“An Evaluation of Effect of Packet Sampling on Anomaly Detection Method”

Takuya Motodate

April 25, 2010
The 3rd CAIDA-WIDE-CASFI Joint Measurement Workshop @Osaka

Background

- Anomaly Detection: Signature-based,Statistical one
- Statistical anomaly detection assumes
  a full-captured dump.
- Traffic of backbone network become broader,
  so, characteristics of it is grasped
  with sampled traffic.

- We have to use sampled-traffic
  as input of anomaly detection.
  What should we do?

Problem Statement

1. Suitable Packet-Sampling Method is not Known.
2. Suitable Anomaly Detection Method is not Known.

   Because of inadequate evaluations.

Purpose

- Evaluate an effect to result of anomaly detection methods with various sampling methods and common traffic data.
  - Sketch and Non Gaussian Multi-Resolution Statistical Detection Procedures as a Anomaly Detection Method.
  - 5 Packet-Sampling Methods.
  - MAWI Dataset as Traffic Data.

Sketch and Non Gaussian Multi-Resolution Statistical Detection Procedures [Dewaele et al. 07]

1.Divide a traffic into some subtraffics.

2. Estimate α and β of each subtraffic, each timescale.

3. Anomalous subtraffic has deviate α or β.

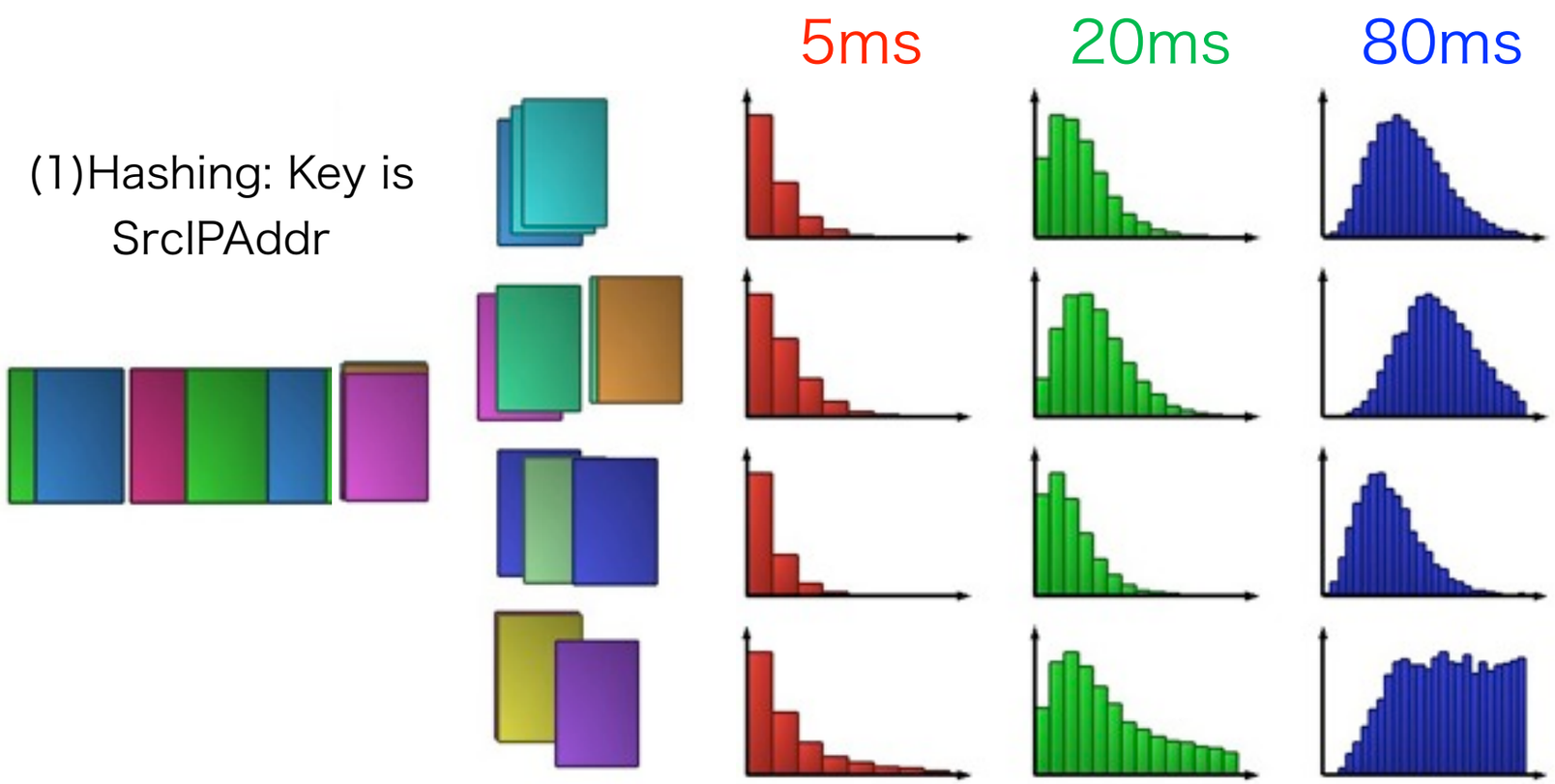
Sketch and Non Gaussian Multi-Resolution Statistical Detection Procedures [Dewaele et al. 07]

1.Divide a traffic into some subtraffics.

2. Estimate α and β of each subtraffic, each timescale.

3. Anomalous subtraffic has deviate α or β.

Sketch and Non Gaussian Multi-Resolution Statistical Detection Procedures [Dewaele et al. 07]

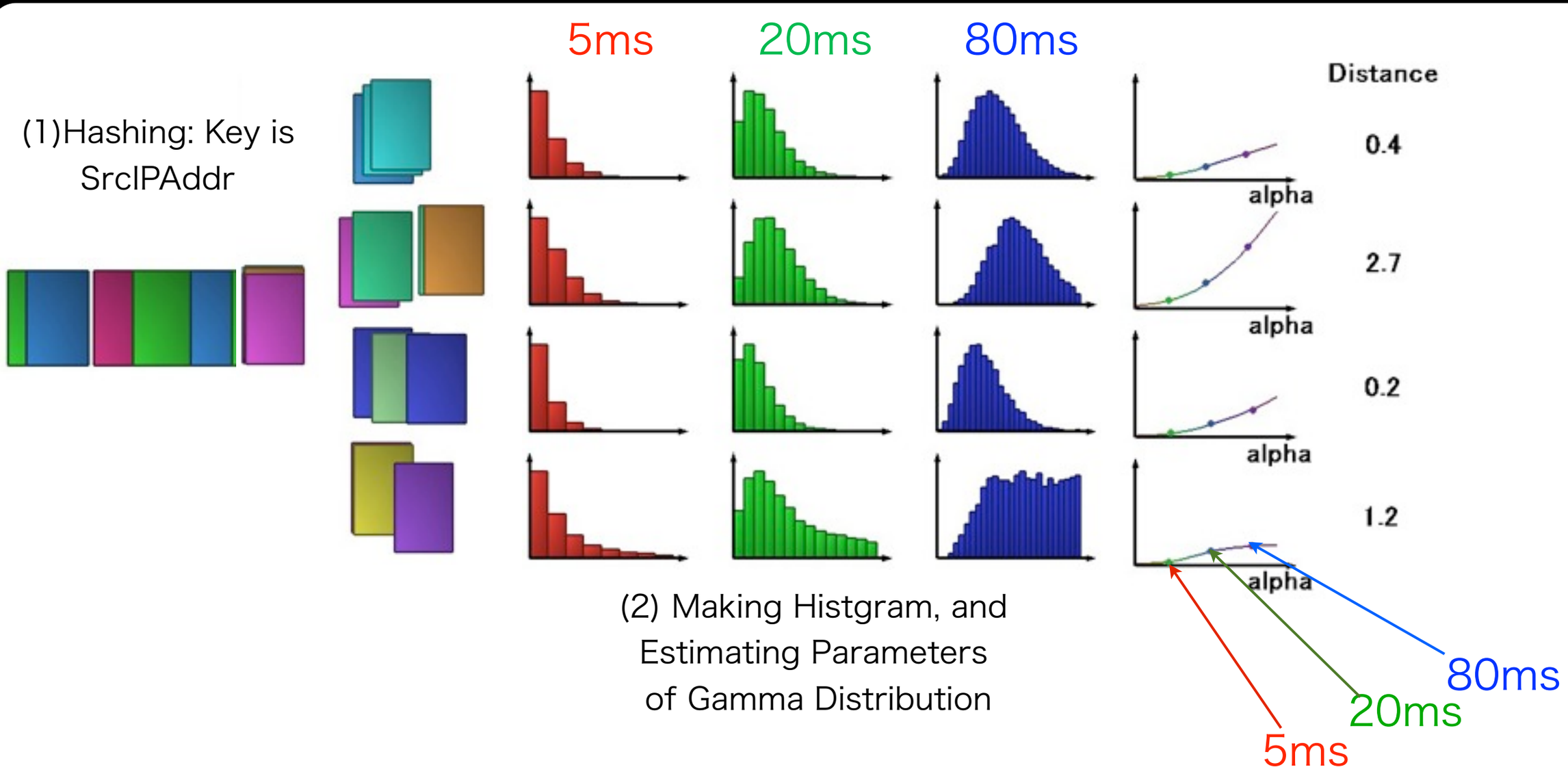


1.Divide a traffic into some subtraffics.

2. Estimate α and β of each subtraffic, each timescale.

3. Anomalous subtraffic has deviate α or β.

Sketch and Non Gaussian Multi-Resolution Statistical Detection Procedures [Dewaele et al. 07]

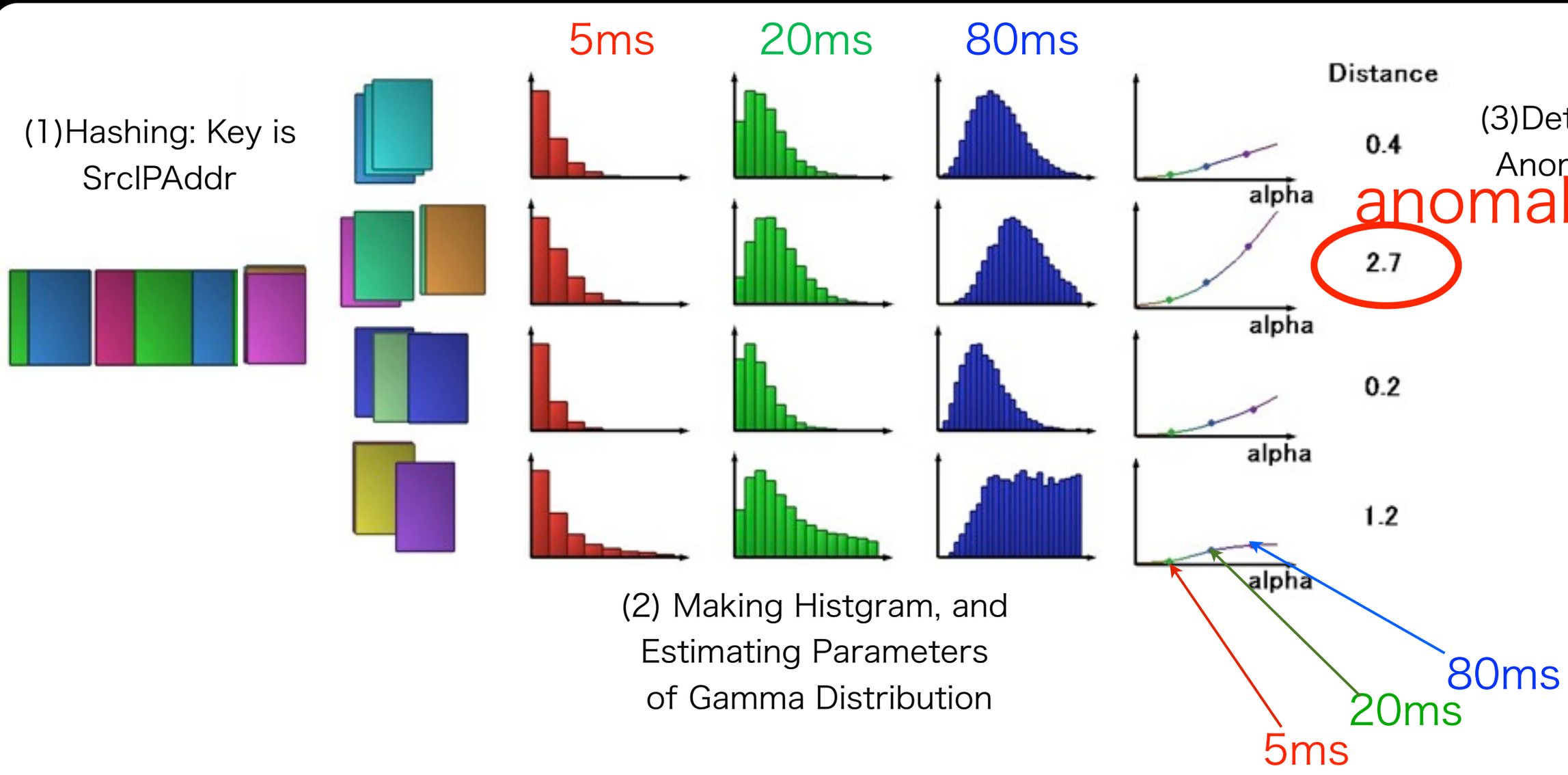


1.Divide a traffic into some subtraffics.

2. Estimate α and β of each subtraffic, each timescale.

3. Anomalous subtraffic has deviate α or β.

Sketch and Non Gaussian Multi-Resolution Statistical Detection Procedures [Dewaele et al. 07]

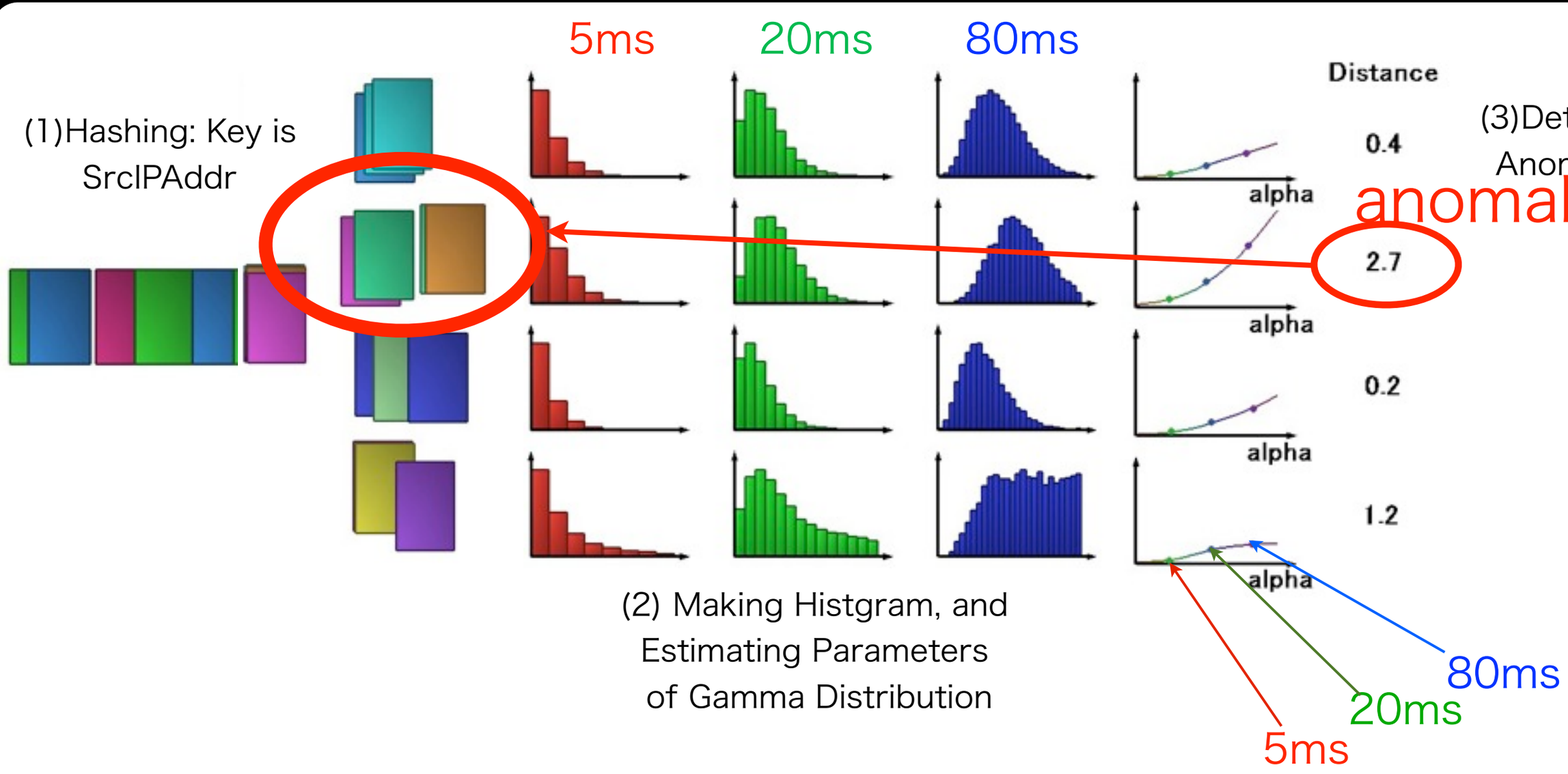


1.Divide a traffic into some subtraffics.

2. Estimate α and β of each subtraffic, each timescale.

3. Anomalous subtraffic has deviate α or β.

Sketch and Non Gaussian Multi-Resolution Statistical Detection Procedures [Dewaele et al. 07]



1.Divide a traffic into some subtraffics.

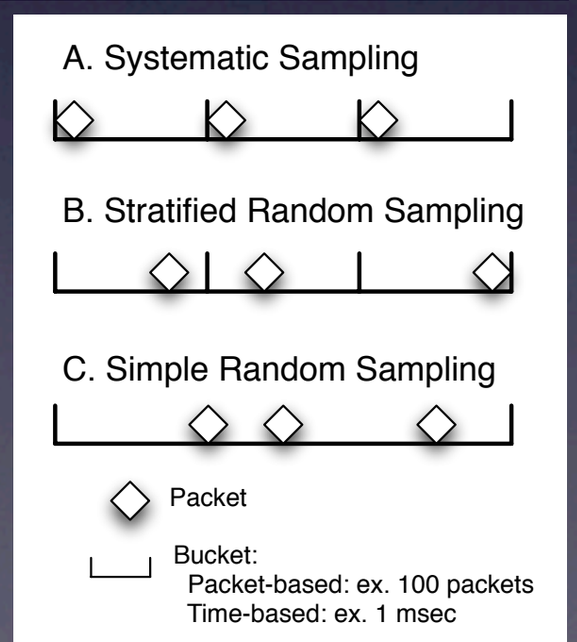
2. Estimate α and β of each subtraffic, each timescale.

3. Anomalous subtraffic has deviate α or β.

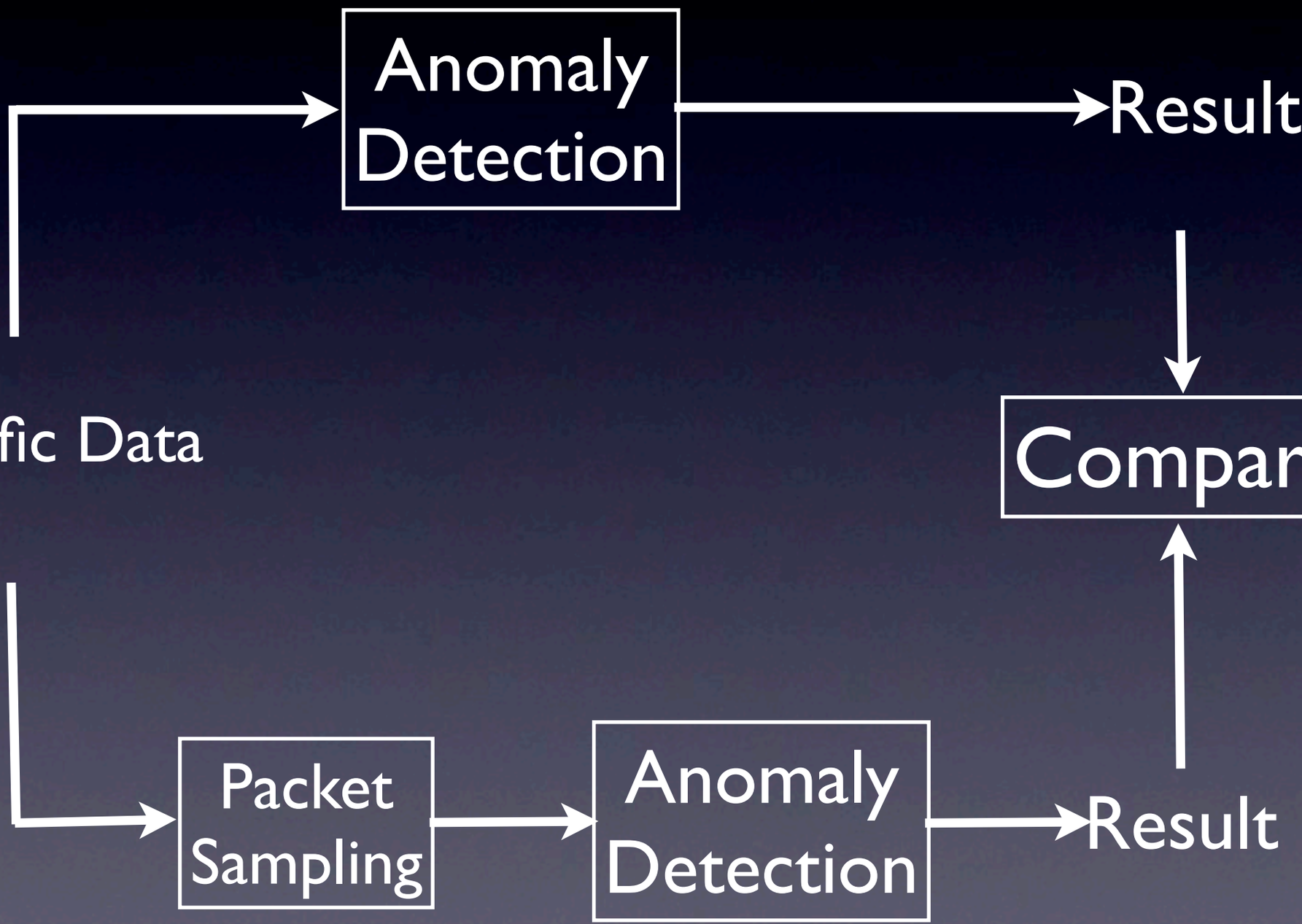
Packet-Sampling Methodologies
[Claffy et al. 93]

	Systematic	Stratified Random	Simple Random
Packet-based	Packet-based Systematic	Packet-based Stratified Random	Simple Random
Time-based	Time-based Systematic	Time-based Stratified Random	

Packet-based :
  Picking up a packet per N packets
Time-based :
  Picking up a packet per M msec



Overview of Evaluation

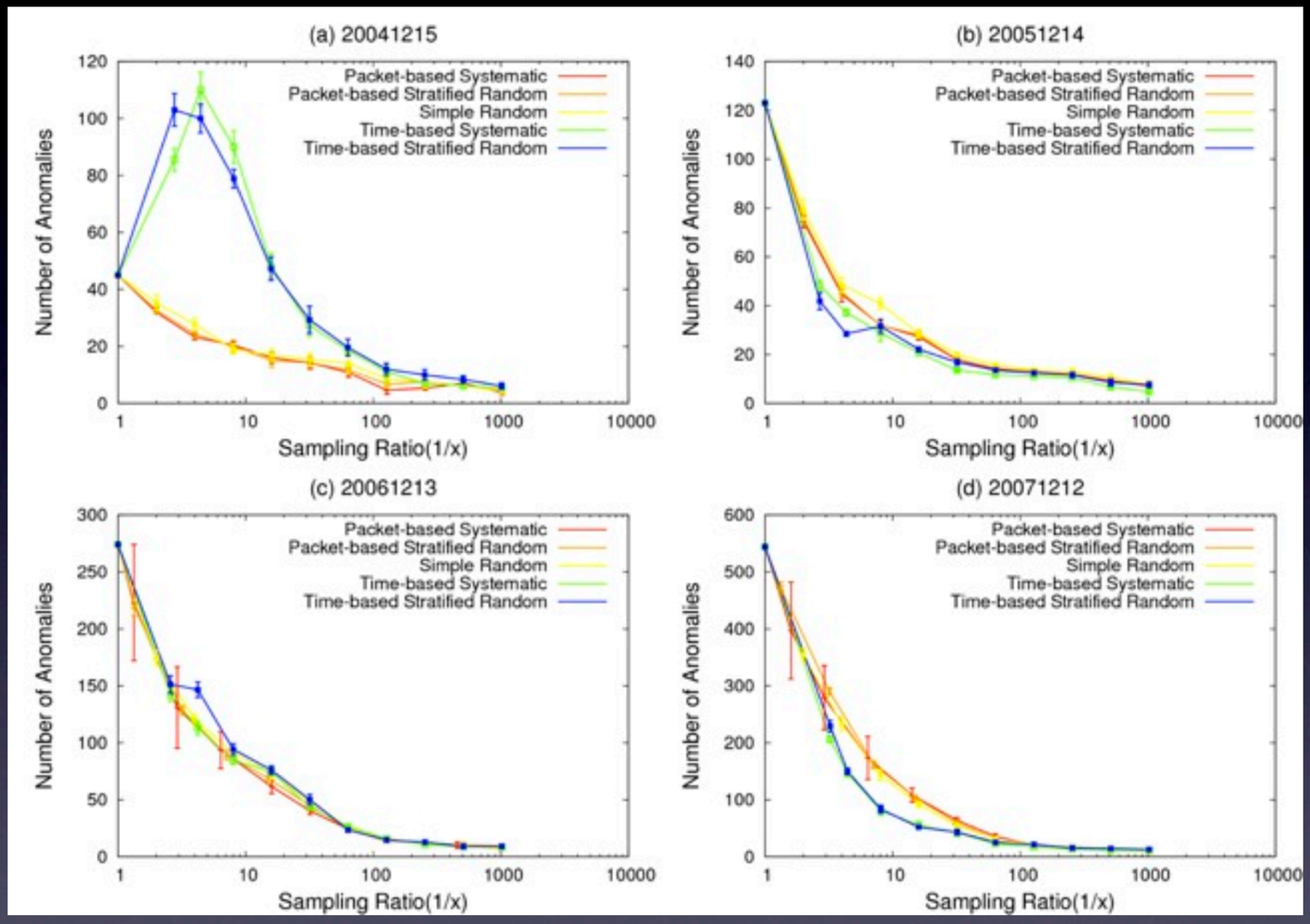


Evaluation

- I apply this evaluation to MAWI Traffic Data at 4days. - A Wednesday in December from 2004 to 2007, sample-Point B or F.

Dec 15, 2004	Dec 14, 2005
Dec 13, 2006	Dec 12, 2007

Numbers of Detected Hosts with each Sampling-Rate

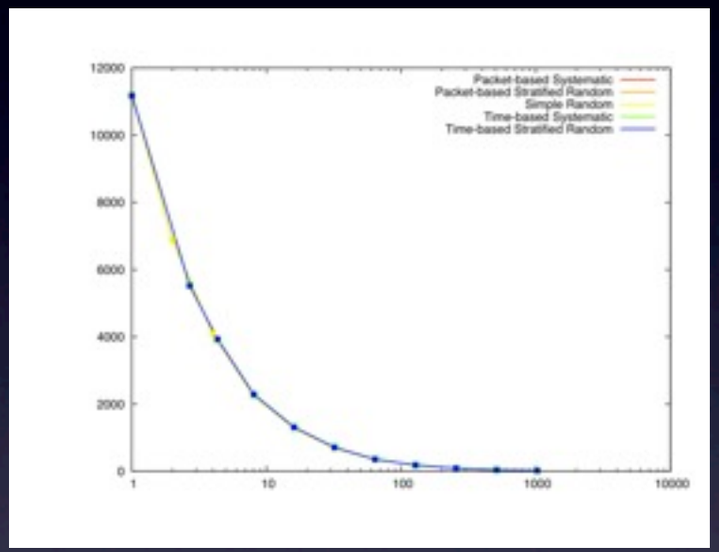


Brief Observation:

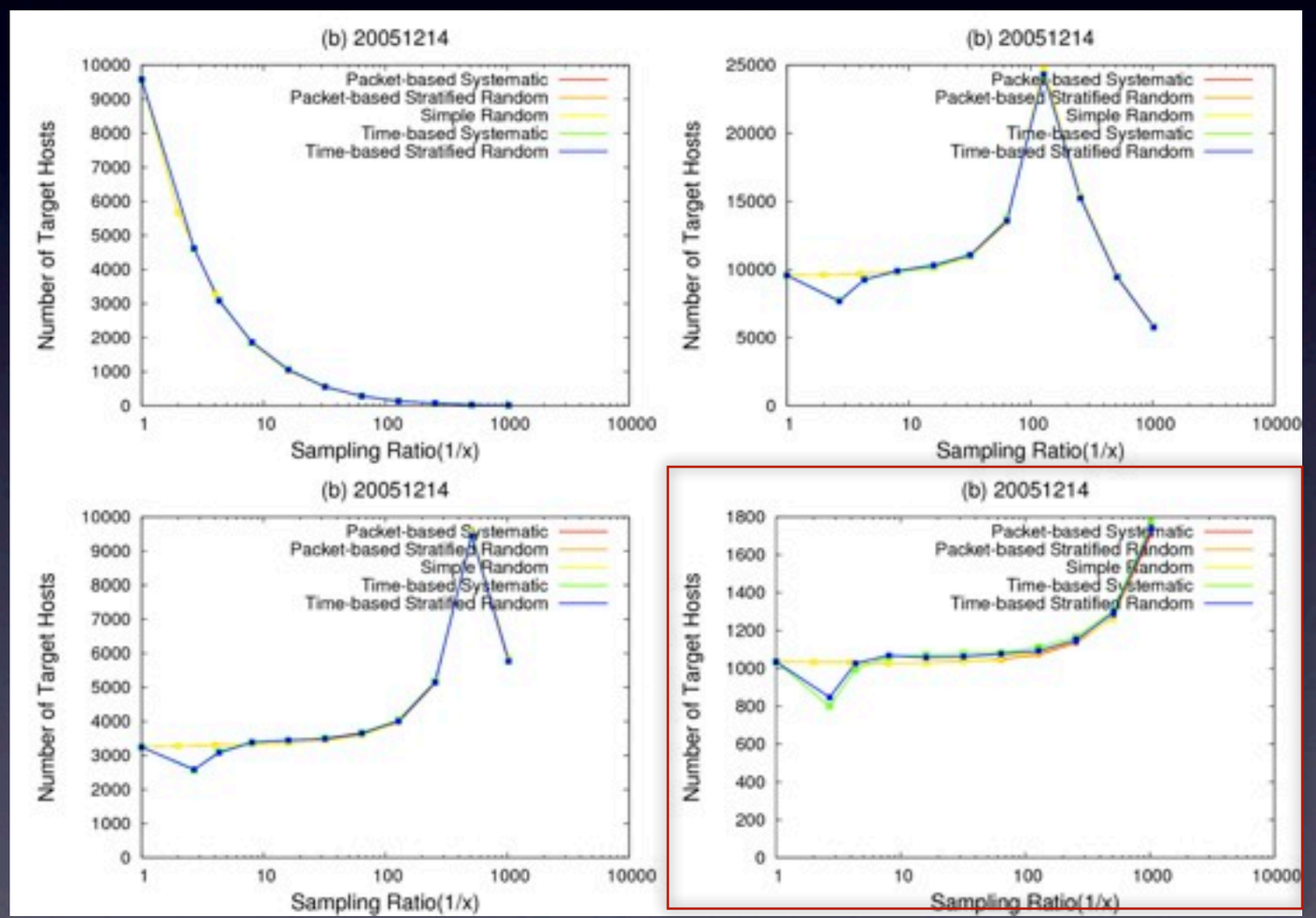
1. Detected hosts decreased as sampling-rate decreased.
2. Rapid increase is observed 2004 with time-based sampling.

Parameter Tuning

Target Hosts

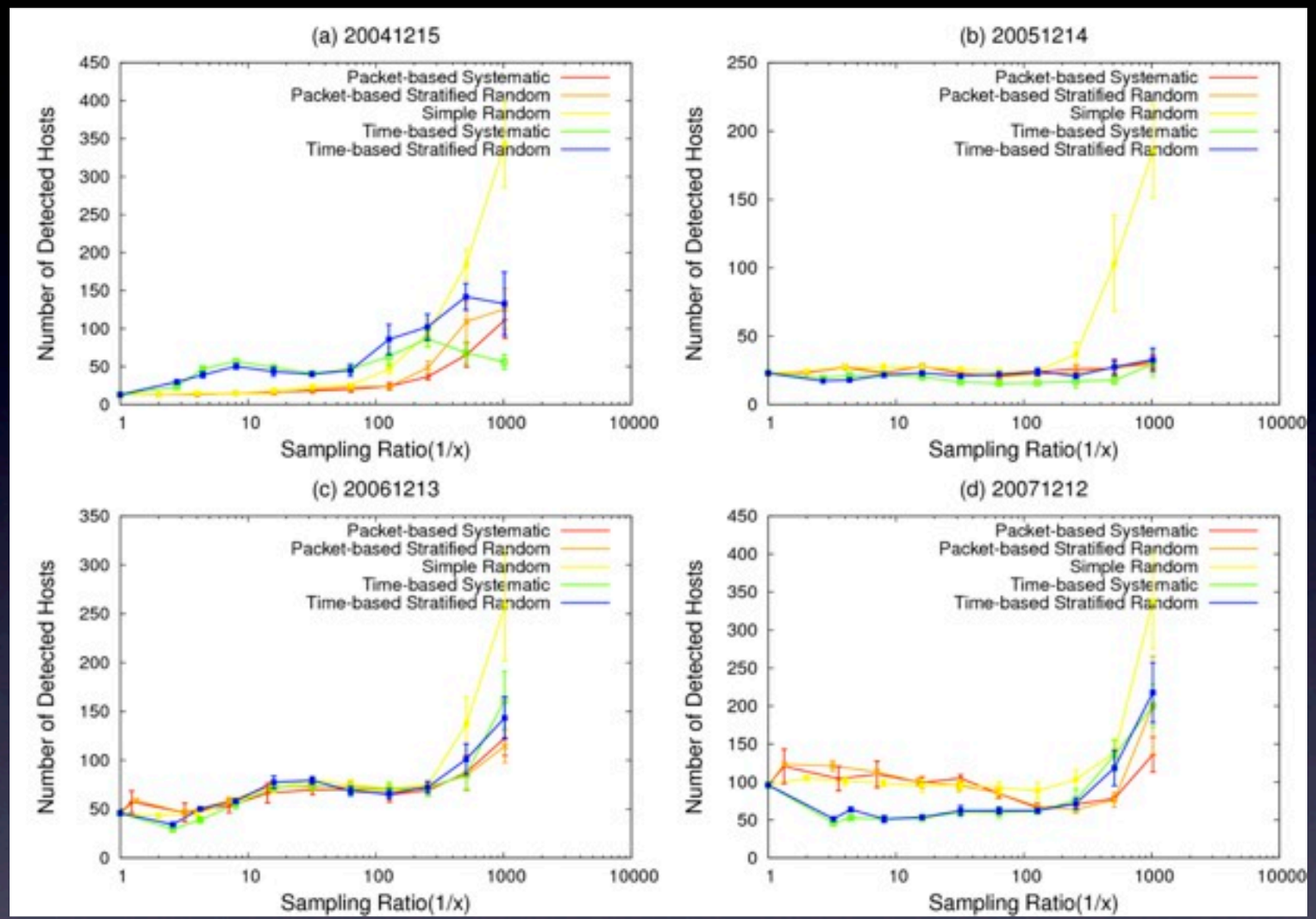


Target Hosts after Parameter Tuning



Trying to make target hosts fixed.

Numbers of Detected Hosts with each Sampling-Rate after normalization



Brief Observation:

1. Different behavior between packet-based and time-based in high sampling-rate
2. Rapid Increase number of Simple-Random in low sampling-rate.

Undergoing Things

- Analysis a reason rapid increase of anomalies with simple-random in low sampling-rate, and difference between result with time-based and packet-based.
- Cross-Validation: with Port-based Categorization.
- Comparison with another Anomaly Detection Method.

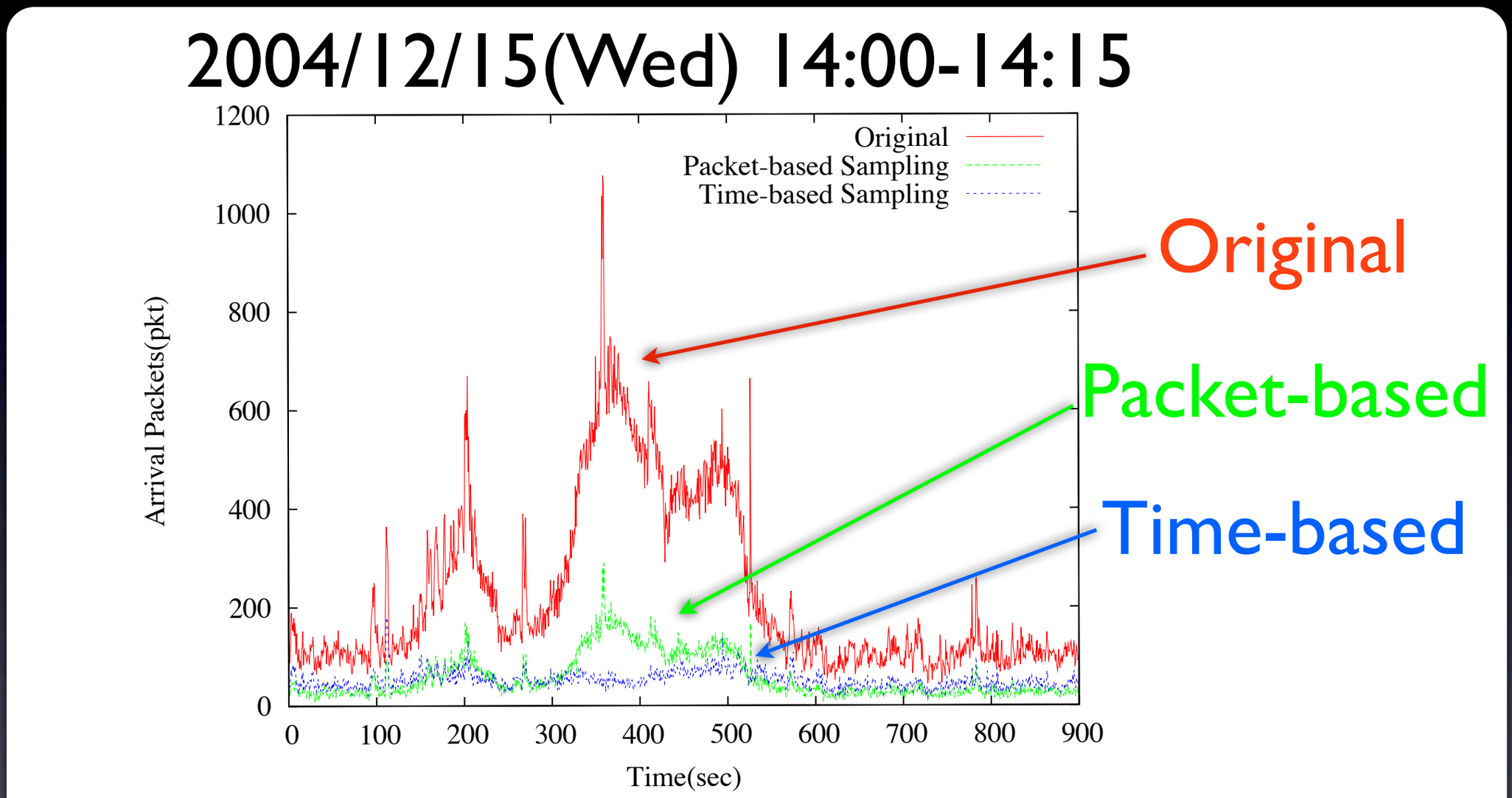
Summary

- Necessary for an evaluation in using anomaly detection with sampled-traffic.
- Evaluating a “Sketch and Non Gaussian Multi-Resolution” with 5 sampling methods.
- Performance Difference between Time-based and Packet-based sampling, simple-random sampling.

Fin.

Thank you for Listening.

Distribution of a number of arrival packets



Pakcet-based Systematic : 1/4 pkt/pkt
Time-based Systematic: 1/4.4 pkt/pkt