Toward a Representative DNS Data Corpus: A Longitudinal Comparison of Collection Methods

1st Calvin Kranig
University of Wisconsin-Madison
Madison, Wisconsin
ckranig@wisc.edu

2nd Eric Pauley
University of Wisconsin-Madison
Madison, Wisconsin
epauley@cs.wisc.edu

3rd Wei-Shiang Wung
University of Wisconsin-Madison
Madison, Wisconsin
wwung@wisc.edu

4th Paul Barford
University of Wisconsin-Madison
Madison, Wisconsin
pb@cs.wisc.edu

5th Mark Crovella
Boston University
Boston, Massachusetts
crovella@bu.edu

6th Joel Sommers
Colgate University
Hamilton, New York
jsommers@colgate.edu

***Abstract*—Domain Name System (DNS) records are frequently used to investigate a wide variety of Internet phenomena including topology, malicious activity, resource allocations and user behavior. The scope and utility of the results of these studies depends intrinsically on the *representativeness* of the DNS data. In this paper, we compare and contrast five different DNS datasets from four different providers collected over a period of 3 months that in total comprise over 10.1B total Fully Qualified Domain Names (FQDNs) of which 3.7B are unique. We process and organize that data into a consistent format that enables it to be efficiently analyzed in Google BigQuery. We begin by reporting the details of the measurement methods and the datasets used in our analysis. We then analyze the relative coverage of each dataset by structural, administrative, and client-behavioral features. We find that while there are significant overlaps in the records provided by each dataset, each also provides unique records not found in the other datasets that are important in different use cases. Our results highlight the opportunities in using these datasets in research and operations, and how combinations of datasets can provide broader and more diverse perspectives.**

I. Introduction

The contents of the DNS (*i.e.,* the records that are managed and maintained within the DNS) determine how most Internet resources and services are accessed. As such, DNS record datasets have been of continued interest to the measurement community [1]–[11]. However, the usefulness of these datasets for downstream studies relies on how *representative* the included data is of the phenomenon analyzed through DNS. In particular, different DNS datasets may offer more complete views of geographies, infrastructure, and service types, and may have differing characteristics and stability over time. In contrast, a dataset may provide insufficient representativeness when it is biased towards specific geographies or simply has too little coverage. A dataset that is representative will lead to conclusions that are more likely to be generalizable for the intended research. When performing studies on the DNS, researchers currently lack an understanding of the relative trade-offs of these DNS datasets that are actionable for dataset curation. The goal of our work is to fill this gap.

In this paper, we report on a comparative analysis that focuses on five DNS datasets collected using different measurement techniques between May–June 2024. Our analysis offers a novel perspective on how each dataset can be useful by itself or in combination with others to provide a more complete and accurate representation of the configurations expressed in the DNS. To this end, we process and organize data that includes over 10 billion records using Google BigQuery. To assess representativeness, we conduct cross-dataset comparisons that include coverage of differing top-level domains, second-level domains, domain complexities, and record types.

We find that DNS records collected actively from a target list of FQDNs compiled from Internet scans have the largest number of unique records and provide deep details on infrastructure-related resources in the Internet. Conversely, records collected via open DNS zone data, applications (*e.g.,* extracted from web crawls), or Certificate Transparency (CT) Logs provide a more comprehensive perspective on all of the records associated with particular domains including those that may not have been actively requested. They also offer an opportunity to seed active DNS measurements and help narrow the space of potentially resolvable domains.

Our findings suggest that studies based on DNS records should consider a union of multiple datasets for a more complete coverage of FQDNs. However, it can be difficult to gather each type of dataset—especially over longer periods of time—and it may not be necessary to use the full union to address certain research and operational questions. A naïve combination of all possible datasets does not inherently imply an improvement in representativeness of the data as some datasets (*e.g.,* CT Logs) contain a large portion of non-active FQDNs (See §V-E2). Additionally, a large portion of the domains for some of the smaller datasets are already contained within larger datasets limiting their added benefit.

Our work aims to inform future research on DNS dataset selection and provides a critical perspective on how represen-

978-3-903176-74-4

tative each dataset is in terms of scale and diversity of records. Finally, our study highlights the strengths and weaknesses of each DNS-related measurement method.

II. Background & Related Work

Owing to the vast scope and public visibility of the DNS, it has continually been the target of study by the Internet measurement community. Analysis of the DNS has yielded insights on the evolution of service deployments [1]–[5], indicators of vulnerability and exploitation [6], [7], vulnerabilities in service management and configuration [8]–[10], end-user behaviors [11], among myriad other findings. In support of such studies, researchers have devised approaches for generating datasets with a broad view of the configurations expressed in the DNS. This is non-trivial because while the DNS is public, it is generally non-enumerable: any user can request a key and, through recursive resolution, obtain the value, but the list of valid DNS keys (*e.g.,* all FQDNs) is not publicly visible, and in many cases may not be fixed or defined *a priori*. As a result, producing a DNS dataset requires approaches for discovering valid keys.

Researchers and practitioners have taken a variety of approaches towards measuring the DNS. At the highest level, these fall into two categories:

- *Passive* approaches record network or transactional data that reference DNS names. For instance, logs from DNS resolver servers directly contain DNS references. Certificate transparency logs refer to domain names when TLS certificates are issued.
- *Active* approaches obtain candidate names from other sources (*e.g.,* zone files of top-level domains provided by operators, or web server responses) and use these as candidate names for querying the DNS. Central to this approach is the ability to make educated guesses on the space of DNS keys, and various datasets have been built around novel sources of candidate names.

Within these categories several techniques have become accepted standards for DNS dataset collection. Characterizing the relative efficacy of these approaches is a key focus of our work.

To the best of our knowledge, the issue of representativeness in *DNS datasets* has not been addressed in prior work. However, several prior studies inform our work. Paxson provides guidelines for sound Internet measurement [12], which include considering errors and bias in Internet measurement datasets. More directly related to our work is the study by Scheitle *et al.* [13], which considers representativeness, potential biases, stability, and overlap between domain ranking lists, including those that use DNS records to create such lists. Zeber *et al.* consider representativeness in the context of web crawls as a proxy for human crawling [14].

A. Dataset Use Cases in Prior Works

Each of the datasets we chose to analyze and compare has been utilized in previous research studies as detailed below. These studies motivated our comparison effort and provide a context for why comparing collection methods is important for further research.

a) Sonar: Rapid7's Project Sonar [15] actively collects various types of records from the global DNS. Sonar data has been used to evaluate the Web's evolution, to inform IPv6 address-space scans, and to analyze potential privacy issues on the Internet. By analyzing the A and AAAA records in Rapid7 and other DNS datasets, Hoang *et al.* [16] reported the highly concentrated nature of web co-location. Related to user privacy, Cangialosi *et al.* [17] evaluated the sharing of website keys with third-party hosting providers using Rapid7's forward and reverse DNS scans in concert with other data sources.

b) OpenINTEL: Since the OpenINTEL project started daily large-scale DNS measurement in 2016 [18], the dataset has been used for a variety of analyses. Sommese *et al.* [19] explored the evolution of anycast adoption in DNS name service from 2017–2021. Abhishta *et al.* [20] evaluated the changes in DNS configurations after DDoS attacks on managed DNS service providers from the NS records in the OpenINTEL dataset.

c) Common Crawl: Common Crawl collects content from 3-5 billion webpages monthly and serves as an important corpus for web-related studies and natural language tasks. For example, Matic *et al.* used Common Crawl data to identify blacklisted or censored URLs/FQDNs at large scale [21]. In addition, due to the availability of multiple languages on many websites, Smith *et al.* [22] proposed that Common Crawl can be used as a data source of parallel texts for machine translation.

d) CT Logs: Recording all SSL/TLS certificates issued by Certificate Authorities, Certificate Transparency (CT) Logs are commonly used for security and privacy analyses but also for DNS-related studies. For example, Sommese *et al.* studied both Common Crawl and CT Log data to compare ccTLD coverage in these public data sources with ground truth DNS zone transfer data from OpenINTEL [23], finding that public data covers roughly 40–80% of the full data, and that coverage is increasing over time. In security-related studies, Gustafsson *et al.* [24] compared the similarities and differences in public CT logs maintained by various entities. Finally, CT logs can be used as a source for phishing attack detection and prevent phishing attacks in real time. Both Fasllija *et al.* [25] and Drichel *et al.* [26] proposed ML-based classifiers to detect potentially malicious domains in CT logs.

e) Reverse DNS: Reverse DNS (PTR) records map IP addresses to hostnames. Names in these records typically relate to servers and access infrastructure on the Internet. As a result, these domain names often include information on geographic locations and service providers. In particular, Lee and Spring studied broadband service providers by analyzing names in reverse DNS entries [27]. Dan *et al.* [28] proposed a method to determine the accurate geolocations of IP addresses at the city level by extracting related information from reverse DNS hostnames. On the other hand, automated PTR record configuration in the global DNS may cause concerns about user privacy. Van der Toorn *et al.* [29] found that automated

changes to DNS from DHCP (Dynamic Host Configuration Protocol) and IPAM (IP Address Management) could reveal the client's positions from IP prefixes.

III. Collection Methodologies

In this section we describe the data sets and respective collection methods studied in this paper [1].

A. Rapid7 Sonar

Rapid7's Project Sonar [15] has conducted active measurement of the DNS since 2014. Rapid7 conducts weekly active measurement campaigns of the DNS to determine which names on the candidate list resolve under a set of query types (`A`, `AAAA`, `NS`, `CNAME`, `MX`, `TXT`, and `CAA`). Their Forward DNS dataset is created by extracting domain names from Reverse DNS (PTR) Records, SSL Certificates, CT Logs, crawled HTTP/S pages, and Zone files from com, org, net, and other TLDs. These domains are then sent an ANY query for each domain. Rapid7's long-standing collection of this candidate domain list gives it unique historical record coverage (*i.e.,* of records created long ago that may not be actively used).

B. OpenINTEL

Another view on the DNS is offered by OpenINTEL [18], which has contractual relationships with DNS TLD operators to provide direct access to lists of the second-level domains (SLDs) in those zones. Using this, OpenINTEL performs active scans of a variety of record types and subdomains [30]. Notably, this approach does not aim to achieve high coverage of subdomains and is limited to specific TLDs. In this work, we consider data from the com, czds, net, opencc, and org datasets that were made available to us by OpenINTEL. OpenINTEL performs scans daily so we create a union of these daily scans to compare them to other datasets as described in §IV.

C. Common Crawl

Common Crawl [31] provides openly available web crawl data which are made available as periodic snapshots. Although the intent of Common Crawl is not to provide DNS data, there are, of course, many DNS FQDNs that are found in the crawled content, which currently consists of around 2.7B web pages. NXDOMAINs encountered by the crawler are logged but are not released as part of the main dataset. They are however present in the WARC and WAT files that Common Crawl utilizes as part of their web graphs [32]. It is important to note that seed domains are updated each month and there is an attempt to keep content overlap between crawls to a minimum. This results in a higher level of churn compared to other datasets which is further discussed in §V-E3. Our comparison uses data from Common Crawl April, May, and June Archives (CC-MAIN-2024-18, CC-MAIN-2024-22, CC-MAIN-2024-26) [31]. We retrieved all of the URL index files for the dataset and extracted all available FQDNs for further analysis.

[1]All datasets used in this study are either publicly available or can be requested from their sources.

D. Certificate Transparency Logs

Certificate Transparency (CT) logs, while originally intended as a public record of TLS certificate issuances [33], also serve as a *de facto* listing of domain names for which TLS certificates have been issued [34] (increasingly a requirement for hosting any Internet services). These logs are publicly available and easily crawled by researchers. They also update in near-realtime as certificates are issued, making them a promising source of candidate names for DNS resolution. Our work uses browser-trusted certificate logs as scanned by Censys [35], though the same data can be collected by researchers without requiring a data access agreement with third parties. Because inclusion in TLS certificates does not necessarily imply that a domain name resolves, we perform DNS resolution on all domains within the CT logs with valid certificates (Certificates that have not expired at the time of DNS Resolution). We use zDNS [36] and public resolvers to resolve this subset using `A` queries. These results are discussed in §V-E2.

E. Reverse DNS

Rapid7 also offers a dataset of reverse DNS (rDNS), *i.e.* DNS `PTR` records [15]. Reverse DNS records have the important distinction that they are (for IPv4 addresses) enumerable. However, the *values* of these DNS PTR records usually refer to conventional names in the (forward) DNS. As a result, they can be used to seed forward DNS queries. We use Rapid7's rDNS dataset as a listing of domains that are discoverable using reverse DNS, though researchers can generate these listings independently with minimal effort.

IV. Comparison Methodology

Due to the different collection methods employed by each data provider, including different time spans of data collection, we create unions of each dataset that equated to one month's worth of data in order to have a fair comparison on the basis of SLD/FQDN coverage. We compare each dataset union, along with pruning methods detailed below, for April, May, and June 2024.

A. Union Methods

a) Rapid7 Sonar and Rapid7 Reverse DNS: Rapid7 conducts their measurements over the course of a week. We took a union of every A record and Reverse DNS dataset that finished collection during the designated month.

b) OpenINTEL: OpenINTEL runs their collection methods every 24 hours to generate the datasets we utilized for comparison (czds, net, opencc, org, and com). We take a union of one month's worth of collection for each of the OpenINTEL datasets for the designated month. For the main comparison, we consider all domain names present in all record types including: query_name, response_name, cname_name, dname_name, mx_address, and ns_address. For the A Record comparison see in §V-C we only consider query_name, response_name, and cname_name.

c) Common Crawl: Since the collection of each common crawl dataset occurs over the course of a month we consider every domain that appears in the URL index files for the given month.

d) CT Logs: For the CT Logs collected by Censys we consider every domain that has a valid certificate for at least some time during the designated month. We consider a valid certificate to be one that has a validity_period.not_after that is greater than or equal to the start of the month and a validity_period.not_before that is less than or equal to the end of the month. We consider all dns_names and common_name values for comparison.

B. Pruning Non-Domains

The union of domains detailed above results in some names with wildcards (Rapid7 datasets), some with IP addresses (OpenINTEL, Reverse DNS), and some invalid domain names. We prune the set of names by removing any wildcards and only consider FQDNs that have non NULL return values when utilizing BigQuery's `NET.REG_DOMAIN(url)` function[2].

Our pruning process results in smaller coverage for the datasets that focus primarily on second-level domains (SLDs). With this in mind we calculated the SLD for each FQDN in each dataset and compared each dataset using only their SLDs (*cf.* §V-B).

V. Evaluation

Our evaluation focuses on three dimensions of representativity: coverage, stability, and uniqueness. We consider these characteristics by examining FQDNs, SLDs, and A records over a period of 3 months. We first examine FQDN, SLD, and A record coverage, then examine aspects of dataset diversity in terms of how datasets evolve over time and distinct characteristics of individual datasets. We also discuss balance between infrastructure-related names and user-facing names.

Our quantitative assessment considers the union of all 5 data sets which we denote as $\bigcup_{i=1}^{5} A_i$ where each A_i is one of our data sets: respectively Sonar, OpenINTEL, Common Crawl, CT Logs, and rDNS. We use the term *coverage* to convey the contribution of any A_i to the union (including overlaps with other sets) *i.e.,*

$$\frac{|A_j \cap \bigcup_{i=1}^{5} A_i|}{\left|\bigcup_{i=1}^{5} A_i\right|} \tag{1}$$

As such, coverage is reported as a percentage. We also report the *unique* contribution of each data set *i.e.,* the elements in A_i that appear only in that data set:

$$\frac{|A_j \setminus \bigcup_{i \neq j} A_i|}{\left|\bigcup_{i=1}^{5} A_i\right|} \tag{2}$$

We use Euler diagrams to illustrate relative data set size and intersections. However, depicting all intersections accurately using only circles or ellipses in Euler diagrams presents mathematical challenges, especially as the number of sets increases [37], [38]. This results in some datasets (notably, Common Crawl) appearing to be completely contained within other datasets despite having some unique contents.

[2]This function returns the registered or registrable domain name given a URL and is a utility function specific to BigQuery.

A. FQDN Coverage

Table I presents FQDN counts across all three months and for each dataset we consider. Figure 1 depicts overlap in FQDNs for each dataset for June 2024. We first observe in the table and figure that the Sonar dataset is the largest source of FQDNs, with coverage averaging 62.3% across all three months. The Sonar dataset is followed by rDNS with coverage of a minimum of 36.4% of domains each month. The intersection between rDNS and Sonar for June contains 769,113,436 FQDNs (22.8 % of all the June FQDNs). The intersections that Sonar has with the remaining datasets are much lower, which speaks to the importance of the rDNS dataset in seeding the Sonar measurements.

CT Logs has the 3rd largest coverage of FQDNs, with an average of 26.0% of all FQDNs per month. Since Censys focuses on collecting certificates utilized in TLS, it omits domains that are not utilizing TLS during communication. Nonetheless, it offers a vast number of unique domains that are not present in the other datasets. Although not all the domains found in CT Logs are active, a significant fraction are indeed resolvable as we discuss below in §V-E2.

OpenINTEL is the other actively collected dataset and has coverage of an average of 16.0% per month. Since OpenINTEL focuses primarily on collecting SLDs [30] its coverage of subdomains is naturally less than other data providers. This is made more evident in Figure 4. OpenINTEL, however, does offer daily snapshots of a variety of DNS resource records (not just A records), which results in a considerable number of unique domains as discussed in §V-E4.

Common Crawl is the smallest dataset by far (average of 1.77% FQDN coverage per month). As a result, there is a small number of unique FQDNs compared with the other datasets (7,301,443 average per month). Since the goal of Common Crawl is to collect content from web pages—domain names are simply collected as a side-effect—this is to be expected. Since Common Crawl seeks to crawl new pages each month it has a higher level of churn (see §V-D), which results in each month having a large portion of new FQDNs and unique FQDNs. This highlights the importance of utilizing crawling in seeding active queries which can be seen by the large coverage that Sonar has of Common Crawl data compared to OpenINTEL. It is also important to note the prevalence of TLS in HTTP/S communication resulting in a high coverage of Common Crawl FQDNs by CT logs.

Table II provides an overview of the top TLDs and SLDs in the union of the 5 datasets we compared in June '24. The FQDN count states the total number of domains that have the given TLD/SLD while the percentage compares that count to the total number of FQDNs for June. These results are similar across all three months. As expected the com and net TLDs make up a majority of all FQDNs. More surprising is the

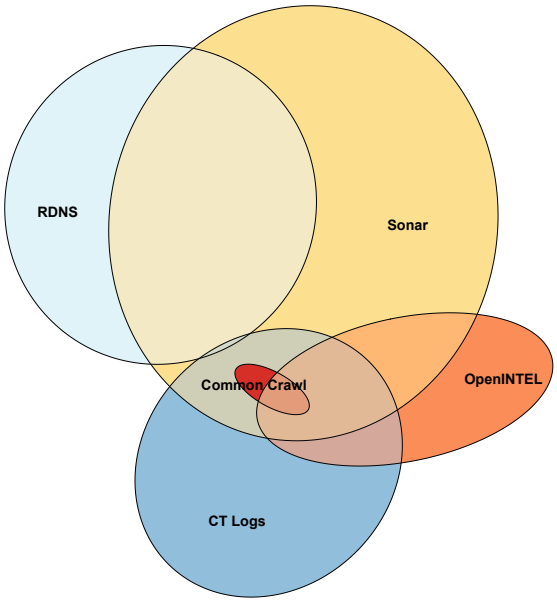

Fig. 1: Overlap relationships for FQDNs for June 2024.

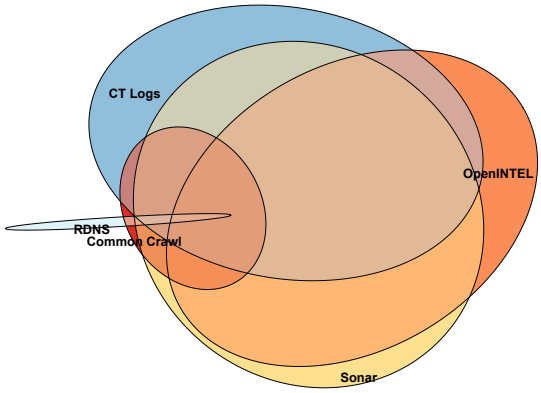

Fig. 2: Overlap relationships for SLDs for June 2024.

TABLE I: FQDN counts for April, May, and June, 2024.

Label	April (% of Total)	May (% of Total)	April to May (% Change)	June (% of Total)	May to June (% Change)
In All	303,956 (0.01)	304,012 (0.01)	56 (0.02)	273,018 (0.01)	-30994 (-10.19)
CT Logs	865,613,944 (25.44)	869,657,868 (25.78)	4,043,924 (0.47)	880,883,137 (26.16)	11,225,269 (1.29)
CT Logs Unique	531,242,365 (15.62)	529,947,195 (15.71)	-1,295,170 (-0.24)	534,843,872 (15.88)	4,896,677 (0.92)
Common Crawl	57,094,127 (1.68)	60,609,268 (1.80)	3,515,141 (6.16)	58,740,058 (1.72)	-1,869,210 (-3.08)
Common Crawl Unique	6,246,748 (0.18)	7,858,220 (0.23)	1,611,472 (25.80)	7,799,361 (0.23)	-58,859 (-0.75)
OpenINTEL	542,440,345 (15.95)	540,344,985 (16.02)	-2,095,360 (-0.39)	540,167,272 (16.04)	-177,713 (-0.03)
OpenINTEL Unique	217,599,838 (6.40)	212,290,246 (6.29)	-5,309,592 (-2.44)	207,935,524 (6.17)	-4,354,722 (-2.05)
rDNS	1,290,813,367 (37.94)	1,229,794,819 (36.45)	-61,018,548 (-4.73)	1,225,731,680 (36.40)	-4,063,139 (-0.33)
rDNS Unique	543,727,497 (15.98)	460,516,810 (13.65)	-83,210,687 (-15.30)	456,324,345 (13.55)	-4,192,465 (-0.91)
Sonar	2,048,058,315 (60.20)	2,104,587,720 (62.38)	56,529,405 (2.76)	2,096,721,826 (62.27)	-7,865,894 (-0.37)
Sonar Unique	900,726,415 (26.48)	935,521,153 (27.73)	34,794,738 (3.86)	929,713,415 (27.61)	-5,807,738 (-0.62)
Total	3,401,924,923 (100.0)	3,373,884,647 (100.0)	-28,040,276 (-0.82)	3,367,361,613 (100.0)	-6,523,034 (-0.19)

TABLE II: Top 10 TLDs and SLDs by count and percentage for June 2024.

Rank	TLD (FQDN Count, %)	SLD (FQDN Count, %)
1	com (1,418,964,927, 42.14%)	amazonaws.com (173,655,967, 5.16%)
2	net (558,235,445, 16.58%)	spectrum.com (62,306,873, 1.85%)
3	de (106,758,827, 3.17%)	comcast.net (46,552,981, 1.38%)
4	ne.jp (61,614,200, 1.83%)	azure.com (38,684,456, 1.15%)
5	com.br (61,344,556, 1.82%)	bbtec.net (34,711,863, 1.03%)
6	org (58,736,129, 1.74%)	sbcglobal.net (31,550,892, 0.94%)
7	fr (50,440,227, 1.50%)	myvzw.com (30,276,580, 0.90%)
8	it (47,277,429, 1.40%)	t-ipconnect.de (25,774,757, 0.77%)
9	io (42,224,536, 1.25%)	hinet.net (23,834,413, 0.71%)
10	ru (41,203,584, 1.22%)	rr.com (22,510,163, 0.67%)

prevalence of infrastructure FQDNs as shown by the top SLDs by FQDN count. Cloud and infrastructure providers make up the vast majority of the top SLDs by FQDN count, highlighting the fact that these providers utilize a large number of unique FQDNs. The coverage of these infrastructure FQDNs differs across the 5 datasets as discussed in §V-E.

B. SLD Coverage

Since SLD's are important in certain research studies (*e.g.*, [39]) we also conducted an analysis of SLD coverage for each dataset. For this analysis, we took the SLDs of every FQDN from our FQDN analysis. For individual datasets we considered an SLD to be in the dataset if any of its FQDNs had a given SLD.

There is a higher level of overlap as shown in Figure 2 and the SLDs coverage by Sonar, OpenINTEL, and CT Logs is more comparable (74.3%, 69.0%, 65.6% average respectively). 125,781,755 (38 %) of the SLDs appear in all three datasets in June. The smaller number of SLDs for OpenINTEL can partly be attributed to the limited datasets we chose to evaluate, but Sonar (and other datasets) still have unique names in the TLDs covered by the OpenINTEL datasets we utilized.

TABLE III: Summary of SLD counts for April, May, and June, 2024.

Label	April (% of Total)	May (% of Total)	April to May (% Change)	June (% of Total)	May to June (% Change)
All	327,330,012 (100.0)	328,415,193 (100.0)	1,085,181 (0.33)	328,684,464 (100.0)	269,271 (0.08)
In All	1,065,599 (0.33)	1,059,357 (0.323)	-6242 (-0.59)	953,395 (0.29)	-105962 (-10.00)
CT Logs	211,039,772 (64.47)	214,009,838 (65.16)	2,970,066 (1.41)	216,848,806 (65.97)	2,838,968 (1.33)
CT Logs Unique	42,040,961 (12.84)	42,893,498 (13.06)	852,537 (2.03)	43,265,742 (13.16)	372,244 (0.87)
Common Crawl	46,264,231 (14.13)	48,276,678 (14.70)	2,012,447 (4.35)	46,690,164 (14.21)	-1,586,514 (-3.29)
Common Crawl Unique	618,383 (0.19)	681,096 (0.21)	62,713 (10.14)	668,554 (0.20)	-12,542 (-1.84)
OpenINTEL	226,798,738 (69.29)	226,827,194 (69.07)	28,456 (0.01)	226,777,911 (68.99)	-49,283 (-0.02)
OpenINTEL Unique	26,880,222 (8.21)	26,393,974 (8.04)	-486,248 (-1.81)	26,314,161 (8.01)	-79,813 (-0.30)
rDNS	6,783,894 (2.07)	6,794,517 (2.07)	10,623 (0.16)	6,782,172 (2.06)	-12,345 (-0.18)
rDNS Unique	2,317,579 (0.71)	2,326,661 (0.71)	9,082 (0.39)	2,333,056 (0.71)	6,395 (0.27)
Sonar	243,396,564 (74.36)	244,182,982 (74.35)	786,418 (0.32)	243,774,794 (74.17)	-408,188 (-0.17)
Sonar Unique	15,643,644 (4.78)	15,130,463 (4.61)	-513,181 (-3.28)	15,002,747 (4.56)	-127,716 (-0.84)

C. A Records

Since A records are the focus of many research studies we analyzed just the A records for the Sonar and OpenINTEL datasets. Full results are shown in Table A1. Sonar has greater overall coverage than the combination of OpenINTEL's datasets that we utilized with roughly 82% of FQDNs being unique to Sonar each month. OpenINTEL had better coverage of SLDs compared to FQDNs but Sonar still has more coverage overall.

D. Longitudinal Stability

As illustrated in Sankey diagram in Figure 3 the relative sizes of the datasets and total number of domains stay fairly consistent over our 3 month period of study with a large number of domains appearing in multiple datasets. Common Crawl has the largest percent change in the number of unique FQDNs from April to May (25% increase) resulting in the overall percentage of unique domains increasing from 0.18% to 0.23% (See Table I). The relatively small change in relation to the total number of domains for each dataset and their number of unique FQDNs shows the consistency in each dataset's ability to provide unique domains. This consistency extends to the relative SLD coverages with the highest percentage changes in the smallest datasets (Common Crawl and SLDs that appear in all five datasets) as shown in Table III.

E. Unique Domain Evaluation

Each of the datasets contain unique FQDNs and SLDs. In this section we provide explanations for why this is the case by examining the distinct attributes of each dataset.

1) Subdomain Length: Assembling datasets by querying the DNS requires a candidate set of FQDNs. Deep coverage of the FQDN name space means an increase in *subdomain length*, which we define as the number of characters including dots after the SLD (*e.g.,* so www.example.com would have a subdomain length of 4). For most data sets (excluding rDNS and Sonar), the subdomain length of FQDNs in a given data set is longer on average for FQDNs that are unique to the data set, as shown in Figure 4 and Table IV. The FQDNs that appear in the union of all datasets have the shortest subdomain lengths, highlighting the fact that it is easier to fully query the space of subdomains that have a shorter length. Since CT Logs are a representation of TLS records, no guessing of subdomains is required. This results in the median subdomain length for unique domains being higher than the other datasets (see Figure 4). In order to extend coverage, datasets assembled by querying the DNS should consult CT Logs to narrow the space of potentially resolvable FQDNs.

2) CT Logs: A large portion of the unique domains contained within the CT Logs can be explained by looking at the domains that are only associated with a *pre-certificate* (precert). Precerts are submitted from a Certificate Authority to CT Logs to obtain the log's signed certificate timestamp (SCT). Censys lists a certificate as a precert "[i]n the case where only the pre-cert was submitted to a CT log and the corresponding certificate has not been observed during a Censys scan of the Internet" [40]. Between 38.53% and 41.4% of the FQDNs that are unique to CT Logs are only associated with a precert each month, and over 98% of all precert domains are unique each month.

The same phenomena can be seen in the percentage of total FQDNs associated with revoked certificates that are unique. When a certificate holder believes that their private key may have been compromised they can initiate a revocation request with the Credential Authority (CA) of their certificate. This revocation is then added to the CT Logs by the CA. These revoked certificates are normally of interest to security research. FQDNs with revoked certificates make up only 0.62% of total FQDNs associated with CT Logs but the majority of these revoked domains are unique to CT Logs. Between 87.7 and 88.4% of all revoked domains are unique each month.

This still leaves over half of the unique domains unaccounted for. In September 2024 we began resolving FQDNs with valid certificates present in the CT Logs. Since our latest comparison was conducted in June 2024 we select FQDNs from June that were present in the CT Logs that also have valid domains in September and check their resolution rates. See Table A2 for full results. For September the resolution rate of all of the FQDNs present was 59.6% with 62.92% of the domains that had a valid certificate in June successfully

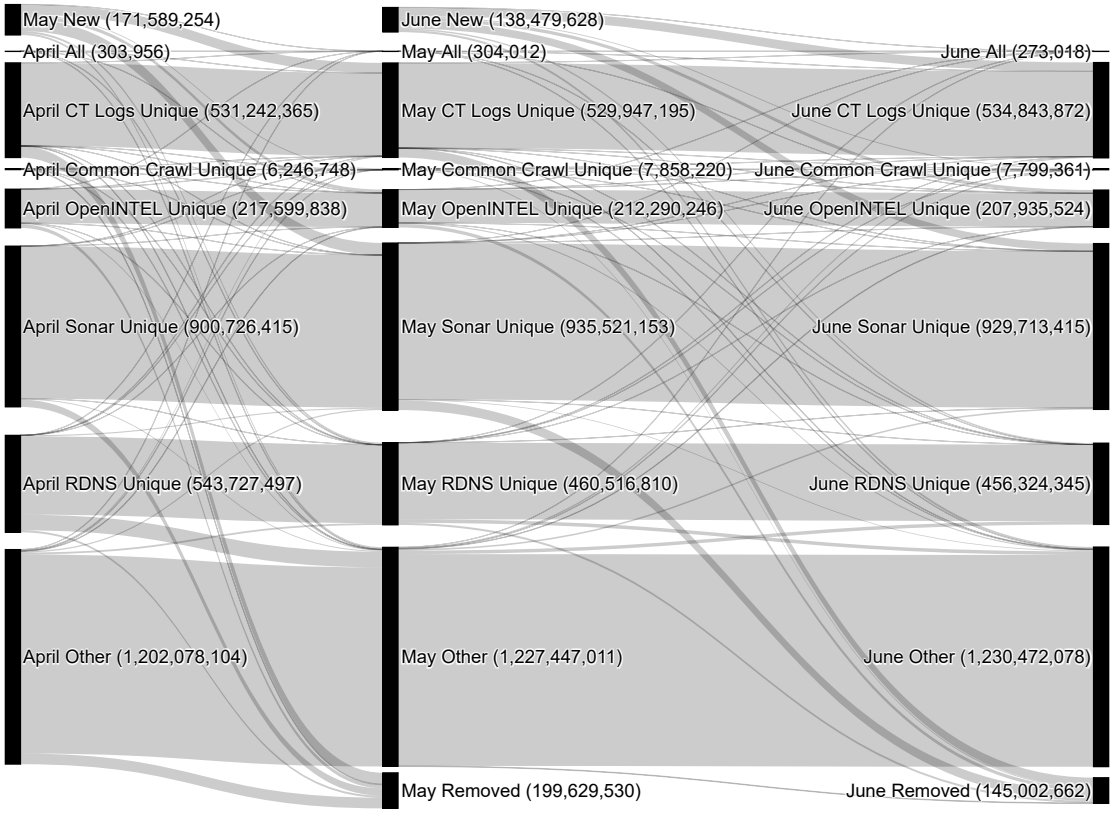

Fig. 3: FQDN hurn from April to June 2024.

TABLE IV: Median subdomain lengths for April, May, and June, 2024.

Category	April	May	June
All FQDNs	20	20	20
In All	4	4	4
Common Crawl	8	8	8
Common Crawl Unique	9	10	10
CT Logs	14	14	14
CT Logs Unique	30	30	27
OpenINTEL	3	3	3
OpenINTEL Unique	3	6	6
Sonar	19	20	20
Sonar Unique	19	19	19
rDNS	21	22	22
rDNS Unique	22	20	20

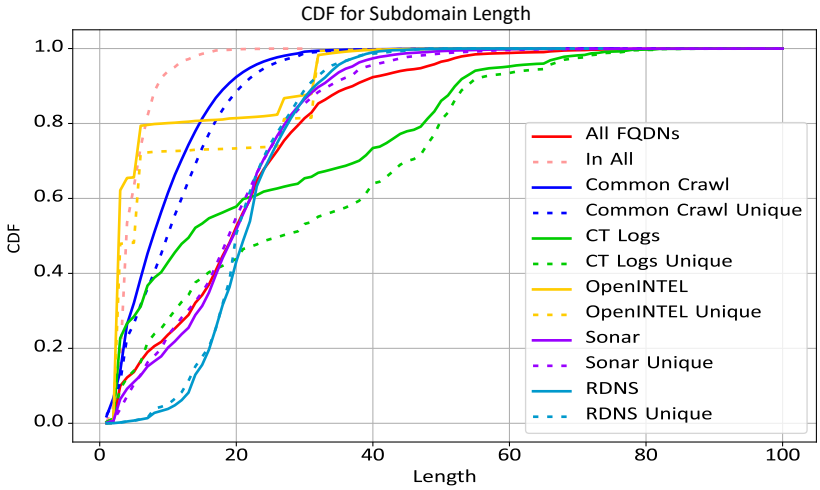

Fig. 4: CDF of Subdomain Lengths in June 2024

resolving. When looking at domains associated with revoked certificates we see that 18.31% of the domains present in both June and September resolve. Not all domains associated with a revoked certificate are malicious and many of these domains

get a reissued certificate after initiating a revocation request. The most interesting observation is that unique domains have a 39.55% resolution rate: 412,034,006 FQDNs are unique to CT Logs and still resolve.

Focusing on unique SLDs, we find that between 19.9% to 20% of SLDs present in the CT Logs dataset are unique each month. The number of SLDs that only have a precert or revoked domain is much smaller than when considering only FQDNs. We find that 85.18% of the June SLDs remaining in September have at least one FQDN that resolves. This decreases to only 56.15% for the unique SLDs. The number of unique FQDNs and SLDs that resolve highlight how CT Logs offer a unique coverage of domains as compared to the other actively collected datasets.

Finally it is important to recognize the coverage that CT Logs have of domains utilized in cloud infrastructure. CT Logs cover 66.3 % of the 173,655,967 FQDNs with the amazon-aws.com SLD versus Sonar (34.1%) and rDNS (33.7%). CT Logs coverage of azure.com FQDNs is even more impressive with 99.4% of the FQDNs with the SLD azure.com in June 2024. Further examination of top SLDs in CT Logs indicates that a large portion are associated with cloud service and infrastructure providers.

3) Common Crawl: Common Crawl is the smallest dataset that we considered and focuses on the Web. It is predominantly covered by the Sonar A records with an average of an 81% intersection between Common Crawl FQDNs and Sonar A records each month. Over half of the remaining domains are unique to Common Crawl with a minimum of 10.94% (From April). Of these unique domains over 83% have at least one page with a 200 HTTP status code indicating that content is being hosted on the domain (or the domain has some active server serving a successful HTTP/S request). This is compared to roughly 84% in the complete dataset. The number of unresponsive domains (domains that only have status codes greater than or equal to 400) is relatively unchanged from the unique to complete FQDN sets (around 8%). This means that unique domains for Common Crawl are as likely to host content as domains that are in other datasets and it presents a motivation for including crawled domains when conducting active DNS measurements.

Common Crawl seeks to minimize re-crawling pages each month. This results in a higher level of churn compared to the other datasets. On average 24% of FQDNs are new each month and 25% of domains are removed. Despite this, there is relative stability in the percentage of unique domains.

When focusing on SLDs not much is different outside of the much smaller portion of unique SLDs compared to the total SLDs present each month in the Common Crawl dataset. It is important to note that SLDs make up a large portion of the FQDNs present within the Common Crawl dataset hinting that most domains hosting HTTP/S content that are crawled by Common Crawl may not have extensive subdomain hierarchies. Finally, when considering the top SLDs, the list is dominated by blogging and personal website-making sites such as wordpress.com, blogspot.com, and weebly.com.

4) OpenINTEL: Over 38% of all OpenINTEL FQDNs are unique each month. Of these unique FQDNs, the majority (over 52% each month) are associated with non-A records. This is also evident by our A record analysis in §V-C, despite FQDNs with A records making up roughly 72% of all FQDNs each month. This highlights the importance of incorporating multiple types of record queries when building a comprehensive DNS dataset based on active measurement.

Despite Sonar's larger coverage of the A record space, over 6% of the FQDNs associated with A records are unique to OpenINTEL. Table A1 shows that Sonar's A record dataset is not fully comprehensive and that there are FQDNs and SLDs not covered by Sonar.

When considering SLDs, we note that OpenINTEL has the 2nd highest percentage of unique SLDs – over 8% each month (see Table III). This is despite having fewer SLDs overall compared to Sonar. This lends credence to their focus on collecting DNS records for SLDs as opposed to FQDNs.

5) Rapid7 Sonar and Reverse DNS: The Sonar A record dataset is the most comprehensive that we studied based on FQDNs, SLDs, and A records. With this in mind, it is no surprise that Sonar has the highest number and percentage of unique FQDNs (over 26.4% each month). This number would be even larger without the inclusion of the rDNS dataset, which Sonar draws from to seed their forward measurements. We find that 44% of all Sonar FQDNs are unique when including rDNS, but when removing rDNS 80% of Sonar FQDNs are unique when compared to the remaining datasets as shown in Table A4.

The same can be said for Rapid7's rDNS dataset. The overlap in FQDNs for rDNS is minimal with all of the datasets except for Sonar as shown in Figure 1 (Also see Table A5). Less than 0.45% of rDNS FQDNs are present in non-Sonar datasets each month.

However, this does not extend to SLDs. The overall coverage for SLDs is drastically reduced for rDNS and a large portion of its non-unique domains are also present in all of the other datasets as shown in Figure 2. The top SLDs (by FQDN count) present in the Rapid7 datasets are dominated by ISP and Cloud providers (Spectrum, Amazon, Comcast, Verizon). These domains are not as widely covered by the other datasets and provide motivation for using Rapid7 datasets when investigating infrastructure domains.

VI. Case Study

To illustrate the benefit of combining DNS datasets we partially reproduce the *combosquatting* collection methodology described in [41]. Combosquatting is a type of domain name abuse where attackers register domain names that are similar to legitimate ones but with additional words, prefixes, or suffixes (*e.g.,* a combosquatted domain for paypal.com might be secure-paypal.com). These domains are designed to deceive users into thinking they are interacting with a trusted site.

We start by gathering the top 500 domains for June 2024 as reported by Tranco [39]. We then look for all combosquatting domains within our combined dataset by finding all SLDs that

TABLE V: Case study of combosquatting for the Tranco top 500 domains identified in each DNS dataset for June 2024.

Category	Count (Percentage) of combosquatting domains
Sonar	454,793 (0.19%)
OpenINTEL	495,304 (0.22%)
Censys	397,597 (0.18%)
rDNS	9,892 (0.15%)
Common Crawl	66,795 (0.14%)
Sonar or OpenINTEL	569,685 (0.20%)
All	665,532 (0.20%)

are in the form .*AD.* (where AD is the Authoritative domain from the Tranco Top 500). We remove exact matches with the original authoritative domain in the subdomain portion of the SLD. The results can be seen in Table V. The combination of both active datasets (Sonar and OpenINTEL) provides an increase of 25% and 15% respectively. Combining the remaining datasets results in an additional 16.8% increase in combosquatting domains whilst maintaining the same ratio of combosquatting domains to total domains.

VII. Selecting Representative DNS Datasets

Our findings on the coverage, stability and uniqueness of the data sets considered in our study lead us to propose the following guidance for assembling representative DNS datasets for research:

1) Studies seeking to examine and analyze the contents of the DNS based on submitting queries for FQDNs should use the full union of all data sets.
2) Studies that seek to examine Internet infrastructure should start with Sonar and rDNS and consider including CT logs for the broadest coverage.
3) Studies that seek to understand details of individual zones would be well-served by using a combination of Sonar and OpenINTEL.
4) Studies focused on Web crawling and content that currently use Common Crawl could be expanded by crawling SLDs from CT logs.
5) The unique features of certain data sets mean that there is no benefit to considering other data sets for certain studies (*e.g.,* CT logs for security and SLD name churn, and OpenINTEL for diverse record types).

VIII. Conclusion

The goal of our work is to improve the understanding of different DNS data sets that are commonly used in empirical research studies. We report on our study that compares and contrasts five different popular DNS datasets from four different providers collected using different methods over a period of 3 months. We describe the methods used to collect each data set and details of each of the data sets that in total comprise over 10B records. We focus on three dimensions of *representativity*: coverage, stability, and uniqueness. We find that data sets collected through active querying of the DNS (*e.g.,* Sonar) provide the broadest coverage and that there is significant overlap between the records provided by each dataset. We also assess how each data set changes over the course of our study and show that all data sets are relatively stable with the most churn in the DNS records extracted from Common Crawl, which is a consequence of its design. We also find that each data set provides unique records not found in the other datasets that are important in different use cases. Our results highlight the opportunities in using these datasets in research and operations, and how combinations of datasets can provide broader and more diverse perspectives. In future work, we plan to examine how DNS data collection can be enhanced to expand coverage *i.e.,* add new FQDNs that do not exist in the union of the current datasets. This will include data collected via passive monitoring that offers the opportunity to gain unique insights on DNS use.

Acknowledgements

The authors thank the Rapid7 Research Team and OpenINTEL for providing data and feedback on this project. This material is based upon work supported by the National Science Foundation under Grant No. CNS2312709, CNS2312710, CNS-2312711, CNS-2319367, CNS-2319368, and CNS-2319369. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] K. He, A. Fisher, L. Wang, A. Gember, A. Akella, and T. Ristenpart, "Next stop, the cloud: Understanding modern web service deployment in ec2 and azure," in *Proceedings of the 2013 conference on Internet measurement conference*, 2013, pp. 177–190.

[2] W. Scott, T. Anderson, T. Kohno, and A. Krishnamurthy, "Satellite: Joint analysis of CDNs and Network-Level interference," in *2016 USENIX Annual Technical Conference (USENIX ATC 16)*, 2016, pp. 195–208.

[3] G. Moura, S. Castro, W. Hardaker, M. Wullink, and C. Hesselman, "Clouding up the Internet: how centralized is DNS traffic becoming?" in *Proceedings of the ACM Internet Measurement Conference*, 2020.

[4] Q. Jacquemart, C. Pigout, and G. Urvoy-Keller, "Inferring the deployment of top domains over public clouds using dns data," in *2019 Network Traffic Measurement and Analysis Conference (TMA)*. IEEE, 2019, pp. 57–64.

[5] T. V. Doan, R. van Rijswijk-Deij, O. Hohlfeld, and V. Bajpai, "An empirical view on consolidation of the web," *ACM Transactions on Internet Technology (TOIT)*, vol. 22, no. 3, pp. 1–30, 2022.

[6] A. Portier, H. Carter, and C. Lever, "Security in plain txt: Observing the use of dns txt records in the wild," in *Detection of Intrusions and Malware, and Vulnerability Assessment: 16th International Conference, DIMVA 2019, Gothenburg, Sweden, June 19–20, 2019, Proceedings 16*. Springer, 2019, pp. 374–395.

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

[8] V. Ramasubramanian and E. G. Sirer, "Perils of transitive trust in the domain name system," in *Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement*, 2005, pp. 35–35.

[9] D. Liu, S. Hao, and H. Wang, "All Your DNS Records Point to Us: Understanding the Security Threats of Dangling DNS Records," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. Vienna Austria: ACM, Oct. 2016, pp. 1414–1425. [Online]. Available: https://dl.acm.org/doi/10.1145/2976749.2978387

[10] E. Pauley, R. Sheatsley, B. Hoak, Q. Burke, Y. Beugin, and P. McDaniel, "Measuring and Mitigating the Risk of IP Reuse on Public Clouds," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, Apr. 2022, pp. 1523–1523, iSSN: 2375-1207. [Online]. Available: https://www.computer.org/csdl/proceedings-article/sp/2022/131600b523/1CIO7rpcgSs

[11] T. Fiebig, S. Gürses, C. H. Gañán, E. Kotkamp, F. Kuipers, M. Lindorfer, M. Prisse, and T. Sari, "Heads in the clouds? measuring universities' migration to public clouds: Implications for privacy & academic freedom," in *Proceedings on Privacy Enhancing Technologies Symposium*, vol. 2023, no. 2, 2022.

[12] V. Paxson, "Strategies for sound internet measurement," in *Proceedings of the ACM Internet Measurement Conference*, 2004.

[13] Q. Scheitle, O. Hohlfeld, J. Gamba, J. Jelten, T. Zimmermann, S. D. Strowes, and N. Vallina-Rodriguez, "A long way to the top: Significance, structure, and stability of internet top lists," in *Proceedings of the ACM Internet Measurement Conference*, 2018.

[14] D. Zeber, S. Bird, W. R. Camila Oliveira, I. Segall, F. Wollsén, and M. Lopatka, "The representativeness of automated web crawls as a surrogate for human browsing," in *Proceedings of the International Conference on World Wide Web (TheWebConf)*., 2020.

[15] "Rapid7 Project Sonar," https://www.rapid7.com/research/project-sonar/, 2025.

[16] N. P. Hoang, A. A. Niaki, M. Polychronakis, and P. Gill, "The web is still small after more than a decade," *ACM SIGCOMM Computer Communication Review*, vol. 50, no. 2, pp. 24–31, 2020.

[17] F. Cangialosi, T. Chung, D. Choffnes, D. Levin, B. M. Maggs, A. Mislove, and C. Wilson, "Measurement and analysis of private key sharing in the https ecosystem," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 628–640.

[18] R. van Rijswijk-Deij, M. Jonker, A. Sperotto, and A. Pras, "A high-performance, scalable infrastructure for large-scale active dns measurements," *IEEE journal on selected areas in communications*, vol. 34, no. 6, pp. 1877–1888, 2016.

[19] R. Sommese, G. Akiwate, M. Jonker, G. C. Moura, M. Davids, R. v. Rijswijk-Deij, G. M. Voelker, S. Savage, A. Sperotto *et al.*, "Characterization of anycast adoption in the dns authoritative infrastructure," in *Network Traffic Measurement and Analysis Conference (TMA'21)*, 2021.

[20] A. Abhishta, R. van Rijswijk-Deij, and L. J. Nieuwenhuis, "Measuring the impact of a successful ddos attack on the customer behaviour of managed dns service providers," *ACM SIGCOMM Computer Communication Review*, vol. 48, no. 5, pp. 70–76, 2019.

[21] S. Matic, C. Iordanou, G. Smaragdakis, and N. Laoutaris, "Identifying sensitive urls at web-scale," in *Proceedings of the ACM Internet Measurement Conference*, 2020, pp. 619–633.

[22] J. R. Smith, H. Saint-Amand, M. Plamada, P. Koehn, C. Callison-Burch, and A. Lopez, "Dirt cheap web-scale parallel text from the common crawl." Association for Computational Linguistics, 2013.

[23] R. Sommese, R. van Rijswijk-Deij, and M. Jonker, "This is a local domain: On amassing country-code top-level domains from public data," *ACM SIGCOMM Computer Communication Review*, vol. 54, no. 2, pp. 2–9, 2024.

[24] J. Gustafsson, G. Overier, M. Arlitt, and N. Carlsson, "A first look at the ct landscape: Certificate transparency logs in practice," in *Passive and Active Measurement: 18th International Conference, PAM 2017, Sydney, NSW, Australia, March 30-31, 2017, Proceedings 18*. Springer, 2017, pp. 87–99.

[25] E. Fasllija, H. F. Enişer, and B. Prünster, "Phish-hook: Detecting phishing certificates using certificate transparency logs," in *Security and Privacy in Communication Networks: 15th EAI International Conference, SecureComm 2019, Orlando, FL, USA, October 23–25, 2019, Proceedings, Part II 15*. Springer, 2019, pp. 320–334.

[26] A. Drichel, V. Drury, J. von Brandt, and U. Meyer, "Finding phish in a haystack: A pipeline for phishing classification on certificate transparency logs," in *Proceedings of the 16th International Conference on Availability, Reliability and Security*, 2021, pp. 1–12.

[27] Y. Lee and N. Spring, "Identifying and analyzing broadband internet reverse dns names," in *Proceedings of the 13th International Conference on emerging Networking EXperiments and Technologies*, 2017, pp. 35–40.

[28] O. Dan, V. Parikh, and B. D. Davison, "Ip geolocation through reverse dns," *ACM Transactions on Internet Technology (TOIT)*, vol. 22, no. 1, pp. 1–29, 2021.

[29] O. van der Toorn, R. van Rijswijk-Deij, R. Sommese, A. Sperotto, and M. Jonker, "Saving Brian's privacy: the perils of privacy exposure through reverse DNS," in *Proceedings of the 22nd ACM Internet Measurement Conference*, 2022, pp. 1–13.

[30] OpenINTEL, "Openintel background information," 2024, accessed: December 01, 2024. [Online]. Available: https://openintel.nl/background/

[31] C. Crawl, "Common crawl dataset," 2024, accessed: August 01, 2024. [Online]. Available: https://commoncrawl.org

[32] ——, "Common crawl web graphs," 2024, accessed: December 12, 2024. [Online]. Available: https://commoncrawl.org/web-graphs

[33] B. Laurie, "Certificate transparency: Public, verifiable, append-only logs," *Queue*, vol. 12, no. 8, p. 10–19, aug 2014. [Online]. Available: https://doi.org/10.1145/2668152.2668154

[34] R. Sommese and M. Jonker, "Poster: Through the ccTLD Looking Glass: Mining CT Logs for Fun, Profit and Domain Names," in *Proceedings of the 2023 ACM on Internet Measurement Conference*, 2023, pp. 714–715.

[35] "Censys search," https://censys.io/, 2024, accessed: November 1, 2024.

[36] L. Izhikevich, G. Akiwate, B. Berger, S. Drakontaidis, A. Ascheman, P. Pearce, D. Adrian, and Z. Durumeric, "Zdns: a fast dns toolkit for internet measurement," in *Proceedings of the 22nd ACM Internet Measurement Conference*, ser. IMC '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 33–43. [Online]. Available: https://doi.org/10.1145/3517745.3561434

[37] G. Stapleton and P. Rodgers, "Drawing euler diagrams with circles and ellipses," in *Proceedings of the IEEE Symposium on Visual Languages and Human-Centric Computing*, 2011.

[38] J. Larsson, "eulerr: Area-proportional euler diagrams with ellipses," Master's thesis, Lund University, 2018.

[39] (2024) Tranco top site ranking list. [Online]. Available: https://tranco-list.eu/

[40] Censys Support, "Certificate transparency and pre-certificates," 2024, accessed: 2024-12-04. [Online]. Available: https://support.censys.io/hc/en-us/articles/13563018212628-Certificate-Transparency-and-Pre-certificates

[41] 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," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 569–586.

APPENDIX

A. Ethics

This work does not raise any ethical issues.

B. Additional Comparison Tables

TABLE A1: Counts and Percentages for OpenINTEL and Sonar A Records

OpenINTEL	Sonar	April FQDN	May FQDN	June FQDN	April SLD	May SLD	June SLD
●	●	243,859,243 (11.10%)	241,773,498 (10.73%)	241,551,746 (10.76%)	186,196,181 (72.73%)	185,285,945 (72.39%)	185,331,300 (72.48%)
●	○	149,802,888 (6.82%)	148,521,222 (6.59%)	148,953,985 (6.63%)	12,603,865 (4.92%)	11,774,830 (4.60%)	11,930,310 (4.67%)
○	●	1,804,199,072 (82.09%)	1,862,814,222 (82.68%)	1,855,170,080 (82.61%)	57,200,383 (22.34%)	58,897,037 (23.01%)	58,443,494 (22.86%)
Total		2,197,861,203	2,253,108,942	2,245,675,811	256,000,429	255,957,812	255,705,104

TABLE A2: CT Logs Data with June Domains Still Valid in September

Type	Count	Remaining (%)	Resolution Rate (%)
September FQDN Count	738,097,343	100	62.92
Revoked FQDN Count	5,336,354	0.723	18.31
Precert FQDN Count	156,573,218	21.21	5.33
Unique FQDN Count	412,034,006	55.82	39.55
September SLD Count	201,965,602	100	85.18
Revoked SLD Count	377,031	0.187	82.64
Precert SLD Count	1,995,153	0.988	28.79
Unique SLD Count	35,214,644	17.44	56.15

TABLE A3: Analysis of OpenINTEL FQDN Data Across April, May, and June

Label	Count (% of All OpenINTEL)			% of Unique OpenINTEL		
	April	May	June	April	May	June
All FQDNs	542,440,345 (100.00)	540,344,985 (100.00)	540,167,272 (100.00)			
All Unique FQDNs	217,599,838 (40.11)	212,290,246 (39.29)	207,935,524 (38.49)	100.00	100.00	100.00
A Record FQDNs	393,662,516 (72.57)	390,295,095 (72.23)	390,506,097 (72.30)			
A Record Unique FQDNs	102,760,072 (18.94)	98,063,279 (18.15)	93,722,682 (17.35)	47.22	46.19	45.07

TABLE A4: Sonar Unique FQDNs

Category	April (% of All Sonar FQDN)	May (% of All Sonar FQDN)	June (% of All Sonar FQDN)
Sonar FQDNs	2,048,058,315 (100%)	2,104,587,720 (100%)	2,096,721,826 (100%)
Sonar Unique FQDNs	900,726,415 (43.98%)	935,521,153 (44.45%)	929,713,415 (44.34%)
Sonar Minus RDNS Unique FQDNs	1,642,298,172 (80.19%)	1,699,268,683 (80.74%)	1,693,608,872 (80.77%)

TABLE A5: RDNS Unique FQDNs

Category	April (% of All RDNS FQDNs)	May (% of All RDNS FQDNs)	June (% of All RDNS FQDNs)
RDNS Total FQDNs	1,290,813,367 (100%)	1,229,794,819 (100%)	1,225,731,680 (100%)
RDNS Unique FQDNs	543,727,497 (42.12%)	460,516,810 (37.45%)	456,324,345 (37.23%)
RDNS and Sonar FQDNs	741,571,757 (57.45%)	763,747,530 (62.10%)	763,895,457 (62.32%)
RDNS and other FQDNs	5,514,113 (0.43%)	5,530,479 (0.45%)	5,511,878 (0.45%)