Orchestrating Open Source Software and WHOIS Newly Registered Domain data feeds to fight the typosquatting plague

Posted on August 4, 2022

Typosquatting and related types of cyber threats, such as domain squatting, phishing campaigns, IDN homoglyph attacks, etc., cause significant harm and incur financial loss, so it is vital to be vigilant and fight against these malicious threats.

In 2020, Andre Tenreiro started the development of an excellent Open Source Intelligence (or simply OSINT, pronounced “O-Sint”) security tool called, openSquat; a "domain squatting and phishing watchdog". It is based on fuzzy text search in the list of newly registered domains in a given day, week, or month, in order to identify domains related to a given list of keywords. Based on these matching domain results, he further extended the list to include known subdomains, check if their web server ports are open, and whether their certificates are valid.. These findings can be validated against VirusTotal and verified for phishing activity. OpenSquat is implemented in Python using well-established open source Python libraries. It is a powerful swiss army knife for brand protection or any kind of typosquatting attack detection.

The primary input of the program consists of a list of Newly Registered Domains (NRDs). Obviously the completeness and coverage of the NRD list has a relevant influence on the number of findings: the more completeness of the input list, the better the results will be. The program downloads a free list automatically by default, leading to some useful results. However, in our experience, using openSquat with the domain lists provided by WhoisXML API's Newly Registered Domains data feed, the results can be drastically improved. In the present blog we demonstrate how. We will conduct our experiment in the popular BASH shell, and all commands are replicable with average Linux CLI knowledge. Our experiment took place on 1 August 2022, after 8 a.m. UTC. We have used the default search search terms provided in the file keywords.txt in the

opensquat package:

- google
- facebook
- amazon
- paypal
- microsoft

In this instance we tried a search with the default settings and no extra or special requirements like subdomain search, lookup in various known blacklists, etc. For example:

```
./opensquat.py -k keywords.txt -o opensquat_results_for_default_keywords_opensquat.txt
```

which resulted in the updating of the file domain-names.txt from opensquat's default data source for the day, and a fuzzy search for the keywords.

The search has found 74 domains like amazonsecure-paymnt[.]com or paxpal[.]club, etc. These are supposed to be daily detections, however, we could not find any documentation about the completeness of the domain list the program uses by default.

Nevertheless the list downloaded by the program contained 86,275 domains altogether, which is less than the number of new domains registered in the ".com" domain alone. Hence, the list, while it is provided for free, is incomplete.

Certainly nobody can provide a complete list of domains registered on a day; in fact, such an accurate list does not even exist. Domain names are kept in a decentralised database (DNS and WHOIS) run by numerous entities. Domain administrations are done in virtually all timezones, and the time between the official registration date (as presented in the WHOIS record) and the actual time of the domain becoming alive also differ. To compile an approximate list of domains registered on a given day is therefore a very involved, arduous technical challenge due to its “big data” processing tasks which are to be accomplished.

In the light of all this, opensquat's offer of a daily list of registered domain names is very generous and appreciative. However, one is eager to check the actual registration dates of the listed

domains. We checked 400 randomly chosen domains from the list using the WHOIS API. We found that the list provided at the time of the experiment, 1 August 2022, 8 a.m. UTC mostly contains domains registered between 25 and 27 July. This is not too bad as it takes time for these domains to become active. But altogether, for the delay in a new domain's appearance in the list and the relatively low number of the domains contained, it is reasonable to try using opensquat with a more comprehensive list.

WhoisXML API's Newly Registered Domains data service offers the probably most accurate list of newly registered domains. In particular, for our experiment we used the file nrd.2022-07-31.ultimate.daily.data.csv.gz as it is was latest one that had been available at the time of the experiment: the files for a given date are uploaded the next day by 02:00 UTC. (Consult the service's website for available data sets and pricing options.)

The data comes together with WHOIS records, so one can readily check the distribution of the registration dates. In this particular case, most of the domains were registered between 27 and 30 July, so this list is definitely more up-to-date. We have 132,234 domains in the list, that is, over 30 percent more than in the free list. The two lists actually share only 10,278 domains as the time frame is different: our file contains more recent domains. Let us describe how to actually do the search after downloading the data file. The list contains not only added, but also dropped, changed, and discovered domains, all with WHOIS data; we only need the names of the added ones. To obtain such a list we can use the csvkit utility along with some standard tools

```
zcat nrd.2022-07-31.ultimate.daily.data.csv.gz | csvgrep -c reason -m added \
    | csvcut -c domainName \
    | tail --lines +2 > domain_names_nrd2_20220731.txt
```

(tail was used to skip the header line.) Having generated our list of domains we can run opensquat, now with our domain list instead of the default one:

```
./opensquat.py -k keywords.txt -d domain_names_nrd2_20220731.txt -o opensquat_results_for_defau
```

Now we get 120 finds instead of the 74 coming from the free list. There are only 8 domains in both lists but this is acceptable: to get a bigger intersection we should use older files from the

WhoisXML API data feed.

In conclusion, while opensquat is a great utility, its efficiency can be beefed up with using a more accurate daily list of newly registered domains. Armed with opensquat's additional functionality not detailed here, one gets a tool with insane power to fight typosquatting and all related types of cyber mischief.