What’s in the Dataset?
Unboxing the APNIC per AS User Population Dataset

Loqman Salamatian
Columbia University
New York, United States

Calvin Ardi
USC/ISI
Marina del Rey, United States

Vasileios Giotsas
Cloudflare
Lancaster, United Kingdom

Matt Calder
Columbia University
New York, United States

Ethan Katz-Bassett
Columbia University
New York, United States

Todd Arnold[*]
Army Cyber Institute
West Point, United States

ABSTRACT

The research measurement community needs methods and datasets to identify user concentrations and to accurately weight ASes against each other for analyzing measurements’ coverage. However, academic researchers traditionally lack visibility into how many users are in each network or how much traffic flows to each network and so often fall back on treating all IP addresses or networks equally. As an alternative, some recent studies have used the APNIC per AS Population Estimates dataset, but it is unvalidated and its methodology is not fully public.

In this work, we validate its use as a fairly reliable user population indicator. Our approach includes a detailed comparative analysis using a global CDN dataset, providing concrete evidence of the APNIC dataset’s accuracy. We find that the APNIC per-AS user estimates closely align with the Content Delivery Network (CDN) per-AS user estimates in 51.2% of countries and correctly identify the largest networks in 93.9% of cases. When we investigate the agreement with CDN traffic volume, the APNIC dataset closely aligns in 36.5% of countries, increasing to 91.0% when focusing only on larger networks. We also evaluate the limitations of the APNIC dataset, particularly its inability to accurately identify user populations for ASes in certain countries. To address this, we introduce new methods to improve its usability by focusing on the statistical representativeness of the underlying data collection process and ensuring consistency across several public datasets.

CCS CONCEPTS

• **Networks → Network measurement**.

KEYWORDS

Datasets, APNIC User Estimates, Traffic Volume

ACM Reference Format:
Loqman Salamatian, Calvin Ardi, Vasileios Giotsas, Matt Calder, Ethan Katz-Bassett, and Todd Arnold. 2024. What’s in the Dataset? Unboxing the APNIC per AS User Population Dataset. In *Proceedings of the 2024 ACM Internet Measurement Conference (IMC ’24), November 4–6, 2024, Madrid, Spain.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3646547.3688411

1 INTRODUCTION

A major challenge for Internet measurement research is obtaining comprehensive and representative datasets. Given this challenge, public datasets are a huge boon to the research community, with RouteViews [61] and AS Rank [14] being two laudable examples.

The lack of a publicly available dataset to identify traffic volumes and networks hosting users is a significant barrier to Internet research [44, 69], leading to a recent call for the Internet research community to estimate relative user activity [46]. A publicly accessible traffic dataset would allow researchers across disciplines to model traffic patterns, study network performance, identify which networks host users, accurately weight Autonomous Systems (ASes) against one another based on user concentrations, or represent accurate traffic volume [69]. Additionally, this understanding would enable policymakers to make informed, data-driven decisions about security and Internet access regulations.

To fill this gap, researchers have recently begun using the Asia-Pacific Network Information Centre (APNIC) per AS User Population dataset [5] (henceforth referred to as the *APNIC dataset*), which estimates the number of users residing within an AS from Google ads (see Section 3.2 for more details). Recent work has relied on the APNIC dataset to determine what percentage of users their measurements represent [7–9, 15, 30, 31, 33, 46, 47, 49–51, 68, 70, 78, 80, 81], to validate techniques in Internet client activity identification [44], or to provide a public service showcasing Internet traffic trends [21]. However, the APNIC dataset has not been extensively validated (at least publicly), and so it is unclear how much it or results that rely on it should be trusted or used.

Contributions. To evaluate the APNIC dataset’s relevancy for weighting ASes in academic research, we compare it to other datasets, some public and some proprietary, that we use to weight ASes by users or user activity in various ways. We start by describing our various datasets and their inherent biases that may influence how user populations are calculated (§3). To validate the APNIC dataset, we begin by comparing aspects of the APNIC dataset with other datasets (§4). A consensus between the APNIC dataset and other

[*] The views expressed herein are those of the authors and do not reflect the position of the US Military Academy, Department of the Army, or Department of Defense.

IMC ’24, November 4–6, 2024, Madrid, Spain

ACM ISBN 979-8-4007-0592-2/24/11
https://doi.org/10.1145/3646547.3688411

representative datasets would reinforce our confidence that it provides a meaningful and publicly available representation of AS user populations. We then compare three specific metrics between the APNIC dataset and proprietary datasets from a large Content Delivery Network (CDN) (AnonCDN): (i) the set of ASes that host users (§4.2), (ii) the most populated ASes in a given country (§4.3), and (iii) the fraction of users and traffic volume associated with each AS in the country (§4.3). In particular, we show that the APNIC dataset is *consistent* across the vast majority of countries for user estimates and Internet traffic volume by quantifying a strong correlation between the APNIC and AnonCDN datasets. Because we cannot share the AnonCDN dataset, we also examine publicly available datasets to see if the APNIC dataset is consistent with them and to help understand the differences (§5.2) and take a data-driven approach to improve traffic volume estimates using IXP capacities (§5.3).

Although the APNIC dataset is mostly consistent, it is important to understand where it may have inaccuracies. We investigate whether there are internal indicators within the APNIC dataset that could be used to predict cases where there are mismatches. Our findings suggest that, when the number of samples (ad impressions) is disproportionately low relative to the predicted number of users, the likelihood of inaccurate results increases significantly (§5.1.1). We also examine the temporal stability of the APNIC dataset's user populations and show that instances of instability in the estimates often signal a lack of reliability in the data generation process itself (§5.1.2). We synthesize these insights into straightforward checks, compiled into an artifact[1], which researchers can use to determine when the APNIC dataset can be reliably employed.

Finally, using our newly validated APNIC dataset, we examine how access networks have consolidated over the past decade worldwide, highlighting an interesting trend of access network consolidation in certain parts of the world (§6). Finally, while we note the irony of being unable to publicly share our proprietary AnonCDN dataset, our validation and analysis code is publicly available.[1]

2 GOALS

The APNIC dataset is actively used as ground truth. Thus, trust and validation of the dataset will benefit the entire community. We also seek to determine under what conditions the dataset should or should not be used and make recommendations to fellow researchers. In "validating", we aim to assess whether the APNIC dataset is a reliable resource for several key tasks:

Does the APNIC dataset identify ASes hosting users? (§4.1 and §4.2) Being able to accurately assess which ASes host users, is a challenge within the community [44], meaning that researchers cannot accurately determine an experiment's impact or a measurement's representativeness from a user perspective.

Does the APNIC dataset accurately estimate the number of users per AS? (§4.1 and §4.3) This is particularly important, as the dataset's primary goal and previous usage hinges on its ability to provide reliable per AS user population metrics [7–9, 15, 21, 30, 31, 33, 44, 46, 47, 49–51, 68, 70, 78, 80, 81]. We explore how well the APNIC dataset's estimates align with other data sources that capture networks' user populations, while also addressing the coverage limitations of these sources.

Table 1: Summary of Datasets

Name	Dates	Data
APNIC	2013-11-01 to 2024-04-21	ASN, samples, user estimates
AnonCDN	2023-07-20, 2023-10-19	HTTP requests
IXP	2023-07-20, 2023-10-19	ASN, network link capacities
M-Lab	2024-01-01, 2024-06-01	ASN, number of speed tests
Broadband	2024-03-01, 2024-03-31	ASN, number of subscribers

Can the APNIC dataset be used to project relative traffic volume per AS? (§4.3) A primary metric for resource allocation and traffic engineering is traffic volume; Cloud/Content Providers/CDNs have ground truth and can accurately weight ASes accordingly, but academic researchers do not have such insights to determine the most significant ASes. Although the APNIC dataset is not designed to reflect traffic, we explore its usefulness in estimating traffic volume per country and AS, as traffic volume likely correlates with the number of users. While a perfect alignment with actual traffic volumes is not expected, our goal is to see if the APNIC dataset can effectively pinpoint the major contributors of traffic volume.

Are there methods to assess and improve the APNIC dataset's accuracy? (§5.1 and §5.2) We ultimately would like to develop a toolkit for researchers to support the utilization of the APNIC dataset in their studies. Our objective is to provide clear guidelines for interpreting the numbers the dataset provides and information about whether and when the dataset can be trusted.

How do we plan to achieve this? By answering these questions, we believe we can enhance the APNIC dataset's usability and make it a staple dataset for measurement studies and research. To do so, we compare the APNIC dataset with several other data sources, including public datasets on Broadband Subscribers (§4.1), M-Lab Speed Tests (§5.2), IXP Fabric Capacity (§5.3), and proprietary AnonCDN `User-Agent` and Traffic Volume data (§4.2, §4.3). Each of these datasets offers different insights into user counts and traffic volume, but they are not without their own limitations, which we discuss in detail (§3). By cross-referencing the APNIC dataset with these alternative datasets, we aim to provide a well-rounded evaluation of its strengths, weaknesses, and areas for improvement.

3 DATASETS AND THEIR BIASES

We next describe the datasets that we analyze and compare in this paper, and their biases: the APNIC dataset, the Broadband Subscriber dataset, the proprietary HTTP request logs from AnonCDN (AnonCDN dataset), the M-Lab Network Diagnostic Tool (NDT) dataset (M-Lab dataset), and the interdomain link capacities at IXPs (IXP dataset). Table 1 summarizes the datasets and their properties.

3.1 Combining Orgs to Compare Datasets

To minimize discrepancies in assigning user populations to specific ASes, we aggregate ASes at the organizational (i.e., sibling) level within each country [13]. Specifically, we combine the relevant ⟨`country, AS`⟩ pairs in each dataset to produce ⟨`country, org`⟩ pairs. This approach allows for a more straightforward and consistent comparison across all datasets.

[1] https://github.com/Burdantes/unboxing_apnic

Table 2: Top 5 ⟨country, AS⟩ in Est. User Population. Dataset: APNIC, 2024-04-21, Window = 60 days.

Country	AS	Users ($\times 10^6$)	% of		Samples ($\times 10^6$)
			Country	Internet	
IN	55836	277.97	46.7	6.61	83.79
CN	4134	265.92	32.8	6.32	29.33
IN	45609	147.06	24.7	3.50	44.33
CN	4837	127.92	15.8	3.04	14.11
CN	9808	123.80	15.3	2.94	13.65

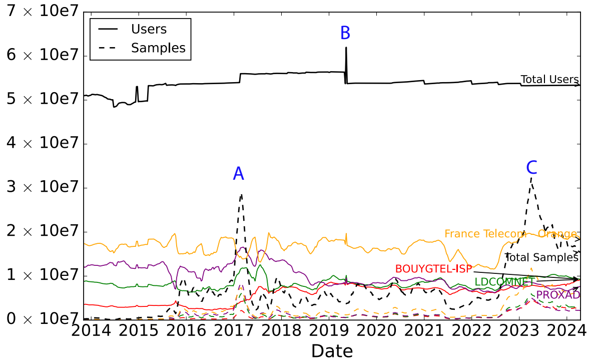


Figure 1: Estimated Users (solid) and Samples (dashed) over time for major ISPs (different colors) in France. While the data is relatively stable, there are several unexplained spikes, labeled A, B, and C. Dataset: APNIC, 2013–2024.

3.2 APNIC per AS Population Estimates

The APNIC dataset [5] is a report, generated daily since 2013-11-01, that provides global estimates on the number of users that ASes host at the ⟨country, AS⟩ granularity over a moving window of 60 days.[2] The dataset includes the following columns: 'Rank', 'AS', 'AS Name', 'CC' (ISO 3166-formatted country code), 'Estimated Users', '% of Country', '% of Internet', and 'Samples'. Table 2 shows partial data of the five most populated ASes on 2024-04-21.

APNIC estimates the per AS user numbers by normalizing the impression count ('Samples') of non-targeted online advertisements via Google Ads with the ITU-T's estimates of Internet users per country [40, 41]. Each sample corresponds to an IP address collected by the ad, and the IP is geolocated using MaxMind with proprietary adjustments. As a result, an AS may correspond with multiple countries—early reports assumed a 1:1 country to AS mapping.

Our interest is understanding whether the number of Samples and Estimated Users correlates with observations in other datasets (§4). Prior work used the Estimated Users at face value. We believe we are the first to publicly take a closer look at the underlying samples, their use in deriving estimated users, and their efficacy.

Biases. There are two biases or skews in the APNIC dataset which may result in inaccurate user estimates: non-uniform ad placement across different countries and the accuracy of ITU-T's Internet users estimates used to normalize the APNIC dataset's ad impressions.

Non-uniform ad placement across different services and countries potentially limits ad impression counts (Samples) as an effective indicator of user population within an AS. The ads the dataset uses for its estimations are served using Google Ads and can be displayed across Google's diverse ecosystem (i.e., search, YouTube, Gmail, etc.) and the Google Display Network, which claims to have a reach of 35×10^6 third-party websites and apps [36]. These ads may not accurately represent user populations in countries where Google is not the dominant search engine, its other services are banned, or third-party sites do not commonly use Google Ads. In general, it is difficult to determine the relationship between ad impressions and the local popularity of the website or Google service in that specific region and its effect on the resulting 'Samples'. In some cases, we can correlate a significant change in Samples with an event: for example, Google pausing all ads in Russia (§4.4).

The second potential bias to AS user estimates is APNIC's use of ITU-T's Internet user estimates to map the Samples to Estimated Users. We observe significant fluctuations in the APNIC dataset's user estimates for some countries. For example, Figure 1 shows the Estimated Users (solid lines) and Samples (dashed) over time for the top 5 Internet Service Providers (ISPs) in France, with several unexplained periods of instability (labeled A, B, and C). A significant change in B on 2019-05-13 may be attributed to fluctuations in the ITU-T's Internet user estimate for France on that week. Specifically, the total number of Internet users reported was 6 million higher than any other week between 2014 and 2024.

3.3 Broadband Subscribers

We gathered a Broadband Subscriber dataset by examining user numbers according to various broadband subscriber surveys and official reports. In some countries, the number of subscribed users is publicly available due to mandatory disclosure requirements [6, 22, 26, 62, 63]. In these cases, we manually compiled the official subscription numbers for access networks from these reports. Where such high-quality datasets were not available, we searched for surveys estimating the number of users per ISP. We manually collected these datasets by browsing different websites for 20 countries across 3 continents [74–77]. We normalized these numbers and projected them as percentages of users, enabling comparison with the APNIC dataset. While this information aligns closely with APNIC's goal of estimating users per AS, it focuses solely on access networks serving end-users. In contrast, the APNIC dataset also considers other types of networks, such as enterprise networks.

Biases. Surveys may be biased toward specific populations who responded to them. In almost all cases where we rely on surveys, there were no specific studies of coverage or representativity, which might lead to an incorrect representation of user distribution across the country. However, we note that all the surveys had enough measurements to recover the underlying distribution with high confidence, assuming perfectly random responses (i.e., the statistical power was sufficiently high for the phenomenon they were studying). Additionally, the number of subscribers in official reports does not precisely map to users, as a subscriber can represent a whole family versus a single user. More generally, collecting broadband data is labor-intensive and difficult to gather at scale. As a result, the dataset is biased toward locations where data collection

[2] The daily values for a specific ASN (Bouygues Telecom for example) can be accessed on the APNIC website at https://stats.labs.apnic.net/ipv4/AS5410?a=5410&c=FR&x=1&s=1&p=1&w=200.

was easier (i.e., countries where information is readily accessible in English or through search engines).

3.4 AnonCDN's HTTP Request Logs

We derive our study's foundational dataset from a major CDN's HTTP request logs. The CDN—AnonCDN—has more than 300 Points of Presence (PoPs) in more than 120 countries, and is estimated to handle a significant portion of global Internet traffic and global websites [56]. The logs are from multiple days (2023-07-20, 2023-10-19, 2024-04-01–2, 2024-05-02–03, and 2024-08-09–12) to capture different times of year and days of the week. AnonCDN logs requests uniformly, randomly sampling 1% of requests received at every server in every PoP. A higher sampling rate is challenging, even for large content providers; prior work found that 1% provides a sufficiently large number of measurements to be representative [71, 72]. In Appendix C, we show that the samples collected on different days in 2024 are consistent.

AnonCDN's request logs contain the following data: client IP addresses, client browser `User-Agent` strings, number of inbound HTTP requests, and outbound network traffic volume (bytes). They derive the client IP's ASN using BGP feeds and perform geolocation using a proprietary, internal tool. For our traffic volume comparisons, we use the outbound traffic volume statistics and are provided total bytes per ⟨`country, org`⟩ pair.

For our user estimate comparisons, we use unique `User-Agent` counts per ⟨`country, org`⟩ pair, as `User-Agent` strings have been shown to be a valid proxy to distinguish multiple hosts per IP address [17, 54, 67] and sufficiently unique to identify individual users across sites and applications [2, 29]. Unlike traffic volume, we do not have the total numbers per ⟨`country, org`⟩ pair as they are considered sensitive. Rather, we are provided with the percentages for each ⟨`country, org`⟩ pair within their country.

Biases. The primary bias we consider is that traffic to AnonCDN is skewed by websites that use it. Specifically, AnonCDN offers robust Distributed Denial of Service (DDoS) defenses and anti-bot detection. Thus, the websites using AnonCDN are more likely to be concerned about cyberattacks or malicious traffic, and that might result in a disproportionate amount of bot or malware traffic.

To mitigate this bias, the dataset provided was filtered for requests that are likely to originate from a human user: each request is labeled by a proprietary bot detection algorithm with a score between 1 (very likely bot) and 99 (very likely human) [28] and we consider scores > 50 to be human-originated.

3.5 M-Lab's Speed Tests

The M-Lab dataset collects speed test data from users across the Internet who voluntarily run tests on their browsers to measure their download and upload speeds[32]. This data is available for analysis and was used to examine trends in network performance across regions and ISPs [20, 43, 53]. We compiled this dataset by focusing on the publicly available test measurements across March 2024. We normalized the number of speed tests at the country granularity to allow for comparison with other datasets, such as the APNIC dataset, which estimates user distribution per ⟨`country, AS`⟩.

While M-Lab's speed test data provides valuable insights into real-world broadband performance, it primarily collects data from users who manually run the test which can lead to data biases.

Biases. Since users voluntarily initiate the tests, the dataset may over-represent users who are more technically-savy or encouraged to check their broadband performance, possibly leading to a skewed representation of the general population's experience. Furthermore, the timing of the tests can introduce bias; users may be more likely to run speed tests during periods of poor performance, which could distort our view of the ASes' populations. Another source of bias is geographical coverage. M-Lab data may be less representative in countries where the M-Lab browser extension is unavailable.

3.6 IXP Peering Capacity

The IXP dataset is an aggregation of ⟨`AS, Capacity`⟩ pairs, where for each AS, capacity is the sum of all port capacities (bit/s) across all IXPs reported in PeeringDB on 2023-07-20 and 2024-08-19 [65].

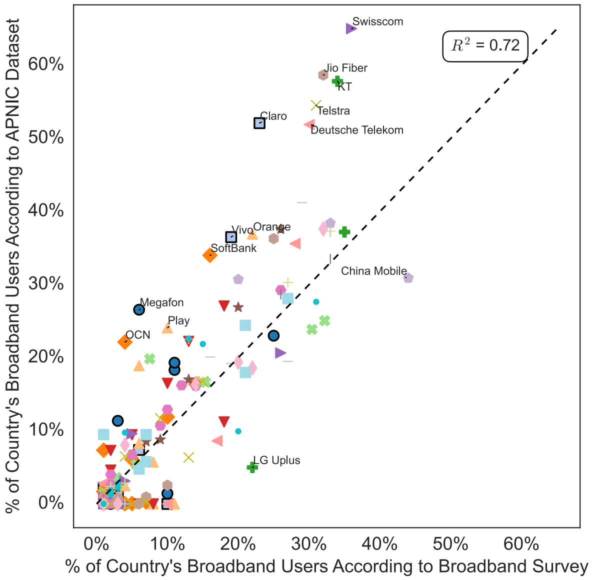
Biases. The primary limitation is that this dataset is incomplete [3]. Prior work found that while PeeringDB data is accurate, it does not contain all of an AS's interconnections [52]. For example, while many hypergiants have an extensive presence across IXPs [11], there are private interconnections between cloud providers, CDNs, and large access networks (ISPs) that are not reflected in PeeringDB, and these are known to carry the vast majority of the traffic [72]. Additionally, the popularity of off-network caches hosted closer to the end-user might have affected the distribution of interdomain capacities [30, 80], which would not be reflected in PeeringDB.

4 VALIDATING POPULATION ESTIMATES

Given the APNIC dataset's frequent usage (§1), we first compare it with the Broadband Subscriber dataset, which directly identifies access networks and their market shares (§4.1). We then compare the APNIC dataset against the AnonCDN datasets to determine which metrics the APNIC dataset is a good proxy for. We first ask whether the APNIC and AnonCDN datasets agree on what constitutes an eyeball network, or an org that hosts users, in each country (§4.2) using AnonCDN's `User-Agent` data. We next look as to whether both datasets agree on the most populated and traffic-heavy orgs at the country-level (§4.3), and conclude this section with an examination on specific outlier organization (§4.4). For all of the examinations, we select the APNIC dataset for the day(s) that align with the AnonCDN dataset.

4.1 Do the APNIC and Broadband Subscriber Datasets Agree on the Number of Users?

We start by validating the APNIC dataset by comparing it to the Broadband Subscriber dataset, which provides a snapshot of different access networks' user numbers based on surveys and official reports. The Broadband Subscriber dataset only covers a limited number of countries and does not include other networks, but we expect it to be very accurate for the countries and networks that it includes. Because end-user-facing businesses often operate under different names than their parent organizations or AS names, we manually matched each company to its corresponding organization and associated ASes. As the APNIC's user population estimates are



	Country | % of Country’s APNIC Estimated Users in Broadband Survey Orgs | R² Comparison of Country Data to 1:1 Model Fit		
●●●	Russia	7.0%	-0.62
■■■	Brazil	29.5%	-1.77
◆◆◆	Japan	55.2%	-2.35
▲▲▲	Poland	60.6%	-1.96
✚✚✚	South Korea	62.3%	-7.23
✖✖✖	USA	64.2%	0.81
▼▼▼	South Africa	68.8%	0.06
◀◀◀	Germany	69.7%	0.13
▶▶▶	Switzerland	77.3%	0.38
⬟⬟⬟	China	78.0%	-0.09
★★★	Sweden	80.8%	0.63
⬢⬢⬢	India	81.6%	0.08
⬣⬣⬣	Canada	82.5%	0.91
♦♦♦	Mexico	83.7%	0.91
| | |	Finland	87.7%	0.96
— — —	France	88.3%	0.90
×××	Australia	88.4%	0.06
+++	Austria	89.8%	0.89
• • •	Italy	94.4%	0.69
■■■	UK	95.4%	0.82

Figure 2: Comparative analysis of `User Estimates` percentages between the Broadband Subscriber and APNIC datasets across 20 countries. Different markers and colors represent the countries. The two datasets generally agree quite closely. The figure labels the organizations where the datasets disagree the most. We highlight, by placing a red rectangle around them, countries with an R^2 is negative between the Broadband Subscriber and APNIC datasets. Additionally, we thicken the borders of organizations where the country’s broadband networks account for less than 50% of the total user estimates according to the APNIC dataset.

not a perfect match for the Broadband Subscription numbers—since the APNIC dataset includes other types of networks like enterprise networks— we renormalize the APNIC data such that it sums to 1 on the subset of organizations that the Broadband Dataset covers. After the normalization, we expect the APNIC and the Broadband Subscriber datasets to be closely matched.

In Figure 2, we compare the user percentages between the Broadband Subscriber and the APNIC datasets across 20 countries. Each point on the scatter plot represents a ⟨`country, org`⟩, with the x-axis showing the percentage of the country’s users hosted by that organization according to the Broadband Subscriber and the y-axis according to the renormalized APNIC dataset. The R^2 fit with the perfect alignment line, where every number from the Broadband Subscriber Dataset is equal to the Broadband Survey, is 0.72, indicating a strong agreement on average between the two datasets. Different markers and colors distinguish between countries, and we highlight the total users in each country, covered by the Broadband Subscriber dataset, according to APNIC. Russia and Brazil have significantly fewer users covered by their country’s broadband in APNIC compared to other countries, which reveals a disagreement between the datasets in terms of the networks that are hosting most of the users. We find strong agreement for over 14 countries, as is demonstrated by the high concentration of points near the diagonal and the corresponding high R^2 fit.

There are significant outliers in the data, with Telstra (Australia), KT (Korea), Swisscom (Switzerland), Jio Fiber (India), Deutsche Telekom (Germany), Claro (Brazil), and Orange (Poland) overrepresented in the APNIC dataset. Notably, all these companies, except KT, are also major mobile carriers in their countries, which may explain the discrepancy, as APNIC includes mobile users, unlike the Broadband Subscriber dataset. APNIC User Estimates in South Korea diverge from the Broadband Subscriber dataset, with KT being overrepresented and LG Uplus underrepresented. According to KT’s latest annual report, KT holds 28.5% of the market, behind SKT with 48.4%, and ahead of LG Uplus, which holds 23.1%. This discrepancy suggests that APNIC’s overestimation of KT’s share may not align with actual market proportions[48]. Similarly, the APNIC dataset diverges from the Broadband Subscriber data in Japan, where Softbank appears significantly larger than NTT Docomo. This observation contrasts with NTT Docomo’s dominance in the Japanese mobile network market [60].

4.2 Do APNIC and AnonCDN Datasets Agree on Orgs that Host Users and Carry Traffic?

Org comparison. Figure 3 (top bar) shows the raw number of ⟨`country, org`⟩ pairs identified in only the AnonCDN (blue) and APNIC (purple) datasets, along with the overlap between the two (green). The APNIC dataset identifies 40% of the ⟨`country, org`⟩ pairs observed by the AnonCDN dataset.

While we do not know the full details of how the APNIC dataset is created, we can postulate a few reasons why APNIC may not identify an org in their dataset. The APNIC dataset uses Google Ads to extrapolate user estimates, which may not be accessible in a country or all parts of it (§3.2). Additionally, a ⟨`country, org`⟩ requires a minimum number of samples to be included, which our empirical observation puts at > 120 samples for inclusion. Interestingly, the same number of samples can correspond to a varying number of users(see §5 for issues with this approach).

User population estimates. Despite overlapping with only 40% of the AnonCDN dataset’s ⟨`country, org`⟩ pairs, we will show that

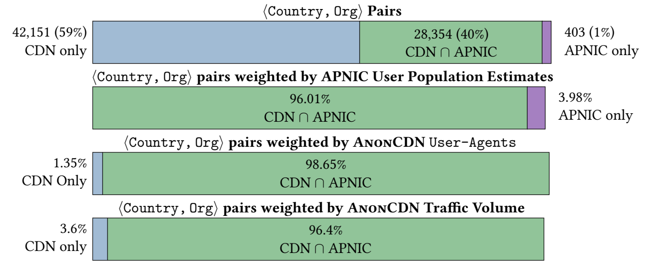


Figure 3: Number (percentage) of ⟨country, org⟩ pairs (top) that are found in only the AnonCDN dataset (blue), in only the APNIC dataset (purple), and in both (green). We then show the ⟨country, org⟩ pairs weighted by the APNIC user population (second bar down), AnonCDN User-Agents (third bar down), and percentage of AnonCDN traffic volume (bottom bar). Even though the datasets only see 40% ⟨country, org⟩ pairs in common, those common pairs cover >96% of the user estimates, User-Agents, and traffic volume.

the missing ⟨country, org⟩ pairs have little impact on the APNIC dataset's ability to estimate user populations, its primary purpose.

We first weight the overlapping ⟨country, org⟩ pairs by the APNIC dataset's user estimates. The ⟨country, org⟩ weighting by user estimates is seen in Figure 3's second bar from the top (overlap in green, APNIC only in purple). The <Country, Org> pairs seen in both datasets account for 96.01% of Internet users according to the APNIC dataset. The population estimates for the vast majority of ⟨country, org⟩ pairs in the APNIC only category are so small—$< 0.01\%$ of their respective country's total users—that they may be missed by the AnonCDN dataset's statistical sampling. The small number of remaining ⟨country, org⟩ pairs only identified in the APNIC dataset are almost entirely from countries have a low Freedom House index [39], such as Yemen, Russia, or Thailand.

User-Agents. We will now demonstrate that the CDN only organizations represent a small fraction of AnonCDN users and traffic, similar to the APNIC only organizations being a small portion of user estimates. This ⟨country, org⟩ weighting is seen in Figure 3's third bar down (overlap in green, AnonCDN only in blue). Similar to how the 40% of ⟨country, org⟩ pairs seen in both datasets account for the vast majority of Internet users as estimated by APNIC, they include 98.65% of the User-Agent counts seen by AnonCDN.

There are a handful of countries in the AnonCDN dataset that do not appear in the APNIC dataset (see Appendix B for the list, countries with 0.0%), mostly small island nations that World Bank population data may bin under another country, so they are not included in the AnonCDN only category. There are two notable exceptions that appear in the AnonCDN dataset and not the APNIC dataset: AnonCDN also classifies Tor exit nodes separately using country code T1 [24] and the Democratic People's Republic of Korea (North Korea) which is likely due to Google banning ads there [37].

Traffic volume. Since Cloud/Content Providers/CDNs use traffic volume as a key metric, a publicly available traffic volume dataset would be valuable to the community [44]. We next focus on the observed ⟨country, org⟩ pairs when weighted by traffic volume. The Figure 3's bottom graph (AnonCDN only in blue, overlap in green) shows the weighting by traffic volume. The graph shows that even though the datasets only agree on 40% of ⟨country, org⟩ pairs, the overlapping ⟨country, org⟩ pairs are responsible for 96.4% of the total traffic volume. According to our AnonCDN dataset, the vast majority of ⟨country, org⟩ pairs missed by the APNIC dataset have far less than 1% of a country's overall volume–with most, 41,428, accounting for $< 0.1\%$ and 34,071 accounting for $< 0.01\%$.

Table 3: The top (left) and bottom (right) 20 countries according the APNIC and AnonCDN dataset's overlapping ⟨country, org⟩ pairs (§4.2), weighted by AnonCDN traffic volume data. We performed a similar analysis for User-Agents (§4.2) and found nearly identical results. The overlapping ⟨country, org⟩ pairs within each country are responsible for over 95% traffic volume at the country level (% Vol), and only 5 have less than 90%. The bottom 20 does not include the countries which have 0%. The complete list is in Appendix B.

Count	Country	% Vol	Count	Country	% Vol
1	Uruguay	100.00	215	Ecuador	98.53
2	Norfolk Island	100.00	216	Kiribati	98.44
3	Comoros	100.00	217	Senegal	97.46
4	Costa Rica	100.00	218	United States	97.39
5	Algeria	99.99	219	Eritrea	97.32
6	Bolivia	99.99	220	Armenia	97.31
7	Tunisia	99.99	221	Vatican City	96.50
8	Togo	99.99	222	Monaco	95.82
9	Oman	99.99	223	Brazil	93.73
10	Burundi	99.99	224	Vanuatu	93.54
11	Chile	99.98	225	Palestine, State of	93.25
12	Macao	99.98	226	Nauru	93.20
13	Uzbekistan	99.98	227	Austria	92.42
14	American Samoa	99.98	228	Russian Fed.	92.35
15	Guinea	99.98	229	Seychelles	91.24
16	Cabo Verde	99.98	230	French Guiana	90.59
17	Mali	99.98	231	Liechtenstein	89.86
18	Guyana	99.98	232	Turkmenistan	88.65
19	Haiti	99.98	233	Saint Barthélemy	85.76
20	Jordan	99.98	234	Tuvalu	79.28

While Figure 3 provides an aggregated view, we further analyze the data to examine whether or not the overlapping ⟨country, org⟩ pairs, when weighted by traffic volume, still achieve a high percentage of traffic *within* a given country.

Virtually every country achieves close to 100% traffic volume using only the overlapping ⟨country, org⟩ pairs. Table 3 shows that the top 20 countries (left side of the table) achieve >= 99.98% traffic volume. The full results, including countries for which neither datset has measurements, are included in Appendix B.

4.3 Do APNIC and AnonCDN Datasets Agree on the Largest Organizations in Each Country?

Our next study explores the consistency across two axes, between the APNIC dataset and AnonCDN (i) User-Agents and (ii) traffic volume datasets for every ⟨country, org⟩ as reported.

Methodology. To assess how well the APNIC and AnonCDN datasets align in identifying the most significant networks within each country and maintaining a consistent ranking for (i) org user populations and (ii) traffic volume, we compare the APNIC dataset against (i) AnonCDN's User-Agent data and (ii) the traffic volume.

We examine three metrics of agreements: (1) Pearson correlation, (2) the coefficient of a linear regression trained on the APNIC data,

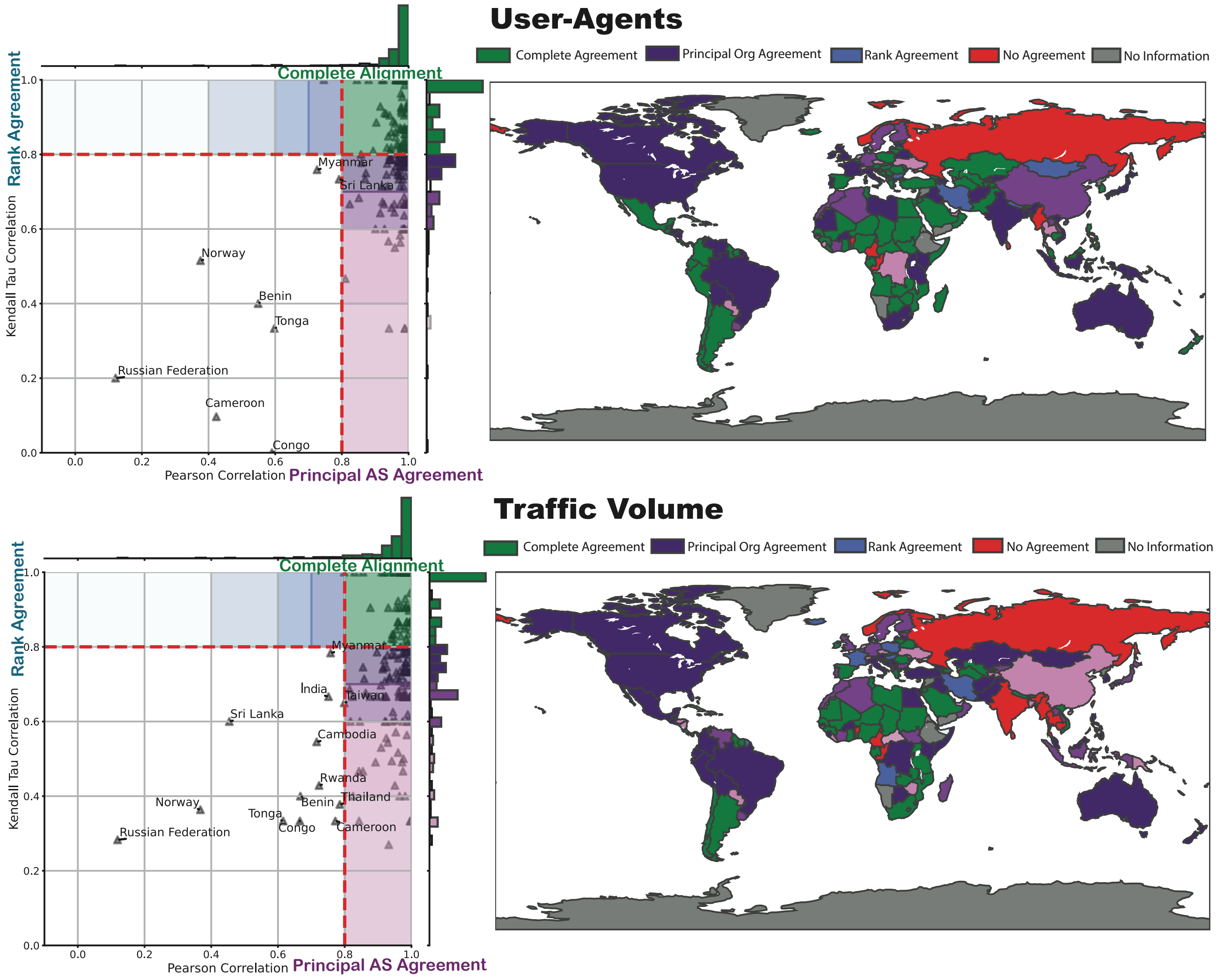


Figure 4: Comparison of Pearson vs. Kendall-Tau correlations between APNIC user estimates and `User-Agent` counts/traffic volumes from AnonCDN. The top figure shows `User-Agents`, while the bottom figure shows traffic volume. Countries are categorized based on their agreement level in each case. In the `User-Agents` comparison, most countries exhibit high agreement, particularly in North and South America, Europe, and Africa. The most significant outliers are found in the African continent showing notably low Kendall-Tau correlations. For traffic volume, APNIC and AnonCDN datasets closely align in most regions, with the largest discrepancies occurring in South and South Eastern Asia.

and (3) Kendall-Tau correlation. The Pearson correlation focuses on the degree to which (i) the `User-Agent` and (ii) traffic volume are linearly correlated with the APNIC dataset. Because the distribution of `User-Agents` and traffic volume is often dominated by a few large networks, discrepancies in smaller networks have minimal impact on the Pearson correlation, with the largest values having the most influence on the overall result. This means that the metric focuses on the agreement between the two datasets regarding the most significant networks within each country. The linear regression fit gives us a predictive model from the APNIC dataset estimates to the AnonCDN dataset values in a way that cannot be obtained from the correlation only. The linear regression coefficient indicates the slope of the linear relationship and the intercept, showing how changes in the fraction of users according to the APNIC dataset are associated with changes in the fraction of AnonCDN (i) `User-Agents` and of (ii) traffic volume. The Kendall-Tau correlation, on the other hand, focuses on rank ordering rather than actual numerical values, offering a different perspective on agreement focused on relative position (as opposed to the linear agreement provided by the Pearson correlation). To mitigate the long tail of very small organizations' impact on the Kendall-Tau correlation, we remove organizations that accounts for less than 0.5% of a country's user population in the APNIC and AnonCDN datasets. This exclusion prevents smaller organizations with negligible user populations from skewing the rank-order agreement. Furthermore, we map organizations not present in one of the datasets to 0.

Table 4: Conditions for dataset agreement across different correlation metrics. A tick (✓) indicates that the condition of being a strong correlation is satisfied (≥ 0.8). See Figure 4

	Correlation Metrics		
	Kendall-Tau	Pearson	Linear Fit
Rank Similarity	✓		
Principal Orgs Agreement		✓	> 0
Complete Agreement	✓	✓	✓

In Table 4, we define categories of agreement based on these three coorelations' values.[3] We follow established terminology, defining strong correlation as values > 0.8 [73]. Strong Kendall-Tau highlights ranking similarities (**Rank Similarity**), i.e., both datasets identify similar organization order, even if their specific user estimates differ. Strong Pearson correlation and a strong positive linear regression coefficient suggest that the datasets agree loosely on the largest orgs and provide similar traffic estimates within the country (**Principal Orgs Agreement**). However, a perfect match between user populations across the two datasets requires both high Pearson and Kendall-Tau correlations, along with a regression coefficient close to 1 (**Complete Agreement**).

Figure 4 illustrates the relationship between the Pearson and the Kendall-Tau correlations across all the countries for User-Agents (on the top) and traffic volume (on the bottom). With regard to the two left figures, countries in the top right corner of the plot exhibit strong agreement in both rank order and linear relationships, while points outside that box correspond to countries with discrepancies in rank order (Kendall-Tau) or principal Orgs agreement (Pearson). All countries are then colored according to their category of agreements on the map on the right.

User Populations. Figure 4's top two figures examine User Estimates and User-Agents. The top left plot depicts that the APNIC and AnonCDN datasets agree on the principal org for 93.9% of countries, on the rank for 54.2%, and completely for 51.2%. Countries in North and South America, Europe, and Africa generally show complete or strong agreement on the principal org. The biggest outliers with low Kendall-Tau and Pearson correlations are Russia, Western Africa (Cameroon, Benin, Congo), and South Asia (Myanmar, Sri Lanka). Overall, the APNIC dataset is reliable for estimating users in most countries according to AnonCDN User-Agents data.

Traffic volume. In Figure 4's bottom row, we perform the same analysis for AnonCDN's traffic volume and the APNIC datasets. AnonCDN and APNIC datasets agree on the principal orgs in 91.0% of the countries, on the rank for 40.5% and completely on 36.5%. Most outliers are now found in Asia, in particular in South Asia. These findings indicate that the APNIC user population data likely reflects a mix of both user counts within Org and traffic volume, influenced by Google's ad serving strategy, which is, in turn, shaped by traffic patterns. Overall, the APNIC dataset is also a surprisingly effective proxy for AnonCDN traffic volume.

4.4 Examining Outlier ⟨country, org⟩ Pairs

In Figure 4, we identified a few outliers where no agreement exists between the User-Agents or traffic volume and APNIC datasets. We take a closer look at a few examples and discuss what limitations in the datasets each may highlight. Figure 5 provides an overview of the difference in User-Agents (top row) and traffic volume (bottom row) between the AnonCDN and APNIC datasets for four countries that our analysis identified as outliers.

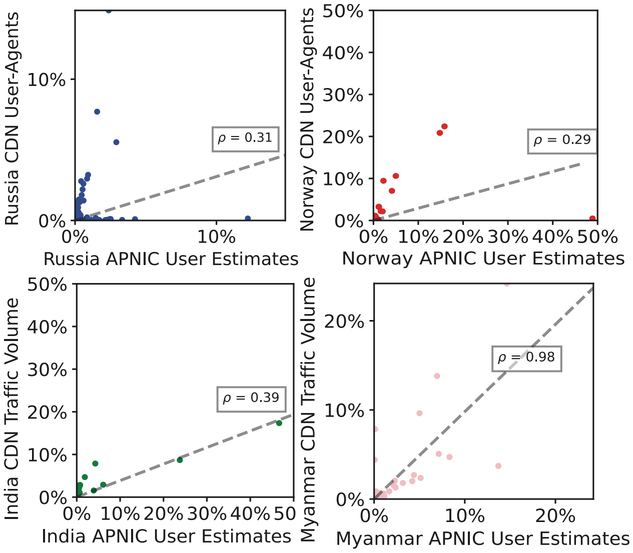


Figure 5: Comparison of the percentage of User-Agent in AnonCDN versus the percentage of user estimates in APNIC for Russia and Norway (top), and the percentage of traffic volume in AnonCDN versus the percentage of user estimates in APNIC for India and Myanmar. A linear regression is applied, with the linear coefficient (ρ) indicated on the plots.

Russia. Russia displays minimal overlap between the APNIC and AnonCDN datasets, an expected finding considering the Yandex Advertising's dominance in the country and Russia's intent to establish its own separate Internet [4]. Our analysis also identified Rostelecom as a network where the APNIC dataset heavily underestimates the user population (§4.1). Additionally, an examination of Google's PoP listings in PeeringDB [65] reveals a reduction of its footprint in Russia from five locations in 2020 to two in 2023, coinciding with the ongoing conflict in Ukraine. This decrease, alongside Google's decision to pause ads on its properties and networks for advertisers based in Russia and its subsequent bankruptcy filing in the region [35], suggests a significant pullback by Google, potentially impacting its hosted Internet services' quality and usability. Thus, we anticipate that Google services will see lower usage, which in turn means their ads are likely to be less representative. Focusing on Figure 5's upper-left graph, the APNIC dataset's inference erroneously attributes relatively minor cloud and content provider in Russia with a high number of customers. Surprisingly, this Org ranks as the 23rd-largest globally in terms of user population according to the APNIC dataset. This observation underscores the importance of understanding the APNIC dataset's potential pitfalls.

Norway. The APNIC and AnonCDN datasets' most significant discrepancies are linked to a specific VPN service, which appears the APNIC dataset disproportionately represents. The overrepresentation is likely due to the VPN's traffic anonymization features, where user traffic is funneled through a limited number of IP addresses

[3] A detailed list of country agreement levels is available in the GitHub repository.

that happen to be geolocated in Norway, although they are likely elsewhere. The APNIC dataset does not account for this type of traffic concentration, it captures many more samples associated with these IP addresses compared to the AnonCDN dataset, which maps the VPN IP addresses to their actual geographical location.

India. The largest discrepancies between the APNIC and AnonCDN datasets occur with major cloud providers and CDNs, which appear more populated in the AnonCDN data. The APNIC dataset is not designed to measure traffic volume or the user populations for these types of orgs, so it is not surprising it weights them less than the AnonCDN dataset does. For these networks, we speculate that they are unlikely candidates for ad targeting because they have few unique users to be served ads. Rather, most traffic likely involves backend services, APIs, and automated systems.

Myanmar. Frequent internet shutdowns are a persistent challenge for Myanmar Internet users [59]. The AnonCDN dataset, which is computed over a much smaller time window than APNIC's 60-day window, is more sensitive to these disruptions and captures short-term fluctuations in network usage more effectively. In contrast, the APNIC dataset, which relies on sampling over a longer period, may not reflect the same variability caused by these short-lived shutdowns. As a result, the AnonCDN data provides a more dynamic view of the network, while the APNIC dataset may present a less detailed picture of the changing network conditions in Myanmar.

5 IMPROVING THE APNIC DATASET'S USABILITY

Section 4 demonstrated that the APNIC dataset is largely consistent with the view provided by the AnonCDN and Broadband datasets. Our objective in this section is to show how the research community can improve its use of the APNIC dataset's estimates by introducing methods to prevent data misuse. We develop techniques that do not require proprietary information and enhance the accuracy and reliability of APNIC user estimates. Specifically, we explore two practical strategies to assess the reliability of the APNIC dataset's estimates: *self-consistency* (analyzing the "`Sample`" data (§5.1) and examining temporal stability (§5.1.2)) and *external consistency* (analyzing the dataset's consistency with the M-Lab (§5.2) and IXP capacity (§5.3) datasets).

5.1 Is the APNIC Dataset Self-Consistent?

5.1.1 Clarifying the importance of samples in the APNIC dataset.

A crucial, yet often overlooked, aspect of the APNIC dataset is the "`Sample`" column. Figure 6 shows that the Sample entry provides vital insights into whether there are sufficient samples to reliably use a country's AS User Population estimates. In some cases, especially when studying user populations at the AS level, the "`Sample`" may offer a more accurate indicator of network activity. Rescaling based on national factors can distort AS-level variations. Using the raw "Samples" data can avoid biases from country-wide normalization, particularly in underrepresented countries where limited samples can skew estimates.

We perform a linear regression on the log-transformed number of samples and users for every country's largest organization, calculating a 95% confidence interval, with points outside this interval marked as outliers.[4] The linear coefficient of the log-log regression, known as the elasticity coefficient β in econometry [12], measures the percentage change in one variable in response to a 1% change in another variable. For example, if β is 2, a 1% increase in the number of "Samples" would lead to a 2% increase in the "User Estimates". In this case, $\beta \approx 0.9$ indicates that on average, an increase of 1% of "Samples" results in an increase of 0.97% of the "User Estimates".

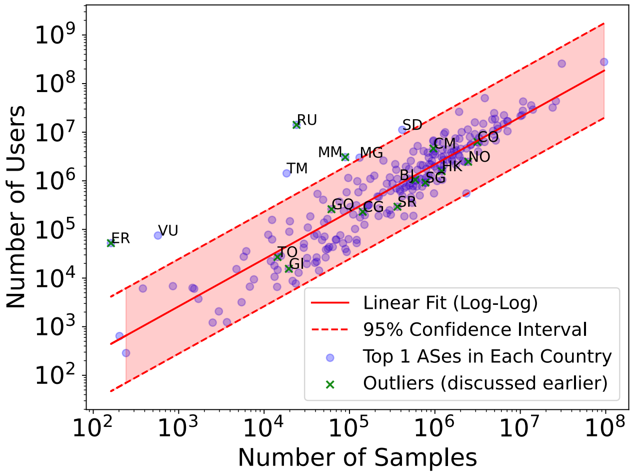


Figure 6: Comparison of the APNIC dataset's "`Samples`" versus "`User Estimates`" on a log-log scale for the top Org in each country. Outliers (marked with a green 'x' and labeled with the country code) represent the countries identified in Section 4 as having discrepancies with the AnonCDN dataset. We also label the countries that deviate from the confidence interval. Data source: APNIC, dated 2024-08-09.

We want to highlight two different sets of organizations. The first set includes countries that fall above the 95th percentile in the log-log regression analysis, where the relationship between Users and Samples is less reliable. These countries include Russia, Turkmenistan (which experiences strict Internet censorship [58]), Eritrea, Madagascar, Sudan, Myanmar, and Vanuatu. In these cases, each sample holds significantly more weight, with 1 sample representing 100 times more users than in a country along the line of best fit. The second set, labeled as "Outliers" with a green '×', consists of countries identified as having 'No Agreement' in Figure 4. We observe that the outliers Russia, Eritrea, and Myanmar appear above the confidence interval, suggesting that APNIC's estimates of AS populations in those countries may be off due to limited sample coverage *and* that a low number of samples relative to population estimates can flag APNIC estimates as suspect. However, the figure reveals that a lack of sample does not explain all outliers.

5.1.2 Avoiding noisy output with further aggregation.

Elasticity evolution. The previous section hints at the fact that if the ratio between "`User Estimates`" and "`Samples`" in a given country is too large, then it is likely to indicate a country with too few samples to trust the APNIC dataset's results. To investigate this property further, we study each country's samples and population

[4] We performed the same analysis for top 5, 10, and 20 and found no differences in the resulting outlier countries. In each case, the data points for a given country were colinear, indicating that the projection factor from Sample to User Estimates remained constant across networks for that country regardless of the number of Samples.

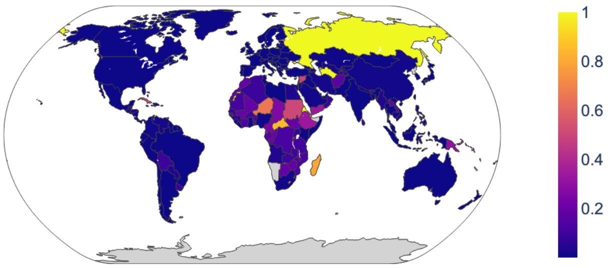


Figure 7: Fraction of days across 2024 where the User-to-Sample ratio did not lie in the confidence interval estimated in Section 5.1.

across all of the daily APNIC datasets from 2024 and focus on the proportion of time the country's User-to-Sample ratio lies above the computed higher-bound for elasticity (§5.1.1). The results can be seen in Figure 7. This analysis is based on the assumption that instances, where the User-to-Sample ratio is above the higher bound, are more likely to provide less trustworthy "`User Estimates`."

Figure 7 reveals three key insights: (1) in some countries, particularly former Soviet states like Russia or Turkmenistan, the User-to-Sample ratio is consistently larger than the upper bound across 2024. This pattern suggests that the underlying process of user estimation is likely to be erroneous. (2) For the majority of countries globally, the coefficient remains consistently below 1 for most of the time. In particular, for these countries, picking the APNIC dataset for any day will likely result in a stable output. (3) In certain countries, primarily in Africa, there are specific dates in 2024 where the User-to-Sample ratio temporarily dives below the threshold, implying that data from different dates may yield more accurate estimates.

Stability over time. The APNIC dataset is averaged across a 60 day window with the intent of smoothing the data. Therefore, we expect that the user estimates for individual organizations are likely to remain reasonably stable over short periods, except during significant events like major outages, company mergers, or takeovers where user populations may be combined. Consequently, APNIC's user estimates per AS are expected to exhibit minimal daily and weekly variation. Therefore, substantial fluctuations in these estimates could suggest issues in the data generation process and raise concerns about the data's reliability.

To assess the stability of the APNIC dataset's distribution of per-AS estimates, we employ the Kolmogorov-Smirnov (K-S) distance to compare user population distributions at consecutive times (t and $t+1$, where t and $t+1$ reflect the data's selected granularity levels: daily, monthly, and yearly). We conduct this analysis at different temporal granularity levels to capture various dynamics (Fig. 8, solid lines). For $\approx$ 10% of the countries, the K-S distance between two successive days is larger than 0.2, meaning that the number of users estimated to be in an organization differs by at least 20% of a country's Internet population across consecutive days for at least one organization (i.e., day t to day $t+1$). The result suggests significant day-to-day variability in user estimates within an Org for more than 10% of (country, day) pairs. Analyzing the monthly and yearly distribution changes also stresses significant temporal dynamics. This observation implies the need to align selecting an APNIC dataset with the timing of measurements to ensure that weighting is relevant.

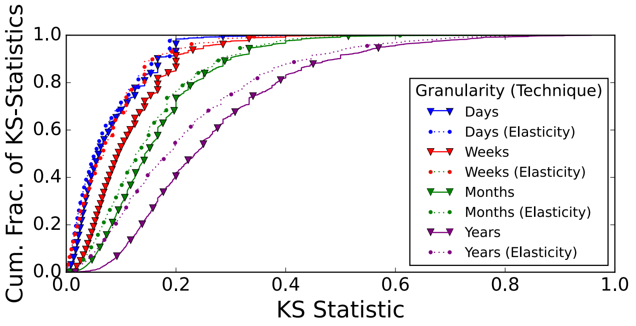


Figure 8: CDF of Kolmogorov-Smirnov (K-S) statistics of the APNIC user population distributions dataset across different temporal granularities (daily, monthly, and yearly). The K-S distance is used to quantify the stability of the dataset, reflecting the difference between user estimates at two consecutive time points. Adjusted curves when picking for each country the day with the smallest β ratio within 60 days are included to capture the increased stability in user estimates induced by our aggregation technique.

Synthesizing both insights. We combine insights from both experiments by replacing the value used in Figure 8 with a new method that selects the date with the smallest elasticity ratio β ratio over a 60-day period. This adjustment corresponds to the dashed lines in Figure 8. This method shows that the evolution of K-S distance is much less sensitive to temporal changes. By using this aggregation strategy, we enhance the dataset's stability, providing more reliable user estimates.

5.2 Using M-Lab Datasets to Identify Mistakes in the APNIC's Datasets

In the previous section, we examined how self-consistency could help identify countries and dates where the APNIC dataset might fail. Now, we turn to analyzing overlaps with an external public dataset. Our goal is to show how discrepancies with a public dataset can help pinpoint regions where the APNIC estimates may be inaccurate, even in the absence of proprietary AnonCDN data. Specifically, we examine the overlap between the APNIC and the M-Lab datasets and how a lack of agreement between these datasets often correlates with a similar lack of agreement between the APNIC and AnonCDN traffic volume datasets (§5.2).

The number of speed tests conducted in a given country can serve as a useful proxy for *traffic volume*. The idea is that more frequent testing often corresponds to higher internet usage levels, which in turn can reflect the overall traffic volume. Despite its limitations—such as being user-initiated and possibly biased towards users who run speed tests more frequently—this metric provides a rough estimate of how traffic is distributed across different networks. As an initial filtering step, we exclude all countries where M-Lab is not integrated into Google Search results [38]. In these countries, only users who actively visit the M-Lab website are likely to run the speed test, which represents a much smaller fraction of the country's total user base.

Methodology. We explore whether M-Lab dataset can help identify cases where the APNIC dataset might inaccurately estimate traffic

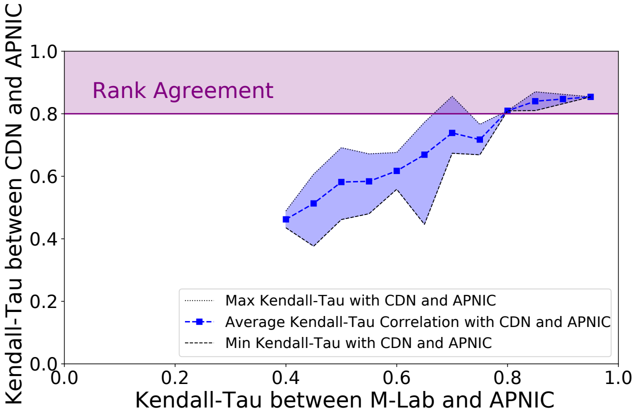


Figure 9: Relationship between Kendall-Tau correlation bins of the APNIC and M-Lab datasets and the average correlations with the AnonCDN dataset. We use purple shading to indicate the correlation thresholds required to achieve Rank Agreement. The plot shows the average Kendall-Tau correlation between the APNIC and AnonCDN datasets for each bin, with shaded regions representing the range between the minimum and maximum correlations observed. For countries where APNIC shows strong agreement with public M-Lab data, it also agrees with private AnonCDN data, providing a public test for confidence.

volume in a given country. Specifically, we examine whether a high correlation between M-Lab and the APNIC datasets predicts a high correlation between the APNIC dataset's user estimates and AnonCDN dataset's traffic volume. We use the Kendall-Tau metric, which is shown to have the largest spread in Figure 4. To explore this question, we group the Kendall-Tau correlations between the M-Lab dataset and APNIC datasets into bins of 0.05 and examine the minimum, average, and maximum Kendall-Tau correlations between the APNIC and AnonCDN datasets within each bin.

Results. Figure 9 illustrates this relationship, highlighting that a strong Kendall-Tau correlation between the APNIC and the M-Lab datasets leads to a stronger Kendall-Tau correlation between the APNIC and AnonCDN datasets. This trend supports the idea that higher agreement between M-Lab and the APNIC datasets is associated with greater accuracy in the APNIC dataset compared to the AnonCDN traffic volume. By extension, focusing on countries where the M-Lab and APNIC datasets have high Kendall-Tau correlation improves the likelihood that the APNIC dataset's population estimates closely match the AnonCDN dataset's traffic volume.

5.3 Better Traffic Volume Estimation by Combining the APNIC and IXP Datasets

An indicator of a network's traffic volume can be its total peering capacity, as higher traffic demand typically leads to expanding peering capacity to compensate. However, detailed peering link data is proprietary and not publicly available. To bypass this limitation, we analyze the capacities of various organizations at IXPs worldwide, using publicly available data from PeeringDB [65] as a proxy for their peering capacity and, by extension, their traffic volume.

IXPs play a crucial role in traffic exchange [1, 10, 16], and we show in Appendix E that the capacity an organization requisitions across multiple IXPs often reflects its interdomain capacity with AnonCDN. By supplementing APNIC user estimates with IXP capacities, we could enhance our traffic volume predictions for different organizations using only publicly available data. This approach would serve as a basis for training an inferential model, which could rely on private data during training while making predictions without needing access to private information in production

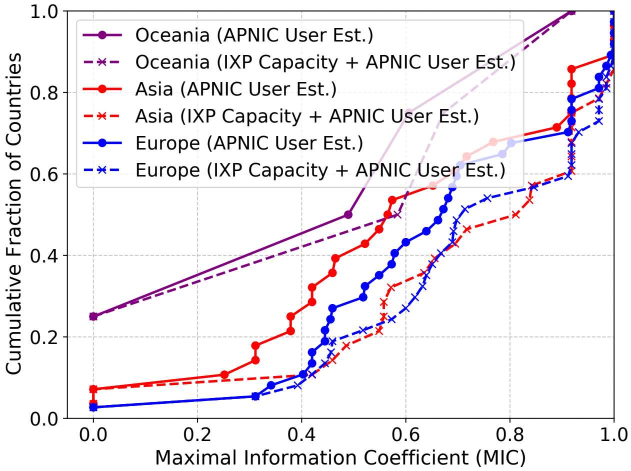


Figure 10: For three continents, CDF across countries of MIC for APNIC user estimates and IXP capacity. The solid line represents MIC values for APNIC User Estimates, while the dashed line shows the combined MIC values for both APNIC user estimates and IXP capacity. Adding IXP capacity data offers more insights into traffic volume than relying solely on the APNIC dataset.

Methodology. To explore the correlation, we calculate the Maximal Information Coefficient (MIC) [66] for each country, assessing how well (i) the APNIC dataset and (ii) the APNIC and the IXP fabric capacity can predict AnonCDN traffic volume. We refrain from using traditional correlation metrics such as the Pearson correlation, suspecting the relationship to not be linear since Private Network Interconnect (PNI) capacities are not visible and capacity may not scale linearly in terms of traffic volume.

Results. Figure 10 plots the MIC for Oceania, Asia, and Europe. The Americas exhibited very similar patterns to Europe, so we removed them to avoid clutter, and Africa has very few IXPs in PeeringDB, adding little information. Each point corresponds to a country, and the distance between the solid and dashed line indicates the amount of information gained on average by adding the IXP capacity information over using the APNIC dataset alone. Combining IXP capacity with APNIC user estimates shows promise for refining traffic volume estimates. While this approach improves accuracy compared to using either dataset alone, it has limitations, particularly in areas where IXPs play a minor role in interconnecting networks. Future work will explore how additional datasets can enhance traffic volume estimations and develop models extracting information from multiple data sources to improve our estimations.

6 ACCESS NETWORKS ARE CONSOLIDATING

Having confirmed the accuracy of the APNIC dataset and identified methods to enhance its reliability, we now explore access networks' user population concentrations, which have significant implications

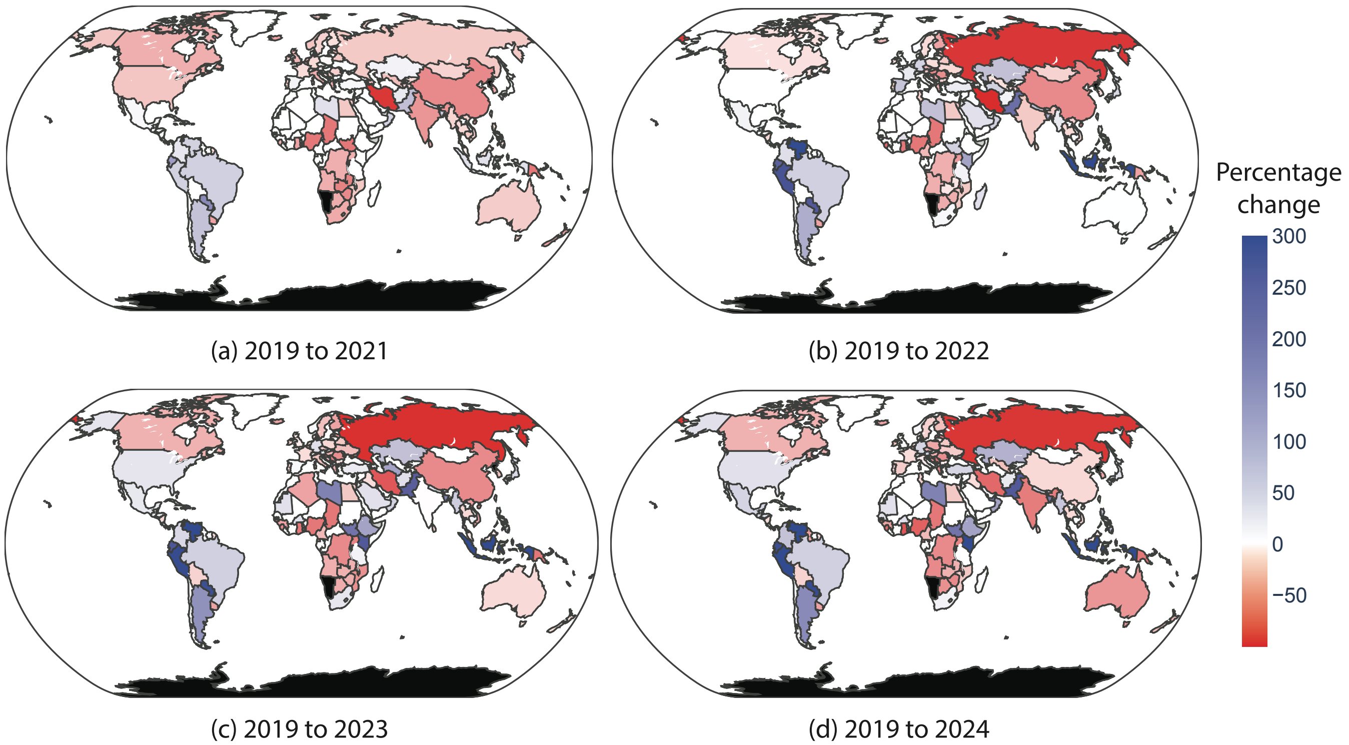


Figure 11: Yearly evolution of the number of Organizations needed to cover 95% of the population per country from 2021 (included) to 2024. The graph shows the percentage change in the number of Organizations required to reach the 95th percentile between 2019 and the labeled date. Locations in black are countries where we could not find any day to get the associated User-to-Sample ratio below the threshold we found in Section 5.1.1.

for industry and policymakers. When few access networks dominate local markets, they gain the power to suppress competition and stifle innovation. Additionally, as user data becomes concentrated within a few organizations, concerns about security and data privacy intensify, making these entities prime targets for cyberattacks and raising the stakes for any potential data breaches [56]. This phenomenon also creates new dependencies and challenges where concentrating control of access networks can also affect how and where content is delivered [79]. As content delivery shifts closer to users, the more centralized ISPs increasingly serve traffic directly within their own infrastructure. These dynamics necessitate new policies by regulatory bodies to address the challenges posed by such centralization and ensure fair access to the Internet while maintaining healthy competition and protecting consumer interests.

To understand this evolution, we analyze how the number of access network organizations needed per country to cover 95% of the population has changed over time. We examine data from 2019 to 2024, where for each year, we select the first day where the elasticity coefficient falls within the range identified in Section 5.1.1. A value of 100% means the number of organizations has doubled, while -50% means it has halved. Figure 11 shows this aggregate evolution from 2021 to 2024 and reveals a clear-cut split between regions. We chose 2019 as our baseline year because 2020 was atypical due to the COVID-19 pandemic.

In Latin America, the number of networks required to reach the 95th percentile has massively increased since 2019. In Table 6 Appendix D, we investigate the average increase and decrease of ASN allocation and announcement in the different regions of the world between January 2019 and January 2024. In particular, the decrease in concentration in Latin America cannot be explained by an influx of new networks and could stress a unique increase in broadband diversity in the region.

In contrast, some countries in Southern Asia, such as India, have seen a drastic decrease in the number of organizations to reach the 95th percentile. Three joint phenomena explain this: (i) an increase of Internet penetration in the country from 33.7% in 2019 to 51.5% in 2023 [23] with (ii) more than 90% of users accessing the Internet via mobile [27] and (iii) the country's users consolidating into the two largest mobile service providers: Jio Fiber and Airtel Bhartia [45]. Most of Europe has experienced a steady decline in the number of organizations, possibly due to smaller ISPs being absorbed by larger companies. For example, two large Switzerland access networks, Sunrise and UPC, merged in 2020 to compete against Swisscom [34] and Vodafone acquired of Unity Media in Germany [57]. Africa has also, on average, seen a decrease in access network diversity. This is especially important in a continent where Internet penetration is the lowest in the world, with only 37% of Africans having frequent access to the Internet [42]. Tracking this evolution across continents and identifying the key players driving access network consolidation is an important topic we plan to explore in future work.

7 RELATED WORK

The need for a proxy for traffic volume has been identified in prior work [69]. They look at creating a weighted graph of the Internet based on the popularity of a network path and posit that a path's popularity can serve as a proxy for traffic volume. The approach showed a strong correlation between traffic volume and the investigated metrics, but the method requires massive traceroute campaigns, which are known to potentially include inaccuracies [55] and biases based on the number and location of sources [25].

More recently, researchers put out a call to action for help creating a traffic map of the Internet [46]. The call did not want traffic volume to rely on proprietary data, they were hopeful that the large Cloud/Content Providers/CDNs could validate data as they did in prior works [8, 30]. In this work, we leveraged a private dataset (§3.4) to validate a public dataset (§3.2) for the benefit of the research community.

Prior work either used a private dataset [18] or peer-to-peer data [19], which are proprietary and out of date, respectively. Another recently leveraged Domain Name Service (DNS) traffic analysis–both Google DNS cache information and root DNS traces–to infer user populations [44]. However, the DNS analysis only identifies the user presence within an AS or IPv4 prefix, and does not infer traffic volume.

8 CONCLUSIONS

The Internet research community can highly benefit from publicly available datasets. One such dataset, the APNIC dataset, has been used extensively without proper scrutiny. In this work, we provided the first validation of its contents for user populations and traffic volume while highlighting its shortcomings.

We hope this can serve as a call to continue to scrutinize publicly available datasets to improve their accuracy as they are a critical resource for the entire community.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their constructive feedback and our shepherd for guiding us through the final submission process. This paper was partially funded by NSF grant CNS-2148275.

REFERENCES

[1] Bernhard Ager, Nikolaos Chatzis, Anja Feldmann, Nadi Sarrar, Steve Uhlig, and Walter Willinger. 2012. Anatomy of a large European IXP. In *Proc. ACM SIGCOMM.*

[2] AmIUnique.org. [n. d.]. Am I Unique. https://amiunique.org/

[3] Scott Anderson, Loqman Salamatian, Zachary S. Bischof, Alberto Dainotti, and Paul Barford. 2022. iGDB: connecting the physical and logical layers of the internet. In *Proc. ACM IMC.*

[4] Daryna Antoniuk. 2023. Russia wants to isolate its internet, but experts warn it won't be easy. https://therecord.media/russia-internet-isolation-challenges. Accessed: 2023-12-03.

[5] APNIC. [n. d.]. Customers per AS Masurements — Visible ASNs: Customer Populations (Est.). https://stats.labs.apnic.net/aspop/

[6] ARCEP. 2024. Fixed Broadband and Superfast Broadband Market Press Release. https://en.arcep.fr/news/press-releases/view/n/fixed-broadband-and-superfast-broadband-market-140324.html Accessed: 2024-09-07.

[7] Todd Arnold, Ege Gürmeriçliler, Georgia Essig, Arpit Gupta, Matt Calder, Vasileios Giotsas, and Ethan Katz-Bassett. 2020. (How Much) Does a Private WAN Improve Cloud Performance? In *Proc. IEEE INFOCOM.*

[8] 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 *Proc. ACM IMC.*

[9] Zachary S. Bischof, Kennedy Pitcher, Esteban Carisimo, Amanda Meng, Rafael Bezerra Nunes, Ramakrishna Padmanabhan, Margaret E. Roberts, Alex C. Snoeren, and Alberto Dainotti. 2023. Destination Unreachable: Characterizing Internet Outages and Shutdowns. In *Proc. ACM SIGCOMM.*

[10] Timm Böttger, Gianni Antichi, Eder L Fernandes, Roberto di Lallo, Marc Bruyere, Steve Uhlig, Gareth Tyson, and Ignacio Castro. 2018. Shaping the Internet: 10 Years of IXP Growth. *arXiv preprint arXiv:1810.10963* (2018).

[11] Timm Böttger, Felix Cuadrado, and Steve Uhlig. 2018. Looking for Hypergiants in PeeringDB. In *SIGCOMM CCR*, Vol. 48. 13–19.

[12] Lyle D Broemeling. 1986. *Econometrics and structural change.* Vol. 74. CRC Press.

[13] CAIDA. [n. d.]. The CAIDA UCSD AS to Organization Mapping Dataset, 2024/01. https://www.caida.org/data/as_organizations.xml.

[14] CAIDA. 2021. The CAIDA UCSD AS Classification Dataset, 2020–2021. https://www.caida.org/catalog/datasets/as-classification

[15] Esteban Carisimo, Alexander Gamero-Garrido, Alex C. Snoeren, and Alberto Dainotti. 2021. Identifying ASes of State-Owned Internet Operators. In *Proc. ACM IMC.*

[16] Nikolaos Chatzis, Georgios Smaragdakis, and Anja Feldmann. 2013. On the importance of Internet eXchange Points for today's Internet ecosystem. *ArXiv* abs/1307.5264 (2013). https://api.semanticscholar.org/CorpusID:384168

[17] Alex Chen and Nate Sales. 2021. Multi-User IP Address Detection. https://blog.cloudflare.com/multi-user-ip-address-detection/ The Cloudflare Blog.

[18] 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 *Proc. ACM IMC.*

[19] David R. Choffnes and Fabián E. Bustamante. 2008. Taming the Torrent: A Practical Approach to Reducing Cross-Isp Traffic in Peer-to-Peer Systems. In *Proc. ACM SIGCOMM.*

[20] David D Clark and Sara Wedeman. 2021. Measurement, meaning and purpose: Exploring the M-Lab NDT dataset. In *TPRC49: The 49th Research Conference on Communication, Information and Internet Policy.*

[21] Cloudflare. [n. d.]. About: Cloudflare Radar. https://radar.cloudflare.com/about.

[22] Australian Communications and Media Authority (ACMA). 2024. Telco Reporting Obligations. https://www.acma.gov.au/telco-reporting-obligations Accessed: 2024-09-07.

[23] DataReportal. 2024. Digital 2024: India. https://datareportal.com/reports/digital-2024-india Accessed: 2024-09-07.

[24] Cloudflare Docs. [n. d.]. Onion Routing and Tor support. https://developers.cloudflare.com/network/onion-routing/.

[25] Damien Fay, Hamed Haddadi, Andrew Thomason, Andrew W. Moore, Richard Mortier, Almerima Jamakovic, Steve Uhlig, and Miguel Rio. 2010. Weighted Spectral Distribution for Internet Topology Analysis: Theory and Applications. *IEEE/ACM ToN.*

[26] Federal Communications Commission (FCC). 2024. Broadband Data Collection. https://www.fcc.gov/BroadbandData Accessed: 2024-09-07.

[27] Data for India. 2024. Living Conditions: Access to Communication Technology. https://www.dataforindia.com/living-conditions-access-to-comm-tech/ Accessed: 2024-09-07.

[28] Blinded for review. 2022.

[29] Electronic Frontier Foundation. [n. d.]. Cover Your Tracks. https://coveryourtracks.eff.org/

[30] 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 *Proc. ACM SIGCOMM.*

[31] Petros Gigis, Vasileios Kotronis, Emile Aben, Stephen D. Strowes, and Xenofontas Dimitropoulos. 2017. Characterizing User-to-User Connectivity with RIPE Atlas. In *Proc. ACM ANRW.*

[32] Phillipa Gill, Christophe Diot, Lai Yi Ohlsen, Matt Mathis, and Stephen Soltesz. 2022. M-Lab: User initiated Internet data for the research community. *SIGCOMM CCR* (2022).

[33] Vasileios Giotsas, George Nomikos, Vasileios Kotronis, Pavlos Sermpezis, Petros Gigis, Lefteris Manassakis, Christoph Dietzel, Stavros Konstantaras, and Xenofontas Dimitropoulos. 2021. O Peer, Where Art Thou? Uncovering Remote Peering Interconnections at IXPs. In *IEEE/ACM ToN.*

[34] Liberty Global. 2020. Liberty Global Completes Acquisition of Sunrise. https://www.libertyglobal.com/liberty-global-completes-acquisition-of-sunrise/ Accessed: 2024-09-07.

[35] Google. 2022. Update related to Russian ads (March 2022). https://support.google.com/adspolicy/answer/11960078?hl=en Accessed: 2024-04-25.

[36] Google. 2024. About Display ads and the Google Display Network. https://support.google.com/google-ads/answer/2404190

[37] Google. 2024. Understanding Google Ads country restrictions. https://support.google.com/google-ads/answer/6163740?hl=en

[38] Google Support. n.d.. Understand your test results - Google's partnership with M-Lab. https://support.google.com/websearch/answer/6283840?visit_id=

638614097442592364-2117357741&p=speedtest&rd=1#zippy=%2Cunderstand-your-test-results%2Cgoogles-partnership-with-m-lab Accessed: 2024-09-07.

[39] Freedom House. 2024. *Freedom on the Net 2024 Scores*. https://freedomhouse.org/countries/freedom-net/scores Accessed: YYYY-MM-DD.

[40] Geoff Huston. 2014. How Big is that Network? https://labs.apnic.net/index.php/2014/10/02/how-big-is-that-network/.

[41] Geoff Huston. 2024. Private Communication.

[42] International Telecommunication Union (ITU). 2023. Measuring digital development: Facts and figures 2023. https://www.itu.int/hub/publication/d-ind-ict_mdd-2023-1/ Accessed: 2024-09-07.

[43] Akshath Jain, Deepayan Patra, Peijing Xu, Justine Sherry, and Phillipa Gill. 2022. The Ukrainian Internet Under Attack: an NDT Perspective. In *IMC*.

[44] 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 *Proc. ACM SIGCOMM*.

[45] Gagandeep Kaur. 2023. India's Top 2 Mobile Carriers Fight for Supremacy in Fixed Broadband. https://www.fierce-network.com/wireless/indias-top-2-mobile-carriers-fight-supremacy-fixed-broadband Accessed: 2024-09-07.

[46] Thomas Koch, Weifan Jiang, Tao Luo, Petros Gigis, Yunfan Zhang, Kevin Vermeulen, Emile Aben, Matt Calder, Ethan Katz-Bassett, Lefteris Manassakis, Georgios Smaragdakis, and Narseo Vallina-Rodriguez. 2021. Towards a Traffic Map of the Internet: Connecting the Dots between Popular Services and Users. In *Proc. ACM HotNets*.

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

[48] DART Financial Supervisory Service Korea. 2024. Financial Report. https://dart.fss.or.kr/dsaf001/main.do?rcpNo=20240320002050 Accessed: 2024-09-07.

[49] Vasileios Kotronis, George Nomikos, Lefteris Manassakis, Dimitris Mavrommatis, and Xenofontas Dimitropoulos. 2017. Shortcuts through Colocation Facilities. In *Proc. ACM IMC*.

[50] Xiang Li, Baojun Liu, Xiaofeng Zheng, Haixin Duan, Qi Li, and Youjun Huang. 2021. Fast IPv6 Network Periphery Discovery and Security Implications. In *Proc. IEEE/IFIP Dependable Systems and Networks*.

[51] Ioana Livadariu, Ahmed Elmokashfi, and Amogh Dhamdhere. 2020. An agent-based model of IPv6 adoption. In *2020 IFIP Networking Conference (Networking)*.

[52] Aemen Lodhi, Natalie Larson, Amogh Dhamdhere, Constantine Dovrolis, and Kc Claffy. 2014. Using peeringDB to understand the peering ecosystem. In *SIGCOMM CCR*.

[53] Kyle MacMillan, Tarun Mangla, James Saxon, Nicole P. Marwell, and Nick Feamster. 2023. A Comparative Analysis of Ookla Speedtest and Measurement Labs Network Diagnostic Test (NDT7). *Proc. ACM Meas. Anal. Comput. Syst.* (2023).

[54] G. Maier, F. Schneider, and A. Feldmann. 2011. NAT usage in Residential Broadband Networks. In *Proc. PAM*.

[55] P. Marchetta, A. Montieri, V. Persico, A. Pescapé, Í. Cunha, and E. Katz-Bassett. 2016. How and How Much Traceroute Confuses Our Understanding of Network Paths. In *Proc. IEEE LANMAN*.

[56] Nick Merrill and Tejas N Narechania. 2023. Inside the Internet. *Duke Law Journal Online* (2023).

[57] Broadband TV News. 2019. Vodafone initiates Unitymedia integration. https://www.broadbandtvnews.com/2019/09/02/vodafone-initiates-unitymedia-integration/ Accessed: 2024-09-07.

[58] Sadia Nourin, Van Tran, Xi Jiang, Kevin Bock, Nick Feamster, Nguyen Phong Hoang, and Dave Levin. 2023. Measuring and Evading Turkmenistan's Internet Censorship: A Case Study in Large-Scale Measurements of a Low-Penetration Country. In *Proc. ACM Web Conference*.

[59] Access Now. 2023. *Internet shutdowns in Myanmar persist as a tool of control, Access Now condemns*. https://www.accessnow.org/press-release/myanmar-keepiton-internet-shutdowns-2023-en/

[60] Ministry of Internal Affairs and Communications (Japan). 2024. Japan Telecommunications Market Data Report 2024. https://www.soumu.go.jp/main_content/000936792.pdf Accessed: 2024-09-07.

[61] University of Oregon. 2024. Route Views Archive Project. http://routeviews.org

[62] Ministry of Science and ICT. 2024. Ministry of Science and ICT - Republic of Korea. https://www.msit.go.kr/eng/index.do Accessed: 2024-09-07.

[63] Ofcom. 2024. Communications Market Report. https://www.ofcom.org.uk/research-statistics-and-data/cmr/ Accessed: 2024-09-07.

[64] Ramakrishna Padmanabhan, Arturo Filastò, Maria Xynou, Ram Sundara Raman, Kennedy Middleton, Mingwei Zhang, Doug Madory, Molly Roberts, and Alberto Dainotti. 2021. A multi-perspective view of Internet censorship in Myanmar. In *Proc. ACM SIGCOMM Workshop on Free and Open Communications on the Internet*.

[65] PeeringDB. [n. d.]. PeeringDB. http://www.peeringdb.com.

[66] David N Reshef, Yakir A Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, and Pardis C Sabeti. 2011. Detecting novel associations in large data sets. *science* 334, 6062 (2011), 1518–1524.

[67] P. Richter, G. Smaragdakis, D. Plonka, and A. Berger. 2016. Beyond Counting: New Perspectives on the Active IPv4 Address Space. In *Proc. ACM IMC*.

[68] Loqman Salamatian, Todd Arnold, Ítalo Cunha, Jiangchen Zhu, Yunfan Zhang, Ethan Katz-Bassett, and Matt Calder. 2023. Who Squats IPv4 Addresses?. In *SIGCOMM CCR*, Vol. 53.

[69] Mario A. Sanchez, Fabian E. Bustamante, Balachander Krishnamurthy, Walter Willinger, Georgios Smaragdakis, and Jeffrey Erman. 2014. Inter-Domain Traffic Estimation for the Outsider. In *Proc. ACM IMC*.

[70] Patrick Sattler, Juliane Aulbach, Johannes Zirngibl, and Georg Carle. 2022. Towards a tectonic traffic shift? investigating Apple's new relay network. In *Proc. ACM IMC*.

[71] Brandon Schlinker, Italo Cunha, Yi-Ching Chiu, Srikanth Sundaresan, and Ethan Katz-Bassett. 2019. Internet Performance from Facebook's Edge. In *Proc. ACM IMC*.

[72] B Schlinker, H. Kim, T. Cui, E. Katz-Bassett, H. V. Madhyastha, I. Cunha, J. Quinn, S. Hasan, P. Lapukhov, and H. Zeng. 2017. Engineering Egress with Edge Fabric: Steering Oceans of Content to the World. In *Proc. ACM SIGCOMM*.

[73] Patrick Schober, Christa Boer, and Lothar Schwarte. 2018. Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia* 126 (02 2018), 1.

[74] Statistica. 2024. Most used internet providers / brands in Austria as of March 2024. https://www.statista.com/forecasts/1001225. Survey conducted in Region, April 2023 to March 2024 with 1307 respondents, aged Age group.

[75] Statistica. 2024. Most used internet providers / brands in Canada as of March 2024. https://www.statista.com/forecasts/998473. Survey conducted in Region, April 2023 to March 2024 with 1240 respondents, aged Age group.

[76] Statistica. 2024. Most used internet providers / brands in Italy as of March 2024. https://www.statista.com/forecasts/1000674. Survey conducted in Region, April 2023 to March 2024 with 1254 respondents, aged Age group.

[77] Statistica. 2024. Most used internet providers / brands in the U.S. as of March 2024. https://www.statista.com/forecasts/997229. Survey conducted in Region, April 2023 to March 2024 with 5561 respondents, aged Age group.

[78] Elisa Tsai, Ram Sundara Raman, Atul Prakash, and Roya Ensafi. 2024. Modeling and Detecting Internet Censorship Events. In *Proc. ISOC NDSS*.

[79] 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 22nd ACM Workshop on Hot Topics in Networks* (Cambridge, MA, USA) *(HotNets '23)*. Association for Computing Machinery, New York, NY, USA, 70–78. https://doi.org/10.1145/3626111.3628213

[80] 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 *Proc. ACM HotNets*.

[81] Zesen Zhang, Jiting Shen, and Ricky K. P. Mok. 2024. Empirical Characterization of Ookla's Speed Test Platform: Analyzing Server Deployment, Policy Impact, and User Coverage. In *Proc. IEEE Computing and Communication Workshop and Conference*.

A ETHICS

This work uses and studies aggregate data about Internet traffic volume and user population estimates only at the AS- or country-level. We believe that this aggregated data does not have any ethical or privacy concerns.

B PER COUNTRY TRAFFIC VOLUME FOR OVERLAPPING ⟨country, org⟩ PAIRS

Table 5: Per country totals for overlapping ⟨country, org⟩ pairs between the APNIC and AnonCDN datasets when weighted by traffic volume and aggregated at the country level. The countries are sorted according to the total traffic volume (% Vol) within the country when the traffic volume for the country's overlapping ⟨country, org⟩ pairs are summed together. The overlapping ⟨country, org⟩ pairs include over 95% of the traffic volume for the vast majority of countries.

Count	Country	% Vol
1	Uruguay	100.00
2	Norfolk Island	100.00
3	Comoros	100.00
4	Costa Rica	100.00
5	Tunisia	99.99
6	Togo	99.99
7	Algeria	99.99
8	Bolivia, Plurinational State of	99.99
9	Oman	99.99
10	Burundi	99.99
11	Chile	99.98
12	Uzbekistan	99.98
13	Macao	99.98
14	Mali	99.98
15	Guinea	99.98
16	Cabo Verde	99.98
17	American Samoa	99.98
18	Haiti	99.98
19	Guyana	99.98
20	Sri Lanka	99.98
21	Jordan	99.98
22	Bahamas	99.98
23	Albania	99.97
24	Belarus	99.97
25	El Salvador	99.97
26	Egypt	99.97
27	Côte d'Ivoire	99.97
28	Niger	99.97
29	Morocco	99.97
30	Jamaica	99.97
31	Kyrgyzstan	99.97
32	Cambodia	99.96
33	Bosnia and Herzegovina	99.96
34	Sint Maarten (Dutch part)	99.96
35	Suriname	99.96
36	Sierra Leone	99.96
37	Nicaragua	99.96
38	Madagascar	99.96
39	Guinea-Bissau	99.96
40	Cameroon	99.96
41	Rwanda	99.96
42	Paraguay	99.96
43	Taiwan, Province of China	99.95
44	Hungary	99.95
45	Honduras	99.95
46	Congo	99.95
47	Zambia	99.95
48	Yemen	99.95
49	Namibia	99.95
50	Lebanon	99.95
51	Dominica	99.95
52	Benin	99.95
53	Kazakhstan	99.94
54	Guatemala	99.94
55	Zimbabwe	99.94
56	Myanmar	99.94
57	Sudan	99.94
58	Qatar	99.94
59	Uganda	99.94
60	Gambia	99.94
61	Saudi Arabia	99.93
62	Saint Lucia	99.93
63	Maldives	99.93
64	Lao People's Democratic Republic	99.93
65	Cook Islands	99.93
66	Venezuela, Bolivarian Republic of	99.92
67	Portugal	99.92
68	Kenya	99.92
69	New Caledonia	99.92
70	Åland Islands	99.92
71	Antigua and Barbuda	99.92
72	Réunion	99.92
73	Lesotho	99.92
74	Grenada	99.92
75	Angola	99.92
76	Liberia	99.92
77	Burkina Faso	99.92
78	Congo, The Democratic Republic of the	99.91
79	Barbados	99.91
80	Tajikistan	99.91
81	Mauritania	99.91
82	Kuwait	99.91
83	Gabon	99.91
84	Lithuania	99.91
85	Israel	99.90
86	Ghana	99.90
87	Nepal	99.90
88	Tanzania, United Republic of	99.90
89	Bonaire, Sint Eustatius and Saba	99.90
90	Curaçao	99.90
91	Malawi	99.90
92	Greece	99.89
93	North Macedonia	99.89

94	Equatorial Guinea	99.89
95	Aruba	99.89
96	Fiji	99.89
97	Nigeria	99.88
98	Croatia	99.88
99	Viet Nam	99.88
100	Cayman Islands	99.88
101	Malaysia	99.87
102	Pakistan	99.87
103	Panama	99.87
104	Northern Mariana Islands	99.87
105	Montenegro	99.87
106	Bulgaria	99.86
107	China	99.86
108	Georgia	99.86
109	Brunei Darussalam	99.86
110	Romania	99.85
111	Philippines	99.85
112	Trinidad and Tobago	99.85
113	Mozambique	99.85
114	Jersey	99.85
115	Guernsey	99.85
116	Colombia	99.83
117	Denmark	99.83
118	Saint Vincent and the Grenadines	99.83
119	Faroe Islands	99.83
120	Andorra	99.83
121	Mexico	99.82
122	Puerto Rico	99.82
123	Türkiye	99.81
124	Slovakia	99.80
125	Gibraltar	99.80
126	Argentina	99.79
127	Belgium	99.79
128	Ireland	99.79
129	Botswana	99.79
130	Mauritius	99.79
131	Belize	99.78
132	Czechia	99.77
133	Thailand	99.77
134	United Arab Emirates	99.77
135	Sao Tome and Principe	99.77
136	French Polynesia	99.77
137	Mongolia	99.77
138	Djibouti	99.76
139	Virgin Islands, U.S.	99.75
140	Japan	99.74
141	Moldova, Republic of	99.74
142	Bangladesh	99.73
143	Korea, Republic of	99.73
144	Finland	99.73
145	Virgin Islands, British	99.72
146	Guadeloupe	99.71
147	Malta	99.71
148	Greenland	99.71
149	Anguilla	99.70
150	France	99.69
151	Dominican Republic	99.69
152	Libya	99.69
153	Sweden	99.68
154	Azerbaijan	99.68
155	Chad	99.67
156	Bermuda	99.66
157	Isle of Man	99.66
158	Australia	99.65
159	Samoa	99.65
160	Netherlands	99.64
161	Solomon Islands	99.61
162	Guam	99.61
163	Saint Helena, Ascension and Tristan da Cunha	99.60
164	United Kingdom	99.59
165	Singapore	99.59
166	Eswatini	99.59
167	Estonia	99.59
168	Ukraine	99.56
169	Italy	99.56
170	Bhutan	99.56
171	Iraq	99.55
172	Spain	99.54
173	Germany	99.54
174	Peru	99.53
175	South Sudan	99.53
176	Bahrain	99.52
177	Timor-Leste	99.50
178	Luxembourg	99.50
179	Papua New Guinea	99.50
180	Turks and Caicos Islands	99.48
181	Poland	99.47
182	Cuba	99.42
183	South Africa	99.41
184	Syrian Arab Republic	99.41
185	Hong Kong	99.39
186	Cyprus	99.39
187	Slovenia	99.39
188	Iceland	99.39
189	New Zealand	99.33
190	Latvia	99.33
191	Somalia	99.32
192	Ethiopia	99.31
193	Central African Republic	99.29
194	Canada	99.26
195	Marshall Islands	99.26
196	San Marino	99.25
197	Martinique	99.23
198	Switzerland	99.21
199	Palau	99.20
200	Indonesia	99.17
201	Micronesia, Federated States of	99.13
202	Iran, Islamic Republic of	99.10
203	Tonga	99.06

204	Saint Kitts and Nevis	99.05
205	Falkland Islands (Malvinas)	98.99
206	Wallis and Futuna	98.90
207	Norway	98.86
208	Afghanistan	98.83
209	Montserrat	98.79
210	Serbia	98.77
211	India	98.73
212	Saint Pierre and Miquelon	98.73
213	Saint Martin (French part)	98.72
214	Mayotte	98.62
215	Ecuador	98.53
216	Kiribati	98.44
217	Senegal	97.46
218	United States	97.39
219	Eritrea	97.32
220	Armenia	97.31
221	Holy See (Vatican City State)	96.50
222	Monaco	95.82
223	Brazil	93.73
224	Vanuatu	93.54
225	Palestine, State of	93.25
226	Nauru	93.20
227	Austria	92.42
228	Russian Federation	92.35
229	Seychelles	91.24
230	French Guiana	90.59
231	Liechtenstein	89.86
232	Turkmenistan	88.65
233	Saint Barthélemy	85.76
234	Tuvalu	79.28
235	United States Minor Outlying Islands	0.00
236	Tokelau	0.00
237	French Southern Territories	0.00
238	T1 [24]	0.00
239	Svalbard and Jan Mayen	0.00
240	Pitcairn	0.00
241	Niue	0.00
242	Korea, Democratic People’s Republic of	0.00
243	British Indian Ocean Territory	0.00
244	South Georgia and the South Sandwich Islands	0.00
245	Western Sahara	0.00
246	Christmas Island	0.00
247	Cocos (Keeling) Islands	0.00
248	Antarctica	0.00

C TIME-SENSITIVITY OF ANONCDN

D1 We conduct a sensitivity analysis to examine the impact of varying the number of days on the fraction of User-Agents observed in a ⟨country, org⟩ pair. To assess the data stability over time, we measured the maximum difference in User Agent percentages for each ⟨country, org⟩ pair between any two dates in 2024 to capture their variability. Our analysis revealed that for more than 93% of the ⟨country, org⟩ pairs, the difference in User-Agent percentages was less than 1%. Only 0.8% of the ⟨country, org⟩ pairs have at least two days where the fraction of User-Agents was above 5%. We establish that these 0.8% are either from very small countries, where percentage differences result in minor absolute changes in the Estimated User Estimates, or from countries with less Internet freedom, where higher variability is expected. In particular, 43.4% of that subset originated from countries with populations under 100,000 and 66.0% from countries with populations under 2 million. Of the remaining countries, 92% are from nations where the Internet Freedom Index is below 30[39] (*e.g.,* Myanmar [64], Turkmenistan [58]), where we expect more volatility in the User-Agent populations. Ultimately, we observe slight variations in only 23 ⟨country, org⟩ pairs (< 0.02%) across 7 countries. This subset is unlikely to have a significant impact on our final results.

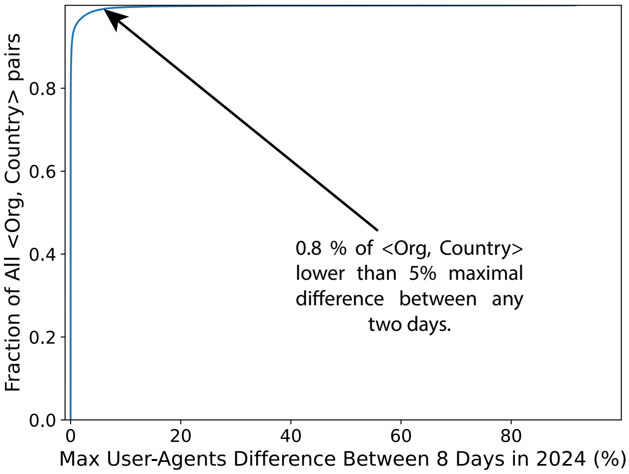


Figure 12: Cumulative distribution of the maximum differences in User-Agents percentages across Organization over 2024 in AnonCDN. The black arrow marks the 5% threshold where there are meaningful differences.

D EVOLUTION OF ASN ALLOCATED AND ANNOUNCED

Table 6: Percentage Increase in Allocated and Advertised ASNs per Region (2019-2024)

Region	Allocated ASN Incr. (%)	Advertised ASN Incr. (%)
Caribbean	20.46	33.14
Central America	7.31	10.17
South America	3.21	8.98
Northern America	-15.13	-12.25
Eastern Asia	62.46	130.34
Asia	42.31	48.99
Southern Asia	55.78	26.93
South-Eastern Asia	27.60	24.37
Eastern Africa	16.94	20.18
Southern Africa	9.50	12.19
Northern Africa	4.06	11.07
Africa	7.93	10.71
Eastern Europe	-28.69	-20.93
Southern Europe	-12.37	-5.01
Northern Europe	-13.46	-10.13
Western Europe	-11.21	-5.32
Australia and New Zealand	-12.87	-10.57
Oceania	-12.29	-10.10

E RELATION BETWEEN THE PNI AND IXP CAPACITIES

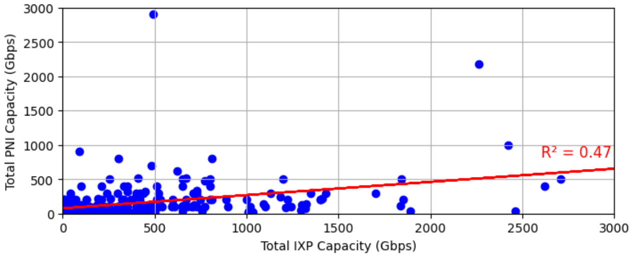


Figure 13: Correlation between IXP Capacity and Private Network Interconnect (PNI) Capacity for AnonCDN Traffic. The plot illustrates a linear regression of IXP peering capacity against PNI capacity for a specific CDN, yielding an R^2 value of 0.47. While not a perfect match, the correlation indicates that IXP peering capacity is a reasonable proxy for estimating PNI capacity, though variations may arise due to factors such as regional user distribution and network-specific routing practices

In this section, we study the relationship between IXP capacity and PNI capacity for AnonCDN traffic to determine whether IXP capacity can serve as a proxy for PNI capacity. This analysis is important because PNI data is often proprietary, and if IXP capacity proves to be a reliable indicator, it could provide insights into traffic volumes for AnonCDN. By understanding this correlation, we can better estimate traffic flows and network behaviors where direct PNI data is unavailable.

In Figure 13, we perform a simple linear regression to evaluate the correlation between an AS's IXP peering capacity and the AS's PNI capacity with AnonCDN. The R^2 value from this analysis shows that, while IXP peering capacity is not a perfect match for AnonCDN PNI, there is a noticeable coarse alignment between the two ($R^2 = 0.47$). This observation suggests that IXP peering capacity can serve as a useful, though approximate, indicator of PNI capacity.