LITERATURE SURVEY

1. Existing solutions

There are many existing solutions deployed for this use case.

- CANTINA

CANTINA is an Internet Explorer toolbar that decides whether a visited page is a phishing page by analyzing its content. CANTINA uses Term Frequency-Inverse Document Frequency (TF-IDF), search engines and some heuristics to reduce false positives. The following procedures are performed by CANTINA to detect phishing websites:

1) TF-IDF of each term on a suspected web page is calculated.

2) Top 5 terms with highest TF-IDF values are taken to represent the document (Named lexical signature in

3) Submit the 5 terms into a search engine (e.g. Google search query:

http://www.google.ae/search?q=t1, t2, t3, t4, t5,), and store domain names of the first returned n entries.

4) If the suspected domain name is found within the n number of returned results, then the site is legitimate.

- Google Safe Browsing API

Google Safe Browsing API enables client applications to validate whether a given URL exists in blacklists that are constantly updated by Google [26]. Although the protocol is still

experimental, it is used by Google Chrome and Mozilla Fire-fox. The current implementation of the protocol is provided by Google, and only consists of two blacklists named goog-phish-shavar and goog-malware-shavar, for phishing and malware respectively. However the protocol itself is agnostic to the list type as well as to the provider of the list.

The API requires client applications to communicate with providers through HTTP while adhering to syntax specified in Protocolv2Spec [27], which is the second version of the protocol; the first version faced scalability and efficiency issues

2. TECHNICAL PAPERS

- Phishing Detection using Machine Learning based URL Analysis
- Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning

3. EXISTING PRODUCTS

- Avanan-https://www.avanan.com
- Proofpoint-https://www.proofpoint.com/us/products/advanced-threat-protection/targeted-attack-protection