Internet Service Usage and Delivery As Seen From a Residential Network

SHUYUE YU, Columbia University, United States
THOMAS KOCH, Columbia University, United States
ILGAR MAMMADOV, Columbia University, United States
HANGPU CAO, Columbia University, United States
GIL ZUSSMAN, Columbia University, United States
ETHAN KATZ-BASSETT, Columbia University, United States

Given the increasing residential Internet use, a thorough understanding of what services are used and how they are delivered to residential networks is crucial. However, access to residential traces is limited due to their proprietary nature. Most prior work used campus datasets from academic buildings and undergraduate dorms, and the few studies with residential traces are often outdated or unavailable to other researchers. We provide access to a new residential dataset—we have been collecting traffic from ~1000 off-campus residences that house faculty, postdocs, graduate students, and their families. Although our residents are university affiliates, our dataset captures their activity at home, and we show that this dataset offers a distinct perspective from the campus and dorm traffic. We investigate the serving infrastructures and services accessed by the residences, revealing several interesting findings: peer-to-peer activity is notable, comprising 47% of the total flow duration; third-party CDNs host many services but serve much less traffic (e.g., Cloudflare hosts 19% of domains but only 2% of traffic); and 11 of the top 100 services that have nearby servers often serve users from at least 1,000km farther away. This broad analysis, as well as our data sharing, pushes toward a more thorough understanding of Internet service usage and delivery, motivating and supporting future research.

CCS Concepts: • **Networks** → **Network monitoring**; **Network measurement**.

Additional Key Words and Phrases: Residential traffic, Residential ISP, Network service, Measurements

ACM Reference Format:
Shuyue Yu, Thomas Koch, Ilgar Mammadov, Hangpu Cao, Gil Zussman, and Ethan Katz-Bassett. 2025. Internet Service Usage and Delivery As Seen From a Residential Network. *Proc. ACM Meas. Anal. Comput. Syst.* 9, 2, Article 41 (June 2025), 30 pages. https://doi.org/10.1145/3727133

1 INTRODUCTION

Residential Internet usage has rapidly increased and morphed in some regions [89], driven by the penetration of broadband home Internet, adoption of Internet-based entertainment, the plethora of Internet-connected devices in many homes, and more people working hybrid jobs in the aftermath of COVID [87]. Given the importance of residential Internet, it is crucial to thoroughly understand which services users access and how they are delivered, as well as to enable future research.

Authors' addresses: Shuyue Yu, syyu@cs.columbia.edu, Columbia University, United States; Thomas Koch, tak2154@columbia.edu, Columbia University, United States; Ilgar Mammadov, im2703@columbia.edu, Columbia University, United States; Hangpu Cao, hc3346@columbia.edu, Columbia University, United States; Gil Zussman, gil@ee.columbia.edu, Columbia University, United States; Ethan Katz-Bassett, ethan@ee.columbia.edu, Columbia University, United States.

2476-1249/2025/6-ART41 $15.00
https://doi.org/10.1145/3727133

In general, the community lacks access to *open residential traces*. Prior studies issued active measurements to map serving infrastructures of large providers [34, 42, 46, 57], but they lacked visibility into how traffic was delivered from the mapped infrastructures. Other studies collected user traffic via crowd-sourcing [28, 61, 91, 95, 96, 99], but what could be measured was limited since they required explicit action from participants who deployed the kits. These datasets usually included few users (~100 households) and were hard to scale and maintain. Most datasets available to researchers were from campus networks and captured behavior in undergraduate dorms, classrooms, and offices [33, 45, 54, 58, 63, 64, 71, 82, 86, 101, 108]. As we show in §3.1.1, even with dorms, on-campus traffic still differs from off-campus residential usage.

Studies using industry traces illustrated the serving infrastructures from their perspectives [44, 93, 104, 107], but they may not apply to services beyond those offered by those providers. Some reports showed global service usage patterns [27, 78, 79, 90], but they did not offer peer-reviewed methodologies and usually lack details. A few academic papers used traces from residential networks [28, 53, 54, 80, 84, 100], but they often did not provide a broad understanding of the Internet use and are now outdated. Further, the datasets were not available to other researchers, limiting the set of questions answered.

One of our key contributions lies in our continuing collection of traffic traces from approximately 1000 off-campus residences of Columbia University, a dataset we will share with other researchers for studies that have undergone IRB review and follow our Acceptable Use Policy. These off-campus residences *are not undergraduate dorms, classrooms, or offices*, but rather house *graduate students, faculty, and their families*. Even though our residents are university-affiliated, our dataset is residential, capturing Internet usage at homes. While our network differs from many regular residential networks operated by large commercial ISPs, our dataset remains a valuable new source.

We have collected anonymized *packet traces* from 400+ residential apartments four hours per day for over three years and ~1000 residences since the beginning of 2024, with plans to scale up further. To the best of our knowledge, this is the residential dataset with the most residences and the highest level of detail ever shared. In §3, we discuss the details of our data collection, sharing, and the limitations and representativeness of our dataset.

Another key contribution is to present a *detailed, broad view of Internet service usage and delivery* for the off-campus residential network. Although some observations mainly uphold the conventional wisdom and are not surprising, we see value in verifying and quantifying the community's understanding, and some results suggest subtle but important differences from our expectations.

- In §4, we explain our methodology for associating flows with services and for classifying remote hosts. While our methodology builds on prior work [28, 37, 50], our novelty lies in our use of temporal correlation via clustering.
- In §5, we investigate the serving infrastructure accessed by our users and discover that (1) peer-to-peer activity accounts for 5% of the total traffic volume and 47% of the total flow duration, due in large part to both file sharing and videotelephony; (2) third-party CDNs host many services but amount to less traffic than we expected (*e.g.,* Cloudflare hosts 19% of domains but only 2% of traffic); (3) some popular services with nearby servers often serve users from much more distant servers—11 of the top 100 services use servers whose average distance (weighted by user traffic) is 1,000 km farther than their closest ones; and (4) the majority of traffic is steered by DNS over UDP to unicast IP addresses, but many connections (14%) use expired DNS records, which may affect a CDN's ability to update paths in the face of failures, overload, or performance changes. Moreover, we observe changes in the serving infrastructure over time—the decommissioning of Akamai cache servers hosted in

Columbia's provider and the reduced use of Lumen (aka Level 3) CDN—highlighting the value of our continuing collection.

- In §6, we study the usage of service and service-types. Our results confirm the intuitive understandings that (1) different metrics of service activity can lead to different service popularity (e.g., iCloud accounts for half the traffic volume of YouTube but three times the flow duration), suggesting that the widely used DNS-based top lists [68, 105] may not fully reflect popularity with respect to traffic or duration; and (2) even small variations in demographics lead to service usage differences as large as the differences in usage across continents. To the best of our knowledge, no other prior work has thoroughly quantified these differences using open methodologies. Moreover, we observe changes in certain services or service-types over both short and long timescales. We also provide a case study on the performance of two very popular Netflix live streaming events, demonstrating that we can infer key performance differences from encrypted traffic and gain insights into service performance. These results highlight the value of our dataset in assessing service popularity and monitoring events, enabling researchers to evaluate their designs in a more realistic and up-to-date context.

Lastly, we summarize the key insights from our findings and the potential uses of our dataset in §7. We will continue our collection, and the latest information about our dataset can be found in: https://wimnet.github.io/CUResidential/ [24]. With our broad analysis and sharing of our residential traces, we hope to enhance the understanding of Internet service delivery and support future research.

2 MOTIVATION AND RELATED WORK

2.1 What We Study: Goals and Non-Goals

Our goal is to provide a thorough understanding of modern Internet use in a residential network:
(1) How is traffic served from various serving infrastructures (content delivery networks, clouds) to our users? What types of remote hosts exchange traffic with our users? Which service providers serve our users, and do they use 3rd-party infrastructures? How far away does traffic come from? How are our users steered to servers via DNS? By answering these questions, we hope to help researchers better understand edge networks and their relationship with service providers.
(2) What services are accessed by our users? An updated understanding of service usage is important in the post-COVID era, as more people work remote and hybrid jobs. While industry reports exist [27, 90], we aim to provide a complementary view with open, detailed methodology and data. This view can highlight services that need further study, and researchers can also request our dataset to investigate emerging applications.
(3) How does the usage of serving infrastructures and services change over time? This question sheds light on temporal trends, and our dataset also enables tracking of future changes.

We also clarify what our goals do not include:
We intentionally choose to not focus on a specific protocol or service. Instead, we present a broad overview of one residential network, enabling us to tie together insights from prior studies that used various traces and examined more focused problems. We also use this broad understanding to introduce our dataset and its potential uses, encouraging researchers to use our dataset for in-depth investigations.
We do not claim that our dataset represents all residential networks. While our traces include the home Internet use of children and adults with no university affiliation besides familial, we acknowledge that our dataset differs from those of other regions or other user types. However, given the limited access to residential data, our study is still useful to the community.

2.2 Comparison with Existing Studies

Given the diversity of networks, services, and users, it is nearly impossible to have a complete understanding of Internet service usage and delivery. Building such a thorough view would require data from most service providers, which is usually proprietary and not shared with researchers. Despite this challenge, researchers have greatly advanced our understanding of Internet usage through available datasets. Each dataset has a unique but limited view, as does our dataset. Together, they tell a more complete story than any individually. This section compares our work with existing studies to show how our analysis and dataset contribute to the community's understanding.

Four main approaches exist among the prior work:

(1) Researchers issue active measurements to understand service delivery. Some studies map serving infrastructures and interconnections of large providers [34, 42, 46, 57], and some studies analyze the reliance of services on 3rd-party CDNs [26, 72]. However, with active measurements, it is difficult to interpret how much the serving infrastructures are used. For example, these studies can identify the existence of a CDN server in a user network, but they cannot measure how often users are served from it. While our study also investigates serving infrastructure, it relies on passive traces, enabling us to capture the volume of traffic traversing serving infrastructures.

(2) Researchers also collect data from residential users via crowd-sourcing [61, 95, 96, 99], which asks volunteers to deploy a measurement script, device, or application to record traffic statistics. The small scale (normally less than or around 100 households) of those experiments suggests how these approaches can be difficult to scale. It is also hard to maintain users and funding over time to enable long-term collection. In contrast, our dataset includes ~1,000 households, and our data collection is continuing.

(3) Most commonly, academic researchers collect data from campus networks [33, 45, 58, 63, 64, 71, 82, 86, 101, 108]. The campus networks primarily represent the Internet usage of undergraduate dorms, classrooms, and offices. A large portion of the traffic is not residential, and the dorm traffic only reflects the devices and activities of undergraduate students. These studies (except one [101]) did not separate dorm traffic from academic traffic (and potentially could not do so for configurations similar to our on-campus WiFi network, where IP addresses are assigned randomly). Our dataset exclusively captures traffic from off-campus residents (including faculty and family). As shown in §3.1.1, our dataset differs greatly from campus datasets. Moreover, these studies often did not share their datasets and focused on specific services (e.g., TCP or video conferencing apps), whereas our study provides a detailed breakdown of the top services accessed by users.

(4) There is limited visibility into residential networks. Industry leaders occasionally publish studies about their serving infrastructure [41, 44, 93, 104, 107]. Our study is not limited to hypergiants and thus complements these studies.

Researchers and companies leverage data collected by monitoring software and platforms from ISPs [39, 53, 54, 78–80, 84, 90, 100]. One set of researchers used to collect traffic traces from a grant-funded residential network [28, 91]. However, such datasets are often one-off, not available to most researchers, and could be made unavailable at any time. As the residential buildings belong to our university and our collection is not tied to a particular grant, we plan to maintain our ongoing collection campaign.

Table 1 compares our study with existing work using residential traces (but not work using other approaches, such as the active measurements, crowdsourcing, academic traces, or CDN data discussed above). Many prior studies investigate a relevant but different research question: they focus on a particular protocol, CDN, traffic demands, etc. [28, 39, 53, 54, 84, 91]. While Labovitz et al. (2010) [80] and Trevisan et al. (2018) [100] examine both serving infrastructure and service usage, our analysis differs from theirs. For example, Labovitz et al. did not study latency to servers,

Paper (Year)	Key Similarities	Key Difference(s)	Open Data?
Our studies	Study serving infrastructure and service usage over time		Yes
Sargent (2014) [91]	Study top services	Focuses on TCP performance; Small scale (100 residences)	No (collection has ended)
Allman (2020) [28]	Study DNS protocol	Only focuses on DNS; Small scale (100 residences)	
Labovitz et al. (2010) [80]	Study serving infrastructure and service usage	Outdated; Only has traffic flow samples but not DNS domains; Does not study 3rd-party dependency, latency to servers, DNS steering, or service usage across demographics	No
Labovitz talks (2019, 2020) [78, 79]	Similar to Labovitz et al. (2010)	Opaque methodology; Limited details (short talks)	No
Sandvine Report (2023) [90]	Study service popularity	Opaque methodology; Does not study serving infrastructure	No
Feldmann et al. (2020) [54]	Study the traffic shifts over time	One-off; Focuses on the effect of the pandemic lockdown on traffic; Exclude traffic from certain popular providers; Does not study 3rd-party dependency, latency to servers, DNS steering, or service usage across demographics	No
Blendin et al. (2018) [39]	Study CDN	Only focuses on Apple CDN	No
Trevisan et al. (2018) [100]	Study serving infrastructure and service usage	Outdated; Focuses on the changes in the serving infrastructure and services (*e.g.*, does not study 3rd-party dependency, DNS steering, or service usage across demographics), while we provide a fine-grained view of how services are delivered	No
Feamster et al. (2016) [53]	Study utilization of interconnections	Does not study at the level of services	Only aggregated statistics of interconnections
Liu et al. (2021) [84]	Study traffic demands and latency over time		

Table 1. Comparison with existing residential Internet studies

and Trevisan et al. did not investigate dependency on 1st-/3rd-party CDNs. Additionally, these two studies were one-off and conducted pre-pandemic, whereas the Internet is ever-changing, making it important to continually refresh our knowledge. The Sandvine report [90] and the Labovitz talks [78, 79] are more recent, but they use opaque methodology and lack details. In summary, our key contribution is to provide an *updated, detailed understanding of serving infrastructure and service usage in a residential network*, as well as *making new data available to researchers upon request* on an ongoing basis.

3 DATA COLLECTION AND SHARING FOR A NEW OPEN RESIDENTIAL DATASET

In this section, we provide details of our dataset, data collection method, dataset limitations and representativeness, and data sharing plan.

3.1 A New Open Residential Dataset

To address the limited visibility into residential networks in academia, we leverage an unusual aspect of Columbia University: most faculty and their families live in university-owned off-campus apartments, as do many postdocs and graduate students. Unlike campus networks that primarily provide WiFi connectivity to classrooms and dorms, these off-campus apartments provide Ethernet access to individual bedrooms and units, resembling the setting of typical residential buildings.

We have been collecting residential traffic of university affiliates from 943 residences in 28 off-campus apartment buildings. We will also make the data available upon request for studies that have undergone IRB review (see more details in §3.4). The buildings accommodate different resident types: (1) graduate students (mainly PhD students and, in some cases, their families), (2) postdocs, and (3) faculty (and a small number of staff) and their families. Among the 28 buildings, 11 only house graduate students, 11 only house faculty and their families, and 6 house different types of residents. We refer to this network as *our network*, residents living in those buildings as *our users*, and addresses outside Columbia's network as *remote hosts*. To reduce university-related effects, we exclude traffic to or from destinations owned by Columbia (<1% of traffic).

3.1.1 ***Our network is significantly different from campus networks.*** Since our users are university-affiliated, one may wonder if our dataset is similar to campus datasets. For the comparison, we collect 7 days (March 21-27, 2024) of NetFlow (v5) at the campus border routers. We divide traffic into three categories: (1) *on-campus academic buildings and undergrad dorms*, (2) *all off-campus residential apartments* (144 buildings, 6687 units), and (3) *off-campus residential apartments within our dataset* (28 buildings, 943 units). Note that (3) is a subset of (2), as our IT department only provides continuous access to traffic from the subset of apartments. We cannot differentiate between academic buildings and undergrad dorms because both use the on-campus WiFi network, which assigns and rotates IP addresses randomly. Other campus networks may experience the same issue, making it hard to isolate the traffic of dorms from available campus datasets [33, 45, 58, 63, 64, 71, 82, 86, 108]. We use RouteViews [5] and ASRank [2] to identify organizations that host remote hosts. For the top 50 organizations which contribute 90% of traffic for all off-campus residential data, we compute the percent of traffic from/to each organization for each category.

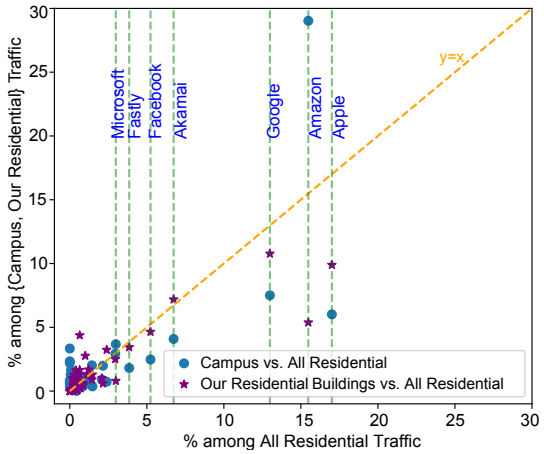


Fig. 1. Columbia University's on-campus network differs from its off-campus residential network, as the R-squared value for fitting the blue dots to the line y=x is 0.4. A subset of the residential network that we have access to is similar to the entire residential network, as the R-squared value for fitting the purple stars to line y=x is 0.7.

Off-campus residential traffic significantly differs from on-campus traffic: We plot a blue point in Fig. 1 for each organization, with the X-coordinate being the percent in *all residences (category 2)* and the Y-coordinate being the percent in *campus data (category 1)* . If an organization serves a similar portion of traffic in both datasets, the corresponding point should be close to the line $y = x$. But, the data for the top 50 organizations does not fit well with the line $y = x$, as the R-squared value is only 0.4. Most traffic is served by organizations that differ greatly between the two datasets—for example, Apple serves 6% of the campus traffic but 17% of the residential traffic, and Amazon serves 29% of the campus traffic but only 15% of the residential traffic. This result shows large differences between our university's residential network and on-campus network.

Our subset of residences is similar to the university's entire residential network: We also plot a purple star in Fig. 1 for each organization, with the X-coordinate being the percent of traffic from/to the organization for *all residences (category 2)* and the Y-coordinate being the percent of traffic from/to the organization for *our subset of residences (category 3).* The data for the top 50 organizations fits with the line y=x with an R-squared value of 0.7. This suggests that our subset of residences displays similarities to the university's entire residential network and is thus distinct from our campus traces (and likely most campus traces).

3.2 Data collection and anonymization pipeline.

Fig. 2 shows the topology of our residential network. Shared graduate student apartments have one Ethernet port per bedroom, while other apartments have one Ethernet port per apartment. The ports connect to a switch in the residential building, which connects to an aggregation switch and then to the Internet via the campus network and a few providers. The aggregation switch mirrors both traffic to and from our residential buildings over 2x10Gbps dedicated fiber to a server in our nearby lab. Since Columbia has not deployed IPv6 in these buildings, we only study IPv4 traffic.

On the server, our pipeline uses gulp [92] to capture packets in the pcap format. It uses tshark [7] to extract useful fields from unanonymized packets, such as TLS SNI and DNS A record. For domain names, it only keeps the last three suffixes (*i.e.,* for domain a.b.c.d, we only keep b.c.d). After saving useful fields into csv files, our pipeline anonymizes the privacy-sensitive fields (*i.e.,* IP addresses and MAC addresses). For IP addresses outside the university, it does not anonymize them; for IP addresses within our residential buildings, it fully anonymizes them with the Crypto-PAn anonymization scheme [106], but it also retains the building IP subnets for demographic information

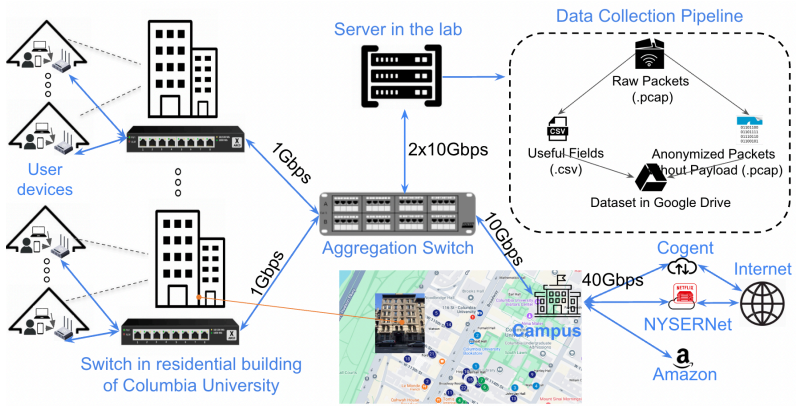


Fig. 2. Our network topology and data collection pipeline. NYSERNet is a Regional Research/Educational Network, and it hosts Netflix cache servers.

(*i.e.,* we know the originating buildings and which packets come from the same unit in a capture, but we do not have IP addresses of the units); for non-residential university IP addresses, it keeps a university subnet. Our pipeline only keeps the Vendor ID unanonymized for MAC addresses. Our pipeline then discards the payload after layer 4 with DagScrubber [17]. After anonymization, it uploads the pcap and csv files to the cloud and deletes the original files.

This work appropriately addresses ethical issues (see more details in Appendix A.1.1). Our data collection/anonymization pipeline follows established practices [73], was approved by Columbia's IT, and received formal review and was declared exempt by our Institutional Review Board (IRB) as it is *not human-subjects research*. Our pipeline anonymizes privacy-sensitive fields and discards personally identifiable information. We do not identify any human or study network usage below the level of buildings.

Due to slow processing speeds and limited storage, we collect traffic for four hours per day. To cover user behavior across different times of day, we space the collection intervals six hours apart. Specifically, we run the pipeline every six hours: we collect *DNS packets* to/from the buildings for the entire six hours and *unsampled raw traces* within the final hour of the six-hour period. Collecting longer DNS traces enables us to map traffic to DNS domains and then to services, even when the DNS response for a flow was cached at the start of our traffic trace collection. We rotate the anonymization key at the end of the period so that one cannot track the long-term activity of a particular unit. We have purchased advanced hardware and plan to scale up to more hours soon.

In January 2022, we started our data collection for more than 400 residences for four non-consecutive hours a day. At the beginning of 2024, we scaled up our dataset to 943 apartments. While we collect a long-running feed, we use the traffic collected to/from the 943 apartments from January 2-26, 2024 as our main dataset for analysis.[1] Our pipeline observes an average of 305GB of traffic and 217 million packets per hour. We justify in Appendix A.3 that we have sufficient days and units for usage trends to converge, and we refrain from using a longer snapshot to limit the impact of changes in user behavior or apartment turnover.

3.3 Data Limitations and Representativeness

We acknowledge the limitations of our data collection. Our users are affiliated with one university and one geographic location, creating biases based on these factors. Our dataset only captures the usage of ~1000 residences. Our pipeline only collects 4 hours a day. These limitations mean that specific details of our datasets and analyses, such as exact quantitative results, may differ from those observed in other residential networks.

Despite these limitations, our study and our data sharing remain valuable to the community, especially given the challenges of accessing residential traces. Our dataset has 10x more residences than some very useful prior studies [28, 91], and we show that we have enough units for some service usage trends to stabilize (Appendix A.3). Given the rise of high-capacity fiber-to-the-home networks [35], the speeds provided to our users are now more typical in residential settings. We also expect that many of our high-level observations are likely to generalize to other user networks, as many trends of Internet usage hold widely, and content providers typically design their deployment and serving strategies based on aggregated user demands. Analyzing the generalizability of our results is an interesting future direction, and our data sharing supports it. Our ongoing collection, coupled with our plan to scale up to more hours and more units, will provide us and other researchers the ability to identify trends over time and emerging Internet behavior.

[1]We miss data for 10 faculty buildings from Jan 10-18 due to a fiber cut.

3.4 Providing access to our existing and future data

We are allowed by Columbia's IT and Institutional Review Board (IRB) to share the datasets with external researchers. While we did not talk about data sharing in the original IRB protocol, we confirmed with our IRB and were informed no further review is required at our institution to share the data because it was deemed non-human-subjects. To protect user privacy and prevent de-anonymization, we will only share data with other researchers after they obtain their own IRB's approval or exemption and they agree to our Acceptable Use Policy (see more details in Appendix A.1.2). These researchers will have access to a longitudinal residential dataset and our future collections, which are not easily available within campus networks and need significant efforts to collect.

We have been collecting since the beginning of 2022, keeping both the anonymized packet-level data (114TB) and the processed flow-level data used for analysis (609GB). We will keep running our data collection and make the new data available. The dataset description and metadata, along with the instructions on requesting our data access, are available on our project website [24]. We further discuss the potential uses of our datasets in §7.

4 METHODOLOGY

Our methodology analyzes traffic at flow level. (We will use the terms flow and connection interchangeably.) We group packets with the same 5-tuple in the IP header (transport protocol, source IP, destination IP, source port, destination port) into a flow. We develop approaches to map traffic to services and to classify remote hosts into servers, unauthorized devices, and peer-to-peer users.

4.1 Mapping Flows to Services

Pair flows with domains. For each flow, we leverage the DN-Hunter algorithm [28, 37] to identify the most recent DNS packet that meets the following criteria: (1) the flow IP and the DNS IP on the client end should be the same, and (2) the flow IP on the remote end should match the DNS A record. For every one-hour capture of traffic, we search the DNS records collected within the last six hours (including the hour of capture) for matched domains. Only 1% of traffic is mapped with DNS packets more than two hours ago. We also conduct an experiment in which we pair a one-hour capture with a week's collection of DNS packets, but it hardly adds any new matches compared to using 6 hours of DNS records. Therefore, we think six hours is sufficient for an hour capture of traffic. For TLS flows, we also associate the flow with its TLS SNI field. For each flow we have a single ⟨`DNS, SNI`⟩ domain tuple, either of which may be blank. We do not use reverse DNS for service mapping, as many IP addresses host multiple services under different DNS names, and the domain returned by reverse DNS is not necessarily the one requested by our users.

*Annotate flows with corresponding services (*e.g., *Netflix, Prime Video).* Our mapping is based on DNS domains, SNI domains, ports, destination IP addresses, and transport protocols. Many services are composed of multiple flows, often to different domains. To improve the mapping rate, we use clustering to learn the temporal correlation between flows. We apply the following steps in order. The details (*e.g.,* mapping keywords, ports, ASes) can be found on our project website [24].

Step 1. Map with keywords within ⟨`DNS, SNI`⟩. For example, domains with 'nflx' are likely Netflix traffic. We compile a list of around 200 keywords for 150 popular services, building off a list in the public traffic inspection tool nDPI [50]. Using these keywords, we map 73% of traffic (by volume) to a service.

Step 2. Map with clusters of ⟨`DNS, SNI`⟩. For traffic with a ⟨`DNS, SNI`⟩ but without a matching keyword, we use an unsupervised learning approach to match it. Our intuition is that multiple domains for a known service are often accessed close in time.

For each pair of ⟨DNS, SNI⟩ pairs (including those with matching keywords) X and Y, we compute a correlation as follows: we count the number of times in our traces that X and Y occur within a configured time threshold of each other, the number of times X occurs but not Y, and vice-versa. We convert these three features into a numeric correlation via heuristically chosen linear weights. The linear map is the same for all ⟨DNS, SNI⟩ pairs. We then cluster ⟨DNS, SNI⟩ pairs using the Louvain method [40].

Clusters consist of ⟨DNS, SNI⟩ pairs with either known services (via keywords) or unknown ones. For each cluster, we identify the most popular service by the associated traffic volume and compute a confidence score as the fraction of traffic associated with the most popular service. For example, if a cluster has 85% Netflix traffic and 15% unknown traffic, the confidence score would be 0.85. We map the unknown traffic to the most popular service if the confidence score for the cluster is $\geq c$, a global confidence parameter. We choose $c = 0.6$ as it offers a good tradeoff between mapping more unknown traffic and being confident that our methodology is correct. On a validation set constructed using keyword mappings (Step 1), $c = 0.6$ leads to 90% of traffic being mapped correctly. This unsupervised ⟨DNS, SNI⟩ to service mapping maps an additional 6.4% of traffic to a service.

Step 3. Map with transport protocol, destination AS, and port. For flows with a ⟨DNS, SNI⟩ where both domains are blank, we manually classify the traffic into a service based on transport protocol, destination AS, and transport port. For example, traffic to destination AS 8075 on port 3480 is likely Microsoft Teams traffic according to public documentation [13]. Using these rules, we map an additional 4.3% of traffic to a service, leading to a total of 83.7% of all traffic mapped.

Step 4. Map all other traffic to its own service. For flows with a ⟨DNS, SNI⟩where either is nonempty, we let that ⟨DNS, SNI⟩ be the service. Otherwise, we let the associated destination AS, port, and protocol be its own service.

Associate each service with a high-level "service-type" (e.g., *video, social media).* We use the service-types from the Sandvine report [90]. Sandvine's methodology is opaque, and so this process introduced some uncertainty as some services could potentially map to multiple service-types (*e.g.,* FaceTime could map to communication or video, and we map it to video).

Evaluation and limitations. Our domain clustering identifies keywords we did not think of. For example, the unsupervised approach mapped aka.warnermediacdn.com to Paramount+. Moreover, our choice of the clustering parameter achieves 90% accuracy on a validation set, which was sampled from traffic with keyword mappings (Step 1).

We acknowledge that our mapping from flows to services may be incorrect or incomplete in some dimensions. Some domains may be used for one service one time and another service another time, but the clustering may associate them with the most popular service. The technique also cannot identify new services that we did not think of.

4.2 Classifying Remote Hosts

A remote host can be a *server* that delivers traffic for a service. It can be a *P2P user* that establishes connections with our users for peer-to-peer (P2P) traffic. It can also be an *unauthorized device* that attempts to connect to our users by sending unsolicited traffic, due to scanning, attacking, or misconfigurations. We refer to these devices as *unauthorized devices* because the traffic is not initiated by users. We consider the *servers* and *P2P users* as *authorized devices.* Our classification of remote hosts is conservative in labeling unauthorized devices and P2P users. We describe our methodology as follows and then discuss its evaluation and limitations.

Step 1. Identify servers if one of the criteria below is met: (1) There was ever a flow with the remote host that was paired with a DNS record; (2) The port used by the remote host was ever 80 or 443; (3) The remote host is identified as an off-net server (*i.e.,* appliances deployed by large providers within

user networks) by Gigis *et al.* [57]; (4) The remote host is within the networks of large content providers, clouds, and CDNs.

Step 2. Identify unauthorized devices. When sending unsolicited traffic, an unauthorized device generally should not elicit a response, but it may occasionally succeed at eliciting one. We say remote hosts that do not generate a response 90% of the time they contact a host in our network are unauthorized devices.

Step 3. Identify P2P users based on three properties of P2P:

(1) Peers tend not to be accessed through DNS. We deem the remote hosts once paired with domains unlikely P2P users.

(2) Peers tend to be human users. We leverage two existing datasets that record IP prefixes with web activity. One dataset comes from a previous paper that identifies web clients by issuing non-recursive DNS queries to recursive resolvers for popular domains from prefixes [70]. A successful DNS reply indicates that the prefix has performed a recent DNS lookup, suggesting that the prefix hosts web clients. We also tabulate prefixes that launched speedtests from a Google search page in January 2023 [59], as these prefixes likely represent web clients which are likely human users.

(3) Peers tend to use non-privileged, high-numbered ports. P2P traffic normally does not have dedicated port numbers and uses ephemeral OS-assigned ports instead, which are automatically and dynamically assigned within a predefined range of port numbers by operating systems. Further, many P2P peers are behind NATs, which assign ephemeral ports. Various systems and standards share different port ranges. RFC 6056 suggests a range beginning with 1024 [81], which is used by some OSes. So, we treat ports ≥ 1024 as a signal.

Overall, for us to label a remote host as *P2P user*, it must meet all the following criteria: The remote host was never seen in a DNS answer, does not use a system/well known port (0-1023), does not belong to a hypergiant or similar AS that predominantly hosts server-based services, and has its prefix identified as hosting web clients by one of the existing datasets.

Step 4. Label the remaining traffic as uncertain.

As our goal is to understand services and serving infrastructure accessed by users, our analysis in the following sections considers *only traffic from/to authorized devices*, except when discussing the classification results of remote hosts (§5.1).

Evaluation and limitations. While it is difficult to find a ground truth dataset for P2P applications, we evaluate our method by comparing our classifications of traffic on ports known to be used by BitTorrent (TCP ports 6881 to 6889) and FaceTime (UDP ports 16384-16387 and 16393-16402) [11, 25]. Our method checks for ports ≥ 1024 but does not classify based on specific port ranges, yet it classified 97% of traffic for the BitTorrent and FaceTime port ranges as P2P. BitTorrent and FaceTime running on the known ports account for 28% of the classified P2P traffic. An additional 15% of the classified P2P traffic uses port numbers that are registered for P2P calls, file transfers, or remote control [6]. Our results also align with our intuition that P2P may result in a larger fraction of outgoing traffic (*i.e.,* from users to remote hosts) than other types of connections (Appendix A.4).

Moreover, many hosts that our method classifies as unauthorized devices are confirmed by an online database of abusive IP addresses [1] (Appendix A.4). It reports 58% of a random sample of our unauthorized devices as sources of scans, spam, or abuse. The remaining may be due to IP addresses not being abusive or because we study different datasets.

Our classification of remote hosts may also be incorrect or incomplete for multiple reasons. While we use port 80 and 443 to identify servers, scanners may also use these ports to bypass firewalls. We use existing datasets that capture web clients [59, 70], which may be incomplete. We also rely on the assumption that web clients are highly likely to be human users, which may be flawed. Moreover, applications can be configured in a way that does not align with the standard

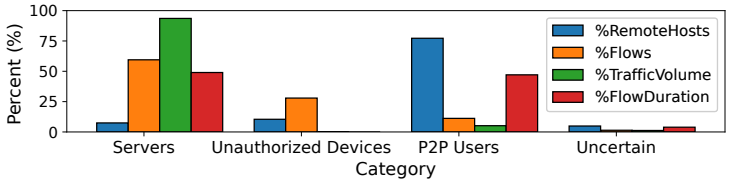


Fig. 3. Breakdown of remote hosts by each category, along with associated flows, traffic, and flow duration. P2P users contribute only 5% of traffic but 47% of flow duration.

settings. For example, P2P applications can choose to use ports below 1024 (*e.g.,* DC++, a P2P file sharing client, uses port 411). Addressing these limitations is an interesting avenue for future work.

4.3 Latency, Organizations, and Off-nets

We use scamper [85] to conduct ping measurements from a wired device in Columbia University to remote hosts seen in our traces. We probe the hosts 10 times non-consecutively and take the minimum of 10 ping round-trip times (RTTs) to approximate propagation delay. To avoid sending a significant amount of traffic, we rank the targets by traffic volume and measure only those contributing to the top 99.9% of traffic. We also limit the target packets-per-second rate to 10000. Additionally, we use RouteViews [5] and ASRank [2] to identify the organization that owns the remote host of a connection. We use an existing dataset from prior work to identify off-nets [57]—appliances within other networks owned by hypergiants [42, 102]. We also geolocate remote hosts using HOIHO [18], RIPE IP Map [52], and custom reverse DNS keywords matches. (More details will be covered when we discuss the related analysis.)

5 SERVING INFRASTRUCTURE

Serving infrastructure deployments impact how efficiently and reliably users can access services. Simply measuring where hypergiants deploy their servers (as some prior studies have done [42, 57, 88]) is insufficient to infer traffic patterns—a rich understanding of the role of serving infrastructures requires associating infrastructures with the user traffic they deliver. In this section, we investigate key aspects of the serving infrastructures that our users access: the categories of remote hosts (§5.1), the organizations hosting services (§5.2), the locality of services (§5.3), and the mechanisms steering users to servers (§5.4). We discuss the insights learnt from our analysis in §7. We present analysis only on the main dataset for clarity, as findings remain the same unless explicitly specified.

5.1 What Are the Remote Hosts?

We classify the remote hosts accessed by our users as *servers*, *P2P users*, and *unauthorized devices* (see the methodology and evaluation in §4.2). Fig. 3 shows the percent of total flows, traffic, remote hosts, and flow duration for each class of remote host. We do not compute total flow duration for unauthorized devices, as the vast majority of those cases are probing/scanning traffic.

Observation 1: The popularity of remote host categories differs greatly by traffic volume, flow count, flow duration, and DNS query count. P2P use contributes 5% of traffic and 47% of flow duration, likely caused by both file sharing and videotelephony.

Servers (7% of remote hosts) account for 93% of traffic, and web servers (i.e., those that deliver traffic via port 80 or 443) account for 91% of traffic. However, many flows (30%) originate from unauthorized devices, and users still connect with numerous P2P users (77% of remote hosts). We do not include the number of DNS responses in which a remote host appears as a metric, because DNS records do not cover the remote hosts used for P2P and unsolicited connections. In addition, 64%

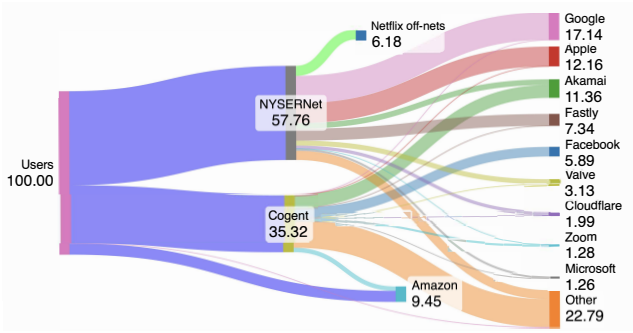


Fig. 4. Key organizations on the routes between our users and remote hosts. NYSERNet is a regional provider for research/educational institutions in New York, including Columbia. Netflix's off-net caches are deployed in NYSERNet and account for 6% of total traffic (incoming plus outgoing).

of remote hosts seen in DNS records were not used by any connection, likely due to applications prefetching DNS records [28] or receiving multiple IP addresses for load balancing or resilience but not using them.

P2P peers account for 5% of traffic and 47% of flow duration. Identifying services driving P2P traffic is difficult—we cannot use domains as P2P peers are not accessed by DNS (§4.2), and many services use dynamic ports. But, we managed to identify some BitTorrent and FaceTime traffic with known ports [11, 25]. The identified BitTorrent traffic accounts for 27% of the total flow duration for peers, while the identified FaceTime traffic represents 0.04%. In terms of volume, BitTorrent traffic is 1%, while FaceTime traffic is 27%. This FaceTime traffic establishes a lower bound on the amount of P2P video calls and suggests that video calls are a large source of P2P traffic, whereas prior studies found that Skype, FaceTime, and Zoom use P2P architectures but could not assess how their traffic volumes compared to other services [86, 109].

5.2 Who Owns the Remote Hosts?

After studying the remote hosts by category, we focus on the traffic from/to authorized devices for the remaining paper. We analyze which organizations own these remote hosts, whether the organizations host their own services or other services, and the changes over time.

Observation 2: CDNs (*e.g.,* Cloudflare) can host many sites but serve low traffic volumes.

Fig. 4 visualizes the flow of traffic between our residential users and the organizations of the remote hosts with a Sankey diagram. For better visualization, our Sankey diagram only depicts Columbia providers and the organizations of remote hosts that account for more than 1.2% of traffic. Amazon directly peers with our university and accounts for 9% of our user traffic. The local Netflix off-nets are deployed within NYSERNet, a regional research and education network, and account for 6% of total traffic, which we will discuss later in this subsection. All other traffic goes through one of Columbia's providers, with much of it from/to popular clouds, CDNs, and content providers.

In addition, Fig. 5 shows the percent of total traffic, flows, remote hosts, flow duration, and domains served by the top organizations by traffic volume (we only display the top 11 for clarity). Combined, these 11 organizations account for 77% of traffic. Google serves the largest share of our user traffic (17%), with YouTube making up the majority (66%) of Google's traffic volume. We observe many sites hosted by CDNs, with Cloudflare hosting 19% of domains. However, Cloudflare (a widely used CDN) only serves 2% of traffic, suggesting that many sites served by Cloudflare are less popular and/or small. Similarly, Amazon hosts 26% of domains but serves a much smaller percentage of traffic (9%), with 7.5% out of that 9% for Prime Video, suggesting that the other

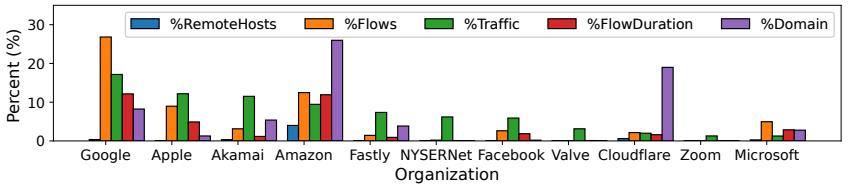


Fig. 5. The percent of flows, traffic, flow duration, and domains served by the top organizations. Noteworthily, while Cloudflare hosts 19% of domains, it only serves 2% of traffic.

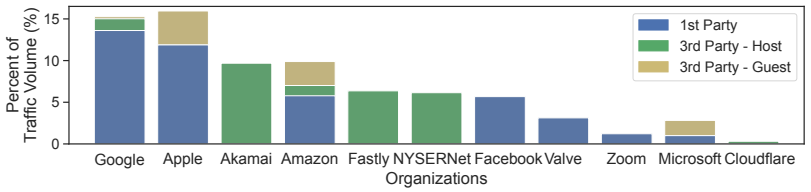


Fig. 6. The breakdown of the top organizations (by traffic volume) based on how their traffic is delivered.

domains make up only 1.5% of total traffic. Akamai and Fastly contribute much more traffic because they host many video services. We are unaware of prior studies with similar findings.

Observation 3: The majority of traffic is served by vertically integrated content providers that both operate the services and the infrastructure the services run on (*i.e.,* 1st-party).

Modern hypergiants often serve different roles for various services. For example, Amazon hosts its own services such as Prime Video as a 1st-party cloud provider and CDN (*1st party*), hosts Netflix web servers as a 3rd-party cloud provider (*3rd party - Host*), and also uses other CDNs to cache content (*3rd party - Guest*).

To quantify how traffic is delivered, we manually associate the mapped services with organizations and check if the infrastructure owner matches the service owner. If so, we consider the connection *1st-party*; otherwise, *3rd-party*. Our manual mapping covers the largest organizations and the most popular services, which accounts for 75% of total traffic volume. While 47% of all the traffic is labeled as 1st-party, 28% is labeled as 3rd-party. In addition to this 75% labeled traffic, 19% is either associated with a less popular service or from the many smaller organizations that we did not check, and 5% is P2P.

A previous study found that 89% of the Alexa top websites depend on 3rd-party services [72], but it relied on actively fetching landing pages and so did not account for pages customized to real users, pages as served to logged-in users [32], internal pages [31], or realistic browsing patterns and page popularities. In contrast, our traces reflect real usage patterns and are dominated by 1st-party traffic, suggesting that 3rd-party dependencies are less prevalent when weighted by traffic volumes.

Fig. 6 shows the compositions of the top organizations (by traffic volume). At one extreme, Facebook, Valve, and Zoom neither rely on 3rd-party CDNs nor deliver traffic for others. At the other extreme, NYSERNet, Akamai, and Fastly only deliver content for others. Most traffic from Google and Apple infrastructure is for their own services, and most traffic from their services comes from their infrastructure. However, Google also hosts other services as a cloud provider, and Apple and Microsoft also rely on other CDNs for their services (including updates). While the figure does not show Microsoft hosting 3rd-party services, it does host a number of services (*e.g.,* auction.housingworks.org, the online auction site of Housing Works, a NY-based nonprofit) with traffic too small to be part of our manual mapping included in the figure. Other factors also explain the limited 3rd-party hosting by Microsoft: many applications use Microsoft for backend cloud

services such as storage, but they either host their web frontends on CDNs (so are classified as hosted by the CDN) or deliver content directly from the Microsoft cloud service domains (so are classified as 1st party); and Microsoft is particularly prevalent in the enterprise market, so we may see less use from Columbia's residential buildings. Amazon has a more balanced hosting for 1st party and 3rd party traffic.

Observation 4: We observe longitudinal changes in the serving infrastructure—the decommissioning of Akamai cache servers hosted in NYSERNet and the reduced use of Lumen (aka Level 3) CDN.

For the top 11 organizations (by traffic volume), we investigate changes in the percent of traffic they deliver. We observe significant and consistent declines in NYSERNet and Lumen CDN. We also observe increases in Akamai and Amazon but do not discuss them here, as they are also related to the increased usage of some services (Fig. 11a).

Off-nets in NYSERNet. An off-net is a server managed by a hypergiant but hosted in another network, while an on-net is a server in its own network. Off-nets serve users in the deployed network or its customer networks, efficiently delivering content and reducing costs. We use data from a prior study that identified off-nets in all ASes for 22 popular hypergiants [57]. It found that our university does not host off-nets but that its main provider (NYSERNet) hosts off-nets for Netflix and Akamai.

We expected that these off-nets would deliver large amounts of the user traffic, but only some results matched our expectation. Two Netflix off-nets deliver 6% of total traffic (one much less than the other), which is 94% of the total Netflix traffic (with the rest from Netflix on-nets) and is similar to Netflix claims that off-nets can deliver at least 95% of a network's Netflix traffic [56]. In the early months of our traces, Akamai off-nets in NYSERNet delivered a small fraction of the Akamai traffic, but they did not show up in our traces since May 2023. These off-nets are now unresponsive to queries on port 443, suggesting these off-nets might no longer be in use. This may reflect a broader trend of Akamai decreasing the number of networks in which it hosts off-nets [57]. NYSERNet confirmed our findings about Netflix and Akamai, including that Akamai asked NYSERNet to decommission the off-nets.

Drop in Lumen CDN usage. Lumen operates a Tier-1 network and a CDN running on it (Lumen was formerly known as CenturyLink, which acquired Level 3 and its CDN in 2017). In December 2022, Lumen served 2.7% of total traffic from/to our network. We observe a gradual drop in Lumen CDN usage, and Lumen only served 0.3% of total traffic in January 2024. Notably, while the shares of Hulu, Disney+, and TikTok remained stable or even increased, the delivery of these services through Lumen decreased. We observe little Hulu and Disney+ traffic hosted on Lumen since April 2023 and little TikTok traffic on Lumen since November 2023. These changes are likely correlated with the sale of CDN service contracts from Lumen to Akamai [19], and we observe Akamai hosting more of those services.

5.3 Where Are the Remote Hosts?

In this subsection, we study the locality of user mappings to deployments, as better locality implies that users can connect to closer deployments and have reduced latency. We use the minimum measured ping round-trip time (RTT) as an estimate of the propagation delay. We measured to targets that account for 99.9% of the total traffic and received ping responses from 73% of the targets. The responsive ones account for 88% of the total traffic. We do not infer RTTs from our traces,

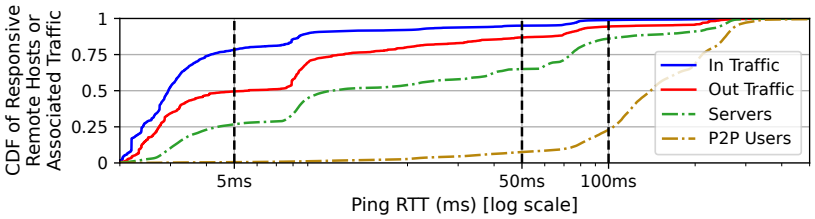


Fig. 7. Most of the incoming traffic (78%) is within 5 ms, indicative of service locality.

as they may include application-layer processing delays. We use ICMP pings, which have RTT measurements close to TCP ping [103].

Observation 5: Most incoming traffic (78%) comes from nearby servers (within 5ms), but only half of the outgoing traffic goes to them.

In Fig. 7, we plot the CDF of RTTs across traffic associated with the responsive remote hosts. We find that remote hosts with RTT within 100ms account for nearly all traffic. The nearby remote hosts with RTT within 5ms deliver 78% of the incoming traffic. However, they only account for 50% of the outgoing traffic, while 87% of outgoing traffic can be delivered within 50ms.[2] This difference arises because services usually have an imbalanced use of incoming and outgoing traffic. Many popular video services are deployed near the users and deliver much more traffic than they receive. For example, the Netflix off-nets are about 2ms away, and they account for 7.5% incoming traffic but only 0.7% of outgoing traffic (considering only responsive remote hosts). On the other hand, we observe many remote hosts that receive much more traffic than they deliver, likely reflecting user uploads and syncs.

Prior studies using data from edge networks also observed the reliance on nearby servers for a few service or content providers [100, 108]. In contrast, we provided an overall distribution across all responsive remote hosts and compared the locality of incoming traffic with that of outgoing traffic, highlighting that much outgoing traffic is not delivered to nearby remote hosts.

We also plot the CDF of responsive remote hosts that are identified as servers and P2P users. We find that 27% of servers and 0.5% of P2P users have RTT within 5ms. The gap between the line of responsive servers and that of responsive P2P users shows that the accessed servers are relatively closer to users.

Observation 6: Some popular services (*e.g.,* FaceTime, Zoom, Ubisoft) are often served from at least 10ms away from the closest remote hosts.

As a service can be delivered by multiple remote hosts, we now turn to quantify how much further those remote hosts are from the closest ones on average (weighted by traffic). Since propagation delay can be inflated for various reasons, we also use great-circle distance. We compute this by issuing traceroutes to remote hosts from a lab machine on our campus and geolocating with HOIHO [18], RIPE IP Map [52], and custom reverse DNS keyword matches (*e.g.,* edgeray-shv-01-iad3.facebook.com has an IAD airport code so is likely near Washington D.C.). For cases where we cannot geolocate remote hosts but can geolocate a hop on the path, we use the hop's geolocation if its RTT is within 1ms of the destination. We can geolocate remote hosts serving 58% of traffic. We convert the distances to the speed of light in fiber to directly compare to latency.

[2]In Jan 2025, we conducted TCP ping measurements by sending TCP-syn packets to the observed <IP address, port> pairs. This methodology was also adopted by prior studies [10, 103]. The conclusions remain similar. For example, 70% of the incoming traffic comes from servers within 5ms, but only 37% of the outgoing traffic goes to them.

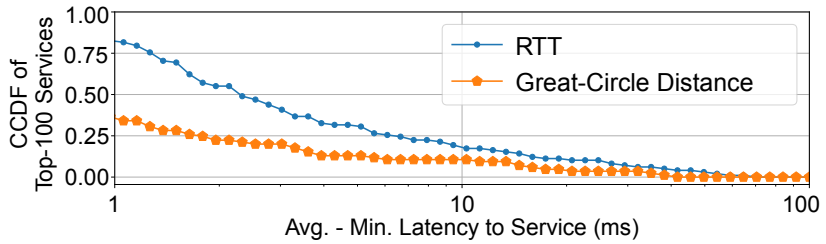


Fig. 8. The gap between the overall *latency* weighted by traffic volume and the minimum *latency* to servers for each of the top 100 services. The great-circle distances are converted to the speed of light in fiber to compare with latency. We find that 19 of the top 100 services have a latency gap of at least 10ms.

For each of the 100 highest volume services, we calculate the gap between the *propagation delay* (and *distance*) to its nearest server and the overall *propagation delay* (and *distance*) with each of its servers weighted by the bytes delivered and received. We exclude anycast addresses found by prior work [97] from this calculation. In Fig. 8, we plot the CCDF of services with respect to these gaps. We find that 19% of these services have a latency gap of at least 10 ms; 11% have a distance gap of at least 10ms, meaning that their traffic travels an additional 1000 km each direction on average, compared to the closest remote host. This finding suggests that although deployments are widespread and mappings can be near optimal, a noticeable volume of traffic still flows to/from distant remote hosts. We are unaware of similar prior studies.

Services with a latency gap of at least 10ms include videotelephony services such as FaceTime and Zoom, VPN services such as WireGuard, gaming services such as Ubisoft, and video streaming services such as Spectrum, Bilibili TV, and some adult video websites. It is not surprising that FaceTime and WireGuard have a large latency gap, as they rely on various peers for different connections. The domains related to Zoom indicate the used data centers (*e.g.*, sjc.zoom.us represents the San Jose data center), and we observe a correlation between the domains and the RTT to remote hosts delivering traffic. For other services, however, the RTT to remote hosts varies even when using the same domain. There could be multiple reasons for directing our users to distant servers, including regulatory requirements to store user data in specific countries [14], system designs of not serving users locally for better load balancing [47], the retrieval of contents unpopular and thus uncached in nearby servers, and the cost of providing services. These examples indicate that the services may prioritize other measures over serving users locally or be limited by other factors.

5.4 How is Traffic Steered to the Servers?

We now examine how our users are steered to servers. The primary steering mechanism is DNS, and so we investigate the prevalence of DNS TTL violations.

Observation 7: Despite the widespread adoption of anycast [48, 49] and recent standards for encrypted DNS [65, 66, 69], we observe the vast majority of traffic is from unicast addresses assigned via unencrypted DNS. We were unable to associate 10% of the traffic with any DNS record due to encrypted DNS, proprietary configurations, or limitations in data collection.

DNS is typically used to redirect users to servers. In our main dataset, 83% of flows and 90% of traffic to servers were steered by unencrypted DNS lookups, as we were able to associate them with DNS responses in our traces. (All percentages in this subsection are based on the total number of flows or total traffic volume directed to servers.) For the remaining 17% of flows, we did not observe an associated DNS lookup, a larger percentage than the 7% of flows in a 2020 study [28].

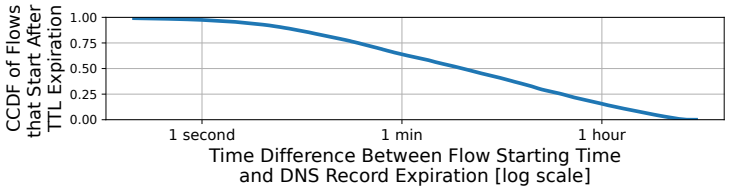


Fig. 9. CCDF of flows that start after DNS record expiration. Among the 14% of flows that start after their corresponding DNS records have expired, 50% of them start over 207 seconds after DNS record expiration.

We identify three possible reasons why these flows were not paired with DNS responses, which might explain the growth from 7% to 17%. The first reason is the use of encrypted DNS protocols. This may only explain a portion of the traffic, since Columbia's DNS resolvers, the default ones that we expect most users to utilize, do not support encrypted DNS. We also only observe an hourly average of 43 packets accessing port 853 (used for DNS-over-TLS). It is hard to identify encrypted DNS traffic, as DNS-over-HTTPS uses the widely used port 443. Even when we identify potential encrypted DNS packets, we lack a way to associate them with specific flows.

Second, some applications do not redirect users in a traditional way. Instead of issuing domain queries to operating systems and waiting for their responses, they adopt their own protocols or specify the use of particular DNS resolvers. An example is the HTTPDNS protocol designed by Tencent, which sends DNS queries to Tencent's DNS resolvers over the HTTP(S) protocol [4]. Some applications even hardcode IP addresses in their implementations for simplicity.

Third, even if users access services with traditional DNS, our methodology will fail when users already have the relevant DNS records cached on their devices. For example, our methodology will miss cases when a user issues a DNS query on a different network, then caches it, moves to the residential network, and contacts the service using the cached record. Another reason could be instances when users access the services before our DNS traffic collection, and their local devices cache the related DNS records through the beginning of our collection. However, as discussed in §4.1, this should only contribute to a trivial amount of traffic.

UDP vs. TCP. As expected, most connections to servers (81% of flows, 89% of traffic) was redirected by DNS over UDP, while only 2% of flows (carrying 1% of traffic) was steered by DNS over TCP.

Anycast vs. Unicast. Anycast refers to advertising the same IP prefix from multiple server deployments, while unicast refers to advertising an IP prefix from one location. We identified anycast IP addresses using the dataset of the Anycast Census [16, 97]. We observe a heavier use of unicast IP addresses (84% of flows, 94% of traffic) than anycast IP addresses (16% of flows, 6% of traffic). Among the top 11 organizations by traffic volume: most Cloudflare traffic is delivered from anycast addresses; Microsoft delivers half of its traffic via anycast and half via unicast; and the others mostly use unicast addresses. Prior work found that content providers and CDNs widely adopt anycast [48, 49, 60]. By directly comparing anycast with unicast in passive traces, we highlight that unicast addresses still account for the majority of traffic.

Observation 8: Many flows (14%) use DNS results after the DNS Time-To-Live has expired, suggesting that CDNs and content providers using DNS to steer clients to unicast addresses may have limited responsiveness to failures, overload, and performance change.

DNS records come with time-to-live (TTL) values, which specify the time interval a record may be cached before it should be discarded. Small DNS time-to-live (TTL) values slow applications (by increasing the fraction of queries that cannot be answered from cache) but still cannot guarantee up-to-date DNS records for clients due to violations when records are used after the TTL expires. The

phenomenon of TTL violations by client software has been documented in technical blogs [3, 12], and prior work found that many connections use outdated DNS information [28].

Our results confirm this, showing that 14% of flows start after their corresponding DNS records have expired. Fig. 9 plots the cumulative fraction of TTL-violating flows that start at least x seconds after the DNS TTL expired. The figure shows that 50% of TTL-violating flows start more than 207 seconds after the DNS record expired. We observe virtually identical trends when limited to domains hosted on major cloud providers (Google, Microsoft, Amazon) and CDNs (Akamai, Fastly, Limelight). In fact, between 20-85% of bytes sent to those cloud providers are sent more than a minute after the TTL expired. This suggests that DNS steering to unicast addresses restricts the content provider's ability to respond to failures, overload, or performance changes.

6 SERVICE USAGE

Understanding service usage is essential, as it provides insights into user needs and highlights services that deserve further study. We provide a fresh and thorough understanding, investigating less studied aspects including service popularity across multiple metrics and demographics (§6.1). We also examine changes in service usage over time (§6.2) and present a short study case on two large Netflix live streaming events to show how our dataset can be used to detect performance differences and degradations even for encrypted traffic (§6.3). We discuss the lessons drawn from our observations in §7.

6.1 Popular Services and Service-Types

We assess service popularity with four metrics: total traffic, number of flows, total flow duration, and number of DNS responses for a service. Every activity metric then has a service distribution associated with it, corresponding to its share of the metric.

The top service-types by traffic volume are Video (50% of traffic that we could map to a service), Cloud (17%), Social Media (10%), and Gaming (9%). When it comes to services, the top 10 include YouTube (11.6% of total traffic), Prime Video (7.5%), iCloud (7.5%), Netflix (6.7%), Instagram (4.5%), Steam (3.5%), Apple Store (3.0%), Hulu (3.0%), Tiktok (2.4%), and Apple Digital Services (2.2%).

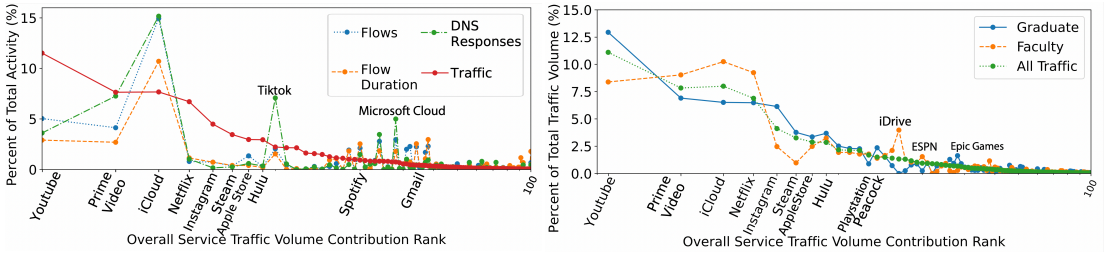
Observation 9: Relative service popularity depends considerably on metrics (*i.e.,* traffic, flows, flow duration, DNS responses).

For the top 100 services (sorted by traffic volume), we compute their share of activity across all metrics in Fig. 10a (where each index represents a service). Flow count and flow duration exhibit the most similar activity patterns, and traffic volume differs greatly from the other metrics (see the spikes in the figure). For example, iCloud accounts for half the traffic volume of YouTube but three times the flow duration. Slack and Gmail have many flows but relatively little traffic volume.

While this point may seem obvious, it is worth analyzing quantitatively. Many measurement studies [29, 30, 72], for example, focus on websites from lists of top sites, but these lists may not always be the best fit for their research objectives. Our findings emphasize this challenge, show how popularity can vary across metrics, and offer a basis for making more informed, cross-metric decisions in future studies.

Observation 10: Even within affiliates of a single university, service traffic usage can vary considerably across demographics.

Demographic information is hard to obtain. A prior study inferred student demographics (domestic vs. international) and compared their traffic usage during the pandemic [101]. Sandvine, a company that monitors access networks globally, reports Internet application usage by region [90].



(a) Relative service popularity by popularity metric. (b) Relative service popularity across resident types

Fig. 10. Service usage differs by metrics and across resident types.

We offer a different perspective on demographics, showing the differences between student traffic and faculty/family traffic. We define the demographic as the resident type of the building from which traffic originates. We acknowledge that this coarse definition misses important information such as ethnicity, gender, and age, but it still highlights the intensity of differences across demographics.

In this comparison, we disregard buildings with mixed resident types and divide the remaining traces into two groups: graduate students and faculty. We plot the percent of traffic for each resident type for the top 100 services in Fig. 10b. Units exclusively housing faculty and families show different service usage patterns from those exclusively accommodating graduate students, highlighting the value of richer data sources to cover a wide variety of demographics.

While video streaming is popular for both groups, graduate students have more YouTube traffic (13%) than faculty (8.4%), but less Netflix traffic (6.5%) than faculty (9.3%). Graduate students also show a preference for Hulu and Peacock compared to faculty, potentially because those video platforms offer student discounts [15]. Another notable difference is that faculty and their families generate more iCloud traffic (10.2%) than graduate students (6.5%). In terms of social media, graduate students use Instagram more frequently (6.1%) than faculty (2.5%). With regard to gaming, the most popular application for graduate students is Steam (3.8%), while it is Playstation for faculty (1.8%).

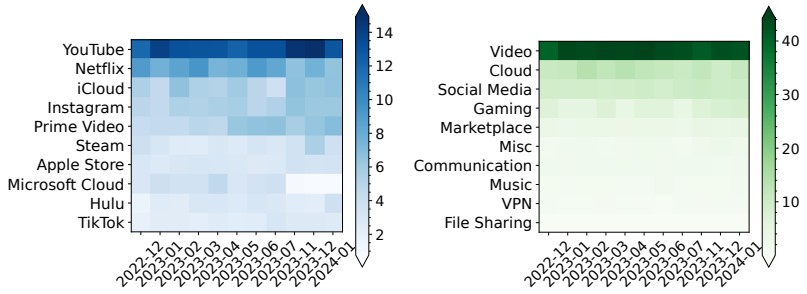
To put these differences into perspective, we compare them against regional differences reported by Sandvine. Sandvine reports application-type (*i.e.,* video, communication) breakdowns by traffic volume per global region: America, Asia-Pacific, and European. We compute Bhattacharyya distances [38] between the service-type usage distributions for these three regions and between the distributions for our graduate students versus faculty. Graduate and faculty distributions, although affiliated with the same university, are as different as American and European distributions and are more different than Asia-Pacific and American distributions. This suggests that relying on only student traffic, whether our graduate student traffic or undergraduate dorm traffic from campus traces, provides limited view, and we need a broader set of data sources.

6.2 Temporal Service Usage Changes

To understand short-term and long-term changes in service usage, we consider only traffic to/from the graduate student residences to limit confounding factors due to user demographics. We used two datasets that span more days than the main dataset (Appendix A.2).

Observation 11: While the service popularity varies notably over time (*e.g.,* 79% drop in Microsoft cloud traffic), the service-type popularity remains stable.

To analyze the longitudinal changes of service usage, we use traces collected over a year (December 1, 2022 to January 26, 2024) from graduate student apartments (to limit confounding factors due to user demographics). Fig. 11 shows the monthly traffic usage for the 10 most popular services



(a) Services (b) Service-Types

Fig. 11. Service popularity changes notably over time, but service-type popularity does not. Note that each box represents the percent of traffic for a service (or service-type) out of all traffic within the month.

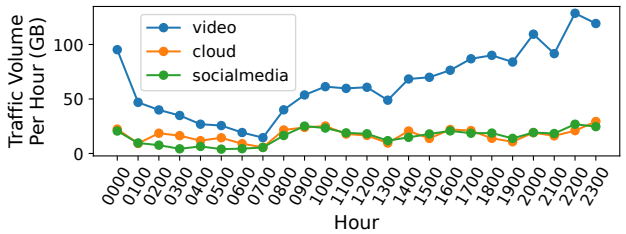


Fig. 12. Percent of traffic for three popular service-types for each collection hour.

and for all service-types. The color intensity of each box for a given service and month represents the percent of traffic for that service *out of all traffic within the month*.

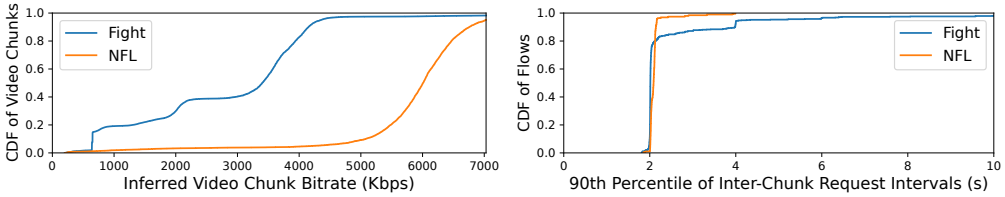
Fig. 11a shows noteworthy changes for some services. For example, Microsoft cloud has become less popular within our network since late 2023, dropping from 4% of traffic in July to 1% in November. Although the decrease is only a small percent of overall traffic, it marks a 79% drop in Microsoft cloud traffic, and this decline is persistent. We found that Microsoft Cloud received more traffic than it delivered, with the decline primarily in the received traffic. We leave more sophisticated investigation for the future. Steam is nearly twice as popular in December 2023 (6%), when our users are enjoying their winter breaks, than in other months (about 3-4%).

Fig. 11b illustrates that the use of service-types has not changed greatly over the last year. Despite the large drops in Microsoft Cloud, the percent of traffic for cloud stays relatively stable, and the relative popularity for service-types does not change.

Observation 12: While overall service usage is lower in the early mornings, the use of cloud applications remains stable.

We varied the hours during which we collected traffic from November 20, 2023 to January 26, 2024 (we collected at least 7 days worth of data for each hour of the day). In Fig. 12, we plot the hourly traffic volume for the top three service-types. Video is unarguably the most popular category, but the use of video applications drops during the early mornings when people are asleep. We observed a similar trend for social media.

In contrast, cloud applications stay roughly the same throughout the day. iCloud is the most used cloud service, and automatic data syncs and updates are likely why cloud services stay active when users are asleep. Hence, cloud takes up a larger percentage of traffic during early mornings (~20% between 2am-9am), whereas the overall fraction of cloud traffic is 13%.



(a) Video Bitrate (b) Rebuffering

Fig. 13. Streaming quality for Paul vs. Tyson fight and the NFL Christmas game. The NFL live event achieved better performance, as its viewers used higher bitrates (Fig. 13a) and had less rebuffering (Fig. 13b).

6.3 Case Study: Two Large Netflix Live Events

Our dataset can be used to observe Internet events, quantify performance for real users (from one network), and provide insights into poor performance. We examine the performance of two large live streaming events hosted by Netflix: *Jake Paul vs. Mike Tyson* (a boxing bout on Nov. 15, 2024) and *NFL Christmas Gameday* (two NFL games on Dec. 25, 2024). They were among the most-viewed live events Netflix has streamed, with tens of millions of live viewers [21, 22].

Video streaming services encode content at various quality levels, divide the content into chunks, and use adaptive bitrate streaming algorithms to select the chunk that best fits the available bandwidth [67, 98]. For example, the live streaming service Twitch encodes content into chunks of 2 seconds, and a new chunk is requested every two seconds [8]. We confirm with active measurements that Netflix also uses a 2-second chunk duration for live streaming (but not for video-on-demand).

We analyze the anonymized packet traces we collected to infer our users' experiences. For flows mapped to Netflix, we identify chunks and chunk requests with an established algorithm [62, 76]. To separate these live events from users viewing Netflix video-on-demand movies and shows, we include only flows for which the majority of time differences between chunk requests is around 2 seconds (*i.e.,* the 25th percentile of time differences is ≥ 1.9, and the 75th percentile is ≤ 2.1). We infer the chunk bitrate as $\frac{\text{chunk size (Kb)}}{\text{2 (s)}}$. We select chunks with bitrate$\geq$ 200 Kbps as video chunks.

For the Paul vs. Tyson event, we use the traces between 9pm-10pm PST, Nov. 15 (which covered the match between the two and was during the semester at our university). During the hour, 2% of total traffic is for Netflix live streaming. For the NFL Christmas Gameday, we use the traces between 2-3pm PST, Dec. 25 (during the second game Ravens vs. Texans and during the inter-semester break at our university). During the hour, 1% of total traffic is for Netflix live streaming.

Anecdotally, viewers had streaming quality issues and experienced rebuffering events during the boxing bout [20, 23], but they had good experiences during the football games. We plot the CDF of inferred video bitrate for both events in Fig. 13a. While 91% of video chunks for the NFL event had bitrates above 5000 Kbps, only 3% of video chunks for the fight reached bitrates over 5000 Kbps. In Fig. 13b, we plot the CDF of the 90th percentile of time differences between chunk requests for each flow. A streaming session without rebuffering should have requests sent every two seconds, and so a large 90th percentile suggests a long freeze in the session. While only 2.6% of flows for the NFL event had the 90th percentile over 2.5 seconds, 15.2% of the flows for the Tyson fight exceeded 2.5s. The results corroborate that the Tyson fight had worse performance and suggest that Netflix has improved its live streaming approaches for better user experience. This case study shows that our dataset can help monitor network events and infer service performance even for encrypted traffic.

7 LESSONS AND POTENTIAL USES OF OUR DATASET

We now summarize what to learn from our findings above and how to potentially use our dataset.

A detailed, broad study helps uncover overlooked aspects and motivate future research [Observation 1, 2, 7, 8, 9]. For example, since we assess the popularity of service architectures with multiple metrics, we observe notable P2P activity for videotelephony, suggesting potential future research on how videotelephony services utilize P2P. Another example is our finding of connections using unicast addresses and outdated DNS records long after they expired, potentially compromising CDN's availability during failures. In fact, this finding motivated a paper to design a new traffic routing system [75].

Even with active measurements, DNS datasets, and campus traces, it remains crucial to analyze residential traces [Observation 1, 2, 3, 5, 6, 9, 10]. As prior studies with active measurements suggest [43, 44, 47, 55, 74, 77, 83, 93, 107], we observe deployments close to users. But, we find that a notable volume of traffic still flows to more distant servers. In addition, we observe less traffic delivered by 3rd-party CDNs than a previous active measurement study suggests [72]. We hope researchers can use our dataset to complement findings from active measurements.

While DNS-based top lists [68, 105] are widely used [29, 30, 110], we show that service usage weighted by DNS queries differs greatly from that with traffic traces. Access to our dataset provides a more accurate understanding of service popularity (at least within our network) and enables researchers to evaluate algorithms in a realistic setting. For example, a recent paper used our dataset to simulate real-world user requests and evaluate the performance of its algorithm [51].

Moreover, we show that service usage varies across demographics and that residential traces are different from campus traces, highlighting the distinctive nature of our dataset. As shown in Table 1, most prior studies do not share their datasets. By providing a new data source, we hope to facilitate comparisons between different types of networks and to enhance the community's knowledge of Internet service usage. We also encourage researchers to explore this dataset in other directions we may not have considered.

Regular monitoring is important [Observation 4, 11, 12]. We demonstrate temporal changes in both the serving infrastructure and service usage, and our findings provide researchers with an updated view of residential Internet usage and help them better design systems. For example, when designing a system for traffic prioritization, it could be reasonable to assign cloud applications a higher priority in the early mornings. We will continue our data collection and update our findings, making our dataset a tool to track Internet changes.

8 CONCLUSION

Using traces from a residential network, we offer a novel perspective distinct from existing campus traces. Our perspective provides a detailed understanding of Internet services and service delivery for our user network, revealing notable P2P activity, less traffic delivered by 3rd-party CDNs than we expected, and the use of more distant servers even for services with nearby ones. Our study—as well as our sharing of our dataset as we continue collecting it—can serve as a step towards a more thorough understanding of residential Internet use and support future research.

9 ACKNOWLEDGEMENTS

We would like to thank our shepherd, Francesco Bronzino, and the anonymous reviewers for their valuable feedback. We are grateful for the assistance of Anthony Robert Acosta, Jarred Lee Buford O'Brien, Joel L. Rosenblatt, Thomas Rom, Yosef Gunsburg, and the security, privacy, and networking teams of Columbia University Information Technology (CUIT). We are thankful for the help of Arpit Gupta, Sanjay Chandrasekaren, Mahshid Ghasem, Todd Arnold, and Loqman Salamatian. We also thank Remi Hendriks and the rest of the MAnycast2 team. This work is supported in part by NSF grants OAC-2029295 and CNS-2212479.

REFERENCES

[1] [n. d.]. AbuseIPDb - IP Address Abuse Reports. https://www.abuseipdb.com/.
[2] [n. d.]. AS Rank. https://api.asrank.caida.org/v2/docs.
[3] [n. d.]. Chromium Project. https://github.com/chromium/chromium.
[4] [n. d.]. HTTPDNS. https://www.tencentcloud.com/products/httpdns
[5] [n. d.]. Routeviews Prefix to AS mappings Dataset for IPv4 and IPv6. https://www.caida.org/catalog/datasets/routeviews-prefix2as/.
[6] [n. d.]. SpeedGuide Ports Database. https://www.speedguide.net/ports.php
[7] [n. d.]. tshark. https://www.wireshark.org/docs/man-pages/tshark.html.
[8] [n. d.]. Twitch Broadcast Guidelines. https://help.twitch.tv/s/article/broadcast-guidelines?language=en_US.
[9] 2012. CAIDA Master Acceptable Use Agreement (AUA). https://www.caida.org/about/legal/aua/. Last modified: 2022.
[10] 2017. Measuring Your Web Server Reachability with TCP Ping. https://blog.apnic.net/2017/10/02/measuring-web-server-reachability-tcp-ping.
[11] 2020. BitTorrent. https://wiki.wireshark.org/BitTorrent
[12] 2020. Microsoft Q&A About DNS Cached. https://learn.microsoft.com/en-us/answers/questions/144963/cloudflare-dns-cached.
[13] 2020. Ports Required for Microsoft Teams. https://answers.microsoft.com/en-us/msteams/forum/all/ports-required-for-microsoft-teams/87c608b5-1650-4a84-a15a-9bd7846cb8bb.
[14] 2022. The European Union (EU) General Data Protection Regulation (GDPR). https://www.hrpo.pitt.edu/european-union-eu-general-data-protection-regulation-gdpr.
[15] 2023. 6 Affordable Streaming Services for Students to Binge the Latest TV Shows and Movies. https://www.billboard.com/culture/product-recommendations/best-streaming-deals-for-students-1235351582
[16] 2023. Anycast-Census Data, Nov 2023. https://github.com/ut-dacs/Anycast-Census/blob/main/dataset/nov2023.csv.
[17] 2023. Dag Scrubber. https://ant.isi.edu/software/dag_scrubber/index.html. Current release: 2023.
[18] 2023. Hoiho - Holistic Orthography of Internet Hostname Observations. https://catalog.caida.org/dataset/hoiho.
[19] 2023. Lumen Announces Sale of Select CDN Customer Contracts to Akamai. https://news.lumen.com/2023-10-10-Lumen-announces-sale-of-select-CDN-customer-contracts-to-Akamai.
[20] 2024. Can Netflix handle live event buffering issues ahead of NFL Christmas Day games? https://news.temple.edu/news/2024-12-17/can-netflix-handle-live-event-buffering-issues-ahead-nfl-christmas-day-games.
[21] 2024. Netflix and Most Valuable Promotions' Jake Paul vs Mike Tyson Mega-Event Makes History With Over 108 Million Live Global Viewers. https://about.netflix.com/en/news/jake-paul-vs-mike-tyson-over-108-million-live-global-viewers.
[22] 2024. NFL Christmas Day Games on Netflix Average Over 30 Million Global Viewers. https://about.netflix.com/en/news/nfl-christmas-day-games-on-netflix-average-over-30-million-global-viewers.
[23] 2024. What resolution are you getting for Tyson VS Paul? I'm at 720p. https://www.reddit.com/r/netflix/comments/1gsc19x/what_resolution_are_you_getting_for_tyson_vs_paul/.
[24] 2025. Columbia University Residential Datasets. https://wimnet.github.io/CUResidential/.
[25] 2025. If You Use FaceTime and iMessage Behind a Firewall. https://support.apple.com/en-us/102036
[26] Bernhard Ager, Wolfgang Mühlbauer, Georgios Smaragdakis, and Steve Uhlig. 2011. Web Content Cartography. In *Proceedings of the ACM Internet Measurement Conference (IMC) 2011.*
[27] Akamai. 2023. State of the Internet Reports. https://www.akamai.com/our-thinking/the-state-of-the-internet
[28] Mark Allman. 2020. Putting DNS in Context. In *Proceedings of the ACM Internet Measurement Conference (IMC) 2020.*
[29] Mario Almeida, Alessandro Finamore, Diego Perino, Narseo Vallina-Rodriguez, and Matteo Varvello. 2017. Dissecting DNS Stakeholders in Mobile Networks. In *Proceedings of the ACM Conference on Emerging Networking EXperiments and Technologies (CoNEXT) 2017.*
[30] Johanna Amann, Oliver Gasser, Quirin Scheitle, Lexi Brent, Georg Carle, and Ralph Holz. 2017. Mission Accomplished? HTTPS Security After DigiNotar. In *Proceedings of the ACM Internet Measurement Conference (IMC) 2017.*
[31] Waqar Aqeel, Balakrishnan Chandrasekaran, Anja Feldmann, and Bruce M Maggs. 2020. On landing and internal web pages: The strange case of jekyll and hyde in web performance measurement. In *Proceedings of the ACM Internet Measurement Conference (IMC) 2020.*
[32] Calvin Ardi and Matt Calder. 2023. The Prevalence of Single Sign-On on the Web: Towards the Next Generation of Web Content Measurement. In *Proceedings of the ACM Internet Measurement Conference (IMC) 2023.*
[33] Martin Arlitt and Carey Williamson. 2005. An Analysis of TCP Reset Behaviour on the Internet. *ACM SIGCOMM Computer Communication Review (CCR)* 35, 1 (2005), 37–44.
[34] Todd Arnold, Jia He, Weifan Jiang, Matt Calder, Italo Cunha, Vasileios Giotsas, and Ethan Katz-Bassett. 2020. Cloud Provider Connectivity in the Flat Internet. In *Proceedings of the ACM Internet Measurement Conference (IMC) 2020.*
[35] Pete Bell. 2025. The Rise of Fiber. https://blog.telegeography.com/the-rise-of-fiber.

[36] Roman Beltiukov, Sanjay Chandrasekaran, Arpit Gupta, and Walter Willinger. 2023. Pinot: Programmable Infrastructure for Networking. In *Proceedings of the ACM Applied Networking Research Workshop (ANRW) 2023*.
[37] Ignacio N. Bermudez, Marco Mellia, Maurizio M. Munafo, Ram Keralapura, and Antonio Nucci. 2012. DNS to the Rescue: Discerning Content and Services in a Tangled Web. In *Proceedings of the ACM Internet Measurement Conference (IMC) 2012*.
[38] A. Bhattacharyya. 1946. On a Measure of Divergence between Two Multinomial Populations. *Sankhya: The Indian Journal of Statistics (1933-1960)* 7, 4 (1946), 401–406.
[39] Jeremias Blendin, Fabrice Bendfeldt, Ingmar Poese, Boris Koldehofe, and Oliver Hohlfeld. 2018. Dissecting Apple's Meta-CDN During an iOS Update. In *Proceedings of the ACM Internet Measurement Conference (IMC) 2018*.
[40] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008, 10 (2008), P10008.
[41] Timm Böttger, Ghida Ibrahim, and Ben Vallis. 2020. How the Internet Reacted to COVID-19: A Perspective from Facebook's Edge Network. In *Proceedings of the ACM Internet Measurement Conference (IMC) 2020*.
[42] Matt Calder, Xun Fan, Zi Hu, Ethan Katz-Bassett, John Heidemann, and Ramesh Govindan. 2013. Mapping the Expansion of Google's Serving Infrastructure. In *Proceedings of the ACM Internet Measurement Conference (IMC) 2013*.
[43] Matt Calder, Ashley Flavel, Ethan Katz-Bassett, Ratul Mahajan, and Jitendra Padhye. 2015. Analyzing the Performance of an Anycast CDN. In *Proceedings of the ACM Internet Measurement Conference (IMC) 2015*.
[44] Fangfei Chen, Ramesh K. Sitaraman, and Marcelo Torres. [n. d.]. End-User Mapping: Next Generation Request Routing for Content Delivery. In *Proceedings of the ACM SIGCOMM Conference 2015*.
[45] Xian Chen, Ruofan Jin, Kyoungwon Suh, Bing Wang, and Wei Wei. 2012. Network Performance of Smart Mobile Handhelds in a University Campus WiFi Network. In *Proceedings of the ACM Internet Measurement Conference (IMC) 2012*.
[46] Yi-Ching Chiu, Brandon Schlinker, Abhishek Balaji Radhakrishnan, Ethan Katz-Bassett, and Ramesh Govindan. 2015. Are We One Hop Away from a Better Internet?. In *Proceedings of the ACM Internet Measurement Conference (IMC) 2015*.
[47] David Chou, Tianyin Xu, Kaushik Veeraraghavan, Andrew Newell, Sonia Margulis, Lin Xiao, Pol Mauri Ruiz, Justin Meza, Kiryong Ha, Shruti Padmanabha, et al. 2019. Taiji: Managing Global User Traffic for Large-Scale Internet Services at the Edge. In *Proceedings of the ACM Symposium on Operating Systems Principles (SOSP) 2019*.
[48] Danilo Cicalese, Jordan Augé, Diana Joumblatt, Timur Friedman, and Dario Rossi. 2015. Characterizing IPv4 Anycast Adoption and Deployment. In *Proceedings of the ACM Conference on Emerging Networking Experiments and Technologies (CoNEXT) 2015*.
[49] Danilo Cicalese and Dario Rossi. 2018. A Longitudinal Study of IP Anycast. *ACM SIGCOMM Computer Communication Review (CCR)* 48, 1 (2018), 10–18.
[50] Luca Deri, Maurizio Martinelli, Tomasz Bujlow, and Alfredo Cardigliano. 2014. nDPI: Open-Source High-Speed Deep Packet Inspection. In *Proceedings of the International Wireless Communications and Mobile Computing Conference (IWCMC) 2014*.
[51] Kahlil Dozier, Loqman Salamatian, and Dan Rubenstein. 2024. Analysis of False Negative Rates for Recycling Bloom Filters (Yes, They Happen!). *Proc. ACM Meas. Anal. Comput. Syst. (POMACS)* 8, 2, Article 21 (May 2024), 34 pages.
[52] Ben Du, Massimo Candela, Bradley Huffaker, Alex C Snoeren, and KC Claffy. 2020. RIPE IPmap Active Geolocation: Mechanism and Performance Evaluation. *ACM SIGCOMM Computer Communication Review (CCR)* 50, 2 (2020), 3–10.
[53] Nick Feamster. 2016. Revealing Utilization at Internet Interconnection Points. *Telecommunications Policy Research Conference (TPRC)* (2016).
[54] Anja Feldmann, Oliver Gasser, Franziska Lichtblau, Enric Pujol, Ingmar Poese, Christoph Dietzel, Daniel Wagner, Matthias Wichtlhuber, Juan Tapiador, Narseo Vallina-Rodriguez, et al. 2020. The Lockdown Effect: Implications of the COVID-19 Pandemic on Internet Traffic. In *Proceedings of the ACM Internet Measurement Conference (IMC) 2020*.
[55] Ashley Flavel, Pradeepkumar Mani, David A. Maltz, Nick Holt, Jie Liu, Yingying Chen, and Oleg Surmachev. 2015. FastRoute: A Scalable Load-Aware Anycast Routing Architecture for Modern CDNs. In *Proceedings of the USENIX Conference on Networked Systems Design and Implementation (NSDI) 2015*.
[56] Dean Garfield. 2021. Red Light, Green Light? No to Network Usage Fees. https://about.netflix.com/en/news/red-light-green-light-no-to-network-usage-fees.
[57] Petros Gigis, Matt Calder, Lefteris Manassakis, George Nomikos, Vasileios Kotronis, Xenofontas Dimitropoulos, Ethan Katz-Bassett, and Georgios Smaragdakis. 2021. Seven Years in the Life of Hypergiants' Off-Nets. In *Proceedings of the ACM SIGCOMM Conference 2021*.
[58] Phillipa Gill, Martin Arlitt, Zongpeng Li, and Anirban Mahanti. 2007. YouTube Traffic Characterization: A View from the Edge. In *Proceedings of the ACM Internet Measurement Conference (IMC) 2007*.
[59] Phillipa Gill, Christophe Diot, Lai Yi Ohlsen, Matt Mathis, and Stephen Soltesz. 2022. M-Lab: User Initiated Internet Data for the Research Community. *ACM SIGCOMM Computer Communication Review (CCR)* 52, 1 (2022), 34–37.

[60] Danilo Giordano, Danilo Cicalese, Alessandro Finamore, Marco Mellia, Maurizio M. Munafò, Diana Zeaiter Joumblatt, Dario Rossi, et al. 2016. A First Characterization of Anycast Traffic from Passive Traces. In *Proceedings of the IEEE Traffic Measurement and Analysis Conference (TMA) 2016*.

[61] Sarthak Grover, Mi Seon Park, Srikanth Sundaresan, Sam Burnett, Hyojoon Kim, Bharath Ravi, and Nick Feamster. 2013. Peeking Behind the NAT: An Empirical Study of Home Networks. In *Proceedings of the ACM Internet Measurement Conference (IMC) 2013*.

[62] Craig Gutterman, Katherine Guo, Sarthak Arora, Xiaoyang Wang, Les Wu, Ethan Katz-Bassett, and Gil Zussman. 2019. Requet: Real-time QoE Detection for Encrypted YouTube Traffic. In *Proceedings of the ACM Multimedia Systems Conference (MMSys) 2019*.

[63] Mackenzie Haffey, Martin Arlitt, and Carey Williamson. 2018. Modeling, Analysis, and Characterization of Periodic Traffic on a Campus Edge Network. In *Proceedings of the IEEE International Symposium on Modeling, Analysis, and Simulation on Computer and Telecommunication Systems (MASCOTS) 2018*.

[64] Tristan Henderson, David Kotz, and Ilya Abyzov. 2004. The Changing Usage of a Mature Campus-Wide Wireless Network. In *Proceedings of the International Conference on Mobile Computing and Networking (MobiCom) 2004*.

[65] Paul Hoffman and Patrick McManus. 2018. RFC 8484: DNS Queries over HTTPS (DoH).

[66] Zi Hu, Liang Zhu, John Heidemann, Allison Mankin, Duane Wessels, and Paul Hoffman. 2016. RFC 7858: Specification for DNS over Transport Layer Security (TLS).

[67] Te-Yuan Huang, Ramesh Johari, Nick McKeown, Matthew Trunnell, and Mark Watson. 2014. A Buffer-Based Approach to Rate Adaptation: Evidence from a Large Video Streaming Service. In *Proceedings of the ACM SIGCOMM Conference 2014*.

[68] Dan Hubbard. 2016. Cisco Umbrella 1 Million. https://umbrella.cisco.com/blog/2016/12/14/cisco-umbrella-1-million

[69] Christian Huitema, Sara Dickinson, and Allison Mankin. 2022. RFC 9250: DNS over Dedicated QUIC Connections.

[70] Weifan Jiang, Tao Luo, Thomas Koch, Yunfan Zhang, Ethan Katz-Bassett, and Matt Calder. 2021. Towards Identifying Networks with Internet Clients Using Public Data. In *Proceedings of the ACM Internet Measurement Conference (IMC) 2021*.

[71] Jaeyeon Jung, Emil Sit, Hari Balakrishnan, and Robert Morris. 2001. DNS Performance and the Effectiveness of Caching. In *Proceedings of the ACM Internet Measurement Conference (IMC) 2001*.

[72] Aqsa Kashaf, Vyas Sekar, and Yuvraj Agarwal. 2020. Analyzing Third Party Service Dependencies in Modern Web Services: Have We Learned from the Mirai-Dyn Incident?. In *Proceedings of the ACM Internet Measurement Conference (IMC) 2020*.

[73] Hyojoon Kim and Arpit Gupta. 2019. ONTAS: Flexible and Scalable Online Network Traffic Anonymization System. In *Proceedings of the Workshop on Network Meets AI and ML 2019*.

[74] Thomas Koch, Ethan Katz-Bassett, John Heidemann, Matt Calder, Calvin Ardi, and Ke Li. 2021. Anycast in Context: A Tale of Two Systems. In *Proceedings of the ACM SIGCOMM Conference 2021*.

[75] Thomas Koch, Shuyue Yu, Sharad Agarwal, Ryan Beckett, and Ethan Katz-Bassett. 2023. PAINTER: Ingress Traffic Engineering and Routing for Enterprise Cloud Networks. In *Proceedings of the ACM SIGCOMM Conference 2023*.

[76] Vengatanathan Krishnamoorthi, Niklas Carlsson, Emir Halepovic, and Eric Petajan. 2017. BUFFEST: Predicting buffer conditions and real-time requirements of HTTP(S) adaptive streaming clients. In *Proceedings of the ACM Multimedia Systems Conference (MMSys) 2017*.

[77] Rupa Krishnan, Harsha V. Madhyastha, Sridhar Srinivasan, Sushant Jain, Arvind Krishnamurthy, Thomas Anderson, and Jie Gao. 2009. Moving Beyond End-to-End Path Information to Optimize CDN Performance. In *Proceedings of the ACM Internet Measurement Conference (IMC) 2009*.

[78] Craig Labovitz. 2019. Internet Traffic 2009-2019. NANOG. https://storage.googleapis.com/site-media-prod/documents/20190610_Labovitz_Internet_Traffic_2009-2019_v1.pdf

[79] Craig Labovitz. 2020. Pandemic Impact on Global Internet Traffic. NANOG. https://storage.googleapis.com/site-media-prod/documents/20200601_Labovitz_Effects_Of_Covid-19_v1.pdf

[80] Craig Labovitz, Scott Iekel-Johnson, Danny McPherson, Jon Oberheide, and Farnam Jahanian. 2010. Internet Inter-Domain Traffic. In *Proceedings of the ACM SIGCOMM Conference 2010*.

[81] M. Larsen and F. Gont. 2011. *RFC 6056 - Recommendations for Transport-Protocol Port Randomization*. Technical Report. IETF.

[82] Michel Laterman, Martin Arlitt, and Carey Williamson. 2017. A Campus-Level View of Netflix and Twitch: Characterization and Performance Implications. In *International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS)*.

[83] Zhihao Li, Dave Levin, Neil Spring, and Bobby Bhattacharjee. 2018. Internet Anycast: Performance, Problems and Potential. In *Proceedings of the ACM SIGCOMM Conference 2018*.

[84] Shinan Liu, Paul Schmitt, Francesco Bronzino, and Nick Feamster. 2021. Characterizing Service Provider Response to the COVID-19 Pandemic in the United States. In *Proceedings of the Passive and Active Measurement Conference (PAM)*

2021.

[85] Matthew Luckie. 2010. Scamper: A Scalable and Extensible Packet Prober for Active Measurement of the Internet. In *Proceedings of the ACM Internet Measurement Conference (IMC) 2010.*

[86] Oliver Michel, Satadal Sengupta, Hyojoon Kim, Ravi Netravali, and Jennifer Rexford. 2022. Enabling Passive Measurement of Zoom Performance in Production Networks. In *Proceedings of the ACM Internet Measurement Conference (IMC) 2022.*

[87] Jelena Mirkovic, Yebo Feng, and Jun Li. 2022. Measuring Changes in Regional Network Traffic due to Covid-19 Stay-at-Home Measures. *arXiv preprint arXiv:2203.00742* (2022).

[88] Nitinder Mohan, Lorenzo Corneo, Aleksandr Zavodovski, Suzan Bayhan, Walter Wong, and Jussi Kangasharju. 2020. Pruning Edge Research with Latency Shears. In *Proceedings of the ACM Workshop on Hot Topics in Networks (HotNets) 2020.*

[89] Heather Morton and Erin Louet. 2024. How State Broadband Offices Are Expanding Internet Access. https://ncsl.org/state-legislatures-news/details/how-state-broadband-offices-are-expanding-internet-access

[90] Sandvine. 2023. https://www.sandvine.com/global-internet-phenomena-report-2023

[91] Matthew Sargent and Mark Allman. 2014. Performance Within a Fiber-to-the-Home Network. *ACM SIGCOMM Computer Communication Review (CCR)* 44, 3 (2014), 22–30.

[92] Corey Satten. 2008. Lossless Gigabit Remote Packet Capture With Linux. https://staff.washington.edu/corey/gulp/.

[93] Brandon Schlinker, Hyojeong Kim, Timothy Cui, Ethan Katz-Bassett, Harsha V Madhyastha, Italo Cunha, James Quinn, Saif Hasan, Petr Lapukhov, and Hongyi Zeng. 2017. Engineering Egress with Edge Fabric: Steering Oceans of Content to the World. In *Proceedings of the ACM SIGCOMM Conference 2017.*

[94] Satadal Sengupta, Hyojoon Kim, and Jennifer Rexford. 2022. Continuous In-Network Round-Trip Time Monitoring. In *Proceedings of the ACM SIGCOMM Conference 2022.*

[95] Ranya Sharma, Tarun Mangla, James Saxon, Marc Richardson, Nick Feamster, and Nicole P. Marwell. 2022. Benchmarks or Equity? A New Approach to Measuring Internet Performance. (2022). Available at SSRN: https://ssrn.com/abstract=4179787.

[96] Taveesh Sharma, Tarun Mangla, Arpit Gupta, Junchen Jiang, and Nick Feamster. 2023. Estimating WebRTC Video QoE Metrics Without Using Application Headers. In *Proceedings of the ACM Internet Measurement Conference (IMC) 2023.*

[97] Raffaele Sommese, Leandro Bertholdo, Gautam Akiwate, Mattijs Jonker, Roland van Rijswijk-Deij, Alberto Dainotti, KC Claffy, and Anna Sperotto. 2020. Manycast2: Using Anycast to Measure Anycast. In *Proceedings of the ACM Internet Measurement Conference (IMC) 2020.*

[98] Bruce Spang, Shravya Kunamalla, Renata Teixeira, Te-Yuan Huang, Grenville Armitage, Ramesh Johari, and Nick McKeown. 2023. Sammy: Smoothing Video Traffic to Be a Friendly Internet Neighbor. In *Proceedings of the ACM SIGCOMM Conference 2023.*

[99] Srikanth Sundaresan, Nick Feamster, and Renata Teixeira. 2015. Measuring the Performance of User Traffic in Home Wireless Networks. In *Proceedings of the Passive and Active Measurement Conference (PAM) 2015.*

[100] Martino Trevisan, Danilo Giordano, Idilio Drago, Marco Mellia, and Maurizio Munafo. 2018. Five Years at the Edge: Watching Internet from the ISP Network. In *Proceedings of the ACM Conference on Emerging Networking EXperiments and Technologies (CoNEXT) 2018.*

[101] Alisha Ukani, Ariana Mirian, and Alex C Snoeren. 2021. Locked-in During Lock-down: Undergraduate Life on the Internet in a Pandemic. In *Proceedings of the ACM Internet Measurement Conference (IMC) 2021.*

[102] Kevin Vermeulen, Loqman Salamatian, Sang Hoon Kim, Matt Calder, and Ethan Katz-Bassett. 2023. The Central Problem with Distributed Content: Common CDN Deployments Centralize Traffic in a Risky Way. In *Proceedings of the ACM Workshop on Hot Topics in Networks (HotNets) 2023.*

[103] Li Wenwei, Zhang Dafang, Yang Jinmin, and Xie Gaogang. 2007. On Evaluating the Differences of TCP and ICMP in Network Measurement. *Computer Communications* 30, 2 (2007), 428–439.

[104] Florian Wohlfart, Nikolaos Chatzis, Caglar Dabanoglu, Georg Carle, and Walter Willinger. 2018. Leveraging Interconnections for Performance: The Serving Infrastructure of a Large CDN. In *Proceedings of the ACM SIGCOMM Conference 2018.*

[105] Qinge Xie, Shujun Tang, Xiaofeng Zheng, Qingran Lin, Baojun Liu, Haixin Duan, and Frank Li. 2022. Building an Open, Robust, and Stable Voting-Based Domain Top List. In *USENIX Security Symposium 2022.*

[106] Jun Xu, Jinliang Fan, Mostafa H Ammar, and Sue B Moon. 2002. Prefix-Preserving IP Address Anonymization: Measurement-Based Security Evaluation and a New Cryptography-Based Scheme. In *Proceedings of the IEEE International Conference on Network Protocols (ICNP) 2002.*

[107] Kok-Kiong Yap, Murtaza Motiwala, Jeremy Rahe, Steve Padgett, Matthew Holliman, Gary Baldus, Marcus Hines, Taeeun Kim, Ashok Narayanan, Ankur Jain, et al. 2017. Taking the Edge Off with Espresso: Scale, Reliability, and Programmability for Global Internet Peering. In *Proceedings of the ACM SIGCOMM Conference 2017.*

[108] Bahador Yeganeh, Reza Rejaie, and Walter Willinger. 2017. A View from the Edge: A Stub-AS Perspective of Traffic Localization and Its Implications. In *Proceedings of the IEEE Traffic Measurement and Analysis Conference (TMA) 2017.*

[109] Chenguang Yu, Yang Xu, Bo Liu, and Yong Liu. 2014. "Can You SEE Me Now?" A Measurement Study of Mobile Video Calls. In *Proceedings of the IEEE INFOCOM 2014.*

[110] Fenglu Zhang, Baojun Liu, Eihal Alowaisheq, Jianjun Chen, Chaoyi Lu, Linjian Song, Yong Ma, Ying Liu, Haixin Duan, and Min Yang. 2023. Silence is Not Golden: Disrupting the Load Balancing of Authoritative DNS Servers. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS) 2023.*

A APPENDIX

A.1 Ethics

To avoid collecting personal identifiable information, we followed established practices [73] to build our data collection and anonymization pipeline. Even if the data contains private information in ways that we did not anticipate (*e.g.,* when combined with other data sources), we protect user privacy by sharing it only with researchers who have received their IRB approval or exemption and agreed to our Acceptable Use Policy.

A.1.1 Data collection and anonymization. Prior to any action, our data collection protocol underwent formal review from the Institutional Review Board (IRB) and was declared exempt as it is not human-subjects research, as the humans are not the subjects of the research under the IRB definitions of subjects. Rather, we are interested in the Internet services and serving infrastructure that delivers the traffic to the residential network in aggregate.

To best protect the privacy of our residential users, we follow established practices [73] that were also used by Princeton University [86, 94] and UCSB [36]. We carefully designed the pipeline to anonymize privacy-sensitive fields and discard personally identifiable information (§3.2). We also rotate our anonymization key to prevent users from being identified across collections. The collected data is securely stored. The data collection methodology and pipeline were approved by the security, privacy, and networking teams of Columbia's IT organization.

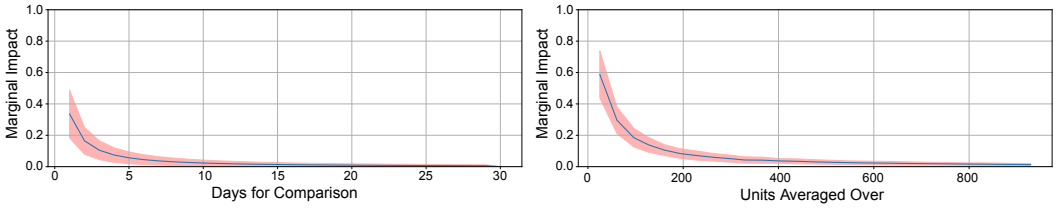
Our collection and anonymization approaches do not keep private information, and our analysis does not attempt to identify a human. We also do not study any network usage below the level of buildings. This work raises no other ethical issues.

A.1.2 Data sharing. We are happy to share the dataset from this paper, as well as data collected in the future, to enable the research of others. Each research team needs to submit a project-specific IRB protocol to their institution's IRB, in which they have to describe the data they need and how they plan to use it. In order to access our data, they need to provide us with their IRB approval or exemption. In addition, we will ask the researchers to agree with our Acceptable Use Policy (which is adapted from CAIDA's Acceptable Use Agreement for traffic traces [9]), including not distributing or deanonymizing data.

A.2 Additional Datasets

As discussed in §3.2, we capture four hours of traffic per day due to limited storage. Between December 2022 and October 2023, we collect traffic during 4-5am, 10-11am, 4-5pm, and 10-11pm. Since November 2024, to cover a broader range of hours, we have shifted the collection hours by 1 hour roughly every 7 days (e.g., after the first shift, the collection hours become 5-6am, 11am-12pm, 5-6pm, 11pm-12am). The four hours are spaced six hours apart, allowing for an average representation of normal user behavior.

Dataset of daily patterns: To analyze usage at different times of a day (Observation 12), we used the traffic to/from the 404 graduate student apartments from November 20, 2023 to January



(a) (b)

Fig. 14. Service popularity converges quickly as data is added from more days (14a) and more units (14b). The blue lines show the changes in normalized median (across random selections) Bhattacharyya distance, and the red shading represents one standard deviation over random selections.

26, 2024. We collected at least 7 days worth of data for each hour of the day. We use the graduate student traces for this analysis to limit confounding factors due to user demographics.

Longitudinal dataset: To study the changes in usage trends over time (Observation 11), we use traffic to/from the 404 graduate student apartments from December 1, 2022 to Jan 26, 2024. The majority of collection occurred during 4-5am/pm, 10-11am/pm.[3]

A.3 Our dataset includes enough days and units

We assess the impact of adding days and units to our dataset, to establish that our main month-long dataset captures the behavior of our users.

To generate Fig. 14a, we randomly order the days of collection and calculate the service usage distributions for the first X days for all X. For example, the service distribution could be 20% Netflix and 80% Prime Video for the first day, and 30% Netflix vs 70% Prime Video for the first two days (if there are only two services). We compute the impact of adding one more day as the Bhattacharyya distance [38] between the service usage distribution for days $1 \ldots X$ and days $1 \ldots X+1$. We repeat this computation over many random orderings of days.

Similarly, to generate Fig. 14b, we randomly order the units rather than the days, calculate the service usage distributions for the first X units, and compute the impact of adding one more unit. We repeat over many random orderings of units.

We use the Bhattacharyya distance here (and elsewhere) to compare two distributions as it offers a reasonable measure of the "closeness" between two distributions. For example, if `unit 1` had service distribution [0.6,0.4,0], `unit 2` had service distribution [0,0.4,0.6], and `unit 3` had distribution [0.333,0.333,0.334], then `unit 3` would be closer to `units 1` and 2 (distance = 0.2) than `unit 1` is to `unit 2` (distance 0.9).

The blue lines in Fig. 14a and Fig. 14b show the changes in normalized median (across random selections) Bhattacharyya distance as the dataset grows (more days or units). The red shading shows one standard deviation over random orderings. Distances converge to zero quickly in both figures, suggesting that our month-long snapshot within our subset of units may be enough to capture the typical use for our users.

A.4 Additional evaluation for methodology

Accuracy in identifying unauthorized devices. We randomly picked 100 unauthorized devices, 100 servers, and 100 P2P users identified by our methodology. We check if the remote hosts' IP addresses were reported as scans, spams, or abuses by abuseIPDB, an online database where network users and system administrators report hacking attempts or malicious behavior [1]. (We did not compare

[3]We miss data for August-October and seven days in April in 2023, due to configuration problems in our device.

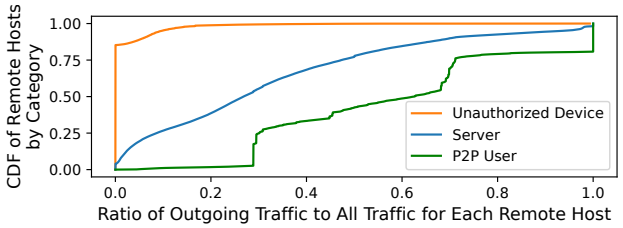


Fig. 15. The fraction of outgoing traffic among total traffic to and from a remote host in each category.

all remote hosts due to the request limits.) 58% of the remote hosts identified as unauthorized devices were reported in abuseIPDB within a year, while only 7.5% of the remote hosts identified as servers or P2P users were detected in the database. The database may include false alarms or fail to receive reports for unsolicited connections caused by misconfigurations, as well as unsolicited (but not malicious) scans by researchers and companies, potentially explaining the differences between the database and our results.

Accuracy in identifying P2P users. Based on the intuition that peer-to-peer connections may result in a larger fraction of outgoing traffic (*i.e.,* from users to remote hosts) than the other types of connections, we performed another evaluation test. In Fig. 15, we plot the fraction of outgoing traffic among total traffic to and from a remote host in each category. An x-value greater than 0.5 means that a remote host receives more traffic than the amount it sends. The figure indicates that the majority of the classified unauthorized devices (85%) only sends traffic to our users but receives no response (*i.e.,* x-value = 0). 77% of the classified servers send more traffic than they receive (*i.e.,* x-value < 0.5). The servers that receive more traffic are used for services such as Google Drive and Apple iCloud. Yet, 57% of the classified P2P users receive more traffic from our users than they send (*i.e.,* x-value > 0.5). The observation aligns with our expectation for all three categories and serves as a supporting evidence for our classification.

Received January 2025; revised April 2025; accepted April 2025