NORTHWESTERN UNIVERSITY

Internet Service Consolidation: Trends, Impact, and an End-User Solution

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Computer Science

By

Rashna Kumar

EVANSTON, ILLINOIS

February 2026

Abstract

This dissertation investigates how the Internet's increasing reliance on third-party infrastructure has reshaped web service provision, user access, and the consequences of concentration. As websites increasingly depend on external providers for foundational services, including content hosting, authoritative DNS, and certificate authorities, a small number of organizations now mediate critical functions of web delivery and access. This work provides an Internet-scale empirical analysis of consolidation in these foundational service layers and examines its implications for resilience, sovereignty, and performance.

The work proceeds in two parts. The first measures consolidation across commercial and public-sector web infrastructure worldwide. Using large-scale measurements collected across countries, it characterizes provider concentration across the service layers and shows that consolidation is substantial, including in government web services. To explain variation in these patterns, it introduces a comparative framework that distinguishes structural consolidation from strategic consolidation. Specifically, it compares government websites to popularity-stratified commercial websites within the same national environment, allowing observed concentration to be interpreted relative to the surrounding provider ecosystem rather than in isolation.

The second part analyzes the implications of consolidation for end users. It shows that concentration expands shared failure domains and often coincides with limited practical redundancy, increasing systemic fragility when widely used providers experience faults. It further examines exposure through both foreign service dependencies and on-path intermediaries involved in access to government websites, focusing on public-sector services because they are a primary channel through which citizens access essential information and state services. This analysis shows that consolidation can increase cross-border and network-level dependencies, with implications for jurisdiction and security. Finally, it studies performance implications, including how consolidation in content delivery and DNS resolver infrastructure can interact to influence latency, especially when CDN replica selection relies on DNS-based signals that are increasingly decoupled from end-user location due to increasing resolver consolidation.

Taken together, this dissertation offers a unified framework for measuring and interpreting Internet infrastructure consolidation and for understanding its technical and governance consequences across sectors and countries.

Thesis Committee

Fabián E. Bustamante	Committee Chair
Northwestern University	
Yan Chen	Committee Member
Northwestern University	
Peter Dinda	Committee Member
Northwestern University	
Aleksandar Kuzmanovic	Committee Member
Northwestern University	
Marcel Flores	External Committee Member
Netflix	
Phillipa Gill	External Committee Member
Google	

Table of Contents

CHAPTER 1

Introduction

Over the past decade, the Internet has increasingly shifted from on-premises infrastructure toward third-party services, driven by scalability, operational efficiency, and improved security. While this transition has enabled rapid deployment and global reach, it has also led to significant consolidation across critical Internet service layers, raising concerns about resilience, control, privacy, and user experience.

This dissertation examines how the Internet's growing reliance on third-party infrastructure is reshaping web service provision, patterns of dependency, and the consequences of concentration. Focusing on foundational service layers, including content hosting, authoritative DNS, and certificate authorities, it first measures how these services are concentrated across providers at Internet scale, across countries, and across different parts of the web, including both public sector and commercial websites. It then explains these patterns by distinguishing broad structural consolidation from more intentional or context-specific forms of centralization through comparisons within shared national and market environments, and finally considers the broader implications of these patterns for resilience, sovereignty, security, and performance.

1.1. Part I: Measuring Consolidation Across Service Layers and Sectors

The first part of this dissertation focuses on the consolidation of infrastructure services. Using worldwide vantage points collected over a two-year measurement campaign, we analyze the reliance of popular websites on third-party providers for content delivery networks, authoritative DNS, and certificate authorities. Our results reveal substantial consolidation within each of these service categories, with a small number of providers serving a dominant share of global web traffic. We extend this analysis to government websites across diverse countries and regions, examining how public sector web services rely on third-party infrastructure for serving content. Our findings highlight widespread consolidation in government serving infrastructure, raising important questions about control over access to public information, national resilience, and dependency on external providers.

1.1.1. Explaining Consolidation: Structural and Strategic Consolidation

Seeing that government websites are often similarly consolidated, we then ask what is driving that consolidation. In some countries, high consolidation may be structural, reflecting a limited domestic provider ecosystem that also serves the commercial web. In others, it may be strategic, where the government chooses to centralize on a narrow set of providers. Without distinguishing these cases, consolidation measurements are ambiguous with respect to network control, failure domains, and user-visible impact. To distinguish these cases, we use a within-country, stratified observational design that treats the domestic commercial web ecosystem as a counterfactual baseline. Across 61 countries,

we compare government websites to popularity-stratified commercial sites operating in the same national network environment and measure consolidation across hosting, authoritative DNS, and certificate authorities.

1.2. Part II: User-Facing Implications of Consolidation

The second part of this dissertation examines how consolidation impacts end users, with a focus on three implications: resilience, exposure, and performance.

1.2.1. Resilience Implications

First, consolidation heightens systemic fragility. When a small number of providers become critical for the operation of large fractions of the web, even routine faults can cascade into widely visible incidents. Recent disruptions in 2025 illustrate this amplification effect: Cloudflare experienced a global outage on November 18, 2025, after an internal change triggered failure in core traffic handling, disrupting access to many sites and services. AWS saw a major US EAST 1 disruption around October 20, 2025, affecting regional service endpoints, with knock-on effects across many dependent services. Microsoft similarly faced a global Azure outage on October 29, 2025, which also impacted Microsoft 365 and other downstream services. Crucially, these incidents show that even the most sophisticated and well-resourced providers are not immune to failures. When so much of the Internet depends on them, a single disruption can ripple outward and impact a large fraction of online services at once. Consistent with this mechanism, our evaluation shows that high concentration often coincides with limited practical fallback: authoritative DNS shows near-zero organizational redundancy (most domains depend on a single

independent DNS provider), and hosting redundancy is limited and often reflects multi-origin complexity rather than true failover capacity.

1.2.2. Exposure Implications

Second, consolidation increases exposure by increasing dependence on infrastructure outside local control. One form is jurisdictional exposure: reliance on foreign providers and hosting locations introduces cross-border dependencies that can complicate accountability, legal jurisdiction, and operational control. In our measurements, structurally consolidated countries exhibit more uniform exposure outcomes, while strategically consolidated countries show substantially greater cross-country dispersion, indicating that when governments diverge from market constraints, exposure becomes more a consequence of policy choice than a fixed feature of the ecosystem. Exposure, however, is not only about where services are hosted, but also about the network intermediaries citizens depend on to reach them. To capture this dimension, we analyze on-path infrastructure dependencies for paths to government websites, including transit networks and exchange points, focusing on public-sector services because they are state-facing, high-stakes, and directly tied to sovereign service delivery. We find that access to government websites often traverses a small set of shared foreign intermediaries, revealing an additional and often overlooked layer of concentration in how citizens reach state-provided services. This pattern also has security implications: countries with greater foreign on-path exposure also tend to exhibit weaker HTTPS adoption, increasing the risk that sensitive interactions are not protected by encryption and authentication and are therefore more

vulnerable to interception or manipulation in transit, particularly when paths traverse third countries outside both the government's country and the hosting country.

1.2.3. Performance Implications

Third, consolidation can potentially affect the performance experienced by end users. Across countries, we observe that structurally consolidated regimes tend to impose higher domestic latency to hosting, DNS, and CA endpoints with tighter government–commercial overlap, consistent with shared ecosystem constraints, whereas strategically consolidated regimes exhibit lower median latency but greater dispersion, consistent with uneven localization and deployment choices.

Consolidation can also shape performance through content delivery steering. As CDNs consolidate, a growing fraction of the web is served through a small number of delivery platforms, centralizing not just where content is hosted, but also how users are steered to replicas. Replica selection is the CDN control plane that maps users to specific points of presence, shaping the latency they experience.

CDNs typically select replicas using one of three approaches. With anycast-based routing, Internet routing delivers a user to a nearby replica, but catchments can be difficult to predict and may shift over time. With regional anycast, DNS first maps the user to a region, and anycast then selects a nearby replica within that region, limiting extreme catchment stretch while retaining anycast within-region dynamics. With DNS-based replica selection, the CDN explicitly chooses a replica based on the apparent location of the user's DNS resolver, using DNS as the primary steering signal.

To study how this centralized steering is implemented in practice, we introduce a methodology that infers a CDN's predominant replica selection approach from the outside, at scale. We use it to characterize which steering mechanisms are most commonly used to serve popular content. We find that DNS-based replica selection is the dominant approach. This result matters because it means that content delivery performance is often mediated by DNS behavior rather than only by Internet routing.

DNS-based steering is not inherently problematic and has long been effective, especially when most users rely on ISP-provided resolvers that are deployed close to access networks, making resolver location a reasonable proxy for user location. The risk emerges because the resolver market is simultaneously consolidating onto a small set of third-party public resolvers, whose deployment footprints are often decoupled from end users. As third-party resolver consolidation grows, both through direct user configuration and through ISP forwarding behavior, resolver location becomes a weaker proxy for where users actually are. When DNS is the dominant steering mechanism, this mismatch can systematically map users to suboptimal replicas, increasing latency even when closer replicas exist. In combination, CDN consolidation, reliance on DNS-based steering, and consolidation in DNS resolvers create an indirect pathway through which concentration at one layer can translate into user-facing performance degradation.

1.3. Summary

Taken together, this dissertation shows how consolidation across the web stack and across sectors can create resilience risks, create jurisdictional exposure that constrains sovereignty and heightens security risk, and can degrade performance directly affecting

end users, with impacts that differ depending on whether consolidation is structural or policy driven. Through large-scale measurements, this work provides a unified view of Internet consolidation and its technical and societal implications.

1.4. Dissertation Organization

The remainder of this dissertation is structured as follows. Chapter 2 presents the thesis statement. Chapter 3 presents related work and puts this dissertation in perspective. Chapter 4 discusses a large-scale longitudinal study of third-party dependency and consolidation trends in the service infrastructure supporting popular commercial websites, with a focus on foundational services including content hosting, authoritative DNS, and certificate authorities. Chapter 5 focuses on web hosting strategies and the extent of consolidation in the infrastructure behind public-facing government websites across countries. Chapter 6 examines what drives observed consolidation patterns and introduces a comparative framework to distinguish structural consolidation from strategic centralization using within-country comparisons between government websites and popularity-stratified commercial websites. Chapter 7 analyzes the resilience implications of consolidation, including shared failure domains and limited practical redundancy across critical services. Chapter 8 examines the exposure implications of consolidation, including foreign service dependencies and on-path infrastructure dependencies involved in access to government websites, and their implications for jurisdiction and security. Chapter 9 analyzes the performance implications of consolidation, including latency effects linked to service concentration and content delivery steering, and introduces a methodology to infer CDN replica selection approaches at scale to study how resolver consolidation

interacts with DNS-based steering to affect end-user latency. Finally, Chapter 10 concludes the dissertation by synthesizing the main findings, discussing their technical and governance implications, and outlining directions for future work.

CHAPTER 2

Thesis

This dissertation argues that consolidation in foundational Internet infrastructure services is a global and measurable phenomenon across web sectors, and that it has direct implications on web users' experience. Through large-scale measurements of popular and government websites, it characterizes consolidation in content hosting, authoritative DNS, and certificate authorities, distinguishes structural consolidation from strategic consolidation, and shows how these dependencies translate into resilience risks, jurisdictional exposure, and performance risks.

CHAPTER 3

Third-Party Dependency and Consolidation in the Public Sector

The shift from on-premise infrastructure to third-party providers has become a global trend, fundamentally reshaping the Internet over the past decade. This shift is driven by the benefits offered by major providers, including access to computing resources in data centers across multiple regions, flexible resource allocation, high service availability, and relatively lower capital and operational costs [**18, 70, 81**].

This shift, however, has also led to increasing concerns about Internet consolidation and centralization – the concentration of traffic, infrastructure, services and users on a handful of providers. The 2019 Global Internet Report [**155**] provides an early overview of this trend in every aspect of the Internet economy, from access provision to service infrastructure and applications. It argues that while consolidation is often seen as an expected result of maturing markets and industries, the combination of society's increased dependency on the Internet, business agility, and the almost total lack of regulation is leading to a handful of platforms in control of much of the Internet's functionality and interoperability. Since that report, several efforts have explored this trend [**182, 6, 116, 91, 64, 113, 96, 77**] and its economic, political and reliability implications [**16, 78, 80, 104, 182, 155, 17**].

The Web provides a relatively accessible environment to characterize these centralization trends in public-facing content and, incidentally, to understand their concerning

implications. Accessing a website depends on several services provided by third parties, including DNS, CDNs, and CAs. To visit a site, a user must interact with at least one DNS authoritative nameserver to retrieve the IP address of the web server(s) hosting the content. These servers may be operated by one or more CDNs to enhance reliability and performance. If the servers use HTTPS, the client may also need to consult one or more CAs to verify the validity of the servers' SSL certificates. In fact, a popular website may rely entirely on third-party providers for these critical services.

Several recent studies have leveraged this observation to assess third-party dependencies in popular websites, though typically from only a single vantage point [**91**, **22**, **54**, **165**, **90**] or only a few vantage points [**172**].

We build on these prior work ***to explore if, and to what extent, third-party dependency and centralization varies across countries and regions of the world***.

Our work is motivated by two simple observations. *First*, while websites could potentially be accessed anywhere, not all websites are popular everywhere. Indeed, we measured that the popularity of websites, beyond a few top-ranked ones, is region specific. *Second*, while many third-party services such as DNS and CDNs have been building global infrastructures, on and off-networks [**64**], their deployment is not (yet) omnipresent, and their relative performance compared to competitors varies across markets.

Motivated by these observations and concerns, we explore third-party and centralization trends in the wider Internet for public web content. We present a methodology that builds on prior work [**91**, **172**] to carry out a large-scale, longitudinal analysis of third-party dependency and centralization around the world. Using this methodology,

we analyze the dependencies of top regional websites in 50 countries, covering approximately 78% of the global Internet population (a total of 16,774 unique websites). We provide results from two consecutive years, offering insights into how these trends have evolved over time. Additionally, we present early findings that explore various factors that may explain the observed differences, including economic conditions, Internet development, trading partnerships, website categories,

Our findings reveal that third-party dependencies and critical dependencies vary significantly across regions. We report that between 19% and as much as 76% of websites, across all countries, depend on a DNS, CDN, or CA third-party provider. Critical dependencies, where a service necessary to ensure access to a website depends on a single provider, while lower are equally spread ranging from 5% (CDN in Costa Rica) to 60% (DNS in China). Interestingly, despite this high variability, the market of third-party providers seems highly concentrated: the top-three third-party providers across all countries serve an average of 92% of all sites and Google, *by itself*, serves an average of 70% of all websites. Perhaps more problematically, we find these values have increased, a year later, by ≈14% on average, for CDNs.

3.1. Motivation

Our work is motivated by two observations. First, the Web is made of 1.9 billion websites as of May 2022[1] and except for a few global sites, most are not universally popular. Second, while the infrastructure of large third-party services continues to expand,

[1]https://www.internetlivestats.com/total-number-of-websites/

it is not yet omnipresent and may fare differently against their competition at different locales.

To illustrate the first point, we measure the overlap between the top-500 regional sites for each of the 115 countries in Alexa[2] and the top 100k sites from the global ranking. For each of these countries, we measured the overlap of the top-500 regional sites with the top 1k, 5k, 10k, 20k, 50k and 100k subsets of the global sites. As we extend the list of global sites considered, there is a higher probability of including a regional popular site in the global ranking and, thus, of finding a higher overlap between global and regional lists.

Figure 3.1 shows the percentage of overlap between these sets. The overall top-ranking websites are clearly dominated by just a few countries, even with the largest global ranking subset, with ≈ 25% of them having an overlap higher than 94%. On the other hand, half of the countries have less than ≈77% overlap even with the full top-100k list of websites.

The second observation has served as motivation for brokering CDN [**23**] and multi-cloud architectures [**181, 153**]. As no single infrastructure is omnipresent, websites popular in specific regions may choose to build on their own services or contract a well-provisioned local service provider. Even globally popular content providers (CP) are known to contract with different CDNs and DNSs in different part of the world based on connectivity, availability or cost [**153**]. For instance Rakuten, the Japanese online retailing company, uses Akamai as its DNS provider for rakuten.com, but a private DNS

[2]Despite concerns with the representativeness and stability of Alexa rankings, these are the longest lists of regional website rankings and considered the appropriate choice for end-user-based analysis [**146**].

for rakuten.jp. Similarly, wikipedia.org uses DigiCert as its CA provider in Germany and Sweden but IdenTrust Inc. in the US and Canada. Likewise, microsoft.com hosts content on both Akamai and Amazon Cloudfront in Canada, but only on Akamai in Sweden.

3.2. Methodology and Dataset

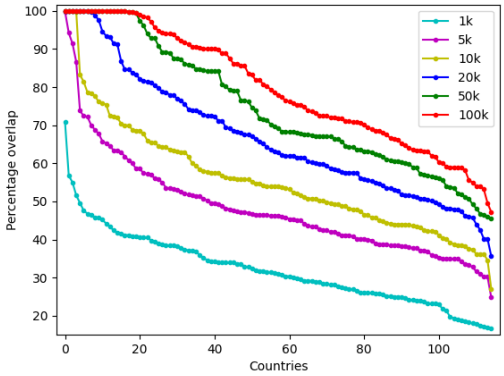


Figure 3.1. Overlap between global and top regional sites. Countries are sorted based on their overlap and plotted left to right.

Overlap Class	Country Codes
High	AE, AR, AU, BR, CA, CN, DE, ES, FR, GB, GR, HK, ID, IN, IT, JP, KR, MX, MY, SG, TR, TW, US, VN
Medium	BE, CH, CL, CR, IL, UA, PL, NL, NO, RO, SE, TH, ZA
Low	AL, BA, BG, CZ, DK, EE, GE, HU, LV, MD, NZ, PT, RS

Table 3.1. Countries grouped by degrees of overlap between top-regional sites and the global ranking.

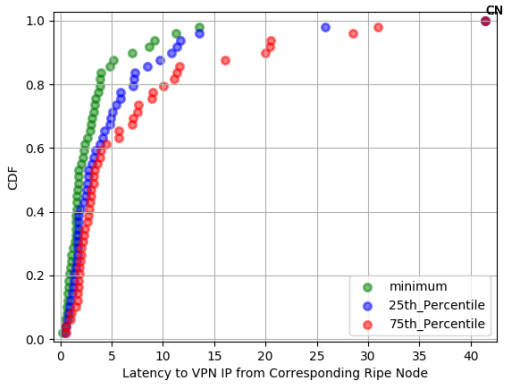


Figure 3.2. Latency (ms) from RIPE Atlas Nodes to corresponding VPN Nodes.

In this section, we describe a measurement methodology that builds on Kashaf et al. [**91**]'s to characterize service dependencies and centralization around the world. Our extensions to Kashaf et al.'s are meant for carrying out a country-specific analysis, including the selection of countries based on the ranking overlap and the use (and validation) of VPN vantage points for measurement; we also analyze the number of websites using OCSP must-staple and quantify the resources that do not have CNAMES and can be mapped to a CDN. We close with a description of the dataset we collect using it.

3.2.1. Country Selection

As Fig. 3.1 shows, the set of top-ranked websites is clearly determined by a handful of countries, while half of all countries have less than ≈77% overlap, even with the full top-100k list of websites. Thus to understand the degree of third-party dependency around the world, we should use a set of countries that together capture the range of possible overlaps.

To this end, we first group all 115 countries with regional rankings in Alexa based on the degree of overlap between their regional popular sites and the list of top global sites. We divide the countries based on this overlap into three groups: *high-*, *medium-* and *low-overlap* for the top, medium and bottom third of countries with the highest overlap. We then select 50 countries from across these three groups ensuring that (*i*) the sample of countries in the analysis captures a sufficiently large fraction of the Internet user population, and (*ii*) there are VPN vantage points in those countries with claimed locations that can be verified. We select countries that have a range of overlap to study the third-party dependency of the popular sites of countries that are not well represented in the global rankings and are therefore not investigated in the broader global analysis.

Figure 3.1 lists the selected countries (using their two-letter codes) which includes 38.4%, 25.9% and 35.7% from the high, medium and low-overlap sets based on regional websites.

In total, our dataset covers all inhabited regions of the world and captures 78.1% of the world's Internet user population [**178**]. We group countries in five different regions: the Americas, Europe, Asia Pacific, Africa, and the Middle East and calculate the percentage of the world's Internet population covered by the corresponding countries in those regions.

In order to make observations about each country, we consider vantage points available via VPN providers. Of the 50 countries, all but China have vantage points available through the Nord VPN [**121**]. To include China we use the Hotspot Shield VPN, since Nord and many other VPNs do not provide service in China [**92**].

To gain confidence in the claimed location of the vantage points, we obtain their public IPs (resolving a domain whose authoritative server we control) and use a set of five RIPE Atlas nodes within the same country to send a sequence of three ICMP pings to the vantage point. We would expect most nodes geographically close to the vantage point's claimed location to have minimum latencies in the 10-20ms range and below 50ms. As a reference, China, the largest country in our set, is 5,250km East-West or ≈40.25ms considering a 2.3x median inflation over *c*-latency [**154**].

Figure 3.2 shows the minimum, 25th and 75th percentile of ping latency from the set of RIPE Atlas nodes to the corresponding vantage point. The minimum latency to all vantage points, with the exception of the one in China, is within 15ms and 86% of minimum latencies are below 5ms, suggesting that these nodes are within the claimed country. We further geolocate the vantage points' IPs using two popular geolocation databases: MaxMind GeoLite2 [**111**] and IP2Location BD11.Lite [**85**]. Past work has shown the geolocation databases to be reliable at the country-level [**136**]. Both databases place the IPs of all 49 Nord VPN nodes in the claimed countries. China proved to be more challenging. The two geolocation databases we use disagree with Maxmind geolocating the node in China, while the IP2Location database placing it in Japan. The ping latency to the VPN node in China, however, are consistent with our estimations and thus we consider the claimed location to be correct.

3.2.2. Data Collection Methodology

Using VPN vantage points in the selected countries, we launch measurements to their country's set of top-500 regional websites, and use a range of heuristics for labeling three major services – DNS, CA and CDN.

3.2.2.1. DNS Measurements. We use a number of heuristics to label all nameservers used by a website as *private* or *third-party*. For instance, the site belgocontrol.be uses two distinct nameservers, i.e. *skynet.be which is a third-party nameserver and *belgocontrol.be which is a private nameserver. For each website, we find all the nameservers used by the website, that belong to different logical entities, by issuing NS queries to the domain name from the selected country's VPN vantage point. Note that we do not perform resolution at this stage so our results are not affected by caching, we just find the unique set of DNS providers used by the website. We start by labeling each nameserver used by a website as of an *unknown* type. We then compare the second level domain (2LD)[3] of the website and the nameserver with a match suggesting this is a *private* nameserver [**95**].

While the *2LD-matching* heuristic works well in most cases, it may result in misclassifications of some nameservers. For instance the nameserver of youtube.com is *google.com and though both belong to the same logical entity, the *2LD-matching* heuristic will classify the nameserver as third-party. To resolve this we make use of an additional heuristic based on Subject Alternative Names (*SANs List*) [**28**]. If the website uses

[3]By 2LD we refer to 2LD + TLD in this work

HTTPS, we find the site's SANs list via the SSL certificate of the website. For each unclassified nameserver, we then look for the presence of the second level domain of the nameserver in the SAN list, whereby the presence indicates a *private* DNS nameserver. This heuristic correctly identifies cases like youtube.com using a private DNS.

We use a third heuristic based on Start of Authority records (SOA) – *SOA-record-matching* – to label the unclassified nameserver [**91**]. In this case, we compare the entity pointed to by the SOA records of the website and the DNS provider pointed to by the SOA records of the nameserver; a mismatch here indicates a *third-party* DNS nameserver. For instance, the SOA record for the website imdb.com is *amazon.com and its nameservers are Dynect and UltraDNS. Since the SOA records of these nameservers do not match the SOA record of imdb.com, we label imdb.com as using two third-party DNS providers.

For the remaining unknown servers, if the concentration of the nameserver (i.e. the number of websites dependent on a given provider) is large, we label it as *third-party*. We set the value of the threshold as >50 (i.e. if an unlabeled nameserver serves greater than 50 websites, we label it as a third-party).[4] For sanity check, we manually investigated the servers that were labeled by this heuristic and they were all, in fact, popular third-party DNS providers, such as Amazon, Akamai, NsOne, Cloudflare, DnsPod and Alibaba DNS. We also performed a sensitivity analysis on this threshold and observed that reducing the value of the threshold resulted in some nameservers that we could not manually determine as third-party with full confidence such as *gandi.net in France and *.hyp.net in Norway.

[4]Following Kashaf et al. [**91**].

The second condition of Algorithm 1 summarizes the heuristic when the service type is instantiated as *DNS*. This basic three-step classification logic, involving 2LD-matching, SANs-List and SOA-record-matching, is described in Algorithm 2 where the *service.url* is provided to the algorithm is the DNS nameserver.

DNS Redundancy. We also measure the percentage of websites that are served by a single DNS provider (i.e., critically dependent on this provider) or served by multiple third-party DNS providers, and the percentage of websites that are served by private and third-party DNS providers.

3.2.2.2. **CDN Measurements**. To find the set of CDNs hosting the targeted website and determine whether the CDNs used are private or a third-party service, we find the CNAME of the internal resources of the website. We start by using the webdriver capabilities of the Selenium library in python to generate a HAR file for each website which gives us all the resources of a website. We filter internal resources from the set of *all* resources by matching the 2LD of the website to that of the resource, checking the presence of the 2LD of the resource in the SAN List of the website, and comparing the SOA records of the website and resource, a match in any of the three cases indicates an internal resource. We additionally use public suffix lists [**102**, **168**] to identify any remaining internal resources.

To find third-party dependence, we find the CNAMEs of all the internal resources of a website by issuing dig CNAME queries on all the internal resources of the webpage. We then obtain the set of CDNs used by the internal resources from our self-populated CNAME-CDN map [**36**, **179**]. An alternative way to identify the CDN hosting an internal resource without a CNAME redirect would be to compare the autonomous system

number (ASN) of the resource with those of popular CDNs [**179**, **109**]. We find an additional 17% of all resources, on average across countries, can be mapped to their CDNs using this approach. However, since the classification algorithm depends on CNAME for the labelling, we can not leverage the AS mapping approach here. Our results, therefore, show a lower bound on the third-party dependency trends for the CDN service. The process is summarized in the third condition of Algorithm 1.

Next, we determine whether each CDN that hosts the internal resources of a given site is a private service or a third-party-provided one. For each *(website, CDN)* pair, we extract the CNAMEs of the internal resources of the website which uses that CDN. Then for each of these CNAMEs, if the 2LD of the CNAME is the same as the 2LD of the website, we classify the CDN as private. If the website uses HTTPS and the 2LD of the CNAME is present within the SAN list obtained from the SSL certificate of the website, the CDN is again classified as private. For example, the website twitch.tv has resources fetched from the CDN Fastly and contains CNAMEs such as *.fastly.net. The 2LD of the CNAME and website do not match but the presence of the 2LD of the CNAME in the SAN list of the website indicates a private CDN in this case. We finally label the remaining websites by comparing the DNS SOA records of both the website and the CNAME; a mismatch here indicates a third-party CDN. For instance, the website reddit.com also has resources fetched from the CDN Fastly. Since the 2LD matching and the SAN list check do not indicate a private CDN, we finally look at the SOA information. In this case, the SOA of the CNAME of the CDN is *.fastly.net, and the SOA of the website is *awsdns.net; the mismatch indicates a third-party CDN. For CDNs that have multiple CNAMEs, we iterate

over all CNAMEs and if any of the CNAME is identified as private, we label the CDN as private. For instance, the website facebook.com uses the CDN Facebook (CNAMEs: "*.fbcdn.net", "*.facebook.com" and "*.cdninstagram.com"). Our heuristic classifies the first two CNAMEs as private so we label the CDN in this case as private. Then, for an unclassified CDN, if any of the CNAME is identified as third-party, we label the CDN as third-party. We manually sampled websites and verified the cases where websites have multiple CNAMEs and find that this heuristic correctly labels all the CDNs. So to classify whether each CDN used by the internal resources of the website is a private or third-party service, we follow the same three-step heuristics of Algorithm 2 using the CNAMEs of the internal resources as the *service.url*.

CDN Redundancy. As with DNS, we measure websites that are (*i*) redundantly provisioned by CDNs (host content from more than one private and/or third-party CDNs), (*ii*) critically dependent on a third-party CDN (host content from that one CDN), (*iii*) use multiple third-party CDNs and (*iv*) use both private and third-party CDNs. We measure CDN redundancy by finding the unique set of CDNs that the CNAMEs of a website map to using our self-populated CNAME-CDN map. For instance, the website zoom.us is redundantly provisioned by CDNs as it uses the CDN Cloudfront (CNAME: *.cloudfront.net), Google (CNAME: googlehosted.com) and Cloudflare (CNAME: *.cloudflare.com).

3.2.2.3. CA Measurements. For each website that supports HTTPS, we want to find its CA and also identify whether the CA is a third-party (e.g. DigiCert used by yahoo.com) or private CA (e.g. Google Trust Services used by google domains or Microsoft Corporation used by microsoft domains). In addition, we want to know if the website has enabled

Online Certificate Status Protocol (OCSP) stapling. This means that before accessing a site, clients do not need to explicitly contact the CA, which manages the Certificate Revocation List (CRL) distribution Points (CDP) and OCSP servers, to verify the validity of the certificate. With OCSP enabled, the certificate's revocation status comes included with the TLS/SSL handshake. This reduces the criticality of the third-party dependency on the CA which means an outage of OCSP responders and CDPs does not translate into the website becoming inaccessible. To this end, we first make a request using OpenSSL to find a website's listed CA. If the request, which is to port 443, fails, then we assume the website is HTTP-only. At this stage, if the request succeeds, we also check if it has enabled OCSP stapling through information in the request response. The second condition of Algorithm 1 summarizes our heuristic when the service type is CA.

Next, we find the CA's url from the name of the CA. To classify the CA's url as third-party or not, we make use of the same three step heuristics described in Algorithm 2 in order to prevent misclassification by using a single approach. If the 2LD of the website and the CA's url match, then we classify the CA as private. If there is a mismatch, but if the 2LD of CA's url is in the SAN list for the website, then we also classify the CA as private. Finally, if neither of the previous two conditions are met, we check if the DNS SOA record for the CA and the website match. If they do not match, then we classify the CA as third-party. If a website does not fit the previous conditions, then we classify the CA as unknown.

Pseudocode 1 ThirdPartyDependence(w)

```
service ::= DNS | CDN | CA
if service = DNS then
    NS ← digNameservers(w)
    for ns ∈ NS do
        nstype ← FindserviceType(w, ns)
        if nstype = unknown ∧ concentration(ns) > 50 then
            nstype ← third
        end if
    end for
end if
if service = CA then
    CA ← findCertificate(w)
    CAURL ← findCAURL(w, CA)
    catype ← FindserviceType(w, CAURL)
end if
if service = CDN then
    IR ← findInternalResources(w)
    cnamesIR ← digCnames(IR)
    CDNs ← findCDN(cnamesIR)
    for cdn ∈ CDNs do
        cnames ← findCnames(w, cdn)
        for cname ∈ cnames do
            cnametype ← FindserviceType(w, cname)
        end for
    end for
end if
```

Pseudocode 2 FindserviceType(w, service.url)

```
service ::= DNS | CDN | CA
service.type ← unknown
if 2ld(w) = 2ld(service.url) then
    service.type ← private
else if isHTTPS(w) ∧ 2ld(service.url) ∈ SANList then
    service.type ← private
else if SOAProvider(w) ≠ SOAProvider(service.url) then
    service.type ← third
end ifreturn service.type
```

3.2.2.4. **Third-party Service Centralization**. We are particularly interested in the degree of service centralization in markets around the world. The hypothesis is that third-party dependencies and centralization are positively correlated (i.e., high degrees of centralizations in markets with a high level of third-party dependencies) as consolidation of

third-party services leads to centralization. However, different markets could be centralized around different or the same set of key service providers.

To measure the degree of centralization across each service, we find the number of third-party websites served by the top-1, top-3 and top-5 providers of each service across the countries and the websites critically dependent on these top providers. We analyze market trends across infrastructures and countries in Sec. 6.5.

3.2.3. Dataset

We collected the final set of regional websites in April 2021 for the 50 countries. The set includes a total of 25,000 websites with 15,774 unique sites. The average number of unique sites across these countries is 280 and China has the most unique set of sites (448 out of the top-500) and Singapore has the least (160 out of top-500). We run the study in April 2021 and again in 2022. In the 2022 snapshot, we find that from 15,774 total unique websites, 11,138 use CDNs and 9,766 use HTTPS. For our CDN analysis, we find a total of 1,339,871 unique resources and use 877,337 unique internal resources. Across countries, we find 68 unique third-party CAs, 60 unique third-party CDNs and 740 unique third-party DNS Providers. Note that, each year for every country, we select three probes and for each probe, we run the measurements thrice. However, in a given year the set of nameservers, CAs and CDNs identified and their classification for each country was the same in all three runs.

3.3. Dependency and Centralization

In the following paragraphs, we use the dataset collected in 2022 to study the degree of third-party dependencies, critical dependency, and market centralization around the world. Our analysis looks to answer the following questions:

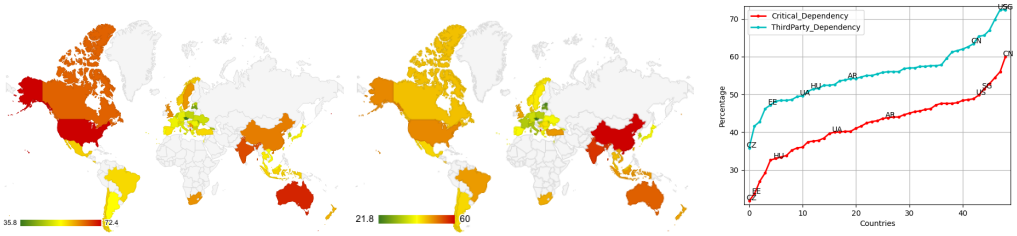
- How common is the third-party dependency of websites around the world?
- How much of this dependency is critical, dependent on a single third-party DNS or CDN provider?
- How concentrated is the market of third-party service providers within a country, region, and globally?

We first look at DNS third-party dependencies, including critical dependencies, around countries and regions.

3.3.1. DNS Findings

Figure 3.3(a) and 3.3(b) plot the map of DNS third-party and critical dependency in each country, with red-colored countries having the highest dependence, yellow-colored countries having moderate dependence and green-colored countries having least dependence. Figure 3.3(c) plots a line graph to show the variation in the degree of DNS third-party dependency and critical dependency across countries. We find an average third-party dependency of 55.4% and most noticeable a wide range of dependency, from as low as 35.8% for the Czech Republic, to as high as 72.4% for Singapore. While critical dependency is

lower than third-party dependency, with an average of 42.0%, the spread is similar ranging from the Czech Republic's lowest critical dependency of ≈21.8% to the critical dependency of China close to 60.0%. While the US and Singapore have the highest third-party dependency, China has the highest critical dependency. Generally, countries that have higher third-party dependency also tend to have a higher critical dependency on a single third-party DNS provider.



(a) Third-Party Dependency (b) Critical Dependency (c) Third-Party and Critical Dependency

Figure 3.3. DNS third-party dependency and critical dependency of each country's top-500 sites.

We also characterize the fraction of websites that use multiple DNS providers (redundant), that use multiple third-party DNS providers, and that use both third-party and private DNS providers. Figure 3.4 plots a map indicating the degree of DNS redundancy in each country, with red-colored countries having the least redundancy and green-colored countries having the highest redundancy. Figure 3.5 plots a line graph to further show the variation in overall redundancy, multiple third-party providers and third-party and private DNS providers across the different countries. We see that, on average, 14% of regional sites have redundant DNS, with Estonia having a maximum redundancy of 24% and China having minimum redundancy of 4.0%. We find that 3.8% of sites, on average,

have multiple third-party DNS providers with the US having a maximum of 13.8% and China having a minimum of 0.8%. On average, 3.2% sites use third-party and private DNS services (maximum of 7.2% for France and minimum of 0.4% for China).

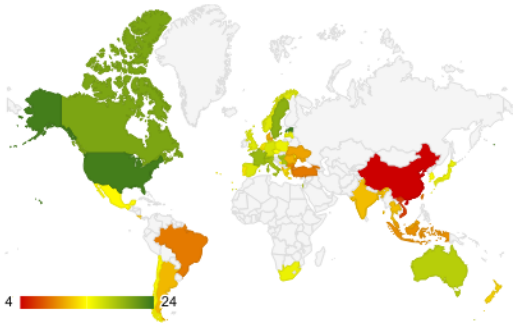


Figure 3.4. Plot of DNS redundancy

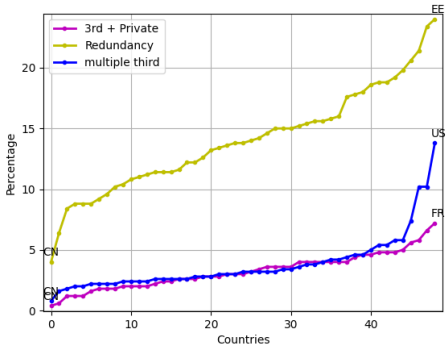


Figure 3.5. Line plot of redundant, multiple third-party, and both third-party and private DNS.

We then identify the level of dependency of websites on the five most popular DNS providers across countries. Table 3.2 shows the average number of websites relying on each DNS provider. We see that Cloudflare and Amazon DNS alone are used by 73% of the websites that use a third-party DNS, on average, across the countries.

We additionally find that only three DNS providers are used by an average of 77.5% of the websites that use third-party DNS across countries. Taiwan has the highest degree of DNS Centralization with the top three DNS providers being used by 88.7% of the country's websites that use third-party DNS providers and the Czech Republic has the least DNS centralization (60.9%). Table 3.3 shows the top-3 popular DNS providers across all regions of our vantage points and the average number of websites dependent on them. We see

DNS Provider	Avg	Std Dev
CloudFlare	**43.5**	**14.2**
Amazon Route 53	**29.3**	**9.6**
NsOne	8.3	3.0
Akamai	7.6	5.0
UltraDNS	4.3	2.1

CDN	Avg	Std Dev
Google	**70.0**	**26.3**
Akamai	26.9	9.7
Fastly	18.7	7.7
Cloudflare	16.6	5.9
Amazon Cloudfront	15.7	8.7

CA Provider	Avg	Std Dev
Digi Cert	**36.3**	**7.3**
Comodo CA Limited	15.2	4.6
IdenTrust Inc.	14.8	7.3
GlobalSign	14.0	4.4
Starfield Technologies, Inc.	6.4	3.3

Table 3.2. Top-5 DNS, CDN and CA providers across countries (average percentage of websites).

that in Europe, the Asia Pacific and Africa and the Middle East, Cloudflare is the most popular third-party DNS provider, whereas, in the Americas, Amazon Route 53 is the most popular provider.

Summary. We observe that more than half of the countries have more than 55% of their regional sites dependent on a third-party DNS provider and more than 43% of their regional sites critically dependent on a third-party DNS operator. Across the countries, most websites have lower redundancy in their use of different DNS providers. When comparing third-party DNS dependency across regions, we learn that Eastern Europe has the lowest third-party DNS dependence, whereas North America and some parts of Asia Pacific have a high third-party DNS dependence. The top-3 providers across all regions are the same (highly centralized) except in Eastern and Southern Europe where NsOne is among the top-3 instead of Akamai as in other regions. Additionally, the top-3 providers are responsible for 70% or more websites in each region. We also see that Cloudflare and Amazon DNS alone are used by 73% of the websites, on average, across the countries.

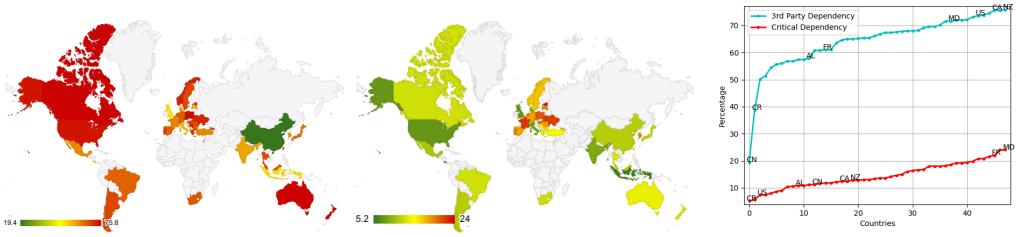
Region	DNS Providers	Avg	Std Dev	CA Providers	Avg	Std Dev	CDN Providers	Avg	Std Dev
The Americas	Amazon Cloudflare Akamai	79.1	4.2	DigiCert GlobalSign Comodo	69.1	2.7	Google, Akamai, Amazon Cloudfront	90.8	9.0
Europe	Cloudflare Amazon Akamai	74.6	6.5	DigiCert IdenTrust Inc. Comodo	70.6	5.3	Google, Akamai, Fastly	92.2	6.7
Asia Pacific	Cloudflare Amazon Akamai	81.6	4.5	DigiCert GlobalSign Comodo	72.8	7.4	Google, Akamai, Fastly	91.7	8.2
Africa and Middle East	Cloudflare Amazon Akamai	80.4	5.5	DigiCert GlobalSign Comodo	70.8	0.8	Google, Akamai, Fastly	86.4	14.9

Table 3.3. Top three DNS, CA and CDN providers with their corresponding market share per region (average percentage of websites).

3.3.2. CDN Findings

Figure 3.6(a) and 3.6(b) plot the map of CDN third-party and critical dependency in each country, again with red colored countries having the highest dependence and green colored countries having least dependence. Figure 3.6(c) plots a line graph to show the variation in the degree of CDN third-party dependency and critical dependency across countries. We find an average third-party CDN dependency across all countries of 64.1%. The country with the lowest dependency of 19.4% is China while the country with the highest dependency of 75.8% is New Zealand. Next, we aim to see the criticality of these CDNs. Critical dependency means when a website is solely hosted on one CDN. Figure 3.6 shows that the country with a maximum critical dependency on a third-party CDN is Moldova with a value of 24.0% and the country with a minimum value of critical dependency is Costa Rica with a value of 5.2%.

We also show the similarity between the CDN usage trends - the percentage of websites using more than one CDN, the percentage of websites using only third-party CDNs,



(a) Third-Party Dependency (b) Critical Dependency (c) Third-Party and Critical Dependency

Figure 3.6. CDN Third-party dependency and critical dependency of each country's top-500 sites.

and the percentage of websites using both private and third-party CDNs. Figure 3.7 plots a map indicating the degree of CDN redundancy in each country, with red-colored countries having the least redundancy and green-colored countries having the highest redundancy. Figure 3.8 plots a line graph to show the variation in overall redundancy, on multiple third-party providers and on third-party and private CDN providers across the different countries. On average 51.2% of Alexa regional sites were redundantly provisioned with CDNs, with the US having the maximum redundancy of 67.2% and China having minimum redundancy of 8.2%. 39.7% sites on average use multiple third-party CDN providers with Canada having a maximum of 59.6% and China having a minimum of 5.0%. 6.6% sites on average use third-party and private DNS with Israel having a maximum value of 43.0% and China having a minimum value of 0.0%.

We observe that more than half of the countries have a third-party CDN dependency greater than 68% and a critical CDN dependency higher than 14%. Interestingly, for some countries such as the US, Canada, and New Zealand we see high third-party dependency but low to moderate critical dependencies. Whereas, countries like France show a higher

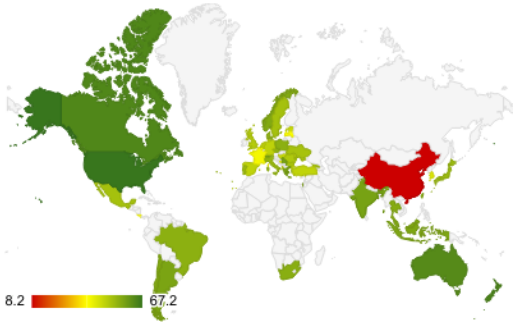


Figure 3.7. Plot of CDN redundancy

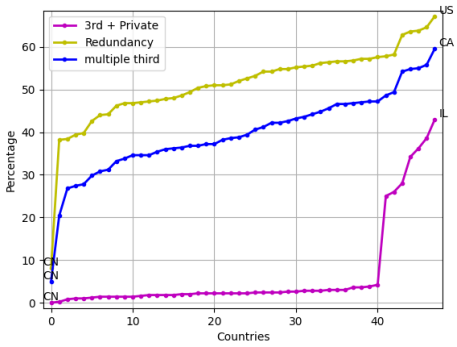


Figure 3.8. Line plot of redundant, multiple third-party, and both third-party and private CDN.

critical dependency but a lower third-party dependency. Overall, we notice lower critical dependency on the CDN infrastructure and higher redundancy since many sites today use content from multiple CDNs.

Table 3.2 shows the top-5 most popular CDNs across the countries we used for measurement, the average percentage of websites hosted on them along with the standard deviation. This shows that of the top-500 Alexa websites that use third-party CDN, 70% sites use Google as their CDN, 27% use Akamai and 19% use Fastly.

We find that only three CDNs are used to host an average of 91.5% of the websites that use third-party CDN across countries. Albania has the most CDN centralization with the top-3 CDN providers hosting 98.3% of the country's websites that are served by a third-party CDN and Denmark has the least centralization with 68.6% of the country's websites served by top-3 third-party CDNs. These results demonstrate a high degree of centralization of the CDN service. Table 3.3 shows the top-3 CDN providers that are popular

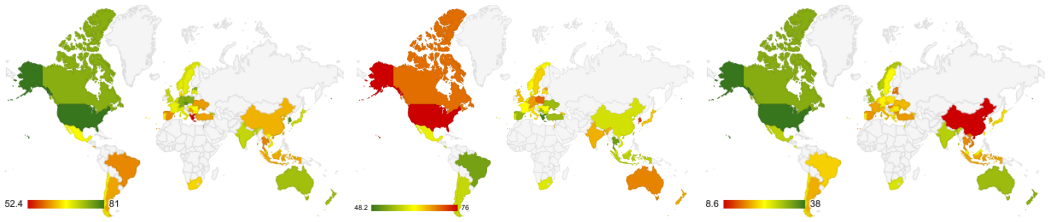
across different regions of our vantage points and the average percentage of websites using them. The results show that Google and Akamai are the top-2 CDN providers across all regions. On average, more than 86% of websites that use a third-party CDN use the top-3 CDN providers across all regions showing a high degree of centralization.

Summary. We observe that more than half of the countries have more than 68% of their regional sites dependent on a third-party CDN provider and more than 14% critically dependent on a third-party CDN. We see more redundancy (and therefore lower critical dependency) on CDN providers compared to DNS providers as more than half of the countries have greater than 53% sites using multiple CDN providers. We see a higher third-party dependency in the Americas and most of Europe and the Pacific regions and the lowest third-party CDN dependency in China. Across all regions, more than 86% of websites are dependent on top-3 CDNs. Google and Akamai are among the top-3 CDNs across all the regions with Google significantly dominating the market (average of 70% of the websites).

3.3.3. CA Findings

In the case of CA, Fig. 3.9 plots the map of HTTPS Support, CA third-party dependency and OCSP Stapling support in each country. For the CA third-party dependency, red-colored countries have the highest value and green-colored countries have the lowest value and vice versa for HTTPS Support and OCSP Stapling.

We find that the average percentage of sites using HTTPS across countries was 67.4%. This average is dragged down by countries in Latin America, and a few countries in Europe and Asia with Greece having the lowest number of websites using HTTPS (52.4%). The US has the highest rate of HTTPS adoption, with 81.0% of the top-500 sites using HTTPS. In terms of average third-party CA dependency across all countries, 61.6% percent of sites within our dataset are using a third-party CA. The results ranged from Albania at the bottom with 48.2% of its top-500 sites, and the US at the top with 76.0% of its top-500 sites using a third-party CA. OCSP stapling is much less popular. On average, 22.3% of countries' top-500 sites use OCSP stapling with China having the lowest usage at only 8.6% and the US having the highest usage at 38.0%. The low popularity of OCSP stapling is perhaps because of the lower OCSP support across web servers and browsers. For instance, the browser with the highest market share, Google Chrome, does not support OCSP stapling [**39**, **159**]. Additionally, since OCSP servers are unreliable so practically all clients implement OCSP in soft fail mode. OCSP must-staple addresses this, however, it is yet to gain widespread adoption [**39**]. None of the websites in our set support OCSP must-staple.



(a) HTTPS Support (b) Third-Party Dependency (c) OCSP Stapling

Figure 3.9. Percentage of each country's sites using HTTPS, third-party CA and OCSP stapling.

Additionally, we find DigiCert is the most popular CA in all countries' top-500 Alexa sites except for Estonia, Latvia, and Moldova. On average, 36.3% of websites in our dataset that used a third-party CA used DigiCert (including Baltimore CyberTrust certificates, which were purchased by DigiCert). The other popular CAs are Comodo CA Limited with 15.2% popularity, IdenTrust Inc. with 14.8% popularity, and GlobalSign with 14.0% popularity. Table 3.2 shows the average percentage of websites relying on each CA provider across the different countries.

Table 3.3 shows the top-3 CA providers across all regions of our vantage points and the average percentage of websites using them. We see that DigiCert and Comodo are in the top-3 CA providers across all regions. On the other hand, GlobalSign is one of the top-3 providers in all regions but Europe, whereas IdenTrust Inc. is one of the top-3 in Europe but not in other regions. Countries, such as the Czech Republic, China, and Serbia tend to be the most centralized around popular CAs, with more than 80% of websites using third-party CA choosing one from the top-3 CAs in their country. Other countries like Taiwan and Switzerland show less centralization: for these countries, less than 60% of websites use third-party CA from the top-3.

Finally, in total, we identified 68 unique CAs, 15% higher than the 59 CAs reported in Kashaf et al. for the top-100K sites. We hypothesize this is because we have a better representation of websites from different countries and thus a better representation of country-specific CAs. Some examples of country-specific CAs that we find include TWCA, a Taiwanese CA, and Microsec Ltc., a Hungarian CA.

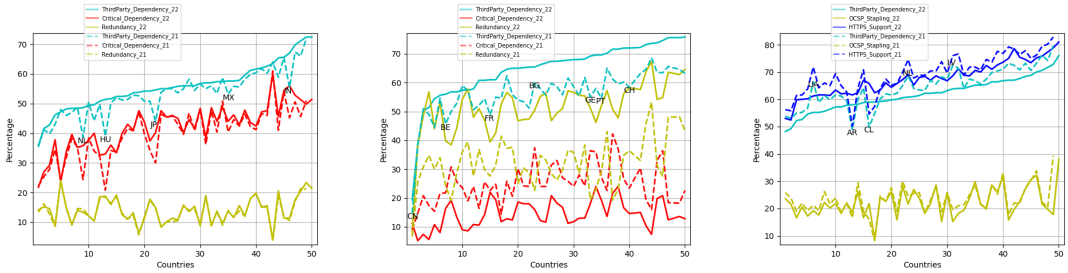
Summary. We observe that more than half of the countries have more than 62% of their regional sites dependent on a third-party CA provider. We see a higher third-party CA dependency in North America and the least dependency in South America, Eastern Europe and most of Asia. Across all regions, more than 65% of websites are dependent on top-3 CAs. DigiCert, GlobalSign, Comodo, and IdenTrust Inc. are among the top CAs across the regions with DigiCert being the most dominant CA, used by 36.3% websites on average, across all the countries except Estonia, Latvia, and Moldova.

3.3.4. Third-party dependency across services

Overall, we observed that some countries have higher third-party dependency across all of the three DNS, CDN and CA infrastructures. These countries, ranked in the order of their third-party dependency, include the United States, Australia, Canada, Singapore, New Zealand, Sweden, Norway, India and Japan(range:58%-76%). Whereas, in Europe, South America and many Asian countries (except for China) only the CDN infrastructure is responsible for higher third-party dependency. We note that China has considerably lower third-party CDN dependency(19.4%) and this may be caused by top CDN providers in our study having almost no deployments in China [**41, 67, 9, 3**].

3.4. Snapshot Measurements

To gain an initial understanding of longitudinal trends on third-party dependency and centralization around the world, we carry out our measurement campaign in two consecutive years, during April 2021 and again in April 2022. Figure 3.10 shows these trends for the different services in our analysis – DNS, CDN, and CA. For each service plot, the



(a) DNS (b) CDN (c) CA

Figure 3.10. Trends across services. Each vertical line corresponds to a country, ordered by 2022 values.

order of countries corresponds to the countries ordered by the third-party dependency of that service in 2022.

DNS. Figure 3.10(a) plots the percentage of websites, for the different countries, and each country's corresponding DNS third-party dependency, critical dependency, and redundancy. We find that on average across all countries, DNS third-party dependency increased by 4% and the critical dependency on a third-party DNS provider increased by 5%, in just one year. For both years, Cloudflare and Amazon remained the most dominant DNS third-party providers across the countries, with on average 73% websites that use a third-party provider dependent on them.

CDN. Figure 3.10(b) plots the percentage of websites, for the different countries, and each country's corresponding CDN third-party dependency, critical dependency, and redundancy in 2021 and 2022. We observe that on average across all countries, the third-party dependency on CDNs increased by 15% and the redundancy increased by 60% in one year. Average critical dependency on a third-party CDN provider, across all countries, however, decreased by 44% in one year. Google significantly dominated the CDN

market across both years (an average of 70% of the websites). Our findings show that more websites are increasingly using multiple CDNs. Though it should be noted that we do not look at criticality in terms of the content served by the CDN.

CA. Figure 3.10(c) similarly shows, for each country, the corresponding change in the country's websites having third-party CA dependency, HTTPS Support, and OCSP stapling over the two-year period. We see that across all countries, on average third-party CA dependency decreased by 4.8% and HTTPS support decreased by 2.9%. The percentage of websites supporting OCSP stapling stayed the same for most countries. The plot shows that the drop in third-party dependency corresponds with the drop in HTTPS support. Upon manually investigating sites that no longer support HTTPS, we found that all these sites return an SSL failure perhaps due to an expired certificate. However, DigiCert remained the most popular CA across the countries (serving an average of 36% websites).

3.5. Discussion

To the best of our knowledge, this is the first large-scale study of global trends towards third-party dependency and centralization across services. Our study, however, is subject to a number of limitations. First, Scheitel et al. [**146**] raises concern about the ranking opacity and stability of top lists, including the Alexa regional ranking list we rely on for our analysis. While the Tranco [**134**] list addresses some of these issues, Alexa's regional ranking offered the largest regional ranking available to the community, and includes in its listing top regional domain aliases (e.g., yahoo.com.jp and yahoo.com). Tranco allows filter-based selection of regional rankings, however, that list contains an

intersection of Tranco's global ranking and domains that also appear in the country-specific Chrome User Experience Report list. This introduces biases in Tranco's regional rankings as Chrome's user Experience list represents only a fraction of users who use Chrome as their browser, have opted in to sync their browsing history, have not set up a Sync passphrase, and have usage statistics enabled [**66**]. Between 1% to 4.5% of websites appearing in Alexa regional rankings have aliases in different countries (e.g., google.com and yahoo.com use google.co.rs and yahoo.co.rs, respectively, in Serbia). These domain aliases can have different dependencies even if they belong to the same entity; for instance, rakuten.com uses Akamai as its DNS provider and rakuten.jp uses a private DNS. Alexa's regional rankings include these domain aliases, while Tranco's regional ranking is derived from its global rankings that may not include these less globally popular, regional versions. We follow recommended best practices [**146**], as Alexa's regional ranking is regarded as the best match for our human-centered study of global third-party dependencies, and we make available the downloaded list (including download date) to enable basic replicability. We are currently exploring alternative sources of regional rankings.

Second, we characterize centralization and third-party dependencies using the list of top-500 regional websites. While we acknowledge this is a relatively limited view of the most popular sites in a region or country, it is however the largest available list.

Third, the study focuses on server-side centralization only thus the implications of centralization and third-party dependence on the end-user side are not the focus of the study. This focus allows us to rely on the use of VPN nodes as vantage points as a proxy

to study the extent of third-party dependency the users of different countries are exposed to when visiting the popular sites of the country. This same method, however, precludes the analysis of services assignments (e.g., CDN replicas) that may depend on users' DNS resolvers.

Fourth, we do not measure physical and network infrastructure dependencies; e.g., physical hosting (content providers, IT operators, common landing point in the submarine network, etc.), routing, or third-party dependencies on the web content.

Last, while we limit our analysis of dependency and criticality to 50 out of close to 200 world countries, the set of countries was selected to ensure it covers the range in terms of overlap between the top-500 regional sites and the global-ranked lists, has vantage points highly likely located within the claimed country, and, together, captures over three-fourths of the Internet user population. We are exploring alternative approaches to expand our analysis, particularly of under-sampled locations such as Africa and the Middle East.

3.6. Conclusions

We presented the first large-scale study of third-party dependency and centralization around the world. We used vantage points in 50 countries, across all inhabited continents, and focused on the most popular regional websites. We find a wide range of third-party service dependencies across countries, partially correlated to economic development, degree of economic freedom and Internet development. We also find that News and Business websites and higher-ranked websites have a higher degree of third-party

dependence across most countries. Additionally, countries that have a higher percentage of third-party dependency also have more sites hosted in the US. Despite this high variability, our results suggest a highly concentrated market of third-party providers and, perhaps more problematically, increasing levels of dependency and centralization only a year later. This work has shown that there is value in a country-level analysis of Internet infrastructure dependencies that a broader global analysis would miss.

CHAPTER 4

A Comparative Analysis of Government Hosting

For governments, however, the growing reliance on third-party infrastructure presents a particularly challenging dilemma. While third-party providers offer specialized content delivery solutions with several benefits, they also introduce significant risks, including a lack of control over data placement [**37**], in addition to multi-tenancy [**5**, **46**, **87**] and centralization [**16**, **78**, **80**, **104**, **182**, **155**, **17**]. This raises critical concerns for governmental authorities, particularly in terms of foreign interference and reliance on foreign governments, tech platforms, and infrastructure [**58**, **137**, **158**, **50**, **56**].

Although early discussions on cyber sovereignty – from data sovereignty to digital privacy and security and Internet governance – followed the 2013 Snowden revelations of widespread surveillance [**106**], various initiatives around the world have since focused on this issue in the context of geopolitical and economic tensions and the growing recognition of the Internet as critical infrastructure [**27**, **62**, **125**, **147**, **150**]. These concerns have motivated the development of legal frameworks, including the European Union's General Data Protection Regulation (GDPR) [**44**], California's Consumer Privacy Act (CCPA) [**124**], and Brazil's General Data Protection Law (LGPD) [**99**]. Together, these frameworks reflect a concerted effort to protect and manage data within the respective jurisdictions, highlighting the increasing importance of data sovereignty.

At the same time some cloud providers have began offering solutions tailored to specific governments. For example, Amazon Web Services [**60**] and Microsoft Azure [**61**] have developed solutions tailored to meet the requirements of the U.S. government. Nevertheless, for the majority of countries, third-party services are foreign-based, forcing them to strike a balance between external expertise and maintaining sovereign control over their digital assets.

Our work aims ***to empirically characterize the various ways in which governments navigate and resolve this emerging dilemma***.

Understanding this is crucial because digital transformation has fundamentally altered how governments communicate, creating new channels for disseminating policies and information while giving citizens direct access to essential services [**21, 170**]. The importance of digital government is evident in cases like federal websites in the US, which attract nearly two billion visits monthly and result in approximately 80 million hours of public interaction [**74**], and in the Asia-Pacific region where 77% of citizens primarily use a digital platform to access government services [**55**]. This transformation underscores the need for understanding the infrastructure behind public-facing government websites.

We present the first comprehensive study of hosting models employed by public-facing government digital services. Our analysis draws on data from 61 countries, covering every continent and representing over 82% of the world's Internet population. We identify government-related sites within these countries, collect resources from the landing pages of government websites and recursively crawl internal pages up to seven levels deep [**152**]. Our dataset comprises over 1 million unique resources, providing a broad

and detailed snapshot of government digital service hosting. Building on this dataset, we conduct an extensive measurement study to analyze government hosting strategies, cross-border dependencies, and the level of centralization in government web services.

Our contributions are threefold. First, we describe a methodology to characterize government approaches to domain hosting by identifying their service infrastructure and geographic location. Second, we apply this methodology to build a comprehensive dataset of government URLs and annotated networks spanning 61 countries. Finally, we present the first extensive measurement study that investigates government hosting strategies, cross-border dependencies, and the degree of centralization in government web services.

Our analysis reveals several key findings regarding governments' reliance on third-party infrastructure for data delivery. Governments predominantly use third-party providers to deliver 62% of URLs and 53% of bytes, though the adoption of these providers varies significantly across and within regions. For example, in North America, 68% of government bytes are delivered via third-party providers, while in South Asia, this reliance drops to just 5%. Neighboring countries also show contrasting patterns: Argentina relies on third-party providers for 90% of its government data, whereas Uruguay's reliance is only 2%.

Despite the overall preference for third-party hosting, 87% of government URLs in our study are served from domestic servers, although this varies regionally. South Asia, East Asia and the Pacific, and North America serve less than 10% of their government URLs from international servers, whereas Sub-Saharan Africa relies on international servers

for 48% of its URLs. Of the government servers located abroad, 57% are based in North America and Western Europe. Notable bilateral relationships include New Zealand, with 40% of its government URLs served from Australia, 79% of Mexico's URLs served from the US, and 26% of China's URLs served from Japan.

Furthermore, consolidation on third-party providers in government services appears more pronounced than in other sectors [**91**, **97**], with Cloudflare serving 49 governments—nearly double the number of the next two providers, Microsoft and Amazon. Diversification also seems correlated with reliance on third-party providers: 63% of countries that primarily use government infrastructure serve the majority of their content from a single network, compared to just 32% of countries that rely mainly on global providers.

4.1. Methodology

To characterize governments' approaches to domain hosting, we (*i*) collect government sites, and (*ii*) identify the resources they rely on, excluding those of external contractors. We then determine (*iii*) the serving infrastructure of those resources and (*iv*) their location. The following paragraphs describe this process in detail.

4.1.1. Gathering Government Websites

The first step in our methodology is to compile a comprehensive list of government sites. In this study, we focus specifically on federal-level (or equivalent) resources, including various segments of the federal administration (e.g., the presidency, ministries, and secretaries), federal agencies, often referred to as decentralized agencies (e.g., the US National

Science Foundation and the US Internal Revenue Service) and state-owned enterprises. To consider State-Owned Enterprises (SOEs), we follow the International Monetary Fund (IMF) guidelines and only include companies where the federal government holds more than 50% of the shares [**63**].

This step requires searching through a country's government sources that may provide insights into the organizational structure, identifying digital directories and authoritative resources that provide details on these structures and links to corresponding government sites. As this information is typically in the country's official languages, we rely on translation tools for this part of the process.

4.1.2. Scraping Government Websites

We scrape the collected government websites to identify the resources they rely on. For this, we use Selenium [**149**], a web automation tool, to capture the URL of each resource that constitutes the queried websites, which are then consolidated into an HTTP Archive (HAR) file. We move beyond the landing pages using the collected HAR files to recursively navigate internal pages up to seven levels deep, a threshold informed by previous work [**152**].

The geographic location of our vantage points can impact website rendering, replica selection, or determine resource accessibility, with some sites restricting access to non-domestic devices.[1] To avoid these and other potential problems, we rely on different VPN

[1]For instance, Mexico's Taxpayers Defense Attorney (in Spanish Procuraduría de la Defensa del Contribuyente, `www.prodecon.gob.mx`).

services including NordVPN [**121**], Surfshark [**164**], Hotspot Shield [**72**], to access these sites from within the target country.

4.1.3. Internal Government URLs

As we scrape seven levels deep into a government domain, we run the risk of leaving the government domain (e.g, into an external contractor's site). After completing data collection we identify internal government URLs and filter out non-government ones following the steps summarized in Table 4.1.

Approach	Method
Government TLDs	All domains including `.gov`, `.govern`, `.government`, `.govt`, `.mil`, `.fed`, `.admin`, `.gouv`, `.gob`, `.go`, `.gub`, `.guv`
Domain Matching	If the hostname of the internal page aligns with those listed in the government websites section (§4.1.1).
SAN	If the hostname is included under domains specified as Subject Alternative Names (SANs) in the TLS certificates of landing pages

Table 4.1. Steps of the methodology to identify government domains.

We first label as government resources those with domains under government top-level domains (TLDs). We adopt the pattern-matching rules defined by Singanamalla

et al. [152], which account for the different government TLDs that vary based on each country's definitions and official languages. This includes TLDs such as `.gov`, `.gouv`, `.gob`, and `.go`, among others, as listed in Table 4.1.

We then identify government resources that do not fall under government top-level domains (TLDs), either because the country does not utilize government TLDs or chooses not to use them for some agencies or state-owned enterprises (§4.1.1). If the hostname of an internal page matches the hostname of any of the sites comprised in our list of government websites, we classify it as a government hostname.

Finally, we identify government resources included under domains specified as Subject Alternative Names (SANs) in the TLS certificates of landing pages [29]. When the hostname of an internal page appears in the SANs list of landing pages, we manually verify that the hostname corresponds to a government resource. This last step allows us to select additional government-affiliated resources that may not be directly evident through their domain names or top-level domains (e.g., `orniss.ro`, `energia-argentina.com.ar`). At this stage, any hostnames that cannot be verified as government hostnames are discarded from our analysis.

4.1.4. Identifying the Serving Infrastructure

We identify the serving infrastructure utilized by government hostnames. This involves determining the IP address, Autonomous System (AS) number, organization, the registered location and the geolocation of the serving infrastructure. Table 4.2 shows an example of the information we collect for a government hostname in Uruguay.

Field	Value
URL	`www.gub.uy`
IP address	179.27.169.201
ASN	6057
Organization	Administracion Nac. de Telecom.
Registration	Uruguay
Geolocation	Uruguay

Table 4.2. An example of the information of serving infrastructure that is collected for each government resource.

To obtain registration and topological data on government website infrastructure, we connect to a VPN within the country and resolve all government hostnames to their IP addresses. Once we have the IP addresses, we determine the corresponding AS number, organization and country of registration using public WHOIS services managed by organizations responsible for IP address registration.

We then determine whether content is served from on-premise infrastructure within government-operated networks. While a recent study has made progress in identifying state-owned Internet providers [**35**], there is no dataset with annotations of government networks. We thus manually examine the entity behind all identified ASes to determine government ownership. It is important to differentiate between state-owned Internet providers – government-controlled companies participating in the Internet market – and government networks used exclusively by government institutions.

We combine various data sources to identify government ownership of networks. We examine PeeringDB records, searching for any indicator of government ownership, which may be revealed in the network's name, associated organization, or note, as in the entry of AS26810 indicating the organization as "U.S. Dept. of Health and Human

Services". We also leverage the website reported on PeeringDB records and investigate whether the associated website reveals any information that could indicate a connection with the government. Given the limitations of PeeringDB's coverage, we use WHOIS records to complement our classification. This involves querying WHOIS databases to check if the organization's name refers to the government (e.g., ministry) or the domain of the contact person's email is linked to a government domain (e.g., ".gov"). Finally, for cases where we are unable to find direct matches, we resort to Google searches. We utilize domain information extracted from WHOIS records to search for these companies' websites. This process also allows us to identify domains of state-owned enterprises that may not always be identifiable as government domains (e.g., AS27655 - Yacimientos Petrolíferos Fiscales).

4.1.5. Server Geolocation

The last step of our process consists of determining the geographic location of the infrastructure serving government websites. Given the limitations of existing geolocation heuristics and databases, we outline our specific methodology to address these challenges.

Step #1: Geolocation databases. We first query IPInfo [**86**], a widely-used open geolocation database, with the addresses of all the collected government hostnames. Darwich et al. [**48**] report that 89% of the geolocation targets in IPInfo have an error of less than 40 km (i.e., within a city).

Step #2: Identifying Anycast addresses. IP Anycast challenges latency-based geolocation. To determine if a server address is anycast, we rely on a recent data snapshot from MAnycast2, generated based on the idea of using anycast IPs as VPs to launch active measurements to candidate anycast destinations [**156**].

Step #3: Verifying country-level geolocation. To enhance the accuracy of our geolocation data, we deploy active-probing measurements to validate the reported geolocations.

For anycast addresses, we select five RIPE Atlas probes situated in the vantage country to send three pings to anycast addresses and calculate the minimum latency to each address. Our methodology integrates active measurements with the country's road infrastructure data to derive a threshold that determines whether a server address is located within a country. When the latency to a specific server address is less than the threshold, we conclude that the anycast address has servers within the country. Anycast addresses with latencies higher than this threshold are excluded from the analysis.

For unicast addresses, we also use five RIPE Atlas probes in the country assigned by IPInfo to send pings to each reported address in that country. To confirm the server's location reported by IPInfo, we calculate the minimum latency to each address and, following [**12**], check if the latency to a specific server address is less than this threshold calculated using the country's road infrastructure data. Discrepancies trigger additional verifications for unicast addresses, explained in Step #4.

Given the different shapes and sizes of countries, rather than settling for a single global threshold, we determine a per-country threshold based on the intercity road distance between the two furthest cities in that country and convert this distance into latency values.

Step #4: Geolocating Unicast Addresses. To verify the location of remaining unicast addresses we use CAIDA's HOIHO methodology [**105**], which leverages geolocation hints found in PTR DNS records, with additional regular expressions (e.g., for NTT). We also consult the cached results from RIPE's IPmap [**141**] and, if not available, we resort to active probing following a single-radius approach for geolocation.

4.2. Government Hosting Dataset

To capture a global view of trends in government hosting, we select a sample of 61 countries across all world regions, and apply our methodology for identifying government approaches to domain hosting. We first describe our criteria for including a country in our sample before providing general statistics on the collected dataset.

4.2.1. A Sample of Countries

We create a representative dataset encompassing countries from all regions worldwide. Regional divisions allow us to identify global and regional trends for governments' digital approaches. We set criteria for sampling countries across these regions, balancing

our scope with technical and logistic limitations (such as the absence of verifiable VPN servers[2] or insufficient information on e-governments).

World's Regional Slicing. To explore regional patterns in government digital strategies, we rely on the World Bank's regional division [**20**]. This division groups countries into seven regions: North America (NA), Latin America and the Caribbean (LAC), Europe and Central Asia (ECA), North Africa and the Middle East (MENA), Sub-Saharan Africa (SSA), South Asia (SA), and East Asia and Pacific (EAP).

Country Selection Criteria. Covering each region, we select countries that, combined, capture a wide range of key development indices, specifically: (1) the E-Government Development Index (EGDI) [**117**], (2) the Human Development Index (HDI) [**138**], and (3) the International Telecommunication Union/World Bank Internet Penetration rates [**161**]. This combination of indices allows us to capture a broad spectrum of countries in various stages of development and digital advancement. We integrate these indices at a regional level and select countries from five different quintiles.

While aiming for uniform coverage across these quintiles, we encounter some limitations. Specifically, the challenge is set by the lack of commercial VPN services in countries from the lower quintile, particularly in regions like Sub-Saharan Africa and Latin America and the Caribbean.

Our final selection of 61 countries from across the globe includes 2 countries from North America, 8 from Latin America and the Caribbean, 29 from Europe and Central Asia, 5 from North Africa and the Middle East, 2 from Sub-Saharan Africa, 3 from South

[2]We gained confidence in the claimed VPN locations of the countries in our set, by validating the VPN vantage points' IPs using the geolocation approach described in (§4.1.5)

Asia, and 12 from East Asia and Pacific (EAP). These countries combined represent 82.70% of the global Internet population. To access government URLs across these countries, we use 3 VPN services: NordVPN (49), Surfshark VPN (10), and Hotspot Shield VPN (2).

4.2.2. Dataset Characteristics

We apply our methodology to the set of countries in our sample. Table 4.3 offers a high-level overview of the extent and scope of our data collection.

Category	Element	Value
Government Websites	Landing URLs	15,878
	Internal URLs	1,017,865
	Total Unique URLs	1,033,743
	Total Unique Hostnames	13,483
Serving Infrastructure	ASes	950
	Govt ASes	347
	Unique IP addresses	4,286
	Anycast addresses	433
	Countries with servers located	68

Table 4.3. Landing URLs, unique hostnames and unique URLs in our dataset.

Government Websites. The dataset includes 15,878 unique landing pages from governments of 61 countries, and 1,017,865 internal government URLs obtained through scraping across seven levels. In total, the dataset comprises 13,483[3] unique hostnames and 1,033,743 distinct URLs. The vast majority of URLs, 84%, were collected directly from the landing pages, with 95% obtained from one additional level below the landing page.

[3]Note that the number of unique hostnames is less than the number of unique landing pages. This is because landing pages can include URLs like https://www.gov.br/secretariageral/pt-br, https://www.gov.br/abin/pt-br, representing different pages with the same hostname.

Internal Government URLs. We apply a set of heuristics (Table 4.1) to identify government URLs and filter out non-government ones from the set of URLs obtained. This step identified 285,767 (27.6%) internal government URLs using the government TLDs, 745,358 (72.1%) using the domain-matching approach and 2,618 (0.3%) using SANs.

Serving Infrastructure. We identified 950 ASes connecting to 4,286 server addresses associated with 13,483 hostnames. We discovered 347 (36.5%) of these ASes are operated by government entities.

We localize the serving infrastructure of the 4,286 addresses. MAnycast2 identified 433 (10.10%) of them as anycast addresses. Active-probing confirmed that 361 anycast addresses (83.37% of all anycast addresses identified) are within the country's borders. We excluded the remaining 72 anycast addresses from the analysis due to insufficient confidence in their location.

Type of Address	AP	MG	UR
Unicast Addresses	0.41	0.57	0.02
Anycast Addresses	0.83	0.00	0.17

Table 4.4. Fraction of unicast and anycast addresses validated by Active Probing (AP) and Multistage Geolocation (MG), or Unresolved (UR).

From the 3,853 unicast addresses, IPInfo identifies 3,349 addresses (86.92%) in the same country as the government they are serving and 504 unicast addresses (13.08%) outside the country borders. To increase our confidence, we tried to confirm IPInfo geolocation. Through active-probing, we confirmed the location of 40.77% (1,571) of these addresses. Through a multistage geolocation approach (§4.1.5) we confirmed the geolocation of an additional 2,198 addresses. In total, we confirmed 3,769 (97.8%) of all unicast addresses.

We exclude 84 instances where the geolocation obtained at this stage conflicts with IP-Info. Table 4.4 summarizes the output of this validation process for unicast and anycast addresses spanning 68 countries.

4.3. Trends in Government Hosting

Building on the collected dataset, in this section, we explore global and regional trends in government domain hosting and compare them with trends among popular websites. We close the section examining the similarities in governments serving strategies across the countries in our study.

4.3.1. Global Trends

We first take a global perspective, exploring governments' preferences in choosing the serving infrastructure powering their websites. *Do governments prefer on-premises or third-party hosting?* For governments opting for third-party providers, we further explore their preferences towards global, regional, or local providers.

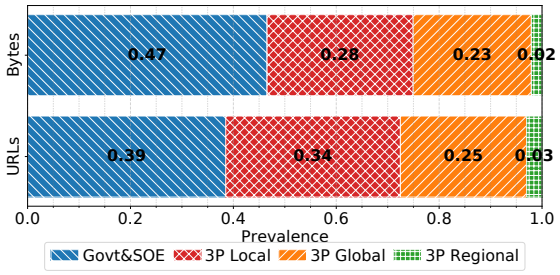


Figure 4.1. Global fraction of URLs and Bytes served by each provider category.

We examine the adoption of on-premises solutions, labeled *Government and State-Owned Enterprises* (Govt&SOE 🔵), versus third-party providers. We categorize third-party (3P) services into three groups: (1) Local (3P Local 🔴), (2) Regional (3P Regional 🟢), and (3) Global providers (3P Global 🟠), with 3P Global 🟠 defined as networks that serve governments across multiple continents, and 3P Local 🔴 as those registered in the same country as the government they serve. The remaining category, 3P Regional 🟢, includes networks registered outside the country they serve, but that do not span beyond one continent.

Using this classification, Figure 4.1 illustrates the global prevalence of each server URL category and quantifies the content by aggregating the total bytes of government URLs to account for variations in URL sizes.

Overall, governments show a preference for 3P infrastructure for data delivery, using them to deliver 62% of URLs and 53% of bytes, compared to only 39% of URLs and 47% of bytes hosted by Govt&SOE 🔵. When focusing on the categories of 3P, the figures show a more balanced reliance on Govt&SOE 🔵, 3P Global 🟠, and 3P Local 🔴 although with a preference for Govt&SOE 🔵 for bytes.

Interestingly, the analysis reveals that governments rarely consider 3P Regional 🟢, preferring to depend on their own infrastructure, collaborate with global partners, or engage with local providers. Utilizing their own infrastructure provides the maximum degree of control, but involves capital and operational expenditures. Global partners, on the other hand, offer the benefit of mature, large-scale infrastructure, while local providers

may combine the benefits of third-party expertise and specialization under government jurisdiction.

Governments vs. Topsites. To compare the hosting strategies of governments and popular websites, we select a subset of 14 countries (described in Table 4.5), including two from each region from different digital development strata and compare the adoption of third-party providers between those countries' governments and regional popular sites. We use Google's Chrome User Experience Report (CrUX) to compile a list of popular websites in these countries. To mirror our methodology, we employ VPNs and limit our scraping to resources one level beyond the landing pages. This depth limit is due to the intensive nature of deeper scraping of commercial sites (i.e., particularly broad trees) and the observation that a significant majority (95%) of government URLs are found just one level down. By leveraging the methodology described in (§4.1.4) and (§4.1.5), we then determine the serving infrastructure and geolocation of the organizations responsible for the infrastructure serving these top sites in each selected country. We also identify the fraction of non-government topsites that use either on-premise or third-party solutions to deliver content. This mirrors our government site analysis and redefines categories as (1) self-hosting, (2) global, (3) local, and (4) foreign providers. To identify self-hosted solutions, we use a heuristic from previous research [**91**, **96**].

This comparison (Fig. 4.2) shows that top sites predominantly rely on 3P Global ●, using them to deliver 78% of URLs and 74% of bytes, more than twice as commonly as government sites with 32% for URLs and 16% for bytes. In contrast, on-premise infrastructure is much more prevalent across governments, with an average of 46% of URLs

Region	Country Code
North America (NA)	Canada United States
Latin America and the Caribbean (LAC)	Mexico Brazil
Europe and Central Asia (ECA)	France Bosnia
North Africa and the Middle East (MENA)	UAE Israel
Sub-Saharan Africa (SSA)	South Africa Egypt
South Asia (SA)	India Pakistan
East Asia and Pacific (EAP)	Japan New Zealand

Table 4.5. Two countries per region were selected to compare content delivery strategies between government websites and top sites. Our selection criteria focus on capturing countries with varying levels of digital development within each region.

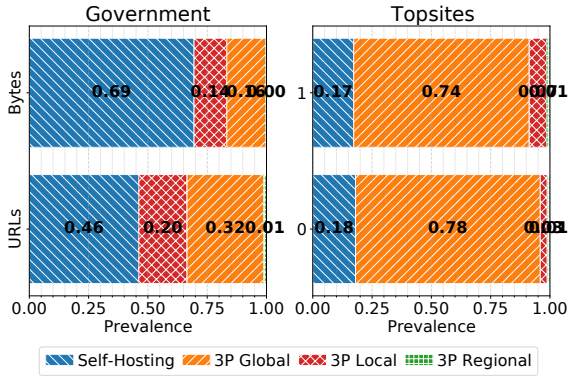


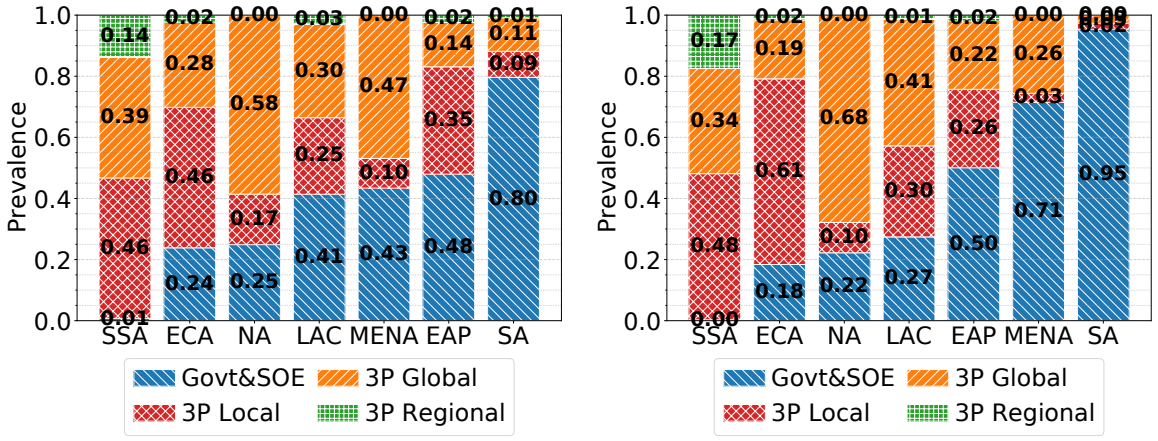
Figure 4.2. Comparison of self-hosting (on-premises) and third-party hosting between government websites and top sites within our selected subset of countries.

and 69% of bytes, compared to only 18% and 17% for top sites. The difference suggests

the relative weight of considerations, beyond market forces, behind government hosting decisions.

4.3.2. Regional Trends

In this section, we replicate our previous analysis now using the World Bank's regional division (§4.2.1) to investigate unique patterns or singularities that might exist in different regions. This regional-focused approach provides valuable insights into how factors like shared geography[4] and common cultural backgrounds may influence government decisions regarding digital infrastructure.



(a) URLs

(b) Bytes

Figure 4.3. Fraction of URLs and Bytes served by each provider category per region.

We assess both on-premises and third-party providers using the same four categories (Govt&SOE 🔵, 3P Global 🟠, 3P Local 🔴 and 3P Regional 🟢) from a regional perspective.

[4]Geographical considerations affect choices to host content with providers whose serving infrastructure is distant from a particular country.

Figure 4.8 illustrates the regional prevalence of each category, represented separately for URLs (Fig. 4.8(a)) and bytes (Fig. 4.8(b)).

Both perspectives consistently reveal a significant variation in adopting Govt&SOE ● or 3P infrastructures across different regions. For instance, in regions like South Asia (SA) and North Africa and the Middle East (MENA) most bytes originate from government infrastructures (95% and 71%, respectively). In the case of North America, most bytes and URLs originate from 3P Global ● (68% and 58%, respectively). Sub-Saharan Africa (SSA), on the other hand, delivers most of their URLs and bytes through a combination of 3P Global ● and 3P Local ● infrastructures (85% and 82%), highlighting the complexity and variability in regional hosting strategies.

4.3.3. Countries' (dis)Similarities

We conclude our evaluation of hosting trends by examining the similarities in governments' serving strategies across the countries in our study.

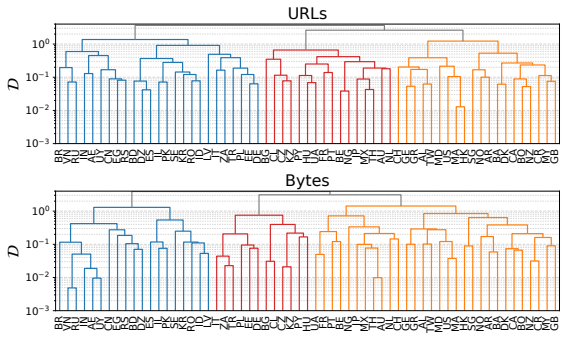


Figure 4.4. Similarities in governments' serving strategies across countries.

We use the same four categories of government hosting options and look at both URLs and bytes. The distribution of URLs and bytes across these sources creates a unique pattern, which represents the signature of a government's digital serving strategy. Our goal is to identify commonalities in these signatures across different countries.

We apply Hierarchical Agglomerative Clustering (HCA) using the Ward distance on a matrix that includes these four categories across countries, each represented by a row. This process results in the two three-branch dendograms shown in Figure 4.4. Each branch corresponds to the principal type of hosting sources (e.g., Govt&SOE 🔵).

The analysis shows the absence of strong regional patterns in government hosting strategies. For example, within the Southern Cone, Argentina, Brazil, and Chile each adopt a different approach, predominantly relying on 3P Global 🟠, Govt&SOE 🔵, and 3P Local 🔴, respectively. A similar diversity is observed in Southeast Asia, where Malaysia primarily depends on 3P Global 🟠 in contrast to Indonesia's reliance on Govt&SOE 🔵. Even more remarkable is the situation within the European Union, a region bound by a common legislative framework yet displaying varied hosting preferences. For instance, Spain, Italy, and the Netherlands each show a distinct inclination, with major dependencies on Govt&SOE 🔵 (64%), 3P Local 🔴 (93%), and 3P Global 🟠 (41%), respectively.

At the same time, it reveals similarities in the hosting strategies of countries from different regions despite having no apparent connections. For example, Brazil, Vietnam, and Russia share the same sub-tree due to their hosting similarities. We note, however, the challenges of generalizing from the observed trends and similarities. Apparent similar hosting practices may be driven by significantly different policies. In this case, Brazil's

hosting choices may be the result of a comprehensive GDPR-like regulation, known as the LGPD [**99**], whereas Russia's [**128**] and Vietnam's [**100**] hosting models may respond to a focus on data localization and state control. France and Canada, though both predominantly rely on global providers (3P Global 🟠) for hosting, differ significantly in the extent of their reliance, with 42% and 79% of bytes, respectively, sourced from these providers. Likewise, Uruguay and Indonesia, primarily depending on government and state-owned enterprises (Govt&SOE 🔵), show considerable variance in their reliance, with 98% and 58% of bytes, respectively, attributed to government sources. These examples highlight the diverse approaches and degrees of dependency on specific hosting types, even among countries with similar strategies.

4.4. Hosting Registration and Server Locations

The previous section focuses on government preferences between on-premise and third-party hosting. Even when opting for third-party service, a government could have its content hosted within its jurisdiction. In this section, we explore this aspect of hosting, specifically answering: *What are the jurisdictions where the organizations serving government content are registered? What is the location of the servers hosting the content of government sites?*

We explore these starting with a global overview (§4.4.1), followed by a regional perspective (§4.4.2), and concluding with an analysis of cross-country dependencies (§4.4.3).

4.4.1. Global Trends

We examine the country of registration and the location of the servers hosting the government URLs in our dataset. Figure 4.1 categorizes this data globally into two distinct groups: (1) Domestic ●, and (2) International ●. While a majority of the URLs, to different extents, are served from servers located within the country (87%) and from addresses allocated to domestic organizations (77%), *23% of URLs are served from internationally registered organizations and 13% are served from servers located outside the country.* Note that foreign-registered organizations of domestically provided services may still need to comply with local legislation.

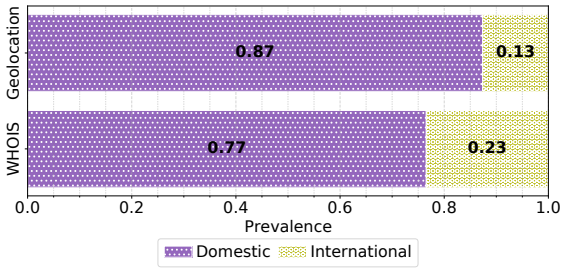


Figure 4.5. Fraction of Government URLs registered and served by Domestic or International Organizations.

Governments vs. Topsites. As in the previous section, we compare the hosting strategies of governments and popular websites, focusing on their use of domestic and international hosting solutions for the 14 selected countries.

Figure 4.6 shows this comparison, displaying: (1) the country of registration of the organization and (2) the server locations serving the URLs in our dataset for this analysis.

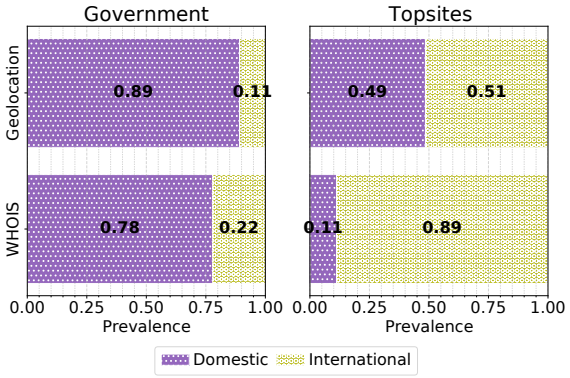


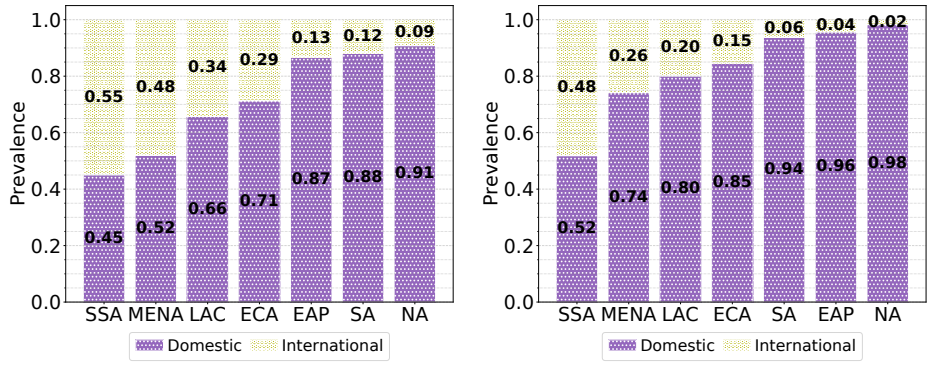
Figure 4.6. Comparison of domestic and international hosting between government websites and top sites within our selected subset of countries.

This comparison (Fig. 4.6) shows that governments predominantly opt for domestic hosting, with 78% of their URLs served by in-country registered organizations and 89% hosted within their borders. In contrast, popular websites prefer domestic hosting less; only 11% of their URLs are from domestically registered organizations, and just 49% of URLs are served from servers within their borders. This comparison highlights the different priorities between government entities, which favor control and jurisdictional autonomy, and popular websites that follow a more varied approach to digital service hosting.

4.4.2. **Regional Trends**

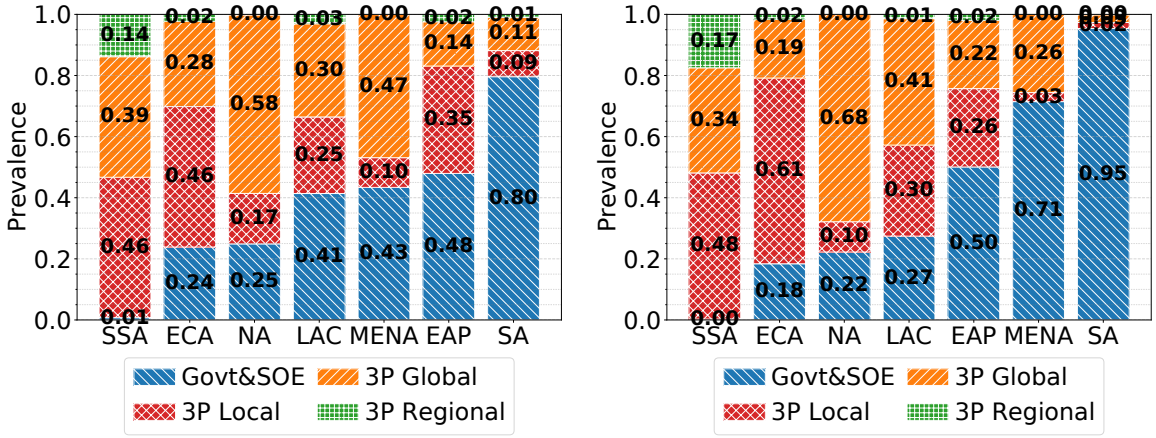
At a regional level, we analyze the country of registration and the physical location of servers hosting government URLs in our dataset.

Figure 4.7 presents this analysis, dividing organizations into two main categories: (1) Domestic and (2) International, with separate plots for their countries of registration (Fig. 4.7(a)) and server locations (Fig. 4.7(b)).



(a) Country of Registration (b) Server Location

Figure 4.7. Fraction of Government URLs registered and served by Domestic or International Organizations per region.



(a) URLs (b) Bytes

Figure 4.8. Fraction of URLs and Bytes served by each provider category per region.

While most URLs in all regions are served from servers within their respective countries, the extent of this adoption varies significantly across regions. For example, in North America (NA), 98% of URLs are served domestically, compared to the Middle East and North Africa (MENA), where this drops to 74% and Sub-Saharan Africa (SSA) where the

number of URLs hosted in the country drops to 52%. These variations are even more pronounced regarding the nationality of registrations. In North America, 91% of content is hosted by domestic companies, while in East Asia and the Pacific (EAP), Latin America and the Caribbean (LAC), Middle East and North Africa (MENA) and Sub-Saharan Africa (SSA), the percentages of URLs served by companies registered domestically are 87%, 66%, 52% and 45%, respectively. This may be partially explained by the maturity of digital markets in the US and Western Europe, where these third-party providers are registered.

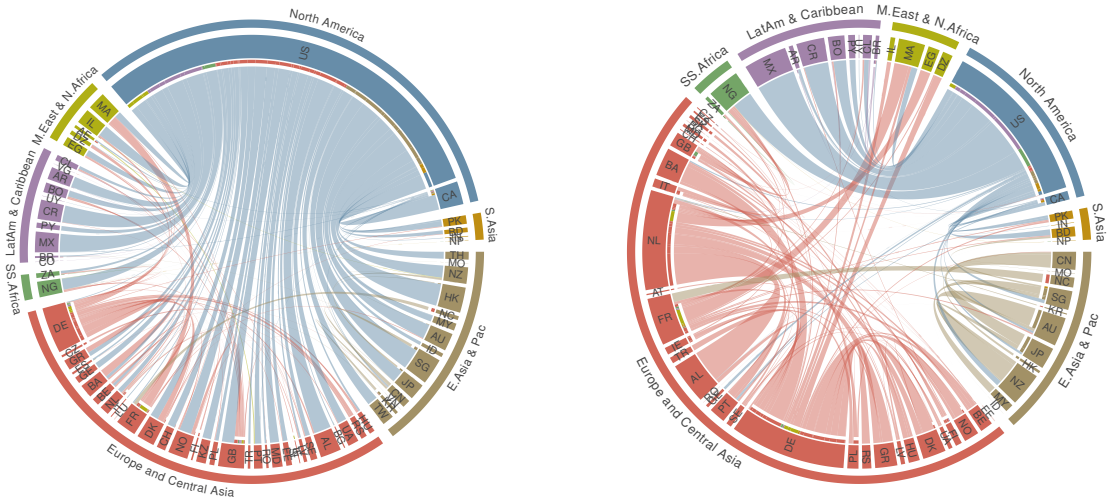
4.4.3. Cross-Border Dependencies

We now explore the cross-border dependencies of government websites to determine whether there are any preferences across the regions when selecting foreign countries from which this content is served.

Our analysis of cross-border dependencies examines both the country of registration and the location of servers from which governments' URLs are served.

Figure 4.9 presents this analysis through two circular Sankey diagrams, where countries are grouped using the World Bank's regional division, with one diagram showing the country of registration for these organizations (Fig. 4.9(a)) and the other showing the server locations (Fig. 4.9(b)). The plots reveal several interesting trends.

Inter-region dependency. This high-level analysis shows a clear trend with most governments largely relying on US-registered organizations in cases of foreign dependence.



(a) Country of Registration

(b) Server Location

Figure 4.9. **Cross Border Dependencies**. Flows represent the fraction of government resources that rely on a foreign country, either because the serving organization is registered there according to WHOIS records (Fig. 4.9(a)) or because the server itself is located there (Fig. 4.9(b)). Colors represent the region of the foreign country, while the color band connecting the flow to the foreign country represents the region of the source country that relies on it.

It also reveals that reliance on servers located abroad is generally confined to the same region; Table 4.6 shows this through interregional percentages.

There are, however, some notable exceptions, such as the Middle East and North Africa (MENA) region relying on servers in Western European countries and Latin America and the Caribbean (LAC) predominantly depending on servers in the US.

In the case of Mexico and Costa Rica, we observe significant reliance on US-based servers with Mexico hosting 79.22% and Costa Rica 49.70% of their government URLs

on servers in the US. In countries like Morocco, Egypt, and Algeria, the percentages of government URLs hosted on foreign servers are 48.38%, 21.1%, and 18.62%, respectively, similarly highlighting a pattern of dependence on international hosting solutions.

In sum, we observe that servers in North America and Europe serve 57% of government URLs crossing their respective country's borders. Brazil stands as the only exception in Latin America and the Caribbean, with only 1.78% of the URLs being served from the US, likely following Brazil's data regulation policy LGPD [**99**].

Region	%
Europe and Central Asia	94.87
East Asia and Pacific	80.79
North America	59.89
Latin America and Caribbean	3.41
Sub-Saharan Africa	2.95
Middle East and North Africa	0.00
South Asia	0.00

Table 4.6. Percentage of the cross-border dependencies that remain in the region.

Regional Affinity. When looking at cross-border dependencies within the same region – resources from other countries within the region – we find that South Africa hosts 100% of regional cross-border dependencies in Sub-Saharan Africa, Brazil hosts 85% in Latin America and the Caribbean (LAC), the US 83% in North America (NA), 76%, Japan hosts 60% in the East Asia and Pacific region and Germany accounts for 36% in Europe and Central Asia.

We also find some specific bilateral cases, such as New Zealand and Australia (with 40% of the URLs in New Zealand served from Australia). In general, we observe that 42%

of government URLs crossing their respective country's borders are served by servers within the same region.

GDPR Compliance. As part of our regional analysis, we explore compliance of government websites with the General Data Protection Regulation (GDPR) [**44**]. This EU regulation establishes that digital content within the European Union must be hosted on servers located within the member countries. Focusing on government websites, which might be more sensitive yet more likely to comply with their own regulations, we find a high level of compliance. Our analysis reveals that 98.3% (41,109 / 41,813) of URLs from EU countries are indeed served from servers within the EU's borders, indicating a strong alignment with GDPR requirements in the governmental digital sphere [**84**].

France and (former) colonies. We find interesting trends involving France with its historical and territorial connections. For instance, Morocco, which was a French protectorate from 1912 to 1956 [**173**], hosts 29.82% of its government URLs (that belong to 6 unique hostnames e.g., social.gov.ma) on servers located in France. On the other hand, 18.03% of the URLs of the French government are hosted on servers in New Caledonia, a French overseas territory in the southwest Pacific Ocean.

While New Caledonia is technically a part of France, its status is unique: it is not part of the European Union [**166**], it is an independent member of APNIC [**13**], listed by the UN as a non-self-governing territory [**171**], and has been engaged in long-standing discussions with France about independence [**167**]. Significantly, all URLs of the French government served from this territory are hosted by New Caledonia's state-owned provider, *Office des Postes et des Telecomm de Nouvelle Caledonie* (OPT-AS18200), and belong to the

hostname *gouv.nc*. This highlights the complex interplay of historical, political, and technological factors in determining the hosting locations of government digital services.

China and India. China and India, two of the world's largest economies, show contrasting trends. Despite both countries predominantly depending on their domestic and government infrastructures, the extent of their reliance varies. For China, despite historical tensions with Japan [**162, 119**], we find 26.4% of its URLs hosted by third-party providers in Japan. India, on the other hand, strongly prefers government hosting, with 99.3% of its URLs served domestically. This approach may relate to India's recent efforts to enhance data privacy, as reflected in the Digital Personal Data Protection (DPDP) Act passed in August 2023 [**82**].

Bilateral relationships and server deployments. The Dutch government adopts a singular approach to domain hosting, deploying servers abroad to support services linked to its bilateral relationships. For instance, *dutchculturekorea.com*, a cultural blog of the Embassy of the Kingdom of the Netherlands in Seoul, is hosted on a server located in Korea. Similarly, *nbso-brazil.com.br*, the website for The Netherlands Business Support Offices in Brazil, is served from a server within this South American country.

4.5. Global providers and diversification

In the last section of our analysis, we focus on the networks responsible for serving government websites. The goal is to understand the role of Global Providers in this context (§4.5.1), and the degree of diversification among government providers (§4.5.2).

4.5.1. The Role of Global Providers

We have seen that governments are also engaged, if to a lesser extent, in the trend towards adopting third-party global providers for their digital services. In the following paragraphs, we characterize these providers, examining their global footprints, and analyzing countries' reliance on them.

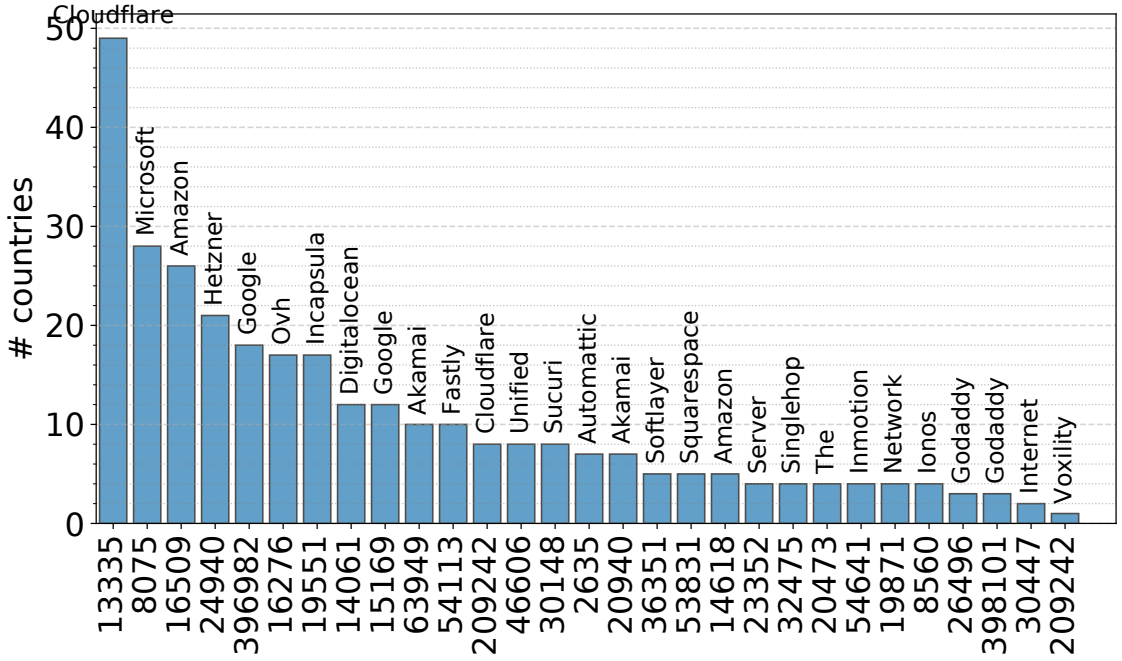


Figure 4.10. No. of countries that rely on Global Providers and CDF of Frac. of bytes served by Global Providers.

Figure 4.10 shows a histogram of the number of countries with government sites relying on one of the 28 global providers we identified. Cloudflare (AS13335) appears as the clear leader, serving content for 49 out of the 61 countries in our study. Cloudflare

is followed by two other major cloud providers, AWS (AS16509, AS14618) and Azure (AS8075), hosting content for 31 and 28 countries, respectively.

To understand the degree of reliance on any given provider, we analyze the proportion of each country's data bytes served by each provider. At the top of the list, Amazon (AS16509) stands out by serving 97% of the bytes for an East Asian country, while Cloudflare (AS13335) is responsible for 72%, 58%, and 56% of the bytes for a country in Eastern Europe, in South America, and a small Asian country, respectively. Additionally, Hetzner (AS24940) delivers 57% of the bytes for the government of a Scandinavian country.

4.5.2. Diversification of Hosting Providers

Diversification in hosting strategies can enhance the resilience of government services by reducing the risk of a digital shutdown caused by organizational failure. It also helps in creating isolation of data access across different domains. We explore whether governments tend to adopt more diversified hosting strategies and how this strategy correlates with their preference for using Govt&SOE 🔵, 3P Local 🔴 or 3P Global 🟠 for hosting their services.

To assess diversification in the networks serving government websites, we utilize the Herfindahl-Hirschman Index (HHI) [**140**], a common measure of market concentration. This index provides a score ranging from 0 to 1, indicating the level of network diversification, where a score closer to 0 indicates high diversification and a score closer to 1 indicates higher concentration. Figure 4.11 illustrates the HHI distribution for both the fraction of URLs and bytes served per network in each country. These are further

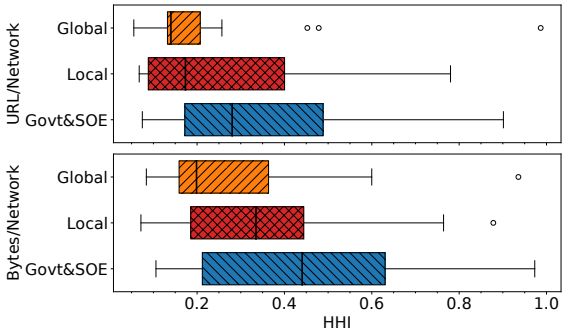


Figure 4.11. HHI distribution for the fraction of URLs and bytes served per hosting category.

categorized into three groups (Govt&SOE 🔵, 3P Local 🔴 or 3P Global 🟠) based on the predominant source of bytes for each country.

While there is some overlap in the boxplots, governments mostly reliant on 3P Global 🟠 tend to adopt more diversified strategies compared to those using 3P Local 🔴, and even more so than those relying on Govt&SOE 🔵. For example, while 63% (12 / 19) of the countries in the Govt&SOE 🔵 category serve over 50% of their bytes from a single network, just 32% (8 / 25) of the countries in the 3P Global 🟠 category depend on a single network for their bytes. Diversification is simpler with third-party providers, as it typically involves just contractual agreements. With on-premises hosting (Govt&SOE 🔵), on the other hand, diversification is more complex and may require significant capital investment.

4.6. Limitations

Our study is subject to a number of limitations. For starters, our compilation of government websites predominantly relies on self-reported information from governments

(§4.1.1). We benefited from a global trend among governments towards developing data repositories to centralize government digital resources. In some countries, this process is part of legislative initiatives, such as in Brazil with the Digital Government Law, while in others, or efforts from the executive branch, such as Argentina's Ministry of Modernization and Spain's Ministry of Digital Transformation. This data is made available in different formats (e.g., HTML items, CSV files) and through different types of resources, from webpages to GitHub repositories, as in the case of the US Cybersecurity and Infrastructure Security Agency (CISA). Despite of this, the criteria for including services on these lists vary, often due to unique governmental administrative structures, legal frameworks, and cultural idiosyncrasies among other factors.

Our findings also reveal a lack of a standard convention for naming government domain names. While numerous countries adopt the ".gov" subdomain (or variations either in English or its equivalent in other languages) exclusively for government services, there are notable exceptions. For instance, state-owned enterprises rarely fall under this categorization and may use different domain structures. Furthermore, certain countries, including Germany, Poland, and the Netherlands, do not adhere to a specific subdomain convention for their government domains, indicating a varied approach to the digital identification of government entities across the globe which may impact our data collection effort.

In addition, our methodology focuses on public-facing services and excludes resources behind login portals, so it remains unclear if the same infrastructure supports publicly accessible and restricted resources. Despite recent advancements in understanding the

potential use of Single-Sign-On (SSO) on top sites [**15**], these heuristics are not applicable to government sites that rarely accept third-party logins.

Finally, while we combine multiple approaches to minimize geolocation inaccuracies, we do not completely solve the problem. For instance, although active probing is the most accurate technique, it depends on factors such as server ICMP responsiveness and proximity of probes. In scenarios where active probing is not feasible, we resort to a multistage geolocation process, which can be costly. We opted for a conservative approach in our analysis, omitting (a small number of) IPs with geolocation from commercial databases that we could not validate.

4.7. Conclusion

We reported on the first comprehensive study exploring the hosting strategies of government digital services worldwide. Drawing from data collected across 61 countries spanning every continent and region in the world, we examined preferred hosting models for public-facing government sites, cross-border dependencies, and the level of centralization in government services. Our work provides the empirical basis for an understanding of hosting approaches in government sectors and can inform national and international policy agendas on digital sovereignty.

CHAPTER 5

Third-Party Dependency and Consolidation of Hidden DNS Resolvers

To assess the impact of consolidation trends on end-users' Internet experiences, particularly in terms of service resilience, privacy, and performance, we narrow our focus on measuring third-party and consolidation trends on the user-facing side of the DNS service.

Originally designed as a simple system to map human-readable names to network-level addresses, DNS has evolved into a cornerstone of modern Internet architecture, integral to everything from web services scalability to security enforcement [**114**]. Today DNS is a key determinant, directly and indirectly, of users' quality of experience (QoE) and privy to their tastes and preferences. It directly determines user performance as, for instance, accessing any website requires tens of DNS resolutions [**30, 25, 26**]. Indirectly, a user's specific DNS resolver determines their QoE as many CDNs continue to rely on DNS for replica selection, on the assumption that the location of a client's DNS resolver provides a good proxy for the client's own location [**73, 31**]

Traditionally, the process of DNS resolution starts with a stub resolver that queries a pre-configured recursive resolver, which then retrieves the answer by querying one or more authoritative DNS servers. As DNS has taken on an increasingly critical role, prior work has shown that the once simple, textbook model of DNS resolvers has also evolved

into a complex infrastructure. This includes ingress or forwarding resolvers, hidden resolvers, and egress resolvers, sometimes organized in cooperating pools [**148, 4**]. Despite these changes, a common assumption persists: the client-side DNS infrastructure is managed by a single organization, typically the client's ISP or, increasingly, a third-party provider.

ISP-provided resolvers are not always optimal in terms of resolution time [**2, 144**], and slower DNS resolutions can significantly impact users' QoE. Besides resolution performance, there are several other potential drawbacks of using these resolvers in terms of reliability, privacy, and censorship [**142, 24, 103**]. Partially in response to these issues, a third-party ecosystem has evolved around DNS over the years. However, this ecosystem is operated by only a handful of providers such as Google, Cloudflare or IBM, strengthening a concerning trend toward DNS consolidation and its implications [**155, 79, 174, 115**].

Earlier studies have highlighted the consolidation in public DNS services [**78, 139**] and DNS traffic market share [**116**]. However, the modern DNS resolution process is more complex, often involving multiple recursive or forwarding resolvers, increasingly managed by third-party providers. In our work, we ***investigate the degree to which end-users—either directly or through their ISPs—depend on third-party DNS recursive resolvers and the resulting consolidation of this service***.

Using all available RIPE Atlas probes we conducted a large-scale measurement campaign to capture clients' ingress and egress DNS resolvers from 803 ISPs around the world. We also complemented our dataset with a crowdsourced experiment that added 243 users in 77 ISPs. The extended dataset comprised 880 ISPs in 113 countries around the world.

Our analysis extends beyond the multilayer structure of DNS resolution to consider additional complexities, such as resolvers located in different networks and owned by separate organizations. We explore various dimensions of this mismatch, including the physical distance between clients and their ingress/egress resolvers and cases where resolvers at the country level differ from the clients' location.

Our findings reveal that approximately 47% of clients use an egress resolver outside the autonomous system (AS) of their configured DNS resolver, and nearly all of these egress resolvers (97%) belong to third-party providers. We also observed a long tail of third-party ingress resolvers, with 145 providers in total. Among these, Google emerged as the dominant provider, accounting for 5.6% of ingress resolvers, followed by Cloudflare at 3.3%. An even higher consolidation trend was evident for third-party egress resolvers, where 489 distinct providers were identified, with Google leading at 19.3% and Cloudflare following at 13.1%. These findings highlight the growing centralization of DNS resolution services.

5.1. Methodology

For our analysis we select all RIPE Atlas probes [**19**] and expand the AS coverage leveraging Amazon Mechanical Turk (AMT) [**11**] and a tool we created to capture ingress/egress resolvers of a volunteer.

Clients resolve a subdomain whose authoritative server we control. For each client, we formulate a unique subdomain that encapsulates the information about the client. This subdomain includes the IP address, country, and ID of the client and the timestamp of the request. We obtain all these fields of the client probe from the RIPE Atlas API. In the

case of crowdsourced measurements, we ask the client to enter their ingress resolver's IP on our tool's website by running a simple command, which is then embedded within the above URL as well. For privacy reasons, we embed the /24 prefix of the client's IP in our subdomain.

Upon receiving the query at our authoritative nameserver, we record the query including the client's IP, the ingress resolver's, and the IP of the egress resolver that directly contacts our nameserver. Subsequently, we use RouteViews BGP dumps [**143**] to find the AS of the client's IP, the ingress resolver's IP, and the egress resolver's IP for each client. This information is then used to map each AS to its organization using the IPInfo API [**83**] and report if there is any organizational level mismatch between the three entities.

Finally, we examine how often DNS resolutions take place in a country different than the client's country. For this, we leverage IPMap [**57**] geolocation. To gain confidence, we also use commercial IP geolocation databases [**83, 111**] which are known to be accurate at the country level [**93**]. We consider a resolver to be in a different country than the client if both geolocation methods agree.

5.2. Dataset

In this section, we briefly describe the dataset we collected using both RIPE Atlas and a crowdsourced experiment run using Amazon Mechanical Turk.

We use a total set of 12,000 connected RIPE Atlas probes and resolve the sub-domain that we control. We were able to identify the AS of all relevant entities for more than 10,000 instances, i.e. unique (client,ingress and egress) tuples and the organization of the client, ingress resolver, and egress resolver for 5,331 probes.

We use this dataset to understand the percentage of probes that use different third-party egress resolvers as opposed to the ingress resolvers selected by the client. We further investigate their probable location and their potential implications. Our Amazon Mechanical Turk experiment added 262 users in 77 ISPs and 80 additional ASes not represented by RIPE Atlas in 7 countries.

Overall, our dataset includes the perspective of 10,432 clients in 880 ISPs, spread over 113 countries around the world.

In addition to information on the client side of DNS, we use RIPE Atlas to gather latency measurements (3 pings) to probes' ingress and egress resolvers to explore the potential overhead of third-party, egress resolvers. We collected latency measurements for about 4,000 clients, ≈40% of those in our dataset.

Finally, using our implementation of IPMap [57] and the public geolocation database IPInfo, we obtained geolocation results for the 7829 pingable IP addresses. There were 5523 pingable egress resolvers and 2306 pingable ingress resolvers. If both of these methods agree on locating the resolver outside the country, we consider the resolver to be outside of the country of the client. We find that this occurs in about 11% of cases. To gain confidence in our geolocation, we also checked our results against a commercial geolocation database Maxmind and found that our results match for all cases.

5.3. Analysis

We now present our analysis of the additional aspects of complexity within the multi-tiered DNS infrastructure.

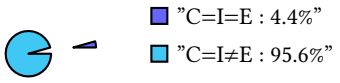
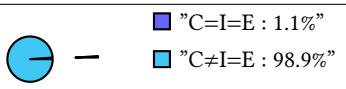
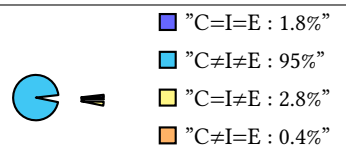
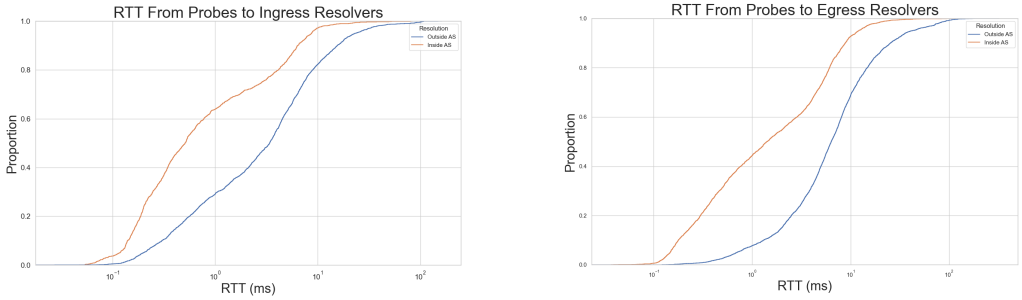
C %	C/I ASes	I/E ASes	Organizational Relationship
38%	=	=	All in the same organization
47%	=	≠	
10%	≠	=	
5%	≠	≠	

Table 5.1. AS and organizational relationship between clients (C), ingress (I), and egress (E) resolvers.

Third Party Resolution. Table 5.1 presents the AS and organizational relationship between clients, ingress, and egress resolvers. We split clients based on the relation between the ASes and organizations of the client and their ingress resolver and that of the ingress and egress resolver. We say that an ingress or egress resolver is a third party if it is in a different AS and organization than the probe.



(a) Ingress Resolver (b) Egress Resolver

Figure 5.1. CDF of the min RTT from probes to their resolvers.

We observe that the case which is commonly assumed to be the majority, that the client-side DNS infrastructure all belongs to the same organization, only happens in 38%

of cases. *In 5% of cases the client, ingress, and egress resolvers all belong to different AS and 47% of the times client and ingress are in the same AS but the egress is in a different AS.* We show that the organizational relationships match the AS relationships at a very high rate as well.

We observe a long tail distribution of usage of third party ingress resolvers with a total of 145 providers. Among these, Google stands out as the most popular choice, accounting for 5.6%, trailed by Cloudflare at 3.3%. Similarly, in the case of third-party egress resolvers, we observed 489 third-party providers, with Google again being the most common choice (19.3%), closely followed by Cloudflare (13.1%). In the cases of ingress and egress mismatch, 12% of the time the ingress resolver belonged to Google and 4% of the times to Cloudflare. Notably, AS47583 (Hostinger) exclusively routes its queries through Google. Deutsche Telekom and Vodafone Germany distribute their queries over Cloudflare and Google, occasionally using their own egress resolvers. Meanwhile, Comcast predominantly uses its own resolver but occasionally diverts its queries to Google or Cloudflare.

Latency Analysis. We analyze the latencies to the ingress and egress resolvers from the clients to understand the impact of using a third-party service. Figure 5.1(a) and Fig. 5.1(b) plot CDFs of ping latencies from the clients to the ingress/egress resolvers when they are in the same AS and outside of the AS of the client. About 5% of the probes using a third party ingress resolver have an RTT value to their resolver of more than 50ms. 15% of clients using a third party egress resolver have an RTT of higher than 50ms to their egress resolver. Overall, both figures show that the RTTs to resolvers are

markedly higher when the client is using a third party resolver. However on average, egress resolvers are slightly further away from clients than ingress resolvers. Using the idea that the speed of light in fiber is equivalent to $c_f = 2/3 * c$ and from the data in Fig 5.1, we can say that for egress resolvers, 15% of them may be more than 10,000 km away from their client. For ingress resolvers, about 5% of them may be more than 10,000 km away from the client.

Geographic Analysis. We used our implementation of IPMap as well as a commercial geolocation database to confirm out-of-country resolution. If both of these methods agree that the resolver is in a different country than the client, we consider the resolver to be outside of the client's country. We find that this occurs in about 11% of cases.

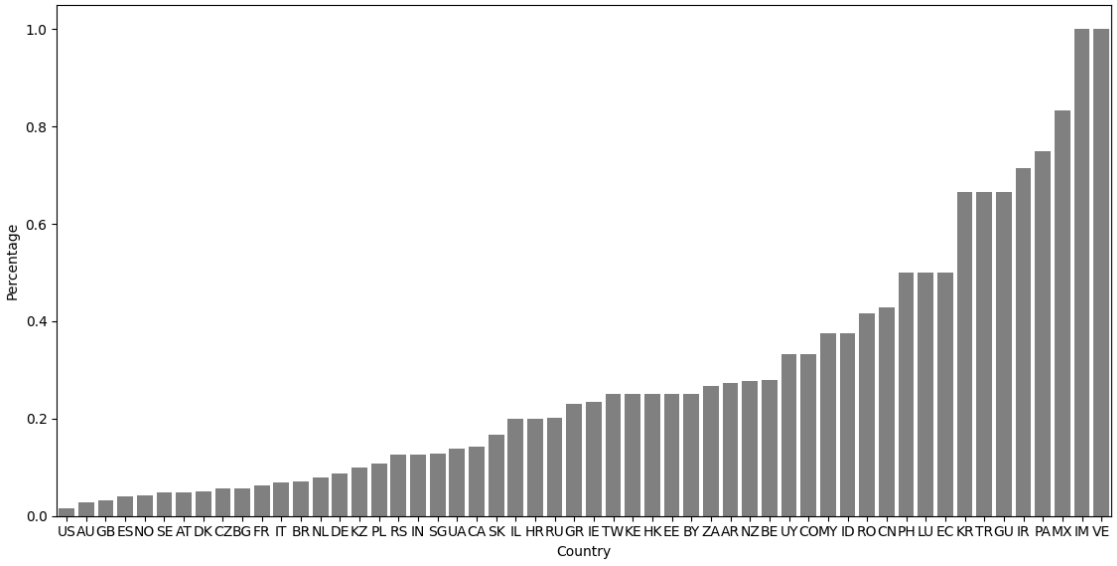


Figure 5.2. Percentage of clients using egress resolvers in other countries.

Figure 5.2 plots the percentage of clients using egress resolvers in another country. Vietnam and the Isle of Man have the highest percentage of out-of-country resolution.

According to the OONI observatory, Vietnam ISPs appear to primarily implement censorship through DNS which may partially explain our findings [**126**]. The Isle of Man is a self-governing British Crown dependency and relies largely on UK Internet infrastructure. As one would expect, large countries with extensive DNS infrastructure like the US have very few resolutions taking place outside of the country. We found several interesting cases where probes were having their DNS resolutions done in countries with which the country of origin has a tense relationship. For example, out of 92 resolvers we were able to geolocate for probes located in Russia, 17 of them were being resolved outside of the country. The most popular destinations were Finland and Germany, with some instances of resolution being performed in the US, Sweden, the Netherlands, Poland, the Czech Republic, and Japan. In the case ofArgentina, several resolutions were being performed in Brazil and Chile. We also found cases where the resolution was taking place on another continent, for instance in the case of Iran majority of resolutions were taking place in Germany and for clients in Singapore, several resolutions taking place in the Netherlands.

5.4. Related Work

There has been some related work over the years that has studied the landscape of DNS resolvers. Schomp et al. [**148**] estimate the number of open resolvers in the wild showing the complex client-side DNS infrastructure of resolver pools, and reported the significant distances DNS messages travel within the DNS infrastructure. Rula et al. [**144**] measured the use of indirect resolution techniques in cellular networks and measure the client latency to internal and external resolvers. A similar related effort has reported

on the distance between the client and their resolvers [**8**, **75**, **107**, **151**]. Vries et al. [**49**] more specifically show that DNS traffic to Google Public DNS is frequently routed to data centers outside the country even though a local data center is available in the country.

Other prior works in this space, have identified transparent forwarders and distinguished them from recursive forwarders. Nawrocki et al. [**118**] measure and analyze transparent forwarders in the DNS resolution path. Niaki et al. [**120**] use internet-wide scans to report that the majority of open DNS resolvers are DNS forwarders. Other works have measured open DNS resolvers manipulating DNS resolutions [**98**, **47**, **129**]. Callejo et al. [**34**] use a lightweight JavaScript-based methodology and the outreach of online advertising networks to distribute and run DNS tests at a global scale from the vantage point of the users and report on the fraction of global DNS lookups resolved by third-party commercial DNS providers rather than by ISP-provided DNS resolvers [**34**]. Doan et al. [**53**] use DNS measurements from RIPE Atlas to quantify the usage of public DNS services and compare their response times and IP and AS path lengths to local resolvers. However, these works do not characterize the use of resolvers forwarding their requests to other resolvers. Our work builds on some of these work and reports additional aspects of this complexity beyond the multilayer infrastructure to include resolvers placed in different networks, owned by different companies, in some instances in different countries as well, across the layers. We use the mismatch observed in our large-scale analysis to propose two implementations that can help reveal the hidden topology of DNS resolution paths.

5.5. Conclusions

We reported on a large-scale study of the complex client-side DNS infrastructure. We showed for the first time additional aspects of this complexity to include resolvers placed in different networks and owned by different companies across the layers, and in different countries as well.

CHAPTER 6

Impact of Third-Party DNS Resolver Consolidation on User QoE

Since virtually all Internet interactions begin with a DNS request, the DNS system provides a convenient mechanism for CDN replica selection. The task of replica selection involves directing users' content requests to the "best" server. This process is crucial, as it significantly impacts users' QoE. Our goal is to assess the impact of third-party DNS resolver consolidation on end-users' QoE.

However, CDN replica selection comes with several challenges. Not all content is available on every server, some servers may become overloaded or temporarily unavailable, and most importantly, the exact location of the client is often unknown, complicating the process of selecting the nearest and most efficient server.

Two popular approaches for CDN replica selection are DNS-based and IP Anycast. Over the years, the problem of CDN replica selection has garnered considerable research attention, with numerous studies proposing and evaluating alternative approaches [**108, 127, 59, 32, 7, 33, 76, 88, 163, 89, 94, 169, 135, 38, 145, 101, 1, 65, 185, 175**].

Yet, while we have gained incredible insights on the approaches' relative benefits and potential challenges, we still lack an understanding of the predominant techniques employed by CDNs for global content delivery.

With DNS-based replica selection, CDNs redirect users to content replicas by using the location of the client's DNS resolver (LDNS) as a proxy for the user's actual location [**123**]. While this method often performs well, the increasing trend from traditional ISP-provided DNS resolvers to third-party DNS providers means that not all users are located near their local DNS servers. This issue, known as the client-LDNS mismatch problem, can result in the selection of a replica server that may not be optimal from the user's perspective [**151**].

Understanding the replica selection approaches used by CDNs to deliver content globally provides valuable insights into the impact of client-LDNS mismatch problem and the impact of the broader trend towards DNS centralization on the Internet.

However, identifying these replica selection techniques is not straightforward, as CDNs do not always disclose them. Experimentally determining these techniques is further complicated by several factors influencing redirection, including CDN and DNS deployment, regional anycast borders, network and server conditions, and even the specific content, as different CDN customers may rely on different replica selection techniques.

As our first main contribution, we present a methodology to experimentally determine the main replica selection approach used by a CDN. Our methodology uses vantage points around the world and a set of DNS resolvers, placed at varying distances from the vantage point. These vantage points resolve resources hosted by each CDN using the different DNS resolvers, and collect latencies to the assigned CDN replicas. We show how the relative differences between the distributions of latencies to these replicas can be used to identify the CDN replica selection approach. We validate our methodology by

leveraging a diverse set of large-scale CDNs that exemplify the prevalent replica selection strategies employed in popular websites: Akamai, Cloudflare, and Edgio. While the redirection mechanisms of these major CDNs are well-documented, there exists over 80 CDNs world wide, many of them small and with no documentation on their redirection approaches. Despite their size, these smaller CDNs play a crucial role in shaping user experience in specific countries and regions.

As our second contribution, we apply our methodology to identify the main replica selection approach used for serving the most popular Web content around the world. To this end, we select 19 countries in every inhabited continent capturing, overall, 66% of the Internet population worldwide (between 50% and 89% of the Internet population of each continent). We apply our methodology to identify the main approach of the set of global and regional CDNs used to serve all resources of the top 1,000 sites for each country[1]

Our analysis shows, among other findings, that 12 out of the 17 CDNs we measure use an DNS-based redirection model. While anycast-based replica selection is used to deliver the largest fraction of resources worldwide (40.8%), the DNS-based approach is the preferred replica selection for delivering the majority of bytes (40.9%) on popular un-logged-in landing pages. Not surprisingly, the predominant approach varies across regions with Anycast being dominant in Europe, and DNS-based dominating in Oceania and North America.

[1]Selected based on Google CrUX dataset [**68**].

6.1. Background

Two popular approaches for CDN replica selection are DNS-based and IP Anycast. With DNS-based replica selection, CDNs redirect users to the content replicas using the location of the user's local DNS resolver (LDNS) as a proxy for the user's location [**151, 108, 127, 51**]. While DNS-based redirection often performs well, not all users are in close proximity of their DNS servers – something referred to as the *client-LDNS mismatch problem* – and thus a replica server chosen based on the user's LDNS may not be the optimal choice from a user's perspective.

End-user replica selection, commonly implemented via a DNS extension, aims to address this problem. The EDNS0 Client Subnet extension (ECS) allows a recursive resolver to specify a prefix of the user's IP when requesting domain name translations on behalf of a user [**45**]. The authoritative DNS can use that client-specific information to make accurate per-client decisions [**45**]. While a few large-scale DNS services and some CDNs have adopted ECS, the adoption of ECS across the Internet is limited, and the volume seen by authoritative resolvers is due to the popularity of a few public resolvers, such as Google Public DNS and OpenDNS [**32, 127**]. For instance, Akamai, one of the most popular CDNs worldwide [**97**], restricts ECS requests to a limited set of public resolvers [**38**], while Cloudflare has stopped supporting it altogether due to privacy concerns [**42**].

The second most common CDN replica selection approach relies on IP anycast [**7, 33**]. Some CDNs avoid the infrastructure and operational costs of DNS-based and rely instead on anycast, announcing the same IP address(es) from multiple distributed locations and

redirect user traffic to nearby CDN sites. While simpler to implement, IP Anycast relinquishes the routing decisions from the CDN operators to the network and prior work have shown may result in users redirected to suboptimal sites, with a significant latency overhead [**33**, **101**].

To leverage the simplicity of operation and control of IP anycast while addressing its limitations, some CDNs have adopted regional anycast [**71**, **112**, **183**]. Using this approach, a CDN partitions its servers into different regions (e.g., loosely following continent or subcontinent lines such as "North and South America" or "Europe, Middle East, and Africa") and announces a distinct IP anycast prefix from each region. When a user makes a request to the CDN, the CDN DNS infrastructure maps the user to a regional IP anycast address based on the – very coarse – user's location.

It is worth noting that a given CDN can potentially use one or a combination of these approaches, selected perhaps based on specific customer, application or location.

6.2. Methodology

In this section, we describe a methodology to experimentally identify a CDN's redirection technique. The methodology uses RIPE Atlas probes [**19**] as clients and a set of strategically selected DNS resolvers. These clients resolve CDN hosted resources, using the different DNS resolvers, and collect latencies to the assigned replicas. The CDN replica selection approach is identified based on the relative differences between the collected latency distributions. The following subsections describe this approach in detail and illustrate its use.

6.2.1. Finding CDN Server Assignments

As a first step, clients perform resolutions of a CDN hosted resource in a given locale, using a set of five (5) selected DNS recursive resolvers. These five resolvers cover different geographic *scopes*, namely: (i) the same metro area as that of the client, (ii) a different metro area in the same country as the client, (iii) a different country within the same region as the client, (iv) neighboring region of the client, and a (v) a non-neighboring region (Fig. 6.1).

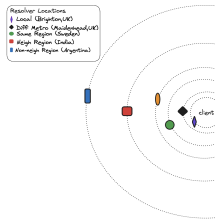


Figure 6.1. DNS resolvers at different distances from the client.

Region	% of Internet Population
North America	89.3
Oceania	75.0
Asia	70.1
South America	60.0
Europe	60.0
Africa	50.1
Worldwide	66.0

Figure 6.2. Internet Population represented by our vantage points.

To illustrate, a client in Brighton, UK could use DNS resolvers located in Brighton, UK (same metro, same country), Maidenhead, UK (same country, different metro), Sweden (same region, different country), India (neighboring region) and Argentina (non-neighboring region).

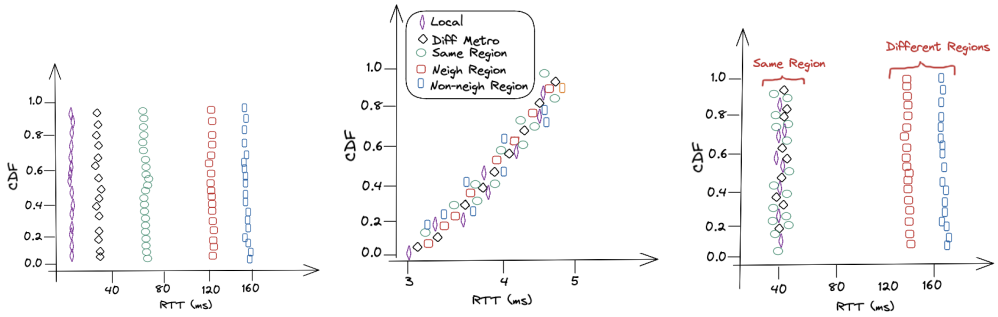
This process is repeated for a large set of resources hosted on the target CDN. Clients collect latencies to the assigned CDN replicas with different DNS resolvers. The resulting

latency distributions, one per DNS resolver scope, are subsequently used to infer the CDN redirection technique.

6.2.2. Canonical Examples for Illustration

Intuitively, if the replica selection approach of a CDN takes into account the specific metro location of a client (Brighton in Fig.6.1), the set of replicas it assigns should be different from (and closer than) what it would assign if the client's DNS resolver were located in a different country (Sweden) or a different region (India).

On the other hand, if the CDN's replica selection approach does not differentiate between clients in different regions, then the set of replicas it assigns when using a client's local DNS resolver, and a DNS resolver in any other neighboring or non-neighboring region (Argentina) should be similar.



(a) DNS-based (b) Anycast (c) Regional-Anycast

Figure 6.3. Expected CDF sets for DNS-based ($CRV[i] \approx 1$ for all i), Anycast ($CRV[i] \approx 0$) and Regional-Anycast replica selection ($CRV[1] \approx 0$, $CRV[3] \approx 1$).

Before detailing our approach, it is useful to visualize a potential set of latency distributions to the replicas assigned by prototypical versions of replica selection approaches. Figure 6.3 show these latency distributions for DNS-based, Anycast and Regional-anycast.

If a CDN uses DNS-based replica selection, the set of replicas assigned by it should be further away from the client as the DNS resolver used also moves away. Correspondingly, the distributions of latencies to the assigned replicas for these resolvers should be spaced and away from the client (i.e., higher RTTs). Figure 6.3(a) illustrates this case, with five distributions of latencies to assigned replicas using the five different DNS resolvers (local, different metro, etc).

When a CDN uses Anycast for replica selection, the location of the client's DNS resolver is not a factor for the selection. Consequently, the distribution of latencies to replicas assigned using the different resolvers should largely overlap. Figure 6.3(b) illustrates this case.

Last, if a CDN uses regional anycast for replica selection, we expect the distributions of latencies to assigned replicas within the same anycast region to be similar, but to the left of (i.e., shorter than) those associated with DNS resolvers in different anycast regions. Figure 6.3(c) shows this case, with the latency distributions of the first three DNS resolver locations – local, in a different metro, and in a different country of the same region – clustered together, while those corresponding to different – neighboring and non-neighboring region – set clearly apart.

6.2.3. Identifying Redirection Techniques

To identify three of the main CDN redirection approaches – DNS-based, Regional Anycast and Anycast – we compare the distribution of latencies to the replicas assigned with different resolvers.

We capture the distance between the latency distributions in the *coefficient of Regionalization Vector (CRV)*. The CRV is a four-tuple vector where each entry, $CRV[i]$, is the distance between the distribution associated with the DNS local resolver and that of the $(i+1)$th resolver scope. For instance, $CRV[0]$ captures the distance between the latency distributions of the local resolver and a resolver in a different metro area, while $CRV[3]$ measures the distance between the local resolver and a resolver in a non-neighboring region.

To compute the distance between two latency distributions we use the Kolmogorov–Smirnov (KS) distance, the maximum difference being taken over all values of x of the absolute differences between two latency distributions F and G ($D = max|F(x) - G(x)|$). [2] Using the cumulative distribution function (CDF) of latencies, each $CRV[i]$ ranges between 0 and 1, being equal to 0 when both distributions are identical and 1 when they are completely different.

We rely on CRV to identify three of the main CDN redirection approaches: DNS-based, Regional Anycast and Anycast. For the prototypical cases of DNS-based replica selection, we expect that at least $CRV[1]$ and $CRV[3] \approx 1$. Other entries in the CDN's CRV could range between 0 and 1 depending on the scale of deployment of the particular CDN.

[2] We choose KS distance as it does not assume any specific distribution, making it particularly useful for latency data, which often do not follow a normal distribution.

For instance, the CDFs of a CDN like Akamai using DNS-based replica selection with a large-scale deployment of over 170,000 servers in 1,300 ISPs [**184**], should look similar to our prototypical case (Fig. 6.3(a)) and have $CRV[i] \approx 1$ for all i. The Anycast-based replica selection approach, on the other hand, should result on all CDFs overlapping, stretching on the latency range from the closest to the furthest away replica, and a $CRV[i] \approx 0$.

Finally, if the CDN uses Regional-anycast for replica selection we would expect the CDFs of latencies to replicas in the same region to be the same, $CRV[1] \approx 0$, and different from the CDFs of latencies to replicas associated with non-neighboring regions, $CRV[3] \approx 1$. Note that the regional divisions used by the measurement and the underlying infrastructure of the CDNs may not match exactly, as the CDN's regions are determined by factors such as the underlying infrastructure and Internet penetration. For instance, while some CDNs differentiate between northern and southern Europe, most CDNs treat all of South America as a single region. This mismatch is however mitigated by our selection of multiple resolver scopes, allowing the system to account for varying regional divisions, while capturing the overall behavior of each system.

We choose to use the distances between latency distributions to the assigned replicas to identify redirection approaches, instead of comparing the set of IP prefix of these replicas. While the latter approach might initially seem simpler, its complexity quickly escalates when addressing necessary details such as determining the right prefix length to use and establishing the correct similarity thresholds for identification. Additionally, although techniques like MAnycast2 [**157**] can determine whether the IP prefixes used

by CDNs are Anycast, they do not provide insights into the redirection model employed by a CDN in a given country.

6.2.4. Checking for ECS Support

Besides relying on one (or more) of the above replica selection approaches, a CDN may support ECS to make per-client decisions and avoid the client-LDNS mismatch problem. For ECS to function, both the client's LDNS and the CDN must implement the ECS specification. Calder et al. [32] examines ECS adoption in LDNSes using DNS queries captured over one month from Microsoft's Azure Cloud platform.

We detect support for ECS by a CDN by issuing ECS resolution queries for its domains and the CNAME chain, using the Google Public DNS resolver – known to support ECS – and the nameserver of the domain. We check for EDNS scope greater than zero to determine if the domains supports ECS. We issue three queries and compare the assigned CDN servers with two distant subnets. If the two vectors do not match, and include responses aligning with the geolocation of the provided subnet, as indicated by geolocation databases, it suggests that the given CDN in fact uses the eDNS0 subnet extension for replica assignment.

6.3. Detecting CDN Replica Selection

We now describe the application of our methodology to experimentally identify the replica selection approaches used by CDNs to deliver Web content around the world. In the following section (§6.4) we validate our approach contrasting the known approaches used by three large CDNs with what our methodology identifies.

We follow the United Nations geo scheme [**176**] and rely on published statistics on Internet penetration [**160**] to select vantage points for our study. We ensure that the set of countries included in our analysis (*i*) capture a sufficiently large fraction of the Internet user population, and (*ii*) host VPN vantage points whose locations can be verified.

We place vantage points in 19 countries that span all inhabited continents and together capture 66% of the world's Internet user population. Table 6.2 shows the regions and their corresponding Internet user population captured by the countries included in our study. For each continent, the selected set of countries account for $\geq$ 50% (and up to $\approx$ 89%) of the continent's Internet population.

For each of these countries, we collect the top 1,000 sites based on the Google CrUX dataset [**68**]. These popular sites should serve as a good proxy of the most commonly accessed content by users in each country, while the aggregate across countries and continents serve as a good starting point for a global study of commonly used replica selection approaches.

We use popular commercial VPN providers [**121, 72, 164**] in each country to gather the resources of these top sites and the CDN(s) hosting them. To collect all resources of each website, we generate and utilizate the HTTP Archive (HAR) file for each website. To find the set of CDNs used by the resources of top sites in each country, we find the Canonical Name (CNAME) records for all website resources and obtain the set of CDNs from our self-populated CNAME-CDN map [**40, 180**]. Additionally, for validation and to identify the CDNs hosting resources without a CNAME redirect, we compare the autonomous system number (ASN) of the resource with those of popular CDNs [**110, 180**].

We validate the claimed locations of our VPN providers, by geolocating the VPN vantage points' IPs using two popular geolocation databases: MaxMind GeoLite2 [**111**] and IP2LocationBD11.Lite [**85**]. Past work has shown that geolocation databases are reliable at the country level [**136**].

We use RIPE Atlas probes as proxies of clients in each of the selected countries. Next, we select resolvers in each geographic scope that our client will use to resolve the CDN hosted resources. For the local resolver scope, we use the client probe's resolver and for the resolvers in other geographic scopes, we select Regional DNS resolvers [**52**] in each of those scopes. [3] We validate the locations of the selected DNS resolvers following a similar approach as with the VPN servers. The clients then issue DNS resolution queries for the resources to the selected set of DNS resolvers. We collect RTTs from the RIPE Atlas probes to the different CDN replicas assigned for each resolver scope by issuing a sequence of three ICMP pings. We record the minimum RTT from the repeated runs to ignore any transient spikes in the latency. Finally, we use the collected latency distributions to compute the CRV of each CDN for different regions.

6.4. Known CDNs for Validation

Before presenting our analysis of CDNs' replica selection approaches around the world, we confirm the findings of our methodology when applied to a set of large CDNs – Akamai, Cloudflare, and Edgio – with known replica selection approaches. We confirm

[3] A sensitivity analysis using various combinations of cities within the client's vantage country, different metropolitan areas and countries for DNS resolvers, and different DNS resolvers within each geographical scope, revealed no variations in the identified redirection models.

the redirection models used by these CDNs by checking their websites, previously published works, personal communication, and third-party websites such as CDN Planet [**131**]. We observe that the redirection models we infer from our analysis coincide with those reported by these CDN providers.

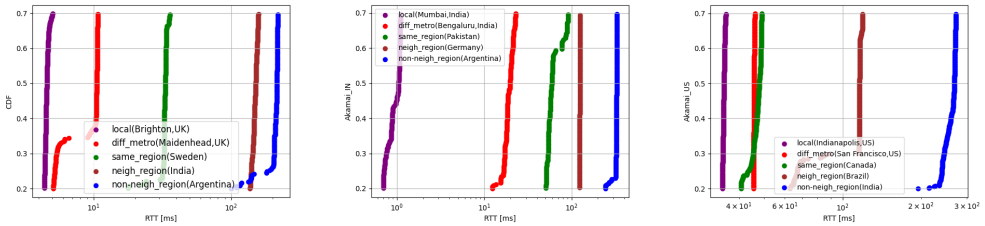
6.4.1. Replica Selection by Akamai

We focus first on the replica selection approach identified by our methodology for Akamai. We run our data collection and analysis using all collected resources hosted by this CDN, and computed the CRV vectors for each country.

Figure 6.4 shows the CDFs of latencies to the set of Akamai replicas assigned for each resolver scope in the UK, India, and the US. We plot the middle percentile for all CDNs across different countries as the tail may contain edge cases, including non-cacheable content (retrieved from the origin). The clean, non-overlapping CDFs at most resolver scopes resembles that of our prototypical DNS-based approach of Fig. 6.3(a). Correspondingly, we find values for $CRV[1]$ and $CRV[3]$, ranging between 0.9 to 1.0, across these countries as shown in Fig. 6.7(a).

Our analysis clearly shows that Akamai, as reported [**38, 122, 130**], relies on DNS-based replica selection across all countries in our dataset. In addition, the Akamai CDN supports ECS for replica selection as we show in Sec.6.5.4.

The CDFs associated with resolvers in the *different metro* and *same region* in the US are interesting, showing the occasional mismatch between infrastructure and geographic scopes (Fig. 6.4(c)). With a vantage point in Indianapolis, US, the replicas assigned with



(a) Akamai-United Kingdom (b) Akamai-India (c) Akamai-US

Figure 6.4. Akamai using DNS-based redirection technique in India, United Kingdom, and the US.

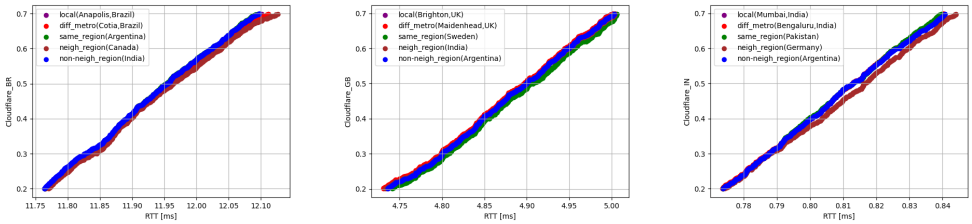
a resolver in Canada are at a similar latency than those assigned with a resolver in San Francisco, the same country but more than 2,200 mi away.

6.4.2. Replica Selection by Cloudflare

The second CDN we use for validation is Cloudflare. Cloudflare is known to rely on anycast for replica selection [**43**, **132**]. As in the case of Akamai, we run our data collection and analysis using all collected resources hosted by Cloudflare and computed the CRV vectors for each country.

Figure 6.5 plots the latency CDFs to the replicas assigned for Cloudflare-hosted resources in Brazil, the UK, and India. Across the three countries, the figure shows similar latencies to replicas obtained at each resolver scope with overlapping CDFs. The three figures are nearly equivalent to those of our prototypical Anycast case (Fig. 6.3(b)).

Figure 6.7(b) shows the $CRV[1]$ and $CRV[3]$ values for Cloudflare in the same countries, each with the expected values ≈0. Our analysis confirms Cloudflare's use of anycast for replica selection.



(a) Cloudflare-Brazil (b) Cloudflare-United Kingdom (c) Cloudflare-India

Figure 6.5. Cloudflare using Anycast redirection technique in Brazil, United Kingdom, and India

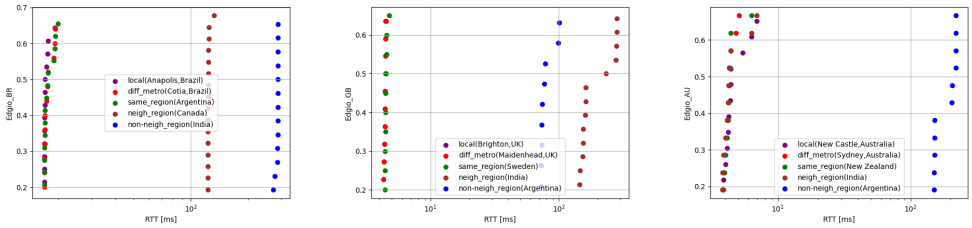
6.4.3. Replica Selection by Edgio

Last we examine Edgio, one of the few CDNs using regional anycast for replica selection [**112, 133**].[4]

Figure 6.6 shows CDFs of latencies to Edgio replicas from clients in Brazil, the UK, and Australia. The figure shows similar latencies for all resolver locations within the same region and non-overlapping CDFs for resolver scopes in a different, non-neighboring region. Figure 6.6(c) for Australia illustrates an interesting case of mismatch between the geographic divisions and the underlying CDN infrastructure. The CDFs associated with a DNS resolver from the same country, but different metro (Sydney) overlaps with the one associated with a DNS resolver in a neighboring region (India). Further examining the responses revealed that for 81% of the Edgio objects we measure, the neighboring

[4]EdgeCast (AS15133) and Limelight (AS22822) became Edgio in 2022; we focus on measurements of the EdgeCast Network.

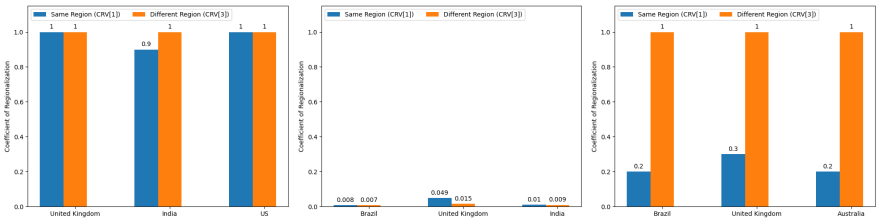
region responses received the same IP address as the different metro, suggesting the responses are part of the same anycast infrastructure region. Personal communication with operators confirmed our findings in this case.



(a) Edgio-Brazil (b) Edgio-United Kingdom (c) Edgio-Australia

Figure 6.6. Edgio using Regional Anycast redirection technique in Brazil, United Kingdom, and Australia

Figure 6.7(c) shows the $CRV[1]$ and $CRV[3]$ for Edgio in these three countries. As expected, $CRV[1]$ has a lower value ranging between 0.2 to 0.3, and $CRV[3]$ has a higher value of 1, confirming that, as reported, Edgio uses Regional Anycast as its CDN replica selection approach.



(a) Akamai (b) Cloudflare (c) Edgio

Figure 6.7. Coefficient of regionalization for same ($CRV[1]$) and different regions ($CRV[3]$) for three larger CDNS.

6.5. Global Analysis of CDN Replica Selection

After confirming the findings of our methodology with those advertised by the CDNs, we now turn to the remaining CDN providers observed in our scan of each measured country's top sites. We analyze the geographical trends in CDN redirection techniques used around the world and report the redirection technique predominantly used by the CDNs in our dataset, based on our measurements. We close the section reporting on our analysis of ECS support.

6.5.1. World CDNs Replica Selection

We identify the replica selection approaches of 14 other CDNs around the world, including large CDNs such as Amazon CloudFront, and Fastly, as well as regional providers such as Azion in Brazil, NGENIX in Russia, and Taobao and Tencent in China. The Google CDN presents an interesting challenge and we discuss it separately. Due to space restrictions, we focus our detailed discussion on a subset of these CDNs.

6.5.1.1. Global CDNs. Amazon CloudFront is a global CDN operated by Amazon Web Services with servers located in Western Europe, Asia, Australia, South America, and Africa, and several major metro areas in the US. According to the company site, CloudFront uses a global network of over 450 PoPs in more than 49 countries.

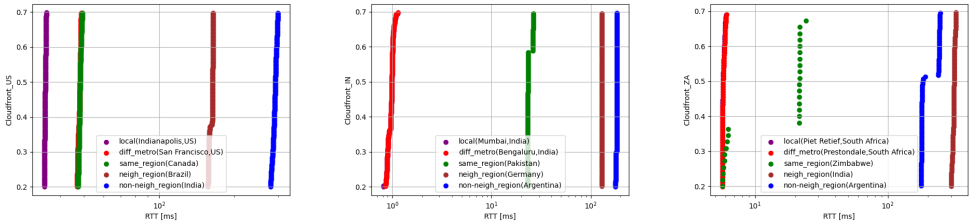
Figure 6.8 shows CDFs of latencies to Amazon CloudFront servers assigned at each resolver scope in the US, India and South Africa. Although different in scale than Akamai, the CDFs for different resolver scopes are mostly non-overlapping indicating the use of DNS-based replica selection. The overlap in the narrower resolver scopes (local and

different metro in Russia, and the metro and same region scopes from the US vantage point) indicates the difference in scale: both are directed to the same or similar replicas.

The South Africa CDF for same region (Zimbabwe) is split in two largely vertical but separate lines, where nearly 40% of the selected replicas are in the same latency range as those assigned with DNS resolvers in other metro areas within South Africa, indicating domain dependent behavior for the region.

To confirm our identification of Amazon CloudFront's replica selection approach, we look at the corresponding CRV. CloudFront shows high values for $CRV[1]$ and $CRV[3]$, ranging between 0.8 to 1, in all three countries (Fig. 6.10(a)).

Furthermore, we also checked for the presence of prefixes associated with Amazon's services beyond CloudFront [**10**], aiming to investigate potential variations in redirection models employed by Amazon across different services and applications. However, we found minimal number of such prefixes in our collected data.



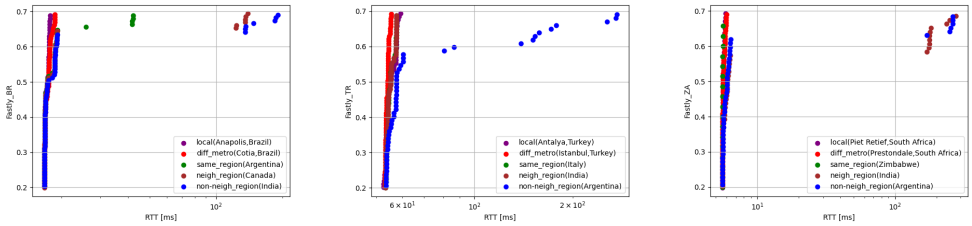
(a) Amazon-US (b) Amazon-India (c) Amazon-South Africa

Figure 6.8. Cloudfront matches DNS-based redirection technique, as shown in the US, India and South Africa.

Fastly is another global CDN, with $\approx$ 80 PoP on 6 continents, known as a real-time CDN it offers a range of other services including streaming media and private CDN.

Figure 6.9 shows the different CDFs of latencies for all DNS resolvers' scope and clients located in Brazil, Turkey and South Africa. It presents, predominantly, the patterns of Anycast replica selection with a large overlap of all CDFs, across DNS resolver scopes and relatively low coefficients of regionalization. We see a few Fastly responses showing DNS-based replica selection, although with a deployment scale significantly smaller than Akamai's. These modes may reflect different customers or configurations.

Fastly's $CRV[1]$ and $CRV[3]$ range between 0.2 to 0.4, across these countries, as shown in Fig. 6.10(c). We observe that the coefficient of regionalization in the case of Fastly is much lower than Akamai but not as low as Cloudflare.



(a) Fastly-Brazil (b) Fastly-Turkey (c) Fastly-South Africa

Figure 6.9. Fastly predominantly matches Anycast redirection, as shown in Brazil, South Africa and Turkey.

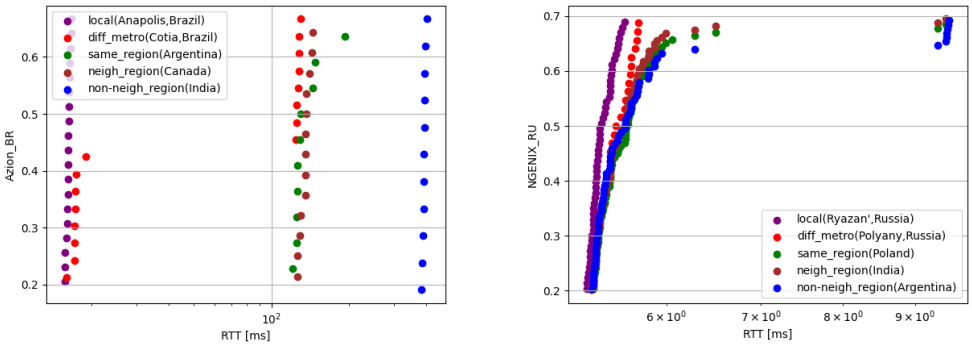
6.5.1.2. Regional CDNs. We focus now on regional CDNs, targeting small, perhaps region or country-specific, markets with correspondingly smaller presence in the top-ranked websites. While not the goal of this work, a regional focused study with larger number of websites in the region could reveal other small, regional CDNs and/or customer-specific replica selection approaches.

(a) CloudFront (b) Google (c) Fastly

Figure 6.10. Coefficient of regionalization for same ($CRV[1]$) and different regions ($CRV[3]$) for other CDNs.

Figure 6.11 shows CDFs for two of the regional CDNs covered in our study: Azion in Brazil and NGENIX in Russia. The latency CDFs for Azion shows the clear separation of DNS-based replica selection, and with $CRV[1]$ and $CRV[3](\approx 1)$.

NGENIX, on the other hand, predominantly matches the Anycast model with overlapping CDFs across resolver scopes (for most hosted content) and low values of $CRV[1]$ and $CRV[3](\approx 0.3)$.



(a) Azion-Brazil (b) NGENIX-Russia

Figure 6.11. Redirection techniques used by regional CDNs.

Country	DNS(%)	Anycast(%)
France	43.8	56.1
Germany	42.1	57.9
Russia	63.7	36.3
Spain	63.9	36.1
Turkey	58.1	41.9
United Kingdom	44.1	55.9
US	45.4	54.6
Argentina	47.4	52.6
Brazil	53.3	46.7
China	46.4	53.6
India	45.3	54.7
Indonesia	62.9	37.1
United Arab Emirates	60.2	39.8
Australia	42.4	57.6
Algeria	52.0	48.0
Egypt	26.7	73.3
Ghana	44.6	55.4
Nigeria	55.1	44.9
South Africa	56.5	43.5

Table 6.1. % of Google resources that match DNS and Anycast.

6.5.2. The Challenge of Mixed Approach

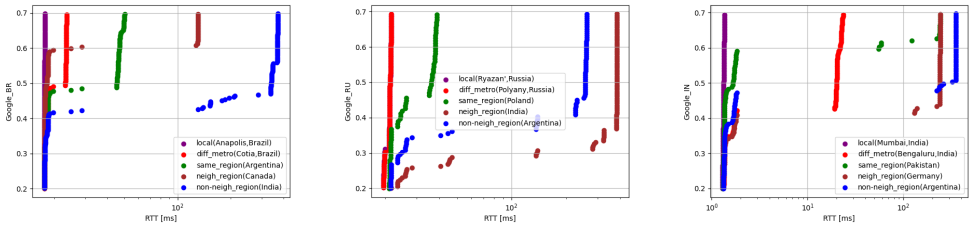
The Google CDN presents a particular challenge to our approach. Figure 6.12 shows the CDFs of latencies to Google replicas in Brazil, Russia and India, for different resolver scopes. These CDFs show two distinct patterns, where a portion of responses (e.g., $\approx$ 50% in Brazil) shows the pattern of Anycast-based replica selection while another segment shows a clear DNS-based pattern.

Correspondingly, the coefficient of regionalization between the resolver scopes show a mid-range of values, with $CRV[1]$ and $CRV[3]$ ranging between 0.52 and 0.85 for these countries (Fig. 6.10(b)).

The graphs and the associated coefficients of regionalization suggest that Google adopts a mixed approach, depending on the customer or resource. Separating resources based on the approaches used shows a clear distinction between Anycast (Fig. 6.13) and DNS-based replica selection (Fig. 6.14). Each of these sub-views clearly follow the expected patterns and the corresponding coefficients of regionalization are similarly consistent, with $CRV[1]$ and $CRV[3]$ of ≈ 0 for the Anycast-based case and $CRV[1]$ and $CRV[3]$ of 1 for DNS-based replica selection.

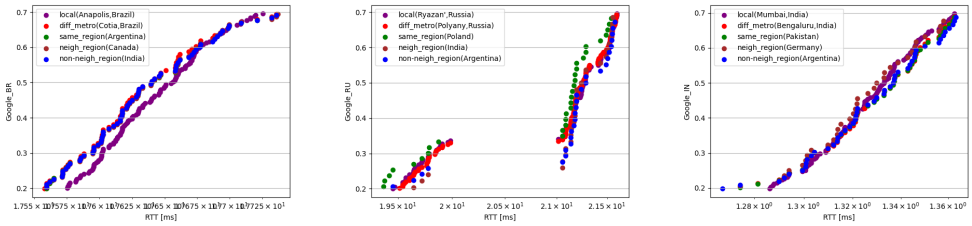
For a more detailed view of the redirection models used by Google, we calculated the percentage of Google resources that are assigned CDN replicas based on the Anycast or DNS-based models for every country and region in our study. Table 6.1 shows that there is an approximately equal split between Google content served using the DNS-based redirection model and the Anycast-based redirection model. Looking at resource domains, we find that the key services of the Google parent company, such as google.com, youtube.com, googleadservices.com and googlesyndication.com rely on DNS-based replica selection. Whereas, apparently external domains, such as snapchat.com and chess.com, are served using the anycast model.

Additionally, we were also curious if Google uses different redirection techniques for different applications. Since Google publicly announces the prefixes it uses for the cloud applications [**69**], we used them to examine the Google cloud replicas. We found that the CDN redirection model Google uses for its cloud applications is indeed exclusively anycast with $CRV[1]$ and $CRV[3]$ of ≈ 0 , whereas for the native applications it uses a mixed approach as shown in Fig. 6.12.



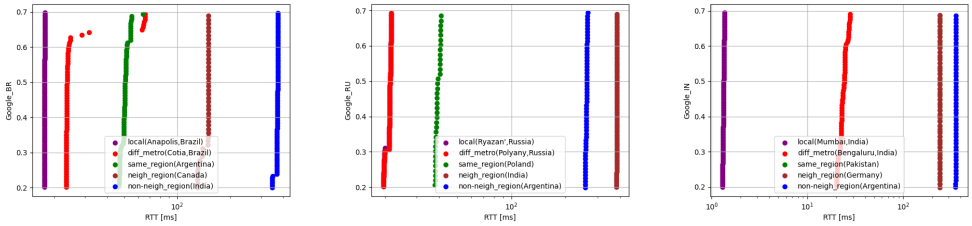
(a) Google-Brazil (b) Google-Russia (c) Google-India

Figure 6.12. Google uses a combination of redirection techniques that matches the DNS-based for some content and Anycast redirection model for other, as shown in Brazil, Russia, and India



(a) Google-Brazil (b) Google-Russia (c) Google-India

Figure 6.13. Google content using Anycast redirection, as shown in Brazil, Russia, and India



(a) Google-Brazil (b) Google-Russia (c) Google-India

Figure 6.14. Google content using DNS-based redirection, as shown in Brazil, Russia, and India

6.5.3. Geographic Trends

We close our analysis of main redirection approaches by looking at at their adoption by CDNs across countries and regions in our dataset. Table 6.2 lists the countries, the percentage of world Internet users they represent, and the percentage of resources in each country that match each redirection technique. The table also aggregates the resources from the countries in each region to show the percentage of resources delivered using each redirection technique, along with the percentage of Internet users potentially affected per region. For instance, for the US (highlighted) 36.8% of resources use DNS-based redirection model, 32.8% use a combination of DNS-based and Anycast, 23.8% use Anycast, and only 6.5% use Regional Anycast.

It shows that in the US and Australia, the majority of resources use DNS-based redirection model, while in much of Europe and China, the percentage of resources using DNS and Anycast redirection models are roughly similar. In the case of all African countries, as well as Russia, Turkey, India, Indonesia, and UAE, a significant percentage of resources rely on Anycast.

Resources, however, are not all of equal size. Table 6.3 shows the percentage of the bytes of content in each country that use each redirection technique. We calculate this by aggregating the total bytes of each resource in the set. Here, interestingly we observe that most content from the countries we measure worldwide (40.9%), and more specifically in North America(34.7%), South America(35.5%), Africa(56.7%) and Oceania(40.1%) uses DNS-based redirection model.

Location	Internet Users(%)	DNS-based(%)	Anycast(%)	Regional Anycast(%)	Combination(%)
Europe	60.0	29.5	39.2	2.0	29.2
France	1.1	33.7	38.8	4.4	23.2
Germany	1.5	34.5	38.1	1.0	26.3
Russia	2.3	10.6	58.0	0.5	30.9
Spain	0.8	41.8	30.1	1.6	26.4
Turkey	1.3	22.7	41.5	0.6	35.2
United Kingdom	1.2	33.0	33.4	4.3	29.4
North America	89.3	36.8	23.8	6.5	32.8
→US	*5.5*	*36.8*	*23.8*	*6.5*	*32.8*
South America	60.0	27.7	35.2	1.3	35.8
Argentina	0.8	24.1	35.1	0.8	40.0
Brazil	3.3	29.4	35.9	1.8	33.0
Asia	70.1	25.0	44.6	1.4	29.0
China	18.8	38.6	39.5	2.8	19.2
India	15.5	32.8	40.4	1.0	25.9
Indonesia	3.9	16.2	48.2	0.8	34.8
United Arab Emirates	0.2	21.1	43.3	1.6	34.0
Oceania	75.0	34.7	28.3	2.4	34.7
Australia	0.4	34.7	28.3	2.4	34.7
Africa	50.1	20.1	49.2	1.5	29.2
Algeria	0.7	21.2	53.7	0.75	24.3
Egypt	1.0	18.0	56.0	0.7	25.3
Ghana	0.3	22.4	43.1	1.9	32.6
Nigeria	2.9	20.8	42.8	3.0	33.4
South Africa	0.6	15.5	55.9	0.6	28.1
World Total	66.0	26.7	40.8	2.0	30.6

Table 6.2. Percentage of resources using a particular redirection approach. For instance, 36.8% resources in the US use DNS-based, while 32.8% use a combination of DNS-based and Anycast.

Differences in the percentage of content served using a given replica selection approach may be attributed to multiple factors, many of them interrelated. Every CDN relies, predominantly, on one replica selection approach and captures different market shares in different regions. Large geographic areas with rich infrastructure deployment would probably lean towards a different replica selection approach than regions with smaller infrastructure footprints and/or fewer users.

Location	Internet Users(%)	DNS-based(%)	Anycast(%)	Regional Anycast(%)	Combination(%)
Europe	60.0	27.6	41.4	2.6	28.4
France	1.1	27.5	38.5	3.9	30.2
Germany	1.5	30.8	42.2	1.6	25.4
Russia	2.3	9.9	63.1	0.6	26.4
Spain	0.8	44.5	27.4	3.1	25.0
Turkey	1.3	26.7	46.0	3.3	24.0
United Kingdom	1.2	28.0	33.8	3.0	35.3
North America	89.3	34.7	32.3	3.8	29.1
→ US	*5.5*	*34.7*	*32.3*	*3.8*	*29.1*
South America	60.0	35.5	33.1	1.4	30.0
Argentina	0.8	34.2	33.7	1.1	31.0
Brazil	3.3	35.7	33.1	1.8	29.5
Asia	70.1	26.3	43.8	1.8	28.1
China	18.8	48.1	30.0	1.3	20.6
India	15.5	28.1	40.1	0.8	31.1
Indonesia	3.9	13.6	53.5	2.7	30.2
United Arab Emirates	0.2	20.5	40.0	1.8	37.7
Oceania	75.0	40.1	27.1	1.8	30.6
Australia	0.4	40.1	27.1	1.8	30.6
Africa	50.1	56.7	26.1	0.9	16.4
Algeria	0.7	22.1	51.0	1.7	25.1
Egypt	1.0	16.9	55.4	0.3	27.4
Ghana	0.3	82.0	9.6	0.4	8.0
Nigeria	2.9	36.5	37.3	2.5	23.7
South Africa	0.6	14.3	52.9	0.9	31.9
World Total	66.0	40.9	33.6	1.7	23.9

Table 6.3. Percentage of total bytes of the resources in each country, region, and worldwide that use the corresponding redirection techniques.

6.5.4. ECS Support

ECS-based replica selection support is, in a sense, orthogonal to the above approaches in that a given CDN could adopt regional anycast as it primary replication approach while, at the same time, respond to clients' ECS requests. For every CDN in our dataset, we also report the number of domains hosted on the CDN that support ECS, i.e. they respond to an ECS-equipped query with a response scope greater than 0. We also check if those domains use ECS when mapping clients, i.e. give a different response based on the client subnet provided in the query.

Most CDNs in our study have over 80% of the domains that support ECS. Given the client subnets of two distant regions, all CDNs except for the 7 ones shown in Table 6.4 give different responses corresponding to the provided subnets, for at least a subset of domains. We confirm the non-random nature of this behavior by verifying that the geolocation of the response matches that of the client subnet provided.

CDN	DNS-based	Anycast	Regional Anycast	ECS
Akamai	✓			✓
Azion	✓			✗
BunnyCDN	✓			✓
CDN77	✓			✓
Cloudflare		✓		✗
Cloudfront	✓			✓
Edgio			✓	✓
Facebook	✓			✗
Fastly	✓	✓		✗
Google	✓	✓		✓
Level3	✓			✓
Medianova	✓			✓
NGENIX		✓		✓
StackPath		✓		✗
Taobao	✓			✗
Tencent	✓			✗
Yahoo			✓	✓

Table 6.4. The predominant model CDNs match based on our measurements[5]and if they respond to ECS.

6.5.5. CDNs and Their Predominant Redirection Model

Table 6.4 lists the set of CDNs found in our dataset and the predominant redirection model they match based on our experiments. We see that 12 out of the 17 CDNs in our dataset match the DNS-based redirection model. Cloudflare, Stackpath, Fastly, and

NGENIX predominantly match the Anycast redirection model. Edgio and Yahoo are the only two CDNs that match Regional Anycast. Google matches DNS-based for some customers and Anycast for others, whereas for Fastly we see a only few responses showing DNS-based replica selection. We note that some CDNs may use different redirection techniques in different regions. Since some CDNs did not have sufficient resources in some regions for a measurement analysis, as a result the redirection models we observe are for the regions we measured. Finally, we find that 10 out of the 17 CDNs in our dataset use ECS.

6.6. Discussion

Experimentally determining the redirection techniques used by CDNs is challenged by differences in the deployment densities of CDNs in each region. This can lead to some clients being assigned the same set of CDN replicas because of a lack of closer alternatives, rather than limitations in the assignment approach. For instance, we can see with the Cloudfront CDN in India (Fig. 6.8(b)), where the latencies to CDN servers using a local resolver and a resolver in a different metro are the same potentially due to the limited footprint of Amazon in India [**177**]. To reduce the effects of this on our inference, our analysis of the coefficient of regionalization, looks at the extremes. That is for the coefficient of regionalization value within the same region, we compare the local resolver and same region resolver ($CRV[1]$) and for the coefficient of regionalization value in a different region, we compare the local resolver and non-neighboring region resolver ($CRV[3]$).

[5]From personal communication we understand that Yahoo uses other approaches in other locales.

Different CDNs may also have different definitions of regional borders affecting our analysis of Regional Anycast, or rely on different approaches in different regions perhaps as a result of the size of the region or the CDN deployment. We note that in our labeling of neighboring/non-neighboring regions we follow the United Nations geoscheme. This, as any regional division will result in mismatches when considering the regional division of services such as CDNs, such as India being a neighboring region of the UK and Argentina being a non-neighboring region. However, examining $CRV[1]$ and $CRV[3]$ and conducting repeated measurements across different countries helps us address such mismatches and identify the primary redirection approach of a CDN. The CDN replicas we obtain from our measurements can also be affected by network conditions and the load on the servers, although repeated measurements should reduce the impact of transient conditions.

Our study conducts experiments using vantage points that cover 66% of the world and the content served by the top sites (reported by the Google CrUX dataset) and popular CDNs in those countries. For instance, we found insufficient resources hosted on the Microsoft and Netflix CDNs. As a result, the most common CDN redirection technique that we report may be biased by the content and the countries we measure as well as by the selected resolver vantage points. We also note that in this study, we measure the content served by only the landing pages of the Web [**14**]. Whereas CDNs using different replica selection techniques may be more popularly used to serve video content or for instance software updates, not readily accessible from such pages.

6.7. Conclusion

This study proposes the first methodology to experimentally identify the predominant replica selection approach used by a CDN and applies it to CDNs serving popular Web Content worldwide. Our findings show that 12 out of 17 CDNs in our set adopt DNS-based replica selection and 10 out of 17 CDNs in our dataset respond to ECS queries. We showed that the predominant approach varies significantly by region, and that while the majority of resources were served by anycast-based systems, the DNS-based approach is the preferred system for delivering the majority of *bytes*. As the central role of CDNs in the current and future Internet continue to grow, our findings provide a context for further research in this space and may help better understand the impact of architectural changes like centralization.

Bibliography

[1] Abley, J., and Lindqvist, K. Operation of anycast services. *www.rfc-editor.org* (2006).

[2] Ager, B., Mühlbauer, W., Smaragdakis, G., and Uhlig, S. Comparing dns resolvers in the wild. In *Proc. of IMC* (2010).

[3] Akamai. Akamai cdn deployment, 2022.

[4] Al-Dalky, R., Rabinovich, M., and Schomp, K. A look at the ecs behavior of dns resolvers. IMC '19.

[5] AlJahdali, H., Albatli, A., Garraghan, P., Townend, P., Lau, L., and Xu, J. Multi-tenancy in cloud computing. In *Proc. of IEEE SOSE* (2014).

[6] Allman, M. Comments on DNS robustness. In *Proc. of IMC* (2018).

[7] Alzoubi, H. A., Lee, S., Rabinovich, M., Spatscheck, O., and der Merwe, J. V. Anycast cdns revisited. In *Proc. of the WWW* (2008).

[8] Alzoubi, H. A., Rabinovich, M., and Spatscheck, O. The anatomy of ldns clusters: Findings and implications for web content delivery. Association for Computing Machinery.

[9] Amazon. Amazon cdn deployment, 2022.

[10] Amazon. Amazon aws, 2023.

[11] Amazon. Amazon mechanical turk, 2023.

[12] Anderson, S., Salamatian, L., Bischof, Z. S., Dainotti, A., and Barford, P. igdb: connecting the physical and logical layers of the internet. In *Proceedings of the 22nd ACM Internet Measurement Conference* (2022).

[13] APNIC. APNIC Service Region, 2024. https://www.apnic.net/about-apnic/corporate-documents/documents/corporate/apnic-service-region/.

[14] Aqeel, W., Chandrasekaran, B., Feldmann, A., and Maggs, B. M. On landing and internal web pages. In *Proc. of IMC* (2020).

[15] Ardi, C., and Calder, M. The prevalence of single sign-on on the web: Towards the next generation of web content measurement. In *Proc. of IMC* (2023).

[16] Arkko, J. Centralised architectures in internet infrastructure. *IETF Internet Draft* (2019).

[17] Arkko, J. The influence of Internet architecture on centralised versus distributed Internet services. *Journal of Cyber Policy* (2020).

[18] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R. H., Konwinski, A., Lee, G., Patterson, D. A., Rabkin, A., Stoica, I., et al. Above the clouds: A berkeley view of cloud computing. Tech. rep., 2009.

[19] ATLAS, R. Atlas console, 2023.

[20] Bank, T. W. The world by region. https://datatopics.worldbank.org/sdgatlas/archive/2017/the-world-by-region.html, 2017.

[21] Bank, T. W. Brief: Digital government for development, 2024.

[22] Bates, S., Bowers, J., Greenstein, S., Weinstock, J., Xu, Y., and Zittrain, J. Evidence of Decreasing Internet Entropy: The Lack of Redundancy in DNS Resolution by Major Websites and Services. *Journal of Quantitative Description: Digital Media 1* (2021).

[23] Biliris, A., Cranor, C., Douglis, F., Rabinovich, M., Sibal, S., OliverSpatscheck, and Sturm, W. *Computer Communications* (2002).

[24] Bortzmeyer, S. DNS censorship (DNS lies) as seen by RIPE Atlas. RIPE Labs, December 2015. https://labs.ripe.net/Members/stephane_bortzmeyer/dns-censorship-dns-lies-seen-by-atlas-probes.

[25] Bottger, T., Cuadrado, F., Antichi, G., Fernandes, E. L., Tyson, G., Castro, I., and Uhlig, S. An empirical study of the cost of DNS-over-HTTPS. In *Proc. of IMC* (2019).

[26] Bozkurt, I. N., Aguirre, A., Chandrasekaran, B., Godfrey, P. B., Laughlin, G., Maggs, B., and Singla, A. Why is the Internet so slow?! In *Proc. of PAM* (2017).

[27] Bustamante, F. E., Doyle, J., Willinger, W., Fayed, M., Alderson, D. L., Low, S., Savage, S., and Schulzrinne, H. Towards re-architecting today's internet for survivability: Nsf workshop report. *SIGCOMM Comput. Commun. Rev.* (2024).

[28] Butkiewicz, M., Madhyastha, H. V., and Sekar, V. Understanding website complexity: Measurements, metrics, and implications. In *Proc. of IMC* (2011).

[29] Butkiewicz, M., Madhyastha, H. V., and Sekar, V. Understanding website complexity: Measurements, metrics, and implications. In *Proc. of ACM SIGCOMM* (2011).

[30] Butkiewicz, M., Madhyastha, H. V., and Sekar, V. Understanding website complexity: Measurements, metrics, and implications. In *Proc. of IMC* (2011).

[31] Calder, M., Fan, X., and Zhu, L. A cloud provider's view of EDNS client-subnet adoption.

[32] Calder, M., Fan, X., and Zhu, L. A cloud provider's view of edns client-subnet adoption. In *Proc. of TMA Conference* (2019).

[33] Calder, M., Flavel, A., Katz-Bassett, E., Mahajan, R., and Padhye, J. Analyzing the performance of an anycast cdn. In *Proc. of IMC* (2015).

[34] Callejo, P., Cuevas, R., Vallina-Rodriguez, N., and Cuevas, A. Measuring the global recursive dns infrastructure: A view from the edge. *IEEE Access* (2019).

[35] Carisimo, E., Gamero-Garrido, A., Snoeren, A. C., and Dainotti, A. Identifying ases of state-owned internet operators. In *Proc. of IMC* (2021).

[36] cdnplanet. Cdn finder, 2023.

[37] Chander, A., and Le, U. P. Breaking the web: Data localization vs. the global internet. *SSRN Electronic Journal* (2014).

[38] Chen, F., Sitaraman, R. K., and Torres, M. End-user mapping: Next generation request routing for content delivery. In *Proc. of ACM SIGCOMM* (2015).

[39] Chung, T., Lok, J., Chandrasekaran, B., Choffnes, D., Levin, D., Maggs, B. M., Mislove, A., Rula, J., Sullivan, N., and Wilson, C. Is the web ready for ocsp must-staple? *Proceedings of the Internet Measurement Conference 2018* (2018).

[40] Cisagov. findcdn, 2023.

[41] Cloudflare. Cloudflare, 2022.

[42] Cloudflare, 2023.

[43] Cloudflare. Cdn · cloudflare reference architecture docs, 2023.

[44] Consulting, I. General data protection regulation (gdpr), 2013.

[45] Contavalli, C., van der Gaast, W., Lawrence, D., and Kumari, W. Client subnet in dns queries. RFC 7871, IETF, 2016.

[46] Coppolino, L., D'Antonio, S., Mazzeo, G., and Romano, L. Cloud security: Emerging threats and current solutions. *Computers & Electrical Engineering* (2017).

[47] Dagon, D., Provos, N., Lee, C., and Lee, W. *Corrupted DNS Resolution Paths: The Rise of a Malicious Resolution Authority.* 2008.

[48] Darwich, O., Rimlinger, H., Dreyfus, M., Gouel, M., and Vermeulen, K. Replication: Towards a publicly available internet scale ip geolocation dataset. In *Proc. of IMC* (2023).

[49] de Vries, W. B., van Rijswijk-Deij, R., de Boer, P.-T., and Pras, A. Passive observations of a large dns service: 2.5 years in the life of google. *IEEE Transactions on Network and Service Management* (2020).

[50] Demchak, C., and Dombrowski, P. Cyber westphalia: Asserting state prerogatives in cyberspace. *Georgetown Journal of International Affairs* (2013).

[51] Dilley, J., Maggs, B., Parikh, J., Prokop, H., Stiaraman, R., and Weihl, B. Globally distributed content delivery. *IEEE Internet Computing* (2002).

[52] DNS, P. Public dns server list, 2023.

[53] Doan, T. V., Fries, J., and Bajpai, V. Evaluating public dns services in the wake of increasing centralization of dns. In *2021 IFIP Networking Conference (IFIP Networking)* (2021).

[54] Doan, T. V., van Rijswijk-Deij, R., Hohlfeld, O., and Bajpai, V. An empirical view on consolidation of the web. *ACM Transactions on Internet Technology 22*, 3 (Aug 2022), 1–30.

[55] Dobson, C. Achieving equity in digital government services, 2023.

[56] Douzet, F., Pétiniaud, L., Salamatian, L., Limonier, K., Salamatian, K., and Alchus, T. Measuring the fragmentation of the internet: The case of the border gateway protocol (bgp) during the ukrainian crisis. In *2020 12th International Conference on Cyber Conflict (CyCon)* (2020).

[57] Du, B., Candela, M., Huffaker, B., Snoeren, A. C., and claffy, k. Ripe ipmap active geolocation: Mechanism and performance evaluation. *SIGCOMM Comput. Commun. Rev.* (2020).

[58] Dönni, D., Machado, G., Tsiaras, C., and Stiller, B. Schengen routing: A compliance analysis.

[59] Fan, X., Katz-Bassett, E., and Heidemann, J. Assessing affinity between users and cdn sites. In *Proc. of TMA Conference* (2015).

[60] for Federal Government, A. C. C. Cloud computing for federal government, 2023.

[61] for US Government, A. Azure for us government — microsoft azure, 2023.

[62] Fruhlinger, J. The opm hack explained: Bad security practices meet china's captain america, 2020.

[63] Fund, I. M. STATE-OWNED ENTERPRISES: THE OTHER GOVERNMENT. `https://www.imf.org/~/media/Files/Publications/fiscal-monitor/2020/April/English/ch3.ashx`, 2020.

[64] Gigis, P., Calder, M., Manassakis, L., Nomikos, G., Kotronis, V., Dimitropoulos, X., Katz-Bassett, E., and Smaragdakis, G. Seven years in the life of hypergiants' off-nets. In *Proc. of ACM SIGCOMM* (2021).

[65] Goel, U., Wittie, M. P., and Steiner, M. Faster web through client-assisted cdn server selection. In *Proc. of ICCN* (Oct 2015).

[66] Google. Chrome user experience report — chrome ux report —google developers, 2022.

[67] Google. understanding-google-cloud-network-edge-points, 2022.

[68] Google. About crux, 2023.

[69] Google. Obtain google ip address ranges, 2023.

[70] Grossman, R. L. The case for cloud computing. *IT professional* (2009).

[71] Hao, S., Zhang, Y., Wang, H., and Stavrou, A. End-users get maneuvered: Empirical analysis of redirection hijacking in content delivery networks. In *Proc. of USENIX Security* (2018).

[72] hotspotshield. Hotspotshield vpn, 2023.

[73] Hounsel, A., Borgolte, K., Schmitt, P., and Feamster, N. D-DNS: towards re-decentralizing the DNS.

[74] HOUSE, T. W. Fact sheet: Building digital experiences for the american people — omb, 2023.

[75] Huang, C., Maltz, D. A., Li, J., and Greenberg, A. Public dns system and global traffic management. In *2011 Proceedings IEEE INFOCOM* (2011).

[76] Huang, C., Wang, A., Li, J., and Ross, K. Measuring and evaluating large-scale cdns. In *Proc. of IMC* (2008).

[77] Huitema, C., Huston, G., Hs, E., Leer, G., and Zhang. *Draft Report of DINRG Workshop on Centralization in the Internet*. 2021.

[78] Huston, G. DNS resolver centrality, 2019.

[79] Huston, G. DNS resolver centrality. APNIC Blog, 2019.

[80] Huston, G. CDN and centrality. APNIC Blog, 2021.

[81] Ibrahim, S., He, B., and Jin, H. Towards pay-as-you-consume cloud computing. In *Proc. of Conference on Services Computing* (2011).

[82] India's MINISTRY OF LAW AND JUSTICE. India's Digital Personal Data Protection (DPDP) Act, 2023. `https://www.meity.gov.in/writereaddata/files/Digital%20Personal%20Data%20Protection%20Act%202023.pdf`.

[83] Info, I. Ipinfo, 2023.

[84] Iordanou, C., Smaragdakis, G., Poese, I., and Laoutaris, N. Tracing Cross Border Web Tracking. In *Proc. of IMC* (2018).

[85] IP2Location. Free IP Geolocation Database, 2023.

[86] ipInfo. ipinfo, 2023.

[87] Jansen, W., Grance, T., et al. Guidelines on security and privacy in public cloud computing.

[88] Johnson, K. L., Carr, J. F., Day, M. S., and Kaashoek, M. F. The measured performance of content distribution networks. *Computer Communications 24*, 2 (2001).

[89] Kangasharju, J., Ross, K., and Roberts, J. Performance evaluation of redirection schemes in content distribution networks. *Computer Communications* (2001).

[90] Kashaf, A., Dou, J., Belova, M., Apostolaki, M., Agarwal, Y., and Sekar, V. A first look at third-party service dependencies of web services in africa.

[91] Kashaf, A., Sekar, V., and Agarwal, Y. Analyzing third party service dependencies in modern web services: Have we learned from the mirai-dyn incident? In *Proc. of IMC* (2020).

[92] Khan, M. T., DeBlasio, J., Voelker, G. M., Snoeren, A. C., Kanich, C., and Vallina-Rodriguez, N. An empirical analysis of the commercial vpn ecosystem. IMC '18.

[93] Koch, T., Katz-Bassett, E., Heidemann, J., Calder, M., Ardi, C., and Li, K. Anycast in context: A tale of two systems. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference* (2021), pp. 398–417.

[94] Krishnamurthy, B., Wills, C., and Zhang, Y. On the use and performance of content distribution networks. In *Proc. ACM IMW* (2001).

[95] Kumar, D., Ma, Z., Durumeric, Z., Mirian, A., Mason, J., Halderman, J. A., and Bailey, M. Security challenges in an increasingly tangled web. In *Proc. of the WWW* (2017).

[96] Kumar, R., Asif, S., Lee, E., and Bustamante, F. E. Each at its own pace: Third-party dependency and centralization around the world. *Proc. ACM Meas. Anal. Comput. Syst.* (2023).

[97] Kumar, R., Asif, S., Lee, E., and Bustamante, F. E. Each at its own pace: Third-party dependency and centralization around the world. In *Proc. of ACM SIGMETRICS* (2023).

[98] Kührer, M., Hupperich, T., Bushart, J., Rossow, C., and Holz, T. Going wild: Large-scale classification of open dns resolvers. *Proceedings of the 2015 Internet Measurement Conference* (2015).

[99] Law, B. G. D. P. Lgpd brazil - general personal data protection act, 2023.

[100] Le, T. V. Vietnam - data protection overview, Nov 2019.

[101] Li, Z., Levin, D., Spring, N., and Bhattacharjee, B. Internet anycast: Performance, problems, & potential. In *Proc. of ACM SIGCOMM* (2018).

[102] List, M. P. S. Public suffix list.

[103] Liu, Y., Song, H. H., Bermudez, I., Mislove, A., Baldi, M., and Tongaonkar, A. Identifying personal information in Internet traffic. In *Proc. of ACM COSN* (2015).

[104] Livingood, J., Antonakakis, M., Sleigh, B., and Winfield, A. Centralized dns over https (doh) implementation issues and risks, 2019.

[105] Luckie, M., Huffaker, B., Marder, A., Bischof, Z., Fletcher, M., and claffy, k. Learning to extract geographic information from internet router hostnames. In *Proc. of CoNEXT* (2021).

[106] MacAskill, E., Dance, G., Cage, F., Chen, G., and Popovich, N. Nsa files decoded: Edward snowden's surveillance revelations explained, 2013.

[107] Mao, Z., Cranor, C., Douglis, F., Rabinovich, M., Spatscheck, O., and Wang, J. *A Precise and Efficient Evaluation of the Proximity between Web Clients and their Local DNS Servers*. 2002.

[108] Mao, Z. M., Cranor, C. D., Douglis, F., Rabinovich, M., Spatscheck, O., and Wang, J. A precise and efficient evaluation of the proximity between web clients and their local dns servers. In *Proc. of USENIX ATC* (2002).

[109] Matic, S., Tyson, G., and Stringhini, G. Pythia: a framework for the automated analysis of web hosting environments. *The World Wide Web Conference* (2019).

[110] Matic, S., Tyson, G., and Stringhini, G. Pythia: A framework for the automated analysis of web hosting environments. In *The World Wide Web Conference* (2019).

[111] Maxmind. Maxmind server ip addresses, 2022.

[112] Mcquistin, S., Uppu, S., and Flores, M. Taming anycast in the wild internet. In *Proc. of IMC* (2019).

[113] Michelinakis, F., Doroud, H., Razaghpanah, A., Lutu, A., Vallina-Rodriguez, N., Gill, P., and Widmer, J. The cloud that runs the mobile internet: A measurement study of mobile cloud services.

[114] Mockapetris, P. Domain names – concepts and facilities. Tech. rep., IETF, 1987.

[115] Moura, G. How centralized is dns traffic becoming?, November 2020.

[116] Moura, G., Castro, S., Hardaker, W., Wullink, M., and Hesselman, C. Clouding up the internet: how centralized is dns traffic becoming? In *Proc. of IMC* (2020).

[117] Nations, U. Egovkb — united nations ¿ about ¿ overview ¿ e-government development index, 2023.

[118] Nawrocki, M., Koch, M., Schmidt, T. C., and Wählisch, M. Transparent forwarders.

[119] News, C. Explainer: What's behind strained china-japan relations, 2022.

[120] Niaki, A. A., Marczak, W. R., Farhoodi, S., McGregor, A., Gill, P., and Weaver, N. Cache me outside: A new look at dns cache probing.

[121] nordvpn. Nord vpn, 2023.

[122] Nygren, E., Sitaraman, R., and Sun, J. The akamai network: A platform for high-performance internet applications. *ACM SIGOPS Operating Systems Review 44*, 3 (July 2010).

[123] Nygren, E., Sitaraman, R. K., and Sun, J. The akamai network: a platform for high-performance internet applications. *ACM SIGOPS Operating Systems Review* (2010).

[124] of California Department of Justice, S. California consumer privacy act (ccpa), 2023.

[125] Office, U. G. A. Solarwinds cyberattack demands significant federal and private-sector response (infographic), 2021.

[126] OONI. MAP State of Internet Censorship Report 2022 - Vietnam. https://ooni.org/post/2022-state-of-internet-censorship-vietnam/, 2023.

[127] Otto, J., and Bustamante, M. S. J. R. F. E. Content delivery and the natural evolution of DNS: remote DNS trends, performance issues and alternative solutions. In *Proc. of IMC* (2012).

[128] Overview, R. D. P. Russia - data protection overview, 2020.

[129] Park, J., Khormali, A., Mohaisen, M., and Mohaisen, A. Where are you taking me? behavioral analysis of open dns resolvers. In *2019 49th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)* (2019).

[130] Planet, C. Akamai cdn, 2023.

[131] Planet, C. Cdn planet, 2023.

[132] Planet, C. Cloudflare cdn, 2023.

[133] Planet, C. Edgio cdn, 2023.

[134] Pochat, V. L., Goethem, T. V., Tajalizadehkhoob, S., Korczynski, M., and Joosen, W. Tranco: A research-oriented top sites ranking hardened against manipulation. In *Proc. of NDSS* (2019).

[135] Poese, I., Frank, B., Ager, B., Smaragdakis, G., Uhlig, S., and Feldmann, A. Improving content delivery with padis. *IEEE Internet Computing* (2012).

[136] Poese, I., Uhlig, S., Kaafar, M. A., Donnet, B., and Gueye, B. Ip geolocation databases: Unreliable? *ACM SIGCOMM CCR* (2011).

[137] Pohlmann, N., Sparenberg, M., Siromaschenko, I., and Kilden, K. *Secure Communication and Digital Sovereignty in Europe*. 2014.

[138] Programme, U. N. D. Human development index, 2023.

[139] Radu, R., and Hausding, M. Consolidation in the DNS resolver market – how much, how fast, how dangerous? *Journal of Cyber Policy* (2020).

[140] Rhoades, S. A. The herfindahl-hirschman index, 1993.

[141] RIPE NCC. IPmap, 2024. https://ipmap.ripe.net/.

[142] Roberts, P. F. Comcast suffers DNS outage, denies pharming link, Apr 2005. https://www.networkworld.com/article/2318771/comcast-suffers-dns-outage--denies-pharming-link.html.

[143] routeviews. RouteViews IPv4 Prefix to AS mappings - coalesced. https://catalog.caida.org/dataset/routeviews_ipv4_prefix2as_coalesced., 2023.

[144] Rula, J. P., and Bustamante, F. E. Behind the curtain – cellular dns and content replica selection. In *Proc. of IMC* (2014).

[145] Saroiu, S., Gummadi, K. P., Dunn, R. J., Gribble, S. D., and Levy, H. M. An analysis of internet content delivery systems. In *Proc. of USENIX OSDI* (2002).

[146] Scheitle, Q., Hohlfeld, O., Gamba, J., Jelten, J., Zimmermann, T., Strowes, S. D., and Vallina-Rodriguez, N. A long way to the top. *Proc. of IMC* (2018).

[147] Schneider-Petsinger, M., Wang, J., Jie, Y., and Crabtree, J. *US-China Strategic Competition The Quest for Global Technological Leadership*. 2019.

[148] Schomp, K., Callahan, T., Rabinovich, M., and Allman, M. On measuring the client-side dns infrastructure. In *Proceedings of the 2013 conference on Internet measurement conference* (2013), pp. 77–90.

[149] Selenium. Seleniumhq browser automation, 2024.

[150] Shad, D. M. R. Cyber threat in interstate relations: Case of us-russia cyber tensions. *Policy Perspectives* (2018).

[151] Shaikh, A., Tewari, R., and Agrawal, M. On the effectiveness of dns-based server selection. In *Proc. of IEEE INFOCOM* (2001).

[152] Singanamalla, S., Jang, E. H. B., Anderson, R., Kohno, T., and Heimerl, K. Accept the risk and continue: Measuring the long tail of government https adoption. In *Proc. of IMC* (2020).

[153] Singh, R., Dunna, A., and Gill, P. Characterizing the deployment and performance of multi-cdns. In *Proc. of IMC* (2018).

[154] Singla, A., Chandrasekaran, B., Godfrey, P. B., and Maggs, B. The internet at the speed of light. In *Proc. of HotNets* (2014).

[155] Society, I. Consolidation in the internet economy, 2020.

[156] Sommese, R., Bertholdo, L., Akiwate, G., Jonker, M., van Rijswijk-Deij, R., Dainotti, A., Claffy, K., and Sperotto, A. Manycast2: Using anycast to measure anycast. In *Proc. of IMC* (2020).

[157] Sommese, R., Bertholdo, L., Akiwate, G., Jonker, M., van Rijswijk-Deij, R., Dainotti, A., Claffy, K., and Sperotto, A. Manycast2: Using anycast to measure anycast. In *Proceedings of the ACM Internet Measurement Conference* (2020).

[158] Stadnik, I. Control by infrastructure: Political ambitions meet technical implementations in runet. *First Monday* (2021).

[159] Stapling, O. The problem with ocsp stapling and must staple and why certificate revocation is still broken - hanno's blog, 2017.

[160] Stats, I. W. World internet users statistics and 2019 world population stats, 2022.

[161] Stats, I. W. World internet users statistics and 2019 world population stats, 2023.

[162] Stokes, B. Hostile neighbors: China vs. japan, 2016.

[163] Su, A.-J., Choffnes, D. R., Kuzmanovic, A., and Bustamante, F. E. Drafting behind akamai (travelocity-based detouring). In *Proc. of ACM SIGCOMM* (2006).

[164] surfshark. Surfshark vpn, 2023.

[165] Sørensen, J., and Kosta, S. Before and after gdpr: The changes in third party presence at public and private european websites. *The World Wide Web Conference on - WWW '19* (2019).

[166] The European Commission. Overseas Countries and Territories, 2024. https://international-partnerships.ec.europa.eu/countries/overseas-countries-and-territories_en.

[167] The French Republic. Accord sur la Nouvelle-Caledonie signe a Noumea le 5 mai 1998, 2024. https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000000555817.

[168] Timlib. Webxray domain owner list.

[169] Triukose, S., Wen, Z., and Rabinovich, M. Measuring a commercial content delivery network. In *Proc. of the WWW* (2011).

[170] United Nations. UN E-Government Survey 2022 – The Future of Digital Government.

[171] United Nations. Non-Self-Governing Territories, 2024. https://www.un.org/dppa/decolonization/en/nsgt.

[172] Urban, T., Degeling, M., Holz, T., and Pohlmann, N. Beyond the front page: Measuring third party dynamics in the field.

[173] US Department of State. A Guide to the United States' History of Recognition, Diplomatic, and Consular Relations, by Country, since 1776: Morocco, 2024. https://history.state.gov/countries/morocco.

[174] Wang, S., MacMillan, K., Schaffner, B., Feamster, N., and Chetty, M. A first look at the consolidation of dns and web hosting providers.

[175] Wang, Z., Huang, J., and Rose, S. Evolution and challenges of dns-based cdns. *Digital Communications and Networks 4*, 4 (Nov 2018), 235–243.

[176] Wikipedia. United nations geoscheme, Jul 2022.

[177] Wikipedia. Amazon cloudfront, May 2023.

[178] world population. internet-users-by-country, 2022.

[179] Xue, J., Choffnes, D., and Wang, J. Cdns meet cn an empirical study of cdn deployments in china. *IEEE Access 5* (2017), 5292–5305.

[180] Xue, J., Choffnes, D., and Wang, J. Cdns meet cn an empirical study of cdn deployments in china. *IEEE Access 5* (2017), 5292–5305.

[181] Yeganeh, B., Durairajan, R., Rejaie, R., and Willinger, W. A first comparative characterization of multi-cloud connectivity in today's internet. In *Proc. of PAM* (2020).

[182] Zembruzki, L., Jacobs, A. S., Landtreter, G. S., Granville, L. Z., and Moura, G. `dnstracker`: Measuring centralization of dns infrastructure in the wild. In *Proc. of AINA* (2020).

[183] Zhou, M., Zhang, X., Hao, S., Yang, X., Zheng, J., Chen, G., and Dou, W. Regional IP Anycast: deployments, performance, and potentials. In *Proc. of ACM SIGCOMM* (2023).

[184] Zhu, G., and Gu, W. User mapping strategy in multi-cdn streaming: A data-driven approach. *IEEE Internet of Things Journal* (2021), 1–1.

[185] Zhu, J., Vermeulen, K., Cunha, I., Katz-Bassett, E., and Calder, M. The best of both worlds. In *Proc. of IMC* (2022).