Empirical Analysis of Domain Blacklists

Tran Phuong Thao*, Akira Yamada†, Ayumu Kubota‡
KDDI Research, Inc., Japan
2-1-15 Ohara, Fujimino-shi, Saitama, Japan 356-8502
Email: {th-tran*, ai-yamada†, kubota‡}@kddi-research.jp

***Abstract*—Malicious content has grown along with the explosion of the Internet. Therefore, many organizations construct and maintain blacklists to help web users protect their computers. There are many kinds of blacklists in which domain blacklists are the most popular one. Existing empirical analyses on domain blacklists have several limitations such as using only outdated blacklists, omitting important blacklists, or focusing only on simple aspects of blacklists. In this paper, we analyze the top 14 blacklists downloaded on 2017/02/28 including popular and updated blacklists like *Safe Browsing* from Google and *urlblacklist.com*. We are the first to filter out the old entries in the blacklists using an enormous dataset of user browsing history. Besides the analysis on the intersections and the registered information from Whois (such as top-level domain, domain age and country), we also build two classification models for web content categories (i.e., education, business, etc.) and malicious categories (i.e., landing and distribution) using machine learning. Our work found some important results. First, the blacklists *Safe Browsing version 3 and 4* are being separately deployed and have independent databases with diverse entries although they belong to the same organization. Second, the blacklist *dsi.ut_capitole.fr* is almost a subset of the blacklist *urlblacklist.com* with 98% entries. Third, largest portion of entries in the blacklists are created in 2000 with 6.08%, and from United States with 24.28%. Fourth, *Safe Browsing version 4* can detect younger domains compared with the others. Fifth, *Tech & Computing* is the dominant web content category in all the blacklists, and the blacklists in each group (i.e., small public blacklists, large public blacklists, private blacklists) have higher correlation in web content as opposed to blacklists in other groups. Sixth, the number of landing domains are larger than that of distribution domains at least 75% in large public blacklists and at least 60% in other blacklists. In addition, we collected and analysed the updated version of 11 public blacklists that we downloaded on 2017/11/09, which is over 7 months after the previous blacklist version downloaded on 2017/02/28, and found some new results such as: the number of malicious domains injected by ransomwares is significantly increased (6.67x larger); or many Top Level Domains (TLDs) which belong to the type of *new generic TLD* such as *.forsale*, *.institute*, *.church*, etc., appear in the new blacklist version. We also discussed several challenges on measuring registration time of malicious domains in each blacklist, how to determine a malicious domain, malicious classification using Whois-document-based text mining, and standardization of Whois-attribute extraction.**

Keywords*—Web Security, Empirical Analysis, Blacklist, Malicious Domain, Whois Information, HTML Document, Text Mining.

I. Introduction

The Internet has become very important to our daily life, and thus, the content of the Web has been growing exponentially. According to a research by VeriSign, Inc. [2], the number of domains is already approximately 12 million as of March 31, 2016. Along with that is a huge amount of malicious domains. Just in 2015, the number of unique pieces of malware discovered is more than 430 million, up 36 percent from the year before [3]. Therefore, nowadays there are many competitive services constructed to detect malicious domains. Each service has its own method, which is often not disclosed and always said to be the best service by its authors. Furthermore, each service also has different definition (ground truth) of the term "malicious". For example, a blacklist A defines a domain D to be malicious if D satisfies a condition set AM while another blacklist B defines D to be malicious if D satisfies a condition set BM, which is a subset, superset or completely different from AM. All of these have brought into **a question: how to measure and compare these services**. Many blacklists are freely available on the Internet (called *public blacklists*). However, some vendors do not want to publish their databases and only provide querying services via APIs or portal applications (called *private blacklists*). Our goal in this paper is to perform a large-scale analysis on popular blacklists including both public and private blacklists. We can then indicate the quality of the blacklists in some specific categories. This research can help the users to determine which blacklists should they choose for some conditions, and also can help the blacklist providers assess and improve their blacklists and methods.

A. Related Work

Sheng et al. [4] analyzed phishing blacklists, which are just subset of malicious blacklists that we are focusing on. A malicious domain's purpose includes all kinds of attacks: spamming, phishing, randomware, etc. Kuhrer et al. [5] analyzed malicious blacklists but only focused on constructing a blacklist parser to deal with varied-and-unstructured blacklist formats rather than researching the blacklists themselves. This is because some blacklists solely include domain names, URLs, or IP address. Other blacklists contain more information, such as timestamps or even source, type, and description for each entry. Therefore, their analysis results have poor information that only contains the entries' registration history in each blacklist, the intersection of every blacklist pair, and the top 10 domains in most of the blacklists. Kuhrer et al. [6] then analyzed blacklists via three measures: (i) identifying parked domains (additional domains hosted on the same account and displaying the same website as primary domain) and sinkhole servers (hosting malicious domains controlled by security organizations), (ii) the blacklist completeness by finding the coverage between each blacklist with an existing set of 300,000 malware samples, and (iii) the domains created by Domain Generation Algorithm. However, 300,000 entries in the second measure are not enough to assess the "completeness" because some large blacklists can contain millions of entries.

Furthermore, the ground truth or definition of their malware samples may be different from that of other blacklists, and thus it is unfair when using them to confirm the completeness of other blacklists. The first and third measures are different for our analysis. Vasek et al. [7] only analyzed Malware Domain Blacklist (malwaredomains.com), which is just one of the blacklists in our analysis. Several other papers also performed empirical analysis but are different from our analysis, which focuses on domain blacklists, e.g., [8] analyzed IP blacklists, [9] analyzed email spam detection through network characteristics in a stand-alone enterprise, [10] analyzed spam traffic with a very specific network, [11] analyzed detections of malicious web pages caused by drive-by-download attack, not blacklist analysis, [12] analyzed whitelist of acceptable advertisements.

B. Our Work

In this paper, we do not aim to figure out the ground truth or definition of "malicious", or the factors affecting malicious domain detection in each blacklist. Instead, we attempt to quantitatively measure and compare the blacklists based on seven important aspects: blacklist intersections, top-level domains (TLDs), domain ages, countries, web content categories, malicious categories, differences between the blacklist versions. To the best of our knowledge, we are the first to achieve the followings:

- We deal with top 14 popular blacklists in which there are two special private blacklists given by Google that are Safe Browsing version 3 and 4 (called GSBv3 and GSBv4). These newest versions are being deployed and used parallelly and independently, and have never been analyzed before. In [5], the old version GSBv2 was analyzed in 2011, which was 6 years ago.
- By designing 7 measures in our analsysis, we not only consider the coverage (intersection) as in previous works, but also compare the blacklists based on Whois (TLDs, countries, domain ages), web content categories using IAB [13], which are an industry standard taxonomy for content categorization (e.g., education, government, etc.), malicious categories (landing and distribution), and the differences between the current and the newest updated blacklist versions (as of the time of writing this paper).
- Our analysis is not straightforward, and not just simple statistics. For the measures of web content categories and malicious categories, we construct two supervised machine learning models using text mining, and a combination of text mining with some specific HTML tags to classify the entries in the blacklists, respectively.
- Last but not least, we filter out the active entries in the blacklists instead of old and useless entries as previous works by finding the coverage between each blacklist with a big live dataset.

Roadmap. The rest of this paper is organized as follows. The methodology of our analysis is presented in Section II. The empirical results are given in Section III. The discussion is described in Section IV. Finally, the conclusion is drawn in Section V.

II. METHODOLOGY

In this section, we introduce our chosen blacklists, how we pre-processed them, and our analysis design.

A. Blacklists

In this paper, we analyze 14 popular blacklists as described in Table I. Since they have different numbers of entries, which can effect the fairness, we categorize them into 3 groups: (I) *small public blacklists* which have smaller than 1,000,000 unique entries, (II) *large public blacklists* which have equal or larger than 1,000,000 unique entries, and (III) *private blacklists*. In the group (III), we consider separately GSBv3 and GSBv4 although they both belong to the same vendor. This is because they are being deployed and used independently. Furthermore, according to our analysis, they have different API and even database.

TABLE I: 14 POPULAR BLACKLISTS.

No	Group	Abbr.	Blacklists	#Domains
1	(I)	MA	malwaredomains.com	17,294
2		NE	networksec.org	263
3		PH	phishtank.com	9,711
4		RA	ransomwaretracker.abuse.ch	1,380
5		ZE	zeustracker.abuse.ch	382
6		MAL	malwaredomainlist.com	1,338
7		MV	winhelp2002.mvps.org	218,248
8		HO	hosts-file.net	5,974
9	(II)	ME	mesd.k12.or.us	1,266,334
10		SH	shallalist.de	1,570,944
11		UR	urlblacklist.com	2,919,199
12		UT	dsi.ut_capitole.fr	1,346,788
13	(III)	GSBv3	Safe Browsing version 3	Unknown
14		GSBv4	Safe Browsing version 4	Unknown

In Table I, the last column indicates the number of unique domains in each blacklist. All the 14 blacklists were downloaded (in case of public blacklists) or queried (in case of private blacklists) on the same date 2017/02/28. Since the blacklists may contain old entries that attackers no longer use, we extract only active entries by finding the intersection between each blacklist with a real-world web access log that we call AL. AL has 3,991,599,424 records from 5 proxy servers, 9,091,980 raw domains with 80,464,378 corresponding URLs accessed by 659,283 users. The intersections between AL and each blacklist are given in Table II. The number of unique domains in the union of 14 blacklists is 50,519. Instead of the complete blacklists, we use these intersections in our analysis. We should mention that the AL focuses on the users in Japan; and thus, the information of the access records is mainly from Japan (e.g., many domains has the top-level domains (TLD) of *.jp*). Using the AL is not generic; and analyses for other countries are recommended to use the different specific access log. Although there exists a bias when using the AL here, the insights found in our analyses can be widely effective to other countries; for instance, most of the domains are registered from United States and created in 2000; or, GSBv4 can detected younger domains than other blacklists, etc.

B. Analysis Design

In this section, we describe the design of our analysis with the following 6 measures.

TABLE II: ACTIVE MALICIOUS DOMAINS IN 14 BLACKLISTS (INTERSECTIONS WITH AL).

No	Group	Intersection	Abbr.	#Domains	Percentage
1	(I)	AL ∩ MA	AMA	77	0.44%
2		AL ∩ NE	ANE	2	0.76%
3		AL ∩ PH	APH	367	3.78%
4		AL ∩ RA	ARA	3	0.22%
5		AL ∩ ZE	AZE	21	5.50%
6		AL ∩ MAL	AMAL	98	7.32%
7		AL ∩ MV	AMV	2,176	1.00%
8		AL ∩ HO	AHO	5,060	84.70%
9	(II)	AL ∩ ME	AME	19,812	1.56%
10		AL ∩ SH	ASH	32,248	2.05%
11		AL ∩ UR	AUR	33,674	1.15%
12		AL ∩ UT	AUT	24,020	1.78%
13	(III)	AL ∩ GSBv3	AGSBv3	189	unknown
14		AL ∩ GSBv4	AGSBv4	639	unknown

The final column indicates the number of filtered samples over that of original samples in Table I.

1) Measure 1 (Blacklist Intersections): For every blacklist pair with the web access log AL, we find the intersection of their domains. In total we found $\binom{14}{2} = 91$ intersection sets. In our previous article [1], we determined the blacklist pair that has the highest correlation in term of overlapping entries based on the number of entries in the intersections (i.e., the blacklist pair that has largest number of entries in their intersection is the one has highest correlation). However, it is unfair to compare between all the blacklist pairs because each blacklist has a different number of entries. Therefore, in this article version, we determine the blacklist pair that has the highest correlation based on the average of the two percentages of the pairs and then choose the largest one. We will explain the example in Section III-A.

2) Measure 2 (Top-Level Domains (TLDs)): A TLD is the domain in the highest level of the hierarchical Domain Name System. For example, the TLD of the domain *kddi.com* is *com*, the TLD of the domain *yahoo.co.jp* is *jp*. To evaluate this measure, we extract the final string after the dot in each domain name. According to ICANN (the Internet Corporation for Assigned Names and Numbers) [14], there are 1,540 different TLDs as of 2017/11/15 categorized into 6 types: (i) infrastructure top-level domain (ARPA), (ii) generic top-level domains (gTLD), (iii) restricted generic top-level domains (grTLD), (iv) sponsored top-level domains (sTLD), (v) test top-level domains (tTLD), and (vi) new generic top-level domains (new gTLD). The *most common type is the generic top-level domains (gTLD)*, which has two sub-types:

- Original TLDs: consist of *.com*, *.org*, *.net*, *.int*, *.edu*, *.gov* and *.mil*.
- Country-code TLDs: consist of the TLDs of each country or region. For example, *.jp* (Japan), *.us* (United States), *.eu* (European Union), etc.

3) Measure 3 (Domain Ages) and Measure 4 (Countries): To evaluate these measures, we firstly extract the Whois information of each domain in all the intersections between the blacklists and the web access log AL as described in Table II. Whois is the registered information of the domains such as creation date, expiration date, organization, address, registrar server, etc. For the measure 3, we extract creation year (from the creation date) and for the measure 4, we extract the country. Note that, although the measure 2 (TLD) includes country-code TLDs, it does not always show correct countries. For example, the TLD of *jp* not only contains domains from Japan, but also another countries such as United States with a non-small portion. This is why we consider the measure 2 (TLD) and measure 4 (country), separately.

4) Measure 5 (Web Content Categories): This measure aims to classify the blacklisted domains into semantic web content categories, such as education, advertisement, government, etc. Although there are several tools (e.g., i-Filter [15], SimilarWeb [16]) which can be used to categorize a domain into semantic content categories, their coverages are low and they cannot label our entire dataset (this will be explained later). Therefore, to evaluate this measure, we construct our own classification model using supervised machine learning with the help of one of the tools for data labelling. Concretely, we first collect 20,000 URLs and label their semantic contents using i-Filter [15]. However, i-Filter cannot label all the samples but only 14,492 samples (72.46%) into 69 categories. Since the number of categories is quite large for the number of classes in our model, we thus generalize these 69 categories into 17 categories using the standardized category set called IAB [13]. We then extract HTML documents of the 14,492 samples and use *text mining* with Term Frequency-Inverse Document Frequency (TF-IDF) as the feature for the training process. We executed nine different supervised machine learning algorithms: Support Vector Machine (including C-based and Linear-based), Naive Bayes (including Multinomial-based and Bernoulli-based), Nearest Neighbors (including Centroid-based, KNeighbors-based and Radius-based), Decision Tree, and Stochastic Gradient Descent. We assessed the algorithms using k-fold cross validation by setting $k = 10$. We pick up the best algorithm, which has highest accuracy and lowest false positive rate. Thereafter, we extract HTML documents of 50,519 domains in our blacklists. Note that, given a domain, we extract the main URL of the domains by adding prefix *http://www* to the domain. For example: the main url of *google.com* is *http://www.google.com*. We use the model computed by the chosen best learning algorithm to classify the 50,519 domains in the blacklists.

5) Measure 6 (Malicious Categories): There are two types of malicious categories. The first type is about the behaviours of attackers such as phishing, spamming or abusing, etc. This type has already been considered in many previous works. The second type is about the behaviours of the domains/URLs themselves such as *landing* and *distribution*, which are very important properties to understand the attacks but have not been widely considered before. Landing domains are what the web users are often attracted to access, and contain some malicious codes (usually Javascript) which can redirect the users (victims) to another malicious domains called distribution domains. Distribution domains are what the victims are redirected to unconsciously, and really install malwares into the victims' computers. To the best of our knowledge, currently there is a unique tool which can be used to classify a malicious domain into landing or distribution, which is GSBv4. GSBv4 not only is a blacklist (i.e., can detect whether a domain is malicious or benign) but also can classify a malicious domain into landing or distribution category. However, its classification rate is too low (this will be explained later); furthermore, it can only classify the domains belonging to its blacklist without being able to classify domains in other blacklists. This is why we construct our own classification model using supervised machine learning and only use GSBv4 for data labelling. Concretely, we first randomly collect 31,507 malicious URLs and label them using GSBv4. We then only have 5,772 samples (18.31%), which can be labelled by GSBv4 (4,124 landings and 1,648 distributions). After that, we extracted HTML documents of the labelled 5,772 samples to use in the training process. For feature selection, at first, adapting the idea of [17], we extracted and counted the following special HTML elements in each type:

- Type 1: eight HTML tags, which are used very often in landing domains including: *<script>*, *<iframe>*, *<form>*, *<frame>*, *<object>*, *<embed>*, *<href>*, and *<link>*. This is because these tags allow to place URLs inside, and thus have potential for the redirection, which is a specific characteristic of landing domains.
- Type 2: three elements which are commonly used in distribution domains including *swf*, *jar* and *pdf*. This is because these elements are mostly potential exploitable contents that distribution domains install into victim's computers.

However, our implementation showed that the accuracy of this method is very low (less than 71% using the 9 learning algorithms and 10-fold cross validation). Therefore, we then combine the 2 methods: the above HTML elements (in which the count of all tags in each type is used as one feature) along with text mining on entire HTML documents (in which the TF-IDF of each unique word is used as one feature). As a result, fortunately, we can get 98.07% in accuracy with merely 2.22% in false positive rate. Finally, we use the model of our combining method to classify 50,519 entries in the blacklists.

6) Measure 7 (New Blacklist Version): This measure aims to the new findings on the differences between the new and old versions of the same blacklists. A simple design is based on the number of unique domains in the previous and new blacklist versions. Note that, in this measure, we will directly analyse the blacklists themselves without getting the intersection between the blacklists with the AL. Since private blacklists (group III) such as GSBv3 and GSBv4 do not disclose their number of entries and also their entries in plaintext format (just in a hashed format), we cannot directly analyse them. For this reason, this measure only focuses on public blacklists including small (group I) and large (group II) public blacklists. Besides the differences between the number of domains in the previous and new blacklist versions, we also found some important findings when implementing the TLDs of the new blacklist version.

III. EMPIRICAL RESULTS

In our implementation, we use two machines: a computer Intel(R) core i7, RAM 16.0 GB, 64-bit Windows 10; and a MacBook Pro Intel Core i5 processor, 2.7 GHz, 16 GB of RAM, OS X EI Capitan version 10.11.6. Since we do not consider the execution time, it does not matter that the two machines have different configurations. They are just used to speed up our evaluation modules, which can be executed parallelly and independently. We execute the 6 measures using Python 2.7.11 programming language with *pandas* library to deal with big data. Furthermore, we use *python-whois* library version 0.6.5 for Whois extraction of measure 3 and 4. We also use *scikit-learn* library for text mining and *BeautifulSoup* library for HTML extraction of measure 5 and 6.

TABLE III: TOP 10 DOMINANT TLDs IN ALL THE BLACKLISTS.

No	TLD	#Domains	Percentage
1	com	32,691	64.71 %
2	jp	4,277	8.47 %
3	net	3,458	6.84 %
4	org	1,856	3.67 %
5	de	726	1.44 %
6	uk	683	1.35 %
7	au	428	0.85 %
8	edu	375	0.74 %
9	tv	366	0.72 %
10	info	310	0.61 %

TABLE IV: TOP 5 DOMINANT TLDs IN EACH BLACKLIST

No	Blacklist	#Distinct TLDs	1st	2nd	3rd	4th	5th
1	AMA	25	com	jp	pl	net	org
2	ANE	2	com	pl			
3	APH	68	com	net	org	ru	pl
4	ARA	3	to	org	cab		
5	AZE	9	net	com	ua	ru	jp
6	AMAL	22	com	net	it	jp	ru
7	AMV	79	com	net	de	ru	org
8	AHO	145	com	net	org	jp	de
9	AME	113	com	net	org	tv	jp
10	ASH	197	com	jp	net	org	de
11	AUR	180	com	net	org	jp	uk
12	AUT	137	com	net	org	jp	tv
13	AGSBv3	34	com	org	jp	net	cn
14	AGSBv4	61	com	net	top	org	biz

A. Measure 1: Blacklist Intersections

In Table V, we computed the intersections of 91 blacklist pairs (with AL); and not only the number of entries in the intersections (overlapping entries), we also computed the corresponding percentages for each of the blacklist pairs. From the results in Table V, we observe some important information as follows:

- Based on the method used to score the correlation as described in Section II-B1, the table show that the blacklist pair that has the highest correlation in term of overlapping entries is (ME, UT) because the intersection AME $\cap$ AUT contains 19,598 entries, which occupies 90.25% in average of the two percentages (98.92% AME and 81.59% AUT). This result is different from the previous result in the paper [1], which showed that the pair that has highest correlation is (UT, UR) due to the highest number of entries in the intersection (23,583 domains). Note that, a blacklist pair in which one blacklist is a/an (almost) subset of the other is not always the pair, which has highest correlation in term of overlapping entries. This is because the percentage of the subset blacklist is very

TABLE V: OVERLAPPING NUMBER OF ENTRIES (INTERSECTIONS) OF EVERY BLACKLIST PAIR

	ANE	APH	ARA	AZE	AMAL	AMV	AHO	AME	ASH	AUR	AUT	AGSBv3	AGSBv4
MA	2 AMA: 2.60% ANE: 100%	7 AMA: 9.09% APH: 1.91%	0 AMA: 0% ARA: 0%	0 AMA: 0% AZE: 0%	0 AMA: 0% AMAL: 0%	0 AMA: 0% AMV: 0%	35 AMA: 45.45% AHO: 0.69%	77 AMA: 100% AME: 0.39%	1 AMA: 1.30% AHS: 0%	13 AMA: 16.88% AUR: 0.04%	1 AMA: 1.30% AUT: 0%	1 AMA: 1.30% AGSBv3: 0.53%	4 AMA: 5.19% AGSBv4: 0.63
NE		0 ANE: 0% APH: 0%	0 ANE: 0% ARA: 0%	0 ANE: 0% AZE: 0%	0 ANE: 0% AMAL: 0%	0 ANE: 0% AMV: 0%	1 ANE: 50.00% AHO: 0.02%	2 ANE: 100% AME: 0.01%	0 ANE: 0% AHS: 0%	2 ANE: 100% AUR: 0.01%	0 ANE: 0% AUT: 0%	0 ANE: 0% AGSBv3: 0%	2 ANE: 100% AGSBv4: 0.31
PH			0 APH: 0% ARA: 0%	6 APH: 1.63% AZE: 28.57%	14 APH: 3.81% AMAL: 14.29%	42 APH: 11.44% AMV: 1.93%	175 APH: 47.68% AHO: 3.46%	15 APH: 4.09% AME: 0.08%	104 APH: 28.34% AHS: 0.32%	100 APH: 27.25% AUR: 0.30%	51 APH: 13.90% AUT: 0.21%	1 APH: 0.27% AGSBv3: 0.53%	1 APH: 0.27% AGSBv4: 0.16
RA				0 ARA: 0% AZE: 0%	0 ARA: 0% AMAL: 0%	0 ARA: 0% AMV: 0%	3 ARA: 100% AHO: 0.06%	0 ARA: 0% AME: 0%	2 ARA: 66.67% ASH: 0.01%	1 ARA: 33.33% AUR: 0%	0 ARA: 0% AUT: 0%	0 ARA: 0% AGSBv3: 0%	0 ARA: 0% AGSBv4: 0%
ZE					2 AZE: 9.52% AMAL: 2.04%	1 AZE: 4.76% AMV: 0.05%	18 AZE: 85.71% AHO: 0.36%	2 AZE: 9.52% AME: 0.01%	6 AZE: 28.57% ASH: 0.02%	6 AZE: 28.57% AUR: 0.02%	1 AZE: 4.76% AUT: 0%	0 AZE: 0% AGSBv3: 0%	0 AZE: 0% AGSBv4: 0%
MAL						21 AMAL: 21.43% AMV: 0.97%	67 AMAL: 68.37% AHO: 1.32%	6 AMAL: 6.12% AME: 0.03%	30 AMAL: 30.61% ASH: 0.09%	36 AMAL: 36.73% AUR: 0.11%	6 AMAL: 6.12% AUT: 0.02%	0 AMAL: 0% AGSBv3: 0%	0 AMAL: 0% AGSBv4: 0%
MV							1,241 AMV: 57.03% AHO: 24.53%	262 AMV: 12.04% AME: 1.32%	1,152 AMV: 52.94% ASH: 3.57%	948 AMV: 43.57% AUR: 2.82%	626 AMV: 28.77% AUT: 2.61%	0 AMV: 0% AGSBv3: 0%	0 AMV: 0% AGSBv4: 0%
HO								754 AHO: 14.90% AME: 3.81%	2,070 AHO: 40.91% ASH: 6.42%	1,733 AHO: 34.25% AUR: 5.15%	1,179 AHO: 23.30% AUT: 4.91%	3 AHO: 0.06% AGSBv3: 1.59%	5 AHO: 0.10% AGSBv4: 0.78
ME									11,736 AME: 59.24% ASH: 36.39%	19,494 AME: 98.39% AUR: 57.89%	19,598 AME: 98.92% AUT: 81.59%	7 AME: 0.04% AGSBv3: 3.70%	28 AME: 0.14% AGSBv4: 4.38
SH										19,495 ASH: 60.45% AUR: 57.89%	14,769 ASH: 45.80% AUT: 61.49%	4 ASH: 0.01% AGSBv3: 2.12%	19 ASH: 0.06% AGSBv4: 2.97
UR											23,583 AUR: 70.03% AUT: 98.18%	7 AUR: 0.02% AGSBv3: 3.70%	29 AUR: 0.09% AGSbv4: 4.54
UT												7 AUT: 0.03% AGSBv3: 3.70%	25 AUT: 0.10% AGSBv4: 3.91
GSBv3													170 AGSBv3: 89.95 AGSBv4: 26.60

high but that of the other blacklist can be very small; and that makes the average percentage is not higher than that of other blacklist pairs.

- The results also indicate that the size of the values in this table is *not only dependent on the size of each original blacklist*. For example, ASH = 32,248 and AUR = 33,674 but ASH $\cap$ AUR = 19,495, which is smaller than AUT $\cap$ AUR = 23,583 even though AUT = 24,020, which is smaller than ASH.
- Most (not all) of the entries in AGSBv3 is listed in AGSBv4 since the percentage in GSBv3-side is very high (89.95%). Note that, GSBv3 and GSBv4 are the different versions of the same Google's Safe Browsing product but are being deployed parallelly and are using different databases. This result can lead to a hypothesis that, GSBv3 probably will be gradually merged with GSBv4 in near future although there has been no official announcement yet.

B. *Measure 2: TLDs*

From 50,519 unique domains in all the blacklists, we found 253 different TLDs in totals in which the top 10 dominant TLDs *for all the blacklists* are given in Table III. We then found top 5 dominant TLDs *for each blacklist* as given in Table IV. The third column is the number of distinct TLDs in each blacklist. The fourth until the eighth columns are the top 5 TLDs in descending order. Similar to the measure 1, the number of unique TLDs (the 3rd column) is *not always dependant on the number of entries* in each blacklist. For example, the blacklist HO belongs to the group I (small public blacklists) and AHO has only 5,060 entries but the number of TLDs is 145; meanwhile, the ME belongs to the group II (large public blacklists) and AME has 19,812 entries, which is almost $4\times$ larger than that of AHO, but its number of TLDs is only 113.

C. *Measure 3: Domain Ages*

Considering the union of all 14 blacklists, there are *34 distinct creation years* (from 1984 to 2017) as given in Figure 1. We can observe that the number of detected malicious domains created *after 1993 increases remarkably* compared to the years before 1993, and *drops down from 2016* (just 1 year before the date that we started our analysis). This indicates that most of the blacklists can *detect the new (young) malicious domains created after 2015 with very low rate*. The top 10 dominant years with corresponding number of domains are given in Table VI. For each blacklist, we also found the top 5 dominant creation years as presented in Table VII. We can observe that the blacklists *MA and GSBv4 can detect younger domains* compared with the other blacklists. Meanwhile, the blacklists *MAL and MV can detect very old domains*.

D. *Measure 4: Countries*

From the union of 14 blacklists, which contains 50,519 domains, we found 173 distinct registered countries. Note that, some domains are registered under one or multiple countries. That is, the registrator's addresses consist of one or multiple countries. For this reason, we consider each different country even in the same domain instead of just randomly choosing one of the countries for each domain when the domain has multiple countries. The top 10 dominant countries throughout the union of 14 blacklists are given in Table VIII. Besides the union of all the blacklists, we also found top 5 dominant countries in each blacklist as presented in Table IX. The third column is the number of distinct countries in each blacklist. The fourth until eighth columns are the top 5 dominant countries described in descending order. From this table, we can observe that ME and UT have highest correlation because their numbers of distinct countries are almost equal, and the order of their dominant countries from the fourth to the eighth column is exactly same.

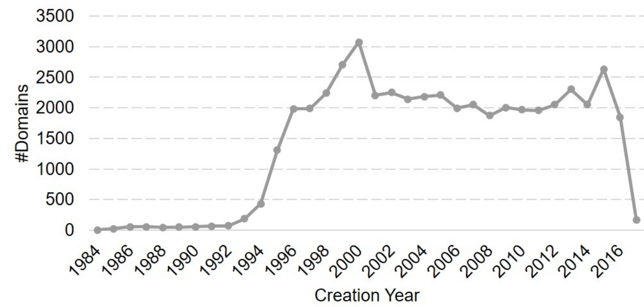

Figure 1: Distribution of Domain Ages (Creation Year).

TABLE VI: TOP 10 DOMINANT CREATION YEARS IN ALL THE BLACKLISTS.

No	Year	#Domains	Percentage
1	2000	3,073	6.08 %
2	1999	2,707	5.36 %
3	2015	2,633	5.21 %
4	2013	2,302	4.56 %
5	2002	2,249	4.45 %
6	1998	2,239	4.43 %
7	2005	2,209	4.37 %
8	2001	2,205	4.36 %
9	2004	2,181	4.32 %
10	2003	2,141	4.24 %

TABLE VII: TOP 5 DOMINANT CREATION YEARS IN EACH BLACKLIST.

No	Blacklist	#Distinct Years	1st	2nd	3rd	4th	5th
1	AMA	16	2016	2015	2014	2013	2012
2	ANE	2	2012	2006			
3	APH	27	2011	2009	2010	1999	2004
4	ARA	3	2014	2013	2008		
5	AZE	12	2007	2004	2001	2008	2006
6	AMAL	25	1999	1997	1998	1996	2005
7	AMV	32	1998	1999	1995	1996	2000
8	AHO	32	2005	2007	2016	1999	2012
9	AME	29	2015	2013	2012	2014	2011
10	ASH	33	2000	1999	2002	2001	1998
11	AUR	33	2015	2013	1999	2000	2007
12	AUT	33	2015	2013	2012	2014	2007
13	AGSBv3	21	2016	2012	2009	2013	2011
14	AGSBv4	21	2016	2015	2014	2012	2013

E. *Measure 5: Web Content Categories*

1) Pre-processing and Determining the Best Algorithm for Classification: After labelling 14,492 samples by i-Filter and

TABLE VIII: TOP 10 DOMINANT COUNTRIES IN ALL THE BLACKLISTS.

No	Country	#Domains	Percentage
1	US	12,267	24.28 %
2	JP	7,959	15.75 %
3	CY	3,988	7.89 %
4	PA	3,207	6.35 %
5	RU	1,194	2.36 %
6	AU	1,172	2.32 %
7	FR	1,072	2.12 %
8	DE	1,072	2.12 %
9	CA	994	1.97 %
10	GB	983	1.95 %

TABLE IX: TOP 5 DOMINANT COUNTRIES IN EACH BLACKLIST.

No	Blacklist	#Distinct Countries	1st	2nd	3rd	4th	5th
1	AMA	28	JP	US	CN	CA	FR
2	ANE	2	PL	CN			
3	APH	54	US	RU	AU	DE	BR
4	ARA	3	TO	DE	CA		
5	AZE	11	US	UA	RU	JP	NU
6	AMAL	28	US	IT	RU	JP	KR
7	AMV	81	US	DE	CA	FR	PA
8	AHO	104	US	JP	PA	CN	DE
9	AME	125	US	CY	PA	JP	RU
10	ASH	153	US	JP	CY	PA	DE
11	AUR	152	US	CY	JP	PA	RU
12	AUT	126	US	CY	PA	JP	RU
13	AGSBv3	39	US	JP	CN	RU	PL
14	AGSBv4	58	US	CN	JP	PL	DJ

IAB as mentioned in Section II-B4, we got 17 categories as described in Table X. Note that, the order of the numbers of samples in these categories does not indicate that of the domains in the blacklists. Even the numbers of samples in the categories are varied, for example, the number of samples of *Tech & Comp.* is double that of *Business* in the training dataset, it does not mean that *Tech & Comp.* always has higher order than *Business* in the applied dataset. We used the 14,492 labelled samples for our training dataset and inputted them to the supervised algorithms. We obtained the accuracy and false positive rate for each algorithm as given in Figure 2. We found that Decision Tree gives the best accuracy (99.58%) and lowest false positive rate (0.04%). We thus choose it to classify the domains in our blacklists.

2) Classification Result Using the Best Algorithm: As explained above, we use Decision Tree for the classification of the web content. For the union of all the blacklists, which consists of 50,519 domains, the web content categories with the corresponding number of domains are given in Table XI. We observe that the top 3 dominant categories are *Technology and Computing*, *Business*, and *Non-Standard content* (such as *Pornography*, *Violence*, or *Incentivized*). For each blacklist, the top 5 dominant categories with corresponding number of domains are presented in Table XII. We found that all the blacklists belonging to the group II (large public blacklists including ME, SH, UR, and UT), have higher correlation in web content categories rather than the other blacklists since the number of distinct categories and the order of dominant categories are exactly the same. Furthermore, MV and HO,

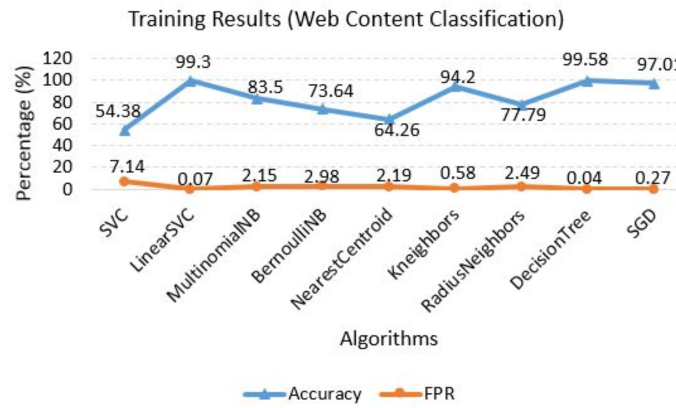

Figure 2: Accuracy and False Positive Rate of Each Algorithm

which belong to the group I (small public blacklists) and GSBv3, which belongs to the group III (private blacklists) also have the same order of dominant categories.

TABLE X: 17 CATEGORIES OF THE WEB CONTENT IN THE TRAINING DATASET.

No	Category	#Samples	No	Category	#Samples
1	Art & Entert.	65	10	Personal Finance	103
2	Automotive	29	11	Real Estate	18
3	Business	4,622	12	Tech & Comp.	7,632
4	Careers	17	13	Society	137
5	Education	15	14	Hobby & Interest	503
6	Shopping	604	15	Non-Standard	490
7	Food & Drink	37	16	News	117
8	Science	8	17	Sports	8
9	Travel	87			

TABLE XI: WEB CONTENT CATEGORIES IN ALL THE BLACKLISTS.

Due to space limitation, we use first three characters in each category as the abbreviation in the 3rd column.

No	Category	Abbr.	#Domain	Percentage
1	Tech & Computing	Tec	13,987	27.69 %
2	Business	Bus	10,259	20.31 %
3	Non-Standard	Non	10,032	19.86 %
4	Shopping	Sho	6,179	12.23 %
5	Hobby and Interest	Hob	2,678	5.30 %
6	Travel	Tra	1,708	3.38 %
7	Education	Edu	994	1.97 %
8	Arts & Entertainment	Art	933	1.85 %
9	Food & Drink	Foo	816	1.62 %
10	Careers	Car	674	1.33 %
11	News	New	628	1.24 %
12	Personal Finance	Per	570	1.13 %
13	Automotive	Aut	446	0.88 %
14	Sports	Spo	231	0.46 %
15	Science	Sci	230	0.46 %
16	Society	Soc	78	0.15 %
17	Real Estate	Rea	76	0.15 %

F. Measure 6: Malicious Categories

1) Pre-processing and Determining the Best Algorithm: Unlike the measure 5, which has 17 labels, this measure only has 2 labels: landing (4,124 samples) and distribution (1,648

TABLE XII: TOP 5 WEB CONTENT CATEGORIES IN EACH BLACKLIST.

No	Blacklist	#Distinct Categories	1st	2nd	3rd	4th	5th
1	AMA	11	Bus	Tec	Non	Sho	Art
2	ANE	1	Bus				
3	APH	16	Tec	Bus	Non	Sho	Hob
4	ARA	3	Sho	Bus	Tec		
5	AZE	5	Tec	Bus	Sho	Hob	Art
6	AMAL	12	Bus	Tec	Non	Sho	Tra
7	AMV	17	Bus	Tec	Non	Sho	Hob
8	AHO	17	Bus	Tec	Non	Sho	Hob
9	AME	17	Tec	Non	Bus	Sho	Hob
10	ASH	17	Tec	Non	Bus	Sho	Hob
11	AUR	17	Tec	Non	Bus	Sho	Hob
12	AUT	17	Tec	Non	Bus	Sho	Hob
13	AGSBv3	14	Bus	Tec	Non	Sho	Hob
14	AGSBv4	15	Bus	Tec	Non	Hob	Sho

samples). We train the dataset using the 9 algorithms and got the results as depicted in Figure 3. Decision Tree gives the best result with 98.07% accuracy and merely 2.22% false positive rate. Therefore, Decision Tree is chosen to classify the entries in the blacklists.

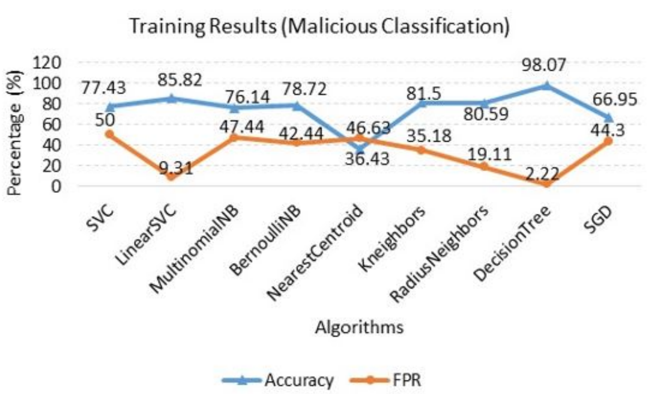

Figure 3: Accuracy and False Positive Rate of Each Algorithm

2) *Classification Result Using the Best Algorithm:* As explained above, we use Decision Tree for the classification of malicious domains. We got the results as depicted in Table XIII. Most of the blacklists contains larger number of landing domains than number of distribution domains **at least 1.5 times**. This is reasonable because a distribution domain may have multiple corresponding landing domains that redirect users to the distribution domain. Concretely, we found that the landing domains occupy at least 60% of total distinct domains in each blacklist. Especially, in the group II (large public blacklists), the landing domains occupy even larger than 75% of total distinct domains in each blacklist.

G. Measure 7: New Blacklist Version

The new updated public blacklists were downloaded on the same date 2017/11/09, which is over 7 months after the date we downloaded and analyzed the blacklists in the previous paper [1] (2017/02/28) as described in Section II-A. The new version of the blacklists is described in Table XIV. Compared

TABLE XIII: LANDING AND DISTRIBUTION IN THE BLACKLISTS.

No	Blacklist	#Distinct Domains	#Landings	#Distributions
0	Total	50,519	37,815 (74.85%)	12,704 (25.15%)
1	AMA	77	55 (71.43%)	22 (28.57%)
2	ANE	2	0 (00.00%)	2 (100.0%)
3	APH	367	234 (63.76%)	133 (36.24%)
4	ARA	3	3 (100.0%)	0 (00.00%)
5	AZE	21	14 (66.67%)	7 (33.33%)
6	AMAL	98	62 (63.27%)	36 (36.73%)
7	AMV	2,176	1,474 (67.74%)	702 (32.26%)
8	AHO	5,060	3,423 (67.65%)	1,637 (32.35%)
9	AME	19,812	15,232 (76.88%)	4,580 (23.12%)
10	ASH	32,248	24,408 (75.69%)	7,840 (24.31%)
11	AUR	33,674	25,508 (75.75%)	8,166 (24.25%)
12	AUT	24,020	18,411 (76.65%)	5,609 (23.35%)
13	AGSBv3	189	134 (70.90%)	55 (29.10%)
14	AGSBv4	639	389 (60.88%)	250 (39.12%)

TABLE XIV: NEW VERSIONS OF THE BLACKLISTS.

No	Group	Blacklists	#Domains	Downloaded on
1	(I)	MA	14,233	2017/11/09
2		NE	243	2017/11/09
3		PH	9,435	2017/11/09
4		RA	9,204	2017/11/09
5		ZE	367	2017/11/09
6		MAL	901	2017/11/09
7		MV	5,497	2017/11/09
8		HO	333,091	2017/11/09
9		ME	410,212	2017/11/09
10		UT	490,829	2017/11/09
11	(II)	SH	1,254,260	2017/11/09

Note that, the fourth column is the number of *unique* domains. Some raw blacklists contain many redundant domains.

with the previous blacklist version, this new version has several big changes:

- *UR is no longer available from 2017/07/25*. The blacklist provider of *urlblacklist.com* “has closed down, shut of its website, and thrown in the towel, they have refunded current subscribers and closed up shop” [18].
- *ME and UT no longer belong the group II (large public blacklists) but now belong to group I (small public blacklists)* since their numbers of entries are significantly reduced: 3.09 times in case of ME (from 1,266,334 entries to only 410,212 entries) and 2.74 times in case of UT (from 1,346,788 entries to only 490,829 entries). Besides ME and UT, the number of entries in MV is also significantly reduced (39.7 times from 218,248 entries to 5,497 entries); however, MV still belongs to the group I.
- On the contrary, *the numbers of entries in RA and HO are significantly increased*: 6.67 times in case of RA (from 1,380 entries to 9,204 entries) and 55.76 times in case of HO (from 5,974 entries to 333,091 entries). *For RA, which mainly focuses on ransomewares, its significant increase in the number of entries probably indicates the significant increase in the number of domains, which are infected by new ransomwares. The most recently serious ransomwares that have widely affected a lot of computers around the world are*

***WannaCry** discovered in 2017/05 [19] and a new variant of **Petya** discovered on 2017/06/27 [20]*. These two dates fall in the period between the dates of downloading our previous and new blacklist versions.

Besides the differences between the previous and new blacklist versions, we also analysed the TLDs of the new blacklist version and have some new findings:

- Unlike the previous analysis, in this analysis we found *many TLDs in the type of new generic top-level domains (new gTLD)* which is the sixth category mentioned in Section II-B2 such as: *.forsale*, *.institute*, *.church*, *.download*, etc.
- The HO is *surprisingly the blacklist having the highest number of distinct TLDs (506 TLDs), which is even much higher than the number of distinct TLDs in the blacklists that have much larger number of entries*. For example, the number of entries of SH (1,254,260) is 3.77x larger than that of HO (333,091) but the number of corresponding TLDs of SH (506) is 1.52x smaller than that of HO (332).

IV. Discussion

In this section, we discuss several issues that can be addressed in future work.

A. Blacklist Extension

In this article, we analyze 14 popular blacklists. We are planning to analyze other private blacklists. The most prioritized candidate is VirusTotal (virustotal.com). VirusTotal checks domains/URLs by referring 40 other antivirus blacklists (however, all blacklists are not always used). VirusTotal also refers the feedbacks/comments from users. Besides the blacklists and user feedbacks, we currently do not know whether it has its own method to classify a domain/URL into malicious or benign. Furthermore, we plan to extend our analysis from domain blacklists to IP, URL and DNS blacklists. Two prioritized candidates are MXTools or also known as Spamhaus (mxtools.com) and Mxtoolbox (mxtoolbox.com), which provide large number of IP entries.

B. Analysis Extension

We plan to extend our current six measures to three other interesting and important measures, which can help to understand the blacklists better:

1) Measuring the registration time of malicious domains in each blacklist: The registration time here means the response time of each blacklist to a malicious domain. For example, when a domain D becomes malicious on 2017/05/01, blacklist A lists D in its dataset on 2017/05/02 but blacklist B lists D in its dataset on 2017/05/03; and thus, A is better than B. The challenge is that, not all blacklists provide this information. A naive method is to download each blacklist periodically to check whether specific malicious domains appear in each blacklist. For example, [21] analysed the blacklist update frequency by monitoring download site. This method requires high communication costs and also cannot deal with private blacklists which do not allow to directly download blacklists. Therefore, better solutions should be investigated to analyse registration time of malicious domains in blacklists.

TABLE XV: RAW WHOIS OF THE DOMAIN 'DNI.RU'

```
% By submitting a query to RIPN's Whois Service
% you agree to abide by the following terms of use:
% http://www.ripn.net/about/servpol.html#3.2 (in Russian)
% http://www.ripn.net/about/en/servpol.html#3.2 (in English).

domain:          DNI.RU
nserver:         ns1.goodoo.ru.
nserver:         ns2.goodoo.ru.
state:           REGISTERED, DELEGATED, UNVERIFIED
org:             OOO "Dni.ru"
registrar:       RD-RU
admin-contact:   https://cp.mastername.ru/domain_feedback/
created:         2000-06-06T14:58:03Z
paid-till:       2018-06-05T21:00:00Z
free-date:       2018-07-07
source:          TCI
```

2) How to decide whether a domain is malicious based on some blacklists when each blacklist has its own ground truth: A naive method is based on *majority rule*. That is, if a domain is detected by larger than 50% number of blacklists, it can be determined as a malicious domain. Another better method is based on the weight of malicious domain in each blacklist. For example, a blacklist A weights a malicious domain D at 80% while another blacklist B weights it at 30%; then we can weight D at 55%, which is the average weight. Similar to the above analysis about registration time, the challenge is that almost all blacklists do not provide the information about malicious weighting. Therefore, finding how to weight domains in each blacklist is a promising approach to label a domain into malicious or benign.

3) Whois-text-based method for the measure 6 (malicious categories): The method used for the measure 6 (malicious categories) in this is based on some HTML elements that are commonly used in landing and distribution pages along with text mining on the entire HTML documents as described in Section II-B5. In future work, we plan to *implement another method, which has been recently published in [22] and compare with the method used in this paper. The method used in [22] is based on the text mining on entire Whois documents of the domains*. Each domain has each own registration information; and instead of extracting its Whois attributes separately (similar to the measure 3, which uses the attribute *creation date* of the Whois, or the measure 4, which uses the attribute *country* of the Whois), this new method retrieves whole raw texts of the Whois and applies text mining on them. Examples of Whois raw texts are given in Tables XV and XVI.

4) Improving the library of extracting Whois attributes: When extracting Whois attributes (i.e., *creation date* in the measure 3 and *country* in the measure 4), along with using the libraries (e.g., *python-whois* 0.6.5 used in our experiment), high manual operational cost and time-consuming computation are required due to the following challenges.

- First, the Whois information stored in different servers is very unstructured, and some Whois attributes are not always available. For example, we can observe that the Whois structures of *.ru* (Table XV) and *.kr* (Table XVI) are very different. The attributes *last update date* or *address* are not available in the Whois of 'dni.ru', but available in that of 'kddi-research.jp'.

TABLE XVI: RAW WHOIS OF THE DOMAIN 'KDDI-RESEARCH.JP'.

```
[JPRS database provides information on network administration. Its use is ]
[restricted to network administration purposes. For further information,  ]
[use 'whois -h whois.jprs.jp help'. To suppress Japanese output, add'/e'   ]
[at the end of command, e.g. 'whois -h whois.jprs.jp xxx/e'.               ]

[Domain Name]        KDDI-RESEARCH.JP

[Registrant]         KDDI R&D Laboratories, INC

[Name Server]        kddfuji.kddlabs.co.jp
[Name Server]        tao.kddlabs.co.jp
[Signing Key]

[Created on]         2016/08/01
[Expires on]         2018/08/31
[Status]             Active
[Last Updated]       2017/09/01 01:05:10 (JST)

Contact Information:
[Name]               KDDI R&D Laboratories, INC
[Email]              email@lan.kddilabs.jp
[Web Page]
[Postal code]        356-8502
[Postal Address]     2-1-15 Ohara, Fujimino-shi, Saitama,
                     356-8502, JAPAN
[Phone]              0492-78-7441
[Fax]                0492-78-7510
```

Second, some attribute values in different Whois servers are very divert. For example, the attribute *country* in different domains is registered as "US", "USA" or "America", which all have the same meaning. Third, we cannot be able to consider all attributes of all domain servers but only some common ones (e.g., creation date, expiration date, registrar, country and organization) since they do not have any standard.

- Even if we use entire Whois text of each domain instead of separately extracting each Whois attribute (as previously discussed in the item 3), there are still some other challenges. First, Whois of a domain can be stored in one or multiple Whois servers. Famous Whois servers can contain Whois of almost all of domains; but for some domains, we need to manually find its corresponding Whois server. Second, English is not always supported in Whois servers. For example, the server *whois.vnnic.vn* only supports Vietnamese, or *ewhois.cnnic.cn* only supports Chinese. Although the Whois are known to be readable-and-understandable by human, semantic language processing is required.

As far as we know, until now there is no library or automatic method which can *completely* standardize Whois information of all kinds of domains. For future work, we plan to improve the existing libraries by adding different patterns for other domains TLDs that the libraries have not supported. Also, we plan to construct English framework for some servers that do not have English Whois.

V. Conclusion

In this article, we analyse 14 popular blacklists downloaded on 2017/02/28 including 8 small public blacklists, 4 large public blacklists and 2 private blacklists by Google. We designed seven important measures including blacklist intersections, TLDs, domain ages, countries, web content categories, malicious categories, and some new findings between the current and new blacklist versions. Especially, we construct our two models using machine learning to analyse the last 2 measures. We finally found several important results: Google is developing GSBv3 and GSBv4 independently; the large public blacklist *urlblacklist.com* contains 98% entries in the blacklist *dsi.ut_capitole.fr*; most of domains in all the blacklists are created in 2000 with 6.08%, and from United States with 24.28%; GSBv4 can detect younger domains compared with other blacklists; (v) *Tech & Computing* is the dominant web content category, and the blacklists in each group have higher correlation in web content than the blacklists in other groups; and (vi) the number of landing domains is larger than that of distribution domains at least 75% in group II (large public blacklists) and at least 60% in other groups. For the final measure, we collected the most updated versions of 11 public blacklists as of this paper (downloaded on 217/11/09), and analysed the differences between the two blacklist versions. We observed some significant changes such as: UR is no longer available from 2017/07/25; ME and UT now belong to group I (small public blacklists) rather than group II (large public blacklists) as in the previous versions; the number of malicious domains injected by ransomwares is significantly increased; and many new-generic TLDs appear such as *.forsale*, *.institute*, *.church* unlike the previous analysis. We also discussed several challenges in analysing registration time of malicious domains, the way to determine a malicious domain, Whois standardization and malicious classification using text mining on entire Whois documents.

Acknowledgement

This research was carried out as part of WarpDrive: Web-based Attack Response with Practical and Deployable Research InitiatiVE, a Commissioned Research of the National Institute of Information and Communications Technology (NICT), JAPAN.

References

[1] T. P. Thao, T. Makanju, J. Urakawa, A. Yamada, K. Murakami, and A. Kubota, "Large-Scale Analysis of Domain Blacklists". In: *Proceedings of the 11th International Conference on Emerging Security Information, Systems and Technologies (SECURWARE'17)*, 2017.

[2] VeriSign, Inc., "Internet Grows to 326.4 Million Domain Names in the First Quarter of 2016". Available: https://investor.verisign.com/releaseDetail.cfm?releaseid=980215. Retrieved: 2016/07/19.

[3] Symantec, Inc. "Internet Security Threat Report". Available: https://www.symantec.com/content/dam/symantec/docs/reports/istr-21-2016-en.pdf.

[4] S. Sheng, B. Wardman, G. Warner, L. F. Cranor, and J. Hong, "An Empirical Analysis of Phishing Blacklists", *6th Conference on Email and Anti-Spam (CEAS)*, 2009.

[5] M. Kuhrer and T. Holz, "An Empirical Analysis of Malware Blacklists", *Praxis der Informationsverarbeitung und Kommunikation*, vol. 35, no. 1, p. 11, 2012.

[6] M. Kuhrer, C. Rossow, and T. Holz, "Paint it Black: Evaluating the Effectiveness of Malware Blacklists", *17th Symposium on Research in Attacks, Intrusions and Defenses (RAID)*, pp. 1-21, 2014.

[7] M. Vasek and T. Moore, "Empirical analysis of factors affecting malware URL detection", *eCrime Researchers Summit (eCRS'13)*, pp. 1-8, 2013.

[8] C. J. Dietrich, and C. Rossow, “Empirical research of IP blacklists”, *Securing Electronic Business Processes (ISSE’08)*, pp. 163-171, 2008.

[9] T. Ouyang, S. Ray, M. Allman, and M. Rabinovich, “A large-scale empirical analysis of email spam detection through network characteristics in a stand-alone enterprise”, *Journal of Computer and Telecommunications Networking*, vol. 59, pp. 101-121, 2014.

[10] J. Jung and E. Sit, “An empirical study of spam traffic and the use of DNS black lists”, *4th ACM SIGCOMM conference on Internet measurement (IMC’04)*, pp. 370-375, 2004.

[11] D. Canali, M. Cova, G. Vigna, and C. Kruegel, “Prophiler: a fast filter for the large-scale detection of malicious web pages”, *20th Conference on World wide web (WWW’11)*, pp. 197-206, 2011.

[12] R. J. Walls, E. D. Kilmer, N. Lageman, and P. D. McDaniel, “Measuring the Impact and Perception of Acceptable Advertisements”, *Internet Measurement Conference (IMC’15)*, pp. 107-120, 2015.

[13] List of IAB Categories. Available: https://www.iab.com/guidelines/iab-quality-assurance-guidelines-qag-taxonomy/. Retrieved: 2015/09/01.

[14] ICANN, “List of Top-Level Domains”. Available: http://data.iana.org/TLD/tlds-alpha-by-domain.txt

[15] Digital Arts. Available: http://www.daj.jp/en/

[16] SimilarWeb. Available: https://developer.similarweb.com/

[17] G. Wang, J. W. Stokes, C. Herley, and D. Felstead, “Detecting Malicious Landing Pages in Malware Distribution Networks”, *43rd IEEE/IFIP Conf. on Dependable Systems and Networks (DSN’13)*, pp. 1-11, 2013.

[18] B. E. Nichols, “July 25 2017 - Urlblacklist.com Is No More”, https://groups.google.com/forum/#!topic/e2guardian/7WeHpD54LE

[19] B. Brenner, “WannaCry: the ransomware worm that didn’t arrive on a phishing hook”, *Naked Security. Sophos* Retrieved 18 May 2017. Available: https://nakedsecurity.sophos.com/2017/05/17/wannacry-the-ransomware-worm-that-didnt-arrive-on-a-phishing-hook/

[20] “Global ransomware attack causes chaos”, *BBC News*. Retrieved 27 June 2017. Available: http://www.bbc.com/news/technology-40416611

[21] Y. Takeshi et al., “Analysis of Blacklist Update Frequency for Countering Malware Attacks on Websites”, *IEICE Transactions on Communications*, vol. E97-B, no. 1, pp. 76-86, 2014.

[22] T. P. Thao, A. Yamada, K. Murakami, J. Urakawa, Y. Sawaya, and A. Kubota, “Classification of Landing and Distribution Domains using Whois’s Text Mining”, *16th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (IEEE TrustCom-17)*, pp. 1-8, 2017.