Proliferation of applications and services that rely on HTTP

Is it possible to automatically mine data to discover potential threats?

GOAL: Develop methodologies and tools to dig into this data and extract useful information

- Looking for similar patterns in traffic
- Clustering similar URLs
- Analyze results

Automatic method that provides aggregated views of URLs
- Simplifies network administrator’s tasks
- Use of passively monitored network traffic
- Transparent for the user
- Completely unsupervised methodology

Workflow

URL Extraction

Network Probe

Batch of collected URLs

Similarity Evaluation

Compute \( d_{URL} \)

Similarity Matrix

DBSCAN

DBSCAN clustering algorithm is applied. Additional statistics are provided for the analyst

Final view & Analysis

Silhouette analysis

an unsupervised methodology to find how well each object lies within its cluster

Can ease the choice

those clusters whose points are very similar inside the cluster, and very different from points in the rest of clusters are likely to be interesting

<table>
<thead>
<tr>
<th>(C)</th>
<th>Main hostname (unique number)</th>
<th>Elements</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.92</td>
<td>skypo_streaming-Lakamahid.net (1)</td>
<td>551</td>
<td>Streaming</td>
</tr>
<tr>
<td>0.91</td>
<td>ad.doubleclick.net (1)</td>
<td>59</td>
<td>Advertising</td>
</tr>
<tr>
<td>0.87</td>
<td>cookeax.amp.yahoo.com (1)</td>
<td>641</td>
<td>Malware</td>
</tr>
<tr>
<td>0.85</td>
<td>static.simply.com (1)</td>
<td>25</td>
<td>File-Hosting</td>
</tr>
<tr>
<td>0.81</td>
<td>d42wbsmhbehld.cloudfront.net (1)</td>
<td>63</td>
<td>File-Hosting</td>
</tr>
<tr>
<td>0.81</td>
<td>mdc001.org (1)</td>
<td>27</td>
<td>Malware</td>
</tr>
<tr>
<td>0.79</td>
<td>adserver.weabads.it (1)</td>
<td>26</td>
<td>Advertising</td>
</tr>
<tr>
<td>0.77</td>
<td>com (3)</td>
<td>37</td>
<td>Malware</td>
</tr>
<tr>
<td>0.75</td>
<td>pixel.quantserve.com (1)</td>
<td>57</td>
<td>Advertising</td>
</tr>
<tr>
<td>0.72</td>
<td>watson.microsoft.com (1)</td>
<td>29</td>
<td>Windows</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Debug</td>
</tr>
</tbody>
</table>

Clusters sorted by silhouette coefficient

Current Developments – questions and experiments

Potential Applications

- Use it to compare activities of different groups of hosts (URLs Groups)
- Detecting the web traffic changes in time for a group of hosts

Experiment

Compare two groups of URLs, the second has the same traffic as the first, plus other injected traffic. Then, look for anomalies.

U \( G_2 \) clusters that contain only new traffic are at a greater distance from U \( G_1 \) clusters.

In the graph some noise is visible, caused by the URLs sampling.