Clustering and Monitoring Edge Behaviour in Enterprise Network Traffic

Chris Schon, N. Adams & M. Evangelou
christopher.schon.16@ucl.ac.uk

University College London & Imperial College London

September 2017
Outline

1. Introduction
2. Aim
3. Data
4. Methodology
5. Examples
6. Conclusion
Outline

1. Introduction
2. Aim
3. Data
4. Methodology
5. Examples
6. Conclusion
Cyber-security has become an increasing threat at an organisational and national level.

The WannaCry ransomware attack affected 230,000 computers in over 150 countries. *Image: Malware Tech.*

Traditional signature-based detection techniques can easily be circumvented by sophisticated intruders.
Aim

- Complement established methods with novel anomaly detection techniques.
- Provide a dashboard-like collection of statistics, interpretable to a network analyst, which give extra support for situational awareness and network cognisance.
- Previous efforts include quantifying ‘normal’ activity at a global level (A. Valdes and K. Skinner, 2000) and at a local level, such as monitoring specific device connections and clusters (J. Neil et al., 2013).
Outline

1. Introduction
2. Aim
3. Data
4. Methodology
5. Examples
6. Conclusion
One week of NetFlow event records from Los Alamos National Laboratory (LANL).

<table>
<thead>
<tr>
<th>Time</th>
<th>Conn. duration</th>
<th>Source comp.</th>
<th>Source port</th>
<th>Destination comp.</th>
<th>Destination port</th>
<th>Protocol</th>
<th>Packet Count</th>
<th>Byte count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>C1065</td>
<td>389</td>
<td>C3799</td>
<td>N10451</td>
<td>6</td>
<td>10</td>
<td>532</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>C1423</td>
<td>N1136</td>
<td>C1707</td>
<td>N1</td>
<td>6</td>
<td>5</td>
<td>847</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>C1423</td>
<td>N1142</td>
<td>C1707</td>
<td>N1</td>
<td>6</td>
<td>5</td>
<td>847</td>
</tr>
</tbody>
</table>
Figure 1: Number of NetFlow events by time window.

Figure 2: Number of unique connections by time window.

Figure 3: Number of unique devices by time window.

Figure 4: Total bytes transferred by time window.
Outline

1. Introduction
2. Aim
3. Data
4. Methodology
5. Examples
6. Conclusion
Our anomaly detection technique follows a three-step procedure:

1. Gather features which describe the active connections (edges) in the network during a 15-minute time window.

2. Group edges into clusters based on their feature vectors.

3. Repeat this process over contiguous windows. Then, derive a series of informative indicators by examining the relationship of edges with the observed cluster structure.
Edge Data
1  C1 C2 0 1 193
2  C1 C2 0 1 200
10 C1 C2 0 2 400

Features: standardised summary statistics based on column values

Repeat over all edges over all time windows

W1

W2

TIME
## Features

| **Time related** | Mean, median, standard deviation and inter-quartile range (IQR) of connection *duration*.  
|                 | Mean and standard deviation of *inter-arrival times* of events.  
| **Bytes related** | Mean, median, standard deviation and IQR of *bytes*.  
| **Packet related** | Mean, median, standard deviation and IQR of *packets*.  
| **Other** | Count of the *number of events* that occurred on the edge.  
|             | Count of events on the *modal destination port* observed on the edge.  

Clustering

- K-means clustering with 5 centroid vectors using standardised edge features over each of the chosen time windows.
C.Schon, N.Adams & M.Evangelou (University College London & Imperial College London)

Data Science for Cyber Security

September 2017 15 / 23
Derived Indicators

Indicator variables for each active edge of each time window are constructed. One each for:

- Whether the edge was present in the previous time window.
- Whether the edge is an outlier.
- If the edge was present in the previous window, whether its cluster assignment has changed.

Where the 1 indicates a shift from the ‘normal’ behaviour. By burning in the algorithm over a day’s worth of data after the edge’s first appearance, control charts can be constructed for the expected frequency of each indicator for each edge.
Cluster change rates show weekday seasonality. Outlier rates are noisy about 0.01.
Outline

1. Introduction
2. Aim
3. Data
4. Methodology
5. Examples
6. Conclusion
This edge shows changing new edge and outlier rates (upper and lower left), but constant cluster change rate (upper right).
If the indicators for the edge sum to two (note: not three), then the edge is considered *anomalous*. Example window:
Outline

1. Introduction
2. Aim
3. Data
4. Methodology
5. Examples
6. Conclusion
Conclusion

- We have developed a methodology that can monitor NetFlow traffic statistics in a real-time environment.
- We can construct different scales of data analysis: individual edges, specific subgraphs, and entire network edge-behaviour.
- This provides a foundation for a complementary enterprise network analysis tool for detecting abnormal edge behaviour.
Thank you

References
