0$.Similarly

$$Pr\{\hat{a}_{ij} > a_{ij} - \epsilon\} \leq exp(-2l_i\epsilon^2)$$

Similar to the previous method by using Hoeffding's inequality, we can say that

$$Pr\{|\hat{a}_{ij} - a_{ij}| > \epsilon\} \leq 2Nexp(-2l_i\epsilon^2) \quad \forall (i,j) \in [N]$$

By using the previous eq. (3.2) we can say that

$$Pr\{|\hat{a}_{ij} - a_{ij}| > \epsilon\} \leq 2Nexp(-2l(\pi_i - \gamma)\epsilon^2) \quad \forall (i,j) \in [N] \tag{3.3}$$

Therefore by combining all these estimates leads to the following final conclusion. Suppose

$$\gamma < \min_i \pi_i$$

Then, for every $\epsilon \geq 0$ (accuracy parameter) and length l, we can say from eq. 3.1 and 3.3 with confidence of at least $1 - r$ that

$$Pr\{|\hat{a}_{ij} - a_{ij}| \leq \epsilon\} \leq 1 - 2Nexp(-2l(\pi_i - \gamma)\epsilon^2) - 2Nexp(-2l\gamma^2) \quad \forall (i,j) \in [N] \tag{3.4}$$

Where r can be written as

$$r = 2Nexp(-2l(\pi_i - \gamma)\epsilon^2) + 2Nexp(-2l\gamma^2)$$

Therefore the final inequality for Markov estimates and true values can be given as

$$Pr\{|\hat{a}_{ij} - a_{ij}| > \epsilon\} \leq 2Nexp(-2l(\pi_i - \gamma)\epsilon^2) + 2Nexp(-2l\gamma^2) \quad \forall (i,j) \in [N] \tag{3.5}$$

This inequality can be extended for a s-step Markov processes as well. As studied earlier in subsection 2.1.2 a s-step Markov process can be represented as one step Markov process. As we have seen that in s step Markov process the dimension of STM becomes $n^s \times n^s$. And stationary distribution also becomes n^s dimensional. To proof Hoeffding's inequality for s-step Markov process lets assume all the notations are same as in Markov chain case. So l represents sample path length. X_t is a Markov process assuming values in same label index $[N]$. Here also we create closed path by using ghost transition $x_{l+1} = x_l$. But here in this case lets say $u = X_{t-s+1}^t$ and $v = X_{t-s}^{t-1}$. And lets say Z_t is a new representation of Markov process X_t as $Z_t = X_{t-s+1}^t$. So in this case we can say Z_t is a Markov process as :

$$Pr\{Z_t = u | Z_{t-1} = v\} = Pr\{X_t = u | X_{t-s}^{t-1} = v\} \tag{3.6}$$

So each index $(u, v) \in N^s$. So we count how many times u occurs in sample path to estimate stationary distribution and how many times u_v pair occurs in sample path. Here $u\_v = X_{t-s}^t$. In symbols,

$$\hat{\pi_u} = \frac{1}{l}\sum_{t=1}^{l} I_{z_t=u} \widehat{b_{u\_v}} = \frac{1}{l}\sum_{t=1}^{l} I_{(z_t z_{t+1})=(u\_v)}$$

Here lets say π and B represents true matrices of Markov process. As we can see (3.1) it does not affected by matrices dimension. So it remains same as above. Because $|\hat{\pi} - \pi|$ represents infinity norm of that vector which is independent of length of the vector. So though length of vector changed because of s-step Markov process, max difference between its element remains unaffected. So we are using the same notions for s-step Markov process except for STM matrix B and its element b_{ij}. So from (3.2) and (3.3) it can be written as:

$$exp(-2l_u\epsilon^2) \leq exp[-2l(\pi_u - \gamma)\epsilon^2] \quad \forall u \in [N^s] \tag{3.7}$$

$$Pr\{|\hat{b}_{u\_v} - b_{u\_v}| > \epsilon\} \leq 2N^s exp(-2l(\pi_u - \gamma)\epsilon^2) \quad \forall (u,v) \in [N^s] \tag{3.8}$$

For $\epsilon \geq 0$ and $\gamma < \min_u \pi_u$. From (3.7) and (3.8) we can say with confidence $1 - q$ that

$$Pr\{|\hat{b}_{u\_v} - b_{u\_v}| \leq \epsilon\} \leq 1 - 2N^s exp(-2l(\pi_u - \gamma)\epsilon^2) - 2N^s exp(-2l\gamma^2) \quad \forall (u,v) \in [N^s] \tag{3.9}$$

Where q can be written as

$$r = 2N^s exp(-2l(\pi_u - \gamma)\epsilon^2) + 2N^s exp(-2l\gamma^2)$$

Therefore from above inequality (3.9), we can write final inequality for s-step Markov process by this equation.

$$Pr\{|\hat{a}_{t-s,t} - a_{t-s,t}| > \epsilon\} \leq 2N^s exp(-2l(\pi_i - \gamma)\epsilon^2) + 2N^s exp(-2l\gamma^2) \tag{3.10}$$

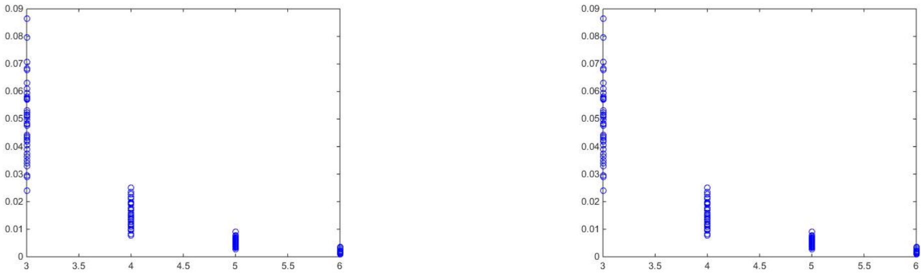
Where $\hat{a}_{t-s,t} = \hat{a}_{t-s}.....\hat{a}_t$ and similarly $\mathbf{a}_{t-s,t} = \mathbf{a}_{t-s}.....\mathbf{a}_t$. But here in the above equation stationary distribution π_i is different from the the previous one. Here our state transition matrix is of $N^s * N^s$. So we can say practically we have N^s states. So stationary distribution can be written as $\pi_i \in N^s$. Eq.(3.6) gives the final representation of Hoeffding's inequality of a s-step Markov process.

3.2 Simulation

We have checked this inequality with simulation as well. We first create a state transition matrix with $N = 4$ and $N = 5$. Then we generate sample path from that state transition matrix. And from the generated samples we would like to see change in error ϵ as length of the samples l varies We have plotted the errors ϵ for 1000, 10000, 100000 and 1000000. We run the simulation for 100 times for each l to see the differences. We have plotted $log(l)$ along x-axis and error ϵ along y-axis. We can see from Figure3.1 that error decreases as no. of samples increases. Figure3.1 gives us an idea of errors Markov processes estimates with 4 and 5 states i.e. $N = 4$ and $N = 5$.

It seems obvious that error reduces as sample path length for training increases. But here in above simulation we have chosen some probability where stationary distribution have some sort of equal probability. That means occurrence of a state i in the process is approximately uniform. But that does not happens in network process. Sometimes some states have much higher probability whereas some states have very less probability. With this assumption a stationary distribution is created as an experiment with 4 states. 3.2a and 3.2b shows the true STM and stationary distribution which generates samples.

As we can see in stationary distribution that 2nd state has very low probability of occurrence. So this probability can be referred as a very rare event. It don't need to be an anomaly. But as we take



(a) For N=4

(b) For N=5

Figure 3.1: error ϵ vs. log(sample path length) $log(l)$

$$\begin{bmatrix} 0.7778 & 0.0087 & 0.1918 & 0.0218 \\ 0.0010 & 0.0010 & 0.9900 & 0.0080 \\ 0.9900 & 0.0010 & 0.0080 & 0.0010 \\ 0.0010 & 0.0010 & 0.0080 & 0.9900 \end{bmatrix}$$

(a) STM

$$\begin{bmatrix} 0.2903 & 0.0032 & 0.0645 & 0.6419 \end{bmatrix}$$

(b) Stationary distribution

Figure 3.2: Example

samples generated by this Markov process shown in fig. 3.2. For $10,000$ training samples we get an estimated matrix as shown in fig. 3.3. As we can see in fig. 3.3 that estimates are not good. Some of the transition probabilities are even zero. There is also lot of variation in stationary distribution probabilities as well. So we estimated these matrices with sample path of lengths 100000, 1000000 and 10000000. As we see in fig. 3.4 number of samples increased to 10000000, We have got a good estimate. Those small transition probabilities like 0.0010 are also estimated approximately. Finally, fig. 3.5 shows plot of the same matrix shown in fig. 3.2. Simulation has been done for this irregular matrix with varying probabilities to check a graph of error vs. sample path length. Here in fig. 3.5 error is plotted w.r.t. $\log(l)$.

As its been given in [2] and [1] that MIT's cyber data consist of 284 event types. That means they have 284 states for Markov process. Then they have 3019 audit events in total from which they have used around 1613 audit events for training of Markov model. As seen earlier that through simulation data were generated from a known Markov process with known parameters depends on the length of training samples. It has been shown that 3019 samples are very small for a 284 state Markov process. In that case Markov process will give many false alarms because it could not separate anomaly with rare events. For an effective detection we need a lot more training data.

$$\begin{bmatrix} 0.7731 & 0.0101 & 0.2001 & 0.0167 \\ 0.0000 & 0.0000 & 0.9773 & 0.0227 \\ 0.9897 & 0.0000 & 0.0091 & 0.0011 \\ 0.0010 & 0.0010 & 0.0107 & 0.9874 \end{bmatrix}$$

(a) STM

$$\begin{bmatrix} 0.3843 & 0.0044 & 0.0876 & 0.5237 \end{bmatrix}$$

(b) Stationary distribution

Figure 3.3: Example:10000 samples estimates

$$\begin{bmatrix} 0.7780 & 0.0087 & 0.1916 & 0.0217 \\ 0.0009 & 0.0012 & 0.9897 & 0.0082 \\ 0.9898 & 0.0010 & 0.0081 & 0.0010 \\ 0.0010 & 0.0010 & 0.0081 & 0.9899 \end{bmatrix}$$

(a) STM

$$\begin{bmatrix} 0.2927 & 0.0033 & 0.0650 & 0.6391 \end{bmatrix}$$

(b) Stationary distribution

Figure 3.4: Example:10000000 samples estimates

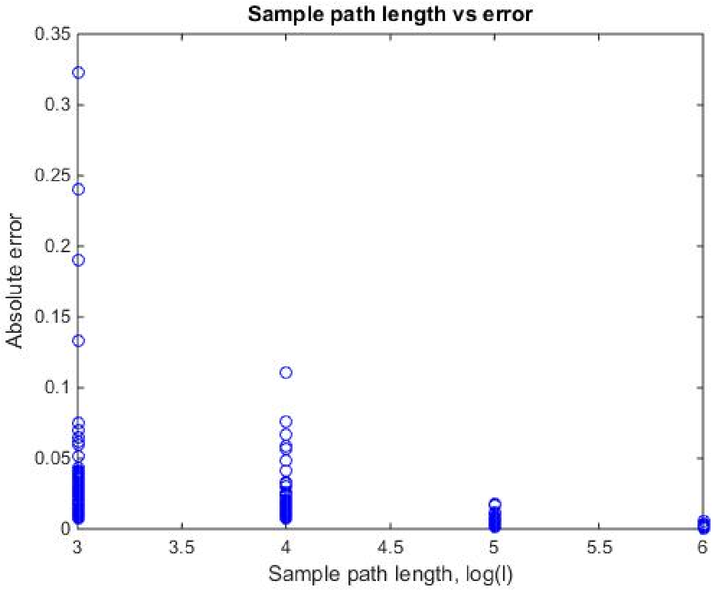


Figure 3.5: Error vs. log of sample length ($\log(l)$) for $N = 4$

Chapter 4

Future work

Machine learning is playing a larger role in cyber security, which can in theory help identifying risks and anticipate problems before they occur. The idea is to create a software which can adapt as per changing attack strategies. Traditional security mechanism have leverage rules, patterns, signatures and algorithms based approaches to detect threats. This is a problem, because these approaches needs constant care and feeding to identify threats. But machine learning can change the game. The benefit of machine learning is it works fast, detection rate is also high and can detect attack vectors that are previously undetectable.

Up till now cyber security systems worked based on creating rules for detecting attacks. These rules were created based on events happened in past. So it is something like ”what we know, not what we don't know”. So what we have till now best describes some machine learning i.e. algorithms learn from data. But in future ”user behavior analytics” will come in picture. Where most of the security breaches will be self discovered. This ”user behavior analytics” can also be known by some other names. In this approach we predict attacks before happening by analyzing changes in user behavior and predicting risks and breaches before they happen. Well anomaly detection methods does the same thing. Because in pattern recognition we use signatures of known attacks to detect attacks or threat to our system. So in that case we could not find the novel attacks or those attacks which are previously unseen. But in anomaly detection approach we keeps user's norm profile. When a system detects something different from normal then it hikes alarm. Till now a few methods of attack detection have been shown for anomaly detection. These methods are giving some false alarms. Which needs to be treated. Statistical methods have low resolution because they can't exactly detect anomalous flows in the stream of data. But they provide stable results and they have high detection rate. Whereas deterministic methods may have higher false alarm rates and gives unstable results. They have high resolution. These observations suggest that as we need all characteristics we have to find a way for handshaking those methods to work together. Because an individual method can't yield better result.

Bibliography

[1] N. Ye et al. A markov chain model of temporal behavior for anomaly detection 166, (2000) 169.

[2] N. Ye, Y. Zhang, and C. M. Borror. Robustness of the Markov-chain model for cyber-attack detection. *Reliability, IEEE Transactions on* 53, (2004) 116–123.

[3] R. P. Lippmann, D. J. Fried, I. Graf, J. W. Haines, K. R. Kendall, D. McClung, D. Weber, S. E. Webster, D. Wyschogrod, R. K. Cunningham, and M. A. Zissman. Evaluating intrusion detection systems: the 1998 DARPA off-line intrusion detection evaluation 2, (2000) 12–26 vol.2.

[4] R. Lippmann, J. W. Haines, D. J. Fried, J. Korba, and K. Das. The 1999 DARPA off-line intrusion detection evaluation. *Computer networks* 34, (2000) 579–595.

[5] I. C. Paschalidis and G. Smaragdakis. Spatio-temporal network anomaly detection by assessing deviations of empirical measures. *IEEE/ACM Transactions on Networking (TON)* 17, (2009) 685–697.

[6] J. Wang, D. Rossell, C. G. Cassandras, and I. C. Paschalidis. Network anomaly detection: A survey and comparative analysis of stochastic and deterministic methods 182–187.

[7] J. Zhang and I. C. Paschalidis. An Improved Composite Hypothesis Test for Markov Models with Applications in Network Anomaly Detection. *CoRR* abs/1509.01706.

[8] M. Vidyasagar. Hidden Markov Processes, Theory and application to biology. Princeton University Press, 41 William Street,Princeton, New Jersey, 2014.

[9] V. Vapnik. Statistical learning theory. 1998 .

[10] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation* 13, (2001) 1443–1471.

[11] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation* 13, (2001) 1443–1471.

[12] W. Hoeffding. Probability Inequalities for Sum of Bounded Random Variables. *Journal of American Statistical* .

[13] P. W. Glynn and D. Ormoneit. Hoeffding's Inequality for Uniformly Ergodic Markov Chain. *Statistics and Probability Lettersl* 56.