Revisiting the Threshold Random Walk Scan Detector

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

Presented for FLOCON 2008
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:
      - Successful: Decreases Connection Ratio
      - Failed: Increases Connection Ratio

<table>
<thead>
<tr>
<th>Connection</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Successful</td>
<td>Decreases</td>
</tr>
<tr>
<td>Failed</td>
<td>Increases</td>
</tr>
</tbody>
</table>

Diagram:
- Scanner Ratio vs. Time
  - Can’t Say
  - Benign
  - Threshold
Threshold Random Walk

- The ratio is calculated as:

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

- Where the probabilities are:

\[\begin{align*}
\Pr[Y_i = 0|H_0] &= \theta_0, & \Pr[Y_i = 1|H_0] &= 1 - \theta_0 \\
\Pr[Y_i = 0|H_1] &= \theta_1, & \Pr[Y_i = 1|H_1] &= 1 - \theta_1
\end{align*}\]

- \(Y\) = success (0) or failed (1) connection attempt
- \(H_0\) = benign hypothesis
- \(H_1\) = scanner hypothesis
- \(\Theta_0\) = probability that the source is benign, for a successful connection attempt
- \(\Theta_1\) = probability that the source is scanner for a successful connection attempt
Threshold Random Walk

- The thresholds are calculated based on
  - desired true positive ($\beta = 0.99$)
  - desired false positive ($\alpha = 0.01$)

\[
\eta_1 \leftarrow \frac{\beta}{\alpha}, \quad \eta_0 \leftarrow \frac{1 - \beta}{1 - \alpha}
\]
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 \{\text{sip, dip}\}
    • HIT = if a corresponding entry \{\text{dip, sip, sport, dport, proto}\} is found within a specified timeout period
    • MISS = if a corresponding entry \{\text{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

Specify Unique Criteria: SP or SDP or SDSP or SDDP or SDSDP

Flow Records → Bloom Filter → Unique Entries → Modified TRW
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

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

"bet_thresholds_final.txt" u 1:2:3
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

Flows per Month

- Number of Flows Original
- Number of Flows SD
- Number of Flows SDP
- Number of Flows SDSP
- Number of Flows SDSDP
Preliminary Results
Preliminary Results
Plot of Likelihood ration for Scanners
Preliminary Results
Plot of Likelihood ration for Can’t Says
Preliminary Results
Plot of Likelihood ration for Benign

![Plot of Likelihood ratio for Benign](image)
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
Thank you

Questions ?