Statistical intrusion detection in modern networks

Michele Pagano

Department of Information Engineering
University of Pisa

XIII International Asian School-seminar
Problems of complex systems' optimization
September 18-22, 2017
Akademgorodok

Outline

1. Theoretical background
2. How to evaluate an IDS
3. Anomaly Detection: Overview
4. Anomaly Detection: Examples

Outline

1. Theoretical background
   - Overview
   - Taxonomy of the Intrusion Detection Systems
   - Examples of the Intrusion Detection Systems
2. How to evaluate an IDS
3. Anomaly Detection: Overview
4. Anomaly Detection: Examples

Why an intrusion detection system?

- Network security mainly means PREVENTION
  - Physical protection for hardware
  - Passwords, access tokens, etc. for *authentication*
  - *Access control list for* authorization
  - Cryptography for *secrecy*
  - *Backups and redundancy for* authenticity
  - . . . and so on

BUT . . .

. . . Absolute security cannot be guaranteed!

What is an Intrusion Detection System?

- Prevention is suitable when
  - Internal users are trusted
  - Limited interaction with other networks
- Need for a system which acts when prevention fails

Intrusion Detection System

An intrusion detection system or IDS is a software/hardware tool used to detect unauthorized access to a computer system or a network

A bit of History

The history of IDSs can be split in three main blocks

1. First Generation IDSs (end of the 1970s)
   - The concept of IDS first appears in the 1970s and early 1980s (Anderson, Computer Security Monitoring and Surveillance, Tech Rep 1980)
   - Focus on audit data of a single machine
   - Post processing of data
2. Second Generation IDSs (1987)
   - Intrusion Detection Expert System (Denning, An intrusion Detection Model, IEEE Trans. on Soft. Eng., 1987)
   - Statistical analysis of data
3. Third Generation IDSs (to come)
   - Focus on the network
   - Real-time detection
   - Real-time reaction
   - Intrusion Prevention System

A taxonomy of the intruders

Intruders can be classified as

- **Masquerader**: an individual who is not authorized to use the computer and who penetrates a system's access control to exploit a legitimate user's account
- **Misfeasor**: a legitimate user who accesses data, programs, or resources for which such access is not authorized, or who is authorized for such access, but misuses his/her privileges
- **Clandestine User**: an individual who seizes supervisory control of the system and uses the control to evade auditing and access controls or to suppress audit collection

Anderson, Computer Security Monitoring and Surveillance, Tech Rep 1980

A taxonomy of the intrusions

- **Eavesdropping and Packet Sniffing**: passive interception of network traffic
- **Snooping and Downloading**
- **Tampering and Data Diddling**: unauthorized changes to data or records
- **Spoofing**: impersonating other users
- **Jamming or Flooding**: overwhelming a system's resources
- **Injecting Malicious Code**
- **Probing**
- **Exploiting Design or Implementation Flaws**: as buffer overflow
- **Cracking Passwords and Keys**

Denning, Cyberspace Attacks and Countermeasures, New York,1997

Host based vs. Network based

Host based IDS

- Aimed at detecting attacks related to a specific host
- Architecture/Operating system dependent
- Processing of high level information (e.g. system calls)
- Effective in detecting insider misuse

Network based IDS

- Aimed at detecting attacks towards hosts connected to a LAN
- Architecture/Operating system independent
- Processing data at lower level of granularity (packets)
- Effective in detecting attacks from the “outside”

Misuse based IDS vs. Anomaly based IDS

Misuse based (Signature based, Rule based) IDS

- Identifies intrusion by looking for patterns of traffic or of application data presumed to be malicious
- Pattern of misuses are stored in a database
- Effective in detecting only “known” attacks
- Encrypted traffic ???

Anomaly based IDS

- Identifies intrusions by classifying activity as either anomalous or normal
- Need a training phase to recognize normal activity
- Able to detect “new” attacks
- Generates more false alarms than a misuse based IDS

Stateless IDS VS Stateful IDS

Stateless IDS

- Treats each event independently of others
- Simple system design
- High processing speed

Stateful IDS

- Maintains information about past events
- The effect of a certain event depends on its position in the events stream
- More complex system design
- More effective in detecting distributed attacks

Centralized IDS VS Distributed IDS

Centralized IDS

- All the operations are performed by the same machine
- More simple to realize
- Only one point of failure

Distributed IDS

- Composed of several components
  - **Sensors** which generate security events
  - **Console** to monitor events and alerts and control the sensors
  - Central **Engine** that records events and generate alarms
- May need to deal with different data formats
- Need of a secure communication protocol (IPFIX)

SNORT

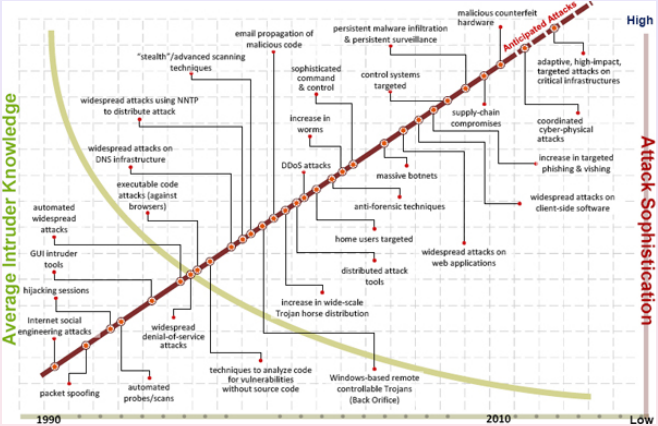
Snort

- The most famous IDS
- Open source software tool
- Network based
- Signature based
- Centralized architecture

Spade, the anomaly detection plug-in for Snort... is not supported any longer

The rules database, as well as the system code, is available for download at the web site http://www.snort.org

Attacks State of the Art



Howard Lipson, CMU Software Engineering Institute CERT®
(Computer Emergency Response Team – by DARPA in 1988)

IDS State of the Art

- Focus is on Network based IDSs (The only ones effective in detecting Distributed Denial of Service - DDoS)
- State of the art IDSs are Misuse Based
  - Most attacks are realized by means of software tools available on the Internet
  - Most attacks are “well-known” attacks

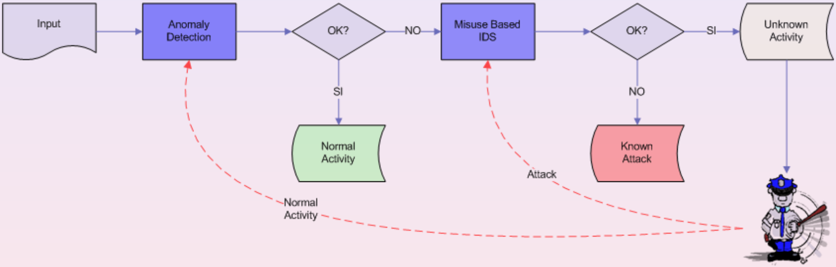
BUT . . .

. . . The most dangerous attacks are those written ad hoc by the intruder!

The best choice?

- Combined use of both
  - HIDS (for insider attacks) & NIDS (for outsider attacks)
  - Misuse IDS (low False Alarm rate) & Anomaly IDS (for "new" attacks)
  - Stateless IDS (fast data process) & Stateful IDS (for "complex" attacks)
- Distributed IDS
  - Not a single point of failure
  - More effective in monitoring large networks

The best choice?



Outline

1. Theoretical background
2. How to evaluate an IDS
   - Receiver Operating Characteristics (ROC) Curve
   - Evaluation Datasets
3. Anomaly Detection: Overview
4. Anomaly Detection: Examples

Basic definitions

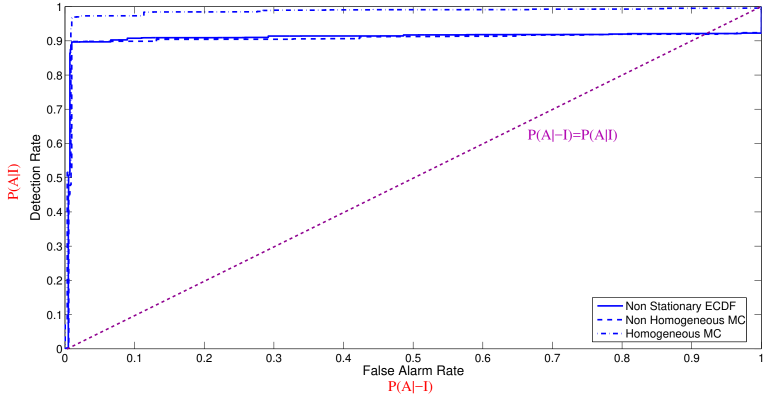
- A = alarm
- $-A$ = not an alarm
- I = attack (intrusion)
- $-I$ = not an attack
- False Positive (FP): the error of rejecting a null hypothesis when it is actually true. In our case it implies the creation of an alarm in correspondence of normal activities

$$P(A| - I) \; = \; \text{False positive (alarm) probability}$$

- False Negative (FN): the error of failing to reject a null hypothesis when it is in fact not true. In our case it corresponds to a missed detection

$$P(-A|I) \; = \; \text{False negative probability}$$

ROC (Receiver Operating Characteristics) Curve



Base Rate Fallacy

Let's suppose:

- $P(A|I) = 0.99$
- $P(-A|-I) = 0.99$
- we have 2 attacks over 10^6 pkts (base rate = 1/500000)

Applying the Bayes theorem:

$$P(I|A) = \frac{P(I) \cdot P(A|I)}{P(A|I) \cdot P(I) + P(A|-I) \cdot P(-I)} =$$

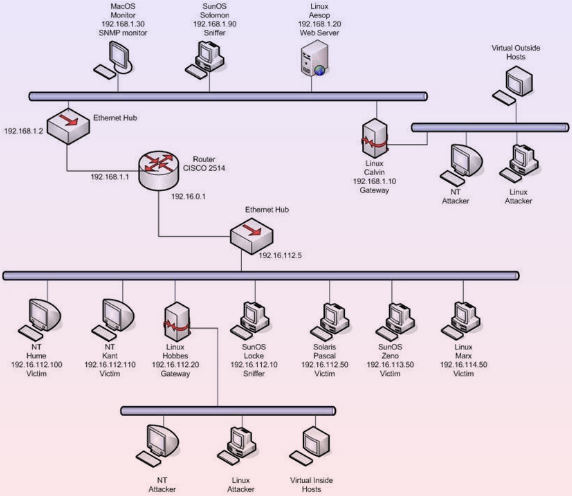
$$= \frac{1/500000 \cdot 0.99}{1/500000 \cdot 0.99 + (1 - 1/500000) \cdot 0.01}$$

Thus $P(I|A) = 0.0002$

DARPA Evaluation Program

- The 1998/1999 DARPA/MIT IDS evaluation program is the most comprehensive evaluation performed to date
- It provides a corpus of data for the development, improvement, and evaluation of IDSs
- Different kind of data are available:
  - Operating systems logs
  - Network traffic
    - Collected by an "inside" sniffer
    - Collected by an "outside" sniffer
- The data model the network traffic measured between a US Air Force base and the Internet

The DARPA Network



The DARPA Dataset

- 5 weeks data
  - Data from weeks 1 and 3 are attack free and can be used to train the system
  - Data from week 2 contains labeled attacks and can be used to realize the signatures database
  - Data from weeks 4 and 5 contains several attacks and can be used for the detection phase
- An Attack Truth list is provided
- Attacks are categorized as
  - Denial of Service (DoS)
  - User to Root (U2R)
  - Remote to Local (R2L)
  - Data
  - Probe
- 177 instances of 59 different types of attacks

KDD99

- Database of features extracted from the 1998 DARPA Dataset
- Used for *The third International Knowledge Discovery and Data Mining Tools Competition*
- Useful for evaluating IDSs
- Some example features
  - @IP source and dest
  - Flags
  - Number of bytes sent/received by a host
  - Duration of flows
  - Protocol
  - . . . and so on
- Can be downloaded at http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

Other Datasets

The DARPA dataset has many drawbacks:

- simulated environment
- not up-to-date traffic
- the methodology used for generating the traffic has been shown to be inappropriate for simulating actual networks

Other Datasets:

- several publicly available traffic traces
- e.g. CAIDA, Abilene (Internet2), GEANT, . . .
- no ground truth is provided!

MAWI-Lab Traffic Traces

- MAWI (Measurement and Analysis on the WIDE Internet) archive (sample-points B and F)
- Anomaly labelling is obtained combining the output of four anomaly detectors
    - Hough transform
    - Gamma distribution modelling
    - Kullback-Leibler divergence
    - Principal Component Analysis

MAWILab : Combining Diverse Anomaly Detectors for Automated Anomaly Labeling and Performance Benchmarking

R. Fontugne, P. Borgnat, P. Abry, and K. Fukuda, ACM CoNEXT 2010

MAWI-Lab Traffic Traces

Traffic taxonomy

- *anomalous*: anomalous with high probability
- *suspicious*: probably anomalous, but not clearly identified by the MAWI classification methods
- *notice*: non anomalous, but reported by at least one anomaly detector
- *benign*: normal

Some information about the kind of anomaly

- *attack*: anomalies representing a well known attack
- *special*: anomalies involving well known ports
- *unknown*: unknown kinds of anomalies

Outline

1. Theoretical background
2. How to evaluate an IDS
3. Anomaly Detection: Overview
   - Statistical models
   - Traffic Descriptors
4. Anomaly Detection: Examples

Profiles

- An activity profile characterizes the behavior of a given subject (or set of subjects) with respect to a given object, thereby serving as a signature or description of normal activity for its respective subject and object.
- Observed behavior is characterized in terms of a statistical metric and model
- A metric is a random variable x representing a quantitative measure accumulated over a period
- Observations x_i of x obtained from the audit records are used together with a statistical model to determine whether a new observation is abnormal
- The statistical models make no assumptions about the underlying distribution of x; all knowledge about x is obtained from the observations x_i

Metrics and Models

Metrics

- *Event counter*
- *Interval timer*
- *Resource measure*

Statistical models

- *Operational model*: abnormality is decided by comparison of x_n with a fixed threshold
- *Mean and standard deviation model*: abnormality is decided by checking if x_n falls inside the confidence interval
- *Multivariate model*: based on the correlations between two or more metrics
- *Markov process model*: based on the transition probabilities
- *Time series model*: takes into account order and inter-arrival time of the observations

Statistical Approach: Traffic Descriptors

To identify some traffic parameters, which can be used to describe the network traffic and that vary significantly from the normal behavior to the anomalous one

Some examples

- Packet length
- Inter-arrival time
- Flow size
- Number of packets per flow
- . . . and so on

Choice of the Traffic Descriptors

For each parameter we can consider

- Mean Value
- Variance and higher order moments
- Distribution function
- Quantiles
- . . . and so on

The number of potential traffic descriptors is huge (some papers identify up to 200 descriptors)

GOAL

To identify as few descriptors as possible to classify traffic with an *acceptable* error rate

Outline

1. Theoretical background
2. How to evaluate an IDS
3. Anomaly Detection: Overview
4. Anomaly Detection: Examples
   - Use of Markov Chains
   - Principal Component Analysis

State Transition Analysis

- The approach was first proposed by Denning and developed in the 1990s.
- Mainly used in two distinct environment
  - **HIDS**: to model the sequence of system commands used by a user
  - **NIDS**: to model the sequence of some specific fields of the packet (e.g. the sequence of the flags values in a TCP connection)
- The most classical approach: **Markov chains**

Markov Chains and TCP

- Idea: Model TCP connections by means of Markov chains
- The IP addresses and the TCP port numbers are used to identify a connection
- State space is defined by the possible values of the TCP flags
- The value of the flags is used to identify the chain transitions
- A value S_p is associated to each packet according to the rule

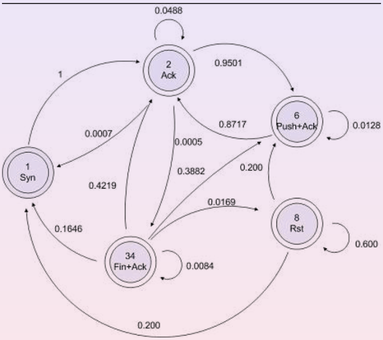
$$S_p = syn + 2 \cdot ack + 4 \cdot psh + 8 \cdot rst + 16 \cdot urg + 32 \cdot fin$$

Markov Chain and TCP - Training phase

Calculate the transition probabilities

$$a_{ij} = P[q_{t+1} = j | q_t = i] = \frac{P[q_t = i, q_{t+1} = j]}{P[q_t = i]}$$

- Server side
- 3-way handshake
- psh flag
- closing



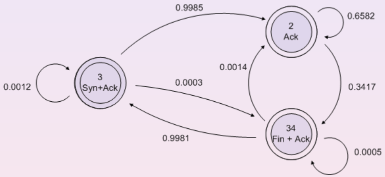
SSH Markov Chain

Markov Chain and TCP - Training phase

Calculate the transition probabilities

$$a_{ij} = P[q_{t+1} = j | q_t = i] = \frac{P[q_t = i, q_{t+1} = j]}{P[q_t = i]}$$

- Client side
- 3-way handshake
- ack flag
- closing



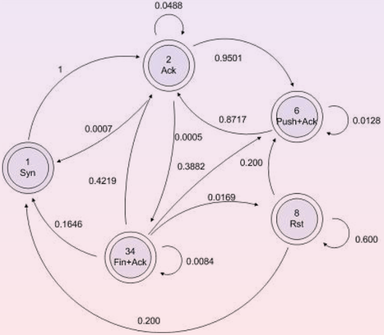
FTP Markov Chain

Markov Chain and TCP - Training phase

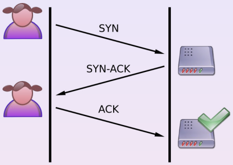
Calculate the transition probabilities

$$a_{ij} = P[q_{t+1} = j | q_t = i] =$$

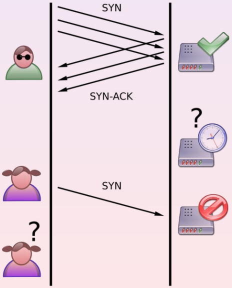
$$\frac{P[q_t = i, q_{t+1} = j]}{P[q_t = i]}$$



SSH Markov Chain



3-Way Handshake



Syn Flood Attack

Markov Chain and TCP - Detection phase

- Given the observation $(S_{R+1}, S_{R+2}, \cdots, S_{R+T})$
- The system has to decide between two hypothesis

$$\begin{aligned} H_0 &: \; \textit{normal behaviour} \\ H_1 &: \; \textit{anomaly} \end{aligned}$$

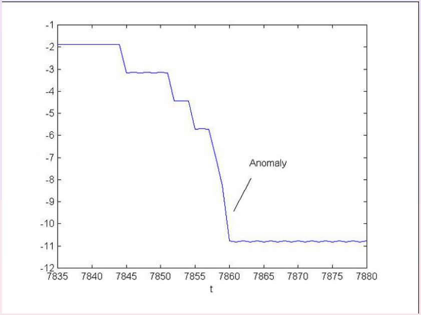
- A possible statistic is given by the logarithm of the Likelihood Function

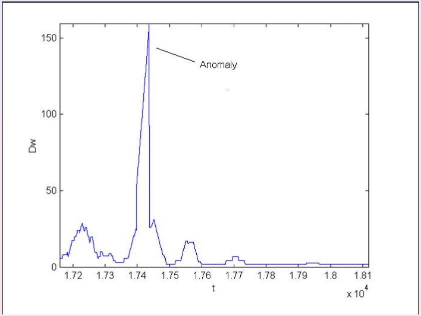
$$LogLF(t) = \sum_{t=R+1}^{T+R} Log(a_{S_t S_{t+1}})$$

- or by its temporal “derivative”

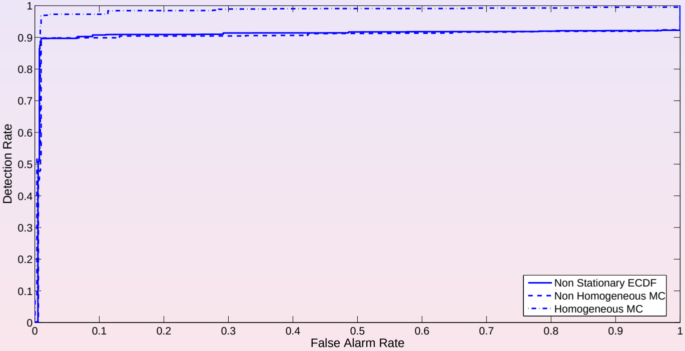
$$D_w(t) = \left| LogLF(t) - \frac{1}{W} \sum_{i=1}^{W} LogLF(t-i) \right|$$

Markov Chain and TCP - Detection phase

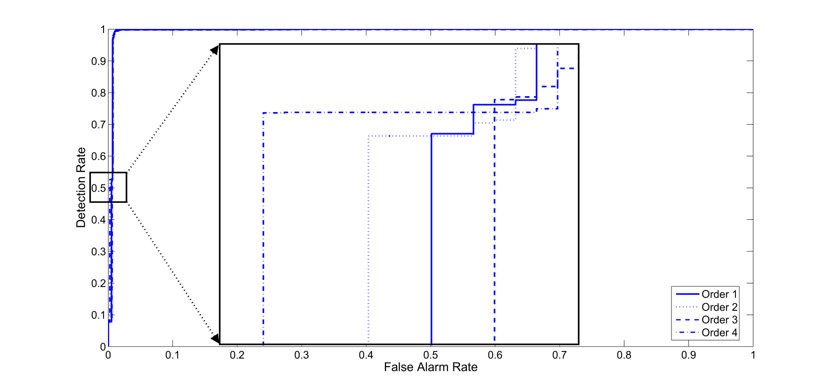




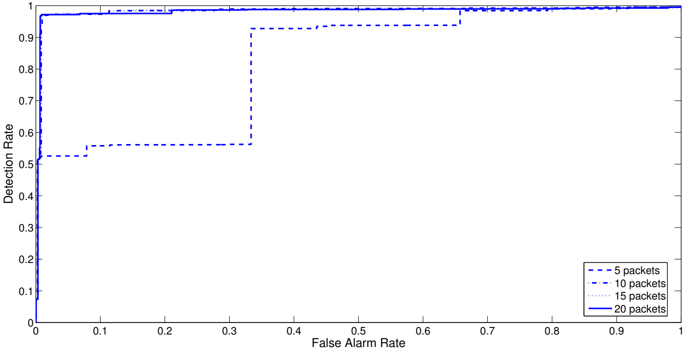
Non-Homogeneous Markov Chain



High Order Markov Chain



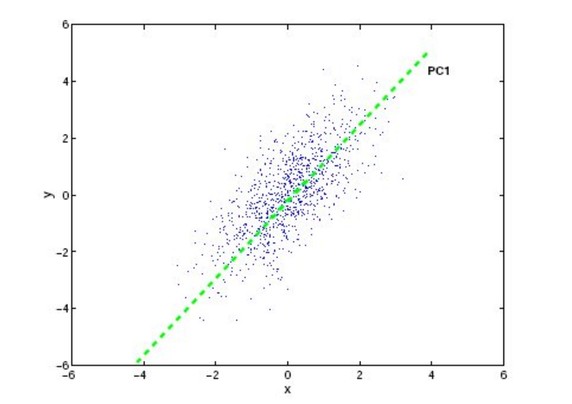
Homogeneous Markov chain — Number of packets



Principal Component Analysis

- Principal Component Analysis (or discrete Karhunen–Loève transform) is the most commonly used techniques to analyze high dimensional data structures
- PCA is an orthogonal linear transformation that maps the measured data onto a new set of axes
    - These axes are called Principal Components
    - Each principal component points in the direction of maximum variation (or energy) remaining in the data, given the energy already accounted for in the preceding components
    - The principal axes are ordered by the amount of energy in the data they capture

Geometric illustration



Linear algebraic formulation

- For a matrix $X = \{x_{i\,j}\}_{I,J}$ (i.e., raws of X are points in $\mathbb{R}^J$), calculating the principal components is equivalent to solving the symmetric eigenvalue problem for the matrix $X^T X$
- Each principal component v_i is the i-th eigenvector computed from the spectral decomposition of $X^T X$:

$$X^T X\, v_i \;=\; \lambda_i\, v_i \quad i = 1, \ldots J$$

where λ_i is the eigenvalue corresponding to v_i

- k-th principal component

$$v_1 \;=\; \underset{||v||=1}{\operatorname{argmax}} ||X\, v||$$

$$v_k \;= \underset{||v||=1}{\operatorname{argmax}} \left|\left| \left(X - \sum_{i=1}^{k-1} X\, v_i\, v_i^T \right) \; v \right|\right|$$

where $||\cdot||$ denotes the L^2 norm

Subspace Method

Subspace Method

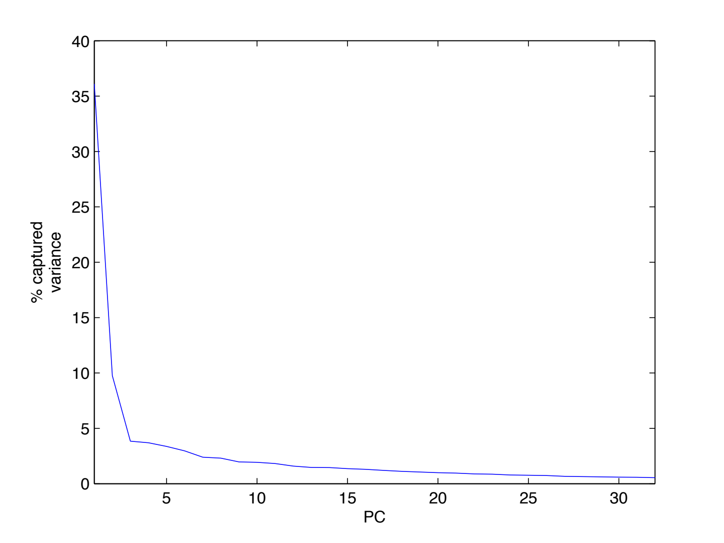
This method is based on a separation of the high-dimensional space into disjoint subspaces corresponding to normal and anomalous behavior

- This separation can be performed effectively by PCA, mapping the dataset onto the new axes
  - Normal subspace: $\hat{S}$
  - Anomalous Subspace: $\tilde{S}$
- $x \in \mathbb{R}^J$ can be written as

$$x = \hat{x} + \tilde{x} = P\,P^T\,x + \left(I - P\,P^T\right)\,x$$

where $P = \{p_{i\,j}\}_{J,R} = (v_1\ v_2\ \ldots\ v_R)$

Example of Scree Plot



СПАСИБО ЗА
ВНИМАНИЕ
И ПОПЫТКУ
ПОНИМАНИЯ