Tracing Covid-19 domain name clusters on the Internet

Alexandru Cristian Olteanu
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
a.olteanu@student.utwente.nl

ABSTRACT

During the Covid-19 pandemic, many new Internet domains appeared which contain keywords related to the pandemic. While part of these domains are just legitimate information sites, a significant number of them are related to all sorts of malicious activity, from disinformation, to phishing, the spreading of malware, and more. Given a dataset from OpenINTEL[14], which includes a set of internet domain names containing Covid-19-related keywords, we will try to find traits to identify trends or common actors that register and run these domains. We determined that the majority of domains were created during the first wave of the pandemic and 28% of domains are malicious. These findings show that it is more demanding than ever to determine which domains are used for legitimate purposes during the Covid-19 pandemic.

Keywords

Covid-19, domain names, DNS, phishing, misinformation

1. INTRODUCTION

As of 2021, the internet has become one of the main sources of information for the majority of people. People are gathering information they need regarding what places to visit, what specialists to consult or hire and even what health treatments to take[5].

Cybercrime is the illegal activity of gaining access, modifying the behaviour or deleting information on information systems of third parties. On the other hand cyber fraud is a subset of cybercrime that involves activities such as phishing, developing malware, scamming, auction fraud, credit/debit card frauds, identity theft, stock market manipulations, investment and pyramid schemes and digital extortion. As more and more people are joining the web, the prevalence of online threats and scams is increasing daily.[1]

During a global disease outbreak, substantial number of new websites are created, websites that sometimes are malicious and they produce situations in which people can get tricked into revealing sensitive information through phishing techniques[3] and probably even more importantly people can get misinformed[6] about crucial aspects regarding their health and the health of people around them.

As the Covid-19 pandemic started the incidence of phishing attacks have skyrocketed up to 220%[3] in March and April 2020, given that the cybercriminals knew the perfect moment to take advantage of people's increased usage of the internet and their reliance on it would be now.

At the same time we already observed that private companies are trying to fight[8] the enormous number of Covid-19 related websites that get created[7] and even the lawmakers[13] are doing their job in trying to stop the situation. This gives us an indication that further research into the problem has to be developed.

Although similar studies[10] have been carried recently in which researchers analyzed the correlation between the Covid-19 outbreak with the new registered Covid-19 domain names and they also tried to find the purpose for which people register these domains; their data includes information from December 2019 to the end of September 2020, so no insight in the second wave of the Covid-19 outburst, which at the moment is making a tremendous amount of victims in Brazil and India.[2] Furthermore we have a bigger keyword set related to the Covid-19 outbreak provided by ICANN, which was specified as a limitation in a previous study[10].

The main research question is:

What are the typical traits between internet domains that contain keywords that could help identify trends or common actors that register and run these domains?

The question can be answered with the following subquestions:

- RQ1. Which of these domains are potentially malicious?
- RQ2. What clusters can we identify based on geolocation or top level domains?
- RQ3. What happens over time with the Autonomous System Number Clusters (ASNC-s in short)?

2. DATA

The data was gathered from OpenINTEL[14], which is a project developed[15] to track and monitor the creation and deletion of new domain names across the Internet and many more aspects. In total we collected 22GB of data. Data that contains information such as: top level domain, domain name, IPV4 address, IPV6 address, prefix for both type of IP, country and autonomous system number. An Autonomous System is one of the independent networks that combined make up the Internet. An autonomous system number is a unique number given by

35th Twente Student Conference on IT July 2nd, 2021, Enschede, The Netherlands.

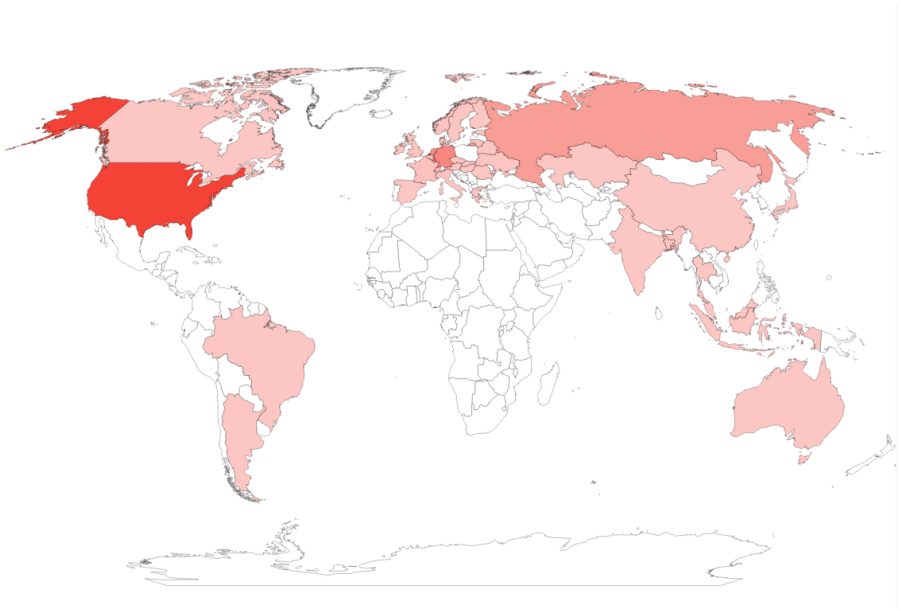

Figure 1. Server locations

IANA used in combination with the Border Gateway Protocol which identifies a network under a single technical responsibility(e.g. 15169 = Google LLC). The data spans from January 2020 to May 2021.

3. METHODOLOGY

3.1 Detecting malicious websites

In order to detect which of the provided domains were malicious, we had to first preprocess the dataset. We choose a random day worth of data that included all domains still active at 21st of April 2021. The total number of entries was approximately 880K. After removing the duplicates, because some domains had multiple servers hosting them, the number of remaining domains was 322K.

A scan using VirusTotal, which is a standard tool for malware detection of websites present in similar studies[10, 12], will be performed to see which websites from these clusters are developed with malicious intent and if there is one group/cluster or multiple that have meaningful results that can be used further for prediction purposes for domains with Covid-19 related keywords. We first had to connect via an API to HybridAnalysis, which internally calls VirusTotal to analyze the domains. The process took two weeks, because of the limited quota of 2000 requests per day and because one week was for requesting the API to analyze it and and another week was used to check the results.

3.2 Cluster based on geo-location and tld-s

In order to understand the data better, we decided to analyze the geo-location of the servers where the domains were hosted. The data for the server's location for each domain was already included in the dataset from OpenINTEL. The same applies for the TLD. The only additional work was to find the right Python libraries to generate the visualisations.

3.3 Autonomous system number clusters across time

Another conducted experiment was to evaluate what happens over time with the ASNC-s in order to determine which was the period when the majority of domains were created. For this task a cluster of 16 servers was used using Apache Server, which makes processing 22GB of .csv files significantly faster. After loading the data in parallel in Hadoop, we had to take a snapshot for each domain at each 1st of the month for the entire period from Jan 2020 up until May 2021. Unfortunately because of time constraints, we took the decision to only have two snapshots: one for 1st of April 2020 and the other one on 1st of April 2021.

Table 1. Overview of detection

Category detected	# of domains	percentage (%)
Phishing	47	<0.01
Malicious	88868	28.07
Malware	5	<0.01
Pending	4	<0.01
No specific threat	218320	68.96
Undetermined	9351	2.95
Total	316595	100

First step was to standardize the domain names from the dataset by removing the *'www'* subdomain in the beginning of their names and to remove duplicate domain names if they have the same host. Essentially we simplified the complexity of the analysis by considering only one variant for each domain name, so if there exists a *www.coronavirus.org* domain then we will not have a new entry for *coronavirus.org* if it has the same host.

Second step was to merge in a new Apache Spark dataframe the ASN from the initial date and the final date, for each domain that exists in the dataset at both times and then group by ASN and count each group for both dates. The key point is that we selected only the rows of counted ASN-s that were top 10 biggest ASNC-s for both the initial date and the final date, which is necessary in order to be able to draw conclusions from a visualisation. At the same time these top 10 ASNC-s account for 65% of migration across clusters.

Finally we generated a Sankey diagram as you can see in Figure 2 that we will analyze in more depth in section 4.3.

4. RESULTS

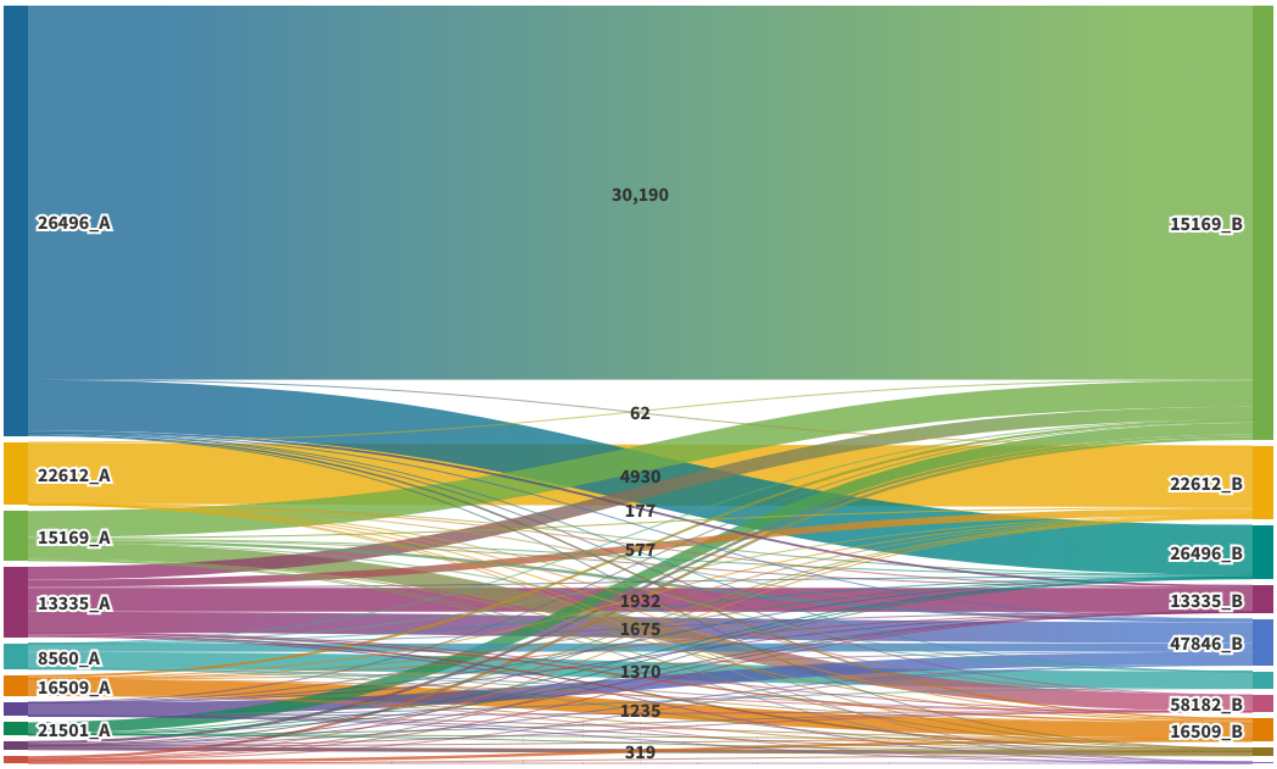

Figure 2. Autonomous system number clusters transition across time

4.1 Detecting malicious websites

Results are summarized in Table 1. The Phishing category contains all domains that were detected as a type of attack which relies on the target to click on a malicious link or run an executable after it was tricked using social engineering in many different variations. VirusTotal classifies a URL in the malicious category if it could not find a more specific category (e.g. malware, phishing). An URL is classified by VirusTotal as being malware if there is an executable hidden in the page. Pending is the category that contains all the URL that could not be analyzed in a timely manner at the moment of publishing this article. The *No specific threat* category contains domains, as the name implies, that VirusTotal considered not to be a threat for the users. Undetermined category is for the domains that our scanner placed in a grey area and needs further manual check.

The results are worrying given the big percentage of malicious detected websites. Even though the method is consistent with other studies that use multiple vendors of VirusTotal[12], some false positives might have been scattered among the results. One reason for vendors to detect some domains as being phishing could be the naive way of using keyword finding. Further research is needed to clarify this aspect, but the **RQ1** has been responded with these results.

4.2 Cluster based on geo-location and tld-s

And as you can see in Figure 1, the majority of the websites are hosted by servers located in United States of America. The results for geo-location do not give us enough information, because many cloud-hosted domains will automatically geo-locate to US, which also seems to be the case in our visualization. Further research is needed to find different results.

The domains were hosted on ccTLDs, but also on a big variety of new gTLDs as we can observe in Figure 3. The bigger a word is in the figure means that more occurrence have been found in the dataset. As we already expected *.com*, *.org* and *.net* are prominent, but we can also observe the large groups of domains in the new gTLDs such as *.online*, *.store*, *.shop* are present and quite prominent. This give us an indication that new gTLDs are both used for informing people, but also to deliver them malware and spam similar to a previous study[11]. These results address **RQ2**.

Figure 3. Word cloud for top level domains

4.3 Autonomous system number clusters across time

The hypothesis before conducting the experiment was that various migrations will happen during the time span of one year in the ASNC-s because new websites related to the Covid-19 pandemic were created during this period, but also because after a period of one year typically domains

are either renewed or are being moved to a parking service and respectively to a new ASNC. The reason why there are 13 categories in Figure 2 is because the rightmost ASNC-s (197695, 58182, 40034) were only present in top 10 biggest ASNC-s on the final date(1st of April 2021) and not during the initial date.

The biggest migration seen in Figure 2 from 26496 ASNC (GoDaddy.com LLC) to 15169 ASNC (Google LLC) is explained by the creation of a hefty amount of new domains related to the Covid-19 pandemic during the end of March and the entirety of April 2020, which also coincides with the peak of the first wave of the Covid-19 pandemic, result which is consistent with a previous study[10]. Even though this result is not offering completely new insights, it does respond to **RQ3**.

5. DISCUSSION

5.1 Limitations

One of the biggest limitations of this study is the appearance of false positives in the domain names. Because the provided keywords from ICANN include words such as 'virus', this can lead to domain names such as 'zikavirus.ch', which most probably contains information about Zika virus and not about the SARS-CoV-2 virus or the Covid-19 disease or the pandemic. Another limitation is that we did not have time to actually verify the content of the websites, which most of the time is more relevant than the domain itself, but we do allow this opportunity for further studies that want to go deep into the problem.

5.2 Ethical considerations

We tried to only use data that does not personally identify people and at the same time we tried to send a limited amount of requests for the API-s we used, but also to use a limited amount of servers for the Apache Spark cluster in our department.

6. RELATED WORK

In order to gather related literature Google Scholar was used. The used search terms were "pandemic", "domain names" and "covid-19".

While there was an increase in traffic on the internet[4] since the pandemic started, some meaningful studies were carried about the big number of Covid-19 domain names that were created.

In one study[10] research was carried by using a large-scale Internet domain name database to look for domain names related to Covid-19. 260 million distinct entries were collected with 1600 distinct top level domains. The following research questions were asked RQ1: Is the number of Covid-19 domain names registrations correlated with the Covid-19 outbreaks? and RQ2: For what purpose do people register Covid-19 domain names? The results were convincing and there is indeed a growth in the number of Covid-19 domain names registrations interestingly, preceded the Covid-19 pandemic by about a month.

The results were convincing as the answer for the second research question showed us that 70% of active Covid-19 domain names websites provided useful information, such as health, tools or product sales related to Covid-19 and a non-negligible number of registered Covid-19 were used with malicious intent. So the aim of our research is to find further trends and even common actors that register these domains. The way we are doing these in a different way is presented in the Problem Statement section.

Previously in another study[9] research showed that we can identify malicious URLs of Covid-19 pandemic using machine learning techniques. The reason behind this study is consistent with the ones mentioned before. Using a large volume of open source data and a tool to generate feature weight, a machine learning model was trained. The tests have shown that the method is a promising mechanism to mitigate Covid-19 related threats, which offers us a guarantee that domain names can offer meaningful information about its malicious property.

Compared to previously mentioned studies, our study is the only one that we are aware of that is using more than two, three words as keywords in the identification of the Covid-19 domains. We use 276 words provided by ICANN as a keyword set which offers more results, but also has some limitations as explained in Section 5.1.

7. CONCLUSION

After the analysis of 316K domains that contain keywords related to the Covid-19 pandemic, we have found that 28% of them are malicious. Additionally we discovered that a very big portion of the websites were created during the end of March and the entire month of April 2020. We conclude that there is a considerable number of websites having malicious intents, another is that it is getting harder and harder to identify and differentiate legitimate domains from the ones with bad intent.

8. FUTURE WORK

One thing that needs to be kept in mind for such a person or group of persons that what to pursue in verifying the content of the websites as explained in Section 5.1 is the fact that a scraping approach can be taken, but it should be taken in combination with a more manual approach of inspecting the exact content of the websites, because sometimes images and similar digital information does convey information that algorithms are not able to reliably identify.

9. REFERENCES

[1] F. P. Bernat and N. Godlove. Understanding 21st century cybercrime for the 'common' victim, 09 2012.

[2] N. Bhowmick. How India's second wave became the worst COVID-19 surge in the world, 05 2021.

[3] M. Bitaab, H. Cho, A. Oest, P. Zhang, Z. Sun, R. Pourmohamad, D. Kim, T. Bao, R. Wang, Y. Shoshitaishvili, A. Doupé, and G.-J. Ahn. Scam pandemic: How attackers exploit public fear through phishing. *ArXiv*, abs/2103.12843, 2021.

[4] T. Böttger, G. Ibrahim, and B. Vallis. How the internet reacted to covid-19: A perspective from facebook's edge network. In *Proceedings of the ACM Internet Measurement Conference*, IMC '20, pages 34–41, New York, NY, USA, 2020. Association for Computing Machinery.

[5] M. Dimock. The role of the internet, 09 2011.

[6] R. Gallotti, F. Valle, N. Castaldo, P. Sacco, and M. De Domenico. Assessing the risks of "infodemics" in response to covid-19 epidemics. *medRxiv*, 2020.

[7] C. Glover. Coronavirus-Related Domain Registrations Rise 6,000 in a Week, 03 2020.

[8] A. Hardaker. How Nominet is using AI to stop fake COVID-19 sites, 04 2020.

[9] J. Ispahany and R. Islam. Detecting malicious urls of COVID-19 pandemic using ML technologies. *CoRR*, abs/2009.09224, 2020.

[10] R. Kawaoka, D. Chiba, T. Watanabe, M. Akiyama, and T. Mori. A first look at covid-19 domain names: Origin and implications. In O. Hohlfeld, A. Lutu, and D. Levin, editors, *Passive and Active Measurement*, pages 39–53, Cham, 2021. Springer International Publishing.

[11] M. Korczynski, M. Wullink, S. Tajalizadehkhoob, G. Moura, and C. Hesselman. Statistical analysis of dns abuse in gtlds final report. 2017.

[12] P. Peng, L. Yang, L. Song, and G. Wang. Opening the blackbox of virustotal: Analyzing online phishing scan engines. In *Proceedings of the Internet Measurement Conference*, IMC '19, pages 478–485, New York, NY, USA, 2019. Association for Computing Machinery.

[13] S. Sabin. In Fight Against COVID-19 Scam Sites, Lawmakers Push for Domain Name Ownership Records, More Accountability, 06 2020.

[14] A. Sperotto, M. Jonker, O. van der Toorn, and R. van Rijswijk-Deij. Active DNS Measurement Project, 04 2015.

[15] R. van Rijswijk-Deij, M. Jonker, A. Sperotto, and A. Pras. A high-performance, scalable infrastructure for large-scale active dns measurements. *IEEE Journal on Selected Areas in Communications*, 34(6):1877–1888, 2016.