Towards a traffic map of the Internet Connecting the dots between popular services and users

Connecting the dots between popular services and users

Koch, Thomas; Jiang, Weifan; Luo, Tao; Gigis, Petros; Zhang, Yunfan; Vermeulen, Kevin; Aben, Emile; Calder, Matt; Katz-Bassett, Ethan; Manassakis, Lefteris

DOI
10.1145/3484266.3487371

Publication date
2021

Document Version
Final published version

Published in
HotNets 2021 - Proceedings of the 20th ACM Workshop on Hot Topics in Networks

Towards a traffic map of the Internet

Connecting the dots between popular services and users

Thomas Koch[⋆] Weifan Jiang[⋆] Tao Luo[⋆] Petros Gigis[♯] Yunfan Zhang[⋆]
Kevin Vermeulen[⋆] Emile Aben[‡] Matt Calder[⋆†] Ethan Katz-Bassett[⋆]
Lefteris Manassakis[3] Georgios Smaragdakis[1] Narseo Vallina-Rodriguez[2]

[⋆] Columbia University [†] Microsoft [‡] RIPE NCC [♯] UCL [1] TU Delft [2] IMDEA Networks / ICSI [3] FORTH-ICS

ABSTRACT

The impact of Internet phenomena depends on how they impact users, but researchers lack visibility into how to translate Internet events into their impact. Distressingly, the research community seems to have lost hope of obtaining this information without relying on privileged viewpoints. We argue for optimism thanks to new network measurement methods and changes in Internet structure which make it possible to construct an "Internet traffic map". This map would identify the 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 with Internet-scale impacts that reach far beyond the research community, and so we hope our fellow researchers are excited to join us in addressing this challenge.

ACM Reference Format:
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 *The Twentieth ACM Workshop on Hot Topics in Networks (HotNets '21), November 10–12, 2021, Virtual Event, United Kingdom.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3484266.3487371

1 INTRODUCTION

We've done it. Heck, half our papers do it, and probably half of the IMC program does it too. 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 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 today's Internet works. 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 largest providers serve traffic from CDN caches in thousands of networks around the world [25] or across private peering links only used for their traffic [64]. 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 weigh and interpret results, and point them to which problems and solutions are most relevant. It would provide security researchers with valuable information to contextualize observed phenomena. It could inform operational decisions and 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. A 2010 paper revealed that most traffic flows between a small number of content providers and user networks [40], and a 2012 paper on the rise of Internet peering found that more than 90% of the IXP's peerings were not visible in public topologies [4]. These papers reshaped the research community's mental model of the Internet's structure and evolution, and both accrued hundreds of citations and shaped research agendas going forward. However, they relied on proprietary data unavailable to the academic community at large and pointed out the inadequacies of existing public datasets and replicable measurement techniques.

HotNets '21, November 10–12, 2021, Virtual Event, United Kingdom

ACM ISBN 978-1-4503-9087-3/21/11.
https://doi.org/10.1145/3484266.3487371

Despite decades of work on Internet mapping, no work captures this sort of traffic-weighted map of the Internet using only public data. Most work focused on IP or network layer maps [1, 9, 27, 42, 55, 62]. Some work focused on physical maps [22]. Other work avoided the need for proprietary data by crowdsourcing measurements and, often, focusing on small slices of the Internet, such as regions [23, 28, 29, 49] or regional cellular networks [46, 50, 51, 61, 65]. These efforts provide valuable insights, but it is difficult to achieve representative coverage, and crowdsourced platforms are challenging to maintain over time. Alexa and other top lists capture aspects of site popularity [54], 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 users per network [33], but the data are coarse-grained, and the approach has not been validated. Other work estimated the amount of traffic on IXP peerings based on traceroutes that crossed them [53], but the approach does not apply to the vast majority of traffic on today's Internet that crosses private interconnects or flows from caches. There is work on estimating traffic matrices [30, 31], but no work we are aware of can answer how much traffic routes carry relative to each other without using proprietary data.

The research community has viewed creating an Internet-wide traffic map using public data as impossible—previous proposals assume (proprietary) measurements from CDN clients [15, 58]—but we bear a message of hope: emerging trends—including content consolidation and increased adoption of TLS and of public DNS services—open up new measurement opportunities that can reveal core components of an Internet traffic map, although many challenges remain.

This paper is a call to action. The research community should develop techniques to generate a traffic map of the Internet, and we should use the 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 posit map components and discuss use cases. For each component, we discuss why previous work does not suffice, sketch possible measurement techniques, and present major open questions. We hope these techniques and questions provide a research roadmap for the community.

2 AN INTERNET TRAFFIC MAP IS VALUABLE & (IF YOU HELP) POSSIBLE

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

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

These components scale back the far-reaching goal of measuring dynamics of the whole Internet to strike a balance between benefit (§2.1), feasibility (§3), and ambition (to push the community forward). With a small number of cloud and content providers responsible for 90% of Internet traffic [25], focusing on popular services provides most of the utility while significantly lowering complexity and scope. These providers employ similar deployment strategies that we have made progress in uncovering [7, 25] and that simplify some challenges we face [18]. By trying to identify routes commonly used between these services and users (§3.3), rather than the exact set of routes in use at a particular point in time, we simplify the problem considerably while still enabling interesting use cases (§2.1).

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 (§3.1.2), and we hope other researchers will develop techniques to fill gaps. For most use cases, relative levels of activity (*e.g.,* "`prefix1` has roughly twice as much activity as `prefix2`") suffice and are easier to estimate (§3.1.3). Machine-to-machine traffic plays an important role on the Internet but is out of our current scope.

Table 1 summarizes desired coverage and precision of components, and what is possible using *current* methods. Due to length restrictions, we point to references for these numbers instead of describing details. Extending the map to measure components at finer granularities with broader coverage will require new measurement methods (§3) and/or more data (§4). The desired precision will vary depending on use case, but being more precise than the desired precision in Table 1 would likely yield diminishing returns.

2.1 The benefits of an Internet traffic map

Benefits to Internet researchers. A study of Internet path lengths demonstrated the impact that an Internet traffic map can have on results [19]. When considering iPlane's paths from PlanetLab to all prefixes [42]—a traditional academic Internet topology—only 2% of Internet paths were two ASes long. However, the paper estimated that 73% of Google queries come from ASes that either host a Google server or connect directly with Google or another AS hosting a Google server. 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 [38]. While only 31% of routes go to the closest site, 60% of users are mapped to the optimal site, providing very different views of routing efficiency. Organizing the components together into one entity (a map) enables us to

ITM Component		Temporal Precision	Network Precision and Network Coverage	
		Desired | Now	Desired | Now	Desired | Now
Where are users? What are relative user activity levels? (§3.1)	Finding prefixes with users	Daily | Weekly [34]	/24 Prefix | Prefix [34]	65K ASes, 8.8M /24s | 50K ASes, 6.6M /24's [34]
	Estimating relative activity	Hourly | Yearly [34]	/24 Prefix | AS [34]	8.8M /24s | 40K ASes [34]
Where are services hosted? What is the mapping from users to these hosts? (§3.2)	Mapping services	Weekly | Monthly [25]	Facility | Server owner [25]	Popular services | Identifying serving infrastructure [25]
	Mapping users to hosts	Hourly | Monthly/Daily [13]	Prefix | Prefix [13]	Client /24s, All services | Routable /24s, ECS-supporting services [13]
What routes are commonly used between services and users? (§3.3)		Daily | N/A	⟨city, AS⟩ | N/A	Commonly used routes | N/A

Table 1: Current and desirable granularities of each component of the ITM, with citations to current precision/coverage. Desired precision is as fine-grained as the ground-truth values would provide significantly more information. Darker shading indicates that a component is more complete given current measurement techniques.

answer rich questions and identify connections among components. For example, to assess the impact of an outage in a ⟨region, AS⟩, the map can tell us which popular services are affected, which prefixes are affected for those services, what fraction of traffic or users are affected, and where the prefixes may be routed instead.

Benefits to industry. Network operators can lack visibility to contextualize network events such as network blackouts, performance anomalies, unusual traffic patterns, or DDoS attacks. Information about users, major services those users are interacting with, and routes users traverse will help operators diagnose problems and plan for the future.

Benefits to other fields. An Internet traffic map can be useful for economists, policy makers and regulators, and sociologists. It can feed better models of the interactions of stakeholders, 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 and digital division.

2.2 Recent results hint it may be possible!

A large barrier to creating a Internet traffic map is mental, not just technical. Building one sounds difficult, and it will be—it will require new measurement methods, many collaborators (we do not have all the answers), and years of research. But given all that the Internet measurement community has accomplished, initial evidence of feasibility (which we provide here), and Internet trends, we do not think it farfetched.

Figure 1 shows initial progress towards a ITM with global coverage of *some* components. Figure 1a demonstrates we were able to identify which prefixes around the world host web clients [34]. The figure is from our contemporaneous IMC paper that presented techniques that identify which prefixes host clients almost as well as a CDN can by looking at its server logs—the techniques identified prefixes responsible for 95% of Microsoft CDN traffic, and less than 1% of identified client prefixes did not contact Microsoft at all [34]. The shading in Figure 1b depicts the percent of a country's Internet users (according to APNIC [33]) that are in ASes that our techniques identified as hosting clients. Although APNIC user counts have not been validated, they likely capture the major eyeball networks in each country, sufficing to show that our techniques uncover most of these networks as well, accounting for 98% of Internet users by APNIC's estimates. We have also been able to uncover the locations of CDN servers globally. The dots in Figure 1b depict the locations of Facebook servers, using techniques from our recent SIGCOMM paper [25]. Locating client prefixes and servers is a promising starting point for an Internet traffic map!

3 TOWARDS MEASURING COMPONENTS OF AN INTERNET TRAFFIC MAP

3.1 What are relative user activities?

3.1.1 Limitations of existing approaches. Prior work used proprietary data to weight analysis [19], which allowed for useful insights but was not reproducible. An existing public alternative is APNIC's network population data [33]. It has been used in studies [6, 7, 24, 39], but APNIC's methodology has not been validated (to the best of our knowledge), and APNIC aggregates data at an AS granularity, which is too coarse-grained for many use cases. Other work achieved broad user coverage by releasing a popular BitTorrent plugin [20], but BitTorrent's popularity has declined, and no recent research projects achieved broad coverage or longevity.

3.1.2 Possible measurement approaches. We summarize our two recently published approaches to determine which prefixes have clients of major services [34]. Together, the methods identify Internet client ASes representing 99% of Microsoft CDN traffic. A key challenge is extending them to find Internet *users* (as opposed to bots and other non-human clients) and to estimate relative user activity levels (§3.1.3).

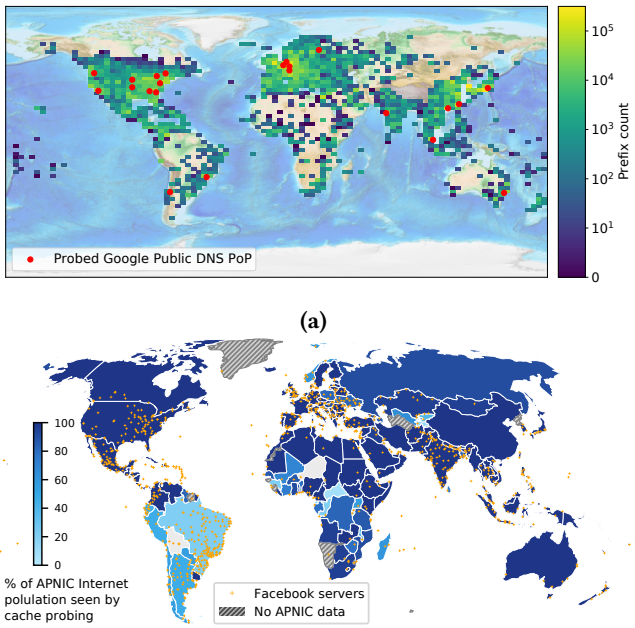

Figure 1: Locations of clients detected with cache probing (1a). Percent of a country's Internet users [33] in ASes that cache probing identified as hosting clients (shading) and locations of Facebook servers detected using TLS scans (1b). The measurements are a promising starting point for an Internet traffic map.

Approach 1: Probing DNS caches. Users issue DNS queries to look up IP addresses for Internet services, and recursive resolvers cache responses. We issued non-recursive queries for popular domains to Google Public DNS (which is responsible for 30-35% of DNS queries [16]) to determine if the popular domains were in the cache [34]. To achieve worldwide coverage, we used the EDNS0 Client Subnet (ECS) option, which enables specifying a client prefix, causing Google Public DNS to only return a result if a client from that prefix recently queried for the domain. By iterating over all routable prefixes, our methods identified client activity in prefixes representing 95% of Microsoft CDN traffic.

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 [59]. 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 [59]. In the same study, we crawled root DNS logs for Chromium queries [34]. Since most queries to the root DNS are from recursive resolvers (rather than clients), crawling root DNS logs gave an indicator of activity by recursive resolver. With the assumption that most users

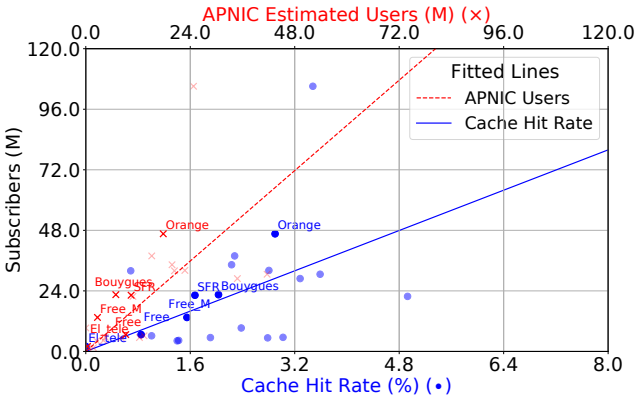

Figure 2: ISP subscriber counts vs. cache hit rate and APNIC user estimates. Preliminary results suggest cache hits may help estimate relative client activity.

are in the same AS as their recursive resolvers, crawling root DNS logs helped us identify the presence of Internet clients in ASes representing 60% of Microsoft CDN traffic.

3.1.3 *Open questions.*

Can we estimate relative activities? Crawling DNS logs provides a proxy for relative client activity (volume). The number of Chromium queries seen at the DNS roots is likely roughly proportional to the number of Chromium clients behind a recursive resolver.

Conversely, crawling Google Public DNS caches provides a *binary* indicator of activity. To extend this binary indication to relative 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 popular web services for one day and recorded cache hit counts by AS. Figure 2 compares (ground truth) subscriber counts of some large eyeball ISPs in France, Japan, South Korea, the UK, and the US[1] with (two unvalidated approaches) relative cache hit rates and APNIC user counts by AS [33]. Figure 2 shows a correlation between cache hits and other measures of activity. Since Google Public DNS adoption and prefix allocation varies by country (among other dimensions), we use French ISPs as a case study (colored darker and named in the figure). Cache hit rate correctly orders French ISPs with respect to their subscriber counts, suggesting there is *some* signal available for estimating relative activities.

An additional methodology that may be useful is measuring IP ID counters. Every packet must include an IP ID value,

[1]Since our list of sources includes links that may eventually be unavailable, we provide the code and references for generating this figure at https://github.com/tkoch96/itm_hotnets_2021_supp.

and many routers source the IP ID values from an incrementing counter. By pinging a router interface, one can monitor the growth of its counter over time, a technique used to infer aliases [8, 37, 55]. 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 ID velocity over time (*e.g.,* at peak time) to estimate the rate at which routers forward *user* traffic.

How can techniques be combined to best overcome biases and limitations and enable fine-grained mapping? The techniques offer different tradeoffs. Probing Google Public DNS enables per-user-prefix, per-service inferences on the granularity of the service's DNS record's TTL, but caches hide the number of queries within a TTL. Crawling root DNS logs provides global coverage and a direct measure of relative activity, but the measurements indicate activity of (unknown) clients 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 coverage. Usage of both Google Public DNS and Chromium may be skewed.

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 UMD), it may be possible to work with them to provide real-time access to logs. Since logs capture the address of the recursive resolver (rather than of the client), we either need to make simplifying assumptions to estimate relative client activity (*e.g.,* clients are in the same ⟨`region, AS`⟩ as their recursive resolver [17]) or deploy techniques to associate recursive resolvers with their clients (*e.g.,* embedding measurements of the associations in popular pages [43]). Such an association would enable joining of resolver-based techniques with client-based techniques. It is possible that (one-off or periodic) logs from organizations (*e.g.,* CDNs) can help understand biases in Chromium usage and/or Google Public DNS usage.

3.2 Where are services located & what are mappings from users to hosts?

3.2.1 Limitations of existing approaches. Many approaches to determining mappings from users to hosts are DNS-based, issuing queries from distributed measurement platforms [52, 57] or open recursive resolvers [32, 60], or by crowdsourcing [3, 44]. These approaches are limited by available vantage points because each only discovers the mapping based on its location and network conditions.

Others are custom techniques for particular services that do not need distributed vantage points, but fail to generalize to other services. Studies have emulated global vantage point coverage by issuing DNS queries using the DNS EDNS0 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 [13, 56]. However, not all services support ECS, and those that do may only reply to ECS queries from allowlisted resolvers [17]. ECS probing of Google Public DNS allows us to infer the users for all services that support ECS (§3.1.2), which suffices for a broad understanding of users but a limited understanding of services. Other work mapped Netflix [12] and Facebook [10, 11] servers by exhaustively trying queries based on patterns in their DNS naming scheme.

3.2.2 Possible measurement approaches.

Approach 1: Identifying infrastructure using TLS scans. TLS certificates validate the owner of a resource. With the recent dramatic increase in web encryption, we used TLS scans to identify the global serving infrastructure of large content providers and CDNs (Figure 1b) [25].

Approach 2: SNI scans for services. We propose using Internet-wide SNI (TLS + hostname) scans to uncover the footprint of popular services by identifying which CDN or cloud IP addresses have the services' TLS certificates.

Approach 3: Locating servers at fine granularities. The first two approaches uncover IP addresses of serving infrastructure hosting a particular service, but many use cases need to know the city/facility of serving infrastructure. Starting points may be client-centric geolocation [13] and constraint-based localization from in-facility vantage points [26, 47].

3.2.3 Open questions: How can we infer client-to-server mappings for services that do not rely on DNS redirection or do not support ECS?. DNS cache probing enables discovery of the client-to-server mapping for services that rely on DNS-based redirection and support ECS, but some services lack ECS support or use anycast [14] or customized URLs to direct a user to a site [5]. How to account for such services in the Internet traffic map remains an open question.

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 [17]. 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 [38]. 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 of edge computing platforms, such as Cloudflare's Workers [2], where CDN customers can execute custom code on CDN PoPs. This may enable use of techniques that infer per-PoP anycast catchments by probing out to the Internet [21].

It is extremely difficult to infer where individual clients are directed when the redirection happens via URLs customized to individual clients. 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. 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 intuition, it is critical to understand the efficacy of these caches. A community-driven project could host caches inside research networks/universities, to measure the cache hit rate under normal operation and during flash events.

3.3 What are routes between users/servers?

3.3.1 Limitations of existing approaches. Approaches to predict routes use measured topologies and AS relationships, coupled with common routing policies [35, 42]. This method only works if the actual routes exist in the measured topology, but available vantage points cannot uncover most peering links for large content providers [4, 48, 63]. When we tried to predict paths from RIPE Atlas probes to root DNS servers, more than half could not be predicted due to missing links.

3.3.2 Possible measurement approaches. Paths between users and popular services are becoming easier to predict, due to the evolving role of large cloud and content providers and Internet flattening. Flattening simplifies prediction, since most users have short, downhill paths to services [19], and simple heuristics accurately predict path lengths between users and the cloud [19]. Measuring out from cloud VMs uncovers most peering links between the cloud and users [7], and Reverse Traceroute can measure reverse paths [36].

3.3.3 Open question: Is it possible to predict missing links to complete the topology? While it is possible for researchers to measure paths to and from cloud providers [7, 36], these techniques require a vantage point within the cloud, so are not suitable for CDNs that do not support VMs running measurements. Is it possible to predict with high confidence which links exist, to feed into a path prediction algorithm? Increasingly many networks 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 publicly available information about networks, such as their peering policy, traffic profile, customer cone size [41], user activity (§3.1), and network type. With the assumption that networks with similar peering profiles are likely to peer with the same networks, one could formulate the problem as a recommendation system [45]—we rate the likelihood that networks (the shoppers) would want to peer with other networks (the items being recommended) and infer the existence of links if the recommendation is strong.

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 available (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 and Microsoft validated recent work uncovering their peers [7] and deployment footprints [25]. Finally, we hope the research community both uses and encourages others to use the Internet traffic map for weighting analysis and conducting Internet research.

Acknowledgements. We would like to thank the anonymous reviewers for their insightful comments. This paper has been partially funded by NSF grants CNS-1351100 and CNS-2028550, by the European Research Council (ERC) Starting Grant ResolutioNet (ERC-StG-679158), BMBF BIFOLD 01IS18025A and 01IS18037A, and the Spanish national project ODIO (grant PID2019-111429RB-C22).

REFERENCES

[1] Archipelago monitor locations, 2015. URL https://caida.org/projects/ark/locations.

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

[3] Bernhard Ager, Wolfgang Mühlbauer, Georgios Smaragdakis, and Steve Uhlig. Web content cartography. In *ACM IMC*, 2011.

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

[5] Zahaib Akhtar, Yun Seong Nam, Jessica Chen, Ramesh Govindan, Ethan Katz-Bassett, Sanjay Rao, Jibin Zhan, and Hui Zhang. Understanding video management planes. In *ACM IMC*, 2018.

[6] 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.

[7] Todd Arnold, Jia He, Weifan Jiang, Matt Calder, Italo Cunha, Vasileios Giotsas, and Ethan Katz-Bassett. Cloud provider connectivity in the flat Internet. In *ACM IMC*, 2020.

[8] Adam Bender, Rob Sherwood, and Neil Spring. Fixing Ally's growing pains with velocity modeling. In *ACM IMC*, 2008.

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

[10] 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.

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

[12] 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. In *ACM SIGCOMM CCR*, 2018.

[13] Matt Calder, Xun Fan, Zi Hu, Ethan Katz-Bassett, John Heidemann, and Ramesh Govindan. Mapping the expansion of Google's serving infrastructure. In *ACM IMC*, 2013.

[14] Matt Calder, Ashley Flavel, Ethan Katz-Bassett, Ratul Mahajan, and Jitendra Padhye. Analyzing the performance of an anycast CDN. In *ACM IMC*, 2015.

[15] 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 *USENIX NSDI*, 2018.

[16] Matt Calder, Xun Fan, and Liang Zhu. A cloud provider's view of EDNS Client-Subnet adoption. In *IEEE TMA*, 2019.

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

[18] Yi-Ching Chiu, Brandon Schlinker, Abhishek Balaji Radhakrishnan, Ethan Katz-Bassett, and Ramesh Govindan. Are We One Hop Away from a Better Internet? In *Proc. ACM IMC*, 2015.

[19] Yi-Ching Chiu, Brandon Schlinker, Abhishek Balaji Radhakrishnan, Ethan Katz-Bassett, and Ramesh Govindan. Are we one hop away from a better Internet? In *ACM IMC*, 2015.

[20] David R Choffnes and Fabián E Bustamante. Taming the torrent: a practical approach to reducing cross-ISP traffic in peer-to-peer systems. In *ACM SIGCOMM*, 2008.

[21] 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 *ACM IMC*, 2017.

[22] Ramakrishnan Durairajan, Paul Barford, Joel Sommers, and Walter Willinger. Intertubes: A study of the US long-haul fiber-optic infrastructure. In *ACM SIGCOMM*, 2015.

[23] Rodérick Fanou, Francisco Valera, and Amogh Dhamdhere. Investigating the causes of congestion on the African IXP substrate. In *ACM IMC*, 2017.

[24] Petros Gigis, Vasileios Kotronis, Emile Aben, Stephen D Strowes, and Xenofontas Dimitropoulos. Characterizing user-to-user connectivity with RIPE Atlas. In *ACM ANRW*, 2017.

[25] 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.

[26] Vasileios Giotsas, Georgios Smaragdakis, Bradley Huffaker, Matthew Luckie, and KC Claffy. Mapping peering interconnections to a facility. In *ACM CoNEXT*, 2015.

[27] Ramesh Govindan and Hongsuda Tangmunarunkit. Heuristics for Internet map discovery. In *IEEE INFOCOM*, 2000.

[28] Enrico Gregori, Alessandro Improta, and Luca Sani. On the African peering connectivity revealable via BGP route collectors. In *EAI AFRICOMM*, 2017.

[29] 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 *PAM*, 2014.

[30] Gonca Gürsun and Mark Crovella. On traffic matrix completion in the Internet. In *ACM IMC*, 2012.

[31] Gonca Gürsun, Natali Ruchansky, Evimaria Terzi, and Mark Crovella. Inferring visibility: Who's (not) talking to whom? In *ACM SIGCOMM*, 2012.

[32] Cheng Huang, Angela Wang, Jin Li, and Keith W Ross. Measuring and evaluating large-scale CDNs. In *ACM IMC*, 2008.

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

[34] Weifan Jiang, Tao Luo, Thomas Koch, Yunfan Zhang, Ethan Katz-Bassett, and Matt Calder. Towards identifying networks with Internet clients using public data. In *ACM IMC*, 2021.

[35] Yuchen Jin, Colin Scott, Amogh Dhamdhere, Vasileios Giotsas, Arvind Krishnamurthy, and Scott Shenker. Stable and practical AS relationship inference with ProbLink. In *USENIX NSDI*, 2019.

[36] 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 *USENIX NSDI*, 2010.

[37] Ken Keys, Young Hyun, Matthew Luckie, and Kim Claffy. Internet-scale IPv4 alias resolution with MIDAR. In *IEEE/ACM ToN*, 2012.

[38] 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.

[39] Vasileios Kotronis, George Nomikos, Lefteris Manassakis, Dimitris Mavrommatis, and Xenofontas Dimitropoulos. Shortcuts through colocation facilities. In *ACM IMC*, 2017.

[40] Craig Labovitz, Scott Iekel-Johnson, Danny McPherson, Jon Oberheide, and Farnam Jahanian. Internet inter-domain traffic. In *ACM SIGCOMM*, 2010.

[41] Matthew Luckie, Bradley Huffaker, Amogh Dhamdhere, Vasileios Giotsas, and KC Claffy. AS relationships, customer cones, and validation. In *ACM IMC*, 2013.

[42] Harsha V Madhyastha, Tomas Isdal, Michael Piatek, Colin Dixon, Thomas Anderson, Arvind Krishnamurthy, and Arun Venkataramani. iplane: An information plane for distributed services. In *USENIX OSDI*, 2006.

[43] 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 ATC*, 2002.

[44] Srdjan Matic, Gareth Tyson, and Gianluca Stringhini. Pythia: a framework for the automated analysis of web hosting environments. In *ACM WWW*, 2019.

[45] Prem Melville and Vikas Sindhwani. Recommender systems. *Encyclopedia of machine learning*, 1:829–838, 2010.

[46] 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.

[47] 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 *ACM IMC*, 2018.

[48] Ricardo Oliveira, Dan Pei, Walter Willinger, Beichuan Zhang, and Lixia Zhang. The (in) completeness of the observed Internet AS-level structure. In *IEEE/ACM ToN*, 2009.

[49] Eduardo E P. Pujol, Will Scott, Eric Wustrow, and J Alex Halderman. Initial measurements of the Cuban street network. In *ACM IMC*, 2017.

[50] John P Rula and Fabian E Bustamante. Behind the curtain: Cellular DNS and content replica selection. In *ACM IMC*, 2014.

[51] John P Rula, Fabián E Bustamante, and Moritz Steiner. Cell spotting: studying the role of cellular networks in the Internet. In *ACM IMC*, 2017.

[52] 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 *USENIX NSDI*, 2013.

[53] Mario A Sanchez, Fabian E Bustamante, Balachander Krishnamurthy, Walter Willinger, Georgios Smaragdakis, and Jeffrey Erman. Interdomain traffic estimation for the outsider. In *ACM IMC*, 2014.

[54] 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 *ACM IMC*, 2018.

[55] Neil Spring, Ratul Mahajan, David Wetherall, and Thomas Anderson. Measuring ISP topologies with Rocketfuel. In *IEEE/ACM ToN*, 2004.

[56] Florian Streibelt, Jan Böttger, Nikolaos Chatzis, Georgios Smaragdakis, and Anja Feldmann. Exploring EDNS-Client-Subnet adopters in your free time. In *ACM IMC*, 2013.

[57] Ao-Jan Su, David R Choffnes, Aleksandar Kuzmanovic, and Fabián E Bustamante. Drafting behind Akamai (travelocity-based detouring). In *ACM SIGCOMM*, 2006.

[58] 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 *ACM HotNets*, 2014.

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

[60] Sipat Triukose, Zhihua Wen, and Michael Rabinovich. Measuring a commercial content delivery network. In *ACM WWW*, 2011.

[61] Narseo Vallina-Rodriguez, Srikanth Sundaresan, Christian Kreibich, Nicholas Weaver, and Vern Paxson. Beyond the radio: Illuminating the higher layers of mobile networks. In *ACM MobiSys*, 2015.

[62] Kevin Vermeulen, Justin P Rohrer, Robert Beverly, Olivier Fourmaux, and Timur Friedman. Diamond-Miner: Comprehensive discovery of the Internet's topology diamonds. In *USENIX NSDI*, 2020.

[63] Florian Wohlfart, Nikolaos Chatzis, Caglar Dabanoglu, Georg Carle, and Walter Willinger. Leveraging interconnections for performance: the serving infrastructure of a large CDN. In *ACM SIGCOMM*, 2018.

[64] Bahador Yeganeh, Ramakrishnan Durairajan, Reza Rejaie, and Walter Willinger. How cloud traffic goes hiding: A study of Amazon's peering fabric. In *ACM IMC*, 2019.

[65] 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 *PAM*, 2014.