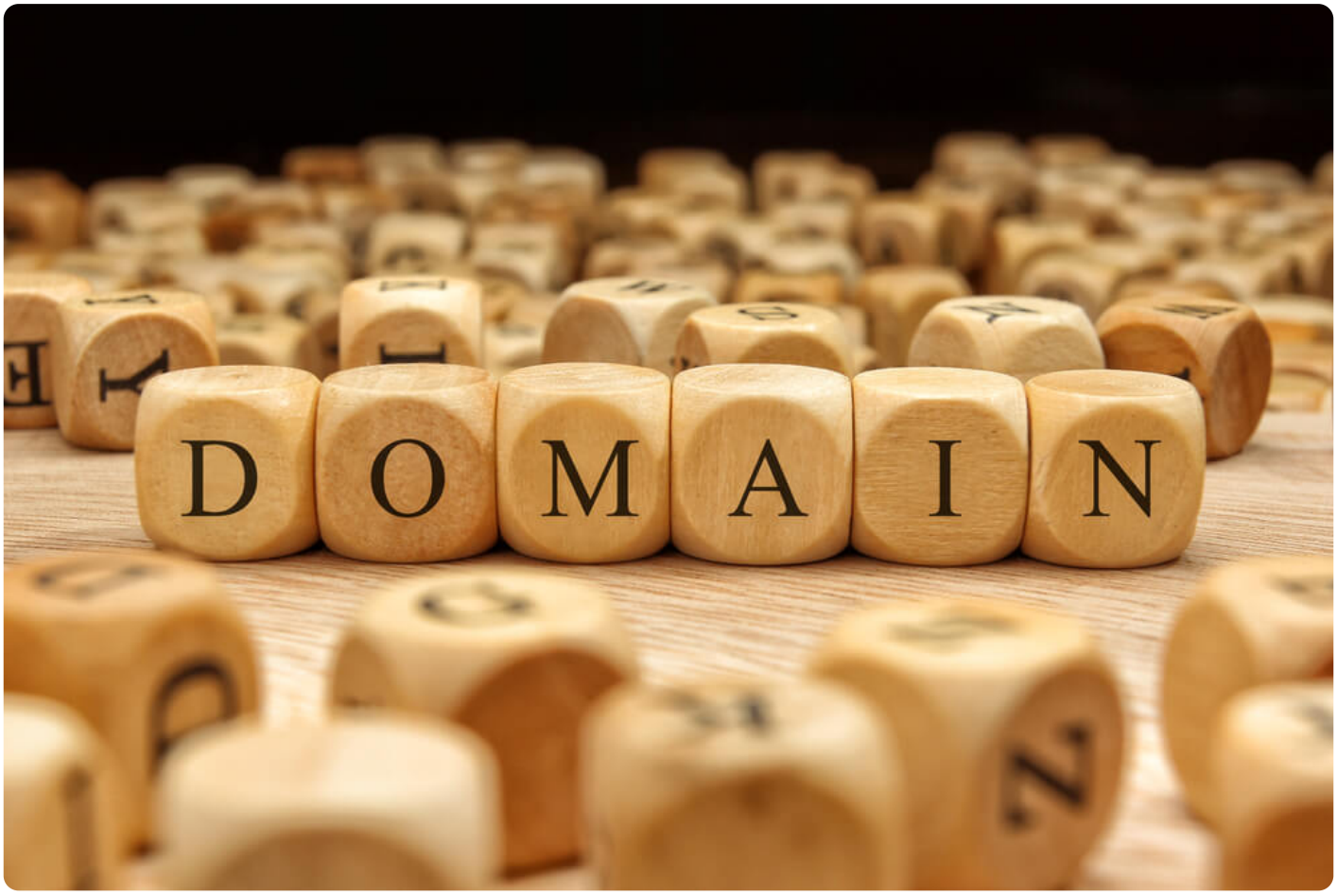
Typosquatting Daily Data Feed: the new enabler in the fight against phishing and malware

Posted on January 29, 2020



One result of our reseach and development is the introduction of the new "typosquatting data feed", an innovative data set based on our long-standing experience with cybersecurity and the Domain Name System. In what follows we will demonstrate how this new resource can be used efficiently in the fight against spam, phishing and malware.

The main idea behind the new data feed is the observation that domain names which were registered on the same day and have similar names have an increased likelihood of being involved in a range of IT scams, including typosquatting attacks, domain name hijacking, and also phishing and malware. So, we have developed a technology for finding these groups of domain names.

The origin of our data are the our newly registered domain daily data feeds, in particular,

- domain_names_new,
- ngtlds_domain_names_new,
- and cctld_registered_domain_names_new

These all have been around for a decade now, and they keep on capturing the new second-level domains on the very day they become technically operational in the Domain Name System, that is, on the day when they are resolved by a browser or other Internet-based applications. These feeds cover all generic top-level domains and several country-code domains, too.

The new data feed lists all groups of at least 3 domains which have names similar to each other in a set and were introduced on the same day. By "similar to each other" we mean similarity with respect to a suitably chosen algorithmically calculable mathematical characterization. These sets are published in .csv files on a daily basis. Importantly, this is not a blacklist in itself, however, the listed domains need special attention because they are potential actors or victims of malicious activity.

As an example, a set looks like this:

- mecedeshvacsettlement.com
- mercedeshvacssettlement.com
- mercedeshvasettlement.com
- mercedeshvacsettement.com
- mercedeshvcsettlement.com
- mercedeshvacsettlment.com
- wwwmercedeshvacsettlement.com
- mercedeshvacsetlement.com
- mercedeshvacsettlement.info
- mercedeshacsettlement.com
- mercedehvacsettlement.com
- mercedeshvacsettlement.org
- mercedeshvcasettlement.com
- mercedesvacsettlement.com
- mercedeshvacsettlement.com
- mercedeshvacsettelment.com
- mercedessettlement.com

While many names like those in the example are human-readable, there are also groups with

apparently machine-generated names. To find out whether these data are related to phishing or malware, a good approach is to correlate them with well-established third-party blacklists. And the conclusion is that in a number of cases when a domain appears on these blacklists, our data reveal additional related domains which are thus likely to belong to the same group of malicious actors and can be expected to be as malicious as the originally blacklisted ones.

In the experiment described in what follows we took data from our feed starting from 1 September, 2019, and have correlated them with the last few weeks data of three blacklists. The actual time range is from the beginning of December 2019 till 17 December. Let us take a good look at the results:

PhishTank

Phishtank is the maybe most reliable resource when it comes to fighting against e-mail phishing. Maintained by a broad and enthusiastic community, they provide accurate lists of URLs involved in phishing activity, e.g. containing phishing web pages. In the examined period we found 90 domain names which were confirmed by PhishTank to be involved in phishing, and appeared in our feeds, too. As an example, secure-access02.top had online and active phishing activity from 2 to 4 December, whereas secure-access05.top had this on 2 and 3 September. In the phishtank data set, there were no other domain names from 2 December with the substring secure-access in their name. In our typosquatting data feed, however, on 30 November there was the following group of 7 domains to be found:

- secure-access01.top
- secure-access02.top
- secure-access03.top
- secure-access04.top

- secure-access05.top
- secure-access06.top
- secure-access07.top

As these have appeared on the same day in the domain-name system, they are very likely to belong to the same actor, and even the name suggests that they might have been all registered for phishing. Of course, further investigation could be done by using, say, the WHOIS API or the Domain Research Suite to gain further confirmation. Nevertheless, it is apparently a good idea to add all these domains to a blacklist, thus adding five others to the original two domains. A similar situation occurs in all of the 90 cases. Amongst these there is e.g. black-oreoo-19.win, a member of a group of 410 domains on 27 November, or typosquatting versions of PayPal like pavypal.com, whose typosquatting data feed also contains significantly more than the raw PhishTank data set.

URLhaus Database

This one is run by abuse.ch to track new malware, and it is possible to browse. Using the same approach as in the case of Phishtank, we also found a number of examples where our feed reveals related domains. For instance, rmailadvert15dxcv.xyz listed there happens to be among the 4 domains on 6 December:

- bmailadvert15dxcv.xyz
- rmailadvert15dxcv.xyz
- pmailadvert15dxcv.xyz

which is unlikely to be just a random coincidence.

Joe Wein's lists

joewein.de LLC is a software company based in Tokyo, Japan, which specializes in solutions to address spam and online fraud. Joe Wein, an excellent developer and anti-spam activist, and his colleagues are doing a really good job. They are the main data provider for SURBL, a respected source of URI reputation data. By correlating Joe Wein's list with ours and using the same methodology, in the given period we found 63 domains which were there on the blacklist and also appeared in the typosquatting data feeds set. Take, for instance, uh-prettygirls.xyz, from the December 8’s blacklist. The typosquatting data feed reveals that this is a member of a group of 14 domains, all appearing in the Domain Name System on 20 November:

- pq-prettygirls.xyz
- gk-prettygirls.xyz
- lg-prettygirls.xyz
- hp-prettygirls.xyz
- zy-prettygirls.xyz
- zq-prettygirls.xyz
- el-prettygirls.xyz
- hg-prettygirls.xyz
- go-prettygirls.xyz
- uh-prettygirls.xyz

- oa-prettygirls.xyz
- gb-prettygirls.xyz
- am-prettygirls.xyz
- bm-prettygirls.xyz

The other 13 are not on the actual blacklist, but if since they were registered in the same burst on a single day and one of them was found malicious, it is likely that the others had been registered for the same purpose by the same actor. Thus, we extended the list of suspicious domains by 13 items.

Quite apparently, the new feed together with certain domain blacklists can enhance the detection of suspicious domains. And those get revealed very early, even at the phase when the potential attack is just being prepared. All you need to do is to check which of the blacklisted domains appear in the typosquatting data feed, and add the members of the groups they belong to. Our unique and innovative new data feed is also, besides its other applications, an enabler in the fight against malware, phishing and spamming. Beef up your IT security system now.