CS 419/DADA

Week 7 - Lab 2

Michael Depuy

Jonathan McNeil

Marie Caswell

Our team approached the lab by first coming up with a set of rules that would indicate a possibly malicious website. Then we assigned a weighted point value to each URL based on our rules. A higher point value indicates a greater degree of potential malicious behavior.

Some URL characteristics we counted as suspicious were:

* A domain age less than 188 days.
* TLD not from a common domain like ‘com,’ ‘net,’ or ‘org’
* An Alexa rank greater than 1000000, or no Alexa rank.
* A domain having greater than three tokens.
* A difference between the registered domain location and the mail exchange location.
* ‘Bad’ file extensions (like .pdf or .exe files).

Our program then classifies any website with a score greater than four as malicious.

We also attempted avoid false positives by awarding ‘safety points’ — meaning that if a URL had characteristics indicating legitimacy (like being located in the same region it was registered) the URL could gain points back.

Our final run against the *classify.json* file found 953 legitimate links and 1,071 malicious links.