**Capstone Project – DETECTION OF PHISHING URLs IN NEAR REAL TIME**

**Problem Statement:**

Web Proxy Server in general, is an application or program that serves as an agent or a middleman, acting on your behalf, to any activity over the internet. There is one more important application of such program in terms of security is to observe the activity over the internet of users.

Web Proxy Server can permit or deny the request sent by the user to access any resource which is available out there in internet. To do this Web Proxy Server is dependent on Whitelist and blacklist features. Those URLs are configured as whitelisted are permitted to access and those are configured as blacklisted are not permitted to access for users. To configure such lists it requires human intervention and this is where it becomes very challenging.

Since it is very difficult to configured all the malicious / phishing urls as blacklisted where there are 9,000 (approx.) malicious / phishing urls found every day.

The challenge is to detect the newly created malicious / phishing urls and deny access to them in near real time so that users system would not get infected.

**Approach to solve the problem:**

* Gather all the logs generated by Web Proxy Servers.
* Extract the relevant information from logs in our case its ‘url’.
* Derived the various features by analyzing each url.
* Lable the benign and phishing url to train the model.
* Provide the train dataset to model to classify the url to benign and phishing category.
* Test the accuracy of model using test dataset.

**Note:** The above approach may change later as currently we have not decided to which technique / method use to classify the urls.

**Dataset:**

**Data Source:** Web Proxy Server

**Total number of data points:** 5002

**Total number of unique benign urls:** 3167 (63% of total data points)

**Total number of unique phishing urls:** 1833 (37% of total data points)

**Dataset contains following variables:** source address, url, label