Traffic Modeling and Analysis

Work done by: Ramsamy, Saifullah, Dinil Mon Divakaran
(M.S Research Scholars)

Guided by

Prof. Timothy A. Gonsalves & Dr. Hema A. Murthy

Outline

- Problem Definition
- Traffic Classification
  - Vector Quantization (VQ)
  - Gaussian Mixture Models (GMM)
  - Test results
- Denial of Service (DoS) attack
- Linear prediction (LP) analysis
- TCP SYN flooding attack detection
- Conclusion

Introduction

- Internet growth has resulted in huge amount of data.
- Data can be used for bandwidth management, traffic prediction, network planning, Quality of Service, anomaly detection etc.
- Modeling and analysis of traffic data give useful information.

Problem definition

- Traffic modeling and classification.

- Linear prediction (LP) analysis for denial of service (DoS) attack detection.

⇒ *Traffic Modeling and Classification*

- DoS attack and LP analysis
- TCP SYN flooding attack detection

Why not port based classification ?

- Firewall related problems - relaying of non-web traffic using port 80.
- Ports are not defined with IANA registration for all applications.
- Non-privileged users run WWW servers on ports other than 80.
- Some well-known ports are used by multiple applications.
- Dynamic allocation of server ports (eg. FTP).

Parameters for modeling

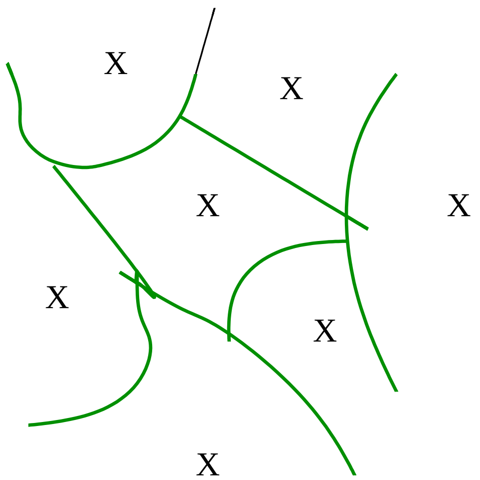
- Traffic characteristics depends on the protocol or service.

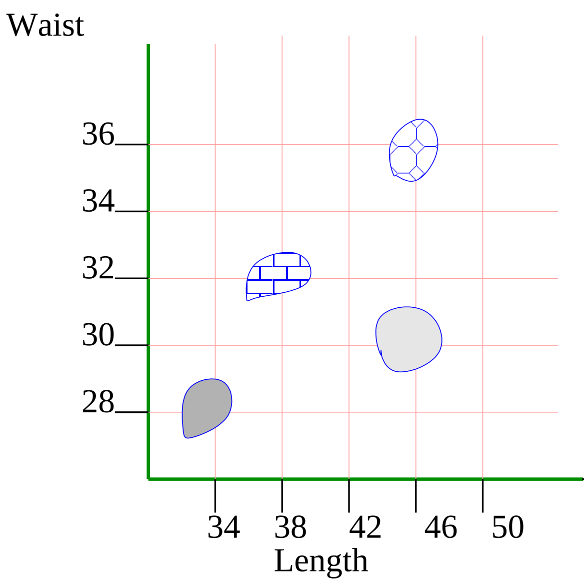
- Commonly used parameters : *packet size, packet inter-arrival time, flow duration, packet train size, packet train length*.

- Packet train: *(src host, src port, dst host, dst port)*.
  Packet train length : Number of packets in a train.
  Packet train size : Number of bytes in a train.

- Input traffic data represented as vectors: $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}\}$
  $\mathbf{x_i} = \{$*packet train length, packet train size*$\}$

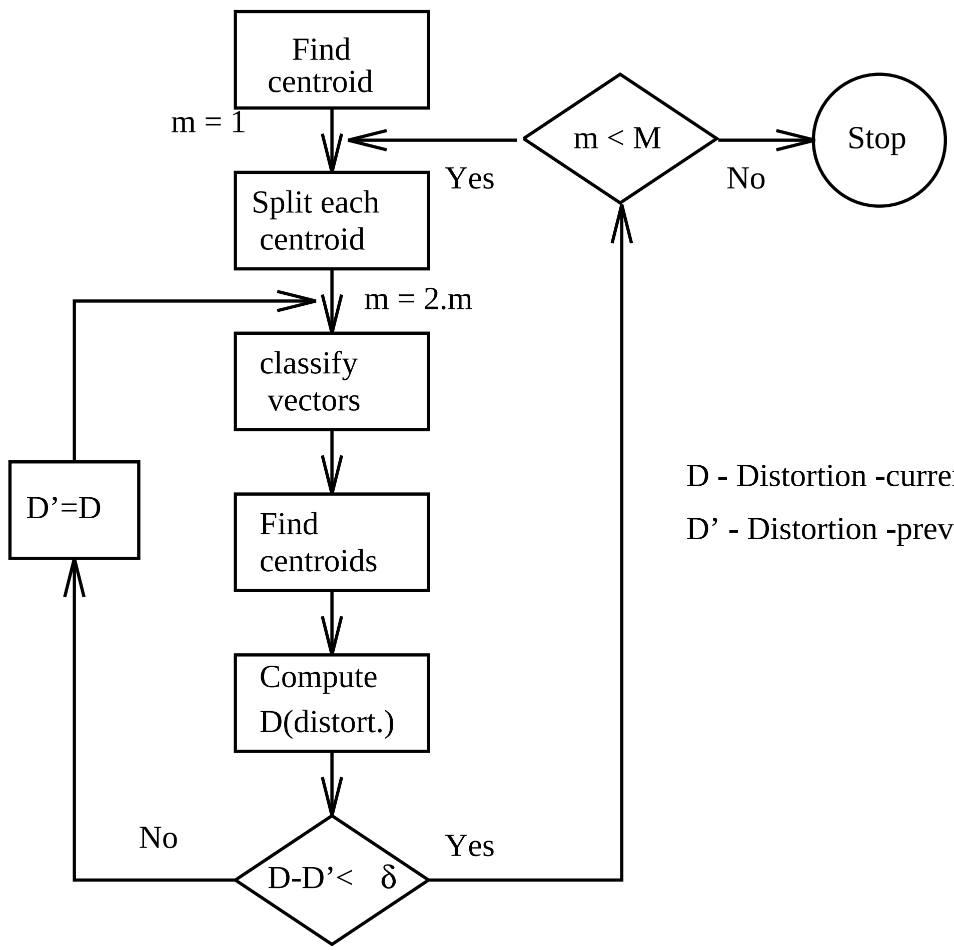
Clusters in Pattern Space (Vector Quantisation)





PARTITIONED VECTOR SPACE   X = CENTROID OF REGION

An Algorithm for Vector Quantisation



D - Distortion -current iter.

D' - Distortion -previous iter.

The average distortion D_i in cell C_i is given by

$$D_i = \frac{1}{N} \sum_{\mathbf{x} \in C_i} d(\mathbf{x}, \mathbf{z}_i)$$

where

- $\mathbf{z}_i$ is the centroid of cell C_i and
- $d(\mathbf{x}, \mathbf{z}_i) = (\mathbf{x} - \mathbf{z}_i)^T (\mathbf{x} - \mathbf{z}_i)$
- N is the number of vectors

The centroids that are obtained finally are then stored in a codebook called the **VQ codebook**.

Modeling Using Vector Quantization

- Traffic data considered as vectors

$$\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}\}$$

where $\mathbf{x_i} = \{$*packet train length, packet train size*$\}$

- $\mathbf{X}$ divided into a set of k clusters, $C = \{C_1, C_2, ..., C_k\}$, such that

$$\bigcup_{i=1}^{k} C_i = \mathbf{X} \qquad and \qquad \bigcap_{i=1}^{k} C_i = \phi$$

Algorithm for obtaining Clusters for Traffic Modeling

- Randomly select k vectors as the centroids of the k clusters.

- A vector $\mathbf{x}$ belongs to cluster C_i, if

$$\left\|\mathbf{x} - \mu_{\mathbf{i}}\right\| < \left\|\mathbf{x} - \mu_{\mathbf{j}}\right\| \qquad \text{for all } j \neq i$$

  $\Rightarrow$ partitions vectors into clusters.

- Recompute centroid after each iteration.

$\Rightarrow$ Obtain clusters for each traffic class.

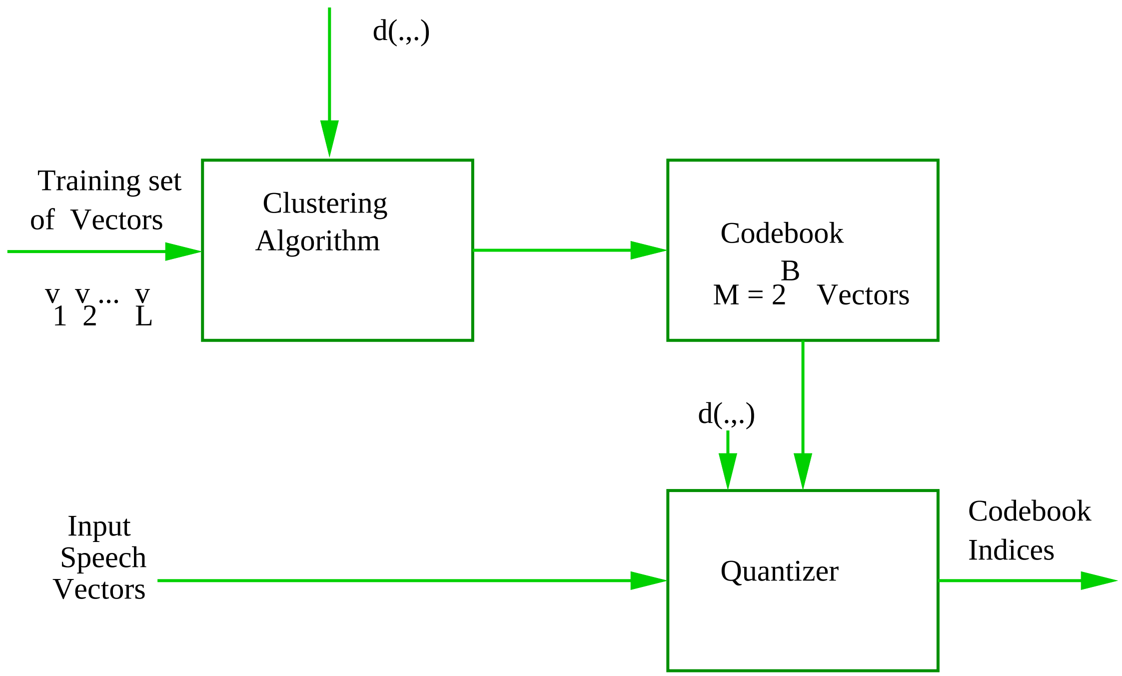


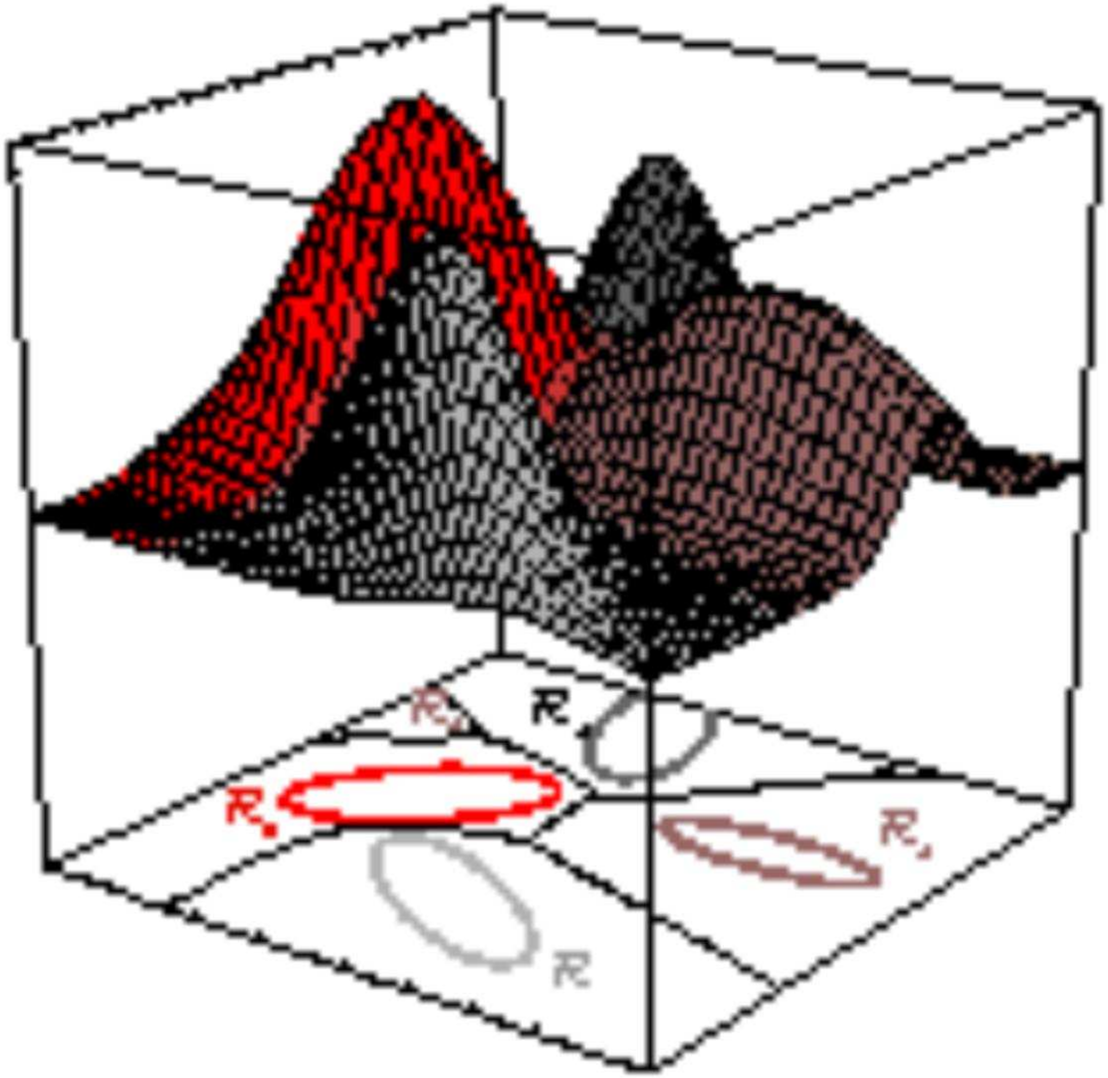
Figure 1: System based on VQ

Classification Using VQ

- For each vector, find the distance from every traffic class.
  $\Rightarrow$ Find the distance to the nearest cluster of a traffic class.

- Find the distance of the input set from each traffic class.

- The traffic class of the given input set is identified as the one to which the distance is minimum.



Gaussian Mixture Models

Bayesian Classification using GMM

- $\mathbf{X}$ divided into k mixtures, $\{m_1, m_2, ..., m_k\}$

$$p(m_i) = \frac{n_{m_i}}{N}$$

$\theta_{\mathbf{i}}$ is the vector with components $\mu_{\mathbf{i}}$ and $\sigma_{\mathbf{i}}$ of the mixture m_i.

- Probability of $\mathbf{x}$ belonging to a mixture m_i [1]

$$p(\theta_{\mathbf{i}}|\mathbf{x}) \approx p(m_i)\frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma_i}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu_{\mathbf{i}})^t \mathbf{\Sigma_i}^{-1}(\mathbf{x}-\mu_{\mathbf{i}})}$$

- A vector $\mathbf{x}$ belongs to mixture m_i, if

$$p(\theta_{\mathbf{i}}|\mathbf{x}) > p(\theta_{\mathbf{j}}|\mathbf{x}) \qquad \text{for all } j \neq i$$

- $\theta_{\mathbf{i}}$ recomputed at the end of each iteration.

Bayesian Classification using GMM (continued)

- Given a set of input vectors $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}\}$ and s being a traffic class, probability that a vector $\mathbf{x_i}$ is of traffic class s

$$p(\mathbf{x_i}, s) = \max_j p(\theta_\mathbf{j} | \mathbf{x_i}) \qquad 1 \leq j \leq n(s)$$

  $n(s)$ is the number of mixtures in the traffic class denoted by s.

- Probability that the given set is of a particular traffic class s

$$P(s) = \prod_{i=1}^{N} p(\mathbf{x_i}, s)$$

  where N is the total number of vectors.

- Input set of vectors belongs to the class having maximum probability.

Test Results

Traffic Type	Accuracy
HTTP	99.27%
SMTP	96.38%
DNS	100%
POP3	90.56%
SSH	88.40%

(a) VQ based

Traffic Type	Accuracy
HTTP	99.60%
SMTP	99.30%
DNS	100%
POP3	97.20%
SSH	96.92%

(b) GMM based

Table 1: Results using one hour data

Test Results (continued)

Traffic Type	Accuracy
HTTP	99.78%
SMTP	99.67%
DNS	100%
POP3	96.28%
SSH	94.53%

Table 2: Results using GMM for 15 minutes data

- Successful in classifying traffic.
- Can not be used for detection of a class of attacks - Denial of Service attacks.

- Traffic Modeling and Classification

$\Rightarrow$ ***DoS attack and LP analysis***

- TCP SYN flooding attack detection

- Legitimate users denied service.

- UDP flooding, ICMP flooding, Smurf attack, TCP Reset attack, TCP SYN flooding etc.

- Distributed DoS.

- Low intensity and high intensity attacks.

- Loss incurred
  - From e-commerce companies like Amazon to small ISPs.
  - Top source of financial loss due to cybercrime in 2004 [2].

- Detection and Prevention

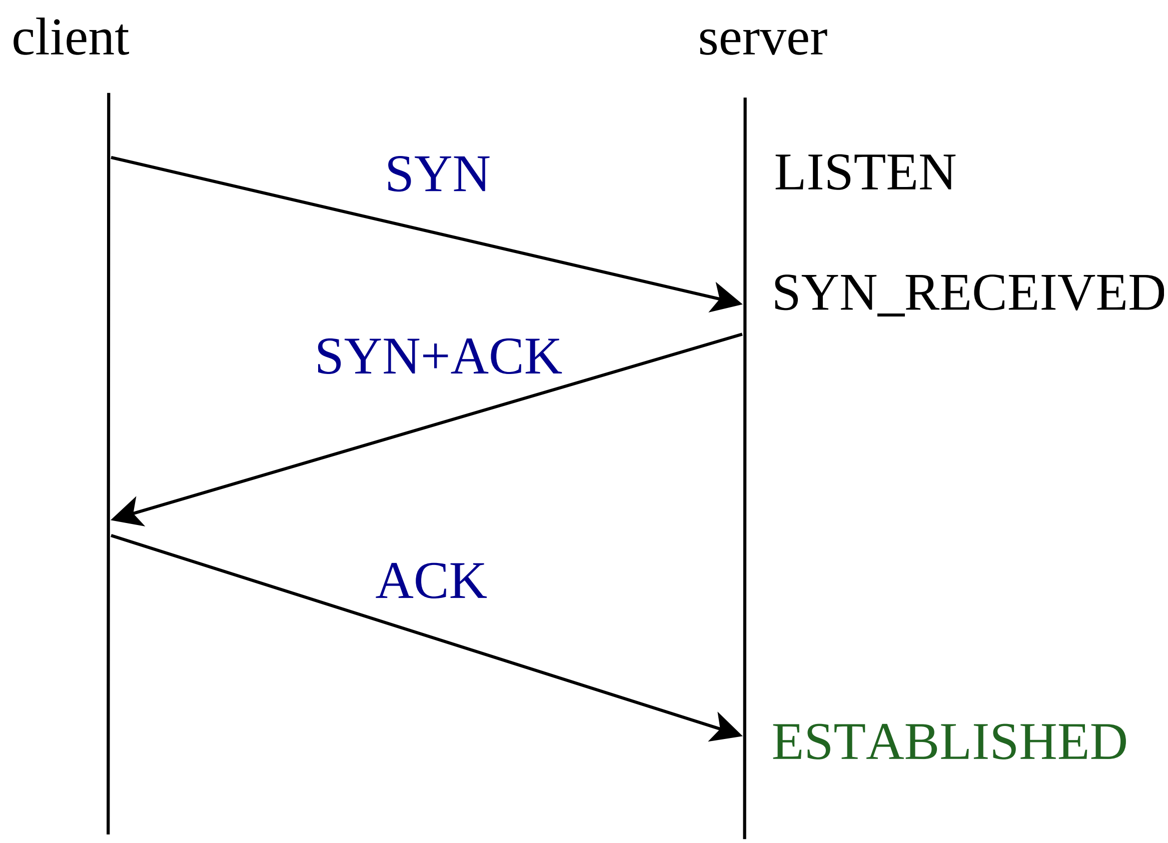


Figure 2: TCP Connection establishment

Properties of TCP SYN flooding attack

- One of the most commonly used attacks [3].
- Easy to launch.
- Achieved by sending less than 20 packets per second.
- Defense mechanisms (*SYN cookies*, *RandomDrop*, *SYN cache*, *SYNkill*, *SYNDefender etc.*) have limitations
- Since most of the applications use TCP, detection becomes all the more important.

Linear Prediction (LP) analysis

- LP approximately estimates a signal, s_n, as linearly weighted summation of past samples [4]

$$\tilde{s}_n \approx \sum_{k=1}^{p} a_k s_{n-k}$$

- Error percentage

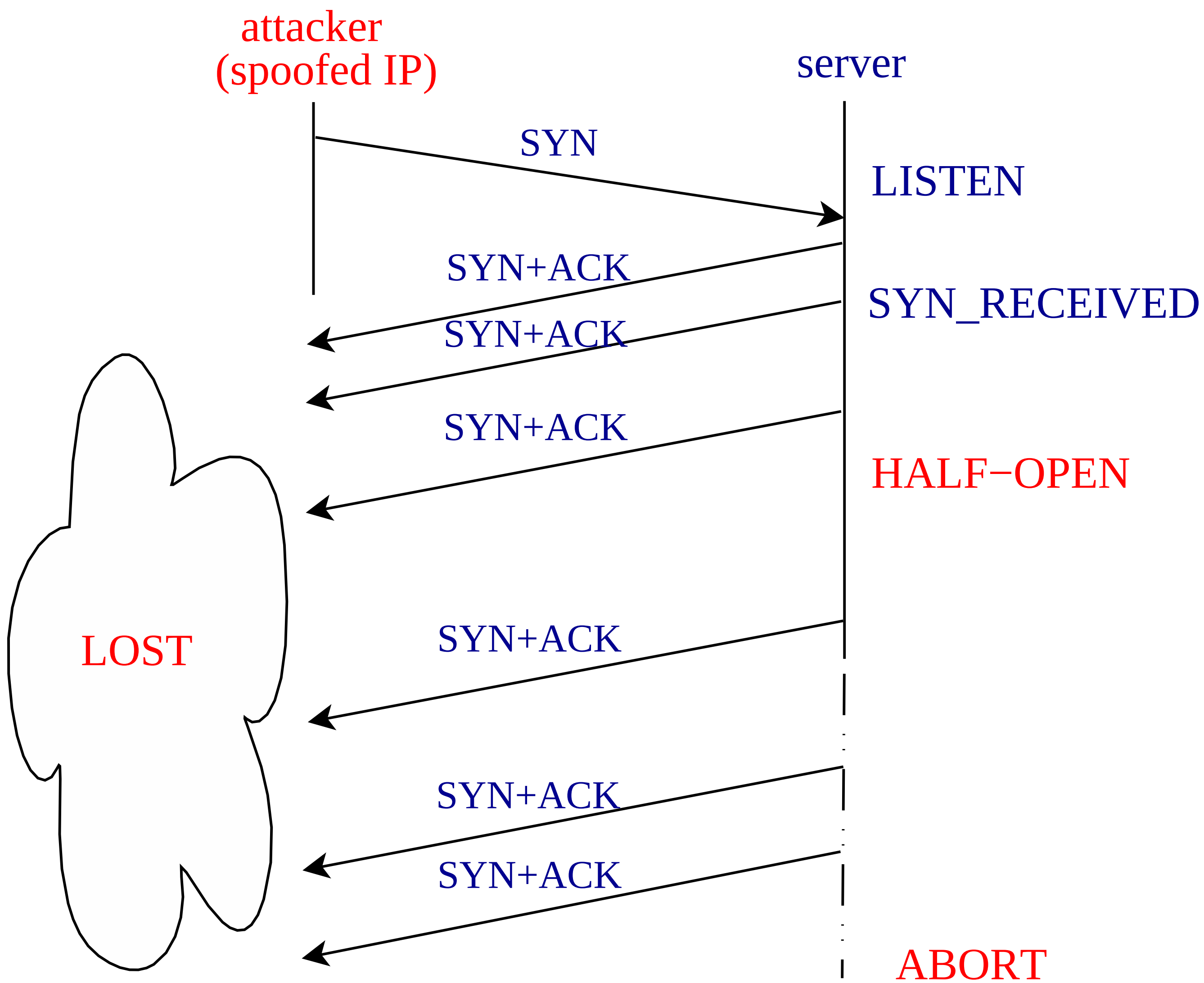
$$e_n = \frac{s_n - \tilde{s}_n}{s_n} * 100$$

- What is s_n ?

- Traffic Modeling and Classification

- DoS attack and LP analysis

$\Rightarrow$ ***TCP SYN flooding attack detection***



A SYN attack scenario

Algorithm for DoS attack detection

$\Rightarrow$ Detect deviation from normal traffic.

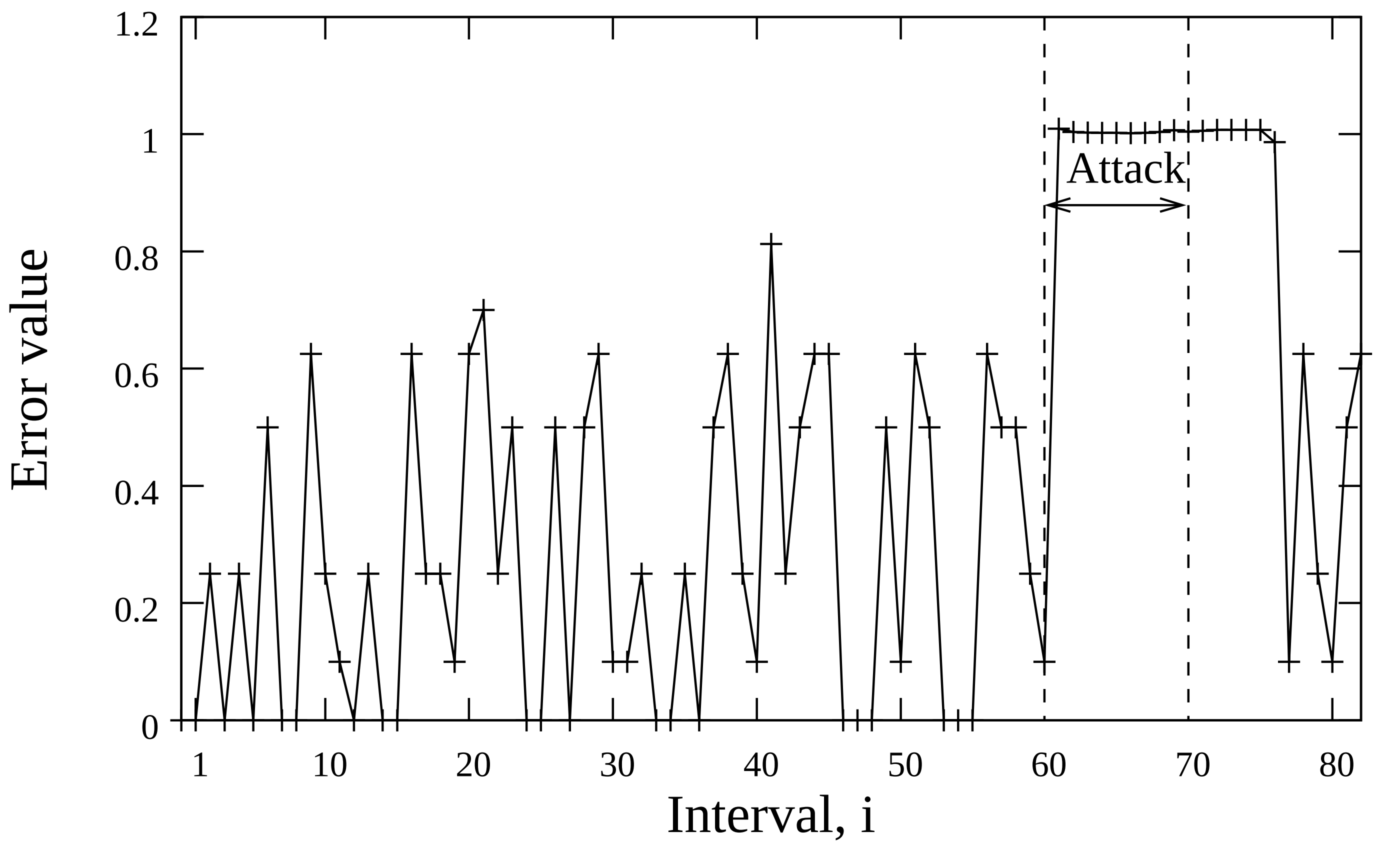
- Initial frame $\{s_1, s_2, s_3, s_4, s_5, s_6\}$
  s_i corresponds to difference in number of incoming SYN and outgoing SYN+ACK in i^{th} time slot.

- Compute the predictor coefficients $\{a_1, a_2, a_3, a_4\}$.

- Predict next signal value. Obtain actual signal value.

- Compute e, if $e > \alpha(threshold)$, then probability of attack is high.

- Recompute LP coefficients periodically using normal frames.



Plot of error for signal values

Power spectrum analysis

- Frame is advanced by one signal value.

$$F_1 = \{s_1, s_2, s_3, s_4, s_5, s_6\}$$

$$F_2 = \{s_2, s_3, s_4, s_5, s_6, s_7\}$$

$$F_3 = \{s_3, s_4, s_5, s_6, s_7, s_8\}$$

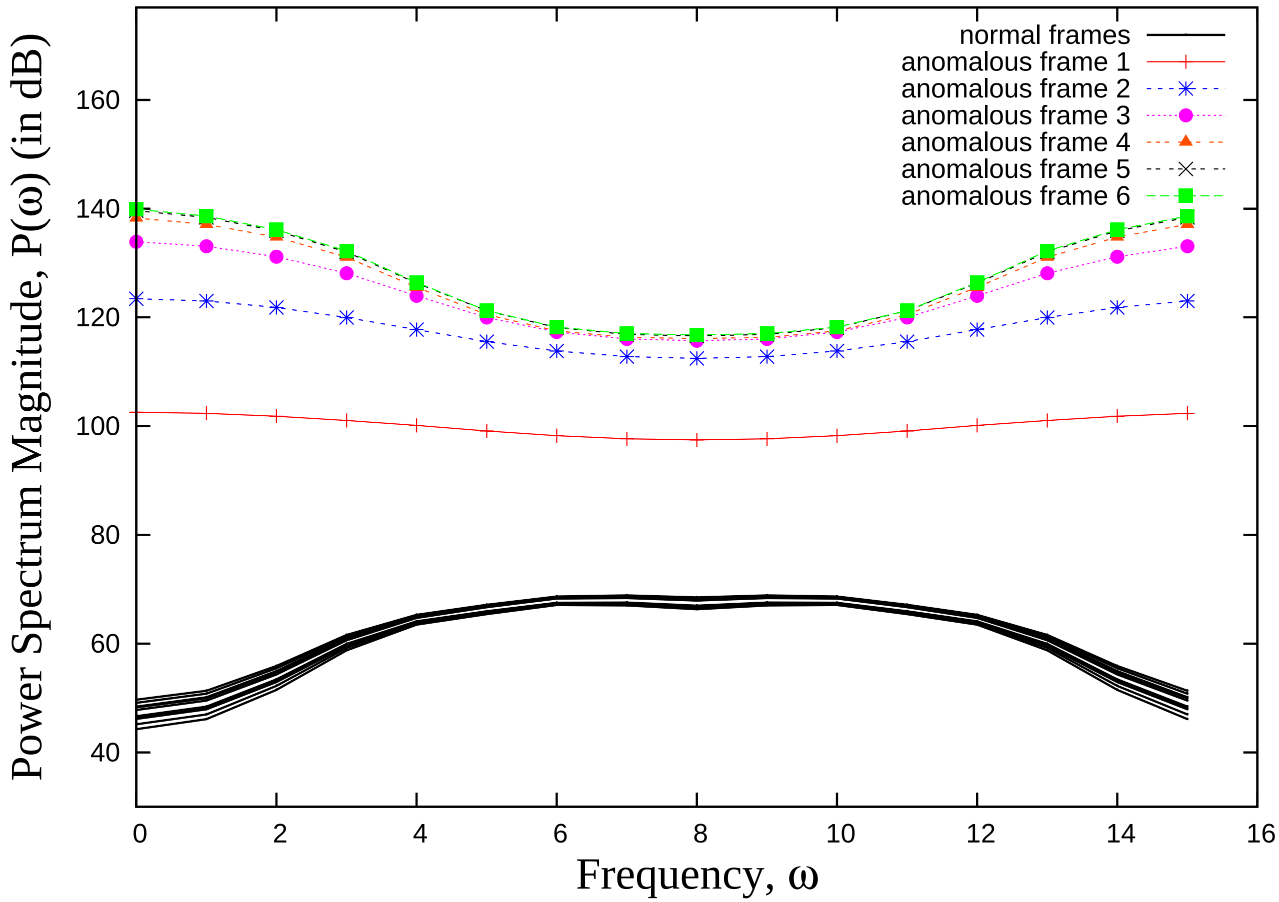
$$\vdots$$

- Transfer function,

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}}$$

where G is the gain factor.

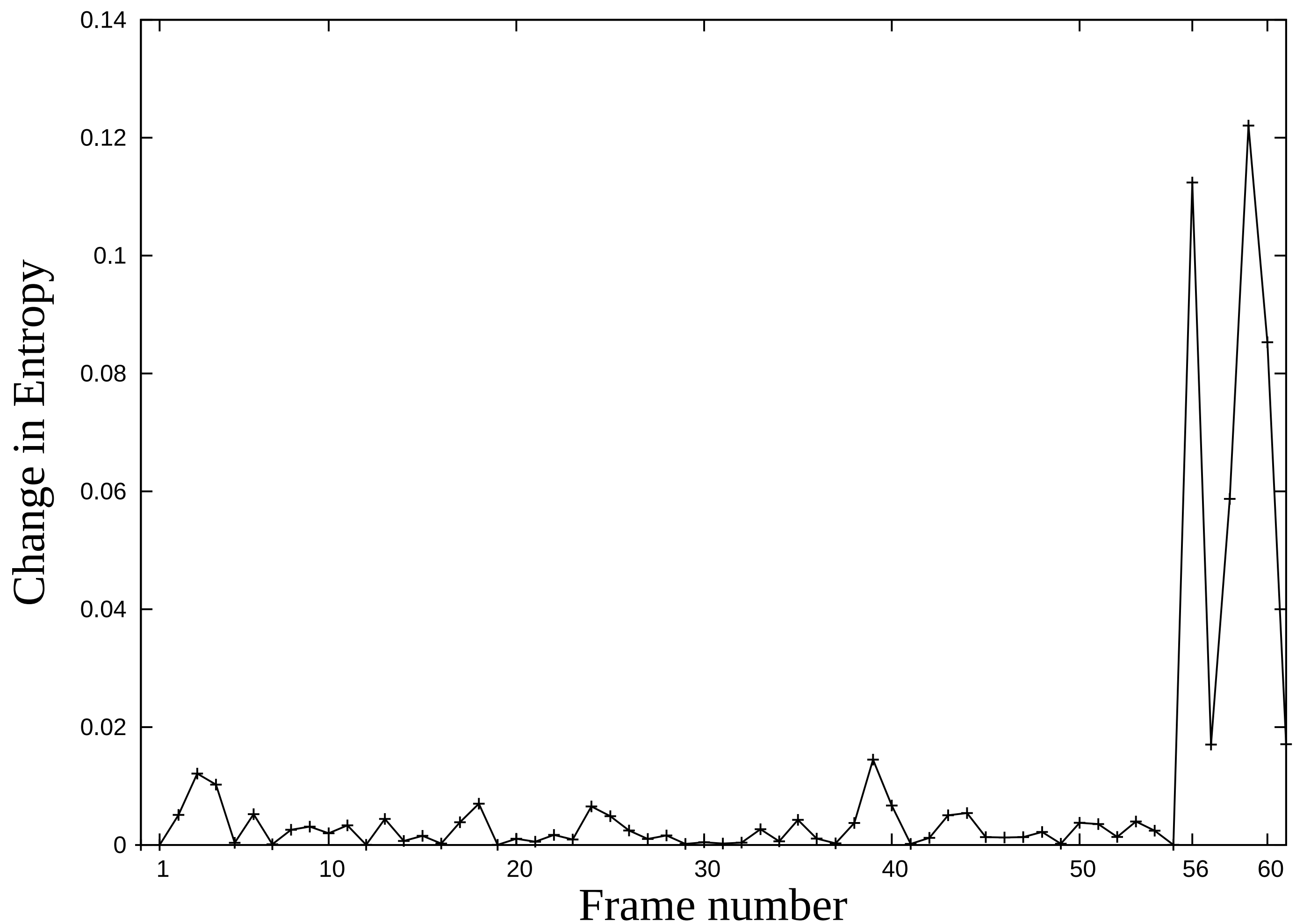
- Compute spectrum for each frame.



Spectrum for half-open connections to TeNeT network

Detection using Entropy estimation

- Entropy is estimated for each spectrum.
- For normal frames, deviation of entropy from frame to frame is of the order of 10^{-3}.
- Entropy of a frame with one anomalous signal value deviates in the order of 10^{-1}.



Entropy difference between adjacent frames

Performance evaluation

SYNs per second	**Best delay**	**Worst delay**
20	8	17
30	7	16
40	6	15
50-90	5	14
≥ 100	4	13

Table 3: Detection delay in seconds for different attack rates

Detection and Defense

- Detection system can be installed on firewalls and leaf router.
- Firewalls $\Rightarrow$ low-intensity SYN flooding attack detection.
- Attack detection $\Rightarrow$ trigger defense mechanism like SYNDefender on firewall.
- Routers $\Rightarrow$ high-intensity SYN flooding attack detection.
- Attacks to paralyze defense mechanism will be detected at the router level.

Conclusion

⇒ Bayesian classification using GMMs gives highly accurate classification.

⇒ *packet train length* together with *packet train size* form better parameters for identifying traffic compared to existing ones in literature .

⇒ Linear prediction analysis of traffic is a new approach to detect DoS attacks.

⇒ The difference in number of SYN and SYN/ACK packets is a new parameter to detect TCP SYN flooding attack.

References

[1] Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*, Wiley-Interscience Publication, second edition, 2001.

[2] L. Gordon et al., *CSI/FBI Computer Crime and Security Survey*, Computer Security Inst., 2004.

[3] David Moore, Geoffrey M. Voelker, and Stefan Savage, “Inferring Internet Denial-of-Service Activity,” in *Proceedings of the 10th USENIX Security Symposium*, 2001, pp. 9–22.

[4] J. Makhoul, “Linear Prediction: A Tutorial Review,” *Proceedings of the IEEE*, vol. 63(4), pp. 561–580, 1975.