An Empirical View on Consolidation of the Web

TRINH VIET DOAN, Technical University of Munich
ROLAND VAN RIJSWIJK-DEIJ, University of Twente, NLnet Labs
OLIVER HOHLFELD, Brandenburg University of Technology
VAIBHAV BAJPAI, Technical University of Munich

The majority of Web content is delivered by only a few companies that provide Content Delivery Infrastructures (CDIs) such as Content Delivery Networks (CDNs) and cloud hosts. Due to increasing concerns about trends of centralization, empirical studies on the extent and implications of resulting Internet consolidation are necessary. Thus, we present an empirical view on consolidation of the Web by leveraging datasets from two different measurement platforms. We first analyze Web consolidation around CDIs at the level of landing webpages, before narrowing down the analysis to a level of embedded page resources. The datasets cover 1(a) longitudinal measurements of DNS records for 166.5 M Web domains over five years, 1(b) measurements of DNS records for Alexa Top 1 M over a month and (2) measurements of page loads and renders for 4.3 M webpages, which include data on 392.3 M requested resources. We then define *CDI penetration* as the ratio of CDI-hosted objects to all measured objects, which we use to quantify consolidation around CDIs. We observe that CDI penetration has close to doubled since 2015, reaching a lower bound of 15% for all `.com`, `.net`, and `.org` Web domains as of January 2020. Overall, we find a set of six CDIs to deliver the majority of content across all datasets, with these six CDIs being responsible for more than 80% of all 221.9 M CDI-delivered resources (56.6% of all resources in total). We find high dependencies of Web content on a small group of CDIs, in particular, for fonts, ads, and trackers, as well as JavaScript resources such as jQuery. We further observe CDIs to play important roles in rolling out IPv6 and TLS 1.3 support. Overall, these observations indicate a potential oligopoly, which brings both benefits but also risks to the future of the Web.

CCS Concepts: • **Networks** → **Network measurement**; *Network structure*; World Wide Web (network structure); Public Internet;

Additional Key Words and Phrases: Web, consolidation, centralization, content delivery

ACM Reference format:
Trinh Viet Doan, Roland van Rijswijk-Deij, Oliver Hohlfeld, and Vaibhav Bajpai. 2022. An Empirical View on Consolidation of the Web. *ACM Trans. Internet Technol.* 22, 3, Article 70 (February 2022), 30 pages.
https://doi.org/10.1145/3503158

1 INTRODUCTION

Today's Web traffic is mostly delivered from large content serving hosts, such as from Google, Amazon, or Facebook [2, 22, 72]. These organizations operate wide-reaching **Content Delivery**

Authors' addresses: T. V. Doan and V. Bajpai, Technical University of Munich, Boltzmannstr. 3, 85748 Garching, Germany; emails: {doan, bajpaiv}@in.tum.de; R. van Rijswijk-Deij, University of Twente, NLnet Labs, Drienerlolaan 5, 7522 NB, Enschede, The Netherlands; email: r.m.vanrijswijk@utwente.nl; O. Hohlfeld, Brandenburg University of Technology, Konrad-Wachsmann-Allee 5, 03046 Cottbus, Germany; email: oliver.hohlfeld@b-tu.de.

1533-5399/2022/02-ART70 $15.00
https://doi.org/10.1145/3503158

Infrastructures (**CDIs**) that are deeply involved in various layers of the Internet and the Web, causing a concentration of control. Such CDIs include both cloud [65] and **Content Delivery Networks** (**CDNs**) [28, 84, 86, 106], along with their DDoS protection [66] infrastructures as dominant but different examples of infrastructures for content delivery. Consequently, they have become essential drivers of the Internet and Web ecosystem, in particular, as the associated benefits (in terms of availability, performance, and security) continuously draw in more and more customers and users. For instance, businesses increasingly decide to externalize technical parts of their operation, such as (internal) communication infrastructure to CDIs [75], which often allows re-allocation of server costs and maintenance to other parts of the business.

In the last decade, previous studies have seen a vast expansion of such infrastructures, primarily hosted by "hyper giants" [22, 43, 72, 92], i.e., large content providers that deliver significant fractions of the overall Internet traffic. As a result, the dominance of these hyper giant CDIs alters the Internet topology substantially: As an effect of their quest for speed, CDIs seek for establishing private network interconnects directly with eyeball networks, which host their customers to avoid having to traverse the tiered hierarchy [43], ultimately flattening the Internet **Autonomous System** (**AS**) topology [72]. A recent study [11] confirms this flattening trend and shows that cloud networks can reach more than 76% of the Internet without having to traverse Tier 1 and 2 ISPs, which makes cloud infrastructure widely reachable with fairly low latency [31]. At the same time, this evolution also eases protocol deployment [108]: E.g., Google, Facebook, and Cloudflare are observed to be main drivers in the deployment of TLS 1.3 [94] among other CDIs in a recent study [55], which brings benefits in terms of latency and security to their end users.

Following these evolutions from recent years, there has been an increased interest in studying the impact and effects of such consolidation trends within the Internet economy [10, 60, 63], in which those CDIs play an important role. In particular, concerns about these trends have been raised with respect to user privacy, security, and legal matters as a consequence of centralization of data and service administration. The Internet Society has proposed an action plan starting 2020 to facilitate the "decentralized Internet way of networking" [61], in which end users are the focus [83]. In the same vein, the European Commission has advocated for digital sovereignty as part of multiple strategies and programs [36, 37] to shape Europe's digital future and to become more independent from major tech companies from the United States and Asia. Similar concerns resulted in a series of antitrust hearings [68] and lawsuits [97] in the United States. Related risks have also been a lead topic of **Internet Architecture Board** (**IAB**) discussions [10, 58], where the trend is discussed from technical, societal, and economic perspectives.

Overall, while there are currently many concerns about increasing consolidation by several important Internet communities [8–10, 17–19, 59], more contemporary empirical studies on Internet consolidation from different views are necessary due to its multifaceted nature (Section 8).

In this article, we analyze recent trends in Web consolidation by leveraging a set of measurement datasets (Section 2) to quantify *CDI penetration* as a metric for Web consolidation: We define CDI penetration (Section 2.3) as the ratio of CDI-hosted objects to all measured objects, which reflects the extent to which webpages rely on CDIs. While we first focus on CDI penetration regarding the hosting of landing pages, i.e., how many landing pages are hosted on CDIs, we later also consider CDI penetration for the delivered page resources, which compose the content of an individual webpage such as images, scripts, or fonts. We begin the study with a high-level analysis based on DNS records of three **Top-Level Domains (TLDs)**, covering more than 140 M unique Web domains (growing to 166.5 M domains over a period of roughly 5 years), and narrow the analysis down to a level of page resources (392.3 M resources) for a snapshot of 4.3 M webpages. Moreover, we consider Web consolidation for IPv6 and carry out case studies to further examine the impact of CDIs on Web content. Our primary findings are:

Landing Pages (Section 3). Using longitudinal DNS measurement data from March 2015 until January 2020 for all Web domains in the `.com`, `.net`, and `.org` TLD namespaces (166.5 M Web domains as of January 2020), we observe that the number of webpages hosted on CDIs has increased by 83%, from roughly 8.2% (2015) to 15% (2020) overall (Section 3.1).

Considering frequently visited webpages through DNS measurements for Alexa Top 1 M toplists, we identify 24.3% of the pages to host their landing page on CDIs over IPv4 as of December 2019, with the penetration being higher among more popular (i.e., higher ranked) pages. When only considering domains with IPv6 support, CDIs even host 81.9% of the webpages (Section 3.2).

In both the TLD and the Alexa Top 1 M datasets, we find only a small set of CDIs to host the vast majority of landing pages, namely, Cloudflare, Google, Amazon, Akamai, Fastly, and Microsoft, which indicates consolidation around these providers in particular.

Webpage Resources (Section 4). With the help of page load data measured for 4.3 M webpages in January 2020, we determine 32.1% of the webpages and 56.6% of all requested 392.3 M resources to be delivered by CDIs. Landing pages that are hosted on CDIs tend to include a higher relative number of resources hosted on CDIs in most cases, with webpages in general using multiple different CDIs simultaneously for resource delivery.

Similar to our results from the DNS measurements, we see Google, Amazon, Cloudflare, Facebook, Akamai, and Fastly as responsible for more than 80% of all CDI-hosted resources, accounting for nearly half of all resources in total. In particular, Google and Amazon together are responsible for more than half of the CDI-hosted resources, which is close to 30% of all measured resources.

While the large contribution of these CDIs indicates strong dependencies for Web content, from an overall perspective, non-CDI hosts still account for a meaningful content share (pages: 67.9%, resources: 43.4%), which suggests variety in Web hosting services still.

Case Studies (Section 5). In a set of case studies, we observe a substantial amount of webpages (40.5%) to include page functionalities through CDI-hosted jQuery scripts, particularly from Google and Cloudflare (Section 5.1). Furthermore, we observe a higher usage of TLS 1.3 among some CDIs (Section 5.3) compared to non-CDIs, which supports observations regarding their potential in pushing the deployment of new protocols and standards. Moreover, we spot consolidation around Google and Amazon in the domain of Web ads and trackers (Section 5.2), and around Google and Facebook for video content (Section 5.4), which also shows consolidation beyond regular content hosting.

The goal of the study is *not* to evaluate individual CDIs to help customers make business decisions. Furthermore, we do *not* take a stance for or against Web content consolidation, as it brings both benefits and risks. Instead, the goal is to study and quantify the involvement of CDIs, especially that of larger players, in the delivery of Web content, which we achieve by mapping dependencies of content to specific CDIs. We remark that CDNs and cloud infrastructures inherently follow different architectural designs, but both consolidate content hosting to few companies; Therefore, we consider both in order to investigate content hosting hyper giants. The presented CDI penetration only represents a lower bound, as we only consider measurements of the surface Web; the actual CDI penetration may be much different for the deeper Web (e.g., social media platforms or paid services), which serve richer and personalized content from dedicated CDIs, or for internal Web pages [7].

Reproducibility and Ethics. To enable reproducibility [12, 13] of our analysis, we share the scripts, Jupyter notebooks, and auxiliary data used in this study.[1] The measurements and analysis do not raise any ethical issues.

[1] GitHub repository: https://github.com/tv-doan/acm-toit-2022-web-consolidation.

2 METHODOLOGY

2.1 Datasets

For our study, we use multiple datasets to provide views on Web consolidation: data based on longitudinal (`.com`, `.net`, `and .org`) and popularity-based (Alexa Top 1 M) DNS measurements, in addition to data based on page loads of common landing pages including their embedded resources.

2.1.1 OpenINTEL Measurements. The OpenINTEL [111] project has performed daily DNS queries from a national research and education network in Netherlands, Europe.

Longitudinal Data. We analyze daily aggregates of active DNS measurements from OpenINTEL for all Web domains (`www.`) of the `.com`, `.net`, and `.org` TLD namespace. The datasets cover measurements for more than 140 M distinct domains in total since March 2015 (166.5 M as of January 2020), i.e., 50% of the global DNS name space [111]. We use aggregates per day due to the large size of the raw dataset and to study the evolution of CDI penetration over the years.

Popularity-based Data. For the most popular `www.` domains based on the Alexa Top 1 M toplist [3], OpenINTEL makes active DNS measurements per day publicly available starting February 2016 [85]. The datasets contain DNS responses for the queried resource records, for which we focus on and distinguish between `A` and `AAAA` records (which represent the IPv4 and IPv6 addresses of Web domains, respectively), which also allows us to examine the influence of CDIs on IPv6 support.

Previous work has shown that the Alexa 1 M list is volatile and experiences frequent changes regarding the included domains [88, 98–100]. While longitudinal data for Alexa Top 1 M are available, the meaningfulness of analyses is uncertain because of the unstable nature of the toplist; in particular, potentially observed changes in the longitudinal Alexa data can reflect changes in the DNS itself (such as a change of the webpage host); however, can also reflect changes in the sampling and composition of the Alexa toplist. As such, we use the measurements of a whole month (15 GB in size, uncompressed) to counteract the frequent daily changes but do not extend the analysis over a longer period of time (for which we use the longitudinal `.com`, `.net`, and `.org` data instead).

2.1.2 HTTP Archive Measurements. The HTTP Archive [57] has performed monthly page load measurements for a list of popular webpages. Since July 2018, it uses URLs from Google's **Chrome User Experience** (**CrUX**) report [46], covering 1.3 M URLs first and later increasing the amount to more than 3–4 M URLs in December 2018 [56]. The CrUX URLs cover popular websites, which are visited by real users of the Google Chrome browser. The HTTP Archive collects data 1–2 times per month over IPv4 from California, the United States.

WebPagetest. The HTTP Archive measurements represent `WebPagetest` [114] runs over desktop and mobile Web browsers, which record various metrics of the base HTML page (i.e., the landing page) and all embedded page resources: The `WebPagetest` takes a URL as input argument and visits the webpage like a regular user would via the configured browser. A test run first fetches the document behind the URL, then parses and loads the resources included in the HTML document. The browser then renders the webpage visually, until the page is fully loaded and network activity stops, which can include execution of embedded scripts that are fired at certain page load events.

Along the way, the `WebPagetest` also collects performance data on the interaction with the webpage, such as the fetching and rendering processes (based on the Navigation Timing specification [112]). Additionally, it records meta data based on HTTP responses for the page and its resources, such as protocol versions or CDN providers. This collected information can then, for instance, be exported as a JSON-formatted HTTP Archive (`.har`) file.

For our analysis, we leverage the HTTP Archive dataset collected from a desktop browser (Chrome) for January 2020, which covers 4.3 M webpages in total (2.5 TB in size, compressed). We also analyze HTTP Archive data for March 2016 (when it started using Chrome [56] for measurements), 2017 and 2018 (when tests still used Alexa Top 1 M domains as input), as well as 2019 in the same way and find similar trends, which we omit for the sake of brevity.

2.2 CDI Identification

Identifying the destinations of resource requests allows us to determine where a resource (either a webpage or a page resource) is fetched from, such as from a CDI host (which includes CDNs and cloud hosts, as well as DDoS protection infrastructures through which the resources are delivered); or from a "non-CDI". In return, this allows us to calculate the respective CDI penetration based on the identified numbers for CDI-hosted objects (see Section 2.3 below).

Note that resources that originate from cloud hosts are treated equally, i.e., we do not explicitly differentiate between the different cloud service models: In each of the three service models, namely, **Infrastructure as a Service** (**IaaS**), **Platform as a Service** (**PaaS**), and **Software as a Service** (**SaaS**), the delivery of content is ultimately managed by the cloud provider, which runs the hosting infrastructure such as the servers, the storage, and the network connectivity.

Set of CDIs and Patterns. We first compile a set of CDIs to consider for the identification. The `WebPagetest` project provides an extensive (though not exhaustive) set of CDIs along with regular expression patterns used for their hostnames [115].

We use the given set, as it covers hyper giants in the context of Web CDIs (which are the focus of this study) as well as other smaller-sized CDI, and curate the respective regular expressions manually. We then look up corresponding **Autonomous System Numbers** (**ASNs**) for each of the CDIs by querying the databases of PeeringDB [87] and RIPEstat [95]. Thus, we obtain a set of CDIs, along with regular expressions they use for `CNAME` DNS records, and their commonly used ASNs for the identification process described in the following.

Identification Process. In order to identify the use of these CDIs in the measurement data, we apply a methodology that considers (1) DNS redirection through `CNAME` records and (2) the content location based on the announcing AS. In this way, we also account for cases in which a CDI redirects the client to the closest content replica through a `CNAME` record, which can point to an ISP content cache for instance, and would therefore be mapped to the ASN of the ISP instead of that of the CDI [29, 43, 77]. If a `CNAME` record matches a regular expression, the second step, which applies ASN-based identification is not applied.

Due to inherent differences of the measurements, we adjust the identification to the different dataset schemes (see Section 2.1):

OpenINTEL Dataset: OpenINTEL determines the ASN using CAIDA's Prefix to AS mappings dataset (pfx2as) [27] (derived from Route Views [96]) at the time of the measurement, mapping the most specific IP prefix, and announcing AS to the IP endpoints resolved via DNS. Therefore, the data already contains the ASN information needed for the mapping of the CDIs.

— *Data for* `.com`, `.net`, `.org` *Web Domains:* Due to the size of the dataset (both in terms of number of domains and days), coupled with the high number of `CNAME` patterns to check, applying the regular expressions to the `CNAME` records of the domains is not a feasible or scalable solution. Thus, the `.com`, `.net`, and `.org` data by OpenINTEL only consists of daily aggregates, i.e., the number of domains observed for an ASN based on the compiled list of CDIs, in addition to the total number of measured domains.

— *Data for Alexa Top 1 M Domains*: The DNS measurements for Alexa Top 1 M are much more compact and contain both the `CNAME` record (if there is one) and the ASN for a Web domain.

Therefore, we first match a domain's CNAME record against the set of regular expressions. In case of no match, we additionally check whether the domain's ASN matches any of CDIs' ASNs.

HTTP Archive Dataset: The WebPagetest uses the same list of regular expressions along with HTTP header patterns [115] to identify CDIs. As such, the HTTP Archive data already provides CDI identification; for entries that do not have a pre-identified CDI by WebPagetest, we only apply the ASN-based identification: For each resource, we extract the IP address information and perform an ASN lookup using Route Views [96] BGP data (January 2020). Note that in most cases, the pre-identified CDIs are identical with the identification based on ASNs, which indicates conformity of the approach.

We acknowledge that an identification of CDIs based on CNAME regular expressions, HTTP header patterns, and ASN matching is not exhaustive. However, the goal of this analysis is not to determine precise website counts for individual CDIs but instead to study Web consolidation from a remote point of view through CDI penetration as a metric, which we define in the following.

2.3 CDI Penetration

We define CDI penetration as the ratio of the number of CDI-served objects to the total number of objects. In the context of this study, an object can refer to either a landing page (Section 3) or a page resource (Section 4). An object is considered to be served by a CDI if it is assigned to a CDI according to the aforementioned identification method.

For example: If 20 out of 50 measured landing pages are hosted on CDIs, the CDI penetration is 40% for all measured pages. On the other hand, if a webpage embeds 1,000 resources, 300 of which are delivered by a CDI, the CDI penetration of that webpage is 30%.

As such, CDI penetration describes the extent to which objects that compose Web content use CDIs. It further allows to determine the contributions of individual CDIs, which can reveal patterns of Web content consolidation around certain providers.

3 LANDING PAGES

We first perform a longitudinal analysis of landing pages, for which we examine measurements over roughly five years, which measure up to 166.5 M Web domains of the .com, .net, and .org TLDs per day. We use this data to study the evolution of CDI penetration over multiple years (Section 3.1).

We then focus on CDI penetration among popular landing pages based on Alexa Top 1 M measurements over a month (Section 3.2); overall, we observe 5.9 M distinct Web domains in total across all daily Alexa Top 1 M measurements over the month due to its volatility. Out of the Alexa domains, we determine 60.5% to be from the .com, .net., and .org TLDs.

3.1 .com, .net, .org Domains over Time

Figure 1 portrays the overall CDI penetration of all .com, .net, and .org Web domains, i.e., relative to the total number of domains within each TLD, with a general upward trend visible for all three TLDs. Note that the drop from March until May 2017 is due to DNS requests from OpenINTEL being blocked by a domain registrar, which was misidentifying measurement requests as malicious attacks. The drop in late 2018 is a result of Amazon domains declining in numbers, whereas the bounceback increase in early 2019 can be attributed to Google. Starting January 2019, we also observe larger fluctuations in terms of CDI penetration, although the overall number of domains increases steadily. Nearly all of these fluctuating domains "entering and leaving" CDIs are assigned to Google's AS15169; more detailed measurement data from OpenINTEL reveals that this is due to *Wix*, a cloud-based website hosting service, changing DNS configurations to balance loads

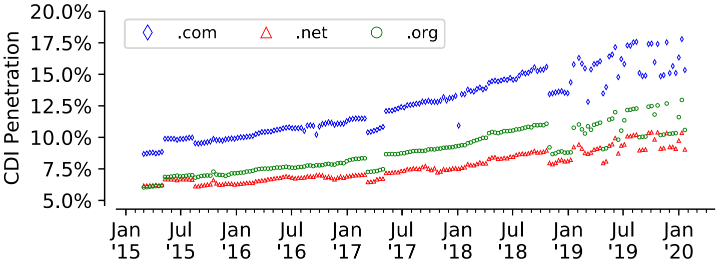


Fig. 1. Time series of CDI penetration for `.com`, `.net`, and `.org` (March 2015–January 2020).

between their own AS58182 and Google's AS15169. These repeated configuration changes affect multiple millions of domains across all three TLDs each time, which has similarly been observed for Wix and Incapsula in previous work [66].

CDI penetration nearly doubles from 2015–2020. The CDI penetration for the three TLDs combined is at around 8.2% in March 2015. Over the course of the years, the overall CDI penetration has increased to 15% −17% as of January 2020, which is an increase by more than 83%. Taking CDI penetration into account, we witness that the penetration for each TLD has increased regardless of the domain space changes: The `.net` CDI penetration experiences the smallest change, as the number of CDI-hosted landing pages grows from 927 k (6.2% penetration) to 1.2 M domains (9.2%). On the other hand, the `.org` domain count for CDIs increases from 634 k (6%) to 1.1 M (10.6%). Yet, both TLDs have experienced declines in their total numbers of domains (`.net`: −1.8 M, `org`: −524 k), showing that despite fewer domains in the dataset overall, the number of domains served by CDI has still increased. Lastly, the absolute number of `.com` webpages hosted by CDIs more than doubles from 10.1 M (8.7%) to 22.4 M (15.6%), which is a significant growth as the number of domains of the `.com` TLD has increased by around 27.4 M domains in total over the period. We find that domains hosted by Amazon cause the largest increase, as they grow from 3.7 M to over 9.6 M `.com` domains alone. Thus, Amazon by itself accounts for around half of the `.com` CDI penetration growth over the years (5.9 M out of 12.3 M additional domains).

Overall, the absolute number of webpages hosted on CDIs increases from 11.6 M to 24.6 M; considering the total number of measured domains has grown by around 25 M domains since March 2015 until January 2020, more than 50% of this amount has migrated to (or is initially hosted on) CDIs. This observation clearly indicates a trend of webpages moving toward the set of CDIs tracked in our study, which suggests a consolidation around these CDIs, in particular, Amazon as of late.

3.2 CDI Penetration by Alexa Rank

CDI penetration is higher among more popular domains. Using Alexa Top 1 M to consider webpage popularity in the analysis, we expect a higher CDI penetration among higher ranked webpages (due to the higher traffic volume they face), which would be dropping off together with popularity toward the lower ranks (cf. [70]). To this end, we consider Alexa ranks 1, 10, 100, 1K, and n·10 K with 1≤n≤100, up to 1 M based on the OpenINTEL data measured in December 2019. Figure 2 (top) shows the distribution of CDI penetration by Alexa rank for both `A` as well as `AAAA` records (webpages over IPv4 and IPv6, respectively).

The top plot in Figure 2 represents the total CDI penetration considering all CDIs combined. As can be seen, the penetration for all Top 1 M domains appears much higher for `A` records (24.3%) compared to `AAAA` records (12.5%). However, note that only 153 k domains of the toplist domains have `AAAA` records (in the median case across the month); 125 k of those are served by a CDI, which

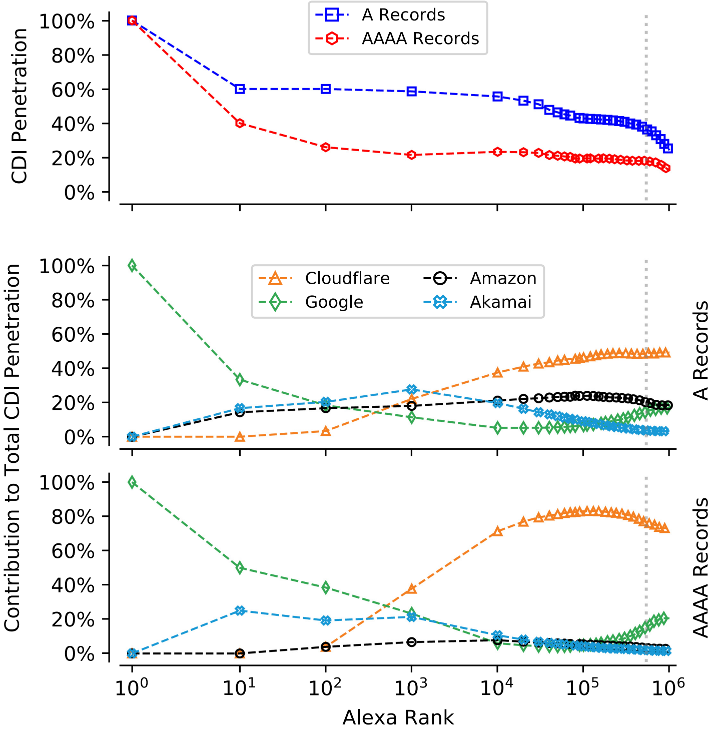


Fig. 2. CDI penetration by Alexa rank in December 2019: overall penetration (top), relative contribution of largest CDIs for A/AAAA records (middle/bottom).

results in a CDI penetration of 81.9% when only considering Alexa domains with AAAA records. This observation indicates a correlation between being hosted on a CDI and having IPv6 support. We find that the daily Alexa 1 M toplists actually report fewer samples than one million domains in December 2019, with some lists already ending at 537 k domains (denoted by the vertical line in the plot). Thus, CDI penetration beyond this rank might be higher than displayed, since we calculate the penetration using the respective Alexa rank as denominator.

Regardless, both curves support the hypothesis that popular content is more likely to be provided by a CDI, particularly for A records. This is expected, as popular pages experience higher traffic and, thus, are more likely to require the dedicated infrastructure provided by a CDI to handle this load. From Top 1 k to Top 10 k, the penetration remains around the same for both record types (A: 58.6% → 55.6%, AAAA: 21.5% → 23.3%). However, from Top 10 k to Top 100 k, the penetration drops significantly to 42.8% for A records and 19.5% for AAAA records. The penetration rates decrease slowly beyond Top 100 k up to rank 530 k (A: 37.0%, AAAA: 18.2%), after which the Alexa aggregations become much more unreliable for December 2019, as mentioned above, although the plots suggest a continued downwards trend.

Google and Cloudflare are main contributors to CDI penetration. To demonstrate the individual contribution of CDIs, we calculate the *relative* median contribution of a CDI to the overall CDI penetration seen at a specific rank. The contributions for the largest CDIs in terms of numbers of hosted webpages are shown in Figure 2 (middle/bottom).

Regarding the higher ranked and more popular webpages, Google dominates, as many domains among the Top 100 are region- and language-specific instances of www.google.* with different TLDs; in particular, we find 12 language-localized domains for www.google.*, as well

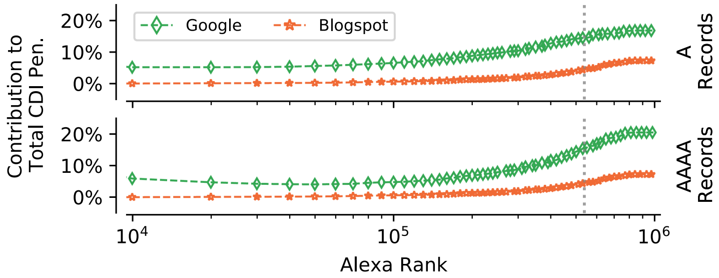


Fig. 3. Contribution of Blogspot domains in relation to contribution of all webpages delivered by Google's CDI within Alexa Top 1 M.

as www.youtube.com, www.blogspot.com, and www.spotify.com to be hosted on Google's CDI among the Alexa Top 100. Furthermore, Google's relative contribution also begins to grow starting from rank 100 k for both A and AAAA records: We observe that this is due to an increase in domains belonging to Blogspot (acquired by Google in 2003), see Figure 3, as well as **Google Hosted Sites** (**GHS**).

We further notice that Akamai peaks at a rank of roughly 1 K (27.6% for A and 21.3% for AAAA records) and overtakes all other CDI providers among A records in terms of contribution, which indicates that Akamai mostly serves webpages of customers with higher popularity. This is also reflected in Akamai's share among the Top 100 domains, to which Akamai contributes 20.3% for A records and 19.2% for AAAA records. We observe that Akamai hosts websites of enterprises that are known to operate their own CDI, such as www.microsoft.com, www.apple.com, and www.amazon.*. However, the Amazon websites also have records that map to Amazon's CDI itself, which indicates the use of multiple CDIs simultaneously [53].

We spot Cloudflare to exhibit a rather steep increase beyond the Top 100 domains, in particular, for AAAA records. Cloudflare begins to surpass other CDIs in the rank range of roughly 1 K–1 M and reaches up to 49.2% contribution to the overall CDI penetration for A records and 83.1% for AAAA records. We speculate that this likely due to Cloudflare offering a free plan to customers, which consists of basic CDI features such as DDoS protection, TLS, IPv6 support, and more recently QUIC [42]. Overall, considering the lower Alexa Top 1 M ranks, this (together with the above observation for Blogspot pages and GHS) strongly indicates that smaller webpages and blogs commonly make use of hosting services (and the associated benefits) offered by CDIs; these pages likely do not have their own dedicated technical infrastructure and, thus, leverage the externalization. As a result, we observe Cloudflare in particular to be responsible for a large percentage of AAAA records for all webpages among Alexa Top 1 M, indicating that such CDIs also play an important and central role in the deployment of upcoming protocols and new technologies.

We witness Amazon's relative contribution to be stable throughout the Alexa ranks, although its contribution increases toward the long-tail for A records where Amazon surpasses Google and Akamai. These identified webpages are primarily making use of **Amazon Web Services** (**AWS**) for hosting purposes, which shows popularity of cloud solutions for webpage hosting. Considering all Top 1 M domains, we further find Fastly to contribute a similar share (A: 2.5%, AAAA: 1.1%) of domains as Akamai (A: 3.2%, AAAA: 1.5%). Microsoft contributes similar shares to the A records (2.7%); however, it does not serve any AAAA record in the dataset.

4 WEBPAGE RESOURCES

Webpages are typically composed of multiple resources (e.g., videos or images), which can be delivered by different CDIs. As the previous analysis only focused on the landing pages based

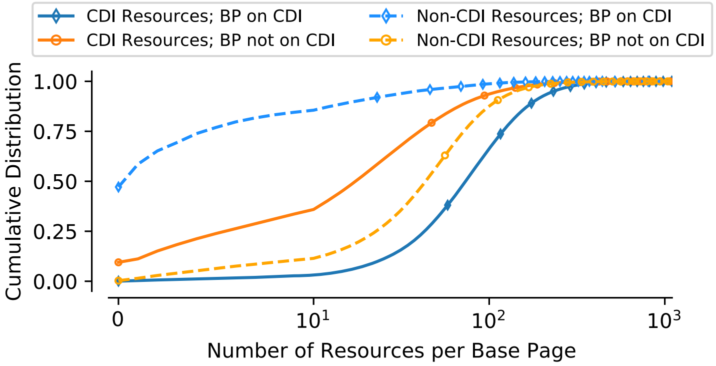


Fig. 4. Cumulative distributions of number of resources per page, split by CDI-hosted and non-CDI hosted pages and resources.

on DNS measurements, we now focus on the involvement of CDIs in the delivery of webpage resources. For this, we extract all resource requests for each page (also referred to as *base page* in the following) measured by the HTTP Archive in January 2020.

In total, the HTTP Archive dataset from that month contains measurements for roughly 4.3 M pages, of which around 1.4 M are identified to be hosted on a CDI, i.e., an overall CDI penetration of 32.1%. We describe this page-centric analysis in Section 4.1. This penetration is comparable to the CDI penetration of `A` records among Alexa Top 1 M (24.3%) observed in Section 3.2, although note that the datasets are inherently different (Section 2).

Further, these HTTP archive measurements cover 392.3 M resources, of which 221.9 M are identified to be served by a CDI, which results in a CDI penetration of 56.6% considering all resources (see Table 2). We focus on resource-level analysis in Section 4.2.

Note: Highlighted cells in tables denote key values and are further discussed in the text.

4.1 Page-Level Analysis

4.1.1 Number of Resources Per Page. We first determine the absolute numbers of resources embedded per webpage in order to describe the compositions of the webpages. Separating between whether the base page is hosted on a CDI or not, we find that pages delivered by CDIs include a slightly higher number of resources overall: The 75th percentile (Q_3) of the webpages hosted on CDIs is at 127 total resources; however, only at 113 resources for pages that are not hosted on CDIs.

We then further distinguish between resources delivered by CDIs and resources not delivered by CDIs. Figure 4 shows the number of resources per base page (denoted as BP in the legend), split by the type of base page, as well as by CDI-hosted resources and resources not hosted on a CDI. We notice that base pages hosted on a CDI have a higher number of CDI-delivered resources (dark blue diamond markers, Q_3=120 resources) compared to the number of resources not delivered by a CDI (light blue diamond markers, Q_3=5 resources). On the other hand, base pages not hosted on CDIs have more resources coming from non-CDI locations (light orange circle markers, Q_3=72 resources), whereas less CDI-hosted resources are used when the base page is also not delivered by a CDI (dark orange circle markers, Q_3=41 resources).

4.1.2 CDI Penetration Per Page. After looking at the absolute numbers of resources, we now turn to relative numbers: For this, we calculate the CDI penetration for each base page, which represents the percentage of the number of CDI-hosted resources to the total number of resources

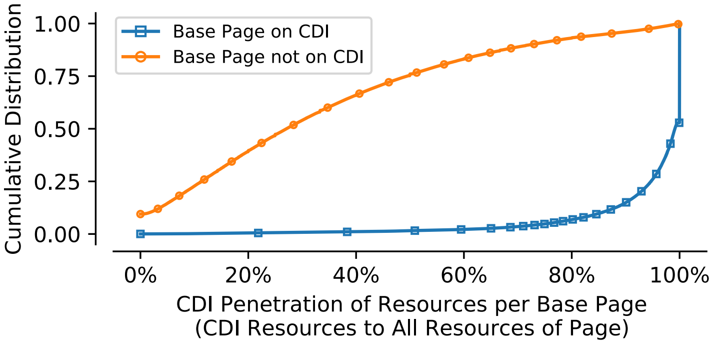


Fig. 5. Cumulative distributions of CDI penetration per page, split by CDI-hosted and non-CDI hosted pages.

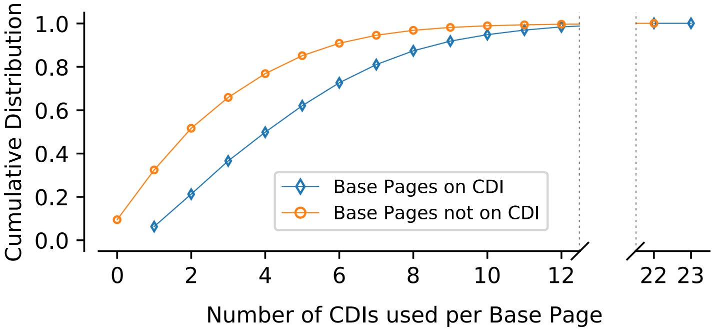


Fig. 6. Cumulative distributions of distinct number of CDIs used for page and resource delivery per page.

on the base page. E.g., on a webpage with 100 resources in total, 20 of which are hosted on CDIs, the CDI penetration of that webpage is 20%.

CDI-hosted pages use a higher percentage of CDI-hosted resources. Figure 5 shows the cumulative distribution for the CDI penetration per page. We find that base pages that are hosted on CDIs exhibit a much higher penetration in terms of their resources as well, since the **interquartile range** (**IQR**) covers penetrations from 95% to 100%. On the other hand, for base pages not hosted on CDIs, the IQR only covers CDI penetrations from 11% to 49%. The heavily skewed distribution of CDI-hosted base pages indicates that most of their resources are hosted on CDIs as well, whereas base pages not hosted on CDIs exhibit a more balanced distribution.

Overall, the previous two analyses have shown that pages which decide to employ CDIs also tend to deliver most of their resources via CDIs, moving the majority of their content to CDIs. Similarly, base pages not hosted on CDIs tend to have more resources coming from non-CDI locations as well, although they also employ a moderate number of CDI-hosted resources, as seen in Figure 4 (dark orange circle markers).

4.1.3 Number of CDIs Per Page. Following the CDI penetration per page, we now determine how many distinct CDIs are involved in the delivery of a webpage and its resources.

Webpages use multiple CDIs for resource delivery. As shown in Figure 6, a base page that employs a CDI is seen to have a higher number of involved CDIs: 72.6% of the CDI-hosted base pages involve up to six distinct CDIs for the content delivery, whereas 76.9% of the pages without CDIs only involve up to four different CDIs. At the 97th percentiles, the former use up to 11 CDIs to serve their content, the latter only make use of up to 8 CDIs.

Table 1. Relative Shares of CDIs Regarding Number of Resources and Object Sizes, Sorted by Number of Resources

	Provider	# Resources (↓)	Sum of Sizes [GB]	Share of CDI Resources by		Share of All Resources by	
				Num.	Size	Num.	Size
(1)	Google	76.6 M	1,494.9	34.5%	24.0%	19.5%	11.1%
(2)	Amazon	38.9 M	1,277.2	17.5%	20.5%	.9%	9.5%
(3)	Cloudflare	27.5 M	956.4	12.4%	15.3%	7.0%	7.1%
(4)	Facebook	17.7 M	423.4	8.0%	6.8%	4.5%	3.1%
(5)	Akamai	15.7 M	496.7	7.1%	8.0%	4.0%	3.7%
(6)	Fastly	10.8 M	411.3	4.9%	6.6%	2.7%	3.0%
(7)	WordPress	4.1 M	109.3	1.9%	1.8%	1.1%	0.8%
(8)	Twitter	4.0 M	65.8	1.8%	1.1%	1.0%	0.5%
(9)	Microsoft	3.8 M	181.0	1.7%	2.9%	1.0%	1.3%
(10)	NetDNA	3.6 M	148.5	1.6%	2.4%	0.9%	1.1%

In rare cases, base pages even involve 23 and, respectively, 22 distinct CDIs out of the 78 identified ones in total, which shows the existence and usage of various players in the CDI landscape. Nevertheless, we find that the majority of resources are delivered by only a small number of (larger and well-known) CDIs, which points to consolidation around these few players.

4.2 Resource-Level Analysis

4.2.1 Resource Shares of CDIs. In order to investigate the individual contributions of CDIs to the delivery of resources, we calculate the relative share of each CDI provider among the number of CDI-hosted resources and the total number of resources; Table 1 presents the top 10 CDIs.

The Top CDIs are responsible for the hosting of most resources. The top six CDIs (namely Google, Amazon, Cloudflare, Facebook, Akamai, and Fastly) together account for 84.4% of CDI-hosted resources (47.7% of all resources). Again, this indicates consolidation around these CDIs.

Individually, Google has the largest share with 76.6 M resources, which is more than one third (34.5%) of CDI-hosted resources and nearly one fifth (19.6%) of all resources. Particularly, Google and Amazon account for more than half of the CDI-hosted (52.1%) and more than a quarter (29.4%) of all resources. As such, while there are many different CDI providers, the results indicate moderate to high concentration around the largest CDIs due to the steep decrease of shares toward less common CDIs, meaning that these CDIs exhibit much smaller footprints in the measured data.

In terms of object sizes, we observe that CDI resources account for 46.3% (13.2 TB) of the overall bytes measured. Note that these measures do not fully represent total traffic volumes delivered by the CDIs, as the object sizes only consider values for a single page load for each base page, whereas individual pages are visited with varying frequency in the wild. Thus, these numbers cannot be used to approximate the total traffic for a webpage or a CDI. However, roughly similar to the number of resources, the top six CDIs account for 81.1% of the measured 13.2 TB by CDIs, albeit with slightly different rankings, i.e., Google (24.0%), Amazon (20.5%), Cloudflare (15.3%), Akamai (8.0%), Facebook (6.8%), and Fastly (6.6%). As such, the resource sizes reflect a comparable view of concentration as the number of resources; nevertheless, we focus on the latter in the following.

4.2.2 CDI Penetration by Resource Type. We further investigate the relationship between the requested resources' content types and CDI hosting. We identify the resource type based on the MIME type provided in the HTTP header of the measured resource; in case the MIME type is empty or does not match any of the common types, we additionally try to classify the resource

Table 2. Distribution of Resources by Type, Sorted by the Types' Shares Relative to All Resources

Resource Type (based on MIME type and file ext.)	**# CDI Resources**	**CDI Pen. of Type**	**# All Resources of Type**	**Share (All) (↓)**
image	82,613,713	46.8%	176,660,130	45.0%
javascript	64,223,345	64.1%	100,195,949	25.5%
text	21,676,628	50.4%	43,017,071	11.0%
html	19,590,470	69.6%	28,148,091	7.2%
other	11,864,834	70.4%	16,847,204	4.3%
font	14,245,056	86.0%	16,569,827	4.2%
application	6,303,607	68.4%	9,220,762	2.4%
video	1,135,211	91.8%	1,236,756	0.3%
audio	265,302	62.2%	426,583	0.1%
Total	221,918,166	56.6%	392,322,373	100.0%

type based on the resource's file extension from its URL. The category *other* denotes resources for which we cannot identify a type with certainty. Table 2 provides an overview of the types along with their absolute as well as relative frequency with respect to CDI hosting.

CDI penetration of Web content is especially high for JavaScript and fonts. We find that images are the most common type of resources, accounting for nearly half (45.0%) of all resources. Moreover, we notice that nearly half (46.8%) of the images are delivered by CDIs. While the number of JavaScript resources are less than the number of images, as JavaScript only accounts for around a quarter (25.5%) of all resources, we observe that the CDI penetration of JavaScript resources is much higher with 64.1%, which indicates that around two out of three scripts are delivered by a CDI. We also see that fonts have an even higher share of being served through CDIs with 86.0%; however, webpages and browsers typically employ fallback fonts, making webpages less prone to breaking when a font cannot be properly loaded in comparison to JavaScript. Considering that images, JavaScript, and fonts are static resources, which are frequently requested and can be easily cached, the higher CDI penetration for these resource types is expected. We also find directly delivered video resources to be rare in the dataset, as most video resources are only linked to and post-loaded dynamically by embedded players, rather than directly embedded into the webpage and delivered along with other resources during a single page load. However, in cases, in which video resources are directly loaded together with the webpage, the videos show a very high dependency on CDIs overall (91.8%). We investigate embedded videos in more detail in a separate case study, which we cover in Section 5.4.

In order to further specify the distribution of types, we determine the most popular CDIs for each resource type, shown in Table 3. We observe that Google is the most prevalent CDI, followed by Amazon, among other popular CDIs that were seen in Table 1 such as Facebook, Cloudflare, Akamai, and Fastly. Again, we primarily find the common CDIs to lead the ranks regarding the different resource types. Nevertheless, non-CDI hosts are the most common sources of resources for 5 out of 9 types, ranging from 31.64% up to 53.24%. Out of the 9 resource types, Google is the leading CDI in six categories, while Amazon leads two categories (*application*, *other*), and Facebook one category (*video*).

Notably, most of the *font* resources (65.26%) are delivered by Google, of which 98.8% are delivered through `fonts.gstatic.com`; the remaining 1.2% are served from miscellaneous sources such as `googleusercontent.com` and WordPress instances hosted on Google's CDI. On the other hand, only 14.03% of fonts are not delivered by CDIs, which indicates a high centralization with potential risks: Google released a statement [47] regarding the use of their Google Fonts API and user privacy, mentioning that logs of font file requests are recorded. In combination with this high

Table 3. Top Content Hosts, Ranked by Relative Contribution to Each Resource Type

Resource Type (↓)	#1	#2	#3	#4	#5	#6
application	— (31.64%)	Amazon (18.59%)	Google (13.78%)	Cloudflare (11.90%)	Akamai (5.04%)	Edgecast (4.72%)
audio	— (37.81%)	Google (33.39%)	Cloudflare (13.59%)	Amazon (7.23%)	CDN77 (2.86%)	Edgecast (2.10%)
font	Google (65.26%)	— (14.03%)	Akamai (7.21%)	Highwinds (2.98%)	Amazon (2.76%)	Cloudflare (2.75%)
html	Google (36.30%)	— (30.40%)	Amazon (10.81%)	Facebook (6.54%)	Akamai (4.80%)	Cloudflare (3.80%)
image	— (53.24%)	Google (11.85%)	Amazon (8.76%)	Cloudflare (7.40%)	Akamai (3.65%)	Facebook (3.19%)
javascript	— (35.90%)	Google (22.54%)	Amazon (9.67%)	Cloudflare (7.68%)	Facebook (7.55%)	Akamai (4.58%)
other	Amazon (30.60%)	— (29.57%)	Google (21.20%)	Fastly (4.43%)	Cloudflare (3.91%)	Akamai (3.34%)
text	— (49.61%)	Google (15.93%)	Cloudflare (7.81%)	Amazon (7.67%)	Facebook (4.07%)	Akamai (2.40%)
video	Facebook (59.20%)	Google (21.11%)	— (8.21%)	Akamai (4.73%)	Amazon (2.40%)	Cloudflare (1.10%)

Non-CDI hosts are denoted by a dash (—) in all tables.

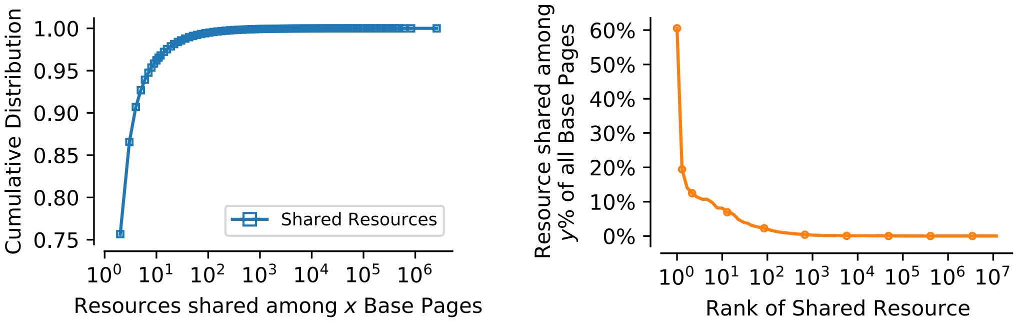
CDI penetration, requests for font resources can potentially be used for profiling or tracking of users as a result. Google further mentions [47] that publish usage statistics for individual fonts and a large scale font analysis, although the public dataset and analysis appear to be discontinued.

4.2.3 Shared Resources. We continue by studying the number of resources that are shared between multiple different base pages (i.e., different base pages load a resource from the same URL), which we will refer to as *shared resources.* Sharing resources (such as jQuery scripts, see Section 5.1) between different base pages can provide benefits in terms of reduced loading times: A resource may already be in the browser cache from visiting another webpage and, thus, does not need to be requested again. As such, consolidating content by leveraging dedicated CDIs can reduce the overall network traffic if resources are shared among a large number of webpages, while providing other benefits related to CDIs such as higher availability and lower latency in addition.

In total, we find 11.7 M resources (3% of all resources) as the baseline number of resources that are shared between at least two base pages. Out of those, 8.9 M resources (75%) are shared between exactly two distinct pages only, as shown in Figure 7(a); the number of resources that are shared between more pages is much lower, as the second highest number of shared resources is seen for three pages (1.3 M or 10.9%) and declines rapidly beyond that (486 k or 4.1% and four pages, 234 k or 2% and five pages). However, this poses privacy risks due to HTTP caching behavior of browsers: If a resource is loaded and cached after visiting website A, a cache hit for that resource in the context of website B can reveal that the user has visited website A before. Sharing resources between a *small* number of webpages amplifies this risk; on the other hand, resources shared between a *large* number of pages makes it more difficult to identify previously visited pages due to a higher anonymity set (but brings risks regarding centralization). While recent browser implementations [69, 113] address this problem by HTTP Cache Partitioning, this fix results in larger traffic volume and higher load times, as cached resources are re-fetched if they are requested in a different context. HTTP Cache Partitioning might shift the ability to track users through shared resources to CDIs instead of individual websites, as the increased number of resource requests to CDIs can potentially reveal browsing behaviors and patterns of users.

More than half of the shared resources are hosted by CDIs. We find that 6 M (51.4%) of the shared resources are hosted by CDIs; Amazon (11.6% of all shared resources), Cloudflare (10.5%), Google (9.6%), and Akamai (5.0%) are the largest contributors.

Although these numbers do not indicate strong consolidation, some resources are shared among a large number of pages, as shown in Figure 7(b): We rank the shared resources in descending order, i.e., the resource that is shared among the highest number of base pages is ranked first. This highest ranked resource, an analytics JavaScript resource from Google (`analytics.js`), is shared among 2.6 M pages (60.5%), followed by an event-related JavaScript resource from Facebook (`fbevents.js`) among 831 k (19.4%) pages. The top 13 shared resources are dominated by Google with 10 resources, while the remaining 3 are assigned to Facebook. The first shared resource from



(a) Cumulative distribution of shared resources by the number of sharing base pages. 75% of the shared resources between distinct base pages are shared across two unique pages.

(b) Shared resources, ranked by number of sharing base pages, and the respective percentage relative to all base pages. The highest ranked shared resource is included in more than 60% of all base pages.

Fig. 7. Distributions of shared resources with respect to unique base pages.

a different CDI is included from Twitter and ranked 14th, being shared among 294 k (6.8%) pages. This observation suggests that the resources that are shared among the highest number of unique base pages originate from Google in particular, with other CDIs besides Google and Facebook being much less prevalent.

5 CASE STUDIES

In addition to the page- and resource-level analyses presented in the previous sections, we now discuss a set of case studies which we explore in order to illustrate potential benefits and drawbacks (among other properties) of hosting content on CDIs.

5.1 CDI Dependency of jQuery

JavaScript is used to dynamically modify content on webpages for a variety of purposes. This includes Web advertisements, for instance, although ads are not essential for webpage functionality and often deliberately blocked by users as a result [91]. In other cases, webpage functionality heavily relies on JavaScript libraries such as *jQuery*; being unable to load the library properly would degrade the intended user experience and may even break the page.

jQuery resources have a moderate dependency on CDIs. Due to the high share of JavaScript resources among CDI-hosted resources, we study this resource type in more detail: We analyze the URLs of the requested resources and determine jQuery scripts (based on the resource name) to account for around 14.5 M (14.5%) among all 100.2 M JavaScript resources. Out of those, 5.4 M jQuery requests are delivered by CDIs, which represents a CDI penetration of 37.4%. Most jQuery scripts are requested from Cloudflare (1.4 M, i.e., 25.3% of all 5.4 M CDI-hosted jQuery scripts) and Google (1.3 M, 23.2%), followed by Amazon (878 k, 16.2%) and the *jQuery CDN* [103] (310 k, 5.7%), which is powered by StackPath (formerly Highwinds). Thus, Google and Cloudflare each serves more than four times the number of jQuery scripts compared to the primary CDI suggested by the jQuery project page [104], despite these third-party CDIs potentially receiving delayed updates to jQuery.

From the 4.3 M measured base pages, we find 1.7 M distinct pages (40.5% of all measured base pages) that include jQuery from CDIs. 21.5% of these pages (371 k) load their included jQuery scripts from multiple CDIs, rather than from a single one exclusively. In order to assess the impact of CDIs on delivery of jQuery scripts, we further examine how many different base pages each

Table 4. Number of Ads (Left) and Trackers (Right) for the Top 10 Providers, Along with Their Share Relative to All Identified Ads and Trackers, Respectively

	Provider	# Ads (↓)	Share (all Ads)	Provider	# Trackers (↓)	Share (all Trackers)
(1)	Google	8,776,465	66.6%	Google	15,995,822	55.3%
(2)	—	2,715,437	20.6%	—	5,073,329	17.5%
(3)	Amazon	401,946	3.1%	Amazon	2,466,341	8.5%
(4)	Akamai	362,619	2.8%	Akamai	1,170,836	4.0%
(5)	Yahoo	291,181	2.2%	Facebook	914,088	3.2%
(6)	Cloudflare	220,693	1.7%	Fastly	680,578	2.4%
(7)	Edgecast	123,498	0.9%	WordPress	598,954	2.1%
(8)	Fastly	116,593	0.9%	Twitter	513,694	1.8%
(9)	Highwinds	32,702	0.2%	Cloudflare	423,429	1.5%
(10)	Internap	21,971	0.2%	Microsoft	323,466	1.1%

More than half of the ads and trackers are delivered by Google, which is one of the most used CDIs for delivery of ads and trackers together with Amazon and Akamai.

CDI serves: Although Google and Cloudflare deliver nearly the same number of resources, Google serves 770 k distinct pages (18%), whereas Cloudflare serves 428 k different pages (10%).

If jQuery scripts from Google's CDI stopped being available or not be delivered properly anymore, this could result in 18% of the 4.3 M measured pages potentially showing a degraded Web experience due to missing page functionality. Similarly, 10% of the webpages would be affected in the case of Cloudflare, and 6% (255 k distinct pages) when jQuery CDN would stop functioning. As such, there is a moderate dependency on CDIs for jQuery. Although, it should be kept in mind that 62.6% of the jQuery resources are not delivered by CDIs, meaning that a large fraction of webpages would be unaffected and not breaking.

5.2 Ads and Trackers

Web advertisements and trackers have been studied extensively by previous work [35, 71, 73]. Lerner et al. [73] find that certain players in the tracking ecosystem have grown in size and with respect to interdependencies, hinting at consolidation in the tracking ecosystem as a consequence.

As the HTTP Archive data contains the resource URL for each resource, we identify ads and trackers based on *EasyList* and *EasyPrivacy* [33] blocklists (from January 2020), respectively. Overall, we find 13.2 M ads (3.4%) and 28.9 M trackers (7.4%) among the 392 M resources in total, with 10.1 M resources classified as both ads *and* trackers.

Google delivers most ads and trackers by far. We observe that Google holds the biggest contribution to ads (66.6%) and trackers (55.3%) by far, followed by non-CDI services (20.6% and 17.5%), Amazon (3.1% and 8.5%), and Akamai (2.8% and 4.0%) in both categories; other CDIs have much slower shares, as shown in Table 4. This corroborates findings of other related studies [15, 26, 35, 76, 79] on trackers and ads, which find that Google has been (and still is) the largest player in online tracking and advertisement through a plethora of different services such as *Google Analytics* or *Doubleclick*, among others. Thus, for the measured surface Web, the majority of the ads and trackers can be attributed to Google, which is responsible for more resources than all non-CDI ad/tracking providers combined.

In contrast, we find that Facebook has a 3.2% share for trackers but only a <0.1% for ads. We explain this due to Facebook providing *Like* buttons and other social plugins to webpage administrators. On the other hand, ads from Facebook are typically only visible on the internal pages after logging in for the most part, which are not measured by the HTTP Archive.

Table 5. Number of Resources, Which Were Requested Over TLS and for Which the TLS Version Was Recorded, Split by TLS Version and Content Provider (Top 10)

	Provider	**TLS 1.0**		**TLS 1.1**		**TLS 1.2**		**TLS 1.3 (↓ %)**		**Identified Resources**
(1)	WordPress	0	(0.0%)	0	(0.0%)	0	(0.0%)	692,339	(100.0%)	692,339
(2)	Facebook	0	(0.0%)	0	(0.0%)	8	(0.0%)	3,053,978	(100.0%)	3,053,986
(3)	Google	152	(0.0%)	16	(0.0%)	783,129	(5.0%)	14,914,626	(95.0%)	15,697,923
(4)	Cloudflare	7	(0.0%)	0	(0.0%)	444,503	(17.6%)	2,083,359	(82.4%)	2,527,869
(5)	Highwinds	0	(0.0%)	0	(0.0%)	302,426	(29.8%)	711,909	(70.2%)	1,014,335
(6)	Akamai	6	(0.0%)	0	(0.0%)	1,672,169	(58.3%)	1,194,278	(41.7%)	2,866,453
(7)	Fastly	1	(0.0%)	0	(0.0%)	1,335,349	(92.1%)	114,748	(7.9%)	1,450,098
(8)	—	291,196	(2.2%)	3,329	(0.0%)	11,711,507	(90.3%)	959,160	(7.4%)	12,965,192
(9)	Amazon	35,941	(0.6%)	85	(0.0%)	6,125,713	(97.3%)	130,728	(2.1%)	6,292,467
(10)	NetDNA	0	(0.0%)	0	(0.0%)	677748	(100.0%)	3	(0.0%)	677,751
	All	**332,835**	**(0.7%)**	**3,609**	**(0.0%)**	**25,225,360**	**(50.0%)**	**24,885,884**	**(49.3%)**	**50,447,688**

The percentage shows the relative frequency of a TLS version for a specific provider. Some CDIs such Wordpress, Facebook, or Google use TLS 1.3 for more than 95.0% of their delivered resources.

Further, Amazon's CDI is also seen to be used in many ads and tracking samples: Around 22.2% of the ads delivered by Amazon (<1% for trackers) are served in the context of their marketplace via `amazon-adsystem.com` and `amazon.com`, e.g., for products sold on the Amazon online store; the remaining ads and trackers are delivered via URLs of customers of Amazon's CDI that use Amazon Web Services (AWS), for instance. This potentially includes advertisers that rent AWS capacity to easily deploy AWS instances, so that they can quickly distribute ads and trackers to end-users to counteract being blocked by popular ad-blockers [62, 82, 91].

Overall, these results indicate a substantial contribution by Google and Amazon to the domain of ads and trackers, although non-CDI providers also account for a significant share. This observation suggests that consolidation occurs at other levels beyond the delivery of what is considered "actual" Web content, in particular as CDIs can make use of their already existing and sophisticated infrastructure close to end-users for the delivery of advertisements. However, note that dedicated platforms such as Facebook might show a much different perspective when going deeper below the surface Web [7], as mentioned above.

5.3 TLS 1.3 Adoption

The adoption of novel protocols on the Internet is typically a lengthy process after standardization. Technical and administrative consolidation presents an opportunity to facilitate the deployment and support of standards and novel protocols [108]. Holz et al. [55] study the deployment of TLS 1.3 and show the impact of large providers on protocol deployment [94]. They find that CDIs such as Google, Cloudflare, and Facebook are main drivers of TLS 1.3 deployment, as those CDIs quickly adopt changes between different standard drafts and apply updates to a wide range of domains.

We complement their analysis with a case study on TLS 1.3 support among webpage resources based on the active measurement data from the HTTP Archive (January 2020), which contains information on the used version of TLS, i.e., TLS 1.0, 1.1, 1.2, and 1.3. Table 5 lists the number of resources (absolute and relative frequency) for which the TLS version was identified by `WebPagetest`. We only find information on the TLS version for 50.4 M resources (12.9%); For the remaining 341.9 M resources, 64.1 M are requested over unencrypted HTTP, meaning that the SSL/TLS version is not captured by `WebPagetest` for 277.8 M resources (70.8%).

CDIs can deploy new protocols such as TLS 1.3 at a large scale. Considering all 50.4 M resources for which the TLS version is identified, we see that 25.2 M (50.0%) are using TLS 1.2, whereas 24.9 M (49.3%) are using TLS 1.3. In comparison with [55], we find similar patterns but different numbers: We find Google to account for 59.9% of all TLS 1.3 secured resources (50.0% of all TLS 1.3 connections in [55]). Similarly, Facebook is responsible for the second highest share with 12.3% (26.8% in [55]). Consistent with their ranking, we observe Cloudflare in third spot with 8.4% (6.9% in [55]) of all TLS 1.3 resources, which supports the findings by Holz et al. [55] from May 2019. However, note that both their monitoring data [55] and the HTTP Archive data are measured from California, the United States, which may introduce some regional bias in the results of both studies.

In the HTTP Archive data, we notice that some of the larger CDIs use TLS 1.3 almost exclusively: WordPress (100%), Facebook (>99.9%), and Google (95.0%) deliver nearly all of their TLS-identified resources via TLS 1.3. In comparison, Cloudflare (82.4%), Highwinds (70.2%), Twitter (60.7%), Akamai (41.2%), and Microsoft (16.9%) have lower TLS 1.3 usage shares. Surprisingly, Fastly (7.9%) and Amazon (2.1%), and all remaining rows in Table 5 use TLS 1.3 for less than 10% of their resources served via TLS, mostly relying on TLS 1.2 instead, similar to non-CDI resources (delivered over TLS 1.3 in only 7.4% of cases).

Despite TLS 1.3 being a relatively novel protocol, we observe some CDIs to already fully support it and make it their standard TLS version. This observation indicates and corroborates that consolidation can facilitate the deployment and adoption of new standards (similar to the observation for IPv6 in Section 3.2), as the decision for adoption is typically propagated across the whole infrastructure after thorough testing.

5.4 Embedded Video Players

Video content accounts for the predominant share of Internet traffic these days [105], being delivered from dedicated large-scale CDIs. Yet, as mentioned in Section 4.2.2 and Table 2, resources of the type *video* only account of 0.3% of all measured resources in total. In most cases, video content on the Web is not directly loaded together with the base page; instead, videos are streamed on demand through embedded video players, which consist of a bundle of resources rather than a single *video* type resource. Note that the following analyses are non-exhaustive examples, as video-related resources may also be served from other streaming portals and CDIs, as well as through other URL patterns, which are not considered.

For instance, in the case of YouTube, one embedded video player typically results in a total of seven different resources delivered by Google (via `youtube.com` and `youtube-nocookie.com`), out of which four are JavaScript resources, with the other three being an HTML, text, and *other* resource type each. Across the whole dataset, we find 369 k unique base pages (8.6% of all measured pages) that embed YouTube videos as part of their page content, embedding a total of 732 k videos from YouTube, whose embedded players are initially served by Google through roughly 5.1 M resources.

A comparison with other video content, such as videos delivered by Facebook, is more difficult, as a straight forward separation between video-related resources and resources related to ads or tracking (see Section 5.2) is less clear. Moreover, Facebook does not only have one dedicated service for video content; instead, they serve videos for multiple of their popular services, such as their social networking service Facebook, or Instagram, among others. As shown in Table 3, 59.2% of the resources of the type *video* are delivered by Facebook (732 k resources), which also covers videos from `cdninstagram.com`, for example. Considering all resources regardless of resource type, we find roughly 662.2 k resources from `facebook.com` and `fbcdn.net` (i.e., from their social networking service) that include `/videos/`, `/video.php?`, or `.mp4` in the resource URL. Out of these, 650 k are classified as type *video* across 37 k base pages overall. Thus, although the number of videos are similar between YouTube and Facebook, the data suggests that videos hosted by Facebook are much less common on the surface Web due to the much lower number of base pages.

As another example for reference, we investigate videos from Vimeo, another popular video on demand streaming platform. We identify Vimeo to deliver content via `vimeocdn.com` and `player.vimeo.com`, which are primarily mapped to Fastly's CDI in the dataset. Videos from Vimeo are only included in 58.4 k unique base pages (1.4% of all measured pages) through roughly 242.4 k resources (primarily HTML, image, and JavaScript resources). Thus, Vimeo shows a much lower presence in comparison with YouTube.

In total, these observations indicate that there is a moderately high dependency of webpages on video content from several CDIs. In particular, videos are often used to enrich a webpage's content; as shown, such videos are most frequently delivered by CDIs such as Google/YouTube and Facebook. An outage of either CDI could cause major parts of the content on a webpage to be missing, resulting in degraded user experience. Thus, replication and high availability provided by dedicated CDIs is essential for video content, especially considering its popularity and volume these days. Similarly, serving a video from a non-CDI host might further impact the user experience negatively, e.g., due to longer loading times and stalls.

5.5 Webpage Performance Metrics

Lastly, we investigate whether hosting a webpage and its resources on a CDI compared to non-CDI hosting has significant differences on the webpage's performance. As mentioned before, the `WebPagetest` [114] records various timestamps for the different navigation timings [112] of the webpage to measure its performance [34]. In this case study, we focus on `domInteractive` timings, i.e., the time until the user is able to interact with the webpage, as well as `visualComplete` timings, i.e., the time until the website is fully rendered above-the-fold by the browser. We choose these metrics for discussion for the performance measurements, as they represent an earlier (`domInteractive`) and a later metric (`visualComplete`) in the webpage parsing and rendering process [112]. However, note that we also compare other navigation timings, namely, **Time to First Byte** (`ttfb`), `visualComplete[85|90|95]` (page visually complete to 85% |90% |95%), time to **First Contentful Paint** (**FCP**), and time to **First Meaningful Paint** (**FMP**), which leads to the same results and are, therefore, omitted from the discussion for the sake of brevity.

For this case study, we first distinguish base pages hosted on CDIs and base pages not hosted on CDIs. We additionally filter base pages with invalid performance measurements, which leaves 1.27 M pages hosted on CDIs and 2.76 M pages not hosted on CDIs for the overall samples. We then *log*-transform the timing data for `domInteractive` as well as `visualComplete` in each of these two groups to normalize the distribution. For the comparison of the group means, we finally apply Welch's t-tests, receiving p-values of $p < 1\text{e-}3$ for the tested metrics, which indicate highly significant differences between the `domInteractive` timings and the `visualComplete` timings of the groups. Therefore, the dataset indicates significant performance differences between base pages hosted on CDIs and base pages not hosted on CDIs (in favor of CDI-hosted pages).

Furthermore, we check whether the calculated CDI penetration of a webpage (cf. Figure 5) correlates with any of the performance metrics. The correlation coefficients r for the CDI penetration and individual performance metrics are close to 0, ranging from −0.081 and −0.012 for CDI-hosted base pages and −0.069 to −0.002 for non-CDI-hosted base pages. Thus, the dataset suggests no correlation between the relative CDI penetration of a page and its performance metrics, likely due to page complexity. Yet, the values for CDI penetration are significantly different between the two groups ($p < 1\text{e-}3$), as indicated by Figure 5.

In conclusion, the CDI penetration per page itself, i.e., the percentage of CDI-hosted resources relative to all resources, shows no correlation with the performance of the website, suggesting that solely including a higher relative number of resources hosted on CDIs does not affect the performance of the website as whole. Nevertheless, the t-tests indicate that there are significant differences for CDI-hosted and non-CDI-hosted base pages, implying that the former likely employ (additional) methods for website optimization to reduce loading times. Such optimizations often go hand in hand with deployment on CDIs due to shifts in resource allocation: For instance, delegating the networking and hosting part of a website to a dedicated CDI can open up opportunities for website administrators to instead shift costs and focus (from maintaining and deploying the infrastructure) to the development and optimization of the Web experience.

However, with such consolidation of Web content and the studied benefits also come certain risks, such as potential single points of failure and risks to data privacy, which we further discuss in the following.

6 IMPLICATIONS

As discussed in the previous sections, our analyses of the measurement data (as well as results of related work, see Section 8) have indicated trends of increasing consolidation of Web content from different angles. While the current extent of concentration does not appear to be excessive on the surface Web, note that the analyzed measurements represent a rough lower bound (see Section 7). Nevertheless, the observed increase in CDI penetration can have various implications: Increasing consolidation of Web content around a few "hyper giants" [22, 72, 92] can bring both benefits and risks to the content delivery process, especially when considering the increasing migration of non-Web services over to HTTP [89], such as **DNS-over-HTTPS** (**DoH**) [52]. Consequently, a high dependency on CDIs should neither be considered inherently good nor bad. The benefits and risks encompass technical [116], societal [54], as well as economic [48] aspects, which also affect other layers of the Internet beyond the Web.

In terms of benefits, CDIs replicate and cache the content at multiple points-of-presence worldwide [40], often close to the edge and with direct peering agreements (between the CDI and the users' access networks), which pushes flattening of the Internet topology [11, 31, 43, 72]. These deployments can result in lower page loading times for the end user due to shorter content delivery paths (cf. Section 5.5), easier deployment of new technologies, as well as higher availability and better load balancing.

For instance, we find improvements of median times to **First Meaningful Paints** (**FMP**) from 3 s down to 2.3 s when comparing non-CDI hosted webpages to CDI-hosted ones using the HTTP Archive data. Such latency improvements (or lack thereof) can, for instance, affect bounce and conversion rates on e-commerce platforms [14, 102], which can be a main incentive when migrating to a CDI. However, note that the complexity of webpages [24, 25, 109] introduce more dependencies and properties that need to be considered, such as the rendering pipeline.

As discussed, the deployment of standards and technologies (such as IPv6, TLS 1.3, QUIC, DNS over HTTPS) can affect much larger scales thanks to CDIs, as they can enforce policies (such as the support of a new protocol) and affect a large number of sites around the globe, although note that operators of larger CDIs also majorly contribute to the standardization process itself. However, policy decisions at single points of trust and control such as a CDI can also have negative effects. In particular, the duality of CDIs acting as infrastructure providers for websites and taking the roles of content moderators is difficult to discern and causes tension as a result.

Centralization of content makes (selective) censorship easier for governments, either by forcing the CDI to block specific content or banning the CDI altogether, making a large amount of content unavailable [5, 78]. Blocking a CDI as a whole to censor a specific piece of content can cause collateral damage since it will block all other content of that CDI [4]; IP address-based blocking, in contrast, only causes low collateral damage [51]. While more sophisticated censorship approaches exist, these can be partially circumvented through domain fronting with CDIs [39, 117]. As some CDIs also police and moderate the content they serve [32, 90], their policies and recommendation algorithms can reduce the diversity of disseminated information (both true and fake information) and foster filter bubbles [81]. E.g., 25% of tweets by news outlets around the 2016 US presidential election were found to propagate false or biased information [23], which is why Twitter has made various efforts to protect elections over the years and the 2020 US Presidential Election period [41]. As another example, briefly before the 2021 Russian federal election [107], Google and Apple were enforced to remove apps of the political opposition from their app stores by authorities.

On the other hand, content replication by CDIs further improves availability, which makes it more reliable against denial of services attacks [44, 66], typically through upscaling of infrastructure. Nevertheless, consolidation around an oligopoly of CDIs can also increases the damage caused by attacks, as it reduces the points of failures in the network to that small set of CDIs, which leaves fewer to even no alternatives in case of an organization-wide attack or outage (despite potential distribution of servers around the globe). Such strong dependencies on a few providers can cause collateral damage: One example for this is the 2016 DDoS attack on DynDNS [6, 67] on October 21, 2016, which resulted in a large number of Web services being unavailable for a day across North America and parts of Europe. Such large-scale network failures can have significant consequences for businesses that rely on highly available and performant Web services provided by a CDI; e.g., unavailability of an e-commerce shop can cause a major loss of sales and, thus, revenue, especially during peak seasons. Similarly, all of Facebook's services suffered an outage of around six hours on October 04, 2021, as a combined result of propagated BGP and DNS failures, which made all Facebook data centers unreachable [38].

From an end user point of view, consolidation of Web services behind one provider comes with convenience and simplicity, as an account on one platform enables them to use a variety of services with that single account through **single sign-on** (**SSO**). Similarly, protocols such as OAuth [49] allows users to use the very same account to interact and share information with a third-party service, which reduces the overhead of managing multiple accounts and password fatigue, for instance. On the other hand, consolidating multiple services into a single one allows easier collection of sensitive user data, which poses a risk to privacy: The collection and storage of user data for millions of users at one actor can be more easily abused and might lead to privacy breaches [16, 45, 64]. In return, such data silos allows the data collector to learn from the user data in order to continue improving the service and user experience, further feeding into this loop: Bundled with a plethora of available services to choose from and high user counts, such content consolidation and similar evolutions can cause network effects, which in return draws an increasing number of users into consolidated systems. This results in a customer lock-in due to the lack of options regarding data migration and interoperability between different centralized systems. Ultimately, this creates a feedback loop that drives market concentration, which can potentially limit technical innovation due to difficulties of reaching critical mass as well as business acquisitions. As hyper giants grow larger and larger, it will become increasingly difficult for new players to grow, making it difficult to regulate the Internet ecosystem.

7 LIMITATIONS

We note that our datasets have limitations resulting from the vantage points from which the measurements have been made: The OpenINTEL datasets (see Section 3) as well as the HTTP Archive dataset (see Section 4) represent a view on CDIs from one vantage point each. Websites might employ different CDIs (or none at all) in different parts of the world; as such, we cannot consider webpages that use different CDIs in different geographical regions. Similarly, content served from CDIs may differ due to language localization for instance. Nevertheless, since we are not looking at the semantic content of the data retrieved, we believe that these limitations do not substantially impact the general conclusions of our work, as we primarily focus on the provenance of CDIs overall.

Moreover, we do not further classify the studied Web domains with different content category labels, as studies [110] have shown that domain classification can vary significantly depending on the used methodology and services. We also do not consider parked domains, i.e., domains that are only registered but not actually in use, as we specifically focus on `www.` domains that resolve to IPv4 and IPv6 addresses (Section 3), or pages with actual Web content (Section 4). Furthermore, as discussed in Section 2.2, the identification of CDIs is not exhaustive, as CDI technologies are involved at different layers internally, which makes it difficult to assess their extent and influence precisely from an outside view. For instance, CDI providers may directly peer with other ASes or host content caches inside an ISP network to bring it even closer to home users at the edge [30, 74]. Furthermore, CDIs are known to serve content in isolated environments exclusively, e.g., only within a specific AS for a specific customer. Consequently, identification of such cases would require auxiliary information at a large scale for individual ASes and CDIs.

As previously mentioned, we only take the surface Web into account, meaning that the actual CDI penetration might be much different from the reported numbers (which represent a lower bound) when considering internal pages [7]. However, we believe that our approach provides a reasonable view of CDI penetration (as a metric for Web consolidation) from multiple perspectives overall, which can be built upon by future studies.

8 RELATED WORK

Recent studies have shown Internet consolidation around a small number of players at multiple layers and from varying perspectives. Overall, the studies conclude at similar observations concerning increasing consolidation around the same set of larger players in the ecosystems as observed in our study. However, the metrics used in the studies as well as their focus are different from our work, in which we focus on CDI penetration (in terms of content hosting) as a metric for Web consolidation (i.e., the application layer) in particular. As such, our overall findings complement other studies by providing observations from different perspectives, namely, the Web in terms of landing pages and page resources.

For instance, Hoang et al. [51] study Web consolidation by using DNS measurements to evaluate website co-location regarding IP addresses and ASes. While their co-location metric based on DNS and AS information is similar to ours, they solely focus on the landing pages of Web domains over a period of two weeks, which they curate from the Alexa and Majestic 1 M toplists [100]—thus, they do not take webpage resources or measurements over a longer period of time into account. The authors perform measurements for 8.6 M domains from nine virtual private servers across the globe, observing high website co-location (and, thus, centralization) at a small number of providers as a result. They then switch their focus to implications of co-location for block lists, showing that severe collateral damage as a result of IP address filtering or censorship is unlikely, contrary to common perception. In comparison, we instead cover 166.5 M domains over a period of nearly five years in our longitudinal analysis, and study 4.3 M webpages along with 392.3 M page resources, although from fewer vantage points.

Kashaf et al. [67] study third-party dependencies of Alexa Top 100 k websites with respect to DNS, CDNs, and **Certificate Authorities (CAs)**. They compare data for the 100 k websites from 2016 and 2020, which they collect from one vantage point in the United States. They investigate dependencies of websites and their internal resources from an infrastructure point of view, similar to our study; however, their scope is on direct and indirect (transitive) dependencies along with their impact on availability due to shared attack surfaces and cascading failures. The authors observe critical amplification and dependency chains in the Alexa Top 100 k websites, where the top providers account for 50% −70% of the dependencies, indicating significant consolidation around these providers. Furthermore, they see slight increases in the dependencies when comparing their datasets from 2016 to 2020 (similar to our comparisons of 2016–2020, see Section 2.1.2). For the identification of CDNs (which partially overlaps with our study on CDIs), they use a heuristic based on various sources that consider `SOA` and `CNAME` records, along with public suffix lists, TLD matching, and subject alternate names of the SSL certificates (if existing). In contrast, our approach takes `CNAME` records, HTTP headers (via `WebPagetest`), as well as IP address prefixes and announcing ASes into account, while also including cloud services such as AWS or Microsoft Azure. Moreover, our analysis covers a larger number of webpages and their resources in total, which therefore provides a broader and simultaneously complementing view, especially given the substantial differences between websites beyond Alexa Top 100 k (see Section 3.2).

Studying centralization in the context of protocol adoption, Holz et al. [55] track TLS 1.3 before, around, and after its standardization by the IETF in August 2018 [94]. They use both active and passive measurement data covering multiple years in total in order to determine the share of TLS 1.3 relative to other TLS versions. They find that the TLS 1.3 deployment is mainly and rapidly driven by larger players such as Cloudflare and Google. Nevertheless, their analysis on the usage of TLS 1.3 by servers (and clients) relies on passive monitoring data, which they capture from North American research networks from 2012 to 2019. As such, the dataset also includes TLS usage for applications such as e-mail or VoIP, among others. Furthermore, their centralization analysis focuses on the second half of 2019. In contrast, our case study (Section 5.3) builds upon active measurements from the HTTP Archive (also located in the United States) at a more recent point in time (January 2020), and focuses on Web content exclusively.

Other recent studies also observe similar trends of centralization through vastly different perspectives and metrics: For instance, Böttger et al. [22] use port capacity and traffic profile properties from PeeringDB [87] to come up with a classification for organizations as “hyper giants” [72] at the network layer. They find that hyper giants further leverage the IXP ecosystem to achieve global reachability in terms of the IP address space. The vast expansion of these hyper giants in recent years is shown by Gigis et al. [43], who find that the hyper giants’ off-nets have tripled from 2013 to 2021, and thus, are able to reach large numbers of end users. Similarly, Todd et al. [11] study the AS-level topology of the Internet, finding that large cloud providers increasingly contribute to the flattening of the Internet. These cloud providers surpass Tier-1 and Tier-2 ISPs in terms of reachability, meaning that other networks (>76%) can reach these cloud providers without traversing any Tier-1 or Tier-2 ISPs. As a result, cloud providers are largely independent from other networks, which suggests consolidation at the routing level. At the same time, this high independence of cloud providers increases the resilience of networks against route leaks, which are often propagated by Tier-1 ISPs. Likewise, Moura et al. [80] observe centralization in the DNS, where they find cloud providers to be responsible for a significant fraction of the DNS traffic between recursive resolvers and authoritative servers. They further also find that centralization facilitates the deployment of new features (similar to TLS 1.3), such as **Query Name** (**QNAME**) Minimization [21], highlighting benefits of consolidation.

9 CONCLUSION

In this study, we analyzed a set of datasets (based on Web domains from the `.com`/`.net`/`.org` TLDs, the Alexa Top 1 M, and Google CrUX) from two distinct measurement platforms (OpenINTEL and HTTP Archive) to empirically investigate Web consolidation through CDI penetration, which covers Web hosting infrastructures such as clouds, CDNs, and DDoS protection infrastructures. Across these datasets, we noticed increasing consolidation around the same CDIs for most CDI-hosted Web content, and supported as well as complemented findings of previous studies to provide an updated discussion on the implications of Web content consolidation. Considering all (more than 140 M) `.com`, `.net`, and `.org` domains, we found that the CDI penetration has nearly doubled from 8.2% to 15% since 2015. For popular webpages based on Alexa Top 1 M and 4.3 M webpages based on Google Chrome User Experience (CrUX), we observed a CDI penetration of 24%–32%. Although 43.4% of the 392.3 M page resources were not delivered by CDIs, we determined a high dependency of webpages on CDIs for fonts and JavaScript, primarily served by Google. In particular, we also observed a moderate dependency of webpages on CDIs regarding jQuery along with ads and trackers.

Our findings highlight the importance of CDIs for Web content, even though content on the surface Web is currently not massively consolidated. Nevertheless, while concerns about consolidation (e.g., loss of control, privacy, competition, and innovation) are currently broadly discussed and considered [1, 9, 10, 17–19, 59, 63], the associated benefits should also be noted: Consolidation can drive the deployment of new standards (as seen by IPv6 or TLS 1.3), user convenience, content availability, and provide avenues for loading time improvements, among other aspects.

Our results revealed the majority of CDI-hosted content coming from an oligopoly of CDIs, which dominate the hosting of Web content. As such, concerns about a consolidated Internet might become reality in the future, should these trends continue. Potential solutions include pushing decentralized solutions or leveraging multiple CDIs simultaneously [20, 53, 101]: This allows content providers to distribute data across multiple CDIs to receive the associated performance benefits, while also counteracting consolidation around and dependence on single companies at the same time. However, note that decentralized Web solutions do not guarantee avoiding consolidation [93] and come with other challenges, such as content moderation [50]. Thus, follow-up studies are required to illuminate the extent and impact of Web consolidation beyond the surface Web.

ACKNOWLEDGMENTS

We thank Justus Fries and Elisa Xiao (Technical University of Munich (TUM)) for their contributions to early explorations of ideas for this study. We also thank the anonymous reviewers, Jörg Ott (TUM), and Leonardo Tonetto (TUM) for their insightful feedback and suggestions.

REFERENCES

[1] Alexander Afanasyev and Matthias Wählisch (Eds.), 2021. In *Proceedings of the Interdisciplinary Workshop on (de) Centralization in the Internet.* ACM. DOI:https://doi.org/10.1145/3488663

[2] Bernhard Ager, Wolfgang Mühlbauer, Georgios Smaragdakis, and Steve Uhlig. 2011. Web content cartography. In *Proceedings of the 11th ACM SIGCOMM Internet Measurement Conference.* 585–600. DOI:https://doi.org/10.1145/2068816.2068870

[3] Alexa. 2021. Top Sites. Retrieved December 9, 2021 from https://www.alexa.com/topsites.

[4] Anon. 2012. The collateral damage of internet censorship by DNS injection. *Computer Communication Review* 42, 3 (2012), 21–27. DOI:https://doi.org/10.1145/2317307.2317311

[5] Anonymous. 2012. The collateral damage of Internet censorship by DNS injection. *Computer Communication Review* 42, 3 (2012), 21–27. DOI:https://doi.org/10.1145/2317307.2317311

[6] Manos Antonakakis, Tim April, Michael Bailey, Matt Bernhard, Elie Bursztein, Jaime Cochran, Zakir Durumeric, J. Alex Halderman, Luca Invernizzi, Michalis Kallitsis, Deepak Kumar, Chaz Lever, Zane Ma, Joshua Mason, Damian Menscher, Chad Seaman, Nick Sullivan, Kurt Thomas, and Yi Zhou. 2017. Understanding the mirai botnet. In *Proceedings of the 26th USENIX Security Symposium.* Engin Kirda and Thomas Ristenpart (Eds.). USENIX Association, 1093–1110. Retrieved from https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/antonakakis.

[7] Waqar Aqeel, Balakrishnan Chandrasekaran, Anja Feldmann, and Bruce M. Maggs. 2020. On landing and Internal web pages: The strange case of jekyll and hyde in web performance measurement. In *Proceedings of the ACM Internet Measurement Conference.* ACM, 680–695. DOI:https://doi.org/10.1145/3419394.3423626

[8] Jari Arkko. 2019. *Centralised Architectures in Internet Infrastructure.* Internet-Draft Draft-arkko-arch-infrastructure-centralisation-00. Internet Engineering Task Force. Retrieved December 9, 2021 from https://datatracker.ietf.org/doc/html/draft-arkko-arch-infrastructure-centralisation-00.

[9] Jari Arkko, Mark Nottingham, Christian Huitema, Martin Thomson, and Brian Trammell. 2017. IETF news: Consolidation. *Internet Architecture Board.* Retrieved December 9, 2021 from https://www.ietf.org/blog/consolidation/.

[10] Jari Arkko, Brian Trammell, Mark Nottingham, Christian Huitema, Martin Thomson, Jeff Tantsura, and Niels ten Oever. 2019. *Considerations on Internet Consolidation and the Internet Architecture.* Internet-Draft Draft-arkko-iab-internet-consolidation-02. Internet Engineering Task Force. Retrieved December 9, 2021 from https://datatracker.ietf.org/doc/html/draft-arkko-iab-internet-consolidation-02.

[11] Todd Arnold, Jia He, Weifan Jiang, Matt Calder, Ítalo Cunha, Vasileios Giotsas, and Ethan Katz-Bassett. 2020. Cloud provider connectivity in the flat Internet. In *Proceedings of the ACM Internet Measurement Conference.* ACM, 230–246. DOI: https://doi.org/10.1145/3419394.3423613

[12] Vaibhav Bajpai, Olivier Bonaventure, Kimberly C. Claffy, and Daniel Karrenberg. 2018. Encouraging reproducibility in scientific research of the Internet (dagstuhl seminar 18412). *Dagstuhl Reports* 8, 10 (2018), 41–62. DOI: https://doi.org/10.4230/DagRep.8.10.41

[13] Vaibhav Bajpai, Anna Brunström, Anja Feldmann, Wolfgang Kellerer, Aiko Pras, Henning Schulzrinne, Georgios Smaragdakis, Matthias Wählisch, and Klaus Wehrle. 2019. The dagstuhl beginners guide to reproducibility for experimental networking research. *Computer Communication Review* 49, 1 (2019), 24–30. DOI: https://doi.org/10.1145/3314212.3314217

[14] Marcus Basalla, Johannes Schneider, Martin Luksik, Roope Jaakonmäki, and Jan vom Brocke. 2021. On latency of e-commerce platforms. *Journal of Organizational Computing and Electronic Commerce* 31, 1 (2021), 1–17. DOI: https://doi.org/10.1080/10919392.2021.1882240

[15] Muhammad Ahmad Bashir and Christo Wilson. 2018. Diffusion of user tracking data in the online advertising ecosystem. *Proceedings on Privacy Enhancing Technologies* 2018, 4 (2018), 85–103. DOI: https://doi.org/10.1515/popets-2018-0033

[16] Hal Berghel. 2018. Malice domestic: The cambridge analytica dystopia. *IEEE Computer* 51, 5 (2018), 84–89. DOI: https://doi.org/10.1109/MC.2018.2381135

[17] Tim Berners-Lee. 2017. Three challenges for the web, according to its inventor. *World Wide Web Foundation.* Retrieved December 9, 2021 from https://webfoundation.org/2017/03/web-turns-28-letter/.

[18] Tim Berners-Lee. 2018. The web is under threat. Join us and fight for it. *World Wide Web Foundation.* Retrieved December 9, 2021 from https://webfoundation.org/2018/03/web-birthday-29/.

[19] Tim Berners-Lee. 2019. World wide web foundation: 30 years on, what's next #ForTheWeb? *World Wide Web Foundation.* Retrieved December 9, 2021 from https://webfoundation.org/2019/03/web-birthday-30/.

[20] Jeremias Blendin, Fabrice Bendfeldt, Ingmar Poese, Boris Koldehofe, and Oliver Hohlfeld. 2018. Dissecting apple's meta-CDN during an iOS update. In *Proceedings of the Internet Measurement Conference 2018.* ACM, 408–414. DOI: https://doi.org/10.1145/3278532.3278567

[21] Stephane Bortzmeyer. 2016. DNS query name minimisation to improve privacy. *RFC* 7816 (2016), 1–11. DOI: https://doi.org/10.17487/RFC7816

[22] Timm Böttger, Félix Cuadrado, and Steve Uhlig. 2018. Looking for hypergiants in PeeringDB. *Computer Communication Review* 48, 3 (2018), 13–19. DOI: https://doi.org/10.1145/3276799.3276801

[23] Alexandre Bovet and Hernán A. Makse. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications* 10, 1 (2019), 1–14. DOI: https://doi.org/10.1038/s41467-018-07761-2

[24] Michael Butkiewicz, Harsha V. Madhyastha, and Vyas Sekar. 2011. Understanding website complexity: Measurements, metrics, and implications. In *Proceedings of the 11th ACM SIGCOMM Internet Measurement Conference.* Patrick Thiran and Walter Willinger (Eds.). ACM, 313–328. DOI: https://doi.org/10.1145/2068816.2068846

[25] Michael Butkiewicz, Harsha V. Madhyastha, and Vyas Sekar. 2014. Characterizing web page complexity and its impact. *IEEE/ACM Transactions on Networking* 22, 3 (2014), 943–956. DOI: https://doi.org/10.1109/TNET.2013.2269999

[26] Aaron Cahn, Scott Alfeld, Paul Barford, and S. Muthukrishnan. 2016. An empirical study of web cookies. In *Proceedings of the 25th International Conference on World Wide Web.* Jacqueline Bourdeau, Jim Hendler, Roger Nkambou, Ian Horrocks, and Ben Y. Zhao (Eds.). ACM, 891–901. DOI: https://doi.org/10.1145/2872427.2882991

[27] CAIDA. 2021. Routeviews Prefix to AS Mappings Dataset (pfx2as) for IPv4 and IPv6. Retrieved December 9, 2021 from https://www.caida.org/data/routing/routeviews-prefix2as.xml.

[28] Matt Calder, Ashley Flavel, Ethan Katz-Bassett, Ratul Mahajan, and Jitendra Padhye. 2015. Analyzing the performance of an anycast CDN. In *Proceedings of the 2015 ACM Internet Measurement Conference.* Kenjiro Cho, Kensuke Fukuda, Vivek S. Pai, and Neil Spring (Eds.). ACM, 531–537. DOI: https://doi.org/10.1145/2815675.2815717

[29] Fangfei Chen, Ramesh K. Sitaraman, and Marcelo Torres. 2015. End-user mapping: Next generation request routing for content delivery. In *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication.* Steve Uhlig, Olaf Maennel, Brad Karp, and Jitendra Padhye (Eds.). ACM, 167–181. DOI: https://doi.org/10.1145/2785956.2787500

[30] Yi-Ching Chiu, Brandon Schlinker, Abhishek Balaji Radhakrishnan, Ethan Katz-Bassett, and Ramesh Govindan. 2015. Are we one hop away from a better Internet?. In *Proceedings of the 2015 ACM Internet Measurement Conference.* Kenjiro Cho, Kensuke Fukuda, Vivek S. Pai, and Neil Spring (Eds.). ACM, 523–529. DOI:https://doi.org/10.1145/2815675.2815719

[31] Lorenzo Corneo, Maximilian Eder, Nitinder Mohan, Aleksandr Zavodovski, Suzan Bayhan, Walter Wong, Per Gunningberg, Jussi Kangasharju, and Jörg Ott. 2021. Surrounded by the clouds: A comprehensive cloud reachability study. In *Proceedings of the Web Conference 2021.* Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia (Eds.). ACM / IW3C2, 295–304. DOI:https://doi.org/10.1145/3442381.3449854

[32] Rachel Dunphy. 2017. Can YouTube survive the adpocalypse? *New York Magazine.* Retrieved December 9, 2021 from http://nymag.com/intelligencer/2017/12/can-youtube-survive-the-adpocalypse.html.

[33] EasyList. 2021. EasyList - Overview. Retrieved December 9, 2021 from https://easylist.to/.

[34] Theresa Enghardt, Thomas Zinner, and Anja Feldmann. 2019. Web performance pitfalls. In *Proceedings of the 20th International Conference on Passive and Active Network Measurement.* David R. Choffnes and Marinho P. Barcellos (Eds.), Lecture Notes in Computer Science, Vol. 11419. Springer, 286–303. DOI:https://doi.org/10.1007/978-3-030-15986-3_19

[35] Steven Englehardt and Arvind Narayanan. 2016. Online tracking: A 1-million-site measurement and analysis. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security.* Edgar R. Weippl, Stefan Katzenbeisser, Christopher Kruegel, Andrew C. Myers, and Shai Halevi (Eds.). ACM, 1388–1401. DOI:https://doi.org/10.1145/2976749.2978313

[36] European Commission. 2021. A Europe Fit for the Digital Age: Empowering People with a New Generation of Technologies. Retrieved December 9, 2021 from https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age_en.

[37] European Commission. 2021. The Digital Europe Programme: Shaping Europe's Digital Future. Retrieved December 9, 2021 from https://digital-strategy.ec.europa.eu/en/activities/digital-programme.

[38] Facebook Engineering. 2021. More Details about the October 4 Outage. Retrieved December 9, 2021 from https://engineering.fb.com/2021/10/05/networking-traffic/outage-details/.

[39] David Fifield, Chang Lan, Rod Hynes, Percy Wegmann, and Vern Paxson. 2015. Blocking-resistant communication through domain fronting. *PoPETs* 2015, 2 (2015), 46–64. DOI:https://doi.org/10.1515/popets-2015-0009

[40] Marcel Flores and Harkeerat Bedi. 2019. Caching the Internet: A view from a global multi-tenant CDN. In *Proceedings of the 20th International Conference on Passive and Active Network Measurement.* David R. Choffnes and Marinho P. Barcellos (Eds.), Lecture Notes in Computer Science, Vol. 11419. Springer, 68–81. DOI:https://doi.org/10.1007/978-3-030-15986-3_5

[41] Vijaya Gadde and Kayvon Beykpour. 2020. Additional Steps we're Taking Ahead of the 2020 US Election. Retrieved December 9, 2021 from https://blog.twitter.com/en_us/topics/company/2020/2020-election-changes.html.

[42] Alessandro Ghedini and Rustam Lalkaka. 2019. HTTP/3: The past, the present, and the future. *The Cloudflare Blog.* Retrieved December 9, 2021 from https://blog.cloudflare.com/http3-the-past-present-and-future/.

[43] Petros Gigis, Matt Calder, Lefteris Manassakis, George Nomikos, Vasileios Kotronis, Xenofontas A. Dimitropoulos, Ethan Katz-Bassett, and Georgios Smaragdakis. 2021. Seven years in the life of hypergiants' off-nets. In *Proceedings of the ACM SIGCOMM 2021 Conference.* Fernando A. Kuipers and Matthew C. Caesar (Eds.). ACM, 516–533. DOI:https://doi.org/10.1145/3452296.3472928

[44] Yossi Gilad, Amir Herzberg, Michael Sudkovitch, and Michael Goberman. 2016. CDN-on-demand: An affordable DDoS defense via untrusted clouds. In *Proceedings of the 23rd Annual Network and Distributed System Security Symposium.* The Internet Society. Retrieved from https://www.ndss-symposium.org/wp-content/uploads/2017/09/cdn-on-demand-affordable-ddos-defense-via-untrusted-clouds.pdf.

[45] Felipe González, Yihan Yu, Andrea Figueroa, Claudia López, and Cecilia R. Aragon. 2019. Global reactions to the cambridge analytica scandal: A cross-language social media study. In *Companion Proceedings of the 2019 World Wide Web Conference.* Sihem Amer-Yahia, Mohammad Mahdian, Ashish Goel, Geert-Jan Houben, Kristina Lerman, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (Eds.). ACM, 799–806. DOI:https://doi.org/10.1145/3308560.3316456

[46] Google. 2021. Chrome User Experience Report. Retrieved December 9, 2021 from https://developers.google.com/web/tools/chrome-user-experience-report/.

[47] Google Fonts. 2020. Frequently Asked Questions. Retrieved December 9, 2021 from https://developers.google.com/fonts/faq#what_does_using_the_google_fonts_api_mean_for_the_privacy_of_my_users.

[48] Éric Gourdin, Patrick Maillé, Gwendal Simon, and Bruno Tuffin. 2017. The economics of CDNs and their impact on service fairness. *IEEE Transactions on Network and Service Management* 14, 1 (2017), 22–33. DOI:https://doi.org/10.1109/TNSM.2017.2649045

[49] Dick Hardt. 2012. The OAuth 2.0 authorization framework. *RFC* 6749 (2012), 1–76. DOI:https://doi.org/10.17487/RFC6749

[50] Anaobi Ishaku Hassan, Aravindh Raman, Ignacio Castro, Haris Bin Zia, Emiliano De Cristofaro, Nishanth Sastry, and Gareth Tyson. 2021. Exploring content moderation in the decentralised web: The pleroma case. In *Proceedings of the 17th International Conference on Emerging Networking Experiments and Technologies.* Georg Carle and Jörg Ott (Eds.). ACM, 328–335. DOI:https://doi.org/10.1145/3485983.3494838

[51] Nguyen Phong Hoang, Arian Akhavan Niaki, Michalis Polychronakis, and Phillipa Gill. 2020. The web is still small after more than a decade. *ACM SIGCOMM Computer Communication Review* 50, 2 (2020), 24–31. DOI:https://doi.org/10.1145/3402413.3402417

[52] Paul E. Hoffman and Patrick McManus. 2018. DNS queries over HTTPS (DoH). *RFC* 8484 (2018), 1–21. DOI:https://doi.org/10.17487/RFC8484

[53] Oliver Hohlfeld, Jan Rüth, Konrad Wolsing, and Torsten Zimmermann. 2018. Characterizing a Meta-CDN. In *Proceedings of the 19th 19th International Conference on Passive and Active Network Measurement.* Robert Beverly, Georgios Smaragdakis, and Anja Feldmann (Eds.), Lecture Notes in Computer Science, Vol. 10771. Springer, 114–128. DOI:https://doi.org/10.1007/978-3-319-76481-8_9

[54] John Holowczak and Amir Houmansadr. 2015. CacheBrowser: Bypassing chinese censorship without proxies using cached content. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security.*, Indrajit Ray, Ninghui Li, and Christopher Kruegel (Eds.). ACM, 70–83. DOI:https://doi.org/10.1145/2810103.2813696

[55] Ralph Holz, Jens Hiller, Johanna Amann, Abbas Razaghpanah, Thomas Jost, Narseo Vallina-Rodriguez, and Oliver Hohlfeld. 2020. Tracking the deployment of TLS 1.3 on the web: A story of experimentation and centralization. *Computer Communication Review* 50, 3 (2020), 3–15. DOI:https://doi.org/10.1145/3411740.3411742

[56] HTTP Archive. 2020. Frequently Asked Questions. Retrieved December 9, 2021 from https://httparchive.org/faq#what-changes-have-been-made-to-the-test-environment-that-might-affect-the-data.

[57] HTTP Archive. 2021. HTTP Archive. Retrieved December 9, 2021 from https://httparchive.org/.

[58] Internet Architecture Board. 2019. Design Expectations vs. Deployment Reality in Protocol Development Workshop 2019. Retrieved December 9, 2021 from https://www.iab.org/activities/workshops/dedr-workshop/.

[59] Internet Society. 2019. Consolidation in the Internet Economy. Retrieved December 9, 2021 from https://future.internetsociety.org/2019/.

[60] Internet Society. 2019. Internet Society Launches Research Project to Understand the Effects of Consolidation in the Internet Economy. Retrieved December 9, 2021 from https://www.internetsociety.org/news/press-releases/2019/internet-society-launches-research-project-to-understand-the-effects-of-consolidation-in-the-internet-economy/.

[61] Internet Society. 2020. Internet Society 2020 Action Plan. Retrieved December 9, 2021 from https://www.internetsociety.org/action-plan/2020/.

[62] Umar Iqbal, Zubair Shafiq, and Zhiyun Qian. 2017. The Ad wars: Retrospective measurement and analysis of anti-adblock filter lists. In *Proceedings of the 2017 Internet Measurement Conference.*, Steve Uhlig and Olaf Maennel (Eds.). ACM, 171–183. DOI:https://doi.org/10.1145/3131365.3131387

[63] IRTF. 2021. Decentralized Internet Infrastructure Research Group (DINRG). Retrieved December 9, 2021 from https://irtf.org/dinrg.

[64] Jim Isaak and Mina J. Hanna. 2018. User data privacy: Facebook, cambridge analytica, and privacy protection. *IEEE Computer* 51, 8 (2018), 56–59. DOI:https://doi.org/10.1109/MC.2018.3191268

[65] Quentin Jacquemart, Clément Pigout, and Guillaume Urvoy-Keller. 2019. Inferring the deployment of top domains over public clouds using DNS data. In *Proceedings of the Network Traffic Measurement and Analysis Conference.* Stefano Secci, Isabelle Chrisment, Marco Fiore, Lionel Tabourier, and Keun-Woo Lim (Eds.). IEEE, 57–64. DOI:https://doi.org/10.23919/TMA.2019.8784472

[66] Mattijs Jonker, Anna Sperotto, Roland van Rijswijk-Deij, Ramin Sadre, and Aiko Pras. 2016. Measuring the adoption of DDoS protection services. In *Proceedings of the 2016 ACM on Internet Measurement Conference.*, Phillipa Gill, John S. Heidemann, John W. Byers, and Ramesh Govindan (Eds.). ACM, 279–285. DOI:https://doi.org/10.1145/2987443.2987487

[67] Aqsa Kashaf, Vyas Sekar, and Yuvraj Agarwal. 2020. 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.* ACM, 634–647. DOI:https://doi.org/10.1145/3419394.3423664

[68] Makena Kelly. 2020. Tech's Four Biggest Companies are Going on Trial. Retrieved December 9, 2021 from https://www.theverge.com/2020/7/28/21344920/big-tech-ceo-antitrust-hearing-apple-facebook-amazon-google-facebook.

[69] Eiji Kitamura. 2020. Gaining Security and Privacy by Partitioning the Cache. Retrieved December 9, 2021 from https://developers.google.com/web/updates/2020/10/http-cache-partitioning.

[70] Balachander Krishnamurthy, Craig E. Wills, and Yin Zhang. 2001. On the use and performance of content distribution networks. In *Proceedings of the 1st ACM SIGCOMM Internet Measurement Workshop.* 169–182. DOI:https://doi.org/10.1145/505202.505224

[71] Deepak Kumar, Zane Ma, Zakir Durumeric, Ariana Mirian, Joshua Mason, J. Alex Halderman, and Michael Bailey. 2017. Security challenges in an increasingly tangled web. In *Proceedings of the 26th International Conference on World Wide Web.* Rick Barrett, Rick Cummings, Eugene Agichtein, and Evgeniy Gabrilovich (Eds.). ACM, 677–684. DOI:https://doi.org/10.1145/3038912.3052686

[72] Craig Labovitz, Scott Iekel-Johnson, Danny McPherson, Jon Oberheide, and Farnam Jahanian. 2010. Internet interdomain traffic. In *Proceedings of the ACM SIGCOMM 2010 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications.* Shivkumar Kalyanaraman, Venkata N. Padmanabhan, K. K. Ramakrishnan, Rajeev Shorey, and Geoffrey M. Voelker (Eds.). ACM, 75–86. DOI:https://doi.org/10.1145/1851182.1851194

[73] Adam Lerner, Anna Kornfeld Simpson, Tadayoshi Kohno, and Franziska Roesner. 2016. Internet jones and the raiders of the lost trackers: An archaeological study of web tracking from 1996 to 2016. In *Proceedings of the 25th USENIX Security Symposium.* Thorsten Holz and Stefan Savage (Eds.). USENIX Association. Retrieved from https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/lerner.

[74] Zhenyu Li, Donghui Yang, Zhenhua Li, Chunjing Han, and Gaogang Xie. 2018. Mobile content hosting infrastructure in China: A view from a cellular ISP. In *Proceedings of the 19th International Conference on Passive and Active Measurement*. Robert Beverly, Georgios Smaragdakis, and Anja Feldmann (Eds.), Lecture Notes in Computer Science, Vol. 10771. Springer, 100–113. DOI:https://doi.org/10.1007/978-3-319-76481-8_8

[75] Enze Liu, Gautam Akiwate, Mattijs Jonker, Ariana Mirian, Stefan Savage, and Geoffrey M. Voelker. 2021. Who's got your mail? Characterizing mail service provider usage. In *Proceedings of the ACM Internet Measurement Conference.* Dave Levin, Alan Mislove, Johanna Amann, and Matthew Luckie (Eds.). ACM, 122–136. DOI:https://doi.org/10.1145/3487552.3487820

[76] Wenrui Ma and Haitao Xu. 2021. A study of the partnership between advertisers and publishers. In *Proceedings of the 22nd International Conference on Passive and Active Measurement* Oliver Hohlfeld, Andra Lutu, and Dave Levin (Eds.), Lecture Notes in Computer Science, Vol. 12671. Springer, 564–580. DOI:https://doi.org/10.1007/978-3-030-72582-2_33

[77] Bruce M. Maggs and Ramesh K. Sitaraman. 2015. Algorithmic nuggets in content delivery. *Computer Communication Review* 45, 3 (2015), 52–66. DOI:https://doi.org/10.1145/2805789.2805800

[78] Allison McDonald, Matthew Bernhard, Luke Valenta, Benjamin VanderSloot, Will Scott, Nick Sullivan, J. Alex Halderman, and Roya Ensafi. 2018. 403 forbidden: A global view of CDN geoblocking. In *Proceedings of the Internet Measurement Conference.* ACM, 218–230. DOI:https://doi.org/10.1145/3278532.3278552

[79] Hassan Metwalley, Stefano Traverso, Marco Mellia, Stanislav Miskovic, and Mario Baldi. 2015. The online tracking horde: A view from passive measurements. In *Proceedings of the 7th International Workshop on Traffic Monitoring and Analysis.* Moritz Steiner, Pere Barlet-Ros, and Olivier Bonaventure (Eds.), Lecture Notes in Computer Science, Vol. 9053. Springer, 111–125. DOI:https://doi.org/10.1007/978-3-319-17172-2_8

[80] Giovane C. M. Moura, Sebastian Castro, Wes Hardaker, Maarten Wullink, and Cristian Hesselman. 2020. Clouding up the Internet: How centralized is DNS traffic becoming?. In *Proceedings of the ACM Internet Measurement Conference.* ACM, 42–49. DOI:https://doi.org/10.1145/3419394.3423625

[81] Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren G. Terveen, and Joseph A. Konstan. 2014. Exploring the filter bubble: The effect of using recommender systems on content diversity. In *Proceedings of the 23rd International World Wide Web Conference.* Chin-Wan Chung, Andrei Z. Broder, Kyuseok Shim, and Torsten Suel (Eds.). ACM, 677–686. DOI:https://doi.org/10.1145/2566486.2568012

[82] Rishab Nithyanand, Sheharbano Khattak, Mobin Javed, Narseo Vallina-Rodriguez, Marjan Falahrastegar, Julia E. Powles, Emiliano De Cristofaro, Hamed Haddadi, and Steven J. Murdoch. 2016. Adblocking and counter blocking: A slice of the arms race. In *Proceedings of the 6th USENIX Workshop on Free and Open Communications on the Internet.* Amir Houmansadr and Prateek Mittal (Eds.). USENIX Association. Retrieved from https://www.usenix.org/conference/foci16/workshop-program/presentation/nithyanand.

[83] Mark Nottingham. 2020. The Internet is for end users. *RFC* 8890 (2020), 1–10. DOI:https://doi.org/10.17487/RFC8890

[84] Erik Nygren, Ramesh K. Sitaraman, and Jennifer Sun. 2010. The akamai network: A platform for high-performance Internet applications. *Operating Systems Review* 44, 3 (2010), 2–19. DOI:https://doi.org/10.1145/1842733.1842736

[85] OpenINTEL. 2021. Data Access. Retrieved December 9, 2021 from https://openintel.nl/data-access/.

[86] John S. Otto, Mario A. Sánchez, John P. Rula, and Fabián E. Bustamante. 2012. Content delivery and the natural evolution of DNS: Remote DNS trends, performance issues and alternative solutions. In *Proceedings of the 12th ACM SIGCOMM Internet Measurement Conference.* John W. Byers, Jim Kurose, Ratul Mahajan, and Alex C. Snoeren (Eds.). ACM, 523–536. DOI:https://doi.org/10.1145/2398776.2398831

[87] PeeringDB. 2021. The Interconnection Database. Retrieved December 9, 2021 from https://www.peeringdb.com/.

[88] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. 2019. Tranco: A research-oriented top sites ranking hardened against manipulation. In *Proceedings of the 26th Annual Network and Distributed System Security Symposium.* Retrieved from https://www.ndss-symposium.org/ndss-paper/tranco-a-research-oriented-top-sites-ranking-hardened-against-manipulation/.

[89] Lucian Popa, Ali Ghodsi, and Ion Stoica. 2010. HTTP as the narrow waist of the future Internet. In *Proceedings of the 9th ACM Workshop on Hot Topics in Networks.* Geoffrey G. Xie, Robert Beverly, Robert Tappan Morris, and Bruce Davie (Eds.). ACM, 6. DOI:https://doi.org/10.1145/1868447.1868453

[90] Matthew Prince. 2017. Terminating service for 8Chan. *The Cloudflare Blog*. Retrieved December 9, 2021 from https://blog.cloudflare.com/terminating-service-for-8chan/.

[91] Enric Pujol, Oliver Hohlfeld, and Anja Feldmann. 2015. Annoyed users: Ads and Ad-Block usage in the wild. In *Proceedings of the 2015 ACM Internet Measurement Conference.* Kenjiro Cho, Kensuke Fukuda, Vivek S. Pai, and Neil Spring (Eds.). ACM, 93–106. DOI:https://doi.org/10.1145/2815675.2815705

[92] Enric Pujol, Ingmar Poese, Johannes Zerwas, Georgios Smaragdakis, and Anja Feldmann. 2019. Steering hyper-giants' traffic at scale. In *Proceedings of the 15th International Conference on Emerging Networking Experiments and Technologies.* Aziz Mohaisen and Zhi-Li Zhang (Eds.). ACM, 82–95. DOI:https://doi.org/10.1145/3359989.3365430

[93] Aravindh Raman, Sagar Joglekar, Emiliano De Cristofaro, Nishanth Sastry, and Gareth Tyson. 2019. Challenges in the decentralised web: The mastodon case. In *Proceedings of the Internet Measurement Conference.* ACM, 217–229. DOI:https://doi.org/10.1145/3355369.3355572

[94] Eric Rescorla. 2018. The transport layer security (TLS) protocol version 1.3. *RFC* 8446 (2018), 1–160. DOI:https://doi.org/10.17487/RFC8446

[95] RIPE NCC. 2021. RIPEstat. Retrieved December 9, 2021 from https://stat.ripe.net/.

[96] Route Views. 2021. University of Oregon Route Views Archive Project. Retrieved December 9, 2021 from http://archive.routeviews.org/.

[97] Dominic Rushe and Kari Paul. 2020. US Justice Department Sues Google Over Accusation of Illegal Monopoly. Retrieved December 9, 2021 from https://www.theguardian.com/technology/2020/oct/20/us-justice-department-antitrust-lawsuit-against-google.

[98] Walter Rweyemamu, Tobias Lauinger, Christo Wilson, William K. Robertson, and Engin Kirda. 2019. Clustering and the weekend effect: Recommendations for the use of top domain lists in security research. In *Proceedings of the 20th International Conference on Passive and Active Measurement.* 161–177. DOI:https://doi.org/10.1007/978-3-030-15986-3_11

[99] Walter Rweyemamu, Tobias Lauinger, Christo Wilson, William K. Robertson, and Engin Kirda. 2019. Getting under alexa's umbrella: Infiltration attacks against Internet top domain lists. In *Proceedings of the 22nd International Conference on Information Security.* 255–276. DOI:https://doi.org/10.1007/978-3-030-30215-3_13

[100] Quirin Scheitle, Oliver Hohlfeld, Julien Gamba, Jonas Jelten, Torsten Zimmermann, Stephen D. Strowes, and Narseo Vallina-Rodriguez. 2018. A long way to the top: Significance, structure, and stability of Internet top lists. In *Proceedings of the Internet Measurement Conference 2018.* ACM, 478–493. DOI:https://doi.org/10.1145/3278532.3278574

[101] Rachee Singh, Arun Dunna, and Phillipa Gill. 2018. Characterizing the deployment and performance of multi-CDNs. In *Proceedings of the Internet Measurement Conference 2018.* ACM, 168–174. DOI:https://doi.org/10.1145/3278532.3278548

[102] Wiktor Stadnik and Ziemowit Nowak. 2017. The impact of web pages' load time on the conversion rate of an e-commerce platform. In *Proceedings of the Information Systems Architecture and Technology: Proceedings of 38th International Conference on Information Systems Architecture and Technology.* Leszek Borzemski, Jerzy Swiatek, and Zofia Wilimowska (Eds.), Vol. 655. Springer, 336–345. DOI:https://doi.org/10.1007/978-3-319-67220-5_31

[103] The jQuery Foundation. 2021. jQuery CDN – Latest Stable Versions. Retrieved December 9, 2021 from https://code.jquery.com/.

[104] The jQuery Foundation. 2021. Using jQuery with a CDN. Retrieved December 9, 2021 from https://jquery.com/download/#using-jquery-with-a-cdn.

[105] Martino Trevisan, Danilo Giordano, Idilio Drago, Maurizio M. Munafò, and Marco Mellia. 2020. Five years at the edge: Watching Internet from the ISP network. *IEEE/ACM Transactions on Networking* 28, 2 (2020), 561–574. DOI:https://doi.org/10.1109/TNET.2020.2967588

[106] Sipat Triukose, Zhihua Wen, and Michael Rabinovich. 2011. Measuring a commercial content delivery network. In *Proceedings of the 20th International Conference on World Wide Web.* Sadagopan Srinivasan, Krithi Ramamritham, Arun Kumar, M. P. Ravindra, Elisa Bertino, and Ravi Kumar (Eds.). ACM, 467–476. DOI:https://doi.org/10.1145/1963405.1963472

[107] Anton Troianovski and Adam Satariano. 2021. Google and Apple Remove App Aimed at Spurring Protest Voting in Russia. Retrieved December 9, 2021 from https://www.nytimes.com/2021/09/17/world/europe/russia-navalny-app-election.html.

[108] Kazuaki Ueda and Atsushi Tagami. 2021. Internet flattening and consolidation considered useful (for deploying new Internet architecture). In *Proceedings of the Interdisciplinary Workshop on (de) Centralization in the Internet.* Alexander Afanasyev and Matthias Wählisch (Eds.). ACM, 11–17. DOI:https://doi.org/10.1145/3488663.3493688

[109] Tobias Urban, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. 2020. Beyond the front page: Measuring third party dynamics in the field. In *Proceedings of the Web Conference 2020*. Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen (Eds.). ACM / IW3C2, 1275–1286. DOI:https://doi.org/10.1145/3366423.3380203

[110] Pelayo Vallina, Victor Le Pochat, Álvaro Feal, Marius Paraschiv, Julien Gamba, Tim Burke, Oliver Hohlfeld, Juan Tapiador, and Narseo Vallina-Rodriguez. 2020. Mis-shapes, mistakes, misfits: An analysis of domain classification services. In *Proceedings of the ACM Internet Measurement Conference*. ACM, 598–618. DOI:https://doi.org/10.1145/3419394.3423660

[111] Roland van Rijswijk-Deij, Mattijs Jonker, Anna Sperotto, and Aiko Pras. 2016. A high-performance, scalable infrastructure for large-scale active DNS measurements. *IEEE Journal on Selected Areas in Communications* 34, 6 (2016), 1877–1888. DOI:https://doi.org/10.1109/JSAC.2016.2558918

[112] W3C Recommendation. 2012. Navigation Timing. Retrieved December 9, 2021 from https://www.w3.org/TR/navigation-timing/.

[113] WHATWG. 2020. Fetch Living Standard. Retrieved December 9, 2021 from https://fetch.spec.whatwg.org/#http-cache-partitions.

[114] WPO Foundation. 2021. Official Repository for WebPagetest. Retrieved December 9, 2021 from https://github.com/WPO-Foundation/webpagetest.

[115] WPO Foundation. 2021. WebPagetest. Retrieved December 9, 2021 from https://github.com/WPO-Foundation/wptagent/blob/master/internal/optimization_checks.py.

[116] Behrouz Zolfaghari, Gautam Srivastava, Swapnoneel Roy, Hamid R. Nemati, Fatemeh Afghah, Takeshi Koshiba, Abolfazl Razi, Khodakhast Bibak, Pinaki Mitra, and Brijesh Kumar Rai. 2020. Content delivery networks: State of the art, trends, and future roadmap. *ACM Computing Surveys* 53, 2 (April 2020). DOI:https://doi.org/10.1145/3380613

[117] Hadi Zolfaghari and Amir Houmansadr. 2016. Practical censorship evasion leveraging content delivery networks. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. Edgar R. Weippl, Stefan Katzenbeisser, Christopher Kruegel, Andrew C. Myers, and Shai Halevi (Eds.). ACM, 1715–1726. DOI:https://doi.org/10.1145/2976749.2978365

Received March 2021; revised September 2021; accepted November 2021