# Evaluating the impact of design decisions on passive DNS-based domain rankings

Victor Le Pochat *, Simon Fernandez §, Tom Van Goethem *, Samaneh Tajalizadehkhoob ‖,
Lieven Desmet *, Andrzej Duda §, Wouter Joosen *, Maciej Korczyński §
* DistriNet, KU Leuven § Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG ‖ ICANN

***Abstract*—'Top sites' rankings of the most popular domains are a core resource for the large-scale measurements that are crucial in Web and Internet research. Recent rankings evolved towards using passive DNS traffic data, but this data's suitability for measuring website popularity is poorly understood. In this paper, we holistically evaluate how design decisions influence the composition and desired properties of passive DNS-based domain rankings. We isolate the effects of these decisions by generating a ranking from the ground up using aggregated "post-recursor" passive DNS data. We evaluate the impact of corrections for resolver caching and CDNs, and confirm that measures such as service classification, bucketing, or long-term aggregation produce more reliable rankings. Our goal is to give transparent insight into the process of using passive DNS data for domain rankings, as a framework for the research community to understand how to develop future rankings that address their needs.**

## I. Introduction

Large-scale Web and Internet measurements play a crucial role in enabling research into how the modern Internet works. Researchers have emphasized that such measurements should be accurate, reproducible, and representative [1–5]. A core element in conducting these measurements according to those expectations concerns the selection of a (representative) sample of websites that form the subject of the study. Most often, researchers use 'top sites' rankings of the most popular domains for this purpose, with hundreds of studies relying on these rankings [6–9]. The academic research community has recently raised awareness on the issues surrounding these rankings that threaten their reliability, such as low agreement [6, 10], stability, and transparency [7, 8]. Nevertheless, they remain a valuable and necessary tool that the research community relies on.

Initially, popularity rankings depended on Web traffic collected through toolbars, extensions, or trackers [8]. However, it is challenging to recruit users who are willing to share their traffic data, and to obtain such data in a privacy-preserving way. Owing to these challenges, the Web-based rankings that researchers rely on are steadily disappearing, with Quantcast and Alexa discontinuing their rankings in 2020 and 2022 respectively. Only the Chrome User Experience Report (CrUX) [11] has emerged as an alternative Web-based ranking since then, which depends on Chrome users opting in to URL sharing [6]. To overcome the privacy challenges and subsequent unavailability of Web traffic data, several providers have recently developed new rankings based on DNS resolver traffic, including Cisco [12], Cloudflare [13], and DomainTools [14]. Despite this shift to DNS traffic, we still lack a good understanding of whether these DNS-based rankings approximate website popularity well. Given the significant impact of ranking properties on research results across all areas of Web and Internet measurement [7], and the need to consider alternatives to the disappearing Web-based rankings, we set out to evaluate the impact of the design decisions made by DNS-based ranking providers in their rankings' designs, to establish whether DNS-based rankings are appropriate and sufficiently reliable for conducting Web and Internet measurements.

To identify which measures contribute to improving the reliability and suitability of these rankings, we design a custom ranking method for passive DNS traffic data, from the ground up. This enables us to holistically and independently evaluate the influence of design decisions on the composition and desired properties of domain rankings, and isolate these effects, as opposed to comparing existing rankings where the underlying data collection and processing are confounding factors [6–8, 10]. We use a feed of "post-recursor" ("above-resolver" [15]) passive DNS traffic data [16] from SIE Europe [17], which is aggregated from multiple "sensors", i.e., recursive resolvers. Its main advantages are the increased coverage and diversity across networks and organizations, and better preservation of user privacy, as data is aggregated and no IP-level data is retained. However, there are significant challenges for measuring popularity: the observed per-domain counts represent cache misses at the recursive resolvers, i.e., not all user requests are visible; ranking methods based on user-level data, such as IP-based voting [18], cannot be used; and inherently, DNS traffic mixes websites and top-level browsing context visits with infrastructural domains and background DNS resolutions.

We assess to which extent design decisions improve the stability of rankings. We evaluate design decisions introduced in recent rankings [8, 11, 13] ('bucketing' ranks and long-term averaging) as well as correcting mechanisms that we develop using additional data available within our feed (accounting for CDNs and classifying services). We demonstrate that the recent design decisions improve ranking stability, that the correcting mechanisms are necessary to avoid that certain domain types overly dominate the ranking, and that service classification caters for the different usage patterns of domain rankings. Finally, while the simplest solution for ranking domains is to sort them by their raw query counts [19], the observation

978-3-903176-64-5 

of only cache misses skews the interpretation of these query counts. Specifically, resolver caching through TTL affects the temporal query patterns and therefore query volumes of each domain, potentially causing domain popularity to be both over- and under-estimated. We weight DNS traffic by TTL, and while after TTL-weighting, traffic distributions closely match those observed on the web, the distribution of observed query counts is more skewed than without weighting, and the suspectibility to ranking manipulation is amplified, meaning that the design decision to incorporate TTL (or not) must be carefully made.

Overall, our contribution confirms that the design decisions introduced in recent (DNS-based) rankings are beneficial and make these rankings suitable for Internet and Web measurements. This improved understanding of the fundamentals of building these rankings allows our community to continue working towards rankings that meet the needs of researchers and that fulfill the soundness and validity requirements that ultimately increase trust in the resulting research findings [20].

## II. Background

### A. Data sources and processing

Rankings of the most popular or "top" domain names or websites use one of three types of data sources. Providers like Alexa, Quantcast, or more recently the Chrome User Experience Report (CrUX) [11] rely on *web traffic* data, gathering it from browser reports or in-page scripts. Providers like Majestic and Common Crawl [21] use the *web link graph* to rank websites that are most often referred to from other websites, similar to how search engines (at their core) rank search results. A third source is *passive DNS traffic*, monitoring DNS traffic at resolvers to derive the most popular domain names. Finally, multiple existing rankings (and therefore sources) can be merged into one aggregated ranking, as is done in Tranco [8].

Given that we evaluate design decisions on passive DNS-based rankings, we look in more detail at the existing publicly available lists and approaches using such data. WebShrinker's *DNSFilter* list [22] uses traffic to their DNS resolvers. Their list ranks domains on the 30-day sum of a daily count of organizations querying each domain, therefore incorporating both the longevity and reach of domain accesses. Cloudflare's *Radar* list [13] uses traffic to their 1.1.1.1 resolvers. They use a machine learning model that predicts each domain's rank based on an undocumented feature set, designed to estimate the "relative size of the user population that accesses a domain" [13]. The *SecRank* list [18] uses traffic from the Chinese 114DNS resolvers. The ranking is based on voting across the domain preferences of individual IP addresses, weighted according to each IP's domain diversity and query volume. Cisco's *Umbrella* list uses traffic to their DNS resolvers [12]. Their list ranks domains on "the number of unique client IPs invoking [each] domain, relative to the sum of all requests to all domains".

A common thread across these DNS-based rankings is that they seek to incorporate the number of distinct entities (users, organizations) querying a domain, often to better approximate genuine human Web traffic. In contrast, aggregated post-recursor data does not contain any information on who requests a domain, so approaches using such data can only be designed to use query volume data. Such data is also used for the *Farsight* list from DomainTools [14], but their methods are not documented, and the list is not publicly available. We transparently explore the challenges brought about by this limitation, and discuss potential solutions in our work.

### B. Related work

Lists of popular domain names have long been used in Internet measurement research (since at least 2005 [23, 24]), but were only thoroughly scrutinized starting in 2018. For three rankings (Alexa, Umbrella, and Majestic), Scheitle et al. [7] described the methods, research usage, characteristics such as stability, and potential impact on Internet measurement research. Le Pochat et al. [8] similarly characterized the three rankings and Quantcast from a security perspective, and showed that all four rankings could be manipulated to insert any domain. Rweyemamu et al. studied aspects such as weekly patterns and domain clusters in detail [25], and refined the manipulation attacks for Alexa and Umbrella [26]. Xu et al. [27] showed how certain open DNS resolvers respond to non-recursive queries in a way that can be abused to manipulate passive DNS-based rankings. Two academic initiatives then sought to design and publish domain rankings that improve upon properties important for research: Tranco [8, 28] and SecRank [18].

Since the appearance of these new lists, further studies have compared top lists and evaluated their accuracy in reflecting popularity. Alby and Jäschke [10] expanded the comparison of top lists to other data sets (e.g., Wikipedia) and search engine results. They find low overlap between data sets, with hosts present or popular in one data set missing from others. They recommend that researchers select a random sample of websites among Common Crawl hosts. Ruth et al. [6] compared top lists to Cloudflare traffic data, concluding that the Chrome User Experience Report [11] most accurately represents the more popular websites, albeit as an unordered set. In general, however, agreement of any list with the 'ground-truth' traffic remains relatively low. Recent work has also shown that popular websites differ significantly between countries and languages [10, 29]. Xie and Li [30] measured real-world top lists use, finding that ranked domains experience an increase in traffic, corroborating the broad appeal of domain rankings.

Beyond these proposals and evaluations for public rankings, several DNS-based ranking methods have been proposed but did not materialize to concrete rankings. Proposals include ranking domains on counts of querying DNS resolvers [31], unique querying clients [32], or per-client DNS query counts [33]. Other proposals estimate client query volumes and therefore popularity through active DNS cache probing [34–37]. Finally, the DNS Observatory [15] tracks top objects including domains in passive DNS post-recursor traffic, similar to our data set.

## III. How to rank

### A. What ranking properties do we expect?

To thoroughly evaluate whether rankings fulfill community needs, we must first understand what the community expects

from these rankings. Inherently, we want the ranks to reflect how *popular* a given domain is. However, what constitutes popularity in and of itself is not necessarily well-defined. Most (passive DNS-based) rankings define popularity as the number of users or organizations accessing a domain, usually counting at most one access per day. However, this disregards how often and how regularly a domain is queried – for example, a search engine or news website might be visited very often throughout one day, yet would not have a better rank than a software update domain that is queried only once a day but by as many users. As one example of a more elaborate metric to quantify popularity, the ranking approach of SecRank [18] integrates query volume, query regularity, and IP address count.

Rankings should also exhibit properties that serve as requirements to make them more suitable for research usage. Prior work evaluated rankings across these dimensions [6–8, 18]:

- *Accuracy*: rankings should correctly capture popularity, i.e., a better ranked domain should genuinely be more popular than a worse ranked one.
- *Agreement* or *similarity*: if all rankings correctly capture popularity, they should place domains at the same ranks.
- *Manipulation resistance*: rankings should be designed such that domains cannot easily be inserted, removed, or shifted.
- *Representativeness*: rankings should correctly reflect Internet-wide distributions and have good coverage across, e.g., site categories, or countries.
- *Reproducibility*: rankings should be easily retrievable and uniquely referenceable such that others can reproduce studies with the exact rankings used in those studies.
- *Stability*: rankings should be sufficiently stable over time, while still incorporating natural changes to popularity.
- *Transparency*: rankings should publish details on the methods used in the ranking for better insight into possible biases.

### B. Design decisions on composition

Certain design decisions can be made to influence the composition of a ranking and ultimately contributed to achieving the previously listed desirable properties of top lists. While these design decisions do not necessarily affect how well a ranking reflects popularity, they still impact how a ranking can be used and how it should be interpreted. Our evaluation (Section VI) is focused on assessing how significantly these design decisions affect a ranking's composition.

- Domains represent *websites*, i.e., resources that humans tend to view in a top-level browsing context, or *infrastructure*, i.e., resources such as nameservers. Depending on its data source, a ranking may naturally tend to contain more domains of one type: Web traffic and link graph lists will likely almost exclusively contain websites, while DNS-based lists will also include infrastructure domains. This distinction will affect what data can be retrieved from a domain and is therefore important to ensure that a ranking fits a research study's purpose. For example, an infrastructure domain may not host any Web content, therefore appearing unreachable in a Web measurement (and potentially skewing its results).
- A ranking can list *root*[1] domains (e.g., Alexa), *fully qualified domain names* (e.g., Umbrella), *origins* (e.g., CrUX), or even *URLs* (e.g., Hispar [9]). This will affect the level of detail in the analysis: e.g., measuring only root domains ignores subdomain-specific properties or content.
- Each entry can have an individual rank, or entries can be 'bucketed', where only a rank bracket is given (e.g., 'top 10000'), as is done in Radar and CrUX. This is meant to better match the 'long-tail effect': human browsing patterns concentrate on a small number of very popular websites [29], while less popular websites (in the long tail) quickly have much less traffic and therefore become more difficult to accurately rank [13]. While most studies ignore individual ranks in existing lists [6], the lack of ranks can still affect the level of detail at which a certain observation can be correlated to a domain's rank.
- A ranking can be computed using data from a certain period of time. This time frame will affect the ranking's stability versus agility, as well as its update frequency. DNSFilter and CrUX are averaged across (and updated every) 30 days, Radar across 7 days, and most other lists across 1 day. Tranco is updated every day but averages across 30 days of data.
- Each of these decisions, as well as the observed traffic volumes and (diversity of) traffic origins in a data source, affects the final length of a ranking. Commonly, rankings have been cut off at 1 million entries. The ranking length will affect how many resources a measurement can (dis)cover.

## IV. Post-recursor passive DNS data

We use post-recursor passive DNS data for our evaluation of ranking design decisions, as it allows us to measure these decisions independently. To readers unfamiliar with (passive) DNS measurement, we recommend the tutorial by van der Toorn et al. [38] as a primer.

### A. Motivation

Passive DNS data is becoming increasingly common as a data source for domain rankings, as it is a useful large-scale repository of network activity that can be used to infer domain popularity. Its advantages compared to other traffic data are:

- Web traffic data is usually collected from individual users. Passive DNS data can be easily collected from (large) recursive resolvers, increasing the user base across which domain popularity is measured.
- Passive DNS data can be aggregated across a diverse range of service providers, therefore smoothing over differing usage patterns, e.g., mixing residential and corporate traffic [7, 25].
- Specifically for post-recursor passive DNS data, it better preserves user privacy as the raw data is aggregated over the users of all resolvers. For Web traffic data and pre-recursor passive DNS data, raw data can be traced back to individual users, potentially revealing their personal browsing habits.

[1]Also called *eTLD+1* – one level above the effective top-level domain – or *pay-level* domains.

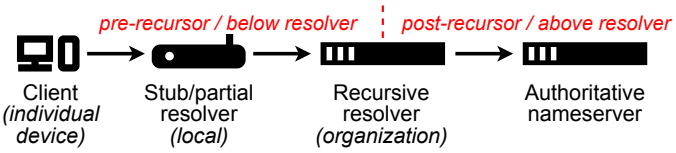


Fig. 1. Parties in a recursive DNS resolution, where we highlight the boundary between pre- and post-recursor traffic.

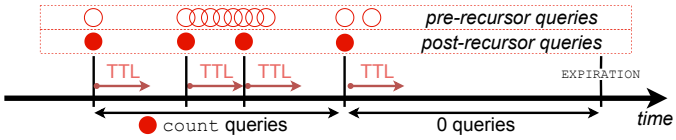


Fig. 2. Pre-/post-recursor DNS traffic have different query patterns, caused by resolver caching and TTL. We only observe the post-recursor query `counts`.

- Owing to these previous limitations, providers may be more willing to share raw (post-recursor) passive DNS data with researchers for the development of new domain rankings.
- The data contains all observed DNS records, enabling their usage for applying corrections (e.g., `CNAME` redirects) or for service classification (e.g., `MX`/`NS` hinting at non-websites).

Nevertheless, passive DNS data has certain disadvantages:

- DNS traffic mixes (human) top-level browsing context visits and (automated) background DNS resolutions. This means that a passive DNS-based ranking will interleave websites and infrastructural domains.
- The selection of recursive resolvers matters for the (global) coverage and representativeness of the ranking. The risk exists that the user base is too small or too skewed to specific user types. Simultaneously, domain popularity is country-specific [29], making the coverage of countries also important to the representativeness of the data.
- Specifically for post-recursor passive DNS data, it precludes the use of ranking methods that incorporate per-user statistics, e.g., the number of distinct users requesting a domain.

The combination of data availability, privacy preservation, and access to raw data leads us to use post-recursor passive DNS data for our evaluation of the influence of design decisions, including the reliability of passive DNS-based rankings for all areas of Web and Internet measurement [7].

### B. Data provenance and format

The input for the rankings that we generate for our evaluation is a continuous live feed of passive DNS entries in the *Passive DNS Common Output Format*, published as an IETF Internet Draft [16, 39] listing the mandatory and optional fields in each feed message. This type of feed is already available from large passive DNS aggregators such as Farsight Security's DNSDB,[2] SIE Europe [17], CIRCL,[3] or mnemonic.[4] Such a feed aggregates DNS requests from multiple 'sensors', i.e., recursive resolvers from organizations or networks that share their data with the passive DNS aggregator. Figure 1 shows how resolvers execute DNS resolution, and where the boundary between pre- and post-recursor traffic lies. The sensors only observe the post-recursor traffic, i.e., the recursive queries to the authoritative nameserver for entries missing from the recursive resolver's cache. Each sensor sends a copy of this traffic to the aggregator that will merge them into a single feed.

The passive DNS aggregator maintains its own cache with a count of observations of tuples of *requested resource* (`rrname` field, i.e., the domain), *record type* (`rrtype`, e.g., `A` or `AAAA`), and *record(s)* (`rdata`, value(s), e.g., IP address(es)). When a tuple is observed for the first time, the aggregator stores it in its cache. Subsequent observations by any sensor increment the count. Entries in this cache expire either when the cache is full or after a set time (for our SIE Europe feed, we estimate this expiration time at 6 hours), upon which an `EXPIRATION` message is emitted to the feed. This message contains the previously described tuple and a `count` field for the total number of observations across all sensors. Since the entry is being removed from the cache, this count is then implicitly reset to 0. Figure 2 shows how the `count` field relates to queries pre- and post-recursor; we observe only the latter. In addition, we rely on the `rrttl` field in this message for the (authoritative) Time-To-Live (TTL) record value. While this field is available in the data feed that we use, the field is not documented in the draft standard, so it is not guaranteed to be present. If necessary, it could be retrieved separately through an active DNS query. Finally, the draft standard has an optional `sensor_id` field for identifying the individual sensor at which the record was seen. It is not present in all passive DNS feeds (this also holds for ours), so we do not expect or use it. In addition, the use of this field could raise privacy issues as it may make it easier to trace observations of specific domains to specific sensors.

## V. Ranking Generation

To conduct our evaluation of the influence of design decision, we use a feed of aggregated post-recursor passive DNS data to generate three rankings: one each for fully qualified domain names (FQDNs), root domains, and domains that likely host websites (i.e., pages visited by regular Internet users while browsing). Figure 3 schematically represents how we process the data feed to ultimately generate our three ranking types.

### A. Processing resource records

*1) Observation counters:* For the final popularity tally, we track the counts of observations per domain (i.e., requested resource) for the `A` and `AAAA` record types. We only observe and count the requests above the recursive resolver — they represent resolver cache misses. Observation counts therefore depend on the resolver caching behavior, which is primarily fixed by the TTL value set by the authoritative nameserver and sent along with the resource records. Next to one variant ranking where we ignore TTL, we generate a second variant where we multiply all observation counts by this TTL value, extracted from the passive DNS data itself. Simply put, we

[2]https://www.farsightsecurity.com/solutions/dnsdb/
[3]https://www.circl.lu/services/passive-dns/
[4]https://docs.mnemonic.no/api/services/pdns/

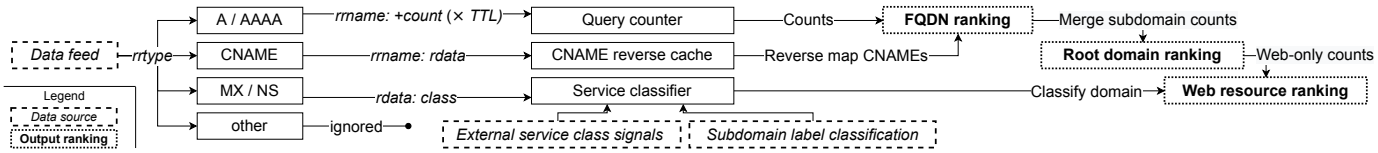


Fig. 3. We process the resource records from the passive DNS data feed to populate the counters and caches, which we use to compute three rankings.

expect a domain with twice the TTL value to have half as many visible requests, so we need to multiple its query count by two. In case the TTL is zero, we retain the original count. Because TTL values of more than 1 day (86400 seconds) are commonly truncated by resolvers [40], and as we primarily use 1 day of data to generate rankings, we truncate all TTL values to 1 day as well. Because of the dominance of nameservers with 2-day TTLs (Section VI-C), this truncation is applied for 48% of queries, although this only affects 3.2% of FQDNs.

*2) CNAME reverse cache:* We track the counts of A/AAAA record observations, i.e., query results instead of the queries themselves. Therefore, if a user resolves a domain that uses a CNAME record to map to the A/AAAA records of a second domain, we count an observation for this second domain, not for the domain the user originally visited. In particular, large CDNs would therefore be over-counted (Section VI-C), because CDN-hosted domains often point a CNAME record to a CDN (sub)domain that hosts the actual website content [41, 42]. Another use case for CNAME redirection is mimicking first-party context for tracking purposes [43].

We therefore maintain a cache of all observed mappings in CNAME records observed in the feed of the ranking's day. We recursively reverse the resulting mapping for all domains in our final ranking. If multiple records map to the same domain, we select the reversed domain for which we observed the highest total query count. We then retain the mapped domain if it is observed often enough (CNAME count $> 1\%$ of the A/AAAA count) to ignore erroneous CNAME records, and if it is not a subdomain of the original domain, to prioritize root domains.

*3) Service classifier:* Our data feed comes from DNS queries, and therefore contains infrastructural domains such as nameservers that are not visited by humans through browsers. However, certain studies focus on user Web traffic and require website-only rankings, akin to Alexa, CrUX, or Majestic.

We therefore need to determine what *service* each domain offers, to then separate domains into service-specific rankings. In our design, we use internal and external *passive* sources to determine the service. Internally, we mark the domains appearing in MX and NS record values as mailservers and nameservers, respectively. We also specifically mark nameservers in the root zone file, as they are not included in the passive DNS feed [44]. We manually develop a service mapping for the most commonly observed subdomain labels, e.g., www for a web service, or ns1 for a nameserver. Externally, we collect websites from three web-focused domain lists (BuiltWith,[5] CrUX [11], and DomainRank[6] [45]), and domains in DBpedia [46].

[5]https://builtwith.com/top-1m

[6]https://www.domainrank.io/download.html

To resolve contradictory classifications from multiple sources (e.g., when a NS record mistakenly contains a web domain), we develop a scoring system that accounts for the reliability and the certainty between and within sources, e.g., awarding higher scores for records observed more often. We first classify individual FQDNs; if there is no known class, we recursively search a class of the parent domain. To determine the class of a root domain, we compare its known class (if any) with that of its subdomains, and heuristically select the most likely class based on the scores and query counts of each. If a domain offers multiple services, we assign it the most common service.

### B. Computing the ranking

We generate daily rankings based on that day's query counts, CNAME reverse cache and service classifier data. To fill these counters and caches, we process 1-day slices of the data feed. For rankings across larger time periods (e.g., 7 days), we combine these processed 1-day slices by summing counts and merging caches. We first compute the base FQDN ranking of fully qualified domain names, mapping each FQDN to its reversed CNAME domain if applicable. We then generate a 'root' ranking across the total counts for each root domain's subdomains. We extract the root domain for each FQDN using the Public Suffix List (PSL) [47]. We use the public section of the PSL such that the root ranking contains only true eTLD+1s. This means that the root ranking has one single entry for domains listed in the private section (a user-contributed subset of domains for which subdomains are maintained by separate users, e.g., *.blogspot.com), and we sum the counts of all subdomains to determine the root domain's rank. A relatively large part of the FQDN ranking (4.95%) belongs to a domain in the PSL's private section, mostly owing to the listing of large CDN domains in the private section. Finally, we extract a 'web' ranking of root domains classified as a website, using only the counts of subdomains classified as a website.

### C. Limitations

Rankings based on post-recursor passive DNS traffic are inherently limited by the traffic representing only cache misses, heavily aggregated across large-scale recursive resolvers. We have no true count of client queries (to the recursive resolver) nor identifiers for individual clients or sensors, and can therefore not incorporate this into the ranking generation process for our evaluation; hence our goal of evaluating whether post-recursor traffic can still be useful for modeling domain popularity despite these restrictions. The aggregator cache also reduces temporal granularity, as counts are aggregated over six hours, which we deem too long to meaningfully integrate in our design, e.g., for client query modeling based on DNS request

inter-arrival times [35]. Evolutions and optimizations in DNS resolution, such as QNAME minimization [48, 49] or encrypted queries [50], may affect the reliability or comprehensiveness of our traffic data. Since we observe DNS queries and not (top-level browsing context) web visits, query counts may also be inflated by website redirects [51], and by the inclusion of third-party resources.

In terms of data quality, we have no reason to believe that the counts provided by the passive DNS aggregator or individual resolvers would be inaccurate. In contrast, we rely on record values set by the domain operators themselves. We sometimes observe inaccuracies that could affect our data processing and subsequent results, e.g., `google.com` incorrectly being set as a `CNAME` or `NS` record value. We account for these errors by (heuristically) setting minimum thresholds for our CNAME reverse cache, and by prioritizing data sources based on reliability for our service classifier. We also use the (authoritative) TTL values, which resolvers may not always respect [18, 19] – e.g., Moura et al. found that values under 1 hour were rarely changed, but TTLs over 1 day were more frequently truncated [40, 52]. For ranking manipulation, incorporating TTL could allow an attack to set high TTL values as an amplifier [15, 52]. Attackers could then issue queries if they can get a sensor to query the domain, e.g., through IP spoofing [53]. We already propose truncating high TTL values, and while we do not have access to sensor-level data, an attacker would still need to discover the unknown sensors and trigger a DNS query from as many of them as possible to achieve a (better) rank. Moreover, the deployment at sensors of proper protections against IP spoofing could prevent attackers from issuing the necessary DNS queries [54, 55].

We monitor, store, and process a live feed of passive DNS entries as the basis of our data set. Such feeds can produce up to tens of gigabytes of aggregated data per hour, requiring a stable and fast network connection with the aggregator and efficient storage systems to preserve all data. During our data collection, we experienced multiple small network disruptions that led to the ingestion of the live feed silently stopping, requiring manual intervention to restart feed processing. These outages has led to gaps in our source data and therefore our daily generated rankings. We verified for the date ranges that we retain (Section VI) that the feed ingestion ran uninterrupted for the entire day; we discarded rankings of days where fewer than 24 hours of data was collected.

The lack of definitive ground truth to evaluate accuracy forms a challenge for our evaluation. Ruth et al. [6] come closest, comparing existing rankings to Cloudflare web traffic (covering around 10% of all websites), but given a lack of open data, we cannot replicate their method. We are also dependent on one specific data aggregator for our evaluation. Our method can be applied to all passive DNS data sources, but our concrete results inherently only hold for SIE Europe data, and are therefore biased to the usage patterns of the users of its European-based resolvers. Due to this bias, we refrain from directly comparing to other top sites rankings, and instead focus on isolating influences within our own ranking data.

## VI. Evaluation

To evaluate the influence of design decisions on rankings using real-world data, we apply our ranking generation method to passive DNS data sourced from SIE Europe [17], aggregated from sensors located at European-based commercial, government, and higher education organizations.[7] We generate the three ranking types (FQDN, root, and web) using the passive DNS feed for 27 June – 4 July, 13–28 July, 24 August – 11 October, 17 October – 8 November, 15 November – 5 December, and 16–30 December 2023 (gaps are due to outages; see Section V-C). When we describe a ranking for a given day and time period, we use the data observed on that day and the days before it for the duration of the time period. We analyze the composition of the rankings, and the impact of the different steps of the resource record processing (Section V-A) on the rankings.

### A. Ranking length

We first quantify the raw lengths of the computed rankings, i.e., the total number of distinct observed domains over the given period. We do not impose any threshold on the observed query counts for inclusion in the ranking; Section VI-B quantifies in detail the distribution of query counts, which could serve to select a threshold that meets requirements for accuracy and representativeness of a domain's popularity yet produces a sufficiently long ranking.

The lengths of the 1-day rankings vary over time between 3.3M and 9.5M FQDNs, 1.6M and 6.3M root domains, and 0.6M and 1.9M web domains (Figure 4); an average reduction of 42.7% and 70.1% respectively. This shows that the choices of excluding subdomains and/or prioritizing websites significantly reduces a ranking's length. In the first half of our measurement period (until 6 October), the length of the web ranking always exceeds 1 million websites, but this no longer always holds afterwards. Figure 5 shows how raw query volumes dropped from a peak of 1.4 to a low of 0.3 billion queries per day.[8]; If the query volume becomes too low, the ranking no longer achieves the threshold of 1 million websites that is commonly used in other rankings, highlighting the need to observe a sufficient traffic volume to capture diverse long-tail domains.

One way of augmenting query volumes is to aggregate traffic data over a longer timespan. Figure 6 shows how the lengths increase for 7-day rankings, yielding an even larger set of domains that can be measured. Owing to the commonality of some ranked domains across time (Section VI-D), these lengths do not increase seven-fold. Instead, the increase in ranking length is between a factor of 2.7 and 6.0 for FQDN rankings, 2.7 and 7.7[9] for root rankings, and 2.2 and 4.7 for web rankings. As a result, the 7-day web ranking contains at least 2.5M websites, comfortably meeting the 1M threshold.

[7]No details are available on the exact set of organizations contributing to SIE Europe. We also make no attempt at deanonymizing them.

[8]We have no immediate explanation for this variance in query volume. Possible causes are an organic (temporary) decrease in traffic, the removal or malfunctioning of sensors or the data feed, or bandwidth limitations.

[9]This outlier (above 7) corresponds to the trough in query volume, where the data for the previous days contain more domains unseen in the 1-day data.

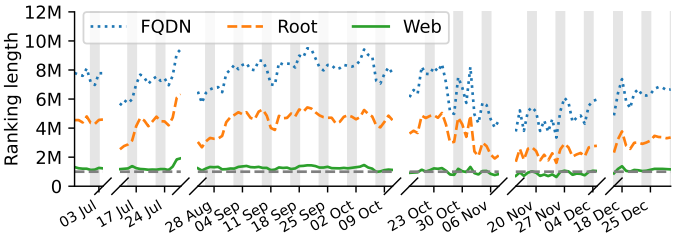


Fig. 4. 1-day FQDN rankings are longer than root and web rankings, at an average reduction of 42.7% and 70.1% respectively. (Weekends are highlighted in gray.)

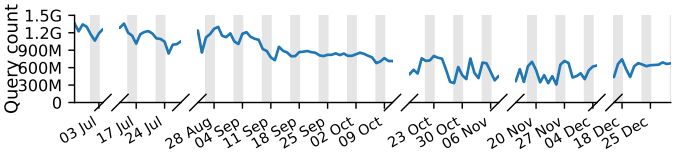


Fig. 5. Observed query counts fluctuate over time, although we do not observe a direct correlation with ranking lengths.

*Takeaway:* It is necessary to observe a sufficiently large traffic volume such that there are enough distinct domains/websites in the ranking. Traffic volume can be increased by aggregating across multiple days, as is done in, e.g., Radar (7 days).

### B. Observation count distribution

Unlike for existing rankings, we have access to the raw scores that determine the final domain ranking. We use this to characterize the estimated traffic distribution across the ranking and its dependence on incorporating TTL, and to understand whether the ranking exhibits patterns that have previously been observed in web traffic.

Figure 7 shows the distribution of the observation counts, both unweighted and weighted by TTL over the domain rank, for the ranking of 1 September 2023. Distributions on other days are similar. Overall, the score distributions appear to adhere to a power law distribution, i.e., the rank and count have an inverse relation. This distribution matches previous observations in website and domain popularity distributions [8, 29, 56, 57]. For the FQDN rankings, there is a stagnation and then drop-off in the head (well-ranked domains). We conjecture that it is the effect of resolver caching: the expected traffic increase (in client queries) is dampened (in cache misses) because caches already serve these requests without going to the authoritative nameserver. There may also be remaining effects from the overcounting of nameservers (Section V-A1). In terms of susceptibility to manipulation, the total unweighted observation counts indicate the query volume required to insert a domain at a given rank. In the TTL-weighted ranking, setting a high TTL amplifies the adversary's ability to perform such manipulation. There is also a drop-off in the tails of all rankings, likely due to shared low discrete query counts. Another view on this ranking tail comes from the absolute differences in query counts from one rank to another. From a rank of around 1,000 onwards, differences sometimes drop to 0 and are generally very low. This effect has also been previously observed in existing rankings, where it was visible through alphabetically ranked clusters [25]. This seems to confirm prior observations [8] that differences between individual ranks in the tail are not (statistically) significant, suggesting that bucketing is also useful for grouping domains with very similar traffic levels.

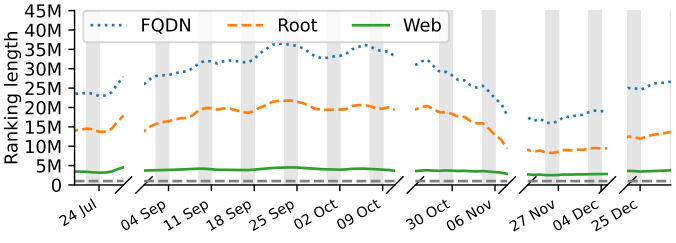


Fig. 6. 7-day rankings, as shown, are much larger than 1-day rankings. Web rankings easily surpass the common 1 million website threshold.

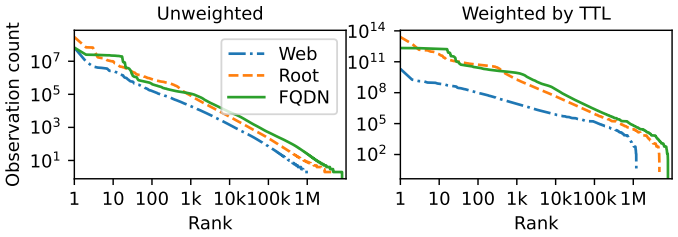


Fig. 7. The rank-size distribution of (un)weighted observation counts appear to adhere to a power law distribution (1 Sep. 2023 rankings).

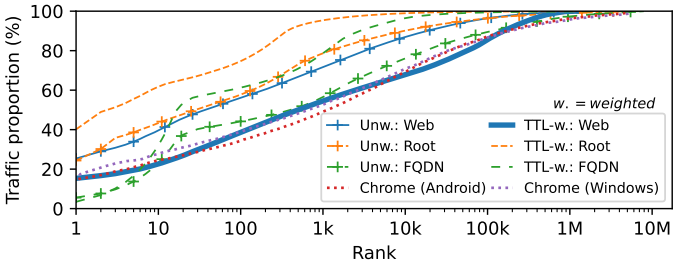


Fig. 8. The TTL-weighted cumulative traffic distribution closely follows the distribution of Google Chrome web traffic [6, Figure 1] (1 Sep. 2023 rankings).

Figure 8 shows the cumulative traffic proportion along the ranking. For the FQDN and root rankings, unweighted rankings attribute more traffic to the head, while for web rankings, the opposite is true. We also compare the web distribution to the Google Chrome traffic distributions traced from Ruth et al. [29]. Here, the TTL-weighted web ranking closely follows this distribution; the unweighted ranking assigns more importance to the head.

*Takeaway:* The observed traffic distributions have a significant effect on the resulting rankings, exacerbated by TTL. The choice to weight query counts by TTL affects the granularity at which domain query volumes can be compared, and how the resulting distributions match those observed empirically.

### C. Correcting mechanisms

We develop two correcting mechanisms as part of our ranking method: CNAME reversal and service classification.

TABLE I
SERVICE CLASS DISTRIBUTION (1 SEP. 2023 FQDN RANKING).

| Class | Percentage | Class | Percentage |
|---|---|---|---|
| Unclassified | 45.79 | IPv4 address | 0.89 |
| Website | 37.65 | CDN | 0.70 |
| Nameserver | 9.31 | Other web service | 0.44 |
| Mailserver | 3.45 | Protocol (FTP, ...) | 0.36 |
| Web admin panel | 1.07 | UUID | 0.34 |

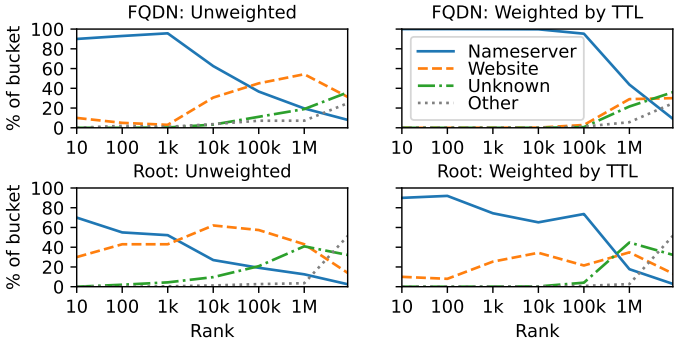


Fig. 9. The distribution of service classes over buckets shows that nameservers dominate the ranking head (1 Sep. 2023 rankings).

On 1 September 2023, across the FQDN ranking, we mapped 261,486 (3.26%) of all subdomains to another subdomain based on CNAME reversal. 90,084 (34.5%) of them mapped to another subdomain within the same root domain. The remaining subdomains are distributed across 24,613 unique root domains and mapped to one of 95,016 other root domains. Across these other subdomains, major CDNs and managed DNS providers are concentrated at the top (e.g., *cloudflare.net*, *azure.com*), who would be overcounted without the CNAME reversal.

54% of the FQDN ranking on 1 September 2023 is classified as a specific service, of which 69.5% as websites (Table I). The rest cannot reliably be classified using our selection of passive sources, and may require additional signals such as an active crawl. Within the root ranking, 33% of domains can be classified, of which 79.7% as a website to be included in the web ranking. Given that our external sources for websites integrate a notion of popularity (for rankings) or notability (for Wiki sources), any unclassified root domains that would be websites are more likely to be unpopular or uninteresting, next to disposable [58] or algorithmically generated [59] domains. In terms of the distribution of service classes over the ranking, nameservers dominate the ranking head, in particular for the FQDN ranking and when weighting by TTL (Figure 9).

*Takeaway:* Using correcting mechanisms, such as CNAME reversal or a service classifier, corrects for otherwise overly dominant domains such as CDNs and nameservers.

### D. Stability

The stability of a ranking represents a balance between rapidly integrating popularity changes and producing a set of domains reusable over time. We measure the stability of the three ranking types by comparing each daily ranking with the previous day's one using two metrics: Spearman's rank correlation, which takes the rank ordering into account, and the Jaccard index, which operates on unordered sets and ignores the ranks. As noted by Ruth et al. [6], these metrics are rather pessimistic; we focus on comparing relative patterns between ranking parameters, and less on absolute results.

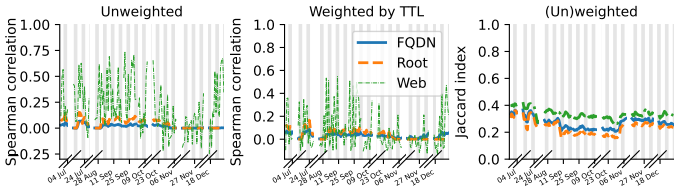


Fig. 10. Between consecutive 1-day rankings, the web ranking is generally the most stable, followed by the FQDN and then root ranking. As the Jaccard index ignores ranks, this metric is equal for the (un)weighted rankings.

Figure 10 shows the patterns that emerge across these stability metrics. The web ranking is generally the most stable, suggesting that web properties exhibit more stable visiting patterns, whereas infrastructural domains may be more volatile (e.g., due to different traffic distribution across servers). Next, the FQDN ranking is more stable than the root ranking, suggesting that our data source sees a similar set of (sub)domains from day to day, but that their combination and re-ranking based on root domains introduces more volatility. Weighting by TTL reduces stability for ranked metrics, suggesting that the unweighted ranking is preferable if stability is desired.

The Spearman rank correlation often exhibits negligible or even negative correlations, suggesting that across the entire ranking (head and long tail), ranks are very unstable over time. This observation supports the preference found among researchers to use rankings as unordered sets [6]. Indeed, the Jaccard index shows higher stability when ignoring rank order. Between using individual ranks and fully ignoring them, using rank buckets may be a good middle ground. Figure 11 shows the evolution of stability per bucket, computed for web rankings using the Jaccard index, between the domain at the first rank after the previous bucket and the domain at the last rank of the current bucket, e.g., 1,001–10,000. First, the metric tends to a much higher value than for the overall ranking (Figure 10) – note that this ranking may be longer than 1 million domains –, suggesting that buckets at the head are more stable in general. Second, the buckets at the head are more stable, with the top 100 bucket being the most stable. (This effect is also most pronounced for the web rankings.) This suggests that at their head, the rankings contain domains that are repeatedly popular.

If a more stable ranking is desirable, e.g., for longitudinal studies, the variability can be decreased by aggregating rankings across a longer period of time [8, 28]. Aggregation of data across 7 days vastly increases the stability for unordered metrics (Figure 12, when compared to Figure 10); for ordered metrics, there is also a positive but much less pronounced effect.

*Takeaway:* Both aggregating across longer data windows and bucketing, as implemented in, e.g., Radar, improve stability. Aggregation achieves this by smoothing over short-term volatility, while bucketing smoothes over the small differences in observed traffic volumes between consecutively ranked domains.

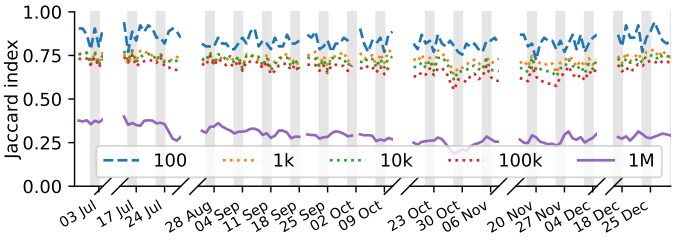


Fig. 11. Between consecutive 1-day web rankings, buckets at the head are more stable.

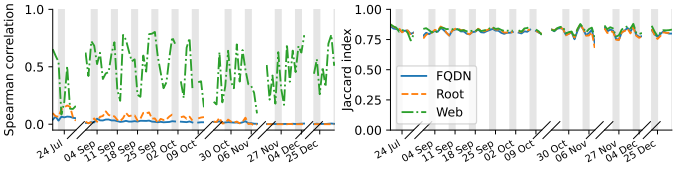


Fig. 12. Between consecutive 7-day unweighted rankings, all rankings become much more stable, in particular on the unordered Jaccard index.

## VII. Discussion and Conclusion

Our goal was to evaluate the impact of design decisions on the development and characteristics of rankings using "post-recursor" passive DNS data, and understand how to leverage such data for generating domain rankings that are suitable for Internet and Web measurements. Based on the design of a ranking method and evaluation on the separate design decisions, we can draw several conclusions on the challenges and necessary steps that turn such data into a usable ranking.

At its base, passive DNS data is usable for generating a website-oriented ranking, primarily after the service classification step; otherwise, nameservers are dominant. This step also lends itself to further development and refinement: we deliberately chose to work only with passive service class indicators to show that this resource-limited approach is already feasible, but service classification may be improved with active indicators (i.e., web crawls that reveal the presence of useful content over HTTP), given the availability of the necessary resources. Correcting factors are necessary: in addition to service classification, we found that without CNAME reversal, CDNs and managed DNS providers would be overrepresented.

Specifically for DNS-based rankings, TTL can be an important factor in accurately assessing domain popularity, as it affects the temporal querying patterns (Figure 2). While incorporating TTL makes theoretical sense, it appears that this does not necessarily improve all ranking properties. For example, nameservers dominate TTL-weighted rankings more heavily, and the observation count distribution is more skewed at the tail. High TTL values can be abused as an amplifier in ranking manipulation. Nevertheless, TTL-weighted web rankings follow known web traffic distributions more closely, so appear more representative in this regard. Incorporating TTL must therefore be carefully considered. For example, Xie et al. [18] decided to omit TTL from SecRank as it was difficult to reliably infer resolver caching behavior from TTL. The choice of including or ignoring TTL is significant, and there is very little similarity between unweighted and TTL-weighted rankings (Appendix C), both when incorporating and ignoring ranks, and even at the ranking head. This direct effect further confirms our indirect observations that TTL affects distributions and stability of rankings. More abstractly, this shows how even for the exact same data set, one design decision can result in wildly different outputs, suggesting that reliably comparing all existing rankings and their variation in data sources and methods is even more challenging. For example, Ruth et al. [6] already had to introduce several normalizations and adjustments, such as filtering on Cloudflare-only sites, to allow for ranking comparison.

Two more recent ranking design elements appear to have a positive impact: rank buckets and long-term aggregation. Buckets yield significant improvements to stability, perhaps unsurprisingly, given the low levels of traffic at the long tail of the ranking, where small differences in rank may, therefore, be meaningless (i.e., statistically insignificant). This effect had already been observed at the heyday of the Alexa ranking [8], and continues in Google Chrome traffic to this day [6]. This also serves as confirmation of Cloudflare's decision to deploy two different models for compiling the head and tail of their Radar ranking [13]. Longer data periods also lead to higher stability, and aggregation over 7 days (like Radar) or even 30 days (like CrUX and Tranco) may therefore be preferable. Also, in this light, while the finding by Ruth et al. that CrUX is more reliable [6] is likely due to Chrome's widespread visibility on Web traffic, these design decisions also play a part in improving CrUX's (and other rankings') accuracy.

In conclusion, (post-recursor) passive DNS traffic can be used to generate a representative domain or website popularity ranking. Despite its limitations, such as the lack of IP-level data, the data is sufficiently rich to generate reliable rankings, as it provides the necessary data for producing traffic volumes, correcting mechanisms, and service classes. Through our evaluation, we are able to confirm the benefits of the design decisions made by recent rankings, specifically the move towards bucketing and longer-term aggregation, although these may not fulfill all use cases, e.g., if more granular ranks are desired. Future directions for evaluating these rankings ideally involve ground-truth data [6], long-term evaluations [28], and an assessment of the impact on research results [7].

## Acknowledgments

We thank the anonymous reviewers and our shepherd Raffaele Sommese for their valuable feedback. We thank SIE Europe for providing access to the passive DNS data feed that enabled this work. This research is partially funded by the Research Fund KU Leuven, and by the Cybersecurity Research Program Flanders. This work has been partially supported by the French Ministry of Research projects PERSYVAL-Lab under contract ANR-11-LABX-0025-01, DiNS under contract ANR-19-CE25-0009-01 and Grenoble Alpes Cybersecurity Institute (ANR-15-IDEX-02).

## References


[1] N. Demir, M. Große-Kampmann, T. Urban, C. Wressnegger, T. Holz, and N. Pohlmann, "Reproducibility and Replicability of Web Measurement Studies," in *2022 ACM Web Conference*, WWW '22, 2022, pp. 533–544. DOI: 10.1145/3485447.3512214.

[2] S. S. Ahmad, M. D. Dar, M. F. Zaffar, N. Vallina-Rodriguez, and R. Nithyanand, "Apophanies or Epiphanies? How Crawlers Impact Our Understanding of the Web," in *The Web Conference 2020*, WWW '20, 2020, pp. 271–280. DOI: 10.1145/3366423.3380113.

[3] D. Zeber *et al.*, "The Representativeness of Automated Web Crawls as a Surrogate for Human Browsing," in *The Web Conference 2020*, WWW '20, 2020, pp. 167–178. DOI: 10.1145/3366423.3380104.

[4] J. Jueckstock *et al.*, "Towards Realistic and Reproducible Web Crawl Measurements," in *2021 Web Conference*, WWW '21, 2021, pp. 80–91. DOI: 10.1145/3442381.3450050.

[5] V. Le Pochat and W. Joosen, "Analyzing Cyber Security Research Practices through a Meta-Research Framework," in *16th Cyber Security Experimentation and Test Workshop*, CSET '23, 2023, pp. 64–74. DOI: 10.1145/3607505.3607523.

[6] K. Ruth, D. Kumar, B. Wang, L. Valenta, and Z. Durumeric, "Toppling Top Lists: Evaluating the Accuracy of Popular Website Lists," in *22nd ACM Internet Measurement Conference*, IMC '22, 2022. DOI: 10.1145/3517745.3561444.

[7] Q. Scheitle *et al.*, "A Long Way to the Top: Significance, Structure, and Stability of Internet Top Lists," in *2018 Internet Measurement Conference*, IMC '18, 2018, pp. 478–493. DOI: 10.1145/3278532.3278574.

[8] V. Le Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczyński, and W. Joosen, "Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation," in *26th Annual Network and Distributed System Security Symposium*, NDSS '19, 2019. DOI: 10.14722/ndss.2019.23386.

[9] W. Aqeel, B. Chandrasekaran, A. Feldmann, and B. M. Maggs, "On Landing and Internal Web Pages: The Strange Case of Jekyll and Hyde in Web Performance Measurement," in *2020 ACM Internet Measurement Conference*, IMC '20, 2020, pp. 680–695. DOI: 10.1145/3419394.3423626.

[10] T. Alby and R. Jäschke, "Analyzing the Web: Are Top Websites Lists a Good Choice for Research?" In *26th International Conference on Theory and Practice of Digital Libraries*, TPDL '22, 2022, pp. 11–25. DOI: 10.1007/978-3-031-16802-4_2.

[11] "Chrome UX Report," Chrome Developers. (2023), [Online]. Available: https://developer.chrome.com/docs/crux/.

[12] D. Hubbard. "Cisco Umbrella 1 Million." (Dec. 14, 2016), [Online]. Available: https://web.archive.org/web/20210502123151/https://umbrella.cisco.com/blog/cisco-umbrella-1-million.

[13] C. Martinho and S. Zejnilovic. "Goodbye, Alexa. Hello, Cloudflare Radar Domain Rankings," The Cloudflare Blog. (Sep. 30, 2022), [Online]. Available: https://blog.cloudflare.com/radar-domain-rankings/.

[14] A. Gee-Clough. "Mirror, Mirror, on the Wall, Who's the Fairest (website) of Them all?" DomainTools. (Apr. 28, 2022), [Online]. Available: https://www.domaintools.com/resources/blog/mirror-mirror-on-the-wall-whos-the-fairest-website-of-them-all.

[15] P. Foremski, O. Gasser, and G. C. M. Moura, "DNS Observatory: The Big Picture of the DNS," in *2019 Internet Measurement Conference*, IMC '19, 2019, pp. 87–100. DOI: 10.1145/3355369.3355566.

[16] A. Dulaunoy, A. Kaplan, P. A. Vixie, and H. Stern, "Passive DNS - Common Output Format," Internet Engineering Task Force, Internet-Draft draft-dulaunoy-dnsop-passive-dns-cof-10, Jun. 2023, Work in Progress, 13 pp. [Online]. Available: https://datatracker.ietf.org/doc/draft-dulaunoy-dnsop-passive-dns-cof/10/.

[17] "A Breakthrough European Data Sharing Collective to Fight Cybercrime," SIE Europe UG. (2021), [Online]. Available: https://www.sie-europe.net/.

[18] Q. Xie *et al.*, "Building an Open, Robust, and Stable Voting-Based Domain Top List," in *31st USENIX Security Symposium*, USENIX Security '22, 2022.

[19] J. St Sauver. "Finding Top FQDNs Per Day in DNSDB Export MTBL Files (Part One of a Three Part Series)," Farsight Security. (Mar. 22, 2019), [Online]. Available: https://www.farsightsecurity.com/blog/txt-record/TopFQDNs-20190322/.

[20] V. Le Pochat, "Reflecting on Research Practices," *Communications of the ACM*, vol. 67, no. 5, May 2024. DOI: 10.1145/3651965.

[21] S. Nagel, `cc-webgraph`, Common Crawl, 2023. [Online]. Available: https://github.com/commoncrawl/cc-webgraph.

[22] P. Lowe, *DNSFilter Top Domains*, DNSFilter, 2023. [Online]. Available: https://github.com/DNSFilter/topdomains.

[23] A. Medina, M. Allman, and S. Floyd, "Measuring the Evolution of Transport Protocols in the Internet," *ACM SIGCOMM Computer Communication Review*, vol. 35, no. 2, pp. 37–52, Apr. 2005. DOI: 10.1145/1064413.1064418.

[24] B. Veal, K. Li, and D. Lowenthal, "New Methods for Passive Estimation of TCP Round-Trip Times," in *6th International Workshop on Passive and Active Network Measurement*, PAM '05, 2005, pp. 121–134.

[25] W. Rweyemamu, T. Lauinger, C. Wilson, W. Robertson, and E. Kirda, "Clustering and the Weekend Effect: Recommendations for the Use of Top Domain Lists in Security Research," in *20th International Conference on Passive and Active Measurement*, PAM '19, 2019, pp. 161–177.

[26] W. Rweyemamu, T. Lauinger, C. Wilson, W. Robertson, and E. Kirda, "Getting Under Alexa's Umbrella: Infiltration Attacks Against Internet Top Domain Lists," in *22nd International Conference on Information Security*, ISC '19, 2019, pp. 255–276.

[27] C. Xu, Y. Zhang, F. Shi, H. Ma, W. Ding, and H. Shan, "Gushing Resolvers: Measuring Open Resolvers' Recursive Behavior," in *4th International Conference on Advanced Information Science and System*, AISS '22, 2022. DOI: 10.1145/3573834.3574533.

[28] V. Le Pochat, T. Van Goethem, and W. Joosen, "Evaluating the Long-term Effects of Parameters on the Characteristics of the Tranco Top Sites Ranking," in *12th USENIX Workshop on Cyber Security Experimentation and Test*, CSET '19, 2019. [Online]. Available: https://www.usenix.org/conference/cset19/presentation/lepochat.

[29] K. Ruth *et al.*, "A World Wide View of Browsing the World Wide Web," in *22nd ACM Internet Measurement Conference*, IMC '22, 2022. DOI: 10.1145/3517745.3561418.

[30] Q. Xie and F. Li, "Crawling to the Top: An Empirical Evaluation of Top List Use," in *25th International Conference on Passive and Active Measurement*, PAM '24, 2024, pp. I-277–I-306. DOI: 10.1007/978-3-031-56249-5_12.

[31] L. Deri, S. Mainardi, M. Martinelli, and E. Gregori, "Exploiting DNS traffic to rank internet domains," in *2013 IEEE International Conference on Communications Workshops*, ICC '13, 2013, pp. 1325–1329. DOI: 10.1109/ICCW.2013.6649442.

[32] A. Mayrhofer, M. Braunöder, and A. Kaplan, "DNS Magnitude - A Popularity Figure for Domain Names, and its Application to L-root Traffic," nic.at GmbH, Tech. Rep., Aug. 5, 2020. [Online]. Available: https://www.nic.at/media/files/nic-report/dns-magnitude-paper-20200805.pdf.

[33] J. L. García-Dorado, J. Ramos, M. Rodríguez, and J. Aracil, "DNS weighted footprints for web browsing analytics," *Journal of Network and Computer Applications*, vol. 111, pp. 35–48, Jun. 2018. DOI: 10.1016/j.jnca.2018.03.008.

[34] C. E. Wills, M. Mikhailov, and H. Shang, "Inferring Relative Popularity of Internet Applications by Actively Querying DNS Caches," in *3rd ACM SIGCOMM Conference on Internet Measurement*, IMC '03, 2003, pp. 78–90. DOI: 10.1145/948205.948216.

[35] M. Abu Rajab, F. Monrose, A. Terzis, and N. Provos, "Peeking through the Cloud: DNS-Based Estimation and Its Applications," in *6th International Conference on Applied Cryptography and Network Security*, ACNS '08, 2008, pp. 21–38.

[36] Z. Wang, "Analysis of DNS Cache Effects on Query Distribution," *The Scientific World Journal*, vol. 2013, pp. 1–8, 2013. DOI: 10.1155/2013/938418.

[37] A. Shimoda *et al.*, "Inferring Popularity of Domain Names with DNS Traffic: Exploiting Cache Timeout Heuristics," in *IEEE Global Communications Conference*, GLOBECOM '15, 2015. DOI: 10.1109/GLOCOM.2015.7417638.

[38] O. van der Toorn, M. Müller, S. Dickinson, C. Hesselman, A. Sperotto, and R. van Rijswijk-Deij, "Addressing the challenges of modern DNS a comprehensive tutorial," *Computer Science Review*, vol. 45, p. 100 469, 2022. DOI: 10.1016/j.cosrev.2022.100469.

[39] R. Edmonds, "ISC Passive DNS Architecture," Internet Systems Consortium, Inc., Tech. Rep., Mar. 2012. [Online]. Available: https://ftp.iij.ad.jp/pub/network/isc/kb-files/passive-dns-architecture.pdf.

[40] G. C. M. Moura, J. Heidemann, M. Müller, R. de O. Schmidt, and M. Davids, "When the Dike Breaks: Dissecting DNS Defenses During DDoS," in *2018 Internet Measurement Conference*, IMC '18, 2018, pp. 8–21. DOI: 10.1145/3278532.3278534.

[41] S. Hao, Y. Zhang, H. Wang, and A. Stavrou, "End-Users Get Maneuvered: Empirical Analysis of Redirection Hijacking in Content Delivery Networks," in *27th USENIX Security Symposium*, USENIX Security '18, 2018, pp. 1129–1145.

[42] A. Kashaf, V. Sekar, and Y. Agarwal, "Analyzing Third Party Service Dependencies in Modern Web Services: Have We Learned from the Mirai-Dyn Incident?" In *2020 Internet Measurement Conference*, IMC '20, 2020, pp. 634–647. DOI: 10.1145/3419394.3423664.

[43] Y. Dimova, G. Acar, L. Olejnik, W. Joosen, and T. Van Goethem, "The CNAME of the Game: Large-scale Analysis of DNS-based Tracking Evasion," *Proceedings on Privacy Enhancing Technologies*, vol. 2021, no. 3, pp. 394–412, Apr. 2021. DOI: 10.2478/popets-2021-0053.

[44] J. St Sauver. "What is a Bailiwick?" Farsight Security. (Mar. 21, 2017), [Online]. Available: https://www.farsightsecurity.com/blog/txt-record/what-is-a-bailiwick-20170321/.

[45] S. Coulondre and B. Sitney, "A Stable and Open Method for Ranking Domains," Profound Networks, Tech. Rep., 2019. [Online]. Available: https://www.profound.net/pages/resources/DomainRank_Whitepaper.pdf.

[46] J. Lehmann *et al.*, "DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia," *Semantic Web*, vol. 6, no. 2, pp. 167–195, 2015. DOI: 10.3233/SW-140134.

[47] "Public Suffix List," Mozilla Foundation. (2022), [Online]. Available: https://publicsuffix.org/.

[48] W. B. de Vries, Q. Scheitle, M. Müller, W. Toorop, R. Dolmans, and R. van Rijswijk-Deij, "A First Look at QNAME Minimization in the Domain Name System," in *20th International Conference on Passive and Active Measurement*, PAM '19, 2019, pp. 147–160. DOI: 10.1007/978-3-030-15986-3_10.

[49] J. Magnusson, M. Müller, A. Brunstrom, and T. Pulls, "A Second Look at DNS QNAME Minimization," in *24th International Conference on Passive and Active Measurement*, PAM '23, 2023, pp. 496–521. DOI: 10.1007/978-3-031-28486-1_21.

[50] C. Lu *et al.*, "An End-to-End, Large-Scale Measurement of DNS-over-Encryption: How Far Have We Come?" In *2019 Internet Measurement Conference*, IMC '19, 2019, pp. 22–35. DOI: 10.1145/3355369.3355580.

[51] I. Sanchez-Rola, D. Balzarotti, C. Kruegel, G. Vigna, and I. Santos, "Dirty Clicks: A Study of the Usability and Security Implications of Click-Related Behaviors on the Web," in *The Web Conference 2020*, WWW '20, 2020, pp. 395–406. DOI: 10.1145/3366423.3380124.

[52] G. C. M. Moura, J. Heidemann, R. d. O. Schmidt, and W. Hardaker, "Cache Me If You Can: Effects of DNS Time-to-Live," in *2019 Internet Measurement Conference*, IMC '19, 2019, pp. 101–115. DOI: 10.1145/3355369.3355568.

[53] M. Korczyński, Y. Nosyk, Q. Lone, M. Skwarek, B. Jonglez, and A. Duda, "Don't Forget to Lock the Front Door! Inferring the Deployment of Source Address Validation of Inbound Traffic," in *21st International Conference on Passive and Active Measurement*, 2020, pp. 107–121. DOI: 10.1007/978-3-030-44081-7_7.

[54] Q. Lone, M. Korczyński, M. van Eeten, and C. H. Gañán, "SAVing the Internet: Explaining the Adoption of Source Address Validation by Internet Service Providers," in *20th Annual Workshop on the Economics of Information Security*, WEIS '20, 2020. [Online]. Available: https://weis2020.econinfosec.org/wp-content/uploads/sites/8/2020/06/weis20-final31.pdf.

[55] Q. Lone, A. Frik, M. Luckie, M. Korczyński, M. van Eeten, and C. Gañán, "Deployment of Source Address Validation by Network Operators: A Randomized Control Trial," in *2022 IEEE Symposium on Security and Privacy*, SP '22, 2022, pp. 703–720. DOI: 10.1109/SP46214.2022.00041.

[56] L. A. Adamic and B. A. Huberman, "Zipf's Law and the Internet," *Glottometrics*, vol. 3, pp. 143–150, 2002. [Online]. Available: https://glottometrics.iqla.org/wp-content/uploads/2021/06/g3zeit.pdf.

[57] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-Law Distributions in Empirical Data," *SIAM Review*, vol. 51, no. 4, pp. 661–703, Nov. 2009. DOI: 10.1137/070710111. [Online]. Available: https://doi.org/10.1137/070710111.

[58] Y. Chen, M. Antonakakis, R. Perdisci, Y. Nadji, D. Dagon, and W. Lee, "DNS Noise: Measuring the Pervasiveness of Disposable Domains in Modern DNS Traffic," in *44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, 2014, pp. 598–609. DOI: 10.1109/DSN.2014.61.

[59] V. Le Pochat *et al.*, "A Practical Approach for Taking Down Avalanche Botnets Under Real-World Constraints," in *27th Annual Network and Distributed System Security Symposium*, NDSS '20, 2020. DOI: 10.14722/ndss.2020.24161.

[60] T. Wicinski, *DNS Privacy Considerations*, RFC 9076, Jul. 2021. DOI: 10.17487/RFC9076. [Online]. Available: https://www.rfc-editor.org/info/rfc9076.

[61] B. Imana, A. Korolova, and J. Heidemann, "Enumerating Privacy Leaks in DNS Data Collected above the Recursive," in *2018 NDSS DNS Privacy Workshop*, 2018. [Online]. Available: https://www.isi.edu/~johnh/PAPERS/Imana18a.pdf.

[62] B. Imana, A. Korolova, and J. Heidemann, "Institutional Privacy Risks in Sharing DNS Data," in *2021 Applied Networking Research Workshop*, ANRW '21, 2021, pp. 69–75. DOI: 10.1145/3472305.3472324.

[63] J. M. Spring and C. L. Huth, "The Impact of Passive DNS Collection on End-user Privacy," in *2nd Workshop on Securing and Trusting Internet Names*, SATIN '12, 2012. [Online]. Available: https://resources.sei.cmu.edu/asset_files/WhitePaper/2012_019_001_57023.pdf.


## Appendix A
## Ethics

Processing DNS query data may come with a privacy risk, as it could be used to identify individual users and the domains that they access [60–63]. We only have access to count data aggregated from post-recursor requests across multiple sensors that mix cache misses originating from many users and institutions. The data does not contain any explicit fields with likely personally identifiable information (PII), such as the client IP address. We already identify domain names (likely) containing IP addresses as a separate class, and would discard them for any publicly available rankings. We further aggregate the raw counts from the passive DNS feed per domain, and for a final public ranking would select a reasonable cutoff (on query count or ranking length) to retain only sufficiently popular domains and avoid publicizing private internal domains. We, therefore, consider that user privacy is preserved.

## Appendix B
## Reproducibility

To enable reproducing our work, we make the ranking generation code, ranking files, and analysis scripts and results available. These resources are available from https://domain-ranking-design-decisions.distrinet-research.be.

## Appendix C
## TTL incorporation: similarity

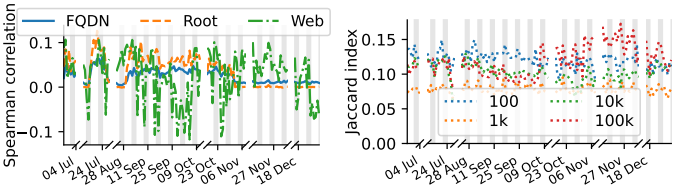


Fig. 13. Unweighted and TTL-weighted rankings are very dissimilar, as measured using the Spearman correlation, and for buckets of the web ranking using the Jaccard index.