Revisiting the Threshold Random Walk Scan Detector

Vagishwari Nagaonkar
Dr.John Mchugh
Faculty of Computer Science
Dalhousie University

Presented for FLOCON 2008

DALHOUSIE UNIVERSITY
Inspiring Minds

Introduction

- Initial Activity in many intrusions
  - Scanning
- Techniques to detect these initial scans
- One of the effective algorithms
  - Threshold Random Walk

Introduction (contd.)

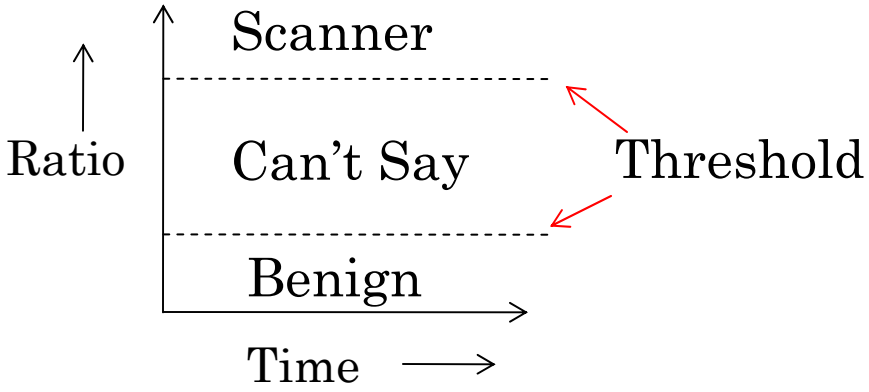
- Challenges when using TRW
  - UDP and ICMP Traffic
  - Repetitive Scanning
  - Slow and Stealthy Scans
- Using Bloom filters
  - eliminate repetitive input to TRW
  - look for reverse matches in time ordered data

Threshold Random Walk

- Scan Detection Algorithm based on sequential hypothesis testing.

- Uses a positive reward based scan detection.
  - For a given host, records connection attempt made :

Connection	Ratio
Successful	Decreases
Failed	Increases



Threshold Random Walk

- The ratio is calculated as :

$$\Lambda(Y) \equiv \frac{\Pr[Y|H_1]}{\Pr[Y|H_0]} = \Pi_{i=1}^{n} \frac{\Pr[Y_i|H_1]}{\Pr[Y_i|H_0]}$$

- Where the probabilities are :

$$\Pr[Y_i = 0|H_0] = \theta_0, \quad \Pr[Y_i = 1|H_0] = 1 - \theta_0$$
$$\Pr[Y_i = 0|H_1] = \theta_1, \quad \Pr[Y_i = 1|H_1] = 1 - \theta_1$$

  - **Y = success (0) or failed (1) connection attempt**
  - **H0 = benign hypothesis**
  - **H1 = scanner hypothesis**
  - **Θ0 = probability that the source is benign, for a successful connection attempt**
  - **Θ1 = probability that the source is scanner for a successful connection attempt**

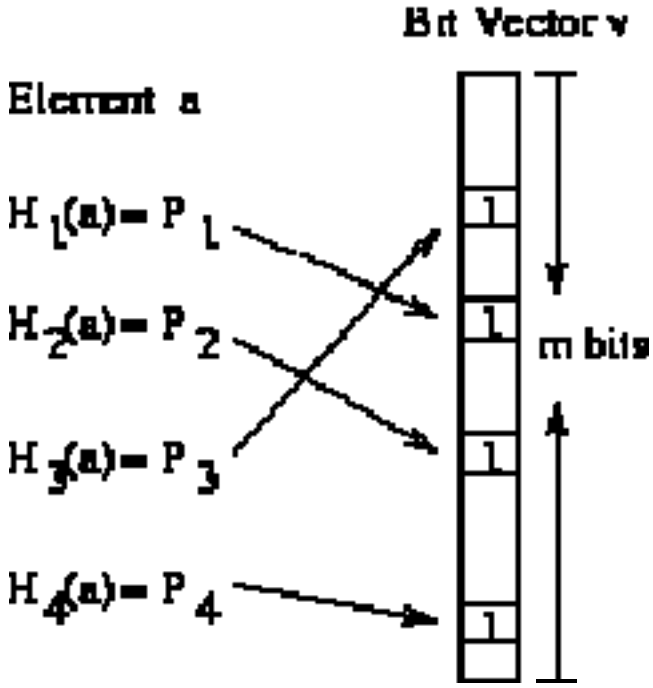
Threshold Random Walk

- The thresholds are calculated based on
  - desired true positive (β = 0.99)
  - desired false positive (α = 0.01)

$$\eta_1 \leftarrow \frac{\beta}{\alpha} \qquad \eta_0 \leftarrow \frac{1-\beta}{1-\alpha}$$

Bloom Filter

- It's a Data Structure
  - test the membership of an element for a given set
- Definition of the Structure
  - bit array of m bits
  - k different hash functions
  - Hash functions maps a key value to one of the m array positions.



Bloom Filter

- Properties :
  - False positives possible
  - No false negatives
  - Elements can be added
  - No deletion possible
  - Greater the number of elements, higher the probability of false positives.
  - Space Efficient
  - Cannot determine the elements present in it.

Modified TRW with Bloom Filter

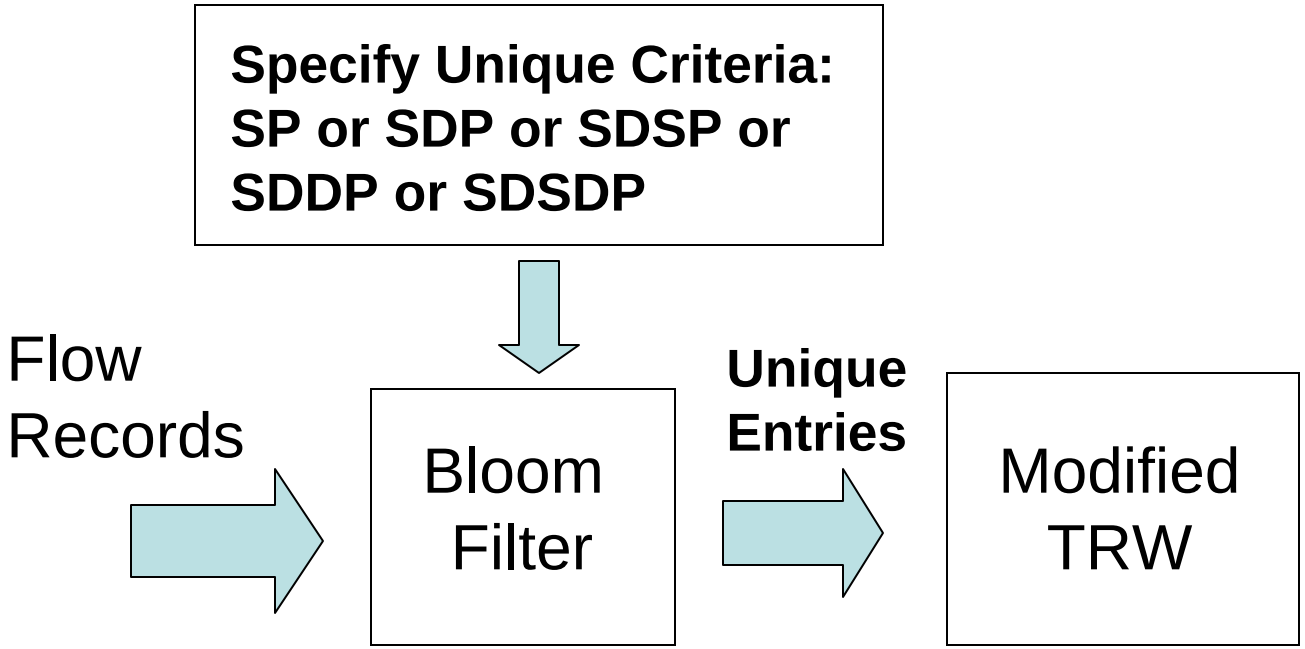
- TRW hit or miss definition
  - For a given pair in the flow record
    eg {sip, dip}
    - HIT = if a corresponding entry {dip, sip, sport, dport, proto} is found within a specified timeout period
    - MISS = if a corresponding entry {dip, sip, sport, dport, proto} is not found within a specified timeout period

Modified TRW with Bloom Filter

- Bloom Filter uses 10 hash functions and a bit vector of size 2^32

- Experiment Set up :
  - Pass the flow records through the bloom filter.
  - Specify selection criteria: {sip, dip}, {sip, dip, proto}, {sip, dip, sport}, {sip, dip, dport}, {sip, dip, sport, dport, proto}
  - Use the TRW scanning algorithm.

Modified TRW with Bloom Filter



DALHOUSIE UNIVERSITY
Inspiring Minds

The Dataset

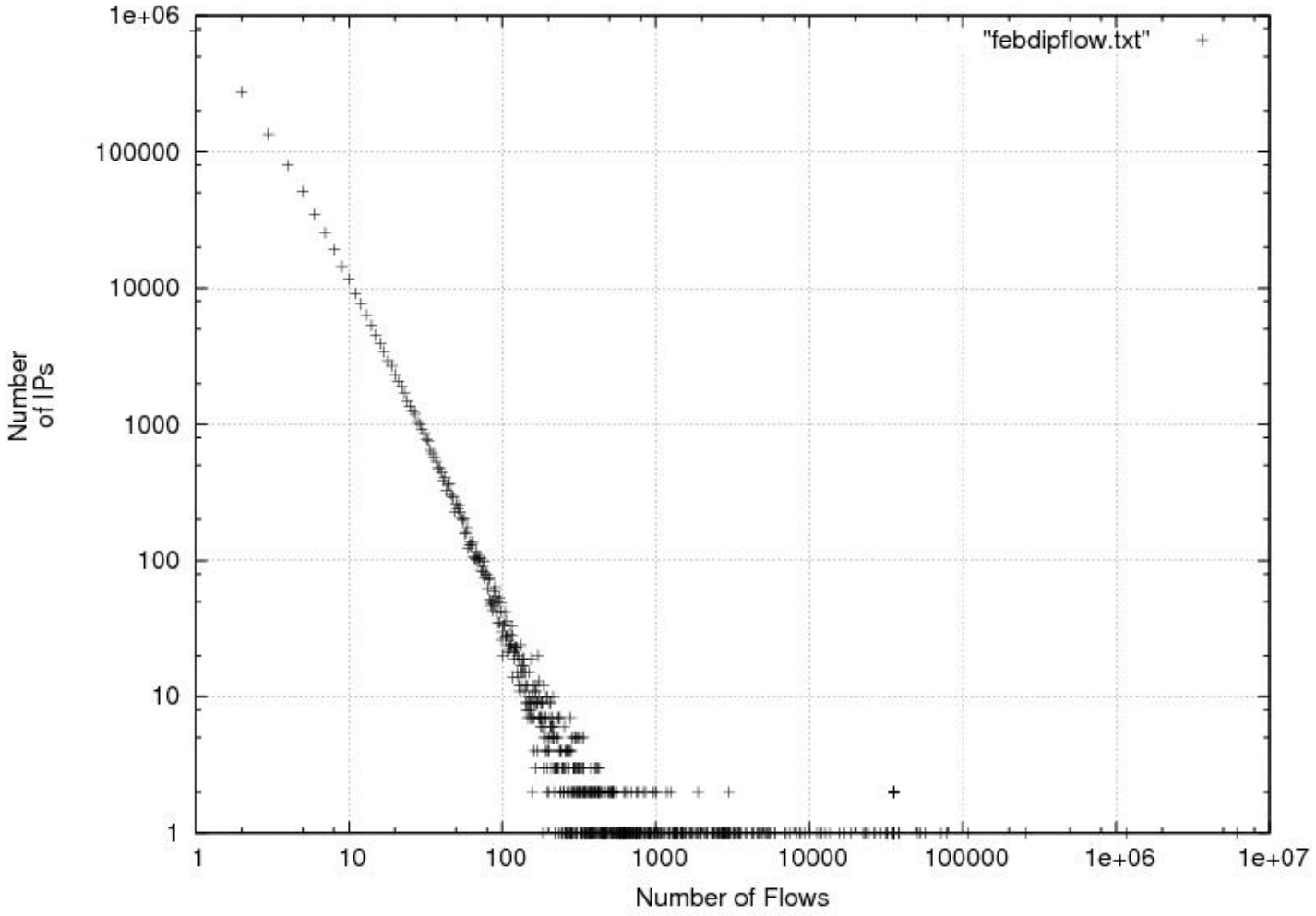
- A year long trace collected on a /22 enterprise network
- Using Silk Tools
- Internal Network Hosts
  - Total Address Space = 1024
  - #Active hosts in a given day = varies between 60-70
  - Active Address Space ~ 6%

The Dataset

	OutIps Seen			
	EtoO	**OtoE**	**Non Responsive Out ips**	**% Non Responsive Out ips**
Feb	26680	7270	19410	72.75112444
Mar	30232	3866	26366	87.21222546
Apr	56126	14576	41550	74.02986138
May	2355612	106893	2248719	95.46219836
June	2847371	283270	2564101	90.05152472
July	2601834	246312	2355522	90.53313932
Aug	30181	29097	1084	3.591663629
Sept	126913	126549	364	0.28681065
Oct	330740	277438	53302	16.11598234
Nov	4050	2932	1118	27.60493827
Dec	2226535	254484	1972051	88.57040199
Total	**10636274**	**1352687**	**9283587**	**87.28232274**

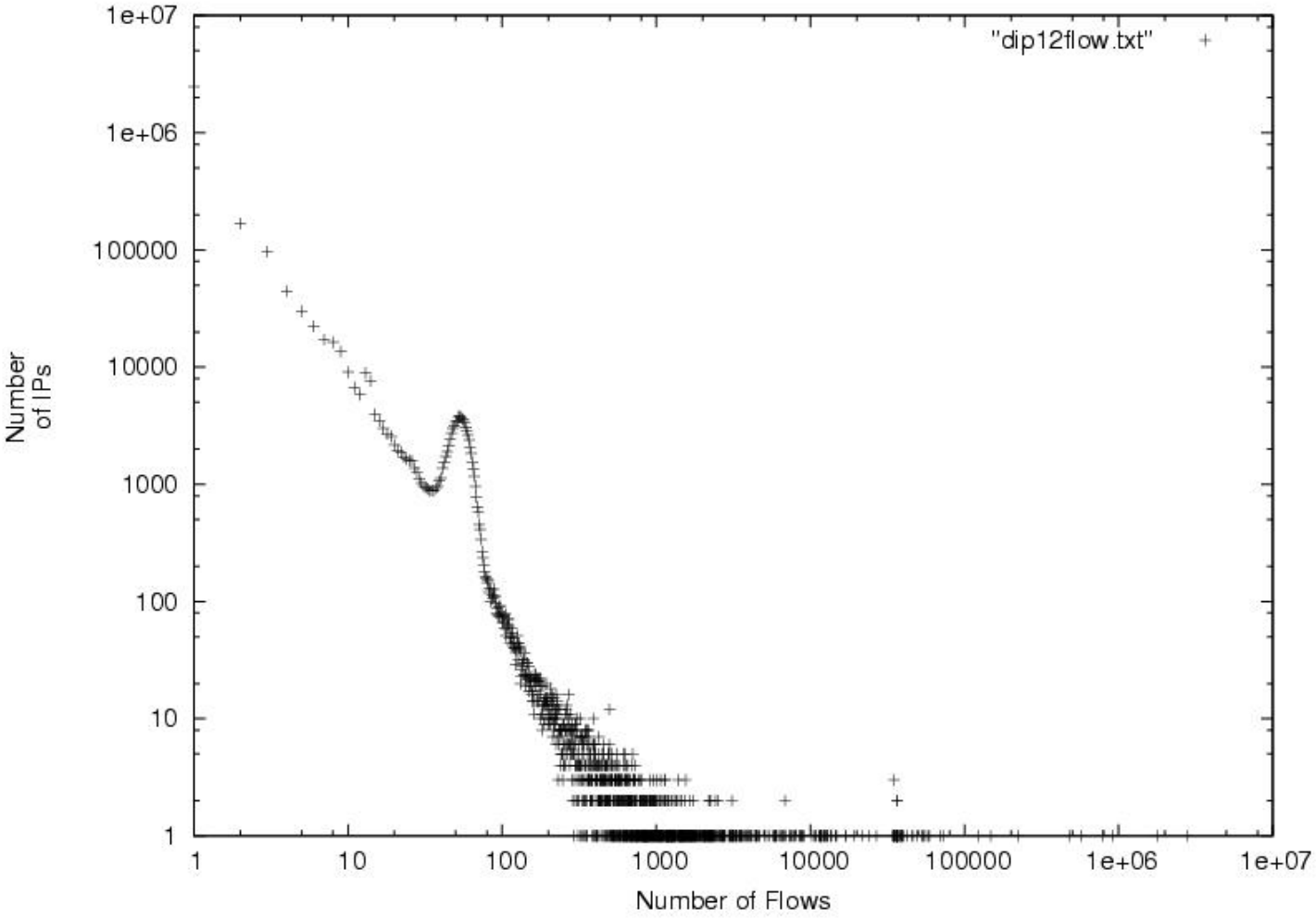
DALHOUSIE UNIVERSITY
Inspiring Minds

The Dataset



DALHOUSIE UNIVERSITY
Inspiring Minds

The Dataset

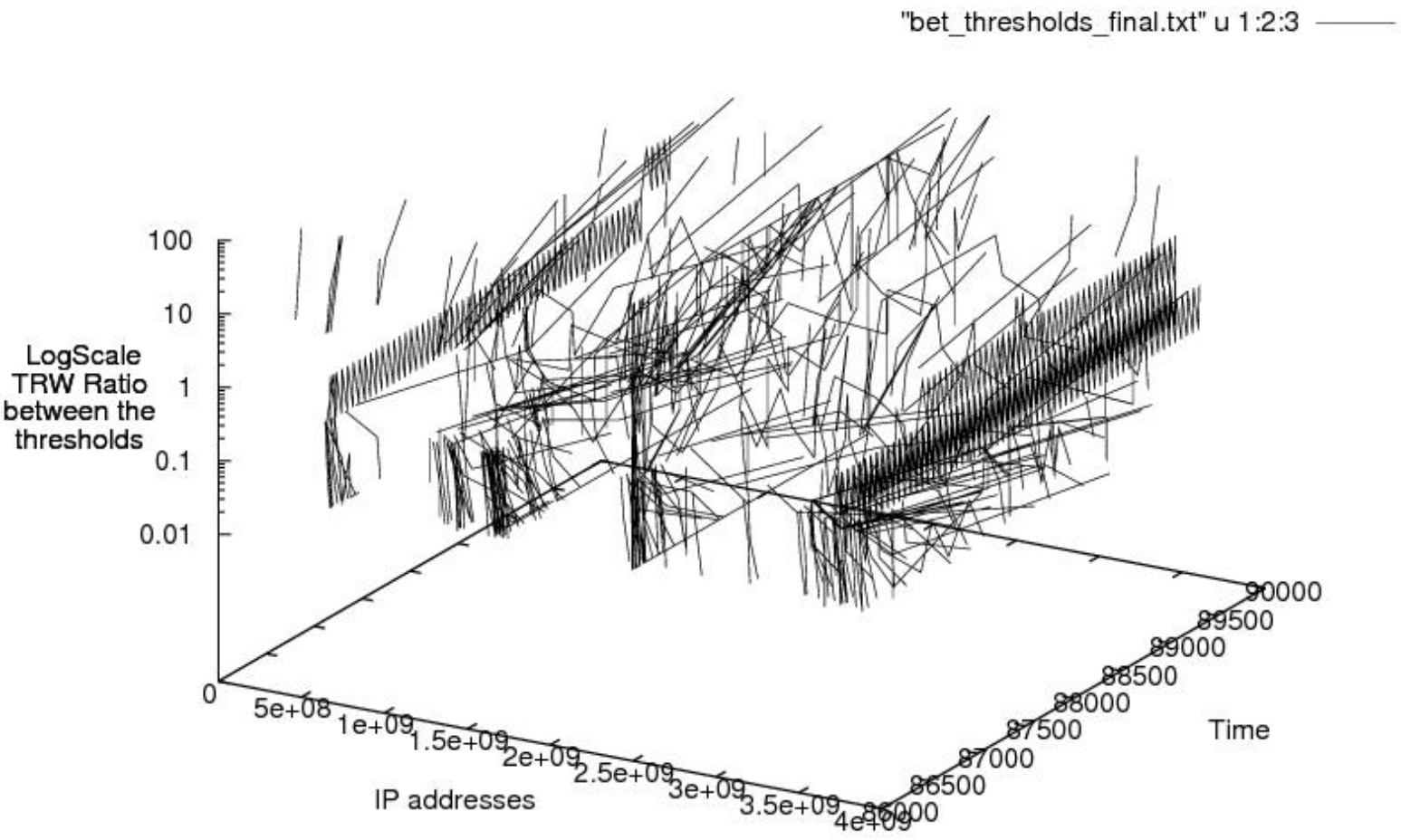


DALHOUSIE UNIVERSITY
Inspiring Minds

Problems faced during Analysis

- Time granularity
  - millisecond not available.
  - The order of flow records for the same second is the outside to inside put first.
- Background noise in the traffic.
- ICMP ping traffic causes false detection.

Problems faced during Analysis

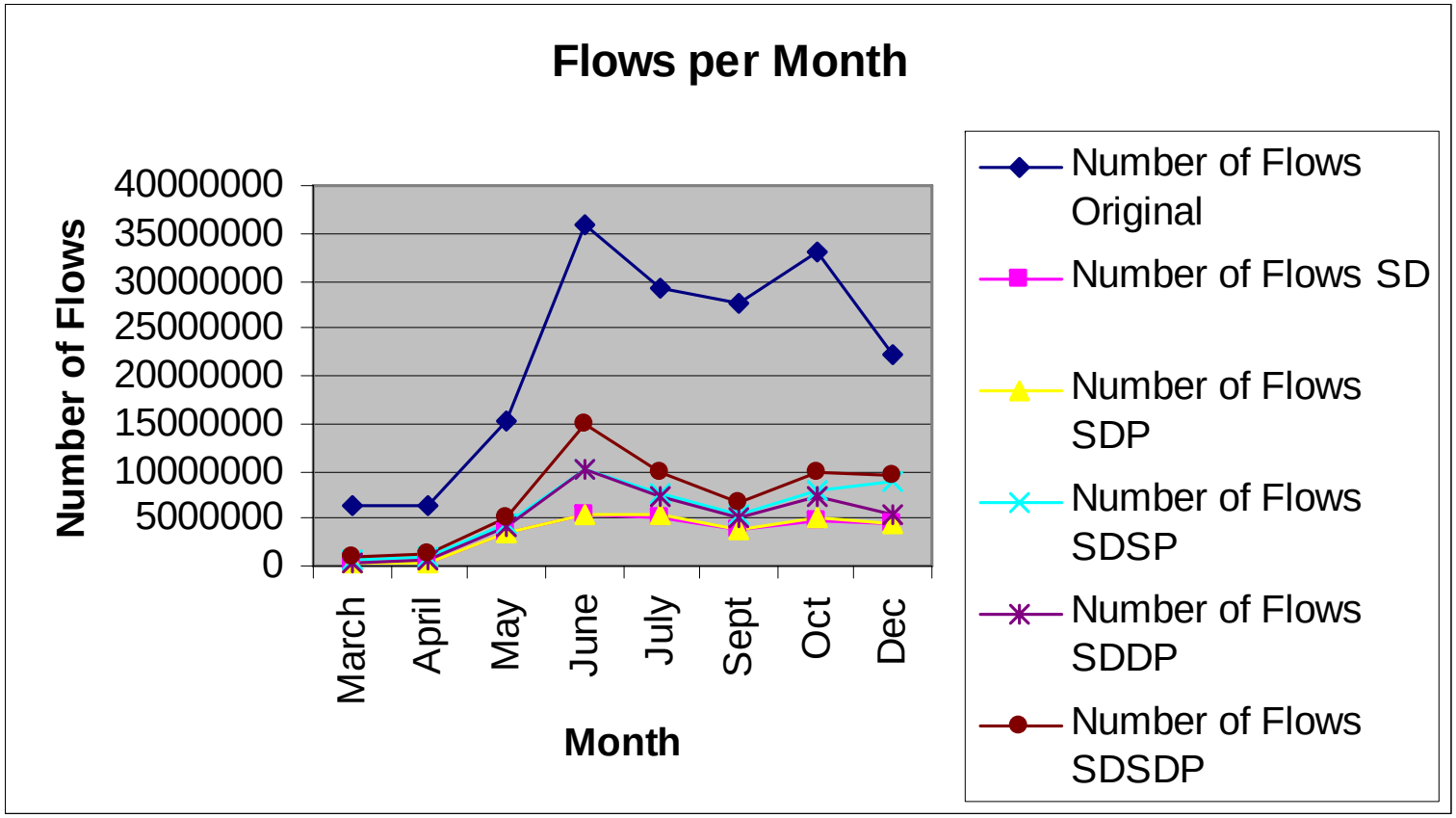


DALHOUSIE UNIVERSITY
Inspiring Minds

Preliminary Results

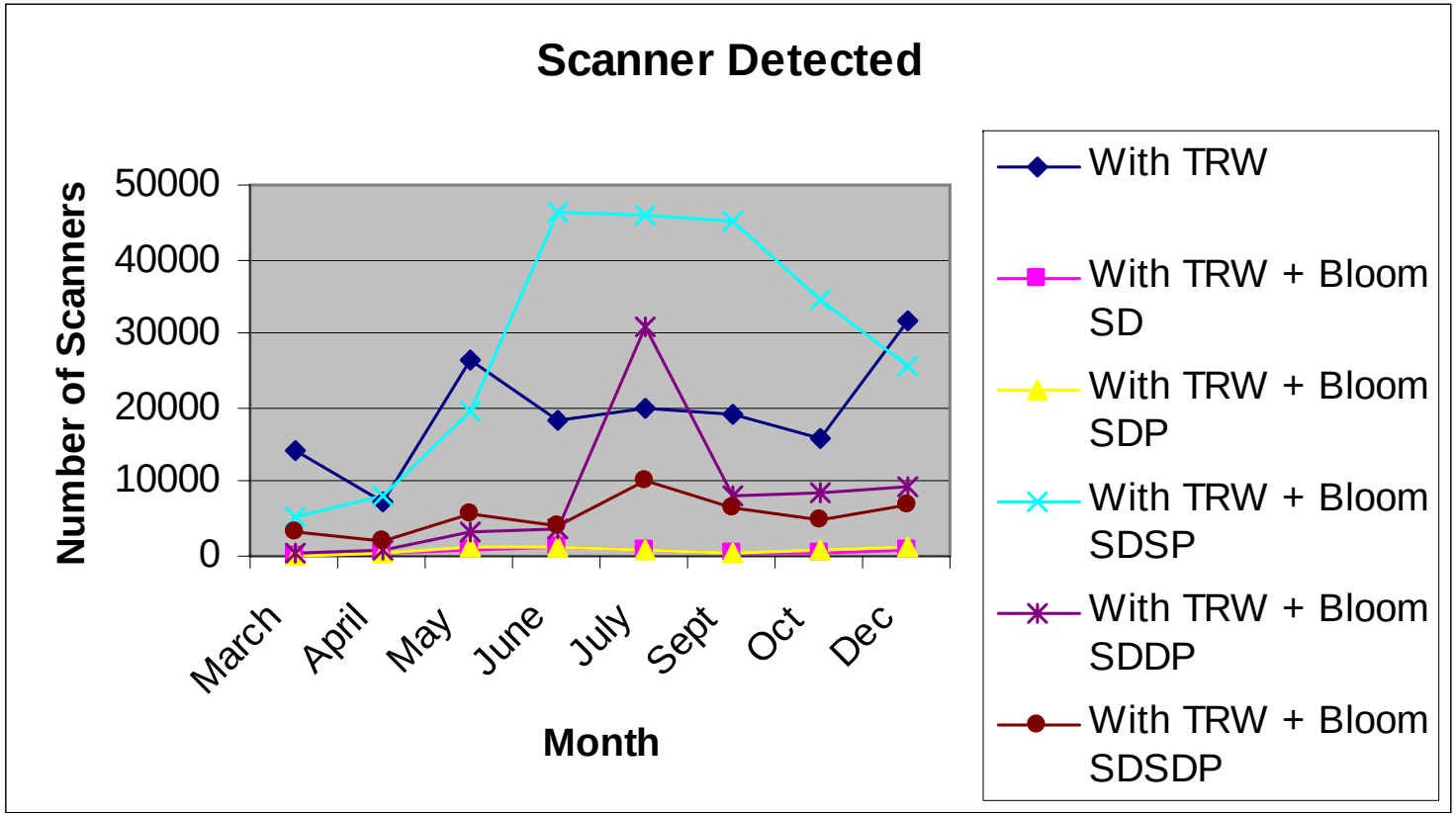
- TRW Parameters used:
  - Theta1 determined based on the %active internal hosts compared to the total address space ~ 0.0654
  - Theta0 ~ 0.8
    - Changed theta0 for benign hosts to hits / (hits + miss)
    - The value of new theta0 ranged from 0.45 to 1.00
    - All benign hosts still classified as benign
  - Alpha (desired false positive) = 0.01
  - Beta (desired true positive) = 0.99

Preliminary Results



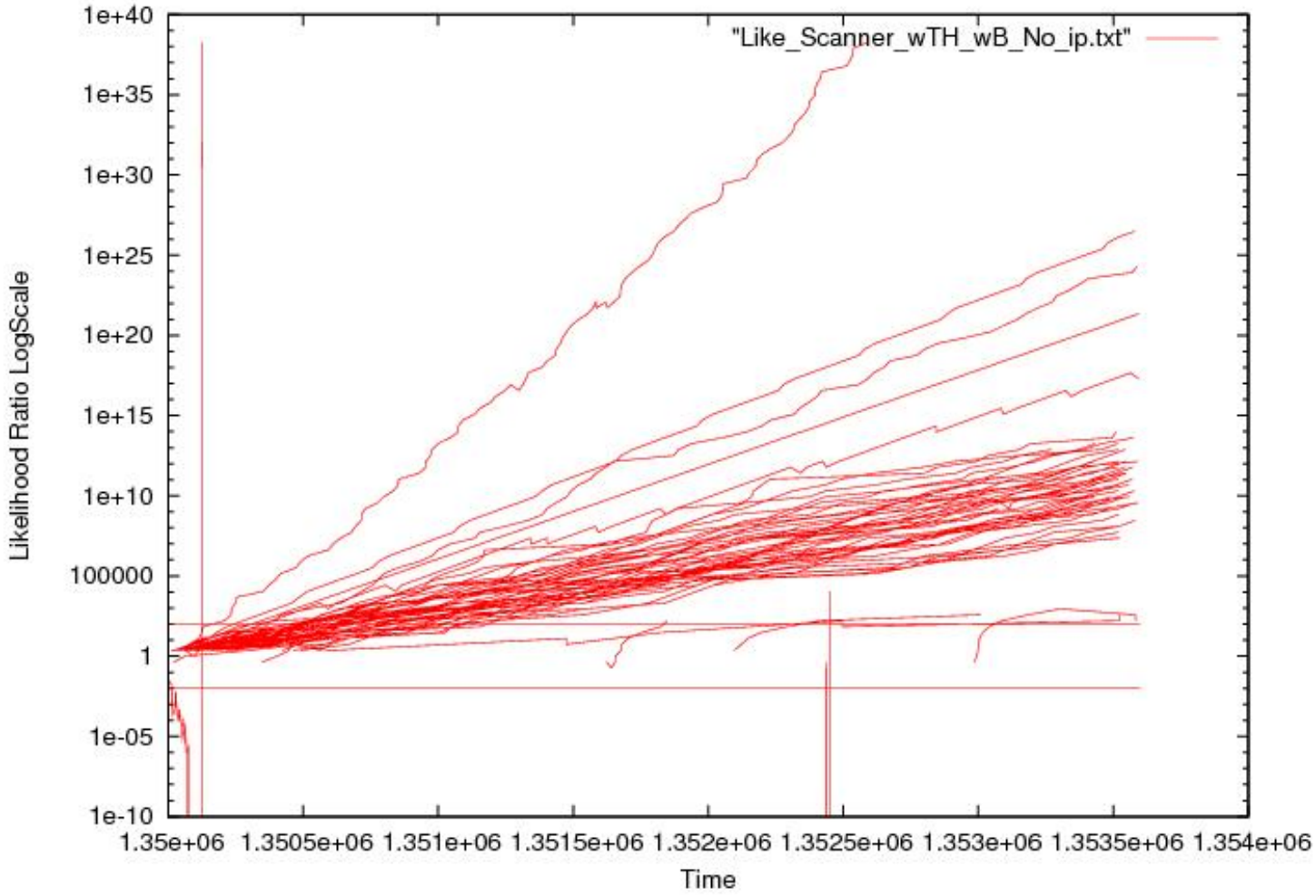
DALHOUSIE UNIVERSITY
Inspiring Minds

Preliminary Results



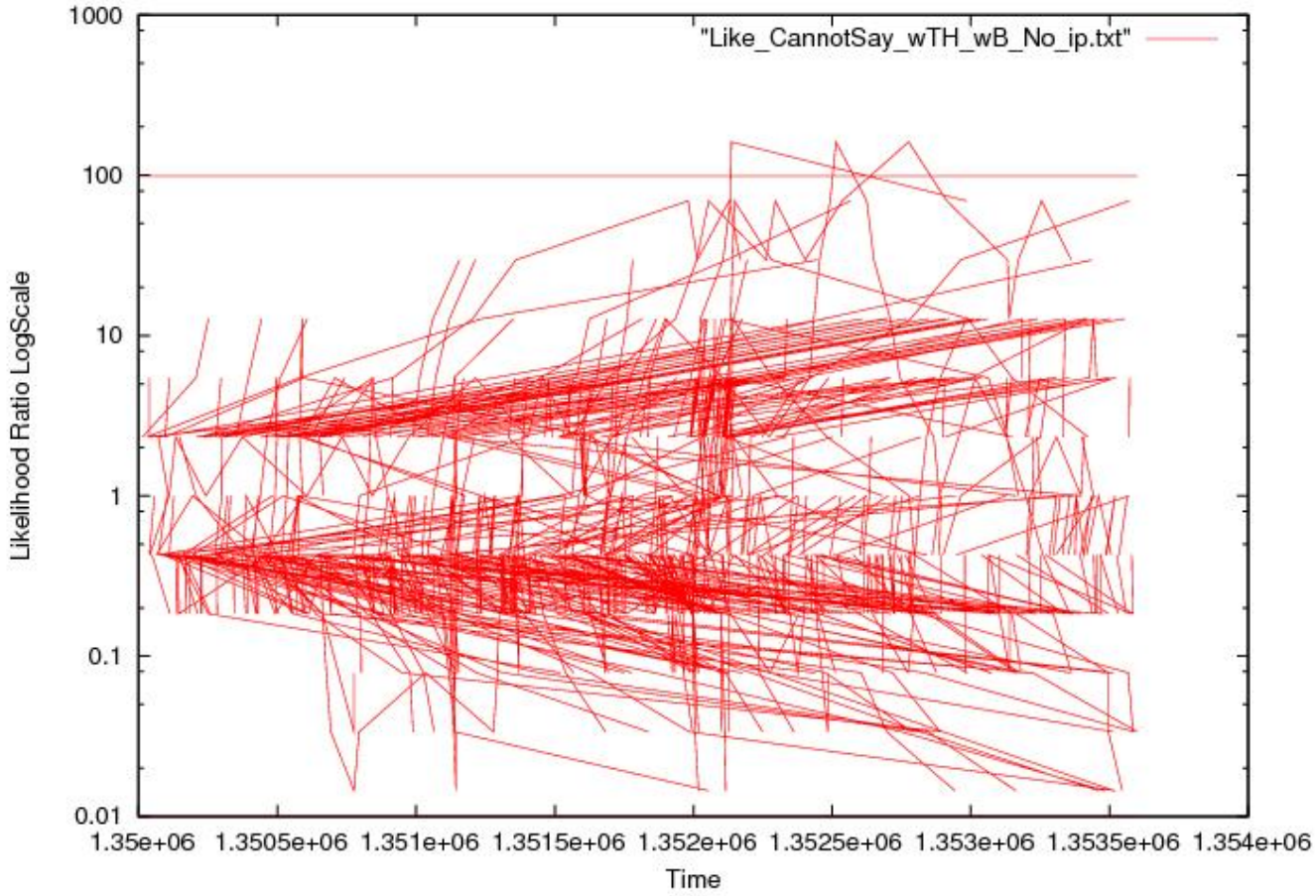
DALHOUSIE UNIVERSITY
Inspiring Minds

Preliminary Results
Plot of Likelihood ration for Scanners



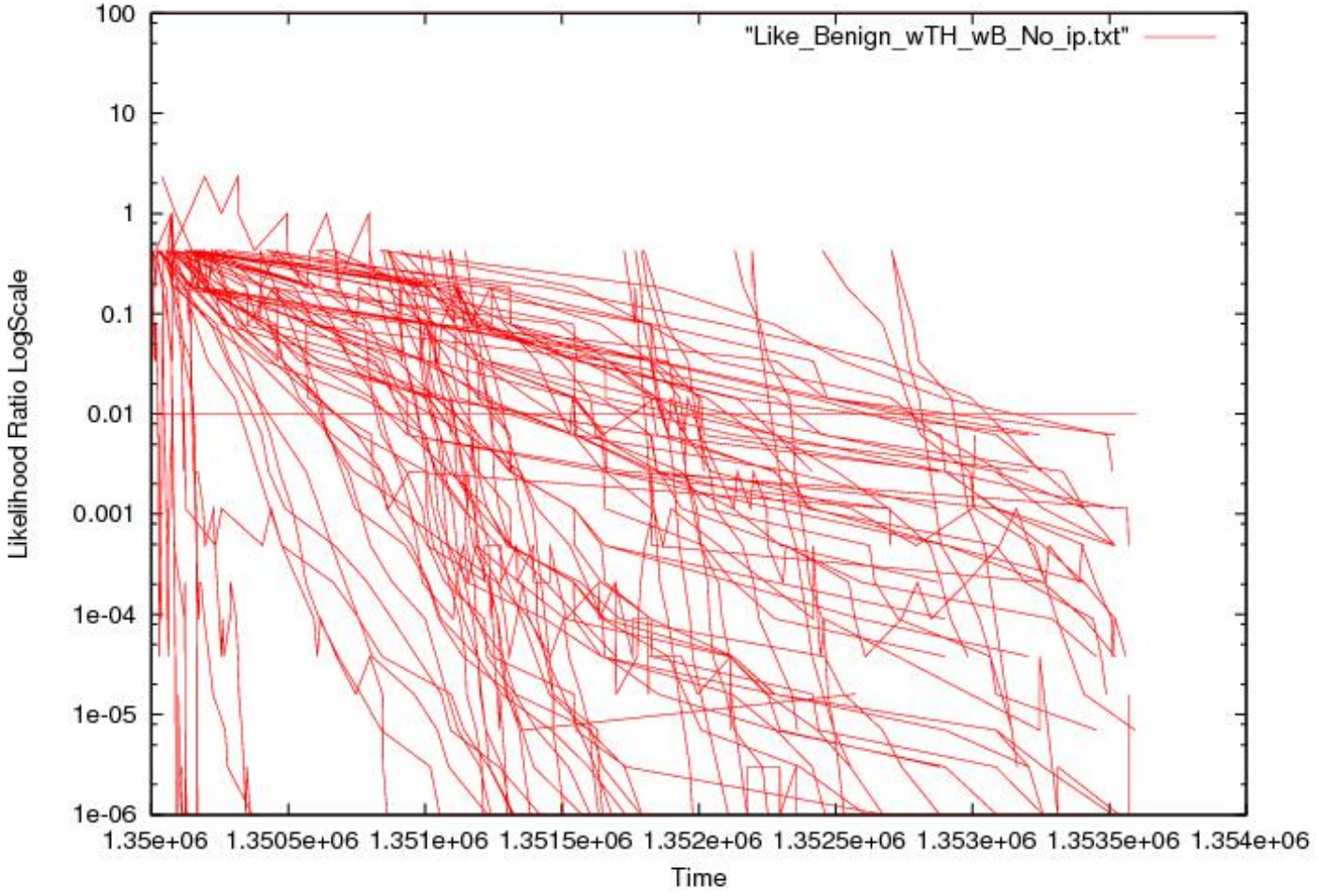
DALHOUSIE UNIVERSITY
Inspiring Minds

Preliminary Results
Plot of Likelihood ration for Can't Says



DALHOUSIE UNIVERSITY
Inspiring Minds

Preliminary Results
Plot of Likelihood ration for Benign



DALHOUSIE UNIVERSITY
Inspiring Minds

Initial Conclusions

- Using Bloom filter, reduces the false positives, (by how much ?)
  - unique entries considered for a given filter criteria
- Using specific filter criteria for the bloom filter
  - detects vertical scanning
  - detects horizontal scanning

Further Work In Progress

- Need to improve the technique by
  - Vary theta0 and theta1 values
  - Effect of timeout period
  - Real time scenario
- Long term analysis of IPs toggling between the three regions
  - Esp. from scanning to Can't say or benign

Acknowledgments

- Ron McLeod
- TARA
- Faculty of Computer Science, Dalhousie University

DALHOUSIE UNIVERSITY
Inspiring Minds

Thank you

Questions ?

DALHOUSIE UNIVERSITY
Inspiring Minds