Real-time Pattern Detection in IP Flow Data using Apache Spark

International Symposium on Integrated Network Management (IM 2019)
May 9, 2019

Milan Cermak, Martin Lastovicka, Tomas Jirsik
Institute of Computer Science, Masaryk University, Brno
Attack Detection in Network Flow Records

challenges that everyone has to deal with
Attack Detection in Network Flow Records
challenges that everyone has to deal with II.
Stream4Flow: Real Time Analysis
distributed data stream processing framework

IP Flow Records

Monitoring Probe

IPFIXcol

Collecting Process

Web Interface

Data Transform

Kafka

Elastic Stack

PatternFinder

Apache Spark Streaming

Patterns Definition

Monitoring Probe

Web Interface

Elastic Stack

PatternFinder

Apache Spark Streaming

Patterns Definition

JSON

JSON

JSON

JSON
PatternFinder
-taking advantage of similarity search

distance_function:
- biflow_quadratic_form

patterns:
- name: anomaly
  request: [23, 8983, 9098]
  response: [24, 1125, 9101]

distribution:
- anomaly:
  intervals: [0, 3, 5, 6, 7, 11]
  weights: [3, 2, 1, 1, 2, 3]
Pattern Definition
discovery of general attack patterns

Dataset
- Only network traffic of interest
- Include attack variations
- Creation
  - Real-world dataset
  - Artificial dataset

Pattern
- Easy to determine from dataset
- Statistical aggregations of attack characteristics
SSH Authentication Attack Use-case

from theory to real-world
Pattern Definition
Hydra, Medusa, or Ncrack?

Dataset Creation
- Virtual environment – attacker and server
- 3 tools, 5 different settings

Derived Patterns – median aggregation

<table>
<thead>
<tr>
<th>Tool</th>
<th>Request</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pkts</td>
<td>Bytes</td>
</tr>
<tr>
<td>Hydra</td>
<td>16</td>
<td>1973</td>
</tr>
<tr>
<td>Medusa</td>
<td>18</td>
<td>2528</td>
</tr>
<tr>
<td>Ncrack-1</td>
<td>13</td>
<td>2860</td>
</tr>
<tr>
<td>Ncrack-2</td>
<td>16</td>
<td>3340</td>
</tr>
</tbody>
</table>
Evaluation comparison with others

Measurement

- one week period
- 478.98 M Flows, 5.54k Flows/second, 9.9k Flows/second in peak
- 21.91 TB data processed

Comparison

- Commercial solution Flowmon Anomaly Detection System
  - More than 30 login attempts in 5 min is an attack
- ADS 264 events from 75 IPs vs PatternFinder 78 events from 42 IPs
  - ADS overlapping events
- Accuracy 39%, precision 82%, recall 43%
Further Results
additional findings worth mentioning
Thank you for your attention

https://stream4flow.ics.muni.cz/
@csirtmu

Milan Cermak et al.
cermak@ics.muni.cz