Sample Paper: <http://cs229.stanford.edu/proj2012/ChongLiu-MaliciousURLDetection.pdf>

Assumptions: -1: Legitimate, 0: Suspicious, 1: Phishing

**Our Research Goal**

Our project aims to identify the most relevant subset of features that can accurately identify phishing URLs, using Decision Tree and Naïve Bayes algorithms together with five different subsets of features.

**Methodology**

We chose to use the classification algorithm, Random Forest, because our dataset is made up of discrete data. This means each feature can only be represented by -1, 1 and sometimes 0. Random Forest also proves to be highly efficient when working with large datasets. We used 10,000 URLs for training and 1,055 for testing. The test data contained 461 phishing sites and 594 legitimate sites. Our belief is that using more features to train the algorithm should result in better predictions, and hence higher accuracy, recall and F1 scores, for the test data.

[FlowChart Diagram here]

**SubSect: Data**

We collected our training and test data from the UCI phishing dataset that is publicly available. This dataset contained 2456 unique URL instances and a total of 11,055 URLs of which 6,157 are phishing and 4,898 are legitimate sites. Each row represented a URL and each URL was previously parsed and represented according to 30 features which could determine whether or not the URL is used for phishing or just identify a specific feature as suspicious for a particular URL. The features considered includes whether or not an IP Address is used instead of a URL, the length of the URL, the presence of link tags to the same domain as the webpage and whether or not the webpage uses IFrames. At the end of each row, there is a result which identifies the true nature of the URL, 1 if it is a phishing site and -1 if it a legitimate site.

**SubSect: Featuers/Sets**

From the 30 features we identified five subsets. These were:

~~Set A: Features we think are most important.~~

~~Having\_IP\_Address, URL\_Length, having\_At\_Symbol, Domain\_registration\_length, links\_in\_tags, submitting\_to\_email, Iframe, age\_of\_domain, having\_Sub\_Domain, Redirect.~~

~~0, 1,3,8,14,16,18,22,23 (A-68.90%)~~

Set A: Web-presence related features: Domain age, Website traffic, Page Rank, Google Index 23,25,26,27 (A-50.42%) 4 features

Set B: Features without 3 possible outcomes {-1,1}. This makes the data more binary and eliminates the possibility of uncertain URLs. 21 features

having\_IP\_Address, Shortining\_Service, having\_At\_Symbol, double\_slash\_redirecting, Prefix\_Suffix, Domain\_registeration\_length, Favicon, port, HTTPS\_token, Request\_URL, Submitting\_to\_email, Abnormal\_URL, Redirect, on\_mouseover, RightClick, popUpWidnow, Iframe, age\_of\_domain, DNSRecord, Page\_Rank, Google\_Index, Statistical\_report.

0,2,3,4,5,8,9,10,11,12,16,17,19,20,21,22,23,24,26,27,29 (A-76.77%)

Set C: Features with 3 possible outcomes {-1,0,1} 8 features

1,6,7,13,14,15,25,28 (A-94.12%) Figure out WHY this is so good!

Set D: All 30 features (A- 95.73%)

Set E: The most important features (those rated 0.01 and up) according to classifier.feature\_importances\_ function.

0,1,5,6,7,8,12,13,14,15,23, 24, 25, 26, 27, 28 (A-96.20%) 16 features

having\_IP\_Address, URL\_Length, Prefix\_Suffix, having\_Sub\_Domain, SSLfinal\_State, Domain\_registeration\_length, HTTPS\_token, Request\_URL ,URL\_of\_Anchor ,Links\_in\_tags ,SFH ,age\_of\_domain ,DNSRecord,web\_traffic,Page\_Rank ,Google\_Index,Links\_pointing\_to\_page

1. SFH { -1,1,0 }
2. Submitting\_to\_email { -1,1 }
3. Abnormal\_URL { -1,1 }
4. Redirect { 0,1 }
5. on\_mouseover { 1,-1 }
6. RightClick { 1,-1 }
7. popUpWidnow { 1,-1 }
8. Iframe { 1,-1 }
9. age\_of\_domain { -1,1 }
10. DNSRecord { -1,1 }
11. web\_traffic { -1,0,1 }
12. Page\_Rank { -1,1 }
13. Google\_Index { 1,-1 }
14. Links\_pointing\_to\_page { 1,0,-1 }
15. Statistical\_repo {-1,1}
16. having\_IP\_Address { -1,1 }
17. URL\_Length { 1,0,-1 }
18. Shortining\_Service { 1,-1 }
19. having\_At\_Symbol { 1,-1 }
20. double\_slash\_redirecting { -1,1 }
21. Prefix\_Suffix { -1,1 }
22. having\_Sub\_Domain { -1,0,1 }
23. SSLfinal\_State { -1,1,0 }
24. Domain\_registeration\_length { -1,1 }
25. Favicon { 1,-1 }
26. port { 1,-1 }
27. HTTPS\_token { -1,1 }
28. Request\_URL { 1,-1 }
29. URL\_of\_Anchor { -1,0,1 }
30. Links\_in\_tags { 1,-1,0 }

**SubSect: Experiment**

Precision: P = TP/TP + FP - all correct classified / all classified

Recall: R = TP/TP + FN - all correct classified / should be classified

F1 weight the Precision and Recall, in other word choose which one is more important, Default is P = R

Libraries used

How we parse the data into the different matrices.

Results/Discussion