Towards a traffic map of the Internet

ABSTRACT

The impact of an outage, congestion, network attack, hijacking, and many other Internet phenomena depends on how many users, queries, or traffic use the route, but researchers lack visibility into how important routes are. Distressingly, the research community seems to have lost hope of obtaining this information without relying on proprietary datasets or privileged viewpoints. We argue for optimism thanks to new network measurement methods and changes in Internet structure which make it possible to construct a living "Internet traffic map." This map would identify the (network) locations of users and major services, the paths between them, and the relative activity levels routed along these paths. We sketch our vision for the map, detail new measurement ideas for map construction, and identify key challenges that the research community should tackle. The realization of an Internet traffic map will be an Internet-scale research effort that will have Internet-scale impacts that reach far beyond the research community, and so we hope our fellow researchers are excited to address this challenge.

1 INTRODUCTION

I've done it. Heck, I'm reviewing 12 IMC papers right now that do it. Maybe you've even done it. That doesn't make it right. We're talking, of course, about graphing a CDF across Internet paths or destinations or networks, giving each path or destination or network equal weight. As if each outage is equally impactful. Or each inflated route impacts the same number of users and services. Or the raw number of customer networks or routes crossing a network is a good measure of the network's importance. Or each congested interconnect impacts the same amount of traffic.

But we all know that isn't how the Internet works. Especially not today's Internet. Most user-facing traffic flows from a handful of large providers. Most other large services are hosted by one of a few large cloud providers. The larger providers serve the traffic from CDN caches in thousands of networks around the world [29], or across private peering links only used for their traffic [68]. The amount of traffic from these services varies greatly across user networks and over time. Compared to these routes between popular services and users, many other Internet routes carry little traffic.

Understanding the relative levels of activity on routes—what we will call an *Internet traffic map*—is crucial to understanding the Internet. An Internet traffic map would give networking researchers and operators meaningful ways to weight and interpret their results, and it would point them to which problems and solutions are most relevant. It would provide security researchers with valuable information to contextualize observed phenomena and to inform operational decisions and inform investments. It could give policy makers and economists a lens into how traffic, content, and money flows.

Two SIGCOMM papers that revealed aspects of an Internet traffic map illustrate the benefits and challenges of creating one. These papers reshaped the research community's mental model of the Internet's structure and evolution, and their citation counts provide a crude estimate of the impact—a 2010 paper that revealed that most traffic flows between a small number of content providers and user networks [43] has 859 citations (according to Google Scholar on June 24, 2021), and a 2012 paper on the rise of Internet peering that found that more than 90% of the IXP's peerings were not visible in public topologies [5] has 353 citations. These studies, relying on proprietary data unavailable to the academic community at large, both shaped the community's understanding of the Internet and research agendas going forward and pointed out the inadequacies of existing public datasets and measurement techniques.

In fact, despite decades of work on mapping the Internet, no existing work captures this sort of traffic-weighted map of the Internet using only public data. Most work has focused on IP or network layer maps [1, 11, 32, 44, 58, 66]. Some work has focused on physical maps [24]. Other work avoided the need for proprietary data by crowdsourcing measurements and, often, focusing on small slices of the Internet, such as regions [25, 33, 34, 50] or regional cellular networks [47, 52, 53, 65, 69]. These efforts provide valuable insights, but crowdsourced platforms are difficult to scale or maintain over time. Some recent work investigates aspects related to an Internet traffic map, although none provides a solution. Alexa and similar top lists capture aspects of site popularity [57], but do not provide a fine-grained understanding of which or how users are being served by those sites. APNIC publishes estimates of the number of eyeball users per network [36], but the data are coarse-grained and the approach has not been validated. Other work estimated the amount of traffic that crossed IXP peerings based on the number of traceroutes that crossed them [56], but the approach is not applicable for the vast majority of traffic on today's Internet that crosses private interconnects or flows from caches. There is also work on estimating traffic matrices, but we are not aware of any that can answer the questions of how much traffic

Internet routes carry relative to each other without access to proprietary data.

We think the research community has viewed creating an Internet traffic map using public data as an impossible goal at Internet scale—previous proposals for an Internet traffic map assume (proprietary) measurements from all clients of a CDN [18, 62]—but we bear a message of hope: emerging trends—including content consolidation, increased adoption of TLS, and increased usage of public DNS services—open up new measurement opportunities that we believe can combine to reveal the core components of an Internet traffic map, although many challenges remain.

This paper is a call to action. The research community needs to develop techniques that can provide a traffic map of the Internet, and we need to use such a map to inform and interpret our research. Let today be the first step towards banishing unweighted CDFs to the dustbins of SIGCOMM history and towards a brighter future full of CDFs (and research!) that reflect the traffic patterns of the Internet. Towards that goal, we make the following contributions:

- We discuss possible uses of an Internet traffic map.
- Informed by the use cases, we posit the high level components and attributes of an Internet traffic map.
- For each component, we discuss why previous work does not satisfy the need, sketch possible measurement techniques, and present major open questions that remain. We intend these techniques and questions to provide a research roadmap for the community.

2 CREATING AN ITM IS IMPORTANT AND POSSIBLE WITH YOUR HELP

2.1 ITM overview

We imagine components to answer these questions (Table 1):

1. *Where are users? What is their (relative) activity level?*
2. *Where are popular services hosted, and what is the mapping from users to these hosts?*
3. *What routes are used between users and services?*

Building these components will require building on existing measurement techniques and addressing remaining challenges (§3). Table 1 summarizes possible time and network granularities. The second (time) and third (network) columns of Table 1 contain granularities which become coarser when read from left to right. The left-most granularity in columns two and three is a reasonable goal for how often or fine-grained we believe that components should be measured – *i.e.,* measurement efforts may be better spent elsewhere after a certain precision is achieved. Shown in bold is the finest granularity at which we believe we can measure each component *right now*, with reasonable coverage, using techniques we sketch in Section 3. Extending our map to include user activity levels at finer granularities with broader coverage will require new measurement methods (§3.1.2) and/or more data (§4).

ITM Component		Time Granularity	Network Granularity
Where are Users & What Are Relative Activity Levels?	Finding Prefixes with Users	Diurnal Pattern, **Daily Changes**, Semi-Regular Average, Yearly	User IP address, Infrastructure Router, **Prefix**, Recursive Resolver, AS
	Estimating Activity per Prefix	Diurnal Pattern, Daily Changes, Semi-Regular Average, **Yearly**	User IP address, Infrastructure Router, Prefix, **Recursive Resolver**, AS
Where are Services Hosted & How Are Users Mapped to Services?	Mapping Services	**O(months)**	Type of Component, **Service (owner)**, Physical Location
	Mapping Users to Services	O(hours) to **O(months)**	IP, **Prefix**, (AS, Location)
What are Routes Between Users and Services?		O(days) to **O(months)**	IP address, Router, (location, AS), **AS**

Table 1: Summarizing the desirable granularities of each component of the ITM. Granularities get coarser when read left to right, top to bottom in columns two and three. We bold the granularity at which we can build the component using *existing* measurement methods, with decent coverage and accuracy.

The desired measure of activity can vary by use case, including number of users, total traffic volume, or volume or query count for a particular service. We sketch techniques that capture some of these, for some services, and we hope other researchers will develop techniques to fill gaps. For most use cases, relative levels of activity (*e.g.,* "`prefix1` has twice as much activity as `prefix2`") suffice and are easier to estimate. Machine-to-machine traffic plays an important role on the Internet but is not within the scope of this submission.

2.2 The benefits of an ITM

Benefits to Internet researchers. We provide specific examples of Internet research where an Internet traffic map would drastically change the conclusions we draw from analysis. Chiu *et al.* demonstrated the impact that access to an Internet traffic map can have on an Internet measurement study [21]. When considering iPlane's paths from PlanetLab to all prefixes with responsive destinations [44]—a traditional academic Internet topology—only 2% of Internet paths were two ASes long. When instead issuing paths from Google cloud (which hosts many popular services), 41% were two ASes long. When considering paths from Google to end-user destinations (the source of most connections for Google cloud), 61% were two ASes long. When weighting by query volume, 66% were two ASes long. When also considering the impact of Google's off-net servers in other networks, 73% of queries come from ASes that either peer with Google, use off-nets hosted in their providers, or themselves host off-nets (Fig. 3 in [21]). This huge swing from most paths being long to most paths being short can inform what problems to work on and what solutions to pursue. Similarly, other work looked whether users of a large CDN experienced routing *inflation* by being directed to a CDN site farther away than the optimal one (Figure 1) [41]. While only 31% of (unweighted) routes take users to the closest site, 60% of users are mapped

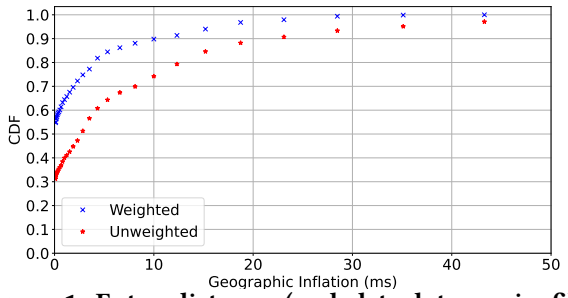


Figure 1: Extra distance (scaled to latency in fiber) queries from a large CDN have to travel over their closest front-end , when weighting and when not. Weighting makes a significant difference – routing optimality is improved by 20%.

to the optimal site, providing very different views of routing efficiency.

We surveyed 2020 IMC papers to assess the potential impact of an ITM. Of 54 papers, 19 would benefit from an ITM, 3 because they used a private dataset that could be replaced or validated [14, 26, 54]. All 19 would benefit from user activity (e.g., [27]), 11 from the service locations (e.g., [51]), and 10 from the routes between users and services (e.g., [31]).

Benefits to industry. Many network operators lack visibility to inform and contextualize operational decisions such as network blackouts, performance anomalies (e.g., path inflation), unusual traffic patterns, or DDoS attacks. Information about users, major services those users are interacting with, and routes users traverse will help network operators diagnose problems and efficiently plan for the future. A multi-vantage point view has the potential to identify vulnerabilities in the Internet structure and prevent or mitigate large scale attacks and network misconfigurations.

Benefits to other fields. The insights gained by creating a Internet traffic map can be useful for economists, policy makers and regulators, and sociologists. An Internet traffic map can feed better models of the interactions of the various stakeholders (and their evolution over time), including the large content providers that are the biggest investors in Internet infrastructure. It can inform assessments of the impact of decisions, e.g., related to network neutrality regulation or monopolies. It can serve as input to assessments of censorship, digital division, and segregation.

2.3 Initial evidence of feasibility

Figure 2 shows initial progress on building a ITM, depicting locations of prefixes where we detected user activity (by probing Google Public DNS (§3.1.2)). To limit probing overhead, we only considered prefixes within 1500 km of a PoP we probed. Figure 2 shows promise, but also that truly global coverage will require refinements to our methods and/or

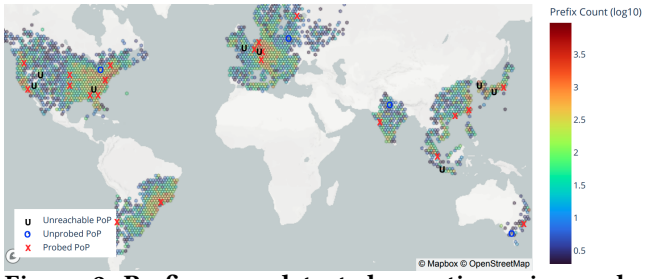


Figure 2: Prefixes we detected as active using cache probing – our measurement methods have good coverage *right now*, suggesting a map is feasible.

more measurement methods. Coupling this uncovering of users with recent work showing the feasibility of uncovering all hypergiant servers [29] gives a useful starting point for an Internet traffic map.

3 TOWARDS MEASURING COMPONENTS OF AN INTERNET TRAFFIC MAP

3.1 What are user activity levels?

As a key to interpreting research results, aiding operations, and weighting analysis, the ITM should indicate the level of user activity within a prefix.

3.1.1 Limitations of existing approaches. Prior work used proprietary data to weight analysis [21], which allowed for useful insights but was not reproducible. An existing, publicly available alternative is APNIC's network population data [36]. APNIC's network population data based on ad impressions has been used in studies [7, 9, 28, 42], but APNIC's methodology has a number of limitations, the key ones being that the approach has not been validated (to the best of our knowledge) and that APNIC aggregates data at an AS granularity, which is too coarse-grained for use cases that require prefix-level information. Other work achieved broad coverage of users by releasing a popular BitTorrent plugin [22], but BitTorrent's popularity has declined, and no recent research projects achieved broad coverage or longevity.

3.1.2 Possible measurement approaches. To overcome the limitations of the APNIC data without resorting to proprietary data, we propose a few approaches to determine which prefixes have users. A key challenge is extending them to estimate relative user activity levels (§3.1.3).

Approach 1: Probing DNS caches. Users issue DNS queries to look up IP addresses for Internet services, and recursive resolvers cache responses. We can issue a non-recursive query for a popular domain (to focus on *user* activity) to a recursive resolver to determine whether users of the resolver queried recently for the domain. Since most ISP-hosted resolvers will not respond to queries from outside the ISP, we instead

propose to use Google Public DNS, a popular global DNS service (it contributed 30-35% of all DNS queries to Microsoft Azure authoritative DNS servers in January 2019 [19]). To achieve worldwide coverage, we can leverage the EDNS0 Client Subnet (ECS) option, which enables specifying a client prefix, causing Google Public DNS to only return a result (for domains that support ECS) if a client from that prefix recently queried for the domain. By iterating over all routable prefixes, we can populate a binary "activity" map by prefix for a domain. Using a combination of cloud VMs and VPNs, our queries can reach 25 out of 33 Google Public DNS sites (Figure 2), representing 99.9% of Google Public DNS queries as seen from a popular authoritative DNS service.

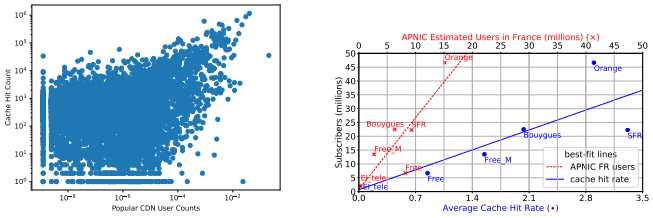
Approach 2: Crawling DNS logs. Many popular web browsers including Chrome, Edge, Brave, and Opera use the Chromium web browser codebase. Chromium browsers use DNS probes to detect DNS interception [63], querying for random strings of 7-15 lowercase letters a-z. Because these queries often have no valid TLD (*e.g.,* COM), they should not result in cache hits at recursive resolvers, so the queries go to a DNS root server [63]. We can count Chromium queries in root DNS logs which are made available yearly as part of the DITL packet captures [2]. Since queries in the root DNS logs are often made by recursive resolvers (rather than end-users), crawling root DNS logs would give us an indicator of activity by recursive resolver.

3.1.3 *Open questions.*

Can we estimate relative activities? Most proposed measurement methods in Section 3.1.2 provide proxies for relative user activity (volume), each of which may be imperfect. For example, the number of Chromium queries seen at the DNS roots is likely roughly proportional to the number of Chromium users behind a recursive resolver.

Crawling Google Public DNS caches provides us with a *binary* indicator of activity. To extend this binary indication to a proxy of activity, we propose looking at cache hit rates over time, with the intuition that prefixes with more activity will populate caches more often. To test our intuition, we used ECS to probe for a day at least once per TTL for popular web services and recorded cache hit counts by AS. We compared the cache hit counts with user counts by AS according to a popular CDN (Figure 3a), and compare relative cache hit rates with both APNIC user counts and subscriber counts of French ISPs (Figure 3b). Figure 3 shows a correlation between cache hits and other measures of activity, suggesting it may be a decent proxy of user activity.

An additional methodology that may be useful is measuring IP ID counters. Any host generating a packet must include an IP ID value, and many routers source the IP ID values from an incrementing counter. By pinging a router



(a) (b)

Figure 3: Comparing cache hit count heuristic with other measures of activity by AS – number of users of a popular CDN (3a), APNIC user estimates and French ISP subscriber counts (3b). Preliminary results suggest the cache hit count is a decent measure of activity.

interface, one can monitor the growth of its counter over time, a technique used to infer aliases [10, 40, 59]. We have observed that the IP ID values of most routers display diurnal patterns, suggesting that the rate at which the routers source packets may be proportional to the rate at which they forward traffic (which is known to follow diurnal patterns), perhaps because they export flow statistics proportional to traffic volume. We propose measuring IP IDs over time to generate time series and use increment velocity (*e.g.,* at peak time) to estimate traffic volume (and changes in traffic volume) traversing the router.

How can techniques be combined to best overcome biases and limitations and enable fine-grained mapping? The techniques offer different tradeoffs. Probing caches of Google Public DNS enables per-user-prefix, per-service inferences on the granularity of the service's DNS record's TTL, but use of Google Public DNS may be skewed, despite its popularity, and it is an open question how well these measurements translate into relative activity levels. Crawling root DNS logs provides global coverage and a direct measure of relative activity, but Chromium usage may be skewed, the measurements indicate activity of (unknown) users using a (known) recursive resolver, the measurements happen only once a year, and more and more root operators anonymize the data in ways that limit the coverage of our proposed analysis.

Realizing the best Internet traffic map attainable will require combining the techniques *and* designing methods to best mitigate their limitations. Since some DNS roots are operated by research organizations (*e.g.,* ISI and the University of Maryland), it may be possible to work with them to provide real-time access to logs with appropriate (*e.g.,*/24) anonymization for the purposes of maintaining the Internet traffic map. Since these logs capture the address of the recursive resolver (rather than of the user), we will either need to make simplifying assumptions (*e.g.,* clients are in the same /24 as their recursive resolver) or deploy techniques

to associate recursive resolvers with their clients (*e.g.,* embedding measurements of the associations in popular pages [45]). Such an association would enable joining of resolver-based techniques with user-based techniques. Similarly, it is possible that logs from organizations can help understand biases in Chromium usage and/or Google Public DNS usage.

3.2 Where are services located & how are users mapped to them?

3.2.1 Limitations of existing approaches. Existing approaches fall into two categories. The first are DNS-based approaches which issue queries from distributed measurement platforms [55, 61], open recursive resolvers [35, 64], or crowdsourcing [4, 46], in order to undercover the IP addresses associated with a particular service hostname (e.g., www.google.com). After IP address discovery, IP-to-AS mapping and geolocation are applied to map the network topologies. Coverage of service IP addresses are inherently limited by the number of vantage points.

The second set are custom tailored techniques per-service or network that get around the need for distributed vantage points, but fail to generalize to other services, and so do not scale with the Internet's growth. Studies have emulated global vantage point coverage by issuing DNS queries using the DNS Extension Client-Subnet (ECS), which allows a DNS query to include the client's IP prefix, allowing researchers to issue queries to a service that appear to come from arbitrary locations/prefixes [16, 60]. However, not all services support ECS, and those that do may only reply to ECS queries from allowlisted resolvers [20]. Other work has mapped Netflix [15] and Facebook [12, 13] servers by identifying patterns in their DNS naming scheme then exhaustively trying queries based on those patterns.

3.2.2 Possible measurement approaches.

Approach 1: Locating infrastructure using TLS scans. In recent years, there is a dramatic increase in encrypted web traffic. Major content stakeholders and software vendors (e.g., browsers) support secure protocols (e.g., HTTPS) to protect their customers' privacy. TLS certificates, by design, validate the owner of a resource, and their ubiquity provides a generalizable mechanism for attribution. A recent measurement study [29] leveraged this and demonstrated the ability to produce high-coverage maps of serving infrastructure of large content providers and CDN networks, including infrastructure hosted in ISP networks (off-nets). According to their findings, networks hosting off-nets have more than doubled from 2013 to 2021, reaching approximately 4,500 networks.

Approach 2: SNI scans for services. Web applications are complex and often depend on many different CDN and cloud providers, and existing approaches [38] have taken a top-down approach to examining web dependencies by starting with HTML and dissecting the hostnames and IP addresses that compose the parts of the whole. The disadvantages to this is the need for distributed vantage points that can only discover network dependencies based on the current location and network conditions. For this we propose to use Internet-wide SNI (TLS + hostname) scans to uncover the footprint of major web properties by identifying CDN or cloud IP addresses serving valid TLS certificates for a particular web property hostname. By requesting from all web servers on the Internet if they are willing to serve content for a particular domain, we can construct a bottom-up view of web property network dependencies with a global view.

Approach 3: Locating servers at fine granularities. The first two approaches will uncover the scale of serving infrastructure by network and the network footprint of a particular service but for many applications and analysis this granularity is too coarse. The Internet traffic map should include a city and facility for serving infrastructure. Previous efforts such as client-centric geolocation [16] and constraint based RTT from in-facility vantage points [30, 48] are starting points.

3.2.3 Open questions: How can we infer client-to-server mappings for services that do not rely on DNS redirection or support ECS?. DNS cache probing enables discovery of the client-to-server mapping for services that both rely on DNS-based redirection and support ECS, but some services lack ECS support or use anycast [17] or customized URLs to direct a user to a particular site [6]. How best to account for such services in the Internet traffic map remains an open question. Next we sketch possible directions.

There is reason to be optimistic for increased ECS adoption in service's authoritative resolvers. Many popular services that rely on DNS-based redirection support ECS, and we expect support to grow, given the demonstrated benefit [20]. Already, 15 of the top 20 sites (according to Alexa toplist) support ECS, representing 35% of Internet traffic and 91% of traffic to the top 20 sites (according to SimilarWeb.com).

Recent work demonstrates that anycast routing is extremely efficient for large services, with 80% of clients directed within 500 km of their closest serving site [41]. So, we anticipate that the main challenge is in inferring in which cases this optimality is likely violated and where clients with suboptimal routing are directed. We expect that some of these cases can be explained using enriched techniques for path prediction (§3.3). Another possibility may come from increased popularity in edge computing platforms, such as Cloudflare's Workers [3], where CDN customers can execute custom code on CDN PoPs. This may enable Verfploeter-style techniques [23] for inferring per-PoP anycast catchments by probing out to the Internet.

It is extremely difficult to infer where individual clients are directed when the redirection happens via URLs customized to individual clients. URL pattern matching approaches similar to what has been done for Netflix [15] may provide some possibilities but this remains a challenging open problem.

However, we anticipate that this challenge will not have a large effect on the Internet traffic map. Such redirection only kicks in after the client has fetched and parsed HTML (or similar content with the URL embedded) from some server reached via an alternate redirection mechanism (typically DNS redirection or anycast), and so it is only worth the switching overhead for long-lived connections, meaning the mechanism is typically used for high-volume connections, many of which are cacheable content, especially video-on-demand. Because custom URLs can be tailored per-client, they enable very precise redirection. As such, we believe that the vast majority of bytes served from sites reached via custom URLs are likely from the optimal site. So, to capture most traffic, it may be enough to *guess* that these services are usually served optimally, without directly observing the custom URLs and the sites they redirect to. An important task in developing the Internet traffic map may be validating this intuition via instrumentation from available vantage points and networks. To refine this guess, it is critical to understand the efficacy of these caches and the flow of traffic. A community-driven project could host such caches inside research networks, and universities, to measure the cache hit rate under normal operation and during flash events.

3.3 What are routes between users/servers?

The ITM should contain routes between users and servers, which is key to understanding topology, infrastructure vulnerability, and Internet economics.

3.3.1 Limitations of existing approaches. Paths are commonly predicted on AS, PoP-level, router, or IP granularity. A way for predicting AS paths is to use BGP data and AS relationships to predict paths using Valley Free routing and Prefer-Customer type assumptions [37]. This method only works well if we have complete BGP data, but BGP feeds lack information about many peering links. We found that when predicting paths from RIPE Atlas probes to the root DNS servers using a BGP simulator, more than half of paths could not be predicted due to a missing link.

More fine-grained path prediction relies on splicing [44], which requires a large number of vantage points.

3.3.2 Possible measurement approaches. Paths between users and servers are becoming easier to predict, due to Internet flattening and the evolving role of large cloud and content providers. Prior work showed simple heuristics could be used to accurately predict path lengths between users and Google servers [21], but that a challenge to predicting routes was lack of topological information. More recent work used new methodology to obtain missing peering links between users and large cloud providers [9], which would aid path prediction. Prediction has been made easier [21] due to Internet flattening since most users have short, downhill paths to services. Traceroute campaigns from cloud providers and tools such as Reverse Traceroute can help us measure reverse paths [39].

3.3.3 Open question: Is it possible to predict missing links to complete the topology? A limitation of (existing) path prediction techniques is that they compose predictions from observed links, but it is well-known that measured Internet topologies are missing many links, especially for the large content providers [5, 49, 67]. While it is possible to measure paths to and from cloud providers [9, 39], these techniques require a vantage point within the provider (*e.g.,* a cloud VM) so are not suitable for CDNs or off-net servers. Is it possible to predict with high confidence which links exist, to feed into a path prediction algorithm? Many networks now indicate in PeeringDB the colocation facilities in which they maintain a peering presence. Given two networks are both present in a facility, it may be possible to develop techniques to predict how likely it is that two networks interconnect at that facility. Such predictions could rely on information that networks publish in PeeringDB, such as their peering policies, traffic profiles, network types, and facilities.

4 CONCLUSION AND A CALL TO ACTION

Will you help us create the Internet traffic map? First, we hope researchers will offer feedback, suggesting modifications to components/definitions/granularities of the Internet traffic map. Second, we hope the research community will work with us to explore the many open challenges to achieve broad coverage and precision (§3). Third, we envision members of the research and operator community making public (to researchers) datasets or vantage points such as root DNS logs (§3.1.3), cache logs (§3.2), and/or aggregated volume reports of networks. Fourth, although we do not want the Internet traffic map to depend on private data, large content providers can help validate it, similar to how Google, Microsoft, and Akamai validated recent work uncovering their peers [8] and deployment footprints [29]. Finally, we hope the research community both uses and encourages others to use the Internet traffic map for weighting analysis and (in general) conducting Internet research.

REFERENCES

[1] Ark. URL https://www.caida.org/projects/ark/locations.

[2] Dns-oarc, 2020. URL dns-oarc.net/oarc/data/ditl.

[3] Cloudflare workers, 2021. URL https://workers.cloudflare.com/.

[4] Bernhard Ager, Wolfgang Mühlbauer, Georgios Smaragdakis, and Steve Uhlig. Web content cartography. In *Proceedings of the 2011 ACM SIGCOMM Internet measurement conference*, pages 585–600, 2011.

[5] Bernhard Ager, Nikolaos Chatzis, Anja Feldmann, Nadi Sarrar, Steve Uhlig, and Walter Willinger. Anatomy of a large European IXP. In *Proc. ACM SIGCOMM '12*. ACM, 2012.

[6] Zahaib Akhtar, Yun Seong Nam, Jessica Chen, Ramesh Govindan, Ethan Katz-Bassett, Sanjay Rao, Jibin Zhan, and Hui Zhang. Understanding video management planes. In *Proceedings of the Internet Measurement Conference 2018*, pages 238–251, 2018.

[7] Todd Arnold, Ege Gürmeriçliler, Georgia Essig, Arpit Gupta, Matt Calder, Vasileios Giotsas, and Ethan Katz-Bassett. (how much) does a private wan improve cloud performance? In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, pages 79–88. IEEE, 2020.

[8] Todd Arnold, Jia He, Weifan Jiang, Matt Calder, Italo Cunha, Vasileios Giotsas, and Ethan Katz-Bassett. Cloud provider connectivity in the flat internet. In *Proceedings of the ACM Internet Measurement Conference*, pages 230–246, 2020.

[9] Todd Arnold, Jia He, Weifan Jiang, Matt Calder, Italo Cunha, Vasileios Giotsas, and Ethan Katz-Bassett. Cloud provider connectivity in the flat internet. In *Proceedings of the ACM Internet Measurement Conference*, pages 230–246, 2020.

[10] Adam Bender, Rob Sherwood, and Neil Spring. Fixing ally's growing pains with velocity modeling. In *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement*, pages 337–342, 2008.

[11] Robert Beverly. Yarrp'ing the Internet: Randomized high-speed active topology discovery. In *Proc. ACM IMC '16*, 2016.

[12] A. Bhatia. Mapping Facebook's FNA (CDN) nodes across the world! https://anuragbhatia.com/2018/03/networking/isp-column/mapping-facebooks-fna-cdn-nodes-across-the-world/, 2018.

[13] A. Bhatia. Facebook FNA node update. https://anuragbhatia.com/2019/11/networking/isp-column/facebook-fna-node-update/, 2019.

[14] Timm Boettger, Ghida Ibrahim, and Ben Vallis. How the internet reacted to covid-19: A perspective from facebook's edge network. In *Proceedings of the ACM Internet Measurement Conference*, pages 34–41, 2020.

[15] Timm Böttger, Felix Cuadrado, Gareth Tyson, Ignacio Castro, and Steve Uhlig. Open connect everywhere: A glimpse at the internet ecosystem through the lens of the netflix cdn. *ACM SIGCOMM Computer Communication Review*, 48(1):28–34, 2018.

[16] Matt Calder, Xun Fan, Zi Hu, Ethan Katz-Bassett, John Heidemann, and Ramesh Govindan. Mapping the expansion of google's serving infrastructure. In *Proceedings of the 2013 conference on Internet measurement conference*, 2013.

[17] Matt Calder, Ashley Flavel, Ethan Katz-Bassett, Ratul Mahajan, and Jitendra Padhye. Analyzing the performance of an anycast cdn. In *Proceedings of the 2015 Internet Measurement Conference*, pages 531–537, 2015.

[18] Matt Calder, Ryan Gao, Manuel Schröder, Ryan Stewart, Jitendra Padhye, Ratul Mahajan, Ganesh Ananthanarayanan, and Ethan Katz-Bassett. Odin: Microsoft's scalable fault-tolerant {CDN} measurement system. In *15th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 18)*, pages 501–517, 2018.

[19] Matt Calder, Xun Fan, and Liang Zhu. A cloud provider's view of edns client-subnet adoption. In *2019 Network Traffic Measurement and Analysis Conference (TMA)*. IEEE, 2019.

[20] Fangfei Chen, Ramesh K Sitaraman, and Marcelo Torres. End-user mapping: Next generation request routing for content delivery. *ACM SIGCOMM Computer Communication Review*, 2015.

[21] Yi-Ching Chiu, Brandon Schlinker, Abhishek Balaji Radhakrishnan, Ethan Katz-Bassett, and Ramesh Govindan. Are we one hop away from a better internet? In *Proceedings of the 2015 Internet Measurement Conference*, 2015.

[22] David R Choffnes and Fabián E Bustamante. Taming the torrent: a practical approach to reducing cross-isp traffic in peer-to-peer systems. *ACM SIGCOMM Computer Communication Review*, 38(4):363–374, 2008.

[23] Wouter B De Vries, Ricardo de O. Schmidt, Wes Hardaker, John Heidemann, Pieter-Tjerk de Boer, and Aiko Pras. Broad and load-aware anycast mapping with verfploeter. In *Proceedings of the 2017 Internet Measurement Conference*, pages 477–488, 2017.

[24] Ramakrishnan Durairajan, Paul Barford, Joel Sommers, and Walter Willinger. Intertubes: A study of the us long-haul fiber-optic infrastructure. In *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*, pages 565–578, 2015.

[25] Rodérick Fanou, Francisco Valera, and Amogh Dhamdhere. Investigating the causes of congestion on the african ixp substrate. In *Proceedings of the 2017 Internet Measurement Conference*, pages 57–63, 2017.

[26] Anja Feldmann, Oliver Gasser, Franziska Lichtblau, Enric Pujol, Ingmar Poese, Christoph Dietzel, Daniel Wagner, Matthias Wichtlhuber, Juan Tapiador, Narseo Vallina-Rodriguez, et al. The lockdown effect: Implications of the covid-19 pandemic on internet traffic. In *20th ACM Internet Measurement Conference*, pages 1–18. ACM, 2020.

[27] Romain Fontugne, Anant Shah, and Kenjiro Cho. Persistent last-mile congestion: Not so uncommon. In *Proceedings of the ACM Internet Measurement Conference*, pages 420–427, 2020.

[28] Petros Gigis, Vasileios Kotronis, Emile Aben, Stephen D Strowes, and Xenofontas Dimitropoulos. Characterizing user-to-user connectivity with ripe atlas. In *Proceedings of the Applied Networking Research Workshop*, pages 4–6, 2017.

[29] Petros Gigis, Matt Calder, Lefteris Manassakis, George Nomikos, Vasileios Kotronis, Xenofontas Dimitropoulos, Ethan Katz-Bassett, and Georgios Smaragdakis. Seven years in the life of hypergiants' off-nets. In *ACM SIGCOMM 2021*, 2021.

[30] Vasileios Giotsas, Georgios Smaragdakis, Bradley Huffaker, Matthew Luckie, and KC Claffy. Mapping peering interconnections to a facility. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies*, pages 1–13, 2015.

[31] Vasileios Giotsas, Thomas Koch, Elverton Fazzion, Ítalo Cunha, Matt Calder, Harsha V Madhyastha, and Ethan Katz-Bassett. Reduce, reuse, recycle: Repurposing existing measurements to identify stale traceroutes. In *Proceedings of the ACM Internet Measurement Conference*, pages 247–265, 2020.

[32] Ramesh Govindan and Hongsuda Tangmunarunkit. Heuristics for Internet map discovery. In *Proc. IEEE INFOCOM '00*, 2000.

[33] Enrico Gregori, Alessandro Improta, and Luca Sani. On the african peering connectivity revealable via bgp route collectors. In *International Conference on e-Infrastructure and e-Services for Developing Countries*, pages 368–376. Springer, 2017.

[34] Arpit Gupta, Matt Calder, Nick Feamster, Marshini Chetty, Enrico Calandro, and Ethan Katz-Bassett. Peering at the internet's frontier: A first look at isp interconnectivity in africa. In *International Conference on Passive and Active Network Measurement*, pages 204–213. Springer, 2014.

[35] Cheng Huang, Angela Wang, Jin Li, and Keith W Ross. Measuring and evaluating large-scale cdns. In *ACM IMC*, volume 8, pages 15–29, 2008.

[36] Geoff Huston. How big is that network, 2014. URL labs.apnic.net/?p=526.

[37] Yuchen Jin, Colin Scott, Amogh Dhamdhere, Vasileios Giotsas, Arvind Krishnamurthy, and Scott Shenker. Stable and practical {AS} relationship inference with problink. In *16th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 19)*, pages 581–598, 2019.

[38] Aqsa Kashaf, Vyas Sekar, and Yuvraj Agarwal. Analyzing third party service dependencies in modern web services: Have we learned from the mirai-dyn incident? In *Proceedings of the ACM Internet Measurement Conference*, pages 634–647, 2020.

[39] Ethan Katz-Bassett, Harsha V Madhyastha, Vijay Kumar Adhikari, Colin Scott, Justine Sherry, Peter Van Wesep, Thomas E Anderson, and Arvind Krishnamurthy. Reverse traceroute. In *Proc. NSDI '10*, 2010.

[40] Ken Keys, Young Hyun, Matthew Luckie, and Kim Claffy. Internet-scale ipv4 alias resolution with midar. *IEEE/ACM Transactions on Networking*, 21(2):383–399, 2012.

[41] Thomas Koch, Ke Li, Calvin Ardi, Matt Calder, John Heidemann, and Ethan Katz-Bassett. Anycast in context: A tale of two systems. In *ACM SIGCOMM 2021*, 2021.

[42] Vasileios Kotronis, George Nomikos, Lefteris Manassakis, Dimitris Mavrommatis, and Xenofontas Dimitropoulos. Shortcuts through colocation facilities. In *Proceedings of the 2017 Internet Measurement Conference*, pages 470–476, 2017.

[43] Craig Labovitz, Scott Iekel-Johnson, Danny McPherson, Jon Oberheide, and Farnam Jahanian. Internet inter-domain traffic. *ACM SIGCOMM Computer Communication Review*, 40(4):75–86, 2010.

[44] Harsha V Madhyastha, Tomas Isdal, Michael Piatek, Colin Dixon, Thomas Anderson, Arvind Krishnamurthy, and Arun Venkataramani. iplane: An information plane for distributed services. In *Proceedings of the 7th symposium on Operating systems design and implementation*, pages 367–380, 2006.

[45] Zhuoqing Morley Mao, Charles D Cranor, Fred Douglis, Michael Rabinovich, Oliver Spatscheck, and Jia Wang. A precise and efficient evaluation of the proximity between web clients and their local dns servers. In *USENIX Annual Technical Conference, General Track*, pages 229–242, 2002.

[46] Srdjan Matic, Gareth Tyson, and Gianluca Stringhini. Pythia: a framework for the automated analysis of web hosting environments. In *The World Wide Web Conference*, pages 3072–3078, 2019.

[47] Foivos Michclinakis, Hossein Doroud, Abbas Razaghpanah, Andra Lutu, Narseo Vallina-Rodriguez, Phillipa Gill, and Joerg Widmer. The cloud that runs the mobile internet: A measurement study of mobile cloud services. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pages 1619–1627. IEEE, 2018.

[48] George Nomikos, Vasileios Kotronis, Pavlos Sermpezis, Petros Gigis, Lefteris Manassakis, Christoph Dietzel, Stavros Konstantaras, Xenofontas Dimitropoulos, and Vasileios Giotsas. O peer, where art thou? uncovering remote peering interconnections at ixps. In *Proceedings of the Internet Measurement Conference 2018*, pages 265–278, 2018.

[49] Ricardo Oliveira, Dan Pei, Walter Willinger, Beichuan Zhang, and Lixia Zhang. The (in) completeness of the observed internet as-level structure. *IEEE/ACM Transactions on Networking*, 18(1):109–122, 2009.

[50] Eduardo E P. Pujol, Will Scott, Eric Wustrow, and J Alex Halderman. Initial measurements of the cuban street network. In *Proceedings of the 2017 Internet Measurement Conference*, pages 318–324, 2017.

[51] Audrey Randall, Enze Liu, Gautam Akiwate, Ramakrishna Padmanabhan, Geoffrey M Voelker, Stefan Savage, and Aaron Schulman. Trufflehunter: Cache Snooping Rare Domains at Large Public DNS Resolvers. In *Proceedings of the ACM Internet Measurement Conference*, pages 50–64, 2020.

[52] John P Rula and Fabian E Bustamante. Behind the curtain: Cellular dns and content replica selection. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, pages 59–72, 2014.

[53] John P Rula, Fabián E Bustamante, and Moritz Steiner. Cell spotting: studying the role of cellular networks in the internet. In *Proceedings of the 2017 Internet Measurement Conference*, pages 191–204, 2017.

[54] Said Jawad Saidi, Anna Maria Mandalari, Roman Kolcun, Hamed Haddadi, Daniel J Dubois, David Choffnes, Georgios Smaragdakis, and Anja Feldmann. A haystack full of needles: Scalable detection of iot devices in the wild. In *Proceedings of the ACM Internet Measurement Conference*, pages 87–100, 2020.

[55] Mario A Sánchez, John S Otto, Zachary S Bischof, David R Choffnes, Fabián E Bustamante, Balachander Krishnamurthy, and Walter Willinger. Dasu: Pushing experiments to the internet's edge. In *10th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 13)*, pages 487–499, 2013.

[56] Mario A Sanchez, Fabian E Bustamante, Balachander Krishnamurthy, Walter Willinger, Georgios Smaragdakis, and Jeffrey Erman. Interdomain traffic estimation for the outsider. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, pages 1–14, 2014.

[57] Quirin Scheitle, Oliver Hohlfeld, Julien Gamba, Jonas Jelten, Torsten Zimmermann, Stephen D Strowes, and Narseo Vallina-Rodriguez. A long way to the top: Significance, structure, and stability of internet top lists. In *Proceedings of the Internet Measurement Conference 2018*, pages 478–493, 2018.

[58] Neil Spring, Ratul Mahajan, and David Wetherall. Measuring ISP topologies with rocketfuel. *ACM SIGCOMM Computer Communication Review*, 32(4):133–145, August 2002. ISSN 0146-4833.

[59] Neil Spring, Ratul Mahajan, David Wetherall, and Thomas Anderson. Measuring isp topologies with rocketfuel. *IEEE/ACM Transactions on networking*, 12(1):2–16, 2004.

[60] Florian Streibelt, Jan Böttger, Nikolaos Chatzis, Georgios Smaragdakis, and Anja Feldmann. Exploring edns-client-subnet adopters in your free time. In *Proceedings of the 2013 conference on Internet measurement conference*, pages 305–312, 2013.

[61] Ao-Jan Su, David R Choffnes, Aleksandar Kuzmanovic, and Fabián E Bustamante. Drafting behind akamai (travelocity-based detouring). *ACM SIGCOMM Computer Communication Review*, 36(4):435–446, 2006.

[62] Yi Sun, Junchen Jiang, Vyas Sekar, Hui Zhang, Fuyuan Lin, and Nanshu Wang. Using video-based measurements to generate a real-time network traffic map. In *HotNets*, 2014.

[63] Matthew Thomas. Chromium's impact on root dns traffic, 2020. URL blog.apnic.net/2020/08/21/chromiums-impact-on-root-dns-traffic.

[64] Sipat Triukose, Zhihua Wen, and Michael Rabinovich. Measuring a commercial content delivery network. In *Proceedings of the 20th international conference on World wide web*, pages 467–476, 2011.

[65] Narseo Vallina-Rodriguez, Srikanth Sundaresan, Christian Kreibich, Nicholas Weaver, and Vern Paxson. Beyond the radio: Illuminating the higher layers of mobile networks. In *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*, pages 375–387, 2015.

[66] Kevin Vermeulen, Justin P Rohrer, Robert Beverly, Olivier Fourmaux, and Timur Friedman. Diamond-miner: Comprehensive discovery of the internet's topology diamonds. In *17th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 20)*, pages 479–493, 2020.

[67] Florian Wohlfart, Nikolaos Chatzis, Caglar Dabanoglu, Georg Carle, and Walter Willinger. Leveraging interconnections for performance: the serving infrastructure of a large cdn. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, pages 206–220, 2018.

[68] Bahador Yeganeh, Ramakrishnan Durairajan, Reza Rejaie, and Walter Willinger. How cloud traffic goes hiding: A study of amazon's peering fabric. In *Proceedings of the Internet Measurement Conference*, pages 202–216, 2019.

[69] Kyriakos Zarifis, Tobias Flach, Srikanth Nori, David Choffnes, Ramesh Govindan, Ethan Katz-Bassett, Z Morley Mao, and Matt Welsh. Diagnosing path inflation of mobile client traffic. In *International Conference on Passive and Active Network Measurement*, pages 23–33. Springer, 2014.