Phish report

Resources:

- Machine Learning method https://github.com/philomathic-guy/Malicious-Web-Content-Detection-Using-Machine-Learning/tree/c86d4709354691d702e0c693f482e1fabd3b172c

- Spelling Corrector

  https://norvig.com/spell-correct.html

Anti- phishing pipeline by priority

- Cache
  - White list
  - Black list
- Regular expressions
- Spelling corrector
- External sites
- ML predication

The priority is important because of the available resources of computation. The best solution is to spend less time on secure sites and use the resources of the computer on unknown sites.

Cache

Check the url in the white/black cache. The most simple method to find if a site is suspicious or not. If a site presents in the white/black list, allow/disallow it to redirect to the site . The cache can be filled from files and updated later by new sites after the anti-phishing pipeline.

Regular expressions

The regular expressions are old-fashioned but still a good tool for searching/matching patterns in the urls.

Spelling Corrector

For example hackers can dispel the site “microsoft” by “m1crosoft”, “microsaft” and etc.
If it is important for the client not to allow this behavior this feature can be included in the pipeline. The spelling is memory consumption (depences on the length of the url).

External links

Available information can be found from other secure known sites by the network. Depends on latence in the network.

ML prediction

From the link (see above “Machine Learning method”) there are methods that allow to extract features from a url to be used for ML algorithms to predict if a site is suspicious or not. The ML has heavy cpu and memory consumption and the best place is the last in the pipeline

The next features presents in the git repo (see feature_extraction.py)

- url length - long url is suspicious
- presence of “@” symbol
- ip address as part of a url
- google index of a site
- double slash
- web traffic
- DNS
- iframe
- abnormal url
- http/https
- age of domain

The above features are the short list of 22 features in the file.