Hosting Industry Centralization and Consolidation

Luciano Zembruzki*, Raffaele Sommese†, Lisandro Zambenedetti Granville*,
Arthur Selle Jacobs*, Mattijs Jonker†, Giovane C. M. Moura‡

*Institute of Informatics (INF) - Federal University of Rio Grande do Sul (UFRGS)
Porto Alegre, Brazil
{lzembruzki, asjacobs, granville}@inf.ufrgs.br

†Design and Analysis of Communication Systems (DACS) - University of Twente
Enschede, The Netherlands
{r.sommese, m.jonker}@utwente.nl

‡SIDN
SIDN Labs
Arnhem, The Netherlands
giovane.moura@sidn.nl

NOMS 2022-2022 IEEE/IFIP Network Operations and Management Symposium | 978-1-6654-0601-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/NOMS54207.2022.9789881

***Abstract*—There have been growing concerns about the concentration and centralization of Internet infrastructure. In this work, we scrutinize the hosting industry on the Internet by using active measurements, covering 19 Top-Level Domains (TLDs). We show how the market is heavily concentrated: 1/3 of the domains are hosted by only 5 hosting providers, all US-based companies. For the country-code TLDs (ccTLDs), however, hosting is primarily done by local, national hosting providers and not by the large American cloud and content providers. We show how shared languages (and borders) shape the hosting market — German hosting companies have a notable presence in Austrian and Swiss markets, given they all share German as official language. While hosting concentration has been relatively high and stable over the past four years, we see that American hosting companies have been continuously increasing their presence in the market related to high traffic, popular domains within ccTLDs — except for Russia, notably.**

I. Introduction

Internet centralization and consolidation refers to the concentration of, *e.g.*, user base, infrastructure, and network traffic in the hands of few, yet large Internet market players. Centralization raises concerns among diverse entities and individuals such as academics, operators, not-for-profit organizations, standardization bodies, the European Commission, the US Department of Justice, policy makers, and civil society [1]–[13]. Internet consolidation usually implies concentration of power too [11] [12] – including political power, as seen in the last US presidential election [14] [15]. From a technical perspective, concentration in the hands of few market players can also create large, single points of failure. Notable examples of this include Denial-of-Service (DoS) attacks against the DNS providers Dyn and AWS, which affected multiple popular websites and services [16] [17], including Twitter, Netflix, and Spotify. Consolidation can also lead to a large attack surface in which "bad decisions or poorly made trade-offs implemented by a company can quickly scale to hundreds of millions of users" [18], as in the case of the Sunburst cyber espionage campaign where Microsoft, a large cloud provider, was hacked.

Quantifying Internet centralization is a challenging task that encompasses questions such as: what exactly to measure and where? Recent studies have either measured *traffic* as a way to assess centralization (*e.g.*, Moura *et al.* [7] found that 1/3 of the DNS traffic towards the country-code Top-Level Domains (ccTLDs) of the Netherlands and New Zealand originated from 5 large, American companies) or quantified centralization in terms of *infrastructure* concentration (Kashaf *et al.* [8] and Allman [19] studied authoritative DNS [20] service infrastructure concentration). Centralization can also be quantified in terms of *user base* or *market share*. Facebook, for example, has 2.9B monthly active users (Jul. 2021 [21]). In general, all these consolidation assessments complement one another and highlight different aspects of Internet centralization.

In contrast with the aforementioned efforts, in this work we focus on analyzing consolidation in the *Web hosting* industry [22], which is a market segment dedicated to hosting websites and services that use HTTP/S [23] as application protocol. In 2020, this market segment was valued at US$ 75B [24]. Yet, the hosting industry has been mostly neglected in state-of-the-art Internet centralization studies. To close this gap, we develop a methodology that employs DNS zone files and active measurements to map *all* domains from 19 Top-Level domains (TLDs), including sixteen ccTLDs and three generic TLDs (gTLDs), to their respective Web hosting companies. In addition, we analyze current and historical measurement data, spanning over five years, from 2017 to 2021, to observe the development (and increase) of Web hosting centralization.

The contribution of this paper is threefold. First, we reveal that *Web hosting* is heavily concentrated by finding that: *(i)* more than 1/3 of 150 million analyzed domains from 19 TLDs are hosted by five large US hosting providers; *(ii)* as a result of *one policy* change by CloudFlare (the largest 2020 hosting provider in our dataset), ~17 M domains were moved to Google Cloud, increasing the concentration of domain names and turning Google into the largest provider in 2021, with 18% of all domains; and *(iii)* although the centralization percentage varies by each TLD, most of them concentrate at least 40% of domains in five hosting providers. In addition, this centralization has been increasing over the past five years. As second contribution, we show that geographical proximity and shared language ties influence the hosting industry; although European ccTLDs have a solid local hosting industry, German-speaking countries in Europe (Austria, Germany, Switzerland,

and Liechtenstein) use each other's hosting provider industry. Canada .ca, in turn, mainly relies on US-based hosting companies. Our third and final contribution is to show that US-based hosting companies have been increasing their market share on popular domains for most TLDs, which poses a challenge for European Union (EU) digital sovereignty goals [13]. The exception is Russia's ccTLDs (.ru and .*рф*), for which most popular domains are hosted by local companies.

The remainder of this paper is organized as follows. In Section II, we present related work on measurements of different aspects of Internet centralization and consolidation, in addition of positioning our work with regard to the state-of-the-art. In Section III, we describe our analysis methodology, limitations and data sets obtained from the OpenINTEL project. Then, in Sections IV and V, we present results and show the market share and top DNS providers by year, TLD, and geographical location. Finally, in Section VI, we draw conclusions and outline future work.

II. Related Work

The centralization and consolidation of the Internet is a known issue. Several efforts have been made to quantify to which extent this is affecting the global hosting ecosystem. However, as maintainers and DNS providers infrequently disclose DNS zone files and rarely share their internal configurations (aggregated or otherwise), measuring concentration is daunting. To that end, authors have taken different approaches to circumvent the general lack of transparency, often relying on active measurement techniques. In addition, identifying industry trends over time requires long-term data collection efforts, which come with a number of infrastructural challenges for researchers. In this section, we first briefly survey the state-of-the-art on DNS measurements, and then visit past efforts to quantify centralization and consolidation of hosting infrastructures.

Bates *et al.* [25] proposed a method to compute to which extent the global DNS infrastructure has maintained its distributed resilience in the face of the development of cloud-based DNS infrastructures. The authors examined the concentration and diversity of DNS over time while considering only .com, .net, and .org domain names on the list of Alexa Top 1k domains. The authors argued that they limited their investigation to the TLDs .com, .net, and .org because these are among the oldest and represent a broad part of the Internet. The authors, however, acknowledge that their findings may vary if additional TLDs, such as .ru and .cn, would be be taken into account.

Kashaf *et al.* [8] studied the presence and impact of third-party dependencies in three infrastructure services: DNS, CDNs, and certificate revocation checks by CAs. They analyzed both direct and indirect dependencies. The authors found that 89% of Alexa's Top 100 k websites rely crucially on third-party DNS, CDN, or CA providers, implying that these websites will experience service disruption if these third-party providers fail. The study also shows that third-party service use is concentrated, with the leading Top 3 providers of CDN, DNS, or CA services affecting between 50% to 70% of the Top 100 k websites. It is worth mentioning that, in their analysis of DNS dependencies, the authors rely solely on NS record labels (*e.g.,* ns1.example.com) to measure concentration. Since multiple name servers can be hosted on a single IP address, the authors' analysis can mask possible cases of centralization.

Moura *et al.* [7] Using an investigation of DNS traffic acquired at a DNS root server and two ccTLDs (one in Europe and one in Oceania), discovered indicators of concentration: over 30% of all requests to both ccTLDs were made via five major cloud providers (Google, Amazon, Microsoft, Facebook, and Cloudflare). Unlike earlier initiatives, the authors underlined one benefit of centralization: whenever the cloud provider improves its infrastructure, for example, in terms of security features like QNAME reduction, a large number of clients benefit immediately.

The aforementioned research efforts inspect and measure different aspects of infrastructure to quantify the centralization and consolidation of DNS. However, they do not consider centralization of the hosting industry, which plays a key role in the Internet ecosystem as well. Although, from a technical perspective, analyzing the DNS infrastructures may provide a general view of the problem space, it is by analyzing the concentration of domains per hosting provider that one can shed light onto the often overlooked issue of hosting industry centralization.

Tajalizadehkhoob *et al.* [26] presented a method to capture the hosting market's complexity. They identify and classify hosting providers by using a technique that combines: (i) passive DNS data to locate hosting infrastructure, and (ii) WHOIS data to address that infrastructure. In doing so, they are able to quantify providers and look at geographic distribution. The authors further exploit features to characterize the hosting provider landscape and reduce hosting market complexity and heterogeneity by using cluster analysis. They discovered 45,434 hosting providers dispersed over 1,517 IP addresses on average. Although the hosting services are commoditized – and thus susceptible to economies of scale, as witnessed in the cloud services sector – the authors surprisingly found little consolidation in the market.

Ager *et al.* [27] proposed an approach to infer and categorize hosting infrastructures. The authors rely on data available to end-users that request hostnames via the DNS and further use information such as IP addresses, prefixes, and AS numbers to create a mapping of hosting infrastructures. Their findings demonstrate that qualitative observations can be made for the establishment of hosting infrastructure and content replication. One of the study's primary findings is that a considerable portion of content is given solely by hosting infrastructures such as Google, or in geographic locations such as China, the latter of which suggests that spoken language could be a factor.

To the best of our knowledge, no prior work has provided a comprehensive overview nor comparison of centralization and consolidation by ccTLDs and gTLDs. Furthermore, the relation between spoken language, hosting provider location, and market division, as we do in this paper, was not previously investigated.

III. Methodological Overview

In this section, we detail the methodology and data sources that we used to analyze both the current landscape of the hosting industry as well as its historical evolution. We also motivate assumptions made and discuss methodological limitations. Finally, we describe the concrete data set used for this work.

A. Methodology

We rely on large-scale DNS measurement data for analysis. Data are provided by the OpenINTEL project [28], which measures roughly 65% of the global DNS namespace (second-level domains) on a daily basis. OpenINTEL primarily takes a set of zone files as seed and enumerates, through active querying, the resource records (RRs) configured under domain names. OpenINTEL covers: generic top-level domains (gTLDs) such as `.com`, `.net`, and `.org`; country-code zones (ccTLDs) for several continents (*e.g.*, `.ca`); and about 1,200 new gTLDs such as `.tokyo`[1]. In addition to zone files, a number of other sources of domains are measured (*e.g.*, the Alexa Top 1 M).

We use other data sources to enrich or complement the OpenINTEL DNS data. For example, we use the Public Suffix List [29], a community-driven initiative of Mozilla, to demarcate top-level domains (*e.g.*, `.co.uk`). We use daily CAIDA prefix-to-AS data [30] to map IP addresses (*i.e.,* `A` and `AAAA` records) to the announcing AS number. We leverage monthly AS-to-organization data [31] to map AS numbers to organizations and countries. Finally, we use an open countries and languages data set to map country codes to (one or more) official/predominant languages [32]. It is important to note that this mapping is different from IP geolocation [33], [34], which maps IP addresses to geographical locations. As we will show later on, we are instead interested in the country to which the announcing AS can be associated.

B. Assumptions and Limitations

Our methodology involves a number of assumptions and also a few limitations:

Web hosting. For simplicity, we assume that the presence of an `A` or `AAAA` record for a domain name (or its `www.` label) signals Web hosting. This assumption will break for domain names strictly used for non Web purposes, but we expect this to have limited impact on our results. As we will show in section V, the vast majority of domain names configure the `www` label, which is a reasonable indicator of Web hosting intent. We also checked this assumption in May 2021 HTTP crawl data of `.nl` and found that: 90.2% of the domains with `A` records have an active website (5.1 M out of 5.6 M, in Table II) and 91.1% with `AAAA` records have an active IPv6 website (2.58 M out of 2.83 M, in Table IIb). Using Internet-wide port-scan data from Rapid7 Sonar, we analyzed the number of domains that have both an IPv4 address and an open HTTP(s) port. The results from this preliminary analysis show that this applies to 90.7% of domains considered, on average per TLD.

[1]Zone file access typically requires specific agreements with registries that operate under varying regulations, which means that covering every possible top-level domain, especially some ccTLDs, is unfeasible.

Year	TLDs	Domains	IPv4	IPv6	ASes(v4)	ASes(v6)
2017	11	161.4 M	149.1 M	8.7 M	35.7 k	3.4 k
2018	13	175.0 M	157.9 M	11.9 M	38.2 k	3.8 k
2019	14	180.6 M	166.3 M	15.0 M	38.8 k	4.0 k
2020	17	188.5 M	173.0 M	17.7 M	39.6 k	4.3 k
2021	19	200.7 M	182.6 M	21.8 M	40.4 k	4.6 k

Table I: Aggregated data sets, measured each year on May 5.

CNAMEs. For cases in which a domain name uses a `CNAME` rather than `A` or `AAAA` records – as typically used with domains hosted in content-delivery networks (CDNs) – we fully expand the `CNAME` chain until it is terminated by `A` or `AAAA` records. We then map the domain name in question to the AS number(s) at the end of the chain.

Parked domains. In case a domain name leads to non-responsive, parking, or redirecting website, we still take into account the infrastructure where the website is supposedly hosted – classifying this typology of websites is left as future work.

Subdomains. Given that OpenINTEL primarily knows zone files and the second-level domain names therein, we mostly infer hosting at the domain apex and for the third-level `www.` label (except for `CNAME` records, where arbitrary labels can be encountered and are expanded).

Single vantage-point. OpenINTEL collects measurements from a single vantage point in the Netherlands, which can introduce a bias towards "nearby" `A` and `AAAA` records when DNS-based load balancing is used. This bias can, in theory, increase the concentration towards European IP address space. However, even if hosting companies operate address space in several countries, the AS number of their headquarter country remains the same.

C. Data set

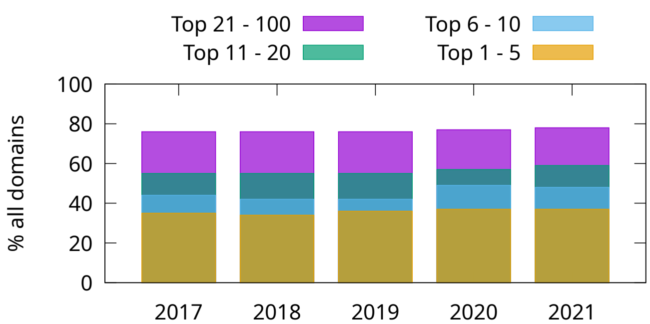
To measure the evolution of web hosting, we use historical data from OpenINTEL – one day per year (on May 5th), covering the last five years (2017 – 2021). Table I summarizes the resulting data set, for all TLDs under consideration combined. We measure 11 – 19 TLDs, covering 161 M – 201 M domains, hosted by 35 k – 40 k IPv4 ASes and fewer than 5 k IPv6 ASes.

IV. Web hosting concentration

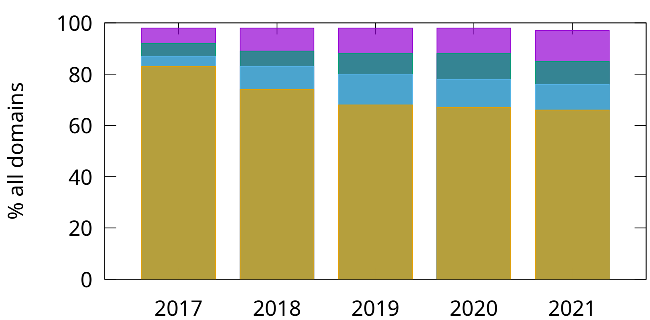
In this section, we focus on determining the largest hosting ASes in our data sets. Table I summarized the aggregate data. Our goal is to determine if there is a significant concentration in the domain namespace. To do that, we compute the percentage of the Top 5, 10, 20, and 100 hosting ASes per year.

A. Concentration on Top hosting providers

Figure 1 shows the result of our analysis. The x-axis contains the considered year (2017 through 2021) and the y-axis shows the percentage of domains hosted in Top hosting ASes. For IPv4 (Fig. 1a), the Top 5 hosting ASes host more than one-third of the 120 M – 150 M domains (depending on the year) that had IPv4 records. We observe that this concentration has been slightly growing over the last five years. Note that our data contains 35 k – 40 k hosting ASes for the evaluated years (Table I), which shows how significant the concentration in top hosting ASes is. For IPv6 (Figure 1b), the Top 5 hosting ASes account for 64–71% of the 10 M – 21 M domain names with IPv6 records. This reinforces the notion that big tech companies have committed to fully implement IPv6 to encourage smaller companies to make the transition to IPv6 [35] and that, despite the still low adoption of IPv6, these companies remain at the forefront until today.



(a) IPv4



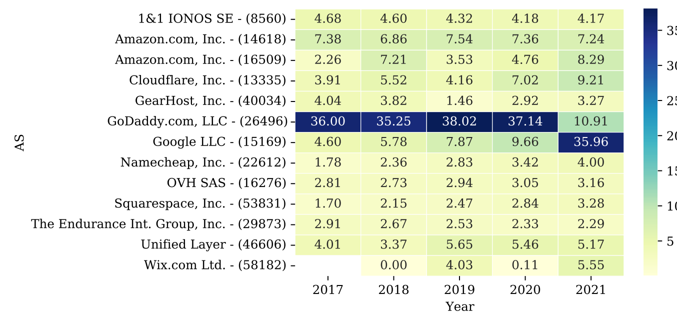
(b) IPv6

Figure 1: Concentration in Top hosting ASes

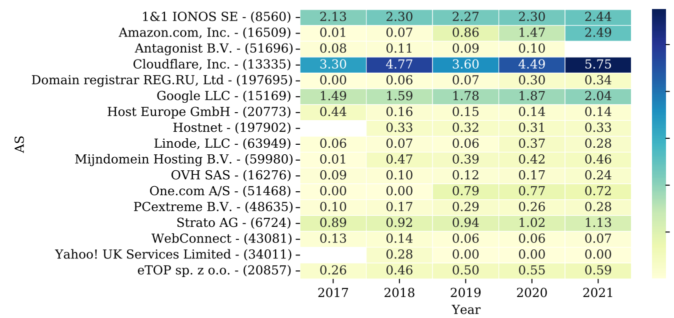
In general we observe that, for IPv4, the concentration has been relatively stable since 2017, and the percentage of the domains hosted by the Top 10 ASes has slightly increased. However, for IPv6, despite being far more concentrated, there appears to be a reverse trend, most likely related to the adoption of IPv6 by more hosting providers (3.4 k to 4.6 k ASes between 2017 and 2021–Table I).

B. Ranking the Top hosting providers

Along the years, the hosting market can present movement of domains, migrating from one hosting provider to another. In order to observe how the market has been dominated by the Top providers, we take into account both the total number of domains hosted by each provider per year, as well as the percentage of the market occupied by each provider.



(a) IPv4



(b) IPv6

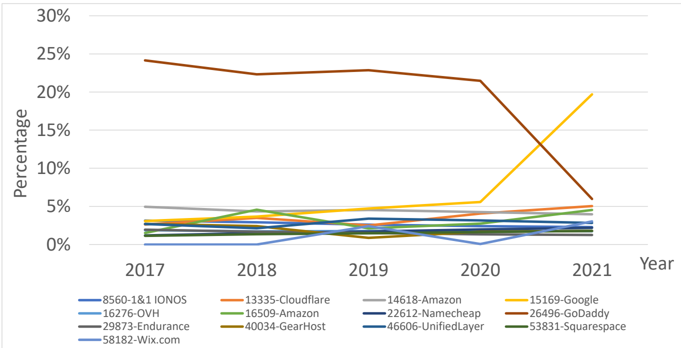
Figure 2: Top Hosting ASes evolution

Google's dominance of the hosting market occurred only in 2021. Figures 2 and 3 show the evolution of the Top hosting ASes observed from 2017 to 2021, in terms of *millions of hosted domain names* and *percentage of hosted domains*, respectively. As can be observed, most ASes grow in terms of hosted domains and ratio. The major surprise is the swap of positions of Google (AS15169) and GoDaddy (AS26496). In 2020, GoDaddy hosted 37.14 million domain names, but only 10.91 M in 2021. In the same period, Google showed growth: from 9.66 M to 35.96 M domains.

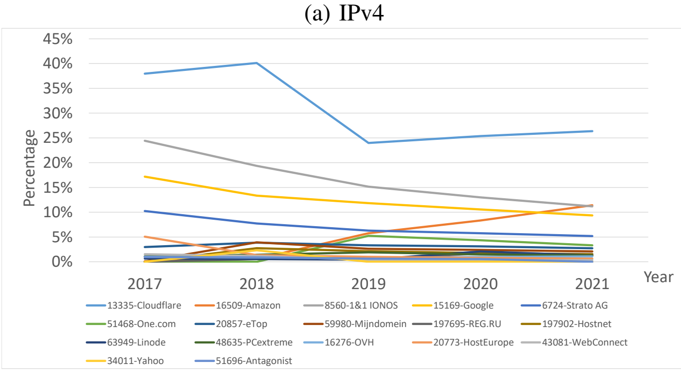
To investigate the reasons for this inversion, we compared both 2020 and 2021 results and found that 17 M domains hosted by GoDaddy in 2020 migrated to Google Cloud. We investigated these domains and found that they are all hosted on the same IPv4 address. This IPv4 address hosts a simple template page for a parked domain; therefore, we marked all the domains as *parked* (parked domains are a part of the domain industry in which domains are parked to reserve, resell, or run ads for profit [36]).

This case exemplifies the amount of power that lies in the hands of a large hosting provider as it can shift the entire domain distribution with a single policy decision.

For IPV6 (Figure 3b), this behavior is not repeated. Cloudflare, which is the provider with the highest number of domains, went from 40% to 26% of domains. We note that despite the providers having shown an increase in the number of hosted domains (figure 2b), they showed a decrease in the percentage of concentration in relation to the total number of domains (figure 3b) from 2017 to 2021.



(a) IPv4



(b) IPv6

Figure 3: Percentage of domains hosted by Top Hosting ASes per Year

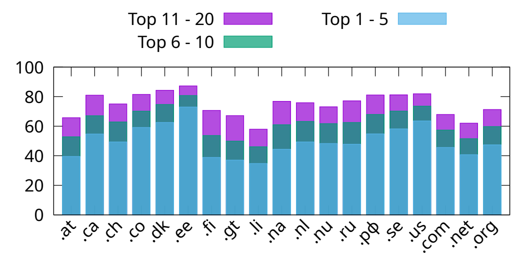
V. Concentration per TLD

Now we move to evaluate the domain concentration per TLD to determine if distinct TLDs have similar concentration patterns. We start by splitting our data set from Table I into TLDs and next evaluate each. Then we compute the concentration of the Top 5, 10, and 20 hosting ASes, similar to Section IV.

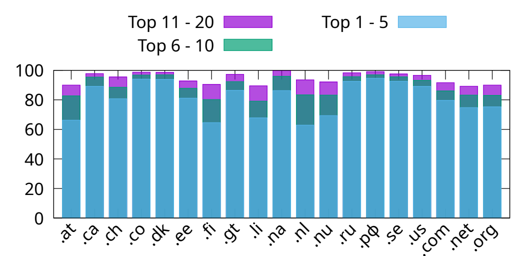
Table IIa shows the data set split per TLD, for the 2021 IPv4 measurements. Table IIb shows IPv6. The tables provide a detailed overview of characteristics of the domains present in each of the TLDs. The row `Domains` shows the number of domains in the zone file. The `Responsive` row represents the domains that respond to *SOA* query (that means well configured). We see that for all zones, the majority of domains have `A` records associated with them (`with hosting` column), either on the `apex` or under `www`. Interestingly, some zones have more `A` records on the *apex* and others more on `www`. For example, for `.at`, 99% of domains with hosting have `A` record under `www`, but only 96.8% on the `apex`. For `.ru`, these values are 99.7% and 96.9%, respectively. The `apex` row is the subset of `with hosting` that have hosting on the `apex`. Row `www` represents the subset that has hosting on `www`. The row with the symbol "∩" represents the domains that have the same `A` records for both `apex` and `www`. The symbol ! = means that the domains have different `A` records for the `apex` and `www`. Finally, the last three rows represent the domains that are under a single, multiple, or unique ASes, respectively.

The `A` records defined at the `apex` and `www` labels may *differ*. For example, the domain `klm.de` has `171.21.122.79` (AS28806) as an `A` record on the apex. At the same time, `www.klm.de` has a `CNAME` [37] that points to `23.45.75.2`, which is a multi-homed address [38], announced by two ASes: AS20940 and AS16625, both belonging to Akamai, a large CDN provider. On a browser, using either name leads the user to the same website. As we show in Table II, the majority of domains have the same `A` records in both `apex` and `www` (row "∩"), but from 3.7% to 10.7% have different (row ! =), depending on the ccTLD. Each domain name may use multiple ASes to announce its `A` records to achieve load-balancing and resilience against failure. However, fewer than 5% of all domains use multiple ASes for hosting for the evaluated TLDs (Table II).

We study hosting concentration per TLD to investigate if the concentration is uniform or dissimilar. We use per-TLD data for 2021 and then compute the concentration of the top 5, 10, and 20 hosting ASes.



(a) IPv4



(b) IPv6

Figure 4: Hosting concentration 2021 per TLD

A. *Web hosting concentration*

Figure 4 shows for each TLD its respective Top hosting ASes the percentage of domain names. A significant concentration is apparent for all TLDs: the Top 5 hosting ASes are responsible for 37–73% of the domains, depending on the TLD, for IPv4 (Figure 4a). Estonia's `.ee` has the highest concentration level: 73%. For IPv6, the concentration is even

	ccTLDs																gTLDs		
	.at	**.ca**	**.ch**	**.co**	**.dk**	**.ee**	**.fi**	**.gt**	**.li**	**.na**	**.nl**	**.nu**	**.ru**	**.pϕ**	**.se**	**.us**	**.com**	**.net**	**.org**
Domains	1.36 M	3.03 M	2.25 M	2.86 M	1.34 M	139.20 k	510.94 k	19.90 k	60.91 k	5.21 k	6.17 M	252.21 k	4.73 M	630.71 k	1.49 M	1.65 M	150.88 M	13.05 M	10.17 M
Responsive	1.31 M	2.84 M	2.07 M	2.68 M	1.22 M	135.52 k	476.80 k	17.39 k	55.11 k	4.84 k	5.84 M	233.21 k	4.47 M	579.72 k	1.41 M	1.49 M	142.21 M	11.99 M	9.58 M
with hosting	1.24 M	2.78 M	1.93 M	2.62 M	1.11 M	130.56 k	442.48 k	16.03 k	50.54 k	3.88 k	5.67 M	221.43 k	4.33 M	557.21 k	1.34 M	1.49 M	138.11 M	11.27 M	9.26 M
apex	1.20 M	2.76 M	1.91 M	2.60 M	1.09 M	129.21 k	431.50 k	15.71 k	49.06 k	3.43 k	5.59 M	217.47 k	4.32 M	556.08 k	1.32 M	1.47 M	136.38 M	11.07 M	9.16 M
`www.`	1.23 M	2.73 M	1.91 M	2.56 M	1.08 M	128.59 k	435.52 k	15.59 k	48.75 k	3.82 k	5.62 M	219 k	4.20 M	536.30 k	1.33 M	1.45 M	135.87 M	11.03 M	9.13 M
∩	1.11 M	2.46 M	1.69 M	2.26 M	1.01 M	12.21 k	386.72 k	13.92 k	43.74 k	3.10 k	5.34 M	206.88 k	3.81 M	511.60 k	1.23 M	1.32 M	118.31 M	9.79 M	8.06 M
! =	77.05 k	258.14 k	187.12 k	282.24 k	53.57 k	5.10 k	37.83 k	1.35 k	3.53 k	267	213.69 k	8.16 k	378.45 k	23.57 k	76.17 k	121.47 k	15.82 M	1.03 M	970.85 k
Single	1.21 M	2.68 M	1.85 M	2.53 M	1.08 M	127.61 k	427.84 k	15.27 k	48.97 k	3.73 k	5.58 M	216.28 k	4.26 M	547.62 k	1.3 M	1.44 M	131.37 M	10.81 M	8.86 M
Multiple	32.97 k	103.53 k	74.53 k	91.33 k	25.04 k	2.95 k	14.63 k	761	1.56 k	153	90.35 k	5.14 k	72.78 k	9.58 k	42.82 k	45.75 k	6.74 M	454.53 k	395.45 k
Unique ASes	4.22 k	4.74 k	4.59 k	6.70 k	3.05 k	1.53 k	2.35 k	631	1.80 k	255	4.77 k	1.89 k	7.93 k	2.72 k	3.69 k	7.64 k	35.15 k	23.41 k	19.51 k

(a) IPv4

	ccTLDs																gTLDs		
	.at	**.ca**	**.ch**	**.co**	**.dk**	**.ee**	**.fi**	**.gt**	**.li**	**.na**	**.nl**	**.nu**	**.ru**	**.se**	**.us**	**.pϕ**	**.com**	**.net**	**.org**
Domains	1.36M	3.03M	2.25M	2.86M	1.34M	139k	510k	19.90k	60.91k	5.21k	6.17M	252k	4.73M	630k	1.49M	1.65M	150.88M	13.05M	10.17M
Responsive	1.31 M	2.84 M	2.07 M	2.68 M	1.22 M	135.52 k	476.80 k	17.39 k	55.11 k	4.84 k	5.84 M	233.21 k	4.47 M	579.72 k	1.41 M	1.49 M	142.21 M	11.99 M	9.58 M
with hosting	197.36k	150.47k	515.70k	336.12k	268.08k	7.97k	54.25k	1.63k	10.80k	247	2.83M	62.37k	895.99k	101.73k	366.84k	107.33k	13.72M	1.12M	1M
apex	190.95k	128.13k	491.77k	309.57k	258.74k	7.21k	48.12k	1.39k	10.29k	206	2.76M	60.78k	887.69k	101.20k	355.05k	96.71k	12.03M	995.07k	874.55k
www.	191.53k	145.81k	501.81k	321.67k	255.91k	7.68k	52.63k	1.54k	10.28k	237	2.77M	61.25k	857.07k	100.01k	362.03k	95.75k	13.19M	1.06M	968k
∩	183.59k	108.94k	474.58k	284.07k	245.32k	6.81k	45.63k	1.25k	9.64k	190	2.69M	59.46k	847k	99.39k	348.89k	73.99k	10.75M	879.61k	767.34k
! =	1.53k	14.54k	3.30k	11.04k	1.25k	112	862	37	131	6	8.17k	198	1.20k	95	1.35k	11.13k	753.03k	58.06k	72.56k
Single	196.66k	148.54k	513.89k	333.53k	267.03k	7.89k	53.43k	1.62k	10.72k	245	2.82M	62.14k	895.01k	101.58k	365.75k	105.94k	13.59M	1.11M	995.60k
Multiple	696	1.92k	1.81k	2.59k	1.05k	81	821	10	81	2	9.98k	236	986	149	1.08k	1.38k	131.11k	7.90k	7.87k
Unique ASes	591	405	678	604	366	178	314	53	266	23	669	330	651	174	522	594	3.4k	2.73k	2.03k

(b) IPv6

Table II: Data sets 2021-05-05

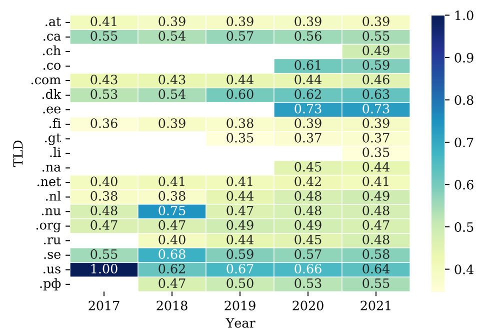
more significant: from 62–94%. We also observe that the difference between the top 20 hosting ASes and the Top 5 is not that significant, which further demonstrates how concentrated the market is. We also see no significant differences between ccTLD and gTLD concentration.

Changes over time: Figure 5a shows the variations in concentration of hosting for the Top 5 hosting ASes, per TLD in terms of the percentage of SLDs hosted. We observe that this concentration has been rather stable for most TLDs for IPv4 (Figure 5a). This is different from the IPv6 case (Figure 5b), which appears to have reduced the concentration of domains in the top 5 providers in some TLDs. For IPv4 some ccTLDs have experienced an increasing concentration: Denmark's `.dk` has gone from 53% to 63% of the domains being hosted by five providers in the last five years. `.nl`, similarly, went from 38% to 49%. For IPv6, we notice that part of the TLDs reduce the ratio of domains hosted by Top 5 ASes. As we can see, Austria's `.at` went from 89% to 66%. As well, `.net` went from 98% to 75%.

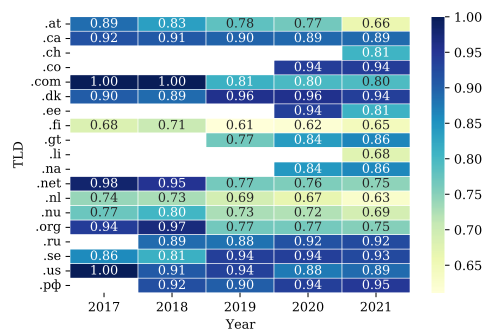
B. Hosting Providers Location

Next, we investigate the locations of hosting companies. For each AS used for hosting, we lookup the company's origin and then map the domain names they host to this country. Note from earlier discussion that this is not the same as geolocating the hosting IP addresses (section III-A).

For each TLD we count the number of domains per country and classify each country into four categories, in the following priority order: *local* (if the country is the same as the TLD), US-based (if the the country is the US, given most of large cloud companies are from the US), same official language (for example, `.ca` and `.mt` also use English as one their official languages), and rest (the remaining domains). Note that each



(a) IPv4

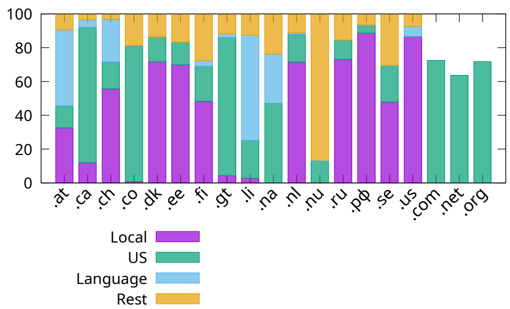


(b) IPv6

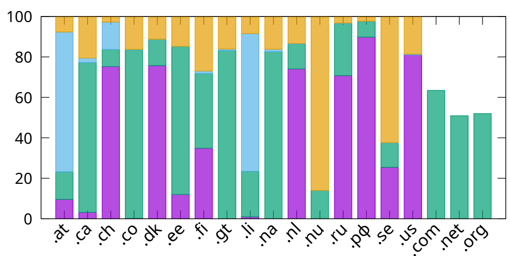
Figure 5: Ratio of domains hosted by Top 5 ASes per TLD and Year

AS/country gets only a label, even if multiple may apply. For example, AS15169 is both local and US hosting for `.us`, but we consider it only as local for `.us`, given that local takes priority in our classification.

Figure 6 shows the AS hosting country classification per TLD. Contrary to what we would expect, the US hosting industry is not dominant for most TLDs. The US hosting companies are popular in `.ca` (same language), `.co` (which is very popular outside Colombia), Guatemala's `.gt` (possible due to geographical closeness to the US and services available in Spanish, which is Guatemala's official language) and Namibia's `.na` (which has English as the official language).



(a) IPv4



(b) IPv6

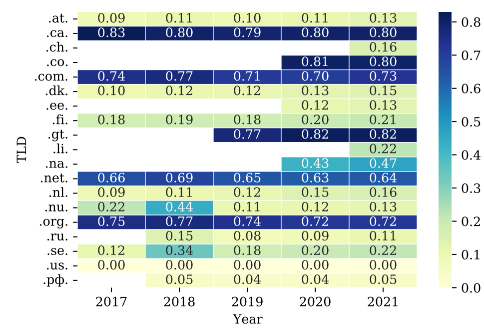
Figure 6: Hosting concentration per type (2021)

Secondly, we see a solid local hosting industry in most European countries and Russia. The respective local companies are dominant in their local market. Interestingly, we see German-speaking countries in Europe (Austria's `.at`, Switzerland's `.ch`, and Liechtenstein' `.li`) that use each other's countries hosting providers – mostly German companies. Figure 6a shows the IPv4 breakdown. Austria hosts 42% of its domains in German hosting companies, while Switzerland hosts 20%. Liechtenstein has 60% of its domains hosted by German or Swiss companies (30% on each). The IPv6 hosting situation differs: we see a reduced presence in the local companies and a larger US companies based-hosting.

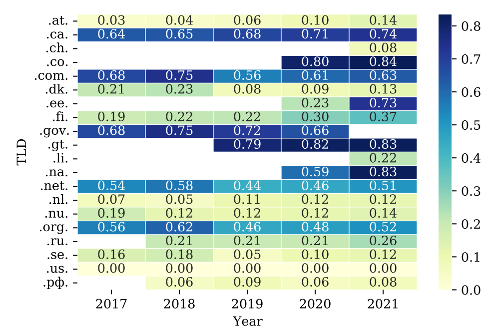
US hosting presence growth: Figure 7 shows the evolution of US hosting presence per TLD. For most TLDs we observe slow but continuous growth over time. Canada's `.ca` has 80% of its domains on US-based hosting companies for IPv4 (figure 7a).

On the other extreme: Russia's ccTLDs (`.ru` and $p\phi$) show a minor proportion of domains hosted by US companies (11% and 4.7%, respectively). Cloudflare, a US company, is among the few US cloud/content companies with a presence in Russia [39], and hosts 6.3% of the `.ru` domain names. Russia is also known for having a controversial piece of legislation on privacy and local data storage [40] and allegedly moving towards building isolated networks [41]. We can only speculate that these two factors contribute to the reduced American presence in Russia's hosting market.

In terms of recent growth, we see Sweden's `.se` US-based hosting increased from 12% to 22% from 2017 to 2021. For IPv6, 73% of Estonia's IPv6 hosting employs US-based companies (Figure 7b). Both results pose challenges for EU policymakers, which have been concerned about Europe's digital sovereignty [13].



(a) IPv4



(b) IPv6

Figure 7: US hosting presence per TLD (IPv4)

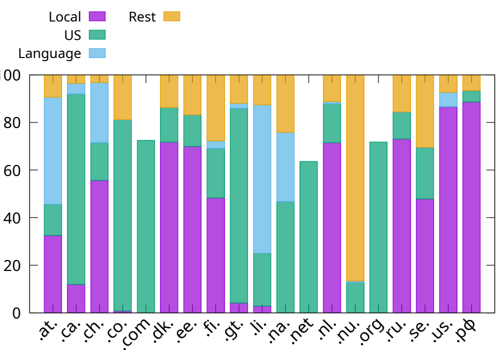
C. Hosting and domain name popularity

The hosting industry has a significant variation in its service: from cheap, shared hosting for small businesses, to large to parking websites to anycast [42] widely distributed services [43] [26].

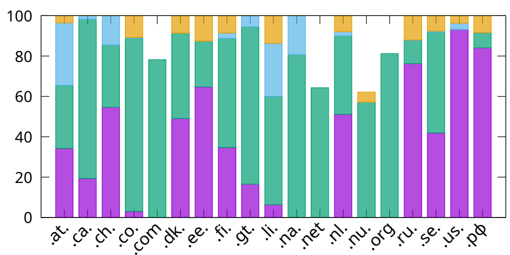
Next, we investigate hosting concentration considering the *popularity* of domain names as a proxy metric for the different segments within the hosting market. The idea is to determine the location of the companies hosting names on the Alexa Top list. It stands to reason that domains on the Alexa list are likely to use robust infrastructure.

To do that, we split each TLD dataset into two categories: if it is present or not on Alexa domain lists [44] on the same day of the measurement.

Figure 8 shows the results for IPv4 in 2021. First, in 8a, we see that the local providers dominate the hosting of domains outside the Alexa list. This, however, differs from what we see in 8b, where we consider domains on the Alexa list. Here we see that most have more significant usage of US-based companies. It is interesting to see that Russian popular domains are mostly hosted by Russian companies.



(a) Not in Alexa



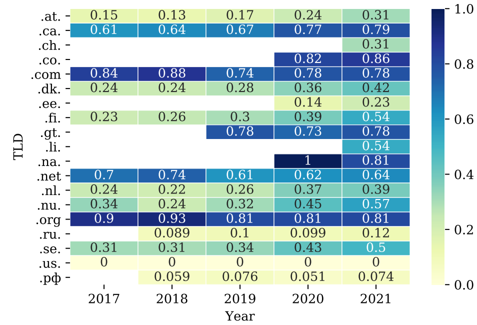
(b) In Alexa

Figure 8: Hosting industry and domains popularity in 2021

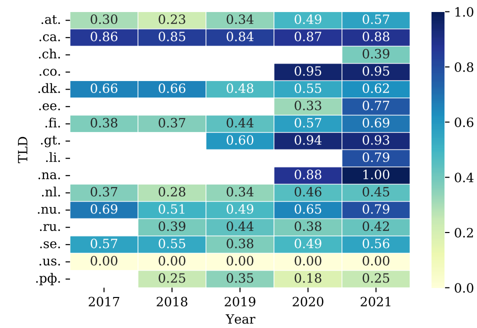
Figure 9 shows the evolution of usage of US hosting companies in Alexa domains for each TLD. Except for the Russian TLDs and Estonia's `.ee`, we observe growth for the US hosting presence in Alexa ranked domains for all TLDs. These results suggest a trend among business of popular websites moving towards large companies.

VI. Conclusions and Future Work

This work provides a new perspective on Internet centralization, complementing previous works. By focusing on the hosting industry, we show how it is heavily concentrated: 10 hosting providers account for most of the hosting for all TLDs considered. We show that European ccTLDs have a strong hosting industry. However, US-based providers have been continuously conquering the market, especially in the high end of it – the popular domain names, which poses challenges for the European Union's goals of digital sovereignty. Russia, on the other end of the spectrum, shows far less reliance on US-based companies.



(a) IPv4



(b) IPv6

Figure 9: US hosting presence on Alexa domains over time

Future work will focus on addressing the limitations of our hosting inferences described in Section III C. In particular, a systematic characterization of website content via web crawling will help further understand possible reasons behind centralization behaviors. Moreover, by using a classifier, we can treat parking domains as a particular case. Finally, coverage of (arbitrary) fully qualified domain names can be increased by considering domain names learned from other data sources (e.g., Certificate Transparency Logs or Common Crawl Data)

VII. Acknowledgements

We thank our reviewers for their constructive suggestions and feedback. Funding for this work was provided in part by Capes PHD scholarship number 88887.480774/2020-00, the EU H2020 CONCORDIA project (830927) and the NWO-DHS MADDVIPR project (628.001.031/FA8750-19-2-0004). This research used data from OpenINTEL,a project of the University of Twente, SURF, SIDN, and NLnet Labs.

References

[1] J. Arkko, "The influence of Internet architecture on centralised versus distributed Internet services," *Journal of Cyber Policy*, vol. 5, no. 1, pp. 30–45, 2020. [Online]. Available: https://doi.org/10.1080/23738871.2020.1740753

[2] J. Arkko, "Centralised Architectures in Internet Infrastructure," Internet Draft, Nov. 2019. [Online]. Available: https://tools.ietf.org/html/draft-arkko-arch-infrastructure-centralisation-00

[3] Arkko, Jari and Tramme, B. and Nottingham, M and Huitema, C and Thomson, M. and Tantsura, J. and ten Oever, N., "Considerations on Internet Consolidation and the Internet Architecture," Internet Draft, Jul. 2019. [Online]. Available: https://tools.ietf.org/html/draft-arkko-iab-internet-consolidation-02

[4] B. Schneier, "Censorship in the Age of Large Cloud Providers," 2018. [Online]. Available: https://www.schneier.com/essays/archives/2018/06/censorship_in_the_ag.html

[5] I. Society, "Consolidation in the Internet Economy," 2019. [Online]. Available: https://future.internetsociety.org/2019/

[6] C. Kang and D. McCabe, "Lawmakers, united in their ire, lash out at big tech's leaders," *New York Times*, July. 29 2020. [Online]. Available: https://www.nytimes.com/2020/07/29/technology/big-tech-hearing-apple-amazon-facebook-google.html

[7] G. C. M. 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*, ser. IMC '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 42–49.

[8] 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 *Proceedings of the ACM Internet Measurement Conference*, ser. IMC '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 634–647.

[9] D. McCabe and D. Wakabayashi, "10 States Accuse Google of Abusing Monopoly in Online Ads," *New York Times*, Dec. 16 2020. [Online]. Available: https://www.nytimes.com/2021/01/09/technology/apple-google-parler.html

[10] A. Satarianoi, "Big Fines and Strict Rules Unveiled Against 'Big Tech' in Europe," *New York Times*, Dec. 15 2020. [Online]. Available: https://www.nytimes.com/2020/12/15/technology/big-tech-regulation-europe.html

[11] S. Zuboff, *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile books, 2019.

[12] T. Wu, *The Curse of Bigness: Antitrust in the New Gilded Age*. Columbia Global Reports, 2018.

[13] S. Erlanger and A. Satariano, "Europe feels squeeze as tech competition heats up between u.s. and china," *New York Times*, Sep. 11 2020. [Online]. Available: https://www.nytimes.com/2020/09/11/world/europe/eu-us-china-technology.html

[14] K. Roose, "In Pulling Trump's Megaphone, Twitter Shows Where Power Now Lies," *New York Times*, Jan. 9 2021. [Online]. Available: https://www.nytimes.com/2021/01/09/technology/trump-twitter-ban.html

[15] J. Nicas and D. Alba, "Amazon, Apple and Google Cut Off Parler, an App That Drew Trump Supporters," *New York Times*, Jan. 9 2021. [Online]. Available: https://www.nytimes.com/2021/01/09/technology/apple-google-parler.html

[16] N. Perlroth, "Hackers used new weapons to disrupt major websites across U.S." *New York Times*, p. A1, Oct. 22 2016. [Online]. Available: http://www.nytimes.com/2016/10/22/business/internet-problems-attack.html

[17] C. Williams, "Bezos DDoS'd: Amazon Web Services' DNS systems knackered by hours-long cyber-attack," https://www.theregister.co.uk/2019/10/22/aws_dns_ddos/, 10 2019.

[18] B. Schneier, "Russia's Hacking Success Shows How Vulnerable the Cloud Is," *Foreing Policy*, p. A1, May. 24 2021. [Online]. Available: https://foreignpolicy.com/2021/05/24/cybersecurity

[19] M. Allman, "Comments on DNS Robustness," in *Proceedings of the Internet Measurement Conference 2018*, ser. IMC '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 84–90. [Online]. Available: https://doi.org/10.1145/3278532.3278541

[20] P. Hoffman, A. Sullivan, and K. Fujiwara, "DNS Terminology," IETF, RFC 8499, Nov. 2018. [Online]. Available: http://tools.ietf.org/rfc/rfc8499.txt

[21] Facebook, "Facebook Reports Second Quarter 2021 Results," *New York Times*, July 28 2021. [Online]. Available: https://investor.fb.com/investor-news/press-release-details/2021/Facebook-Reports-Second-Quarter-2021-Results/default.aspx

[22] R. Prodan and S. Ostermann, "A survey and taxonomy of infrastructure as a service and web hosting cloud providers," in *2009 10th IEEE/ACM International Conference on Grid Computing*, 2009, pp. 17–25.

[23] M. Belshe, R. Peon, and M. Thomson, "Hypertext Transfer Protocol Version 2 (HTTP/2)," IETF, RFC 7540, May 2015. [Online]. Available: http://tools.ietf.org/rfc/rfc7540.txt

[24] F. B. Inisghts, "Web Hosting Services Market," *New York Times*, Dec. 15 2020. [Online]. Available: https://www.fortunebusinessinsights.com/industry-reports/web-hosting-services-market-100863

[25] S. Bates, J. Bowers, S. Greenstein, J. Weinstock, and J. Zittrain, "Evidence of decreasing internet entropy: The lack of redundancy in dns resolution by major websites and services," National Bureau of Economic Research, Tech. Rep., 2018.

[26] S. Tajalizadehkhoob, M. Korczyński, A. Noroozian, C. Ganán, and M. van Eeten, "Apples, oranges and hosting providers: Heterogeneity and security in the hosting market," in *NOMS 2016-2016 IEEE/IFIP Network Operations and Management Symposium*. IEEE, 2016, pp. 289–297.

[27] B. Ager, W. Mühlbauer, G. Smaragdakis, and S. Uhlig, "Web content cartography," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, 2011, pp. 585–600.

[28] 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.

[29] Mozilla, "Public suffix list," https://publicsuffix.org/. [Online]. Available: https://publicsuffix.org/

[30] CAIDA, "Routeviews Prefix to AS mappings Dataset for IPv4 and IPv6," 2020. [Online]. Available: http://www.caida.org/data/routing/routeviews-prefix2as.xml

[31] CAIDA, "Inferred AS to Organization Mapping Dataset," https://www.caida.org/data/as-organizations/, 2017.

[32] A. Studio, "Countries, languages & continents data," https://github.com/annexare/Countries. [Online]. Available: https://github.com/annexare/Countries

[33] J. A. Muir and P. C. V. Oorschot, "Internet geolocation: Evasion and counterevasion," *ACM Comput. Surv.*, vol. 42, no. 1, pp. 1–23, 2009. [Online]. Available: http://portal.acm.org/citation.cfm?id=1592451.1592455&coll=ACM&dl=ACM&CFID=102306505&CFTOKEN=56109667

[34] I. Poese, S. Uhlig, M. A. Kaafar, B. Donnet, and B. Gueye, "IP Geolocation Databases: Unreliable?" *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 2, pp. 53–56, Apr. 2011.

[35] W. I. Launch, "Ipv6 is the new normal," https://www.worldipv6launch.org/, Jun 2012.

[36] T. Vissers, W. Joosen, and N. Nikiforakis, "Parking sensors: Analyzing and detecting parked domains," in *Proceedings of the 22nd Network and Distributed System Security Symposium (NDSS 2015)*. Internet Society, 2015, pp. 53–53.

[37] P. Mockapetris, "Domain names - concepts and facilities," IETF, RFC 1034, Nov. 1987. [Online]. Available: http://tools.ietf.org/rfc/rfc1034.txt

[38] J. Abley, K. Lindqvist, E. Davies, B. Black, and V. Gill, "IPv4 Multihoming Practices and Limitations," IETF, RFC 4116, Jul. 2005. [Online]. Available: http://tools.ietf.org/rfc/rfc4116.txt

[39] X. Gong, "Moscow, Russia: CloudFlare's 83rd data center," Jun. 16 2016. [Online]. Available: https://blog.cloudflare.com/moscow/

[40] E. Moyakine and A. Tabachnik, "Struggling to strike the right balance between interests at stake: The 'Yarovaya', 'Fake news' and 'Disrespect' laws as examples of ill-conceived legislation in the age of modern technology," *Computer Law & Security Review*, vol. 40, p. 105512, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0267364920301175

[41] J. Sherman, "Russia and Iran Plan to Fundamentally Isolate the Internet," *Wired*, Jun. 19 2019. [Online]. Available: https://www.wired.com/story/russia-and-iran-plan-to-fundamentally-isolate-the-internet

[42] D. McPherson, D. Oran, D. Thaler, and E. Osterweil, "Architectural Considerations of IP Anycast," IETF, RFC 7094, Jan. 2014. [Online]. Available: http://tools.ietf.org/rfc/rfc7094.txt

[43] A. Noroozian, M. Korczynski, S. Tajalizadehkhoob, and M. van Eeten, "Developing security reputation metrics for hosting providers," in *8th Workshop on Cyber Security Experimentation and Test ({CSET} 15)*, 2015.

[44] Alexa, "Alexa: Keyword Research, Competitive Analysis & Website Ranking," May 2021. [Online]. Available: https://www.alexa.com/