CS 3546: Intro to Security Analytics
Course Project: Data-driven Identification of Malicious Versus Bening Domains
Spring 2023

Description:

Websites have long been abused as a medium to propagate various cyberattacks such as malware, Phishing, online scams, and frauds. There has been a lot of advancement in identifying malicious websites in the wild and blocking them through putting them in a blacklist or blocking the connection from your internal network so that no users can interact with them. Moreover, there are browser level interventions to block a website and prevent users from interacting with it such as Google Chrome shows alerts when someone tries to visit a website which has potential for Phishing or other maliciousness. In this project, we want to learn about these groups of websites and domain names in more detail and figure out ways to identify them or separate them based on statistical analysis.

Initially, you are given **2 lists** of website domains (as ground truth) where one of the lists contains good or reputed websites on the internet, while the other list contains website domains that are either labeled as Phishing urls. Make sure to clean the Phishing lists and make the comparisons between domains and not include the URL paths of the protocols. However, there is also a third list of websites and domains where there is a mixture of good and bad web domains. Now, your end goal is to understand and identify the domains in this **third** list as either good or bad domains.

In this project, your task is to first determine the characteristics of good websites versus the characteristics of bad websites from the given ground truth lists (e.g., first 2 lists). You can take inspiration from lab-08. But, you need to be more creative than that.

In order to learn characteristics, you can check

- their FQDN (Fully Qualified Domain Name) length
- FQDN digit percentage
- FQDN unique characters
- If FQDN is a twisted name or 'typosquatting' for some reputed brand (e.g., Go0g1e.com).
- If FQDN contains a famous brand name (e.g., Amazon, ebay, google etc.)
- FQDN number of '.' (dot) characters
- Number of hyphen '-' characters
- What is most prevalent TLDs among bad websites versus good websites (e.g, TOP 10 TLDs between these groups),
- Get TLD reputation/badness score from spamhaus (https://www.spamhaus.org/statistics/tlds/)

- What is the rank of that FQDN in the 'Tranco rank list' (given), What is the avg. ranking in terms of malicious domains and benign domains.
- Level of subdomains (e.g., "my.web.sites.com" for this FQDN the domain is sites.com but there are additional 2 level of subdomains "my.web")
- Can we extract any content from the homepage and match if it is similar to domain name. For example, if you visit "https://www.wayfair.com/" there should be some text in the homepage related to wayfair. In case of malicious websites, oftentimes there is no content or no related contents.
- Is the website/domain name using any page redirection. Meaning if you want to visit a page then it takes you to another landing page automatically. If yes, then what is the number of redirects meaning how many hop the before going to the landing page. Use the redirect count as a characteristic and see if malicious websites are more likely to use a redirect than the benign ones. Also, if the redirect domains are completely different from the one you requested.
- You can also use other characteristics like what is the IP address for that domain and where that IP is residing or web scraping to get the size of the homepage (if 404 or 403 status code then size should be 0). You can also use Internet WayBack Machine (https://archive.org/web/) to collect information about a website. Other than that, if you feel any other characteristics can help you differentiate these groups of domains, this part can get your extra credit.

Note: If you feel like some of these characteristics are not useful then you can discard them and get more creative with other characteristics based on your research/findings.

Now, once you have learned about these characteristics for both good and bad domains. Based on these findings, can we go to the third list and make some prediction/identification of good versus bad websites by using takeaways from the characterics extraction part. If two lists are of different sizes, then analyze the graphs in terms of percentages rather than exact values.

Datasets to be used for this project:

- Tranco full list for ranking (tranco_full_list_for_ranking.csv)
- Malicious Domain List (malicious_URLs.csv)
- Benign list (benign_domain_list.csv)
- Mixed List (mixed_domain_list.csv)

By the end of this project, you need to **submit a PDF report** of your findings with gap analysis, bar graphs, pie charts, differences in means, and other creative ways to differentiate two groups. You have to also report how many if the 3rd list (mixed list without labels) is malicious vs, benign. You also need to justify why for each of these identification. If you feel something is false positive (wrong detection as a malicious class) or false negative (wrong detection as benign class) based on the statistical outcomes, report any example cases of such. Finally, you need to record a short presentation (8-10 minutes) to highlight your findings about the malicious/benign lists, mixed list, and takeaways.

Tools or Python libraries you can leverage:

- TLDextract (https://pypi.org/project/tldextract/)
- Dnstwist (https://pypi.org/project/dnstwist/)
- from newspaper import Article ##(pip install newspaper3k) [For web scraping]
- Waybackpy(https://pypi.org/project/waybackpy/) [For Querying wayback machine]
- Other related libraries we used in the class might help