Blacklist, do you copy? Characterizing information flow in public domain blacklists

Jesse van der Velden
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
j.vandervelden@student.utwente.nl

ABSTRACT

In this paper, we will analyse the information flow of public domain blacklists. Various vendors maintain a list of public domain blacklist to prevent access to domains containing malware, phishing, and counterfeit/ fake webshops. Both malware and phishing can have a disastrous impact on society when critical companies or infrastructure are affected. We will explore the information flow in public domain blacklists to make good decisions which blacklist to use, to prevent access to as many malicious domains as possible and not prevent access to benign domains. Research into the overlap between blacklists was already a focus of a couple of studies. However, there was not much attention into the information flow between blacklists, and if there are occurrences of blacklists that copy from each other. We created several metrics to identify occurrences of copying behaviour of blacklists: we will do a pairwise comparison using data from crawled public domain blacklists, looking at intersections, correlations, and finding interesting overlapping domains. In this research, we have identified that it is indeed possible to show that some blacklists copy from another blacklist. We verify this by using data from blacklists which openly mention that they copy from another blacklist.

Keywords

public domain blacklists, information flow, copying of blacklists

1. INTRODUCTION

Malware, botnets, phishing, and webshops selling counterfeit products are important problems on the web these days. As browsers and operating systems are not always up to date especially in corporate environments or poorly designed Internet of Things devices [4, 7], those computers become an easy target for malware or botnets. Phishing is also a problem that malicious actors use to gain access to online accounts or bank accounts with privileges. These can have a disastrous effect on society when critical companies or infrastructure are affected [7, 15]. What malware, botnets, phishing, and fake webshops all have in common, is that they are largely served through the internet by domain names or IP addresses. Browser developers, email providers and applications, and other software developers use public domain blacklists to reduce the number of successful malware and phishing attacks. This is done by preventing access to navigate to those domains on a blacklist that may contain malicious/ harmful content. Choosing a good blacklist is essential for preventing access to malicious domains, as there are many public domain blacklists from different vendors. There is still more information needed to determine the information flow between blacklists because it will help and support the choice for a particular blacklist. Therefore, we will analyse the information flow between blacklists in this research project.

32th Twente Student Conference on IT Jan. 31th, 2020, Enschede, The Netherlands.

1.1 Research Goal and Questions

The objective and goals of this research are to get a better overview of the information flow in public domain blacklists. This will be done by data analysis (the domain listed, the timestamp when it got listed) from public domain blacklists. For example, finding similar domains across multiple blacklists and keeping track of the time when they were added. For each blacklist also the activity (number of additions and removals) will be measured. In the end, the usability of a particular domain blacklist can be determined with the metrics defined in Section 3.

By being able to tell if one blacklist copies from another, one can determine which blacklist lists first the malicious domains.
The main research question to answer is: **"Do blacklists copy from each other?"**

To get an answer to this question, we will explore also the following sub-question:

1. What is a good metric to identify copying-behaviour for blacklists?

Online results and charts:.

The results will come available in the research. We aim at the reproducibility of our research by sharing the code that will be used for the data analysis in this research on GitHub. Additionally, there is also the aim to also publish abbreviated daily statistics data that is obtained during this research on GitHub. A site using HTML and a JavaScript charting library is made to provide a dynamic insight into the results of this research. The results will consist of several tables that compare the amount of overlap pairwise intersections between two different blacklists, and a conclusion can be made about which metrics perform the best at identifying copying behaviour.

Results are available at:

https://megacookie.github.io/blacklist-research

The paper is organised as follows: Section 2 provides an overview of the background of blacklists and relevant literature is reviewed. In Section 3, we will explain the methodology of this research and the data set is explained. Section 4 provides a detailed analysis of the data using the defined approaches in Section 3. Section 5 will close the research with a conclusion, and we will discuss Future Work in Section 6.

2. BACKGROUND

2.1 Blacklist Vendors

There are private and public domain blacklists, where the public domain blacklists can all accessed by whomever it wants. The private blacklists, on the other hand, require some kind of subscription to access the data. Some examples of public blacklists are: VXVault, URLHaus, Squid-Blacklist and some examples of private blacklist are from antivirus vendors such as Bitdefender, ESET, Norton, and McAfee. For this research, only publicly available blacklists will be used. The different vendors of the blacklists all have different details on which they focus. Some are specifically related to spam campaigns by email, other vendors can have a focus on malware or phishing only, while other vendors list all malicious domains in their blacklist. An effective way to obtain malicious domains is to set up a so-called email spam account honeypot [19]. This email account will receive lots of emails as they were an easy target for spam senders in the past, and as such these emails can be automatically scanned on its content. Most of the times there is a certain pattern in spam-related emails that can be used to identify a malicious domain or IP address.

Blacklists will list malicious domains, while at the same time they must not list benign domains. Otherwise, it can cause an undesired consequence of not being able to access that domain. This is the result of false-positive domains on a blacklist. With a so-called *ground truth*, somebody can identify false-positive results. The ground truth will also be of interest during the literature review.

2.2 Review of literature

Scientific studies into the blacklists is not a very well-researched area, but the overlap of blacklists was already a focus of a couple of studies. However, there is not that much research into the characterization of the information flow between blacklists.

Jhaveri et al. (2017) [9] did a case survey into abuse reporting of domain names. As there are multiple parties involved, it describes the relation between and incentives of security companies and the operators of internet infrastructure (such as hosting companies). It also explains how abusive data can be contributed eventually distributed, and the delisting process to remove domains from a blacklist. Besides they also discussed the incentives to operate a blacklist. This is also important to know because for choosing a blacklist the incentives of the vendor can play an important role. Some do it for a moral duty, or for protecting their customers. Many ISP's or large autonomous systems also make use of blacklists for botnets to prevent large amounts of traffic that are used for DDoS attacks by botnets.

Pitsillidis et al. (2012) [19] analysed email spam and email spam data feeds. This was done by having some honeypot email accounts where they receive lots of spam. Blacklists were also used as a feed. Spam on a honeypot was then analysed for its content and all the links in the email were checked. This was done by checking the purity, coverage, proportionality, and timing. They check for each feed if the listed domain is active, if it is listed across multiple blacklists using pairwise comparisons, the ratio of spam detection across a feed, and how fast each feed was with the listing of a malicious domain. To check for false positives, they use Alexa's top list and the Open Directory lists. When a site is in such a list, it is considered to be a benign domain.

Kührer et al. (2014) [11] analysed 15 public domain blacklists and 4 private blacklists by AV vendors. They looked at the completeness of a list and analysed the domains itself on parked domains and sinkholed domains by looking at DNS entries and HTTP responses. To check the completeness coverage ratio of a blacklist it was checked against SANDNET, pDNS, and VirusTotal.

As this research is mostly a data analysis study on domain lists, such research was also done by Scheitle et al. (2018) [22]. They did data analysis on several top lists. It compares the top lists from Alexa, Umbrella and, Majestic. This dataset contains millions of domains that are processed every day. They looked at the significance, the structure, stability, ranking mechanisms of the different top lists, where they look at the impact of top lists in the academic world and other online places. What is of interest for this study as well is that they measured daily changes and intersections between top lists. This is similar to finding changes and pairwise intersections of blacklists. The researchers even open-sourced their toolset and code to let everyone verify their results. The code is using Python Pandas and NumPy for the comparisons in the dataset. It can also be used for other large domain lists, such as blacklists as it is the same kind of data structure. As such a similar tool can be more easily made for usage in blacklists instead of starting from scratch. Even to this day (January 2020) their site automatically reports daily updates of the comparisons of these data sets.

3. METHODOLOGY - TOWARDS FINDING COPYING BEHAVIOUR

In this Section, we will discuss the data set and the metrics for this research.

3.1 Data set

There are several distinct sources of public domain blacklists analysed. This research only used data that was crawled from public available blacklists and does not contain any non-publicly available blacklists. Those blacklists were crawled every day with some small exceptions. The maximum frequency of crawling was one day at almost at the same time. The earliest crawl is going back to the 6th of July 2016, where just six blacklists were crawled at the time. The last crawled data for this research came from the 20th of November 2019. During this period, at the 10th of October of that 2016, two more blacklists were added to be crawled for in the data set. In October 2017, one blacklist was added and on New Year's Eve 2018 six more blacklists were added. In March of 2019, nine more blacklists were added. A summary can be found in Table 1 and Section 4.2. In the end, twenty-four blacklists have been crawled, where two blacklists have been stopped crawling earlier as those blacklists were being discontinued.

The scope of crawling those public domain blacklists is not covered in this research and was given by members of the DACS group of the University Of Twente which started crawling since July 2016.

The format of the data that was crawled from every blacklist had the same structure. A file with the day name was created. On each line, a different domain is listed with a separator that separated the domain and all the blacklists that listed that domain. Furthermore, an additional separator was added that put redundantly the date of crawling.

3.2 Measurement of the data set & Tools

As the data set was not in the right format to do statistical analysis, it had to be converted for the different metrics that were defined during this research. The different day to day files were parsed and the number of domains for each blacklist was calculated. This result was added to a table with the date as the index and in each column the number of domains per list. This schema was designed to perform statistical measurements of the blacklists that were researched which can be seen in Table 1. The tools that will be used in this research will be on data analysis. Python in combination with Pandas is used to make the pairwise comparisons between the different blacklists and make graphs out of the data.

We will perform comparisons between the several lists. In particular by looking at the pairwise intersections of the different blacklists using the tools described. To be able to tell if one blacklist copies from another, we will use metrics so the listed domains will be checked for intersections on a two-day basis between two lists.

3.3 Metrics

To find out if there is indeed copying-behaviour between one blacklist to another blacklist, we will define three different approaches. A good first metric is to identify if a list is included by another list. Therefore, all the number of common domains between all the possible combinations of two blacklists will be calculated. A smaller blacklist can be included by a larger blacklist. This way, if a blacklist is included by another blacklist for a consecutive time period the percentage of common domains must be a high number (or of course 100% if *every* domain is listed by another bigger blacklist). Choosing 80% as the offset will be high enough to make sure that copying can be determined, but also allows room for the miss-measurements in the data set, as the frequency of updates of blacklists might not match.

Next, we will define a metric to look at the number of domains for a certain time window. We will perform a statistical analysis by using the Pearson coefficient, commonly known as Pearson r correlation coefficient. This way we will measure the correlation of two blacklists. The correlation coefficient will be high if the same order of the number of domains amongst two blacklists is similar, and will be low if the number of domains is dissimilar.

Lastly we can compare the number of added and removed domains of a blacklist. If one blacklist is always listing about the same new domains or removing about the same domains but one day later than another blacklist this can be also seen as copying-behaviour.

3.3.1 Verification of the metrics

We will perform verification by using the blacklists that publicly announce that they copy from another blacklist. This is also described in the next Section with brief descriptions of each blacklist that was used in this research. In the end, also non-publicly announced copying-behaviour of blacklists can be identified. For this verification, we will use the blacklists of *DShield*, *SquidBlacklist*, and the *University of Toulouse* as those blacklists copy from another blacklist that is also crawled in this research. An important caveat that must be considered is: several independent blacklists might have the same methods to come to the same list of domains for a blacklist. A case where it might happen is of two vendors with a focus on spam in a particular area of the world. They both set up a honeypot, and they will largely receive the same spam emails. This might lead to the conclusion that those lists might copy from each other, in particular when there is a large overlap between the two blacklists with a certain time delay. However, it could also be a coincidence as one vendor just updates its list more frequently than the other blacklist vendor.

Lastly, we will do a comparison of these three metrics.

4. ANALYSIS

In this section, we analyse the data set using our defined metrics. We first begin with some general observations and a description of each blacklist. Next, we will analyse our three defined metrics on the data set.

4.1 General observations

There can be some general observations be concluded by looking at Table 1 and Figure 4.1. The observations that stand out are:

1. Some dates do not contain any data, either to malfunction of the crawling program or network issues. However, this should not influence the results that much as every blacklist has plenty of data points to make some general conclusions, as will discussed during the in the next Sub-Section.

2. There are two blacklists stopped crawling. At the end of this research, even some more blacklists have been discontinued as seen as a stable line in the graph. This was due to either a specific focus on just one malware that has stopped or because they were succeeded of more general blacklists of the same vendor or the blacklist was being split up into different categories.

3. The number of domains differs a lot even within some blacklists, the graph is presented in logarithmic scale to have a meaningful comparison and a compact graph.

4.2 Analysed Blacklists

In this Section, we will perform a detailed analysis of the gathered data. First, every blacklist will be described with details about the blacklist itself and the number of listed domains.

C2-domains is a list from Bambenek Consulting [5] that is listing all command and control domains that are used in the major malware threats. Crawling started recently. It has a relatively stable number of domains over time by looking at the statistics, the quantiles in specific. It can, however, have quite some variance. Looking at the graph it shows that since 21nd of March 2019 it is relatively stable in the number of domains listed, but it still has frequent updates.

Table 1. Data set: dates (in *Y-m-d* format), mean and standard deviation of the number of domains listed ($\mu \pm \sigma$), quantiles (Q_1, *Median* (Q_2), Q_3), minimal number of domains (*min*), maximal number of domains (*max*), mean of daily change (μ_Δ)

List	Crawled	$\mu \pm \sigma$	Q_1	Q_2	Q_3	min	max	μ_Δ
C2-domains	2018-12-31 - 2019-11-20	617 ± 308	690	720	747	26	1881	2
CyberCrimeTracker	2018-12-31 - 2019-11-20	10192 ± 220	9997	10125	10331	9923	10702	2
DNS-BH	2018-12-31 - 2019-11-20	23062 ± 12	23053	23062	23073	23033	23084	0.05
DShield	2019-03-21 - 2019-11-20	2032 ± 3	2030	2033	2033	2022	2043	0.03
hostfile	2016-07-08 - 2019-11-20	81266 ± 78304	13203	42883	160175	100	209037	187
hphosts	2016-10-01 - 2019-11-20	252680 ± 45850	248874	248878	248878	6715	396314	156
JoeWein	2016-07-08 - 2019-11-20	1151 ± 672	770	893	1344	387	5666	-2
Malc0de	2016-07-08 - 2019-11-20	80 ± 75	36	49	105	5	333	-0.03
MalwareDomainList	2016-07-08 - 2019-11-20	896 ± 48	860	893	906	73	994	-0.03
OpenPhish	2017-10-28 - 2019-11-20	1279 ± 806	823	985	1332	428	5181	-8
Ponmocup	2019-03-21 - 2019-11-20	96 ± 7	89	96	101	83	107	-0.07
RansomwareTracker	2016-07-08 - 2019-11-20	1493 ± 329	1462	1664	1667	23	1668	1
SquidBlacklist-Malicious	2019-03-09 - 2019-11-20	130401 ± 44118	89245	138570	152788	299	229743	724
ThreatExpert	2016-10-01 - 2017-12-14	251 ± 10	244	248	255	231	283	-0.12
Toulouse-Crypto	2019-03-21 - 2019-11-20	8363 ± 1635	7191	7598	8327	6889	11314	-0.3
Toulouse-DDoS	2019-03-21 - 2019-11-20	285 ± 4	278	287	287	278	287	0.03
Toulouse-Malware	2019-03-21 - 2019-11-20	3140 ± 1308	2883	3000	3777	879	6449	4
Toulouse-Phishing	2019-03-21 - 2019-11-20	25491 ± 28878	3377	3845	63251	1019	63253	-247
URLhaus	2018-12-31 - 2019-11-20	54035 ± 12255	44793	58648	62032	29627	72306	131
URLVir	2019-03-21 - 2019-11-20	402 ± 259	242	278	427	173	1285	-2
VXVault	2018-12-31 - 2019-11-20	61 ± 16	47	59	75	31	95	-0.07
ZeusTracker	2016-07-08 - 2019-07-08	362 ± 27	343	353	371	336	431	-0.06

CyberCrimeTracker [6] is a recently started crawled list that is also listing different domains that are or were used for command and control malware servers and distribution. Looking at the statistics and the graph, the blacklist is only adding new domains that were of the result of malware analysis.

DNS-BH is also a recently-started crawled blacklist that is maintained by RiskAnalysis [21] that also focuses on malware. The list is relatively stable as seen by the low standard deviation. This list also removes domains, in contrast to *CyberCrimeTracker*, when they are not used for malware anymore or are offline, as can be seen by the graph. This list is also being used by VirusTotal

DShield is also a relatively new crawled blacklist from the Internet Storm Center [8] that aggregates malware results from every security researcher who likes to contribute. The list itself is stable with a low standard deviation. As seen by the quantiles and the graph, it is not that active in listing or removing domains. They say they are listing domains from also *RansomwareTracker*, *DNS-BH*, *MalwareDomainLists*, *ThreatExpert* and *VirusTotal*. It would be interesting to see with the analysis of the common domains if this is the case. Looking at the graph, the blacklist seems to be stable since the 10th of July 2019. This list is the successor of *ThreatExpert*.

hpHosts is a blacklist from HostFile that is powered by MalwareBytes [13]. This list was crawled from 2016 till the end. However, the list was stable since March 2018. So the analysis is only measured until 2018. Looking at the graph, it has some predefined dates were lots of new domains were added and removed, but the rest of the time it stayed relatively stable.

hostfile is a subset list from *hpHosts* of the same vendor HostFile [13]. This list contained more frequent updates, and later on, was transformed to be the main list to be maintained after March 2018. Looking at the graph, it seems like they removed once in one/ two months lots of domains to less than 1000. In the following ten - twenty days it increased to more than 10 000 domains. Since March 2018, they did not do this anymore and the list steadily increased from then on. Since the end of August 2019, the number of domains did not increase as the full list is deprecated and it is now split into sublists which were not crawled during this research. It is also being used by VirusTotal.

JoeWein is a list made by the anti-spam activist Joe Wein [10]. It was crawled from the beginning and it lists largely domains used by email spammers and phishers. It is used by other companies and email providers. By looking at the statistics and the graph, it shows lots of fluctuations in the number of domains, which can also be seen by the large standard deviation and large differences in the quantiles and min. and max. number of domains listed.

Malc0de [12] is also a blacklist that is focused on malware. Looking at the graph and the statistics, it had some inactive periods. At the beginning of 2017 and from June 2019 the list seemed to be stable. Besides, the list fluctuated a lot between 36 and 105 domains but it had some highs of 333 domains listed. It is one of the smaller blacklists that is also used by VirusTotal.

MalwareDomainList is a blacklist that lists malware [14]. Looking at the graph and the statistics, it is a stable blacklist with not that much updates. It raises the question if it is actively maintained or that there are not that much new malware families since February 2019. This list is also being used by VirusTotal.

OpenPhish is a blacklist that lists largely domains of phishing sites [17]. It is updated twice a day for free usage, or every 5 minutes for paid subscriptions with more details about the listed site. This blacklist was added in October 2017. It also lists the path of the URL, which is truncated for this research. Every 14 days, domains that are offline are removed from the list, which can also be observed by looking at the graph. This was also stated at their website. Looking at the statistics and graph, it shows it is a highly active list that contained between 1200 and 5000 domains

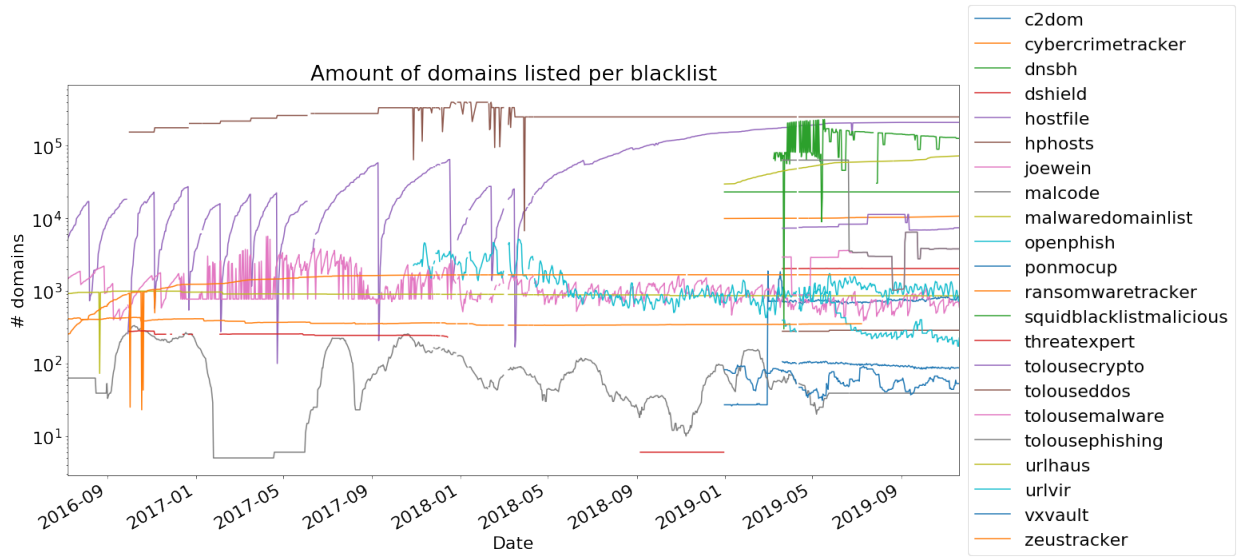

Figure 1. The number of listed domains per blacklist. Details can be viewed on GitHub.

till June 2018. From then on, it contained between about 450 and 1700 domains.

Ponmocup is a recently added blacklist maintained by DynDNS.org [18] that contains a small set of domains listed. Looking at the graph, it is relatively stable with the number of domains listed, but it is still updated regularly. It contains the domains of the still active Ponmocup botnet/ malware. It seems it is a casual cat and mouse game between the operators and the blacklist vendor as new domains are added regularly.

RansomwareTracker was a blacklist maintained by Abuse.ch [2] that was listing malware domains. It was crawled until the end but was discontinued listing domains in December 2019. Looking at the data, the number of domains was almost always steadily increasing and was relatively stable from 2018. This can also be seen by looking at the last two quantiles.

SquidBlacklist-Malicious is also a recently added blacklist [23] that aggregates domains from *DNS-BH hostfile/ hpHosts*, *C2-domains*, *DNS-BH*, *CyberCrimeTracker*, *OpenPhish*. The number of domains moves with the domains listed by those other providers. It will be of particular interest to see if our metrics will identify all the aggregated blacklists.

ThreatExpert was also a discontinued list of the Internet Storm Center. It also contained data from *MalwareDomainsList*, *DNS-BH*, and from the also discontinued Abuse.ch trackers: *RansomwareTracker* and *ZeusTracker*. The successor is *DShield*. Looking at the graph and statistics it was a not that large list, with a low standard deviation.

Toulouse-Crypto, **Toulouse-DDoS**, **Toulouse-Malware** and **Toulouse-Phishing** are recently added blacklists from the *Université Toulouse 1 Capitole* [20] that aggregates and categorizes domains from *Abuse.ch* (*URLhaus*, *ZeusTracker*, *RansomwareTracker*), *Squidblacklist-Malicious*, *DNS-BH*, *CyberCrimeTracker* and others. Categorizing of the domains is done by hand and scripts. Looking at the statistics, only *Toulouse-DDoS* is relatively stable in the number of domains listed, while the other blacklist-categories are updated regularly but also relatively stable in the number of domains listed. Those lists show some fluctuations because on some set intervals, domains are listed or removed in larger numbers with small updates in the meantime.

URLhaus is a relatively large and a active blacklist from *Abuse.ch* [1] that lists domains for malware distributions. URLhaus lists entire URL paths to the malware, and new URLs can easily be added by contributors. Looking at the statistics and graphs, the number of domains steadily increases. This list is also used by email providers (from the same company's Spamhaus) and Google's Safe Browsing list.

URLVir blacklist is maintained by NoVirusThanks [16]. It mainly lists URLs to malicious executables and was recently added to be crawled. It is updated a lot, as can be shown by looking at the statistics and graphs.

VXVault also lists malicious URLs to executables [24]. It is a relatively small set but with quite some updates over time. This list is also used by VirusTotal.

ZeusTracker is also a discontinued blacklist by *Abuse.ch* [3] it listed mainly domains that were used for the Zeus malware with a stable number of domains listed. It was updated often.

4.3 Activity Analysis - Splitting the data set

As we have been shown in Figure 4.1 the number of domains per blacklist varies a lot. It can be seen that some blacklists have a constant number of domains listed. This is because some blacklists have been deprecated, but were still being crawled. It resulted in a constant number of domains listed by a blacklist. We split up the data set to do meaningful analysis on the day to day data. This split up in the data is been gathered by looking at the day to day data, and making a list in the number of changes by a blacklist. So in specific, we gathered the data of the additions and removals of a blacklist. This way even the blacklists that appear to have the same amount of domains listed can be checked if they add and remove the same amount of domains each day. We split up the data set based on the following observations and activity analysis:

- **DShield** is been crawled from March 21, 2019, till November 2019, however, it showed to be steady in

the number of domains listed since July 2019.

- **hpHosts** is been crawled from October 2016 till November 2019. It showed it has been stable in the number of domains listed since March 2019.
- **Hostfile** is been deprecated since August 2019. It was split up in individual lists, which are maintained even to this day (January 2020), but were not included in the data set.
- **MalwareDomainList** is also been crawled from the start of October 2016 till November 2019. This blacklist does not show a change in the number of listed domains since February 2019.
- Then on New Year's Eve 2018m six new blacklists were added to the data set.
- In March 2019 ,nine more blacklists were added.

4.4 Common domains - Intersection between lists

We first study the metric of the number of common domains (intersection) between lists for each day in the data set. The percentage of the number of common domains vs. other domains is given below in Table 2. In the end 24 blacklists were crawled, so there can be 202 permutations made and even more if all the blacklists would have an overlap in time.

It is important to note that this table is aggregated of multiple iterations of this metric, because pairwise intersection comparisons only made sense when blacklists are actively maintained. There were blacklists deprecated and added during the crawling period as described in Section 4.3. The threshold in this table is set to 50%. This percentage of intersection between two lists is measured for each day that there was data available from this list. Table 2 contains the median values of these percentages.
If we look at Table 2, we observe the number of intersections between blacklists can be very high and even be 100%. This means the second blacklist listed (included) 100% of the first blacklist. There are no surprises here, all the listed blacklists in the table mention that they copy from the other blacklists. This way we proved using the first metric that it is indeed possible to identify copying-behaviour. However not all copying behaviour that is stated from the blacklists analysed can be identified. And not all of the blacklists that mention that they aggregate domains from another blacklist are 100% included. This can mean that those blacklists have additional checks and algorithms to decide which domain to list from another blacklist.

4.5 Correlation between lists

It is also interesting to look at the correlation between lists. The Pearson correlation coefficient is calculated for each combination of the crawled blacklists. An entire correlation matrix would take up much space because of the large number of combinations that can be made of the twenty-four crawled blacklists. Therefore, the results are listed in Table 3, similar to the common domains analysis table with a threshold of 75%. The results of the correlations table are not surprising, most of the blacklist combinations are publicly admitting that they copy from another list. It is, however, difficult to make conclusions about copying-behaviour from a correlation matrix. A correlation only looks at the number of domains listed and if the changes in the number of domains are more or less the same. It can, however, be a good start to identify possible combinations of copying-behaviour blacklists. We can then use the other approaches to research them further.

Table 2. Median percentage of the number of common domains between two blacklists. It shows how many domains of the first blacklist are included in the second blacklist. A percentage of 100% shows the first list is fully included by another blacklist.

Blacklist1	Blacklist2	Perc.
ransomwaretracker	dshield	100%
tolousemalware	tolousephishing	100%
ransomwaretracker	tolousemalware	100%
ransomwaretracker	tolousephishing	100%
coinblocker	tolousecrypto	99%
zeustracker	dshield	97%
malcode	tolousemalware	97%
malcode	tolousephishing	97%
vxvault	urlhaus	92%
urlvir	urlhaus	83%
dshield	tolousemalware	82%
dshield	tolousephishing	82%
malcode	urlhaus	79%
c2dom	squidblacklistmalicious	76%
squidblacklistmalicious	hostfile	76%
zeustracker	tolousemalware	68%
malwaredomainlist	squidblacklistmalicious	64%
malcode	hostfile	64%
tolousemalware	dshield	56%

4.6 Activity - Analysing added domains

Next, we identify copying-behaviour using our third metric: by looking if one blacklist structurally adds domains a day after another blacklist. We only looked for domains that were listed on one blacklist the first day, and the additions of blacklists on the second day. In this case, we could show that the following blacklists are regularly adding the same domains from another blacklist. This does not necessarily mean that one blacklist copies from each other. It just shows that one blacklist is regularly and structurally later in listing the domains that were first listed by a former blacklist.

- **OpenPhish** is adding domains from HostFile/ hpHosts. This is interesting as they do state that they are adding resources from its *global partner network* [17] but not specifically from which sources they get their domains. It seems like the blacklists from Hostfile are included by OpenPhish.
- **ThreatExpert** is adding domains from Hostfile/ hpHosts.
- **HostFile/ hpHosts** is adding domains from ZeusTracker.
- **JoeWein** is adding domains from Hostfile/ hpHosts.
- **SquidBlacklist**-Malicious is adding domains from Hostfile and C2-Domains
- **CoinBlocker** is adding domains from Toulouse-Crypto

4.7 Takeaway - Which metric to choose?

We explored the three different approaches to identify copying-behaviour. First, this was done by comparing the number of common domains listed each day between two blacklists. The median value was used to identify copying-behaviour. We have shown and verified that this is indeed possible to

Table 3. The Pearson correlation coefficient of a combination of two blacklists.

Blacklist1	Blacklist2	Perc.
hostfile	urlhaus	95%
hostfile	threatexpert	95%
urlhaus	zeustracker	95%
dnsbh	ponmocup	93%
cybercrimetracker	ponmocup	93%
c2dom	hphosts	92%
openphish	threatexpert	90%
cybercrimetracker	urlhaus	89%
cybercrimetracker	zeustracker	88%
ponmocup	urlhaus	87%
ransomwaretracker	zeustracker	87%
ransomwaretracker	urlhaus	86%
c2dom	malwaredomainlist	85%
cybercrimetracker	tolousephishing	83%
dnsbh	tolousephishing	83%
tolousecrypto	zeustracker	83%
ponmocup	tolousephishing	81%
tolouseddos	urlhaus	79%
hostfile	tolouseddos	79%
hphosts	urlhaus	79%
tolouseddos	tolousephishing	77%
tolousephishing	urlhaus	77%
cybercrimetracker	hostfile	77%
malwaredomainlist	urlhaus	76%
dnsbh	tolouseddos	75%

identify that one blacklist copies from another blacklist.
The next metric, the correlation between blacklists is not a good way to specifically identify copying-behaviour because of the nature of correlation. Correlation is only looking at the number of domains listed and it can not measure if one blacklist lists the same domains of another blacklist. It can, however, be used to measure trends and to identify potential copying-behaviour of blacklists where a high correlation exists.
The third approach is the most interesting as it specifically looks at the number of domains that were added one day after it was listed by another blacklist. It can show that some blacklists regularly and structurally listed a domain after it was added by another blacklist one day before.

Therefore the combination of the three approaches can be used to characterize the information flow between blacklists.

5. CONCLUSION

In this research, we have shown it is indeed possible to identify blacklists that copy from another blacklist using our three defined approaches. The research revealed that some blacklists indeed fully include another blacklist by looking at the high percentages of the number of common domains. The second approach of looking at the correlations can be used to measure patterns and trends, and it gives a list of blacklists that can be researched further.

We revealed that even some blacklists which do not publicly state that they copy from another blacklist still can be identified using our third approach of looking at the added domains from a blacklist that were listed before another blacklist. Still, the caveat that it is not a real proof that those blacklists are copying from another blacklist must be considered. This is because different blacklists still can have the same methods of identifying malicious domains, but one blacklist is structurally one day later in listing than another blacklist.

Therefore we can answer the research question and subquestion, that it is indeed possible to show that one blacklist copies from another, while considering the caveat. The combination of our three approaches can be used to characterize the information flow and identify copying-behaviour amongst blacklists.

6. FUTURE WORK

There are more topics of interest to gain more knowledge about blacklists that can be explored in the future to characterize the information flow and to help software developers and network administrators in choosing a blacklist.

We only analyzed day to day data, as that was the most frequent update of the data set. However in the data set there were also domains which were listed by multiple blacklists on the same day. In order to find out if those blacklists show copying-behaviour or not, there is need to crawl the blacklists more frequently.

Some blacklists are using entire URL paths instead of listing domains only. It would be interesting to split these blacklists and do our analysis again on the full URLs and see if any differences emerge.

7. ACKNOWLEDGMENTS

I would like to thank my supervisor R. Sommese of the Design and Analysis of Communication Systems (DACS) Group from the University of Twente, for the support while doing data analysis and writing this paper and for always being available to answer my questions. Besides, I would to thank my fellow students and the supervisor, dr. S. Bayhan, of the Dependable Networks track for providing regular feedback and insights for this research.

8. REFERENCES

[1] Abuse.ch. URLhaus. https://urlhaus.abuse.ch.
[2] Abuse.ch. RansomwareTracker. https://ransomwaretracker.abuse.ch.
[3] Abuse.ch. ZeusTracker. https://zeustracker.abuse.ch.
[4] K. Angrishi. Turning internet of things(iot) into internet of vulnerabilities (iov) : Iot botnets. *CoRR*, abs/1702.03681, 2017.
[5] B. Consulting. OSINT Feeds from bambenek consulting. https://osint.bambenekconsulting.com/feeds.
[6] CyberCrimeTracker. http://cybercrime-tracker.net/about.php.
[7] J. M. Ehrenfeld. Wannacry, cybersecurity and health information technology: A time to act. *Journal of Medical Systems*, 41(7):104, May 2017.
[8] InternetStormCenter. Suspicious Domains. https://dshield.org/suspicious_domains.html.
[9] M. Jhaveri, O. Cetin, C. H. Ganan, T. Moore, and M. Eeten. Abuse reporting and the fight against cybercrime. *ACM Computing Surveys*, 49:1–27, 01 2017.
[10] JoeWein.de. joewein.de - fighting spam and scams on the internet. https://www.joewein.net.
[11] M. Kührer, C. Rossow, and T. Holz. Paint it black: Evaluating the effectiveness of malware blacklists. volume 8688, pages 1–21, 09 2014.
[12] Malc0de. http://malc0de.com.
[13] MalwareBytes. hpHosts online. https://hosts-file.net.
[14] MalwareDomainList. https://malwaredomainlist.com.

[15] S. Mccombie, J. Pieprzyk, and P. Watters. Cybercrime attribution: An eastern european case study. *Australian Digital Forensics Conference*, 01 2009.

[16] NoVirusThanks. URLVir Monitor Malicious Executable Urls. http://www.urlvir.com.

[17] OpenPhish. Openphish. https://openphish.com.

[18] Oracle. DynDNS.org Malware Feeds. http://security-research.dyndns.org/pub/malware-feeds.

[19] A. Pitsillidis, C. Kanich, G. Voelker, K. Levchenko, and S. Savage. Taster's choice: A comparative analysis of spam feeds. pages 427–440, 11 2012.

[20] F. Prigent. Université Toulouse 1 Capitole blacklists. https://dsi.ut-capitole.fr/blacklists/index_en.php.

[21] RiskAnalytics. DNS-BH – Malware Domain Blocklist by RiskAnalytics. http://www.malwaredomains.com.

[22] Q. Scheitle, O. Hohlfeld, J. Gamba, J. Jelten, T. Zimmermann, S. D. Strowes, and N. Vallina-Rodriguez. A long way to the top: Significance, structure, and stability of internet top lists. *Proceedings of the Internet Measurement Conference 2018 on - IMC '18*, 2018.

[23] SquidBlacklist.org. http://www.squidblacklist.org.

[24] VxVault. http://vxvault.net.