What does a Domain Name say

Jelle Smits
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
j.j.smits@student.utwente.nl

ABSTRACT

Over the past decades, the amount of internet traffic has grown exponentially. In the beginning, only large organizations and businesses would have a website, nowadays the threshold to register a domain name and start a website is lower than ever before. With the fast increase in the online presence of companies, and the ease of use the internet offers, the number of phishing attacks are at a peak with an average 88,792 active phishing websites in the 3^{rd} quarter of 2019. This is the highest amount of attacks since early 2016.[4]

This research aims to identify the categories in which SLD names can be divided. In order to analyze this, three datasets with a different level of reliability are used. The focus of this research is on English SLD names, therefore Unicode domain names are filtered out and not studies in the analysis. Digit-only domain names are taken into account, although they are considered not to contain linguistic value, and therefore are not given a language score. We find that is a correlation between the length of a domain name and the likelihood of that domain name to have a malicious intent. In addition, SLDs consisting of only digits are much more likely to have malicious intent than domains which consist exclusively of letters or a combination of letters and digits. Further, we find that malicious domain names mostly have an English SLD, similar to the representation of the English language in the SLDs of domain in the Alexa top domain list.

Keywords

Domain Names, DNS, SLD, Malicious, Phishing, NLP

1. INTRODUCTION

In order to access a website, a domain name is used. These domain names are often easy to remember and, in case of a company website, they often contain the name or abbreviation of the company. However, domain names are often abused by people with malicious intent. Though methods as well as reasons to create a malicious website vary, one of the first steps of creating a malicious website starts with the domain name. A user with malicious intent could, for example, register a domain name which looks like a benign domain name or is likely to be typed in by mistake, making it likely for users to land on the malicious page. This research aims to identify patterns in the name of Second Level Domain (SLD) names. This is the part of the domain name that can be chosen freely by the registrant. These patterns can be used to help in ranking the websites based on their legitimacy and likelihood of malicious intent. This increases online security, as current anti-malware and anti-virus software focuses on a combination of blacklists, user rating, age of a website and scanning the contents of a website before accessing it.[11] Therefore, by adding another measure of page safety, this research gives the ability to classify a website as safe or unsafe, based on its domain name. Actively analyzing DNS data allows us to preemptively blacklist domains before a user visits them, ultimately making the internet a safer place.

According to research amongst 1600 full-time IT security professionals, the most pressing concerns to address regarding IT security are: preventing malware and ransomware (22% of respondents), and preventing phishing attacks (13% of respondents).[12] One of the reasons why users get to be a victim of a phishing attack is that they do not pay enough attention to the domain names of the pages they visit. Therefore, a user might enter their data on a copycat website, which has an identical body but a sole intent of obtaining login information. In the U.S. in 2018, 26% of people use mostly identical passwords for all online logins.[2] This means that when hackers have an email and password combination for one platform of a user, they are likely to be able to access the accounts of that user on other websites as well. Therefore, obtaining username-password combinations on a seemingly irrelevant website, may give malicious users access to more sensitive platforms as well.

Not all can be put on the users not paying attention, however. The creators of malicious websites are increasingly shifting their attention to abiding by industry standards to look as if they are a benign website. In the 3^{rd} quarter of 2019, 68% of phishing websites were hosted on the HTTPS protocol. This is the highest amount this measure was first started in 2015. [4] This emphasizes that it is not enough for a user to check for general security measures.

This paper is structured as follows. In section two, we will first give a background of the topic, were the relevance and actuality is illustrated. In section three, we will look at previous work on domain names, where we focus on at their approach of the analysis and their findings. In section four, we will summarize the aims of the research. Section five describes the used data, and gives a detailed description of the operations we performed on the data. In section six, we discuss and illustrate the findings. Section seven analyzes the choices made within the research and their possible effects on the results. Finally, we will present

32^{nd} Twente Student Conference on IT Jan. 31^{st}, 2020, Enschede, The Netherlands.

Figure 1. Parts of a domain[5]

the conclusions in section eight.

2. BACKGROUND

2.1 Domain Names

As the internet began to grow, it quickly became difficult to keep track of all individual IP addresses. Therefore, the Domain Name System (DNS) was created. This gave uses the means to navigate to the desired page or server using an easy-to-remember name rather than a numeric IP address. One of the important functions of the DNS is to translate the domains back to IP addresses. Another main advantage of using domain names instead of IP addresses, is that the IP address of a resource can change without the user noticing anything.[3]

The Domain Name System is effectively hierarchic database for domain names. Each *dot* signifies a new tier in the hierarchy. At the root domain, the Top Level Domain (TLD) is stored. Some examples of TLDs are the *.com*, *.org* and *.edu* domains. Each of these TLDs have references to their sub domains. Every sub domain or child can in turn have further sub domains.[3] The string left of the dot of the TLD is the Second Level Domain (SLD). The SLD together with the TLD is typically know as a Domain Name. An overview of what a domain name looks like, is shown in *Figure 1*. Instead of *subdomain*, as mentioned in the figure, this can also be regarded as a Third Level Domain.

2.2 Risks of Domain Name scam

Scammers are always trying to find new means to lure people into giving our their personal information in order to make a profit to themselves. There are different kinds of scam such as fake login pages, landing pages which ask for a verification means by credit card, pages who lure one into a mobile subscription, and a wide variety of other methods.

In multiple occasions, phishing could be prevented if users would check the domain name of the website they are on before entering personal information such as login names, passwords or credit card details. For example, if one would like to log in to their Google account in order to access their email, the user should be on a website with the domain name *google.com*, or possible one of its ccTLD alternatives, such as *google.nl*. One should be extra cautious to the position of the domain name. Scammers often use a website which has a legitimate domain within their domain name, but the actual domain of the website is not related to the domain the user expects. An example of this is shown on the image below.

In contrast to the method shown in the image above, another way users are tricked into logging in into a copycat website, is to register a domain name which looks almost identical to the domain name of the site the user intends to visit, or to register a domain name with a typing error in it which the user is likely to make, e.g. *gogle.com* instead of *google.com*, or *twiter.com* instead of *twitter.com*.

3. RELATED WORK

There is little research which is very closely affiliated with the topic of categorization and classification of domain names, with the purpose of analyzing the malicious intent. Related to the topic of Domain Categorization is *domain squatting*, or *cybersquatting*. This is the act of registering a domain name to profit from the reputation of another domain or brand name. An example of domain squatting would be to look for a TLD name for which SLD name *google* is not yet registered. One could then purchase the domain with no other intent than to sell it to Google with a profit.

Other related work is on the algorithmic generation of domain names with malicious intent. This work is more related to this research, as it also uses forms of Natural Language Processing (NLP)[14]. In their research, datasets different from the datasets for our research were used, as they used a validated dataset of non-malicious domains form an ISP, and had different criteria for malicious and non-malicious, as their focus was only on malicious intent by botnets, rather than individual domains.

Also with a research focus on Domain Squatting, research by Kintis et al, focuses on *combosquatting*, which they describe as a form of *combosquatting*, which contains a domain name of an existing website, combined with another word.[6] In order to do the research, several domains were excluded, such as *apple.com*, for example, as these kinds of domains were not exclusive because other words are formed where they are a part of (e.g. *applejuice.com*. Their research uses similar methods as intended for this research, with Python libraries for language recognition. Interesting in the research is the discussion, regarding the domain registrars, who are, according to them, in the unique position to have insight into the domains users want to register, whereas this is not available for other parties.[6] Therefore, registrars could use active fraud prevention software, helping them request more (personal) information from users who are about to register a domain which is likely to be used for any malicious intent.

Abuse of domain names has been an issue for considerable time. In 2006, Wang et al published their research on *typosquatting*. In this research, they found five types of *typosquatting*: a missing dot between parts of the domain name, characters that were left out, characters that accidentally swapped places, characters that were accidentally replace by another (mostly adjacent to the intended character) and finally characters that were wrongfully or double inserted.[13] Within the data used for the research of this paper, we can expect to find domain names which could be categorized as one of these *typosquatting* domains as well. A major issue with *typosquatting* is that larger companies who are the victim of a *typosquatted* domain might buy domain names which are likely to be entered by mistake, and forward potential customers to the intended page, in order to protect customers and their company name and reputation.[8]

Research by Nikiforakis et al, which is on the topic of Domain Squatting as well, with their focus on the topic of *soundsquatting*, which was a new topic in the literature at the time of writing.[9] They describe it as comparable to *typosquatting*. Whereas *typosquatting* was one of the first forms of domain squatting,[9] which exploits typing mistakes in domain names which users are likely to make, already mentioned in section 2.2 of this paper, *soundsquatting* is the method which focuses on domain names with words. An exploiter would then choose a word which is pronounced in the same way, but written differently. Therefore, a user who is not sure about the name of a website, could land on the page of the exploiter.

4. OBJECTIVES

The aim of this research is to identify patterns in the naming of Second Level Domain names in order to categorize them. Based on these categories, the risk of visiting a page can be assessed based on its domain name. Therefore, there will be another measure for page safety, in addition to the page reputation, which is currently often used to assess websites.[11] By implementing this measure into current anti-virus or anti-malware software, this could further improve safety on the internet for its users by preventing them from visiting pages with malicious intent.

In order to be able to assess a domain name, we need to divide them into categories. These categories will aid in comparison between the datasets. In order to be able to identify which domain belongs to a certain category, we will firstly look at the properties of the domain names that helps us distinguish them.

5. METHODOLOGY

In the following section, the research approach towards the data collection and modification will be addressed.

5.1 Dataset description

In order to be able to have reliable results and identify meaningful categories, we need sizable data. Therefore, we use three datasets of domain names. The datasets originate from the OpenINTEL project. This project is a cooperation of the University of Twente, Surfnet, SIDN (operator of the *.nl* ccTLD) and NLNet Labs. The project captures daily snapshots of sizable parts of the global DNS, in order for DNS operators and academic researchers to use.[10] The OpenINTEL project offers more than only the domain names. As the focus of this research is exclusively on the domain names, this data could also be collected from other sources, such as zone files.

The following three datasets have been used in the research:

- **Alexa**: Alexa ranks webpages based on their popularity, which they calculate based on the averages of the daily time on the site, the daily pageviews per visitor, external sites linking to the site, and others. Generally, webpages which are ranked in the Alexa dataset can be considered a safe page.
- **OpenCC**: This dataset consists of resolved data from ccTLDs which make their zone files publicly available. In this data, there is no certainty whether the sites are benign or have malicious intent. The ccTLDs in this dataset are primarily Scandinavian ccTLDs.
- **RBL**: Contains resolved DNS data from blacklists. Pages on a blacklist generally have a bad reputation. Therefore, this data-set is likely to contain the highest amount of malicious webpages.

These three datasets combined give us 3.515.446 unique domain names. Within these datasets, the OpenCC set is the largest, followed by Alexa and finally the RBL dataset. The size comparison is shown in *Figure 2*.

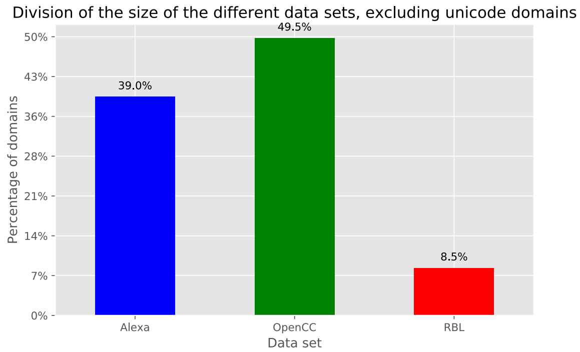

Figure 2. Amount of data in each dataset

In order to be able to conduct the analysis within the time restrictive component of this research, the datasets used are a snapshot of the day of the start of the research, November 13th 2019.

5.2 Cleaning the data

All three datasets have the same format. This format is as shown in *table 1* below. The origin will be used at a later point to be able to compare between the different datasets. The full-domain name, however, has to be split into the SLD and the TLD, as the focus of this research is upon the SLD.

Table 1. Format of the datasets

	origin	domain
0	alexa	startbzworld.com.
1	alexa	mirndv-chudesa-sveta.ru.
2	alexa	mbo99.xyz.
3	alexa	sekonsalt.ru.
4	opencc	cez.se.
5	opencc	messes.se.
6	opencc	fixmylaptop.se.
7	rbl	pandorajewelleryvip.top.
8	rbl	paxful.cf.
9	rbl	ekros.com.tr.

In addition, the domain names which will be excluded in this research can be marked accordingly. For this research, the focus will be on English domain names, and the domain names will be compared to an English dictionary. Therefore, the Unicode domains will be excluded from the research, as they have no meaning in a dictionary comparison. Unicode domains can easily be recognized by their prefix *xn–*. Although the amount of Unicode domains is increasing and more TLDs are now offered in a non-Latin alphabet extension, the domains are currently barely represented in the Alexa and RBL dataset. In the OpenCC set, however, they make up 5.5% of the domains. This is logical, as Scandinavian languages, such as Swedish and Finnish use many punctuation marks. The overall amounts of Unicode domain names throughout the datasets, is shown in *figure 3*. In *figure 2*, the Unicode domains are already substracted from each dataset, to represent the data that will be used in the research.

Apart from the difference in amount of Unicode domain names in the datasets, it became apparent that the domains in the RBL dataset are, on average, one character

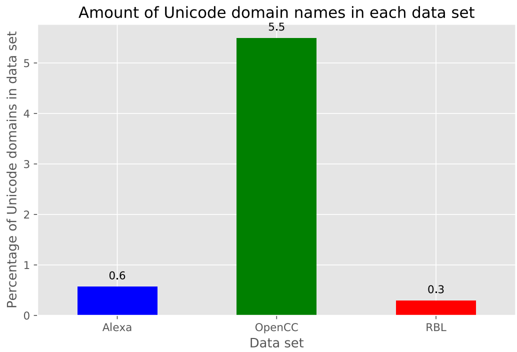

Figure 3. Amount of Unicode SLDs in the datasets, as percentage

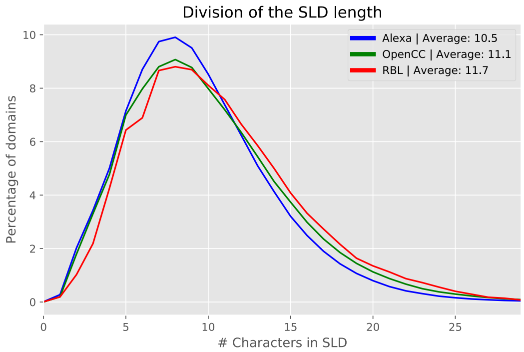

Figure 4. Length of SLD between the datasets

longer than those in the other datasets, as is shown in *figure 4*.

5.3 Data modification

After the initial cleaning of the data and familiarizing its content, several steps are taken to analyze the data in order to be able to visualize it.

To be able to focus on the SLD, the TLD is separated from the SLD in the first step. Then, the total length of the SLD is calculated and put into the data, in order to be able to compare length. As it is not possible to conduct language analysis on digits, the digit-only domain names should be labelled as the first category. Although, these do not represent a large portion of the dataset, they do not have any meaning as they are in contrast with one of the aims of DNS: using names instead of numbers, as they are easier to remember. In order to be able to categorize these domains, the amount of digits in each domain name is calculated and compared to the total length of the SLD. This gives the first insights into the data: how many domains are *Digit-only*, and how many contain any digits.

In order to assess the language of the domain names, it is necessary to separate the SLD strings into words, and clear them of any punctuation marks, as domain names do not contain spaces, but might contain punctuation marks such as '-' to separate words. The mixed domains names, containing both letters and digits, should also be taken into account. The digits will be removed from these domain names and will not be considered any further in the language analysis. The splitting algorithm will convert the SLD string into a list of most probable words. This algorithm aims to find the most likely words in the string, aiming to have as little excess characters as possible, but it respects the positions of punctuation marks and numbers. This has both advantages and disadvantages. For example, domain names who use a digit in order to replace a letter, will most probably be split into non-English words. Domains who use punctuation marks, however, will benefit from this, as the algorithm is most likely to split their domain name into the words they intended, provided the domain name is in English.

After splitting the SLD into an unfiltered list of words, separate characters, digits and punctuation marks, this list is filtered for the first time. In this first filter, we remove all punctuation marks and digits, as they have no value for the dictionary search. The resulting list is called the *unfiltered* list. This list is then filtered once again, now removing all one and two letter entries, as they have no dictionary value and are most likely to give false positives: one letter will always return true, although in almost every case, except for some cases for *a* and *I*, they have no dictionary value. This list will be addressed as the *filtered* list. This unfiltered list aims to increase the overall language score, but it affects the non-English domains in an undesired way. Therefore, both the filtered and unfiltered list will be used.

After splitting the domain names into lists words, the words are stemmed and compared against a dictionary. For each recognized word, we denote *True* or *False* in a new list, with the index identical to the word lists. This procedure is repeated for both the filtered and unfiltered lists. For the unfiltered list, which does contain one-letter entries, any one letter entry will obtain a *False* value assigned to it.

When the English word identification is completed, we give the domain name an *Language Score*, which is given by dividing the total amount of words in the domain name by the amount of words that had a *True* entry, multiplied by 100. The score is assigned in two ways, which are called *filtered score* and *unfiltered score*. The *filtered* score is obtained by only considering parts of the domain name which are more than two characters. The reason to do so is because firstly, one-character words have no meaning, as the splitting algorithm only creates these elements because it cannot form a word out of them. In addition, two-character entries are disregarded because they are most often meaningless, and in addition mostly likely to be formed by remnants which the splitting algorithm could not form a word out of. This method is called filtered, because the list of words, resulting from the splitting algorithm has the one- and two-character entries filtered out of it. The language score is calculated by taking the total amount of words recognized as an English word, divided by the length of the filtered list. The second method is an unfiltered variant of the list described above: obtaining the one- and two-character entries in addition to the longer words. The reason to analyze using both the *filtered* and *unfiltered* list is two-fold: the language score is more reliable when applying the unfiltered method when addressing non-English domain names. However, the filtered method is more reliable when looking at brand-domain names. Brands are often not a word, but rather decomposition of an existing words. Using the algorithm, it would split up the brand name. For longer brand names, this would result in a lot of excess characters, comparable to rows three, six and eight of Table 2. This would critically lower our language score.

To illustrate how the language score is calculated, we illustrate an example using a Russian domain name in table 2.

Table 2. Example operation of the splitting algorithm for SLD *mirndv-chudesa-sveta*

1st split	Filtered	Unfiltered	Is English
mir	mir	mir	Yes
nd	-	nd	No
v	-	v	No
chu	chu	chu	No
desa	desa	desa	No
s	-	s	No
vet	vet	vet	Yes
a	-	a	No
Final Score	$2/4 = \mathbf{0.5}$	$2/8 = \mathbf{0.25}$	-

In the column *1st split*, the result of applying the splitting algorithm is shown. Each row of the table represents an element of the list. The *unfiltered* list is identical to the splitting result. For the *filtered* list, all elements with a length of two or lower are removed. This is represented by a '-'. In the row *Final Score*, the language score for both the *filtered* and *unfiltered* list is calculated. As explained before, this is done by dividing the amount of elements which is are an English word (thus: have a *Yes* in column *Is English*) by the total amount of elements in the list. The table illustrates that the language score for this non-English domain is more reliable when using the unfiltered calculation method.

6. RESULTS

The data analysis, as described in section 5, has given insights in the contents of all datasets and in the relations between them. This chapter will summarize the most important findings.

As shown in *figure 4*, the domain names within the RBL dataset tend to be longer. This suggests the length of the domain name is an indicator of its malicious intent.

Within the first identified category of *Digit-only* SLDs, we found that the amount of domain names that only consist of digits is extremely limited. Within the Alexa and OpenCC dataset, respectively 0,13% and 0,18% of the domains consisted of only digits. For the RBL dataset, this percentage is more than double, with 0,42% of the domains consisting only of digits. This suggests that domain names consisting of numbers are very likely to have malicious intent.

Although not analyzed to any detail, the second category is the category of the Unicode domain names. These barely occur in the Alexa and RBL datasets, compared to the OpenCC set. The least occurrences were in the RBL dataset. This is not surprising, as users with malicious intent generally put their effort into building a method that appeals to a broad public, to increase their possibility of success.

In contrast to the occurrence of SLDs with only digits, the SLDs who occur at least one digit is a lot higher. Within the Alexa and OpenCC domains, the percentage of domains which contain at least one digit, are 7,0% and 5,85%, respectively. Within the RBL domains, this number is higher, with 8,6%. This can be explained by *typosquatting* practices, as introduced in section three. Domain names who intent to have users accidentally visit their page, might replace a character of the top row of the keyboard. This would most likely be one of the characters of the top row of the QWERTY keyboard-layout, as this is a major global standard.[1]

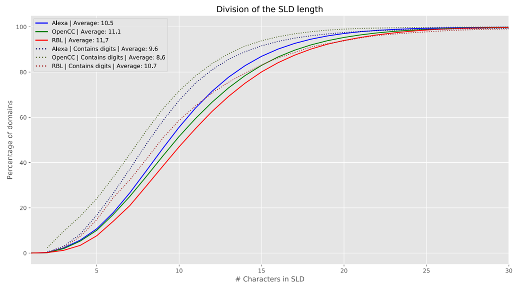

Figure 5. Length of SLD between the datasets (Cumulative)

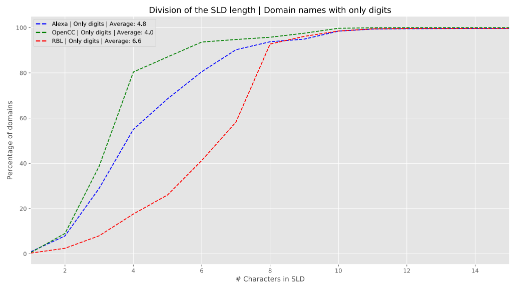

Figure 6. Length of SLD between the datasets, for digit-only SLDs (Cumulative)

As mentioned previously, the domain names have been sorted into three initial categories: *digit-only*, *containing both digits and letters* and *letter-only*. Looking at the lengths of the domain names, we can remark that the domain names containing both digits and letters are, on average, one full character shorter than their counterparts consisting only of letters, as is shown in *figure 5*. In addition, the digit-only domain names are even shorter, with the average SLD length between 4.0 and 6.6 characters. However, for the digit-only domain names, there is a significant difference between the length of the SLDs of RBL domains and the length of the SLDs in the Alexa and OpenCC datasets. Where we can confirm from *figure 5* that the domain names are shorter when the SLD consists of both digits and letters, we see that the length of SLDs consisting of only digits is, on average, half the length of a letter-only SLD. However, even though the domain names are shorter when they contain digits, the domain names within the RBL dataset are still the longest domain names in each of the three categories, as shown in *figure 6*

The Alexa dataset contains the most popular domains, which consists mostly of the most popular international TLD *.com*, and additionally of a large number of domains from large ccTLDs, such as Russia, Iran and India. These three ccTLDs occur only within the Alexa dataset in these large amounts and represent 10% of the domains total domains, where the most popular TLD *.com* represents 45%, in the Alexa dataset. For the OpenCC dataset, the ccTLD of Sweden alone makes up over 80% of its domains. For the RBL dataset, in contrast, 50% of the domains belong to the *.com* or *.net* TLDs. These two TLDs generally represent English websites, so they are more prone to have an English SLD. Because of the large representation of the general TLDs in the Alexa and RBL dataset, we see a similar pattern, where both have a high English score.

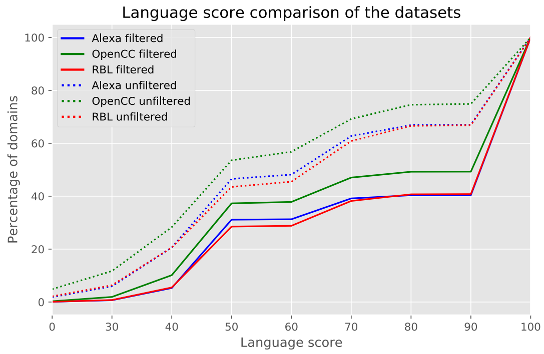

Figure 7. Language score given to the SLDs (Cumulative)

As for the amount of English detected in the SLDs, the results were relatively consistent through the database. Most language scores were at or around 0.5, 0.7 and 1.0. Interestingly, though the datasets combined are of considerable size, there were barely (less than 1%) scores of 0.0 - 0.2, 0.6 and 0.9 in all datasets. The data is shown in figure 7. As can be seen, the lines are relatively close together. It can be marked that the OpenCC dataset has the lowest mean language score. This is no surprise, as the OpenCC dataset consists of domains from ccTLD DNS providers from countries such as Sweden and Finland, whose native language is not English. Therefore, many domains in their native language can be expected in the OpenCC dataset. In contrast, however, the RBL dataset has most entries with a language score of 1 or above. This, once again, can be explained by the people who want to exploit a domain name for malicious use, tend to appeal to a public as wide as possible. Since English is one of the most common languages on the internet, with 25% of users [7], domains in English are prone to be the number one choice for malicious web pages.

7. DISCUSSION

Within the analysis, several design choices have been made to prepare the data. Taking into account these choices, their effect on the research is discussed point by point below.

String splitting method.

In order to extract words from the SLD, an algorithm has been used which tries to split the string in order to maximize the amount of word. Characters which it cannot place within a word, remain in the split result as a separate character, and are discarded afterwards. In addition, numbers are removed, i.e. *goog1e* would be split into *goog*, *1* and *e*, giving an overall language score of zero. Therefore, this research does not take domain names into account, as a human user would do in their browsing experience. In further research, this could be taken into account, by checking the domain name against a list of known trademarks and brand names.

Blacklist bias.

Within this research, domains in the *RBL* dataset are considered to have malicious intent. The main reason to use blacklists, is that these are best available way to assess the domain names in this research, without having to assess every individual page manually. It should be noted, however, that blacklists do contain false positives, and therefore benign domains might be in the blacklist, as is generally hard for a domain to remove themselves from blacklists one a domain or IP address has been blacklisted []. This should be taken into account when the *RBL* dataset is addressed as containing mostly malicious domains.

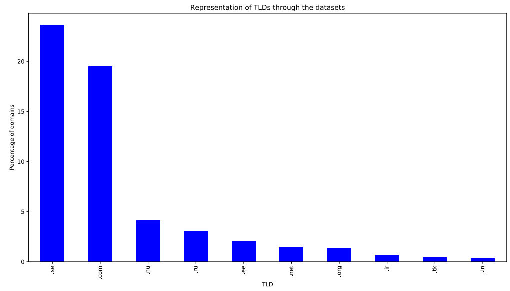

Figure 8. Top 10 TLD representations in the datasets

Dataset sizes.

The three datasets that have been used had a large number of data. Over 50% of this data, however, was within the OpenCC dataset. Within the OpenCC, more than 80% were Swedish domains. The RBL dataset was far smaller, making up only 8.5% of the total amount of data. Therefore, the results might look different if the data would focus on either more international domains, compared to specific, non-English ccTLDs, or if more ccTLDs in general would be considered. In addition, adding the amount of domains with a *bad* reputation, could confirm the claims made in this research. Future work could focus on data sets with domain names with a larger representation of other TLDs. As a comparison, the representations of the TLDs used in this research is shown in *figure 8.*

Unicode domains.

In this research, the Unicode domains have not been taken into account. Unicode domain names are still relatively new and therefore little represented in the datasets, and even less in the Alexa most popular domains. Malicious Unicode domains can be expected to see an increase when the overall amount of Unicode domains rises. One of the main advantages for the exploiter would be that they could offer a domain in the language of choice of its targets, though the targeted group has to be very specific: generally country-bound.

Digits as letters.

As mentioned in the methodology part, this research focused on the domain names as they were given, and assumed that digits did not have any linguistic meaning. However, it is common in English language to abbreviate some words by a digit, or to use a digit instead of a letter. For example, a *1* could replace a lowercase *L*, or the word *for* can be replace by *4*. Future research can take these meanings of digits in the word strings into account, as this might increase the amount of English domains that are recognized.

8. CONCLUSION

From the data analysis and its results, there are several lessons to be learned. As for now, Unicode domain names do not occur in significant numbers among the most popular web pages, and only have a limited representation in the data of malicious web pages for this reason.

We found a cohesion between the length of the SLD and the presence of digits: if the domain name contains digits, it tends to be shorter. If a SLD consists exclusively of digits, it is much shorter, with an average of 5 characters, as opposed to the average of 11 characters for letter-only domain names. In addition, we found the digit-only SLDs within the RBL dataset are significantly longer than the digit-only SLDs in the Alexa and OpenCC datasets, with the RBL average being 2 characters above the average of both Alexa and OpenCC. Therefore, the data suggests that a longer domain name, its SLD consisting either exclusively or partially out of digits, is more likely to have a malicious intent.

Additionally, we found that a SLD consisting exclusively out of digits, regardless of its length, is twice more likely to have malicious intent than a domain name consisting of only letters, or a mixture of letters and digits, as the share of digit-only SLDs is twice as high in the RBL dataset.

Other SLDs, who contain at least one number, have a slightly higher representation in our RBL dataset than in the other datasets. This could be explained by *typosquatting* practices, where a domain name with a number, instead of a symbol on the top row of the keyboard, is used in order to try people to get to visit the page by making a typing error in trying to visit a known benign page.

Regarding the language of the domain names, we see that both the RBL and the Alexa datasets have the highest average English score. For the RBL dataset, this can be explained by the exploiters of the domains trying to appeal to the broadest possible public. For the Alexa dataset, it is also logical for it to have a high language score, as it lists the global most popular webpages, and English is a world language.

For this topic, further research should be done into the brand names within domain names. In this research, they have been disregarded and analysed with the English language criteria, but assigning them to a different category could shine a light on the duplication of brand domains in the RBL dataset, and additionally give more reliable language score, by not trying to score them as a word.

In addition, the numerical domains, as well as the domains who contain numbers, could be regarded as containing more information than has been done in this research. As discussed previously, a number can be used to replace a letter, taking this into account could change the language score. It could also be possible that this practice is currently done by malicious domains, as we did notice an increased amount of domain names which contain a letter in our current data.

As for this research, there was no information available about the domain, other than its name and what dataset it originated from. For future work, one could check the domain and its designated IP-address against several blacklists, to be able to more reliably tag the intent of the domain name as malicious or benign.

9. REFERENCES

[1] W. B. Arthur. Comment on neil kay's paper—'rerun the tape of history and qwerty always wins'. *Research Policy*, 42(6):1186 – 1187, 2013.

[2] J. Clement. U.s. online password repetition 2018, Nov 2018.

[3] M. Dooley and T. Rooney. Introduction to the domain name system (dns). *DNS Security Management*, page 17–29, 2017.

[4] A. Gregg and APWG. Phishing Activity Trends Report 3rd Quarter 2019. Technical Report November, APWG, 2019.

[5] Hover. *An image outlining the parts of a domain name.* Hover, Dec 2014.

[6] P. Kintis, N. Miramirkhani, C. Lever, Y. Chen, R. Romero-Gómez, N. Pitropakis, N. Nikiforakis, and M. Antonakakis. Hiding in plain sight: A longitudinal study of combosquatting abuse. pages 569–586. Association for Computing Machinery, 2017. cited By 21.

[7] Nielsen. Internet: most common languages online 2019, Jul 2019.

[8] N. Nikiforakis, S. Acker van, W. Meert, L. Desmet, F. Piessens, and W. Joosen. Bitsquatting: Exploiting bit-flips for fun, or profit? May 2013.

[9] N. Nikiforakis, M. Balduzzi, L. Desmet, F. Piessens, and W. Joosen. Soundsquatting: Uncovering the use of homophones in domain squatting. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8783:291–308, 2014. cited By 15.

[10] OpenINTEL. Background, 2019.

[11] Symantec Corporation. Browse the Internet securely with Norton Safe Web, 2019.

[12] Trustwave. Which it security tasks are you facing the most pressure to address? [graph]. Technical report, Statista, May 2018. Retrieved November 29, 2019.

[13] Y.-M. Wang, D. Beck, J. Wang, C. Verbowski, and B. Daniels. Strider typo-patrol: Discovery and analysis of systematic typo-squatting. 07 2006.

[14] S. Yadav, A. K. K. Reddy, A. N. Reddy, and S. Ranjan. Detecting algorithmically generated malicious domain names. *Proceedings of the 10th annual conference on Internet measurement - IMC 10*, 2010.