**Project Topic: Detecting Malicious URLs**

**Students: Ashraful Tauhid, Surendra Pathak, Joshua Menefee**

**Team: 11**

1. **Papers Covered:** We’ve read a few papers on malicious URL detection. For instance, a paper named “Finding effective classifier for malicious URL detection” discussed on the character features of malicious URLs by statistical methods to obtain char distribution features and structural features and used the WEKA library for construction of models. Another paper named “Malicious URL Detection Based on Associative Classification” achieved an accuracy of 95.8% with low false positive and negative rate using associative classification. We tend to read a few more papers to clear the fog before we jump into the implementation. Besides publications, we are also reading a few only blogs and articles related to Malicious URLs detection. For instance,

* <https://www.kdnuggets.com/2016/10/machine-learning-detect-malicious-urls.html>
* <https://medium.com/@ismaelbouarfa/malicious-url-detection-with-machine-learning-d57890443dec>
* <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7648114/>
* <https://towardsdatascience.com/predicting-the-maliciousness-of-urls-24e12067be5>
* <http://www.sysnet.ucsd.edu/projects/url/>
* <http://athena.ecs.csus.edu/~narendrp/>
* <https://github.com/philomathic-guy/Malicious-Web-Content-Detection-Using-Machine-Learning>

1. **Datasets:** We haven’t collected the dataset yet but since we cannot single handedly reproduce a huge dataset of malicious URLs, we will be taking help from the noted sites:

* Phising URLs

1. Phishtank - <https://www.phishtank.com/developer_info.php>
2. Open Phis - <https://openphish.com/>

* Spam URLs

1. JWSPAMSPY - <http://www.joewein.de/sw/blacklist.htm>

* Malware URLs

1. DNS-BH - <http://www.malwaredomains.com/wordpress/?page_id=66>
2. <https://www.malwarepatrol.net/my-account/>
3. <http://www.malwaredomainlist.com/>

* Block listed URLs:

1. <https://zeltser.com/malicious-ip-blocklists/>

* Safe URLs

1. <https://majestic.com/reports/majestic-million>

**3. Our next tasks:**

* Select a set of base papers for review and analysis
* Accumulate the malicious URLs with the help of datasets from the above sites and add a few on our own in a CSV format
* Pre-process the dataset so that the implementation stage doesn’t face hiccups
* Implement and analyze the results with respect to the base papers which already had proven to be legitimate
* Write a paper on the findings