

# MultiResolution Anomaly Detection for Long Range Dependent Time Series

Lingsong Zhang  
lszhang@email.unc.edu  
University of North Carolina at Chapel Hill

March 21, 2007

## Outline

- 1 Background and Introduction
- 2 Method
- 3 Theoretical properties
- 4 Simulation
- 5 Discussion

This work are advised by my advisors [Dr. Zhengyuan Zhu](#) and [Dr. J. S. Marron](#).

Internet data sets are collected by the [UNC Internet Study Group](#).

Thanks  
[Jeff Terrell](#), [Kevin Jeffay](#), [Don Smith](#), [Haipeng Shen](#), [Andrew Nobel](#)

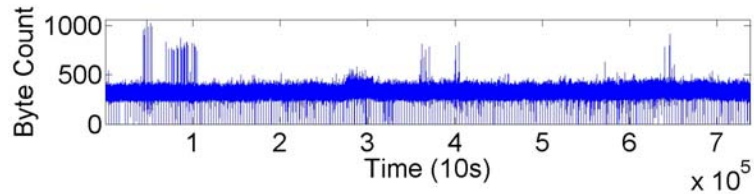
## Background and Introduction

- Internet intrusion detection
  - Intrusions become a large problem in Internet
  - Detection methods (see McHugh (2001))
    - Pattern matching : Signature based-misuse method
    - Anomaly Detection
- Anomaly Detection
  - "Normal" traffic
  - Network anomalies
- Statistical Outlier Detection

## Internet Traffic Data

- Features of data collected at a single location
  - packet count, byte count, etc. at a given time interval
  - time series of counts

*10ms Byte-Count data, 13:00 -15:00 April, 10, 2002, UNC main Internet Link*



- UNC Internet Data Study Group
- Spikes, other possible anomalies

## Internet Traffic Data (continued)

- Time series collected in Internet
  - Self-similarity (Willinger et al, 1996)
  - Long range dependence (Leland et al, 1994)

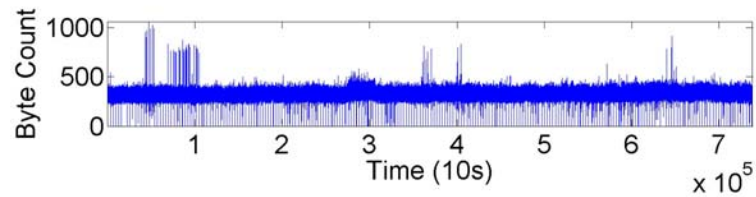
## Outliers in Time Series

- Outlier Types
  - Additive Outlier, Innovation Outlier, Level Shift, Variance Change, etc.
  - Fox (1972), Tsay (1988), etc.
- Previous Detection methods
  - Tsay (1988), Chang et al. (1988)
  - Robust Time Series Estimation

## Applicability of Previous Method to Internet traffic

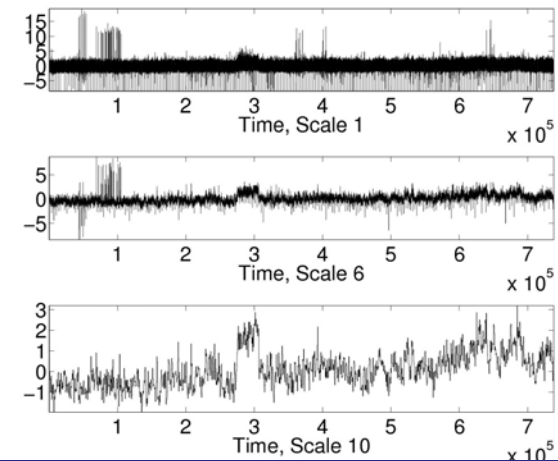
- Don't apply
  - Internet time series are **Long Range Dependent**
    - Classical time series - Short Range Dependent
  - Internet time series are more **bursty**
    - Natural artifacts of LRD might be misidentified
- **Multi-scale** property
  - Normal traffic has self-similarity
  - Anomalies exist in different time scales (Barford et al, 2002).

## Motivating example - One UNC bytecount trace

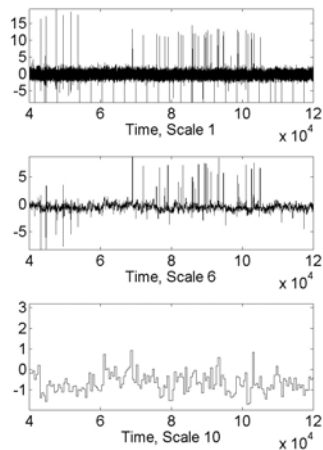


- $\gamma(k) \sim k^{2H-2}$ ,  $H$  is called **Hurst parameter**
- Self-similarity  $Z(at) \stackrel{d}{=} a^H Z(t)$
- Average estimation of Hurst parameter  $\hat{H} > .9$

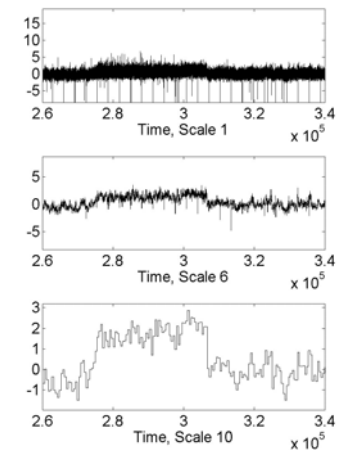
## Time series at different scales



## Time series at different scales



## Time series at different scales



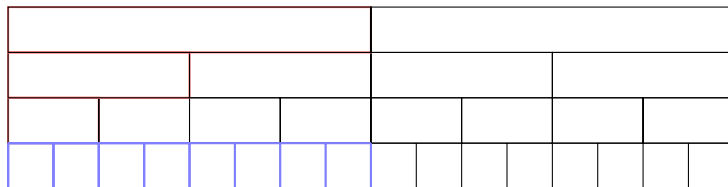
# Anomaly detection for the Internet

- Multiscale approach is important
  - View all scales
    - Kernel methods
    - Wavelet methods
    - Simple aggregation
- Multiple features
  - Different anomalies are described by different features, (Lakhina et al, 2004).
    - Port Scan : flow counts
    - ALPHA: Byte count or Packet count
  - Combination of features (Terrell and Zhang et al, 2005)
  - Simplest case : one feature
- Detection goal : close to real time.

# Simple aggregation methods

- Non-Overlapping Window Aggregation
- Sliding Window Aggregation

# Non-Overlapping Window Aggregation



- natural way to form time series of different scales
- good statistical feature for time series in different scales

# Non-Overlapping Window Aggregation

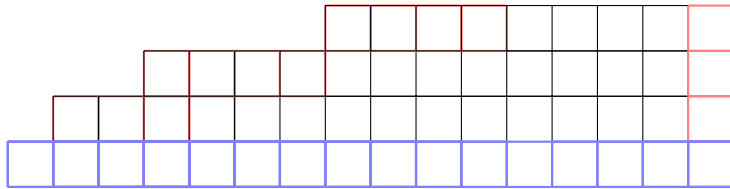
- Aggregate time scales at scale  $L$ :

$$Y_L(i) = \frac{1}{C} \sum_{j=1}^L Y_1((i-1)L + j)$$

i.e.,

$$\underbrace{Y_1(1), Y_1(2), Y_1(3), Y_1(4), Y_1(5), \dots}_{Y_2(1), Y_2(2), \dots}$$

# Sliding Window Aggregation



- one location, different scales
- only past information used.
- can be implemented online

# Sliding Window Aggregation

- Aggregate time scales at scale  $L$ :

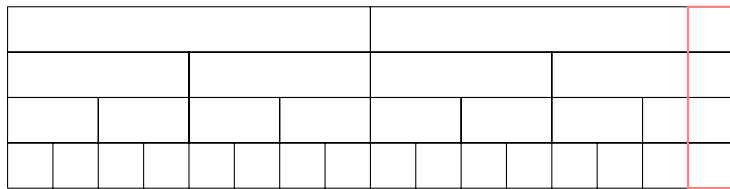
$$Y_L(i) = \frac{1}{C} \sum_{j=i-L+1}^i Y_1(j)$$

i.e.,

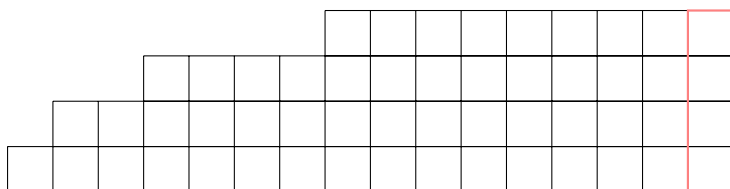
$$Y_1(i-L-1), Y_1(i-L), \underbrace{Y_1(i-L+1), \dots, Y_1(i-1), Y_1(i)}_{Y_L(i)}$$

# Common underlying statistical test

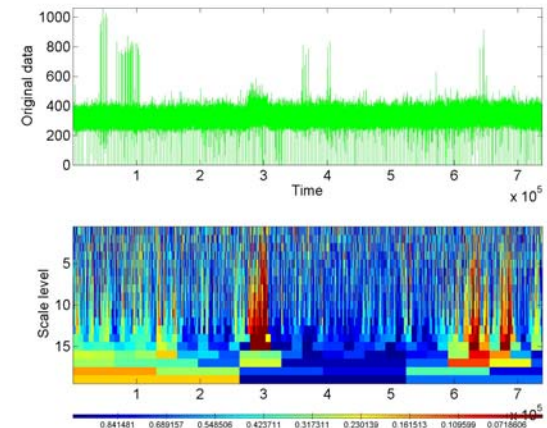
Non-overlapping window aggregation



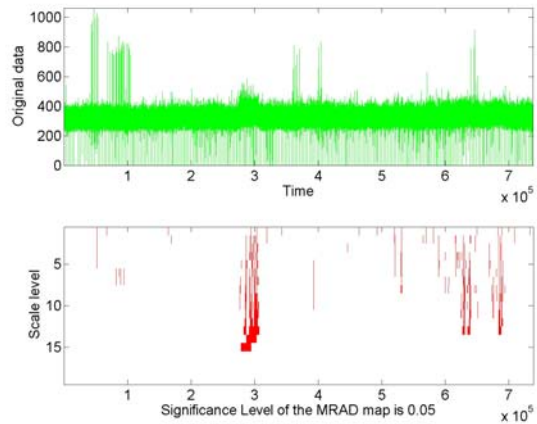
Sliding window aggregation



# MRAD outlier map



## MRAD threshold outlier map



## Formulating the problem

- Description of the problem
  - $Y_1(i)$  - the  $i$ th observation, finest scale,
  - $Y_L(i)$  - the corresponding observation at scale  $L$ .
  - Target: **whether  $Y_1(i)$  is an outlier or not**
  - Naive rejection region at Scale  $L$ :

$$R : |Y_L(i)| > C_\alpha$$

- MRAD rejection region

$$R : \max_L |Y_L(i)| > C_\alpha^M$$

## Expected theoretical properties

- More conservative
  - $C_\alpha^M \geq C_\alpha$ , at a given significance level  $\alpha$ .
  - fewer false alarms at a given scale
- Larger power on average over scales
  - Theorem proven: 2-scale MRAD procedure
  - Conjecture: more scales

## Theoretical properties

- Model setting
  - Let  $Y_1(i) = X_1(i) + \delta I_{i \in [a,b]}$
  - $X_1(i)$  - fractional Gaussian noise,

$$\gamma(k) = \frac{1}{2} \{ |k-1|^{2H} + |k+1|^{2H} - 2|k|^{2H} \}$$

Hurst parameter -  $H$ .

- Hypothesis testing
  - For the  $i$ -th time point,

$$H_0 : Y_1(i) = X_1(i), \text{ vs. } H_1 : Y_1(i) = X_1(i) + \delta$$

i.e., to test whether the  $i$ th observation has a mean level shift.

## The MRAD procedure

- Multi-scale time series
  - $\{Y_1(i)\}$  - the time series at the finest scale
  - $\{Y_L(i)\}$  - scale  $L$

$$Y_L(i) = \frac{\sum_{j=1}^L Y_1[(i-1)L+j]}{L^H}$$

- Rejection Region

$$R : \max_L |Y_L(i)| > C_\alpha^M$$

## MRAD is more conservative

### Proposition

If  $P_0(\cup_L \{|Y_L(i)| > C_\alpha^M\}) = \alpha$ , and  $P_0(|Y_L(i)| > C_\alpha) = \alpha$ , we have  $C_\alpha^M \geq C_\alpha$ .

### Proof.

Due to the fact that  $\{|Y_L(i)| > C_\alpha^M\} \subset \cup_L \{|Y_L(i)| > C_\alpha^M\}$ , we have

$$P\{|Y_L(i)| > C_\alpha^M\} \leq P\left(\cup_L \{|Y_L(i)| > C_\alpha^M\}\right) = P\{|Y_L(i)| > C_\alpha\},$$

which yields that  $C_\alpha \leq C_\alpha^M$ .  $\square$

## Two-scale MRAD procedure

### Proposition

For a 2-scale MRAD method, let  $C_\alpha^M$  be the testing threshold of significance level  $\alpha$ , we have

$$C_\alpha^M = C_0 - \frac{\phi(C_0)C_0^2 H^2}{2\sqrt{1-\alpha}} L^{2(H-1)} + o(L^{2(H-1)}),$$

as  $L \rightarrow \infty$ . Here  $C_0 = \Phi^{-1}\left(\frac{1+\sqrt{1-\alpha}}{2}\right)$ .

### Idea of Proof

Use Taylor series of bivariate normal distribution, as scale grows up.

## Two-scale MRAD procedure (continued)

### Theorem

Let

$$\begin{aligned} \beta_{(1,L)} &= P_1(\max_{l=1,L} |Y_l(i)| > C_\alpha^M), \\ \beta_1 &= P_1(|Y_1(i)| > C_\alpha), \\ \beta_L &= P_1(|Y_L(i)| > C_\alpha). \end{aligned}$$

For any  $\delta > 0$ , there exists  $\alpha_\delta > 0$  and  $L_\delta > 0$ , when  $\alpha \in (0, \alpha_\delta)$  and  $L > L_\delta$ , the following inequality holds:

$$\beta_{(1,L)} \geq \frac{\beta_1 + \beta_L}{2}.$$

## Two-scale MRAD procedure (continued)

### Idea of Proof

The power at scale 1 and L are given by

$$\beta_1 = 1 - [\Phi(C_\alpha - \delta) - \Phi(-C_\alpha - \delta)],$$

$$\beta_L = 1 - [\Phi(C_\alpha - \mu_L) - \Phi(-C_\alpha - \mu_L)].$$

where  $\mu_L$  is the mean under the alternative hypothesis. When  $K = (b - a)$  and  $\delta$  are fixed, let  $L \rightarrow \infty$ , we have  $\rho \rightarrow 0$ ,  $\mu_L \rightarrow 0$ , and

$$\beta_L = \alpha + O(L^{-2H}),$$

$$\beta_{(1,L)} = 1 - [\Phi(C_0 - \delta) - \Phi(-C_0 - \delta)]\sqrt{1 - \alpha} + O(L^{-1}),$$

## Two-scale MRAD procedure (continued)

### Idea of Proof (cont.)

Define  $f(\alpha, \delta)$  as the leading term of  $2\beta_{(1,L)} - (\beta_1 + \beta_L)$ . It can be proven that

$$\frac{\partial f(a, \delta)}{\partial a} \Big|_{a \rightarrow 0^+} > 0,$$

and  $f(\alpha, \delta) = 0$  when  $\alpha = 0$ .

The above leads to that, for any  $\delta > 0$ , there exists  $\alpha_\delta$ , such that for any  $\alpha \in (0, \alpha_\delta)$ , we have

$$f(\alpha, \delta) > 0,$$

hence the theorem holds.

## Asymptotic threshold for $m$ -scale MRAD

### Theorem

Let  $M_m = \max_L Y_L(i)$  for a  $m$ -scale MRAD procedure,  $M_m$  has a limiting distribution of the double exponential type

$$P\{a_m(M_m - b_m) \leq x\} \rightarrow \exp(-e^{-x}) \quad \text{as } m \rightarrow \infty,$$

with

$$a_m = \sqrt{2 \log m}, \quad \text{and} \quad b_m = a_m - \frac{\log \log m + \log 4\pi}{2a_m}.$$

## Asymptotic threshold for $m$ -scale MRAD (continued)

### ■ Remarks:

- $P(M_m \leq x) \rightarrow \Phi^m(x)$ , when  $m \rightarrow \infty$  (here  $\Phi(x)$  is the cdf of  $N(0, 1)$ ).
- significance level -  $\alpha$ ,  
Asymptotic threshold:  $C_0^M = \Phi^{-1}((1 - \alpha/2)^{1/m})$ .
- multiple testing in the scale space: Bonferroni test threshold:  
 $C_b = \Phi^{-1}(1 - \alpha/2m)$ .  
 $(1 - \alpha/2)^{1/m} = 1 - \alpha/2m - (m - 1)\alpha^2/(8m^2) + o(\alpha^2)$   
 $\implies C_0^M \leq C_b$ . i.e., **larger power than the Bonferroni procedure.**

## Asymptotic threshold for $m$ -scale MRAD (continued)

### Idea of Proof

- $\{Y_L(i)\} (L = 1, \dots, m)$  is also a stationary process
- The autocovariance function

$$\begin{aligned} \gamma_h(k) &= \text{Cov}(Y_1(i), Y_{k+1}(i)) \\ &= \dots \\ &= H[h^{(1-H)}]^{-k} + \frac{h^{-kH}}{2} - \frac{H(2H-1)}{2} h^{k(H-2)} + o(h^{k(H-2)}). \end{aligned}$$

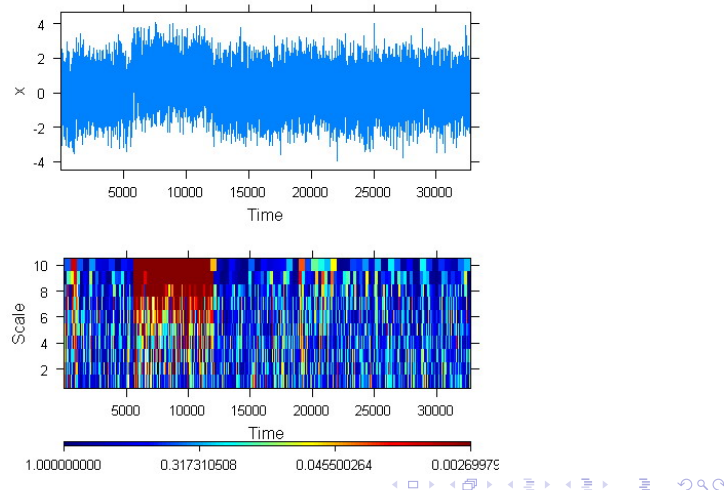
acf decays exponentially.

- Use the results in Berman (1964) (see also Leadbetter et al. (1983)) to prove this theorem.

## Simulation - fGn + local mean level shift

- Background - fGn
  - $H = .9, n = 2^{15}$
- Starting point of the level shift
  - $a \sim U(0, 2^{14})$
  - This simulation  $a = 5644$ .
- Duration of the level shift
  - $K \sim \exp(\lambda)$ , where  $\lambda = 4000$  ( $\approx 12\%$ )
  - This simulation  $K = 6465$ .
- 100 different realizations.

## The MRAD map for one realization



## Simulation evaluation of the MRAD method

- True Outlier Proportion
- False Discovery Rate
- False Negative Rate
- Total Discovery Rate
- Outlier Proportion

$a$	$K$	$\delta$	TOP	FDR	FNR	TDR	OP
5644	6465	1	19.73%	5.36%	0.08%	94.64%	20.80%

Table: Summary of simulation evaluation over 100 realizations

## More Evaluations

$a$	$K$	$\delta$	TOP	FDR	FNR	TDR	OP
10223	11835	1	36.11%	1.26%	0.26%	98.74%	36.41%
1946	3760	1	11.47%	6.26%	0.24%	93.74%	12.03%
9572	4407	1	13.45%	6.47%	0.25%	93.53%	14.17%
3173	4491	1	13.71%	6.35%	0.03%	93.65%	14.62%

Table: More simulation evaluations

## Future work

- 1 Real trace evaluation
- 2 Robust estimation of  $H$  and normalization.
- 3 Other background setting
- 4 Other types of outliers
- 5 Other way to form multi-scale time series
- 6 Multiple Comparison in the time domain
- 7 Spatial extreme value theory

## Summary

- MRAD method and MRAD outlier map
- Theoretical properties
- Empirical evaluation of our method

Thanks!

## Other work in my thesis

- Functional Singular Value Decomposition
- New visualization methods for PCA and SVD
- Different types of mean for a data matrix
- Comparisons of different types of SVD.
- Applications
  - Internet traffic data set
  - Chemometrics data set
  - Demographic data set
- see Zhang et al (2006) JCGS accepted. Won 2005 Student paper award.