Domain Name Classification via Web Search Results Mining
Tiffany Cauthen and Roberto Perdisci

Goals
- Learn the differences between malicious and legitimate domain names.
- Create a tool to classify the legitimacy of unknown domain names.
- Help unsuspecting victims detect malicious domain names and prevent infections.

The Problem
- Domain names are mnemonic addresses to different Internet sites.
- Malicious domains are often used to spread malicious software or remotely control infected machines.
- Legitimate domain names do not host harmful content and provide useful services.
- We want to create technologies that can help us automatically distinguish between malicious and legitimate domain names.

Our Approach
For each domain, we query the domain name string on public search engines to learn the “Internet sentiment.”

We consider the following features:

Domain Results
The hypertext links in the search results

Web Content
We visit each link and grab the related web page content

Domain Name
Syntax-based analysis of queried domain (e.g., length, letter frequency, etc.)

Preliminary Results

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Correctly Classified Instances</th>
<th>Incorrectly Classified Instances</th>
<th>Weighted Accuracy Average</th>
<th>Confusion Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domain Results</td>
<td>1071</td>
<td>93</td>
<td>0.92</td>
<td>351</td>
</tr>
<tr>
<td></td>
<td>92.0103 %</td>
<td>7.9897 %</td>
<td></td>
<td>85</td>
</tr>
<tr>
<td>Web Content</td>
<td>1013</td>
<td>151</td>
<td>0.87</td>
<td>291</td>
</tr>
<tr>
<td></td>
<td>87.0275 %</td>
<td>12.9725 %</td>
<td></td>
<td>145</td>
</tr>
<tr>
<td>Domain Names</td>
<td>1148</td>
<td>33</td>
<td>0.972</td>
<td>726</td>
</tr>
<tr>
<td></td>
<td>97.2058 %</td>
<td>2.7942 %</td>
<td></td>
<td>19</td>
</tr>
</tbody>
</table>

Conclusion, Next Steps
- Accurate identification of malicious domain names by leveraging “Internet sentiment” about domains seems possible.
- Next steps: Collect more data, combine the three classifiers into one, and improve overall accuracy.