Towards more rigorous domain-based metrics: quantifying the prevalence and implications of "Active" Domains

1st Siôn Lloyd
ICANN
sion.lloyd@icann.org

2nd Carlos Hernandez-Gañan
ICANN
carlos.ganan@icann.org

3rd Samaneh Tajalizadehkhoob
ICANN
samaneh.tajali@icann.org

***Abstract*—The Domain Name System (DNS) is a critical component of the internet infrastructure. As such it is often the subject of various measurements with a view to quantifying different aspects of its use. Some of these measurements cover legitimate uses; however, identifying any threats associated with domain names has also become a vital task in enhancing DNS security. Current abuse metrics used for identifying malicious domains typically rely on the count of domains listed on Reputation Blocklists and are normalized by the size of the zone for registries or domains under management for registrars. However, these metrics are imprecise and do not account for whether the domain name is resolvable or serves active content. In this paper, we propose a novel approach to identify active domains, which account for domains that serve actual content under the control of the registrant. We demonstrate the proportions of inactive, active, and non-resolving domains across different samples of the name space. Our findings suggest that current normalized metrics are not necessarily giving a true picture of the underlying situation. By introducing a more precise classification system for domains, we show how this can lead to more reliable and robust metrics that can, for example, enhance DNS security by enabling a more thorough analysis of active domains. We also discuss the implications of these findings for registries and registrars, highlighting how they can use this information to combat domain abuse more effectively.**

1. Introduction

The Domain Name System (DNS) plays a vital role in the functioning of the Internet by providing naming services that allow users to access specific online resources using easy-to-remember domain names. As of December 2022, there were more than 349 million domain names registered across all top-level domains (TLDs) [32]. Despite the significance of DNS, it has been increasingly exploited by cybercriminals to launch various malicious activities such as phishing, malware distribution, and other forms of cyberattacks. Therefore, accurately identifying the nature of a domain name has become a critical task for enhancing DNS security.

Multiple studies design metrics to report on the concentration of abusive domains within the scope of different actors who share responsibility in combating abuse such as registries or registrars [5], [12], [13], [25]. The goal of such studies and practices is to use metrics to highlight actors who perform better in terms of fighting abuse and identify their effective policies.

That said, existing abuse metrics typically take the count of domains listed on Reputation Blocklists (RBLs) [4], [15], [17], [28] and, in order to make comparisons between populations of different sizes, are normalized by the size of the zone for registries, or domains under management (DUM) for registrars. We argue that such metrics might not be very precise, and do not account for whether the whole name space is resolvable and whether the domains in question serve active content under the control of the registrant or non-specific content from a service provider. As a simple example of why this might be an important distinction, we could imagine a study to see the proportion of domains which have a valid SSL certificate. It could be argued that domains with no active content gain little benefit from a certificate and so could reasonably be excluded from any metrics (or at least be considered separately).

The distinction can also have a significant operational impact as, assuming that a domain has been reported for some form of abuse, registrars or hosting providers may be required to take action to mitigate the issue. Removing inactive domains from any investigations can save time and allow for a more thorough analysis of the remaining domains.

The main objective of this paper is to propose a method for creating more rigorous domain-based metrics for domain name registries and registrars. The proposed approach involves developing a rule-based classifier that can differentiate between three distinct categories of domains: no-IP, inactive, and active. The first category refers to domains where no IP address can be found, while the second category includes domains that display generic or partially targeted advertisements (a.k.a. "parked" pages), suspended pages, directory listings, or error pages. The third category comprises domains that do not fall into the first two categories and can be accessed via HTTP requests without any errors.

Note that while we concentrate on the effect on abuse metrics as these tend to be the most serious, the method we propose can equally be applied to any measurements based on populations of domains.

In short, the main contributions of this paper are:

- Design of a method to classify domain names into three categories –no-IP, active, and inactive– using DNS-based and URL-based markers.

- Analysis of different samples of the domain name space, providing insights into the proportions of each category observed.
- Demonstration of the impact on metrics and rankings for registries and registrars before and after the identification of categories introduced in this paper.

2. Related Work

Parked domains, which form the vast majority of inactive domains, have been topic of multiple previous studies ranging from mechanisms for discovering them [33], [34], [36] to analyzing the content for security threats [2], [19]. Other groups have looked at the temporal characteristics of parked domains [29] to see how many become malicious after being parked. There have also been studies going the other way, starting with domains claimed to contain security threats and looking at the proportion which are parked or unregistered [16]. Note that some work uses a different definition of parking, for example [18] describes parking as where a domain "resolves to an IP not controlled by the domain owner". This is different to the definition we, and other groups, have used which considers "non user-centric content" [36].

Spamhaus publishes a league table of the most abused TLDs [25]; the data is normalized by looking at abusive and legitimate total seen in their data.

The most relevant work with respect to our research is [36] as we have incorporated the markers which they have published into our discovery model. The authors report 23% of domain names as parked (looking at the general population of registered domains). This is slightly lower than the 28% we see (§5.1), possibly due to our use of final URLs catching more examples. We build on that work to look at how these classifications can impact metrics involving groups of domains.

It is also worth stating that is this paper we use the definition of DNS Abuse used in the "DNS Abuse Framework" as referenced by the OECD [20]

3. Data Collection

Our analysis is based on data collected from a variety of sources, including domain zone files as well as data from our own active DNS measurements. To quantify the implication of our findings, we also collect various Reputation Block Lists (RBLs). In this section, we provide an overview of the data collected and their corresponding sources.

3.1. Zone Files

When a registrant purchases a new domain from a registrar, the registrar sends a request to the registry with the relevant domain and name server information. Subsequently, the domain appears in the corresponding TLD's zone file, which, at a high level, reflects a DNS server's anticipated answers to DNS queries. For a domain to resolve, it must have name server information in the zone file and those name servers must be configured to answer appropriately for that domain.

Internet Corporation for Assigned Names and Numbers (ICANN) requires most generic top-level domain (gTLD) registries to provide access to zone files for research and other purposes. However, some registries, and all country code top-level domains (ccTLDs), are exempt from this requirement. In anticipation of the rapid expansion of TLDs, ICANN developed a solution called the Centralized Zone Data Service (CZDS)[1], through which third parties can apply for accounts to access multiple zone files. Once the registry approves access, users can download the zone file through a simple API call up to once per day. We downloaded zone files from CZDS on February 14th and 15th 2023.

3.2. Active DNS Measurements

As part of our data gathering process, we used the `dnspython` library to perform a set of DNS queries for each domain. This allows us to actively gather DNS data for each domain, following CNAME and NS records until we either locate an A or AAAA record or determine that no such record exists. We also keep a record of NS and CNAME records found for a domain.

3.3. Registration Data

Registry operators for most TLDs are required to publicly provide accurate domain registration data using the WHOIS protocol, which is actively maintained. However, imposed quotas on queries make using this service for bulk lookups prohibitive. We used WHOIS lookups where we had lists of domains of a reasonable size; however for finding the full size of a registrar we used the data provided by the Domain Name Stat website [26].

3.4. Reputation Blocklists

We collected domain reputation blocklists (RBLs) to quantify the impact of different types of domains on abuse metrics. We utilize multiple abuse feeds that are provided to us by reputable organizations such as Spamhaus [24], APWG [11], SURBL [27], WMC Global [7], Phishtank [23], openphish [21], URLHaus [30], Global Cyber Alliance [1] and phishstats [22]. These blocklists contain domains that are highly likely to be malicious, while some other lists include domains detected through more experimental techniques. They consist of data related to various types of abuse, including malware, phishing, and spam.

More information on these sources can be found in appendix A and on their websites.

4. Methodology

Our method consists of two steps: (i) we characterize each domain by extracting a set of DNS features, and (ii) we use these features against a rule-based classifier to determine the category of each domain. The rule-based classifier leverages markers that split these domains into three categories:

1. https://czds.icann.org

- **no-IP:** domain where no IP address can be found;
- **Inactive:** domain that resolves to an IP address, but the website associated with the domain displays generic or partially targeted advertisements ("parked" pages), suspended pages, pages showing directory listings, or error messages such as a `404` error page;
- **Active:** domain that has not been classified as no-IP or inactive. An active domain indicates that an IP address was found and some data could be retrieved through an HTTP request. This means that the website associated with the domain is accessible and operational, however it does not necessarily indicate "useful" content;

4.1. Discovering Domain Markers

Previous work [36] shows that it is possible to define a set of markers that can be used to classify a domain as parked without looking at the actual content of any pages retrieved. This is mainly because there are relatively few services responsible for hosting[2] a large proportion of parked domains. These services use fixed infrastructure and so aspects of the information that define a domain, for example the name servers which it has configured, can associate the domain to the service. We have also discovered markers which indicate other inactive states like suspended or domains being auctioned.

The important aspect is that the marker can unambiguously classify a domain into one of the categories we define in this paper. Where, for example, infrastructure hosts both active and inactive content then we can not use that marker without seeing false positives.

DNS-Based Markers. The DNS-based markers we use are the name server records (known as "NS" records), the IP addresses retrieved (known as "A" records when referring to IPv4 addresses and "AAAA" records when they are IPv6 addresses) and a DNS redirection record known as a "CNAME". For NS records we look for those associated with parking services, for example, a domain with name servers belonging to *parkingcrew* (*e.g.* "ns1.parkingcrew.net") is under the management of their parking service and will display their adverts. Other forms of inactive site can be discovered via the NS records, for example some registrars suspend domains by moving them to specific name servers, often with terms like "verification" or "suspended" in them. For IP addresses we found some services publish information on how to park a domain with them by setting the IP address to a specific value. For example, GoDaddy require the IP address to be set to 34.102.136.180 [8]. CNAME records act as a redirection in DNS and so are a convenient way to set multiple parameters with a single record. Some domains are parked this way, for example setting a CNAME of "parkingpage.namecheap.com" will cause that domain to inherit all the records of the namecheap parking infrastructure.

Most of our markers were discovered through searching the documentation of well known services and manual inspection, over a year, of domains that we observed to be parked. Where we can confirm that the infrastructure only hosts parked domains then we keep those markers, where we see potential false positives we remove them from our list. Some care needed to be taken compiling this list, for example it is not possible to just search for the term "parking" in the NS records because of the false positives seen.

Many of these DNS-based markers have been discovered and made available by other groups [33], [36]. In particular we had substantial overlap with the markers published by [36] and added the remaining markers to our list; while we are not ready to publish our code yet, their data can be can be downloaded as a json file [35].

URL Markers. The use of just DNS markers is sufficient if infrastructure contains only inactive domains; however, where an IP or a name server is responsible for a mixture of content, then DNS markers alone are not enough. In addition to the DNS based markers, but still without looking at actual content, information can sometimes be obtained from the final URL after any redirections are followed. Oftentimes the lexical features of this URL itself can provide useful data, in particular where domains are redirected to a page indicating the domain is available for sale or suspended. For example, a domain being sold by the service "dan.com" will redirect to "https://dan.com/buy-domain/[DOMAIN]"; a page displaying details about how the domain can be bought. It is also useful to put this final URL back through the DNS gathering process where the domain has changed from the original, even the addition of a "www" subdomain can change the markers seen. Note though that a simple "get" request (*e.g.* using the python "requests" library) is often not enough. Javascript redirects require something more like a full headless browser to work; in our work the final URL was found by using URLScan.io [31] to scan the domain.

4.2. Domain Characterization

Given a domain name, we first collect DNS data for it. Minimally the set of name servers and IP addresses (both v4 and v6) are collected along with any CNAME records.

Where we found at least one IP address, classification could be run on this DNS data; however, better results are obtained if i) the final URL is determined (preferably via a mechanism that can follow javascript re-directions) and ii) the DNS data for the domain part of the URL can be gathered (if it is different from the starting domain).

Note that at no point is the actual content of a webpage used in the process, the basic data points that need to be gathered for domain are thus DNS-based:

- The name servers
- The resultant IP addresses
- CNAME records

These data points are straightforward to capture and don't require any code to be downloaded and so are safe to collect, even for RBL entries believed to be malicious. Additional data can also be collected, covering the URL:

- The endpoint URL after any redirects
- DNS data for final URL (name servers, IP addresses and CNAMEs)

These data points are then passed on the the classifier.

2. either via a configured IP address or by assigning specific nameservers

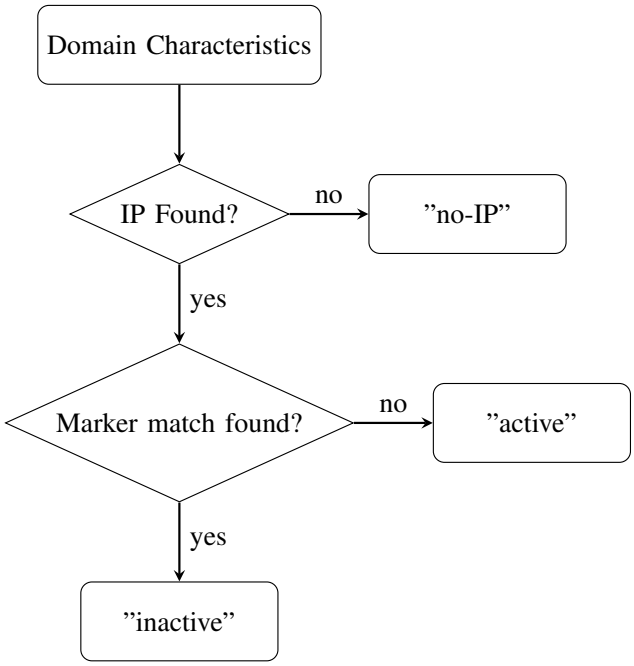

Figure 1. Flowchart depicting the domain classification process.

4.3. Rule-Based Classifier

The first rule of our classifier is to determine whether a domain resolves. If we were unable to find at least one IP address of either variety then the domain will be classified as “no-IP”. Note that other DNS records may exist for the domain, such as ”MX” records used for email, although in our data this was only seen in around 2% of cases. Note also that in our current setup we only check once and from a single vantage point; so it is possible that we are seeing temporary resolution failures.

Next, the features of each resolvable domain as described in section 4.2 are compared against the sets of markers discovered in section 4.1. In our system a domain is classified as inactive if any match is found, *i.e.* we don’t require all observed data points to be in our lists of markers.

The classifier can thus be visualised as shown in Figure 1.

4.3.1. Classifier Evaluation. We manually evaluate the performance of our rule-based classifier by using a random set of newly registered domain names. Taking a sample of 250 recent registrations as a test set from the zone files, we use the rule-base classifier to categorize them and then manually validate the results from what URLScan shows.

This is not always simple, as some cases require a translation of language in order to understand, and some cases could be argued either way (for example, is a page showing an empty blog template active or not). In the end 32 of the 250 were not categorized due to not showing any content to allow a validation and 38 did not resolve at all. For the remaining 180 the confusion matrix looks like Table 1

This gives us an accuracy of 0.87, a sensitivity of 0.74 and a precision of 0.99 as shown in table 2. The low sensitivity indicates the relatively large number of inactive

TABLE 1. CONFUSION MATRIX

	Prediction		
Detection	Inactive	Active	Total
Inactive	66 (True Positive)	23 (False Negative)	89
Active	1 (False Positive)	90 (True Negative)	91

TABLE 2. CLASSIFIER EVALUATION

Accuracy	**Precision**	**Sensitivity**	**Specificity**	**G-Mean**
0.87	0.99	0.74	0.99	0.86

pages which are classified as active. The high precision indicates the low number of false positives seen.

The one false positive in our sample is from a domain which has two IP addresses, visiting one gives a parked page while the other gives content; so it is explainable and suggests a highly conservative approach of insisting that all markers evaluate to true. Or maybe all markers of a particular type, so for example if all IP addresses are on our list. It is to be seen how much that would change the numbers; but it would remove this particular false positive, presumably at the cost of reducing the sensitivity.

5. Analysis of Domain Categories

Here we present the results of applying our method to different populations of domains. Firstly, we look at a set of registrations selected at random from the gTLD zone files, *i.e.*, our baseline measurement. We then look at them grouped by TLD. Finally we look at measurements of domains listed on RBLs, grouped by TLD and registrars.

For the last category we demonstrate the application of our classifier, showing how abuse metrics would be before identifying their active domain name space and after. The goal of this analysis is to show how different populations have different proportions of domains in the different categories discussed, and so accounting for these different categories can have a significant impact on metrics, abuse being one of many possible.

5.1. Baseline Measurements

We look first at a set of domains chosen at random from the gTLD zone files (data collected on January 23rd 2023). By downloading the zone files it is possible to obtain a snapshot of all the gTLD zones as they appeared in the previous 24 hours. From this data we took ten random samples of 1,000 domains each and classified them. The results can be seen in Figure 2 which shows that only around 57% show active content. We see about 29% appear to be inactive (28% classify as parked, the other 1% being suspended, *etc.*) and 14% do not resolve at all. As a side note it is slightly surprising to us that we see so many domains which do not resolve as we start from copies of zone files so we know these domains all have name server records configured. We see around 2% with other (non-IP) DNS records, the remaining cases could be due to the time difference between the zone file data and

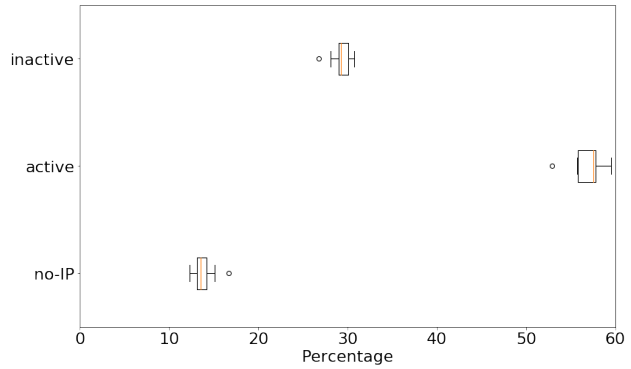

Figure 2. Percentages of domains classified into each category from ten batches of gTLD domains selected at random.

TABLE 3. MEASUREMENTS FROM A SELECTION OF GTLDS

	Status (%)		
TLD	no-IP	Active	Inactive
.berlin	5.6	90.3	4.2
.art	14.2	60.1	23.3
.finance	15.5	42.1	41.9
.llc	31.5	36.4	31.3
.earth	10.4	63.7	25.6
.life	10.6	50.6	37.5
.cyou	54.0	34.0	12.0

the DNS measurements being made, although we also see instances of name servers returning an error response or not responding at all (timeout).

5.2. Domain Metrics for Registries

When looking at individual TLDs, an obvious metric to measure is the number of registrations that TLD contains [6], [9]. These measurements can be used to normalize other metrics to give them as proportions of the zone; this allows comparisons between TLDs of very different sizes.

While the above approach is entirely valid, it might also be useful to normalize to the number of active domains in each zone. If different TLDs show different proportions of active domains, then their relative normalization factors to one another will change. For example, if we look at Table 3, we see zones with high percentages of resolving domains (*i.e.* a low percentage of no-IP) but very different percentages of active domains (.art vs .finance). We see zones of similar sizes but very different percentages of resolving and active domains (.earth vs .llc); zones of very different sizes, similar resolving percentages but very different active percentages (.earth vs .life).

5.3. Abuse Domain Metrics for Registries and Registrars

Another interesting population of domains that get examined regularly is that of domains appearing on an RBL [10], [13], [15]–[17]. These are domains which have been reported as being involved in some form of abusive activity, although the focus and collection methods vary between providers.

This data is often looked at per TLD [5], [12], [14]; again, commonly using the full zone size to normalise the data. We put current RBL data through our classifer, and again we see marked differences in the numbers of non-resolving, active and inactive domains covering most scenarios; see table 4. This will change the observed abuse proportions in the different TLDs.

TABLE 4. MEASUREMENTS FROM A SELECTION OF TLDS, WITH MEAN AND STANDARD DEVIATION FROM THE FULL POPULATION

	Status (%)		
TLD	no-IP	Active	Inactive
buzz	73.9	23.4	2.8
website	59.6	37.4	3.0
site	51.5	38.8	9.8
link	49.1	49.7	1.2
shop	27.9	67.7	4.5
xyz	23.1	59.0	17.9
cyou	19.9	79.1	1.0
com	13.1	73.6	13.3
store	11.4	67.6	20.6
org	9.8	61.2	29.0
club	3.1	90.7	6.2
From Full Sample (%)			
Mean	29.8	62.0	8.2
S.D.	19.6	18.2	7.9

Another common angle for work on RBLs is to group the entries by the sponsoring registrar to see which have more abusive domains under management than the average. However, the number of reported domains does not tell the whole story. Registrars take action on abusive domains in different ways, some will suspend domains from DNS, others will redirect to a holding page (maybe requesting additional owner verification or saying the domain is suspended). Allowing for these categories may, to some extent, correct metrics for domains which have already been dealt with by the registrars.

Shown in table 5 are the percentages for each category seen in domains appearing on RBLs for a sample of registrars. We again see a wide variety of cases; from almost half of the domains not resolving to nearly 99% appearing to show active content.

TABLE 5. MEASUREMENTS FROM A SELECTION OF REGISTRARS, WITH MEAN AND STANDARD DEVIATION FROM THE FULL POPULATION

	Status (%)		
Registrar	no-IP	Active	Inactive
01	17.9	80.4	1.7
02	33.1	51.4	15.6
03	19.9	78.1	2.0
04	25.5	68.0	6.5
05	13.4	86.6	0.0
06	47.2	46.7	6.1
07	25.5	67.7	6.7
08	0.4	98.6	1.1
09	26.0	58.4	15.6
10	3.8	59.3	36.9
From Full Sample (%)			
Mean	16.6	73.8	9.6
S.D.	11.4	13.3	11.4

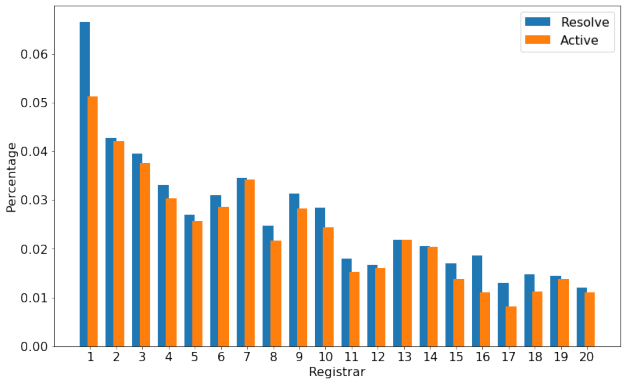

Figure 3. Percentages of registrar DUM on RBLs ordered by Raw, showing Resolving and Active

As with the corrections made to gTLDs; if we use the above normalization when creating metrics or league tables of registrars there is change compared to using the raw count. Although not always large there is movement in/out of the top 5, top 10 and top 20 which could change which registrars are called-out as having more abusive domains under management. Consider for example the data shown in Figure 3 which shows the resolving and active percentages of RBL reported domains for a group of registrars. The registrars are ordered by the raw percentages (not shown), but we can see that if the ordering were done by resolving or active percentages (shown in blue and orange respectively) then the order would change. For example, the registrar in position seven would move up into fourth, that in fifth would move to ninth and so on. This could effect any prioritisation of where resources should be used to most efficiently address any issues.

6. Ethical Considerations

The data used in this study was obtained through active and passive internet measurements, which were carried out in compliance with the principles outlined in the Menlo Report [3]. We took steps to ensure that our measurements did not interfere with the normal functioning of any network, and did not compromise the privacy or security of any operators involved.

7. Limitations

As has already been mentioned, the queries to determine if a domain resolves are made once and from a single vantage point. This means that we are recording a single observation and so it is possible that we are catching a domain during a temporary interruption to its service. This will raise the proportion seen as "no-IP"; however, it is not clear how large an effect his might be without further investigation.

We also do not look further into other DNS record types like "MX" which may indicate a domains' non-web use. In our data around 2% of domains had no IP address but did have an MX record, often a default record which we have not validated for its actual use. Similarly other "non-web" uses of domains or HTTP content not served from the default port or at the root URL of a domain will be miss-classified as inactive. We aim to understand these cases, quantify how large their contributions might be and hopefully introduce ways to classify them correctly.

Our classifier also presents several limitations. Firstly, it does not leverage content-based markers. It is not clear how quickly adding content-based identification would improve the sensitivity; it may be that a small number of indicators would cover a significant number of cases, or it may require many new indicators in order to make a noticeable improvement.

Secondly, it is not yet clear if the DNS markers are static in nature or if they will change over time. New services may become popular, or services may change the nameservers or IP ranges used; which, in the event of these markers then being used to host active content, would result in false positives. If this happened to one of the more popular services then the effect would be significant.

We validated our classifier over a set of recently registered domains as that was the focus of our initial work. While we have monitored for false positives since then we could also validate our classifier over more data sets to make sure it is consistent across more diverse inputs.

A final issue to consider is whether different clients would see different behaviors. It is possible, for example where javascript re-directions occur, that the source IP address or user-agent string is used to direct certain geographies to different end points. We have anecdotal evidence of this technique being used to hide targeted phishing attacks, making them harder to discover.

8. Conclusions & Future Work

In this paper, we have proposed and evaluate a rule-based classifier to assess a domain into one of three categories; namely no-IP, active and inactive. Based on this classification, we have analyzed and compared the implications of normalizing populations of domain names. Specifically, we have examined how using active domains as a baseline can change the way different populations are compared and how corrections can be made to account for differences in registrar practices.

When measuring the size of a population of domains, removing those that do not resolve or those that do not contain active content can change the picture of the scale of any problems.

Furthermore, we have examined how different registrars take action on abusive domains in different ways, and how normalizing metrics or league tables by accounting for these differences can change which registrars are listed as having more abusive domains under management.

Looking ahead we have begun to make measurements on new domain registrations looking at how their classifications change over time. While the overall trends are likely dominated by registrar practices like grace periods *etc.* the behavior of individual domains may contain interesting information of the use of that domain. This aspect requires more investigation before useful conclusions can be drawn.

References

[1] Global Cyber Alliance. GCA Domain Trust website. https://www.globalcyberalliance.org/domain-trust/.

[2] Sumayah A. Alrwais, Kan Yuan, Eihal Alowaisheq, Zhou Li, and Xiaofeng Wang. Understanding the dark side of domain parking. In *USENIX Security Symposium*, 2014.

[3] Michael Bailey, David Dittrich, Erin Kenneally, and Doug Maughan. The menlo report. *IEEE Security & Privacy*, 10(2):71–75, 2012.

[4] Jan Bayer, Yevheniya Nosyk, Olivier Hureau, Simon Fernandez, Ivett Paulovics, Andrzej Duda, and Maciej Korczyński. Study on domain name system (dns) abuse: Technical report. *arXiv preprint arXiv:2212.08879*, 2022.

[5] CSC David Barnett. The highest threat tlds - part 1. https://circleid.com/posts/20230112-the-highest-threat-tlds-part-1.

[6] DomainTools. Domain count statistics for tlds. https://research.domaintools.com/statistics/tld-counts/.

[7] WMC Global. WMC Global website. https://www.wmcglobal.com/phishfeed/.

[8] GoDaddy. Park a domain registered with godaddy. https://uk.godaddy.com/help/park-a-domain-registered-with-godaddy-23936.

[9] greenSec GmbH. new gtld summary. https://ntldstats.com/.

[10] Harm Griffioen, Tim Booij, and Christian Doerr. Quality evaluation of cyber threat intelligence feeds. In Mauro Conti, Jianying Zhou, Emiliano Casalicchio, and Angelo Spognardi, editors, *Applied Cryptography and Network Security*, pages 277–296, Cham, 2020. Springer International Publishing.

[11] Anti-Phishing Working Group. APWG website. https://apwg.org/.

[12] ICANN. Domain abuse activity reporting. https://www.icann.org/octo-ssr/daar.

[13] LLC Interisle Consulting Group. Phishing landscape 2022. https://interisle.net/PhishingLandscape2022.pdf.

[14] DomainTools John "Turbo" Conwell. Using domaintools threat profile to identify risky tlds. https://www.domaintools.com/resources/blog/using-domaintools-threat-profile-to-identify-risky-tlds.

[15] Maciej Korczynski, Maarten Wullink, Samaneh Tajalizadehkhoob, Giovane CM Moura, Arman Noroozian, Drew Bagley, and Cristian Hesselman. Cybercrime after the sunrise: A statistical analysis of dns abuse in new gtlds. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, pages 609–623, 2018.

[16] Marc Kührer, Christian Rossow, and Thorsten Holz. Paint it black: Evaluating the effectiveness of malware blacklists. In *Research in Attacks, Intrusions and Defenses: 17th International Symposium, RAID 2014, Gothenburg, Sweden, September 17-19, 2014. Proceedings 17*, pages 1–21. Springer, 2014.

[17] Sourena Maroofi, Maciej Korczyński, Cristian Hesselman, Benoit Ampeau, and Andrzej Duda. Comar: classification of compromised versus maliciously registered domains. In *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 607–623. IEEE, 2020.

[18] Leigh Metcalf, Dan Ruef, and Jonathan M. Spring. Open-source measurement of fast-flux networks while considering domain-name parking. 2017.

[19] Leigh Metcalf and Jonathan M. Spring. Domain parking: Not as malicious as expected. Technical Report CERTCC-2014-57, CERT Coordination Center, 2014.

[20] OECD. Security of the domain name system (dns). https://www.oecd-ilibrary.org/content/paper/285d7875-en, 2022.

[21] Openphish. Openphish website. https://openphish.com/.

[22] Phishstats. Phishstats website. https://phishstats.info/.

[23] Phishtank. Phishtank website. https://phishtank.org/.

[24] Spamhaus. Spamhaus website. https://www.spamhaus.org/.

[25] Spamhaus. The World's Most Abused TLDs. https://www.spamhaus.org/statistics/tlds/.

[26] Domain Name Stat. Domain name registrations, by registrar. https://domainnamestat.com/statistics/registrar/others.

[27] SURBL. SURBL website. https://www.surbl.org/.

[28] Samaneh Tajalizadehkhoob, Tom Van Goethem, Maciej Korczyński, Arman Noroozian, Rainer Böhme, Tyler Moore, Wouter Joosen, and Michel Van Eeten. Herding vulnerable cats: a statistical approach to disentangle joint responsibility for web security in shared hosting. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 553–567, 2017.

[29] Takayuki Tomatsuri, Daiki Chiba, Mitsuaki Akiyama, and Masato Uchida. Time-series measurement of parked domain names. In *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, pages 1–6, 2020.

[30] URLHaus. URLHaus website. https://urlhaus.abuse.ch/.

[31] URLScan. URLScan website. https://https://urlscan.io/about/.

[32] Verisign. The domain name industry brief – q3 2022 data and analysis. https://blog.verisign.com/domain-names/verisign-q3-2022-the-domain-name-industry-brief/.

[33] Thomas Vissers, Wouter Joosenand, and Nick Nikiforakisy. Parking sensors: Analyzing and detecting parked domains. In *NDSS '15*, February 2015.

[34] Peng Yang, Chao Shan, Dongan Wang, Lei Su, Juan Li, and Xinxin Wan. Mechanism of parked domains recognition based on authoritative dns servers. In *Proceedings of the 2nd World Symposium on Software Engineering*, WSSE '20, page 128–134, New York, NY, USA, 2020. Association for Computing Machinery.

[35] Johannes Zirngibl, Steffen Deusch, Patrick Sattler, Juliane Aulbach, Georg Carle, and Mattijs Jonker. Domain parking: Largely present, rarely considered. https://tma22-parking.github.io/.

[36] Johannes Zirngibl, Steffen Deusch, Patrick Sattler, Juliane Aulbach, Georg Carle, and Mattijs Jonker. Domain parking: Largely present, rarely considered! In *Proc. Network Traffic Measurement and Analysis Conference (TMA) 2022*, June 2022.

A. RBL Providers

Spamhaus [24] provides data on domains with low reputation that are collected from spam payload URLs, spam senders and sources, known spammers, phishing, virus, and malware-related websites.

APWG [11] contains blocklisted phishing URLs that are submitted by accredited users through the eCrime Exchange (eCX) platform.

SURBL [27] provides different blacklists containing malicious domain names. Their lists include phishing domains, spam domains, and domains used for malware. The data in these lists comes from various sources such as MailSecurity, PhishTank, and participating mail servers.

WMC Global [7] provide data on phishing which has a focus on mobile phishing including phishing via SMS (smishing).

Phishtank [23] is a community driven source of phishing URLs, including a verification system.

openphish [21] receives unfiltered URLs from multiple sources which it then uses to detect live phishing URLs.

URLHaus [30] is a project operated by abuse.ch. It collects, tracks and shares malware URLs.

Global Cyber Alliance [1] is an initiative to allow sharing of threat data and aggregates data from a number of contributors.

phishstats [22] is another phishing feed which gathers data from a number of sources before making it available.