# Anomaly detection and sequential statistics in time series

Alex Shyr

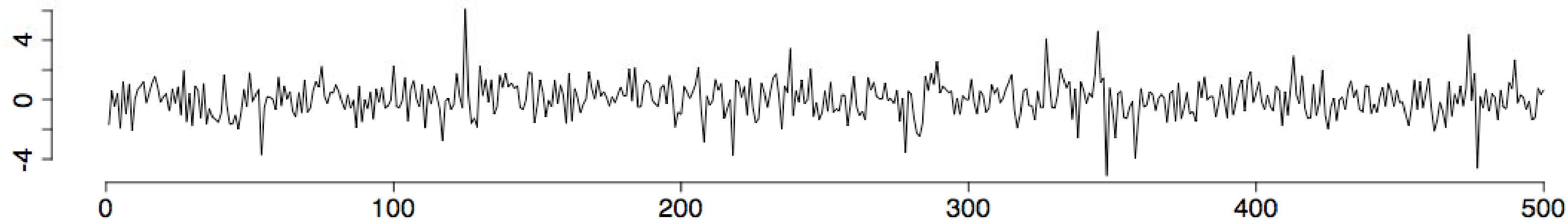
CS 294 Practical Machine Learning

11/12/2009

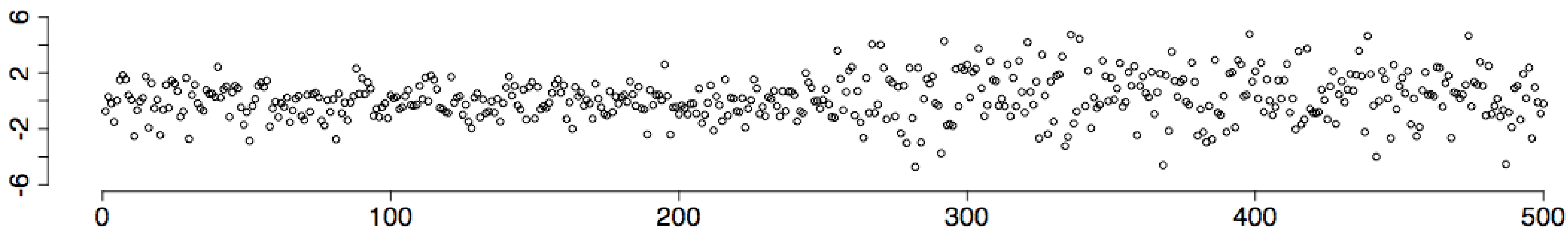
(many slides from XuanLong Nguyen and Charles Sutton)

# Two topics

## Anomaly detection

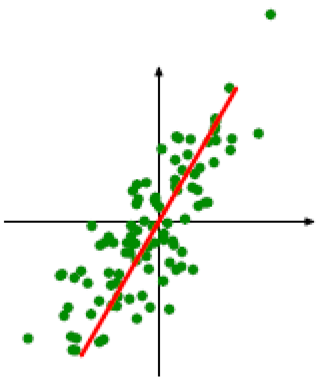



## Sequential statistics

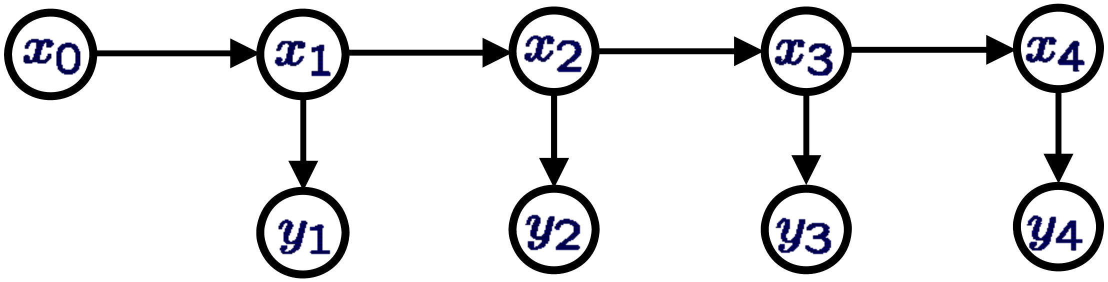


# Review

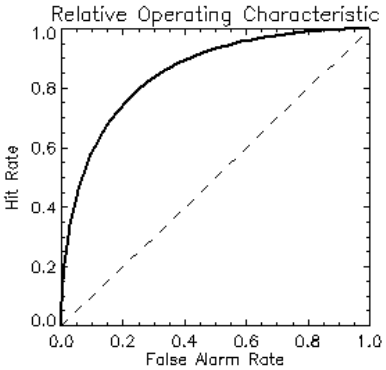
- Dimensionality Reduction
  - e.g. PCA



- HMM




- ROC curves



# Outline

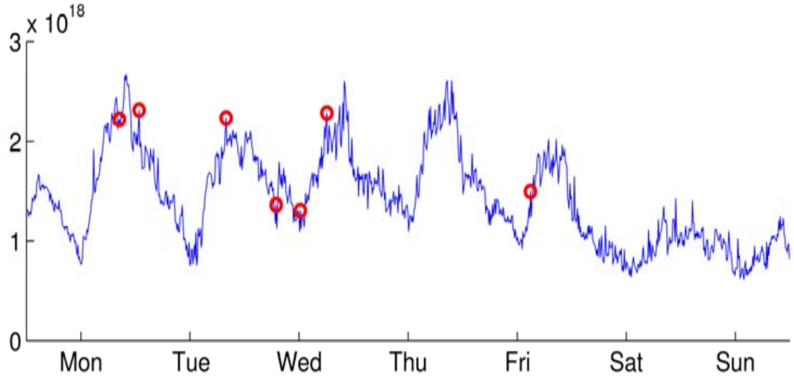
- Introduction
- Anomaly Detection
  - Static Example
  - Time Series
- Sequential Tests
  - Static Hypothesis Testing
  - Sequential Hypothesis Testing
  - Change-point Detection

# Anomalies in time series data

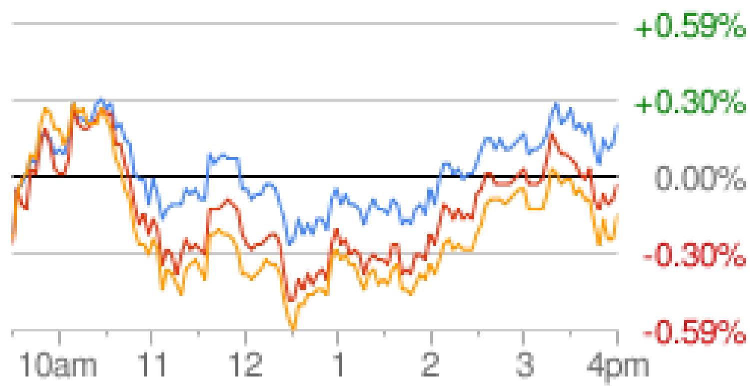
- Time series is a *sequence of data points*, measured typically at successive times, spaced at (often uniform) time intervals

- Anomalies in time series data are data points that *<u>significantly deviate</u>* from the *<u>normal pattern</u>* of the data sequence

# Examples of time series data




Network traffic data




Finance data

Human Activity data

# Applications

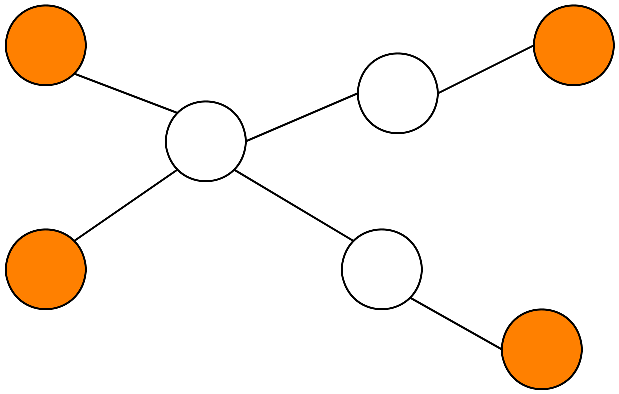
- Failure detection
- Fraud detection (credit card, telephone)
- Spam detection
- Biosurveillance
  - detecting geographic hotspots
- Computer intrusion detection

# Outline

- Introduction
- Anomaly Detection
  - **Static Example**
  - Time Series
- Sequential Tests
  - Static Hypothesis Testing
  - Sequential Hypothesis Testing
  - Change-point Detection

# Example: Network traffic

***[Lakhina et al, 2004]***



Backbone network

Goal: Find **source-destination** pairs with high traffic (e.g., by rate, volume)

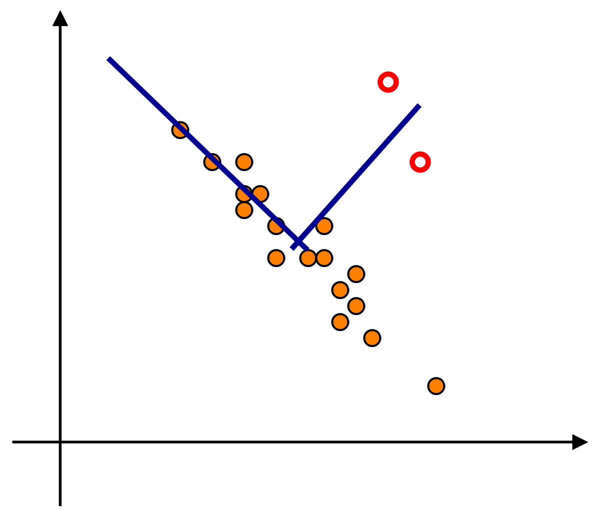
$$\mathbf{Y} = \begin{bmatrix} & & & \cdots & & \\ 100 & 30 & 42 & 212 & 1729 & 13 \\ & & & \cdots & & \end{bmatrix}$$

# Example: Network traffic

Data matrix

$$\mathbf{Y}=\begin{bmatrix} \cdots\cdot \\ 100\quad 30\quad 42\quad 212\quad 1729\quad 13 \\ \cdots\cdot \end{bmatrix}$$

Perform PCA on matrix $\mathbf{Y}$



Low-dimensional data

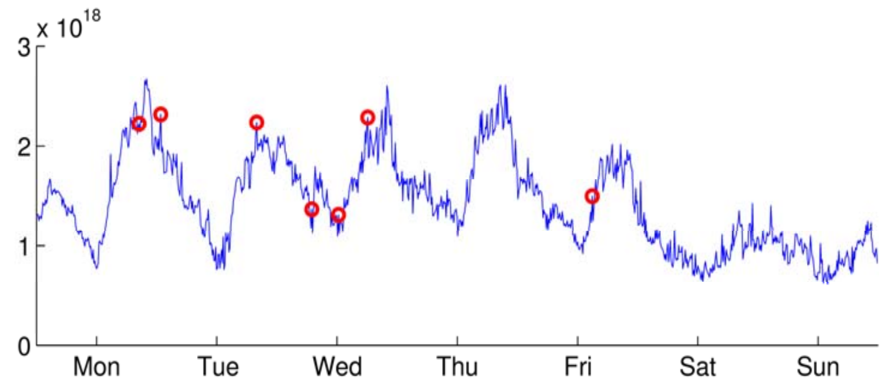
$$\mathbf{Yv}=\begin{bmatrix} \cdots\cdot \\ \mathbf{y}_t^T\mathbf{v}_1 \quad \mathbf{y}_t^T\mathbf{v}_2 \\ \cdots\cdot \end{bmatrix}$$

Eigenvectors

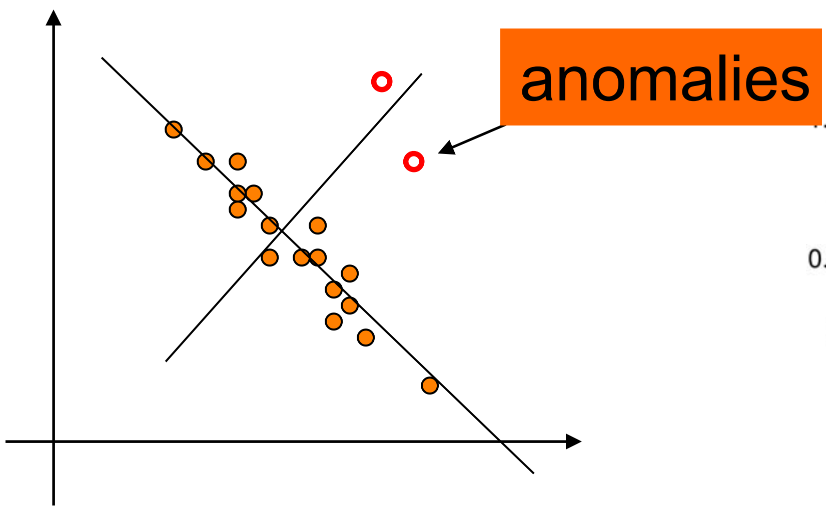
$$\begin{bmatrix} \mathbf{v}_1 \quad \mathbf{v}_2 \quad \ldots \end{bmatrix}$$

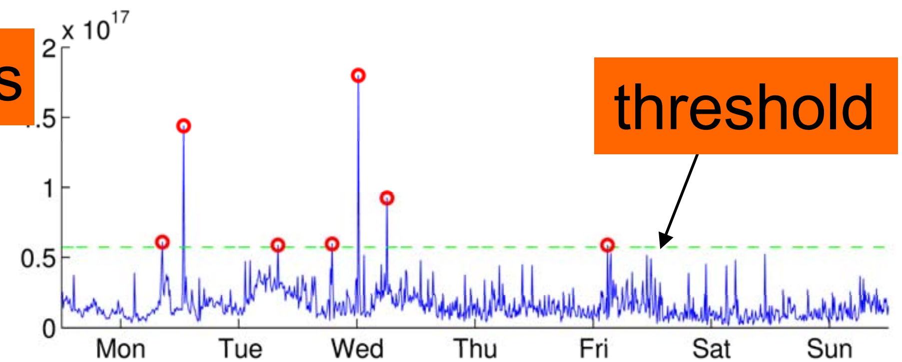
# Example: Network traffic

Abilene backbone network traffic volume over 41 links collected over 4 weeks




Perform PCA on 41-dim data
Select top 5 components

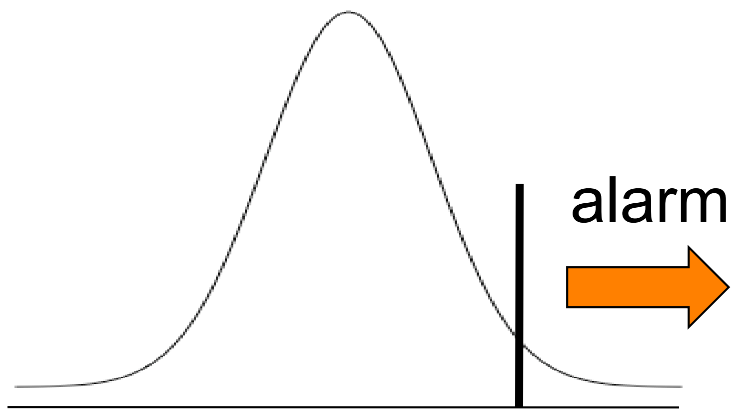



Projection to residual subspace

# Conceptual framework

- Learn a <u>model of normal behavior</u>
- Find outliers under some statistic



# Criteria in anomaly detection

- False alarm rate (type I error)
- Misdetection rate (type II error)
- Neyman-Pearson criteria
  - minimize misdetection rate while false alarm rate is bounded
- Bayesian criteria
  - minimize a weighted sum for false alarm and misdetection rate
- (Delayed) time to alarm
  - second part of this lecture

# How to use supervised data?

***D***: observed data of an account
***C***: event that a criminal present
***U***: event controlled by user
***P(D|U):*** model of normal behavior
***P(D|C):*** model for *attacker profiles*

$$\frac{p(C \mid D)}{p(U \mid D)} = \frac{p(D \mid C)}{p(D \mid U)} \frac{p(C)}{p(U)}$$

By Bayes' rule

*p(D|C)/p(D|U)* is known as the *Bayes factor*
(or *likelihood ratio*)

Prior distribution p(C) key to control false alarm

# Outline

- Introduction
- Anomaly Detection
  - Static Example
  - **Time Series**
- Sequential Tests
  - Static Hypothesis Testing
  - Sequential Hypothesis Testing
  - Change-point Detection

# Markov chain based model for detecting masqueraders

***[Ju & Vardi, 99]***

- Modeling "signature behavior" for individual users based on system command sequences
- High-order Markov structure is used
  - Takes into account last several commands instead of just the last one
  - Mixture transition distribution
- Hypothesis test using generalized likelihood ratio

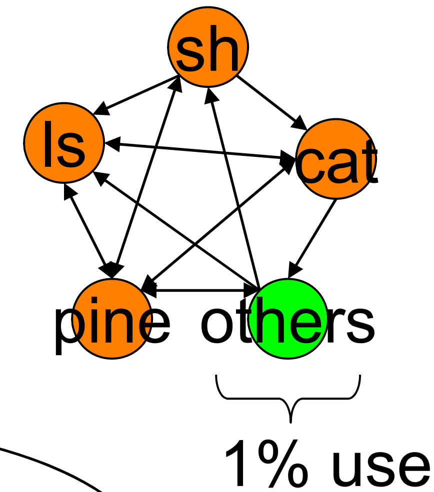
# Data and experimental design

- Data consist of sequences of (unix) system commands and user names
- 70 users, 150,000 consecutive commands each (=150 blocks of 100 commands)
- Randomly select 50 users to form a “community”, 20 outsiders
- First 50 blocks for training, next 100 blocks for testing
- Starting after block 50, randomly insert command blocks from 20 outsiders
  - For each command block i (i=50,51,...,150), there is a prob 1% that some masquerading blocks inserted after it
  - The number x of command blocks inserted has geometric dist with mean 5
  - Insert x blocks from an outside user, randomly chosen

# Markov chain profile for each user

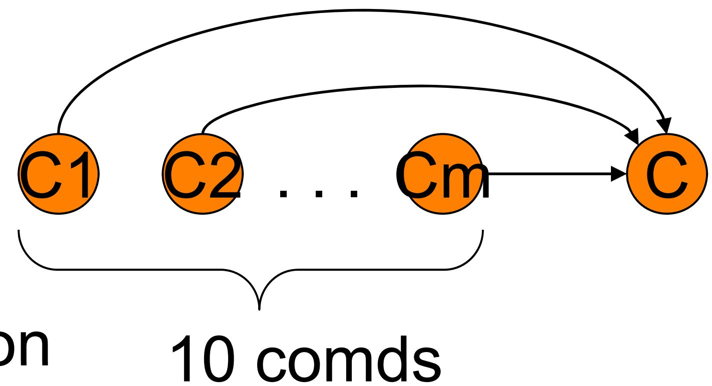
Consider the most frequently used command spaces to reduce parameter space

K = 5




Higher-order markov chain

m = 10




Mixture transition distribution

Reduce number of params from K^m to K^2 + m (why?)

$$P(C_t = s_{i_0} \mid C_{t-1} = s_{i_1}, \ldots, C_{t-m} = s_{i_m})$$

$$= \sum_{j=1}^{m} \lambda_j \, r(s_{i_0} \mid s_{i_m})$$

# Testing against masqueraders

Given command sequence $\{c_1, \ldots, c_T\}$

Learn model (profile) for each user *u* $(\Lambda_u, R_u)$

Test the hypothesis: H0 – commands generated by user u
H1 – commands NOT generated by u

Test statistic (generalized likelihood ratio):

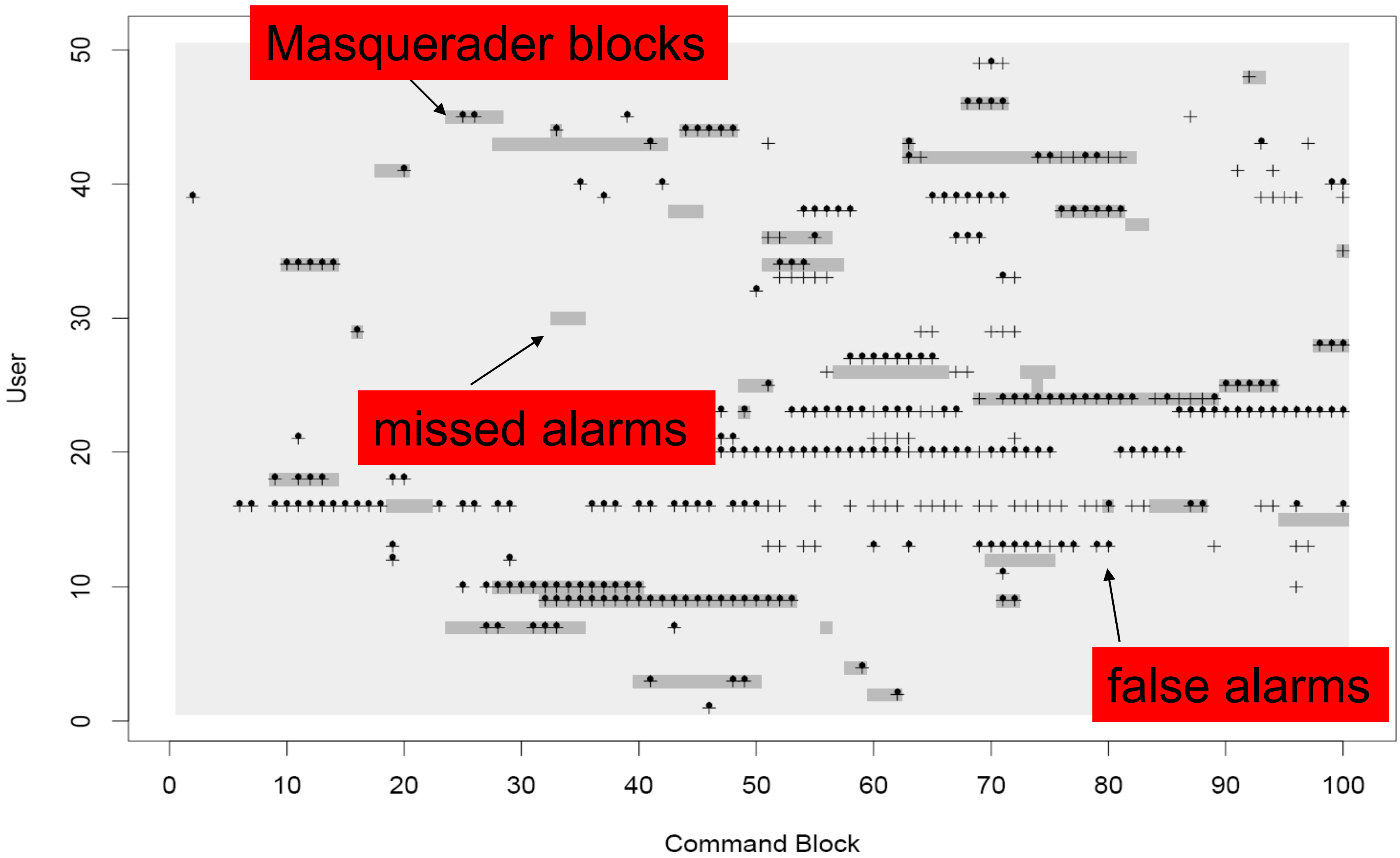
$$X = \log\left( \frac{\max_{v \neq u} P(c_1, \ldots, c_T \mid \Lambda_v, R_v)}{P(c_1, \ldots, c_T \mid \Lambda_u, R_u)} \right)$$

Raise flag whenever

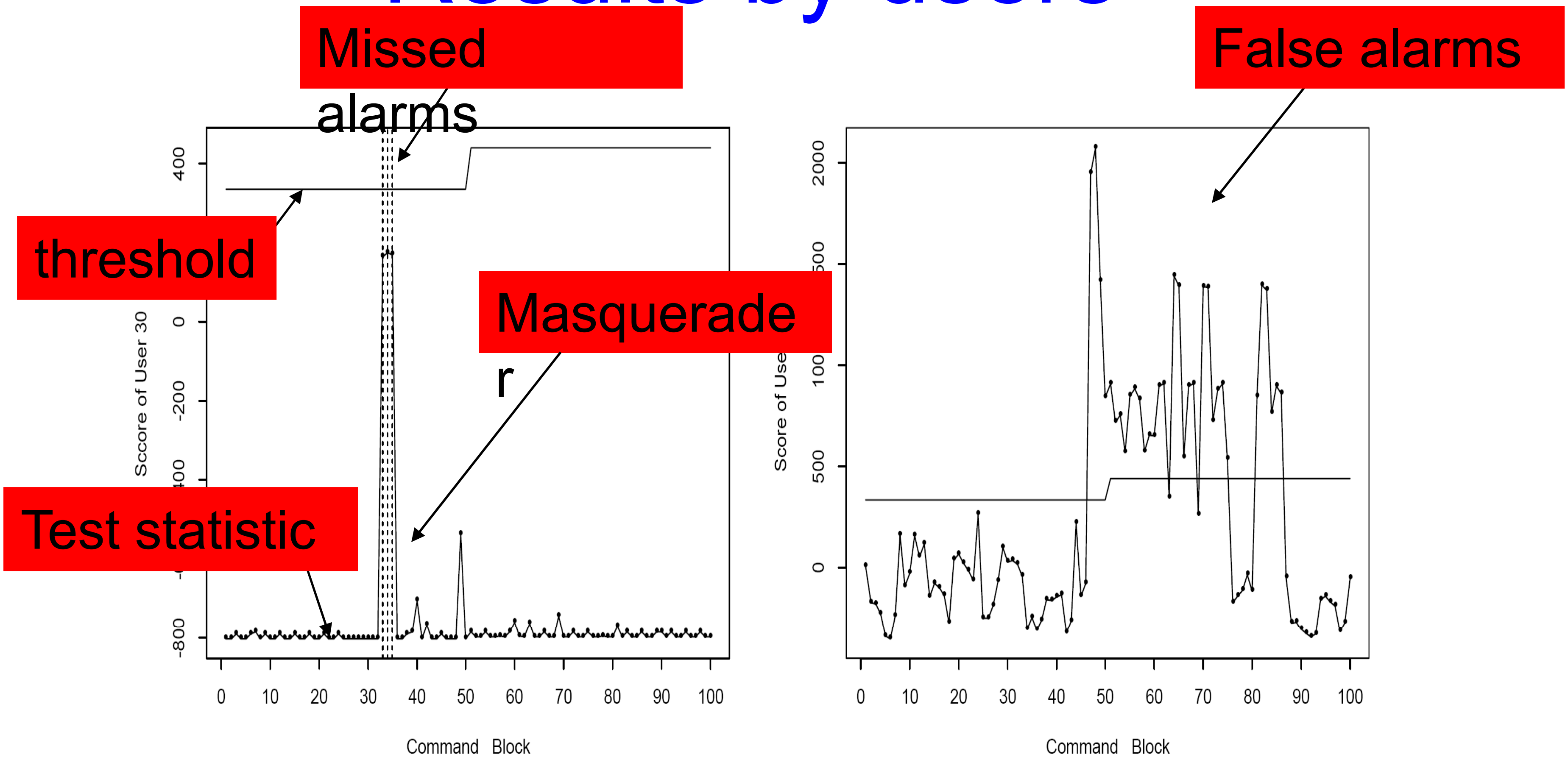
*X* > some threshold *w*

● with updating (163 false alarms, 115 missed alarms, 93.5% accuracy)

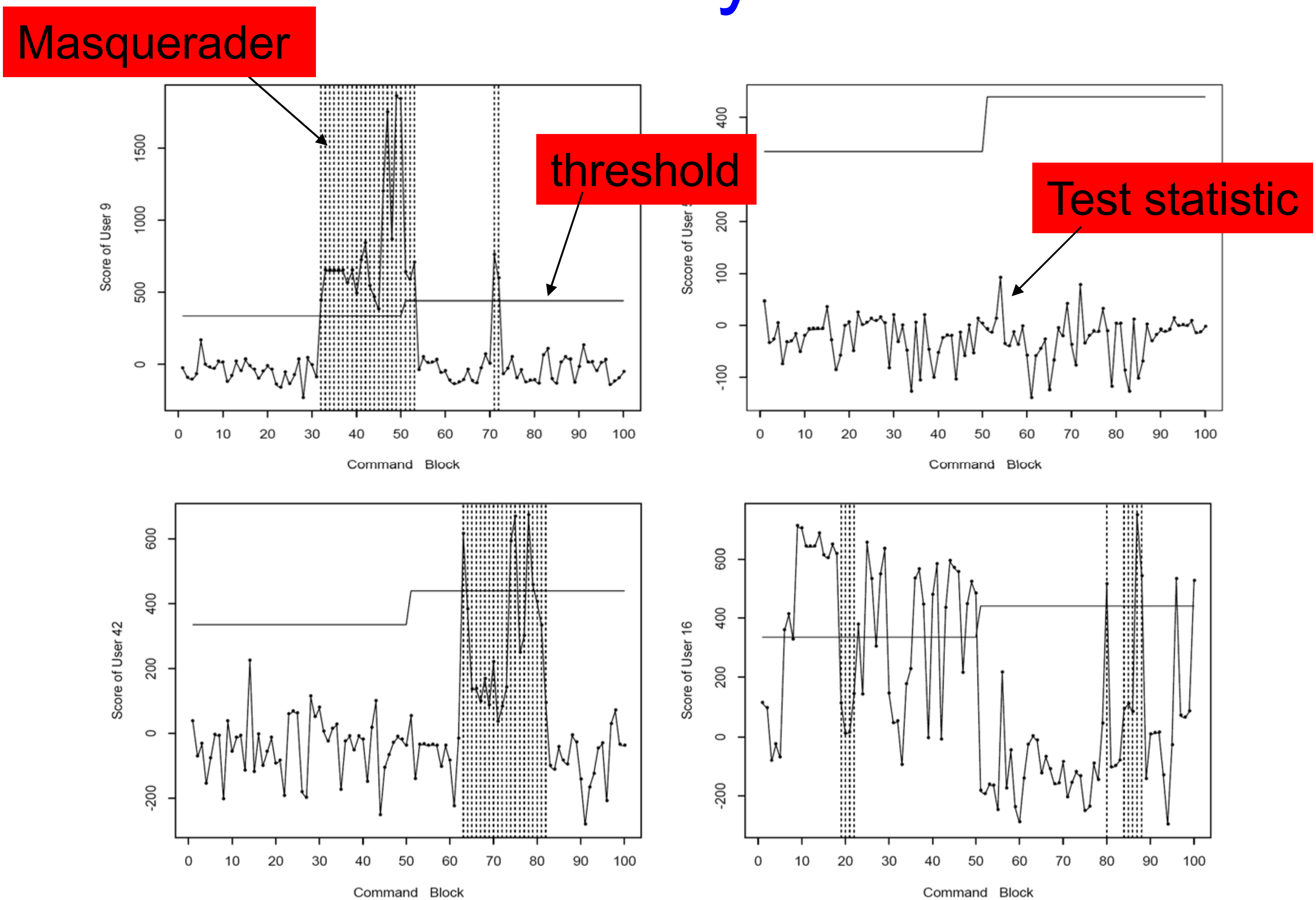
\+ without updating (221 false alarms, 103 missed alarms, 94.4% accuracy)



# Results by users



# Results by users



# Take-home message

- Learn a model of normal behavior for each monitored individuals
- Based on this model, construct a suspicion score
  - function of observed data

    (e.g., likelihood ratio/ Bayes factor)
  - captures the deviation of observed data from normal model
  - raise flag if the score exceeds a threshold

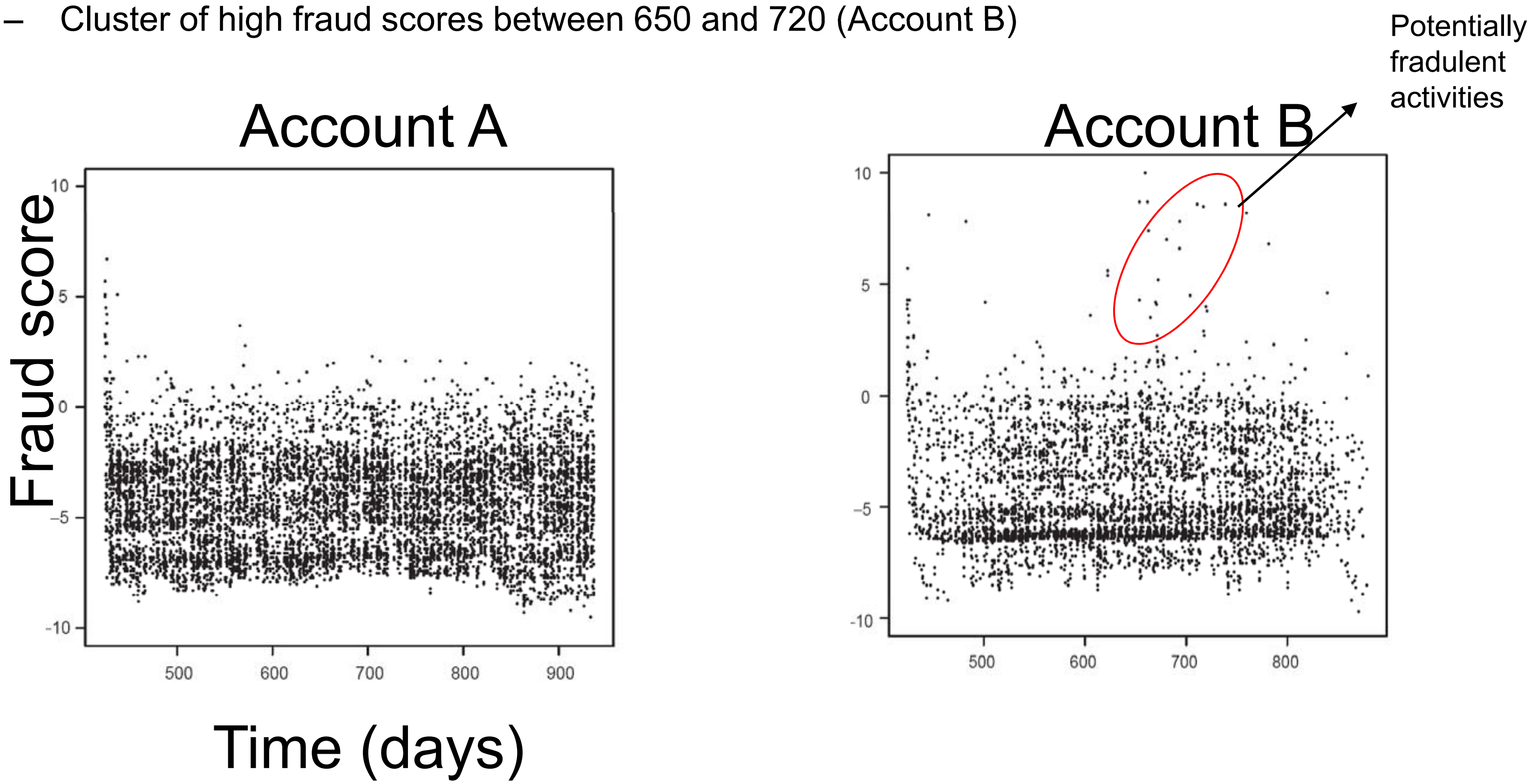
# Other models in literature

- Simple metrics
  - Hamming metric [Hofmeyr, Somayaji & Forest]
  - Sequence-match [Lane and Brodley]
  - IPAM (incremental probabilistic action modeling) [Davison and Hirsh]
  - PCA on transitional probability matrix [DuMouchel and Schonlau]
- More elaborate probabilistic models
  - Bayes one-step Markov [DuMouchel]
  - Compression model
  - Mixture of Markov chains [Jha et al]

- Elaborate probabilistic models can be used to obtain answer to more elaborate queries
  - Beyond yes/no question (see next slide)

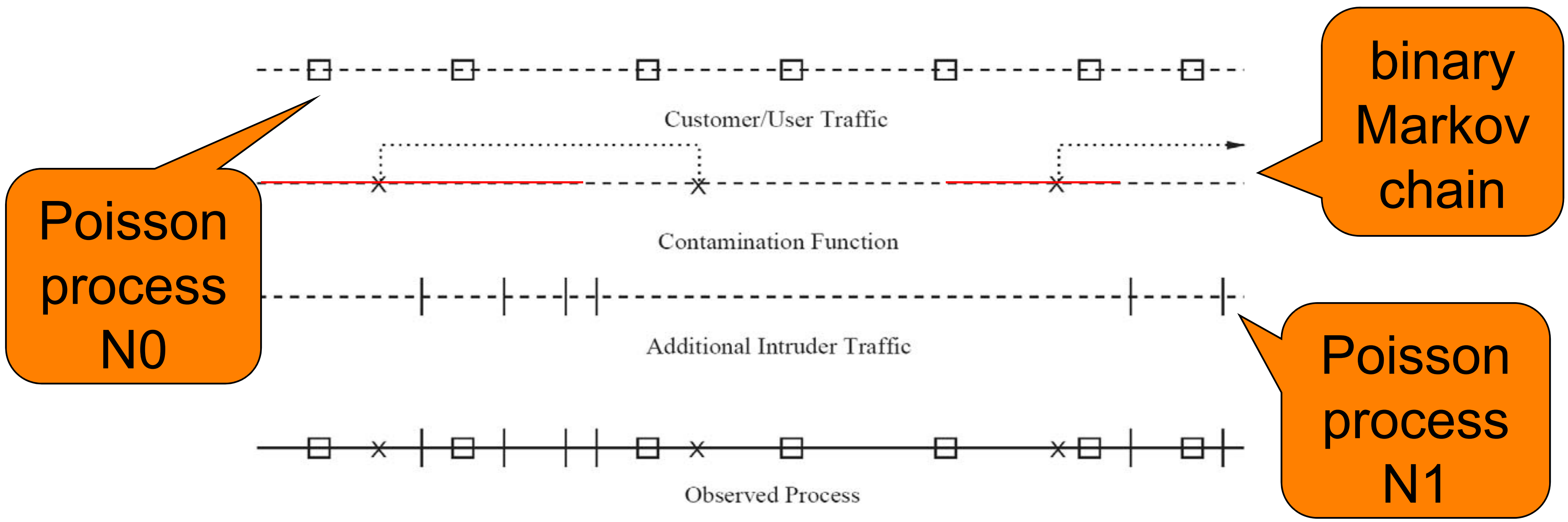
# Example: Telephone traffic (AT&T)

- Problem: Detecting if the phone usage of an account is abnormal or not **[*Scott, 2003*]**
- Data collection: phone call records and summaries of an account's previous history
  - Call duration, regions of the world called, calls to "hot" numbers, etc
- Model learning: A learned profile for each account, as well as separate profiles of known intruders
- Detection procedure:
  - Cluster of high fraud scores between 650 and 720 (Account B)



# Burst modeling using Markov modulated Poisson process

[Scott, 2003]




- can be also seen as a nonstationary discrete time HMM (thus all inferential machinary in HMM applies)
- requires less parameter (less memory)
- convenient to model sharing across time

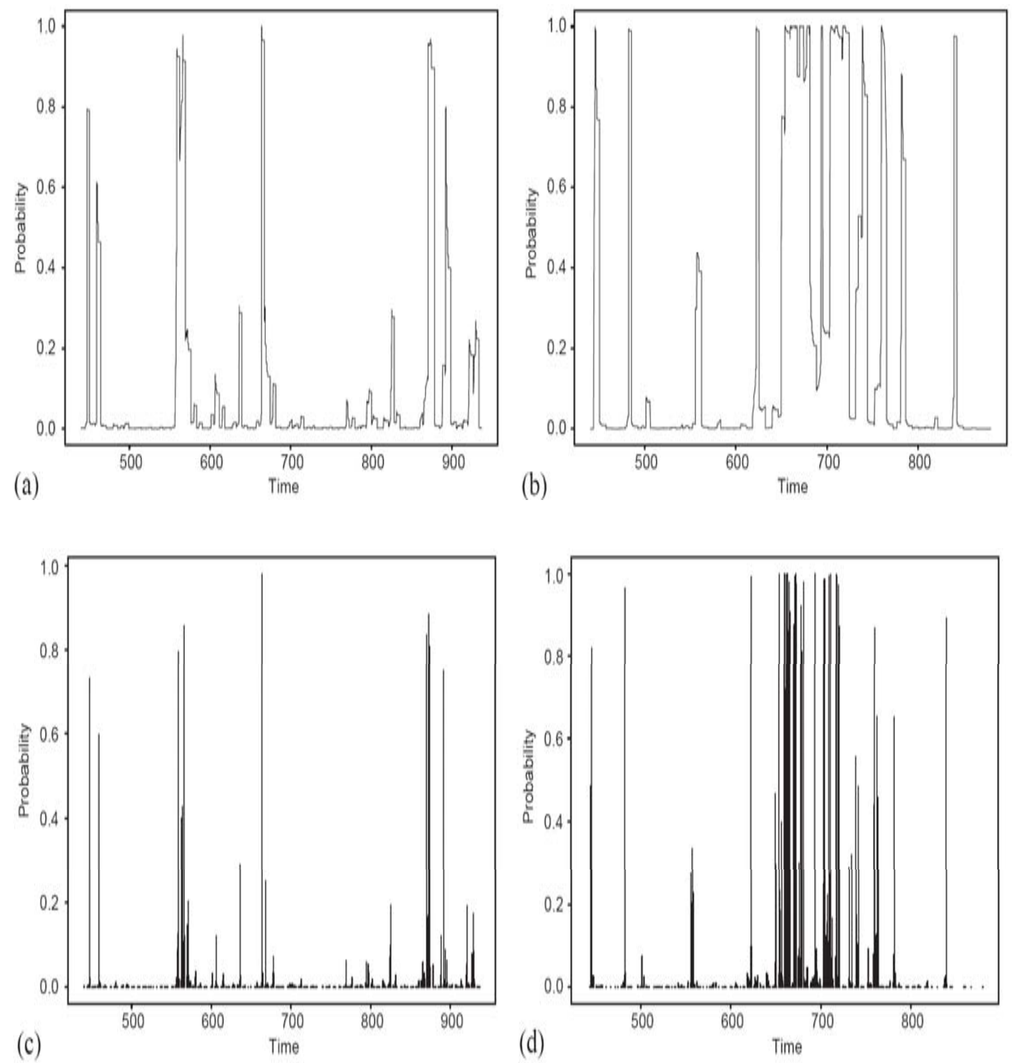
# Detection results

probability of a criminal presence

probability of each phone call being intruder traffic

Uncontaminated account

Contaminated account

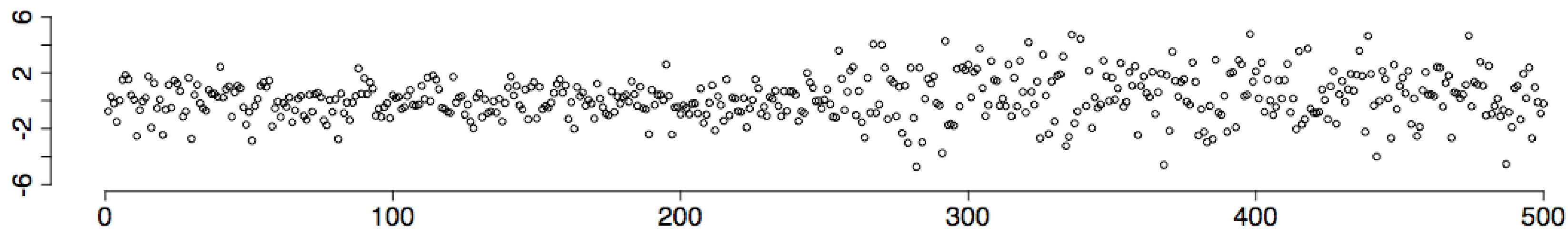


# Outline

- Introduction
- Anomaly Detection
  - Static Example
  - Time Series
- Sequential Tests
  - Static Hypothesis Testing
  - Sequential Hypothesis Testing
  - Change-point Detection

# Sequential analysis outline

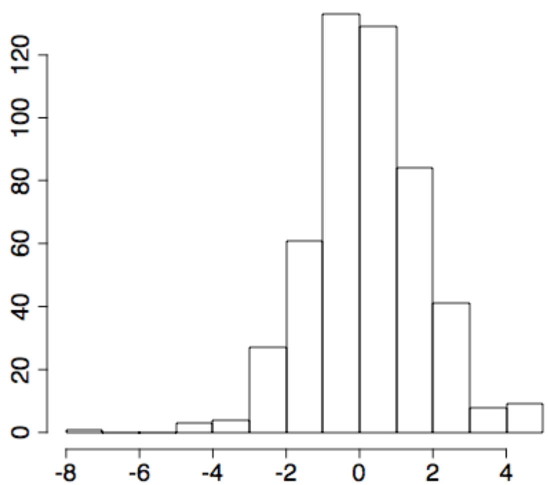
- Two basic problems
  - sequential hypothesis testing
  - sequential change-point detection
- **Goal**: minimize detection delay time



# Outline

- Introduction
- Anomaly Detection
  - Static Example
  - Time Series
- Sequential Tests
  - **Static Hypothesis Testing**
  - Time Series

# Hypothesis testing




(same data as last slide)

$H_0 : \mu = 0$ null hypothesis

$H_1 : \mu > 0$ alternative hypothesis

Test statistic:

$$t = \frac{\overline{X}}{s}$$

Reject $H_0$ if $t > c_\alpha$

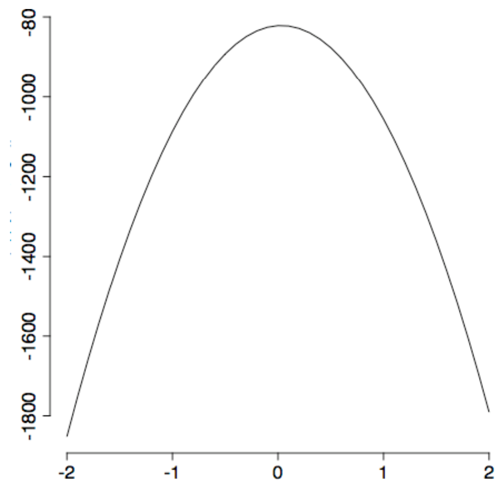
for desired false negative rate α

# Likelihood

Suppose the data have density $p(x;\mu)$

$$p(x;\mu)=\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}x^2\right\}$$

The **likelihood** is the probability of the observed data, as a function of the parameters.



# Likelihood Ratios

To compare two parameter values $\mu_0$ and $\mu_1$ given independent data $x_1 \ldots x_n$:

$$\Lambda = \log \frac{l(\mu_1)}{l(u_0)} = \sum_{i=1}^{n} \log \frac{f(x_i;\mu_1)}{f(x_i;\mu_0)}$$

This is the likelihood ratio. A hypothesis test (analogous to the t-test) can be devised from this statistic.

What if we want to compare two *regions* of parameter space? For example, H0: μ=0, H1: μ > 0. Then we can maximize over all the possible μ in H1.

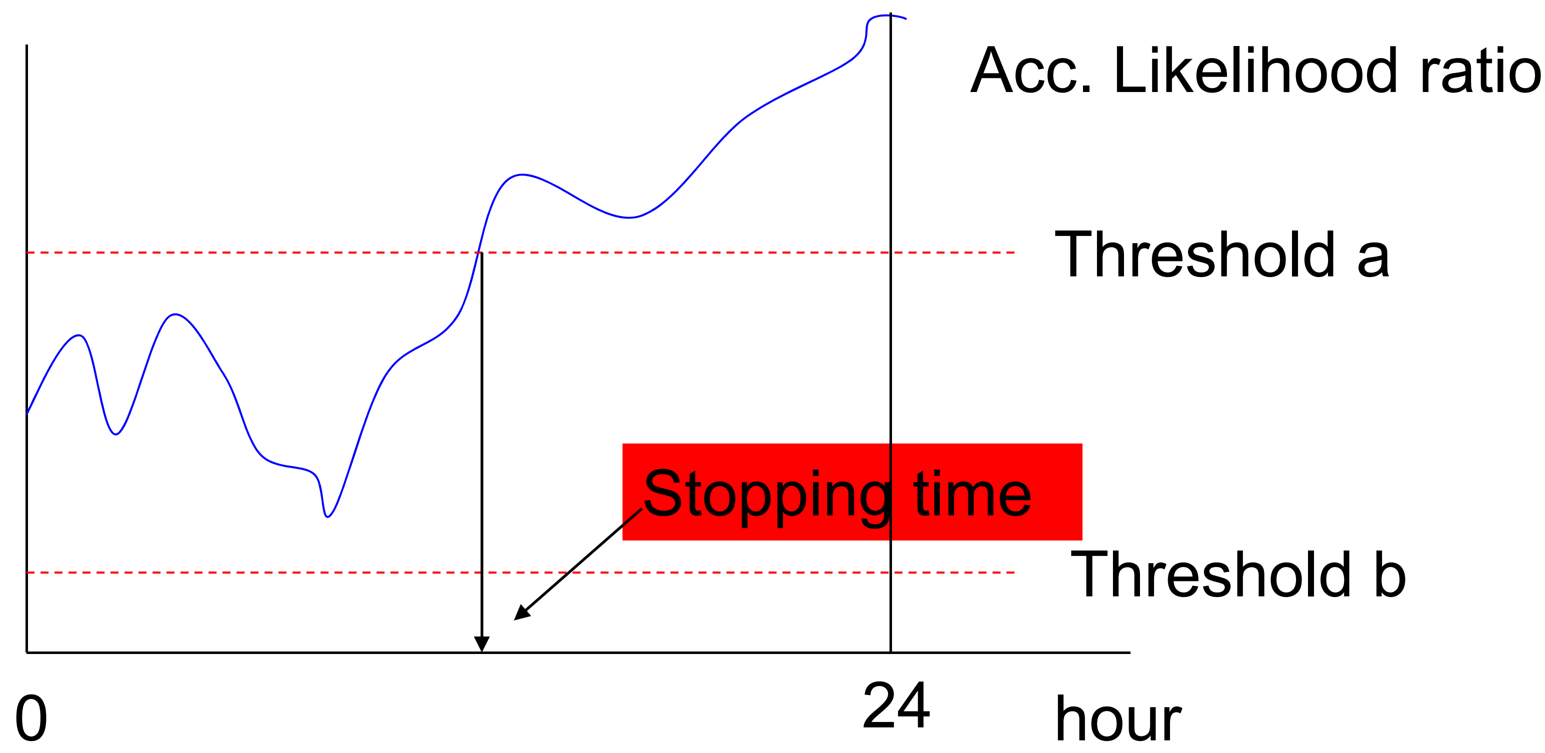
This yields the generalized likelihood ratio test (see later in lecture).

# Outline

- Introduction
- Anomaly Detection
  - Static Example
  - Time Series
- Sequential Tests
  - Static Hypothesis Testing
  - **Sequential Hypothesis Testing**
  - Change-point Detection

# A sequential solution

1. Compute the accumulative likelihood ratio statistic
2. Alarm if this exceeds some threshold



# Quantities of interest

- False alarm rate $\alpha = P(D=1 \mid H_0)$
- Misdetection rate $\beta = P(D=0 \mid H_1)$
- Expected stopping time (aka number of samples, or decision delay time) $E\ N$

**Frequentist formulation:**

Fix $\alpha, \beta$

Minimize $E[N]$

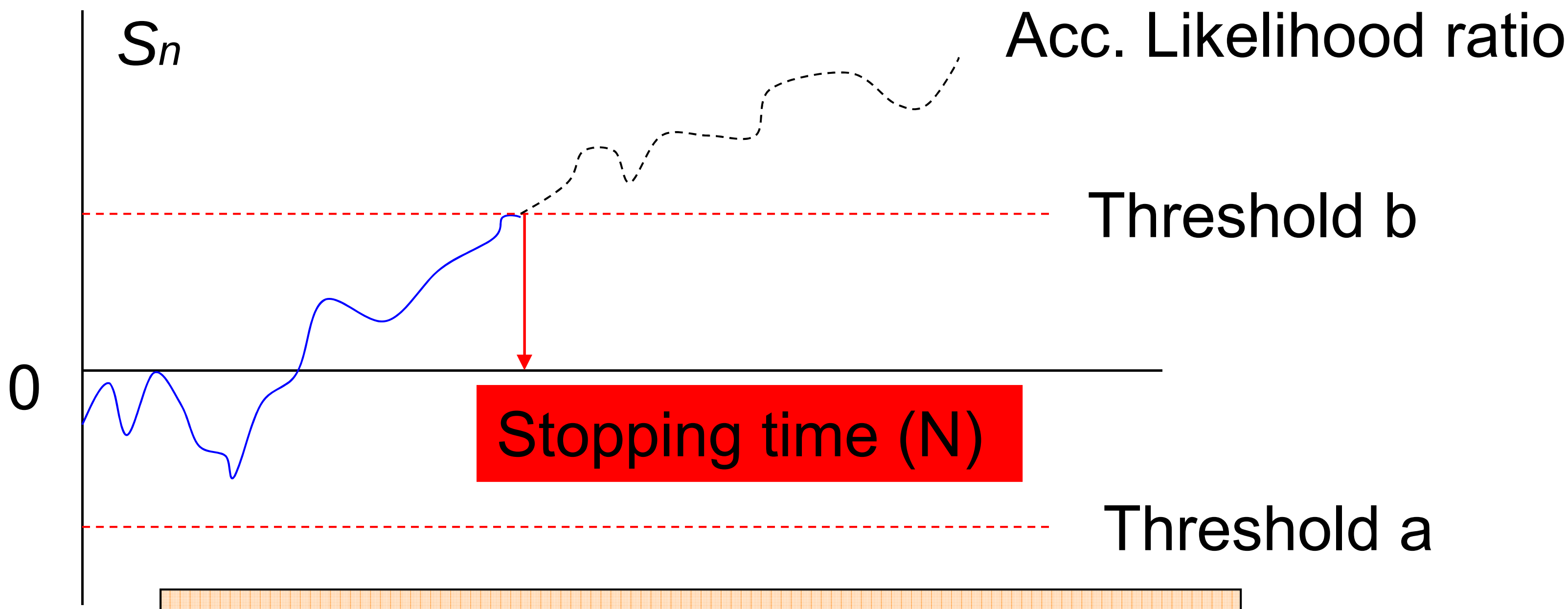
wrt both $f_0$ and $f_1$

**Bayesian formulation:**

Fix some weights $c_1, c_2, c_3$

Minimize $c_1\alpha + c_2\beta + c_3 E[N]$

# Sequential likelihood ratio test




Wald's approximation:

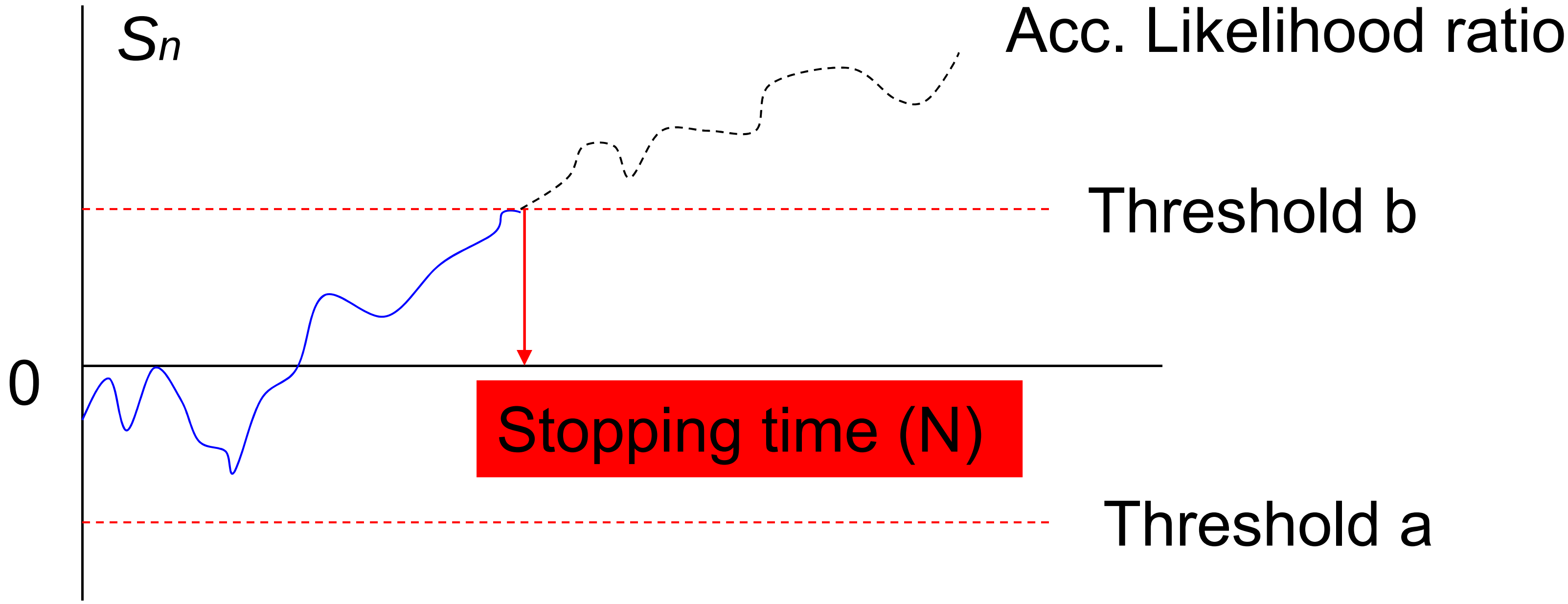
$$a \geq \log \frac{\beta}{1-\alpha} \quad \Rightarrow a \approx \log \frac{\beta}{1-\alpha}$$

$$b \leq \log \frac{1-\beta}{\alpha} \quad \Rightarrow \quad b \approx \log \frac{1-\beta}{\alpha}$$

$$\text{So, } \alpha \approx \frac{1-e^{a}}{e^{b}-e^{a}} \quad \text{and} \quad \beta \approx \frac{e^{-b}-1}{e^{-b}-e^{-a}}$$

Exact if there's no overshoot!

# Sequential likelihood ratio test




Choose α and β

Compute a, b according to Wald's approximation

$S_i = S_{i-1} + \log \Lambda_i$

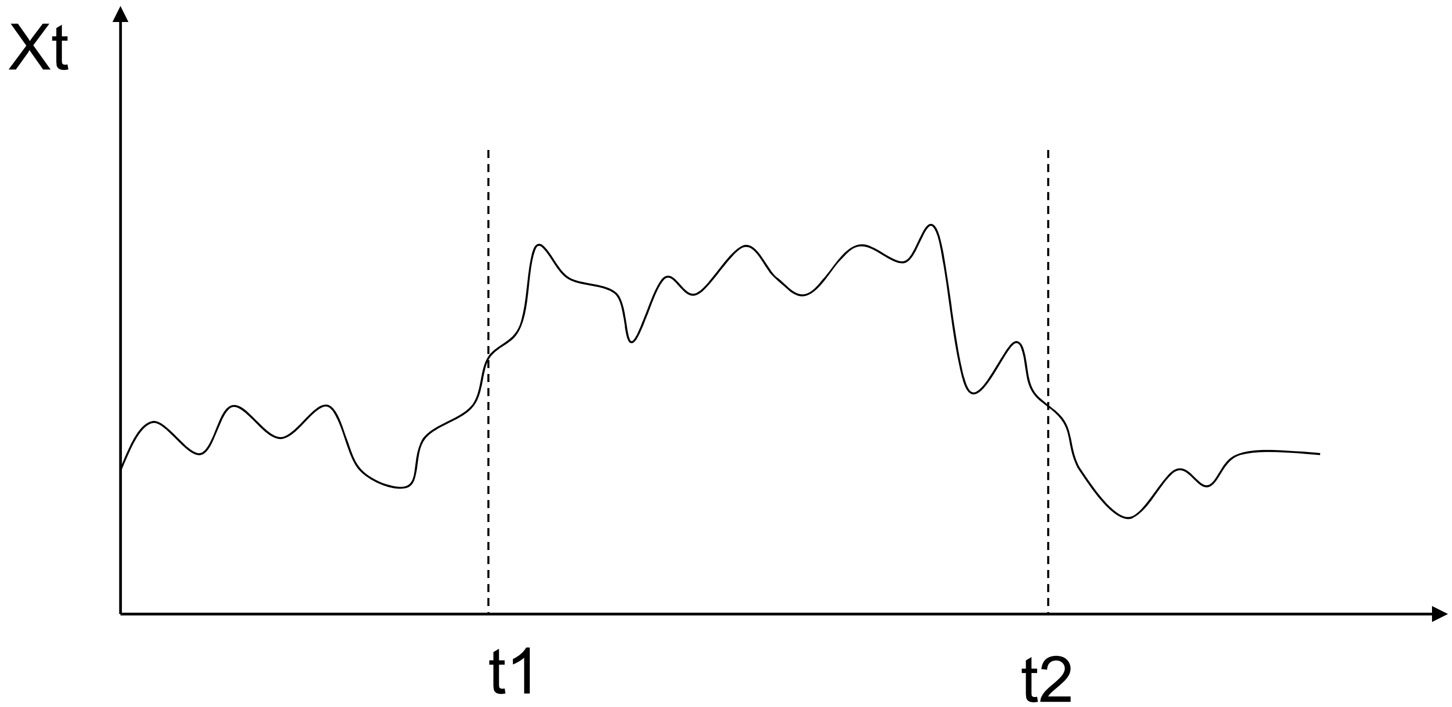
if $S_i >= b$: accept $H_1$

if $S_i <= a$: accept $H_0$

# Outline

- Introduction
- Anomaly Detection
  - Static Example
  - Time Series
- Sequential Tests
  - Static Hypothesis Testing
  - Sequential Hypothesis Testing
  - **Change-point Detection**

# Change-point detection problem




Identify where there is a change in the data sequence

- change in mean, dispersion, correlation function, spectral density, etc…
- generally change in distribution

# Motivating Example: Shot Detection

- Simple absolute pixel difference





# Maximum-likelihood method

[Page, 1965]

$X_1, X_2, ..., X_n$ are observed

For each $v = 1,2,..., n$, consider hypothesis $H_v$

$v$ is uniformly dist. $\{1,2,..., n\}$

Likelihood function corresponding to $H_v$ :

$$l_v(x) = \sum_{i=1}^{v-1} \log f_0(x_i) + \sum_{i=v}^{n} \log f_1(x_i)$$

Hv: sequence has density f0 before v, and f1 after

H0: sequence is stochastically homogeneous

MLE estimate : $H_v$ is accepted if

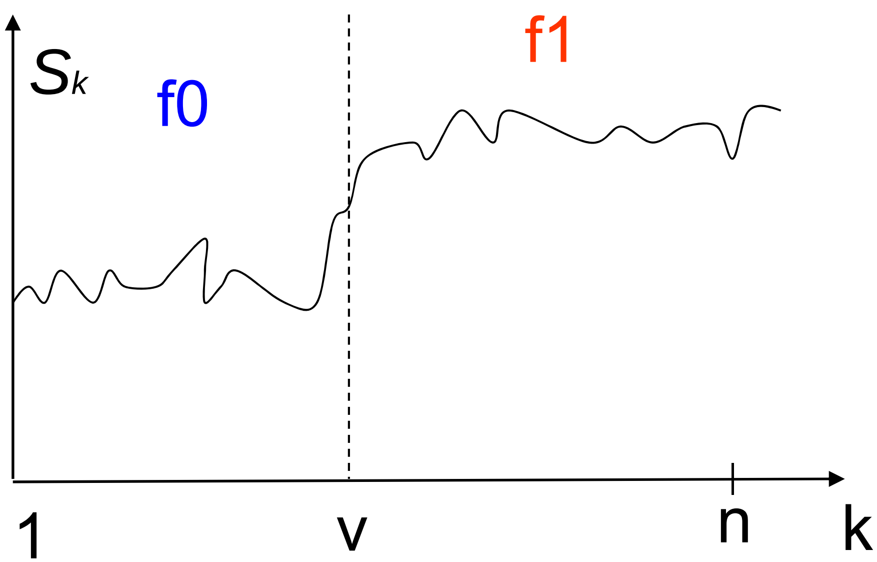
$l_v(x) \geq l_j(x)$ for all $j \neq v$

Let $S_k$ be the likelihood ratio up to $k$,

$$S_k = \sum_{i=1}^{k} \log \frac{f_1(x_i)}{f_0(x_i)}$$

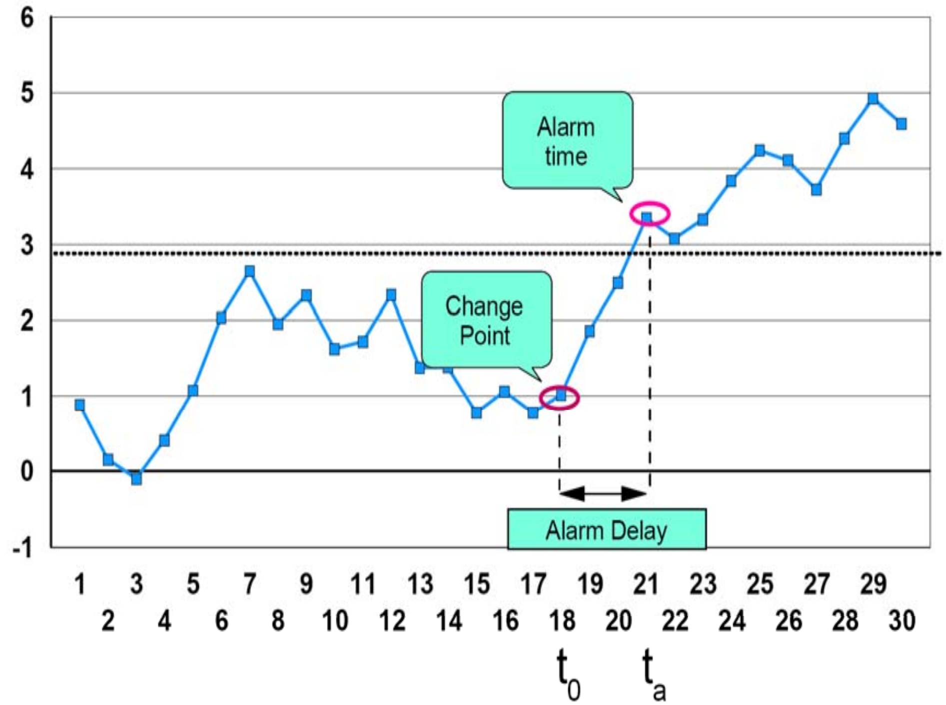
then our estimate can be written as

$v := k \mid S_k \leq S_v \forall k \leq v, \quad S_k \geq S_v \forall k \geq v$



# Sequential change-point detection

- Data are observed serially
- There is a change in distribution at ***t0***
- Raise an alarm if change is detected at ***ta***




Need to minimize

Average observation time before false alarm $E_{f_0}[t_a]$

Average delay time of detection $E_{f_1}[t_a]$

# Cusum test (Page, 1966)

Likelihood of composite hypothesis $H_\nu$ against $H_0$:

$$\max_{0\le k\le n}(S_n - S_k) = S_n - \min_{0\le k\le n} S_k,$$

where

$$S_0 = 0;\ S_k = \sum_{j=1}^{k} \log \frac{f_1(x_j)}{f_0(x_j)}$$

Hv: sequence has density f0 before v, and f1 after

H0: sequence is stochastically homogeneous

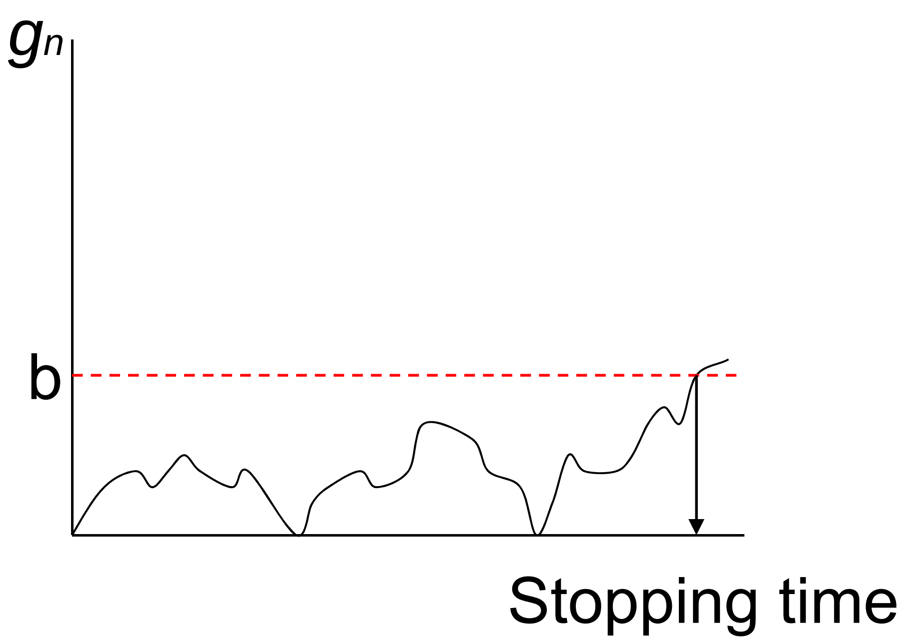
Stopping rule:

$$N = \min\{n \ge 1 : g_n = S_n - \min_{0\le k\le n} S_k \ge b\}$$

for some threshold $b$

$g_n$ can be written in recurrent form

$$g_0 = 0; g_n = \max(0, g_{n-1} + \log \frac{f_1(x_n)}{f_0(x_n)})$$



# Generalized likelihood ratio

Unfortunately, we don't know $f_0$ and $f_1$
Assume that they follow the form

$$f_i \sim P(x \mid \theta_i) \mid \quad i = 0,1$$

$f_0$ is estimated from "normal" training data
$f_1$ is estimated on the flight (on test data)

$$\theta_1 := \arg\max_\theta P(X_1, \ldots, X_n)$$

Sequential generalized likelihood ratio statistic:

$$R_n = \max_{\theta_1} \sum_{j=1}^{k} \log \frac{f_1(x_j \mid \theta_1)}{f_0(x_j)}$$

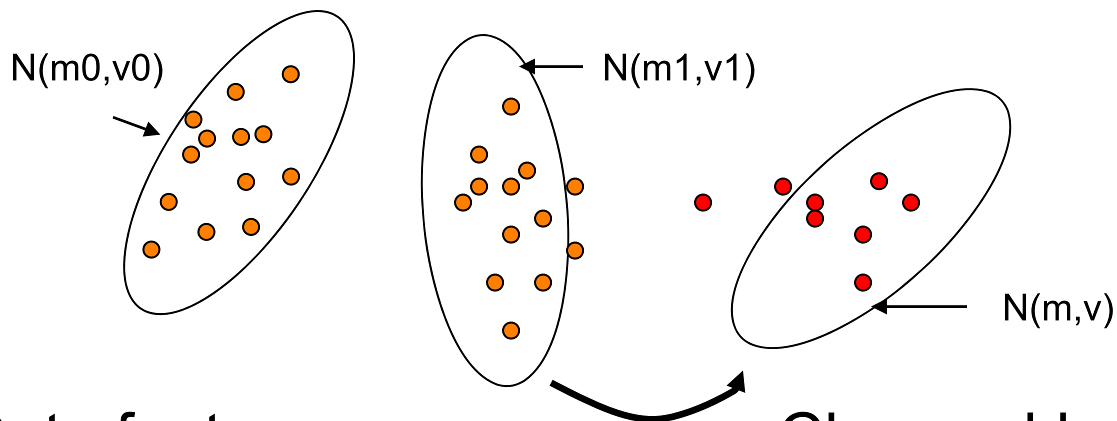
$$S_n = \max_{0 \le k \le n} (R_n - R_k)$$

Our testing rule: Stop and declare the change point at <u>the first n</u> such that

$S_n$ exceeds a threshold $w$

# Change point detection in network traffic

[Hajji, 2005]




Data features: Changed behavior

number of good packets received that were directed to the broadcast address

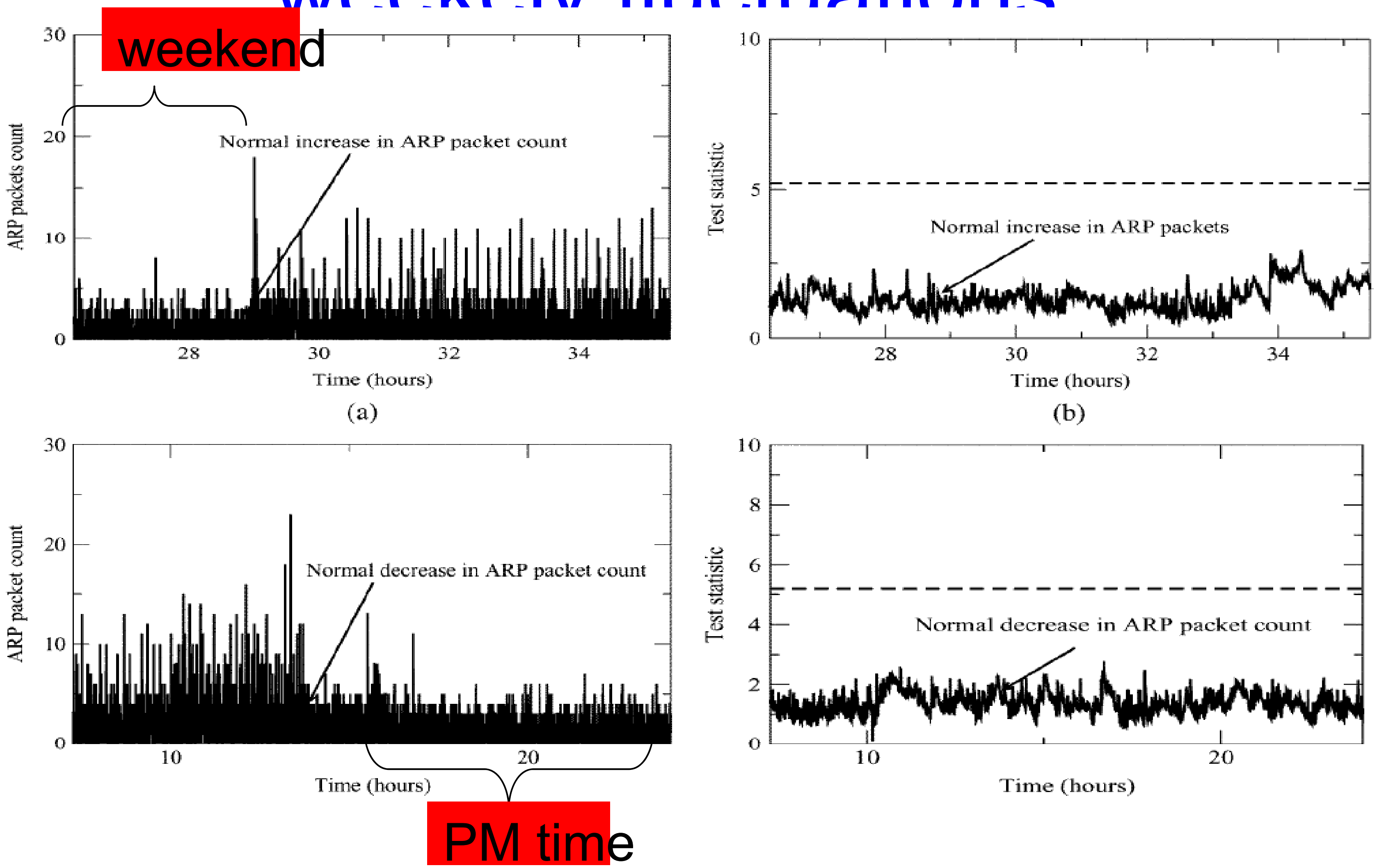
number of Ethernet packets with an unknown protocol type

number of good address resolution protocol (ARP) packets on the segment

number of incoming TCP connection requests (TCP packets with SYN flag set)

Each feature is modeled as a mixture of 3-4 gaussians to adjust to the daily traffic patterns (night hours vs day times weekday vs. weekends….)

# Adaptability to normal daily and weekely fluctuations

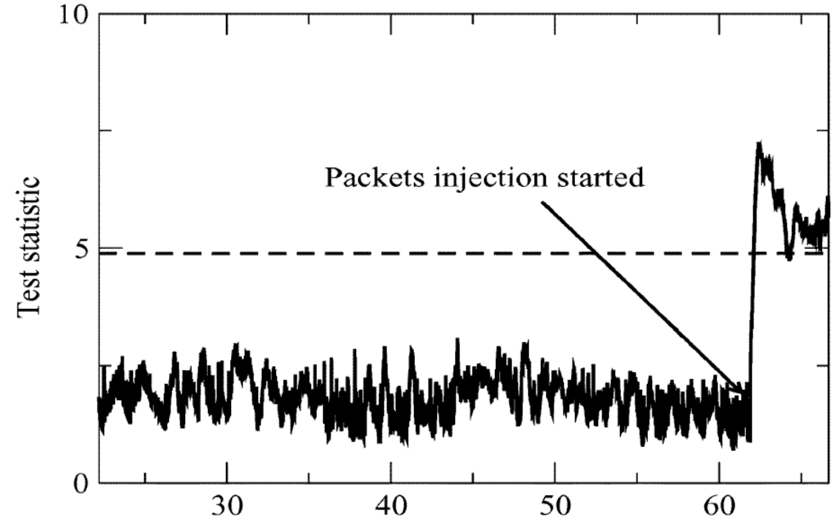


# Anomalies detected

Broadcast storms, DoS attacks

*injected 2 broadcast/sec*

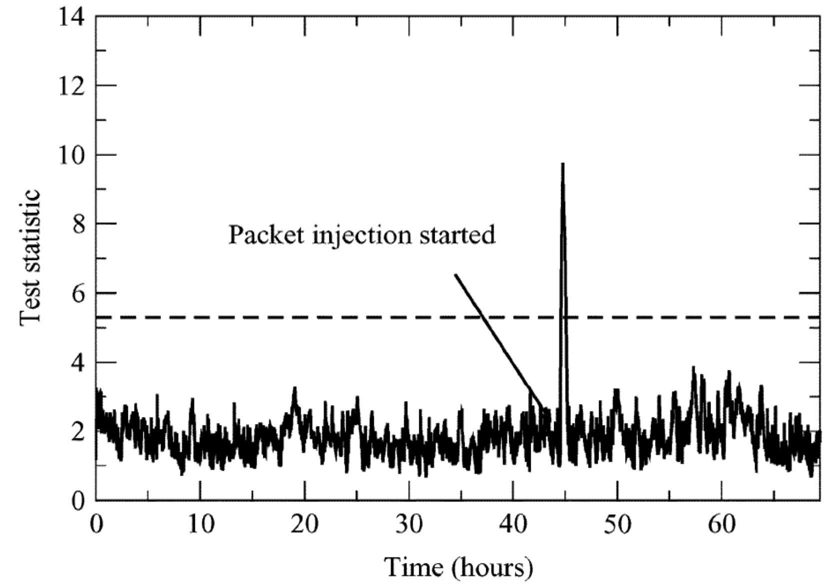
16mins delay




Sustained rate of TCP connection requests

*injecting 10 packets/sec*

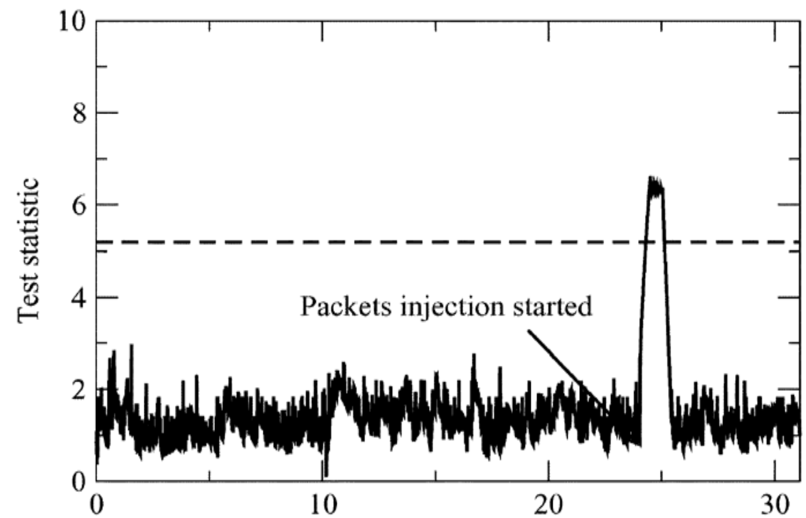
17mins delay



# Anomalies detected

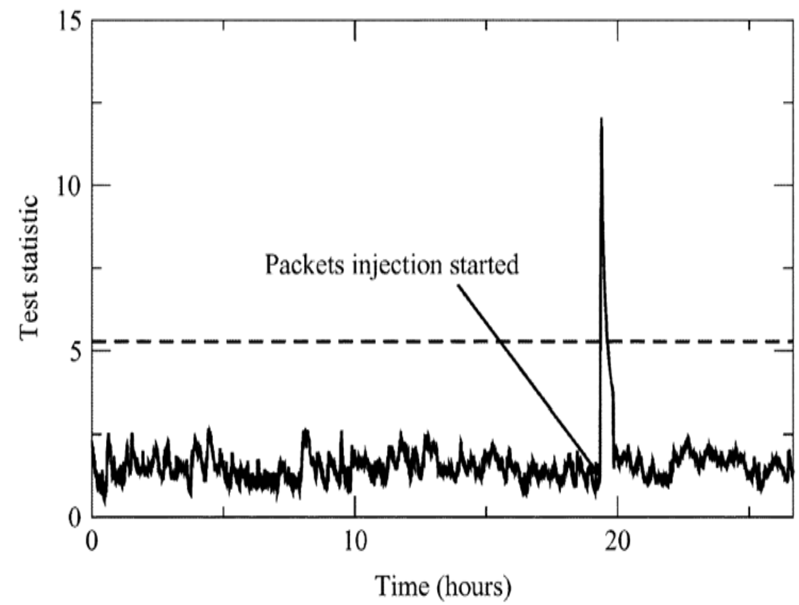
ARP cache poisoning attacks

16 min delay




TCP SYN DoS attack, excessive traffic load

50s delay



# References for anomaly detection


- Schonlau, M, DuMouchel W, Ju W, Karr, A, theus, M and Vardi, Y. Computer instrusion: Detecting masquerades, *Statistical Science*, 2001.
- Jha S, Kruger L, Kurtz, T, Lee, Y and Smith A. A filtering approach to anomaly and masquerade detection. Technical report, Univ of Wisconsin, Madison.
- Scott, S., A Bayesian paradigm for designing intrusion detection systems. *Computational Statistics and Data Analysis*, 2003.
- Bolton R. and Hand, D. Statistical fraud detection: A review. *Statistical Science*, Vol 17, No 3, 2002,
- Ju, W and Vardi Y. A hybrid high-order Markov chain model for computer intrusion detection. Tech Report 92, National Institute Statistical Sciences, 1999.
- Lane, T and Brodley, C. E. Approaches to online learning and concept drift for user identification in computer security. *Proc. KDD*, 1998.
- Lakhina A, Crovella, M and Diot, C. diagnosing network-wide traffic anomalies. *ACM Sigcomm*, 2004

# References for sequential analysis


- Wald, A. Sequential analysis, John Wiley and Sons, Inc, 1947.
- Arrow, K., Blackwell, D., Girshik, *Ann. Math. Stat*., 1949.
- Shiryaev, R. Optimal stopping rules, Springer-Verlag, 1978.
- Siegmund, D. Sequential analysis, Springer-Verlag, 1985.
- Brodsky, B. E. and Darkhovsky B.S. Nonparametric methods in change-point problems. Kluwer Academic Pub, 1993.
- Lai, T.L., Sequential analysis: Some classical problems and new challenges (with discussion), *Statistica Sinica*, 11:303—408, 2001.
- Mei, Y. Asymptotically optimal methods for sequential change-point detection, Caltech PhD thesis, 2003.
- Baum, C. W. and Veeravalli, V.V. A Sequential Procedure for Multihypothesis Testing. *IEEE Trans on Info Thy,* 40(6)1994-2007, 1994.
- Nguyen, X., Wainwright, M. & Jordan, M.I. On optimal quantization rules in sequential decision problems. *Proc. ISIT*, Seattle, 2006.
- Hajji, H. Statistical analysis of network traffic for adaptive faults detection, *IEEE Trans Neural Networks*, 2005.