Operating a DNS-based Active Internet Observatory

Oliver Hohlfeld
RWTH Aachen University

ABSTRACT

The Internet is subject to constant evolution. Its improvement requires understanding its current properties, a perspective provided by measurement studies. A key challenge in broadly studying evolution is to *i)* cover multiple protocols *ii)* with longitudinal measurements. In this poster, we present an Internet observatory that performs active measurements of multiple protocols (e.g., DNS, HTTP2, QUIC) regularly since 2016. Its measurements cover both *i)* the entire IPv4 address space and *ii)* $>$ 50% of the domain name space to provide a new perspective on Internet evolution. The goal of this poster is to present its extensible architecture and capabilities, thereby aiming to foster collaboration.

CCS CONCEPTS

• **Networks → Naming and addressing**;

KEYWORDS

DNS, CDN, Cloud, HTTP2, QUIC, active measurements

ACM Reference Format:
Oliver Hohlfeld. 2018. Operating a DNS-based Active Internet Observatory. In *SIGCOMM Posters and Demos '18: ACM SIGCOMM 2018 Conference Posters and Demos, August 20–25, 2018, Budapest, Hungary.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3234200.3234239

1 THE NEED FOR MEASURING DOMAINS

Since the Internet is an entirely man-made system, it could be assumed that its properties and usage is fully understood. Yet, this is not the case. Well understood are the complex building blocks ("DNA") of the Internet, such as its protocols (e.g., TCP/IP), technologies (e.g., Wifi), and applications (e.g., the World Wide Web). Due to its size (e.g., ≈ 1B end-systems in $>$ 70k ASes) and the many individual components (e.g., 6K IETF documents), the Internet can be considered one of the largest and most complex man-made systems. Optimizing it requires understanding its properties and evolution, a quest that has motivated countless measurement studies.

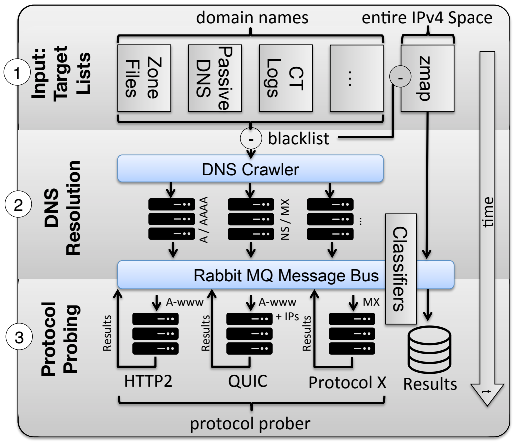


Figure 1: Observatory Architecture

Yet, *typical* active measurement studies study *single* protocols over *short* time frames. Internet operation often involves multiple protocols (e.g., TLS security can be improved by DNS CAA records [12]), which is why a multi protocol focus can enrich the understanding of Internet evolution. Besides time, achieving a coverage of the probed *addresses* is a further challenge. Only recently have regular scans of the entire IPv4 address space become feasible with the release of `zmap` [1, 8] in 2013. Since then, Censys [2, 7] conducts regular multi-protocol IPv4 scans—a valuable resource for many research projects. However, Internet evolution renders IP-only scans incapable of covering services by infrastructures that require requests to provide valid host names (e.g., CDNs). To account for this new reality, an Internet observatory must cover both the *i)* domain name and the *ii)* IP address space.

However, obtaining domain name lists is notoriously hard since names cannot be (efficiently) enumerated and many registries do not provide lists (zone files)—if, they often require NDAs. Further, DNS resolving large domain lists alone is complex. This is demonstrated by the OpenINTEL project [13] that resolves ≈ 60% of the domain name space daily, enabled by a substantial infrastructure of many workers. No further protocols are probed—an additional complexity challenge.

2 OBSERVATORY ARCHITECTURE

Goal. The ambitious goal of our Internet Observatory is to provide a new application/protocol-level perspective on Internet evolution by *regularly* (daily / weekly) probing both

SIGCOMM Posters and Demos '18, August 20–25, 2018, Budapest, Hungary

ACM ISBN 978-1-4503-5915-3/18/08…$15.00
https://doi.org/10.1145/3234200.3234239

i) the *entire IPv4 address* space and *ii) more than 50% of the domain name space* for *multiple protocols*. The observatory is based on an extensible architecture enabling scans for new protocols to be added easily. Currently, it probes domains/IPs for HTTP2 [14], QUIC [11], TLS support, and TCP IW settings [10]—besides multiple DNS records per domain.

We next describe its architecture shown in Figure 1.

Target lists ① The first challenge involves achieving a large domain name coverage. To tackle this, we obtain (daily) zone files for .se, .nu, .gov, .com, .net, .org, .name, .fi, and about 1k newgTLDs (e.g., .london) through various agreements (e.g., as OpenINTEL [13]). As of May, these TLDs account for 195M domains (58.7% of the overall domain name space).

Beyond current efforts such as OpenINTEL, we argue to increase the domain coverage by complementing zone files with additional data sources. This way, we feed a passive DNS feed obtained from a campus network in our system. We further incorporate a Certificate Transparency (CT) Log live feed from which we extract domain names.

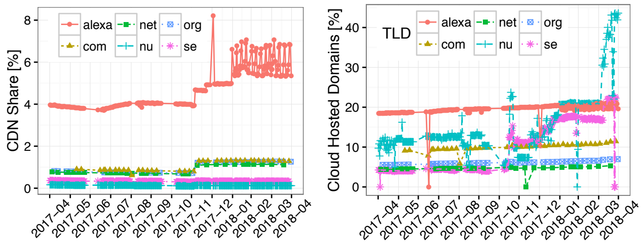
For example, .com as the largest TLD currently accounts for 132M domains of which we find 44M to be TLS-capable. Using historic up to recent CT Log data, we could extract 29.3M com domains. This approach, enables to identify domains for which we have no zone file access—e.g., 26.4M additional ccTLD domains—highlighting its value.

Before scanning, we drop blacklisted domains/ IPs.

DNS resolution ② We query multiple DNS resource records for every domain (i.e., A, AAAA for both the domain and an www prefix, ANY, SOA, CAA, MX, NS, including A/AAAA for every MX/NS record). To perform the DNS resolution, we distribute it to a set of worker nodes, similar to OpenINTEL [13]. To scale the system to our hardware resources, we resolve small TLDs daily and large ones (e.g., .com/net/org) only weekly. This resolution can be scaled by adding hardware.

Protocol probing ③ They key addition of our observatory is an extensible multi protocol probing architecture. It uses a RabbitMQ [3] message broker to distribute workload to subscribed workers via per-protocol measurement queues. That is, *i)* IP list from zmap address space scans and *ii)* DNS resolution of domain lists are fed into queues identified by a routing key (e.g., "(A-www, com)" indicates A lookups of www prefixed .com domains). The routing keys enable workers to select the workload (e.g., to only scan selected TLDs). Scan results are sent back to the message queue and are written to disk by a dedicated worker. This way, many workers can work in parallel and the number of workers (per measurement) can be scaled while the measurement is running. Workers can also be in remote locations to perform multi-vantage point measurements (currently not used). In addition, classifier scripts enrich the results with IP geolocation, ASN data from current RIBs, and cloud or CDN usage.

3 CASE STUDIES



(a) CDN usage on www. **(b) Cloud usage on www.**

Figure 2: Domains' `www.` hosts infrastructure analysis

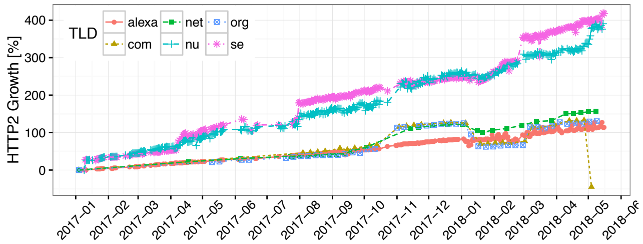


Figure 3: HTTP2 Adoption (data: http2.netray.io)

We demonstrate the measurement abilities of our observatory with two simple longitudinal case studies run from our vantage point in Aachen. First, we analyze hosting *infrastructures* by analyzing information returned to A record queries of the domains www host. By matching CNAME records against pattern of 77 CDNs [9], we show the prevalence of CDN full site hosting in Figure 2a. We observe that domains in the Alexa Top 1M list have a higher CDN adoption than domains in the shown TLD zones. Yet, only few domains point their www record directly to a CDN. Note that CDNs are often used for assets (e.g., images.domain.tld), which this analysis omits but which can be added by analyzing the HTML payload from our TLS scanners. In contrast, more domains point their www to Google [5], Amazon [4], or Azure [6] cloud hosts as identified by matching public cloud IP prefixes in Figure 2b. To study the *evolution* of a new Internet protocol, we probe domains for HTTP2 and show the HTTP2 adoption in Figure 3. Beyond simple studies, our data set enables new perspectives by correlating properties of multiple protocols in future work; a discussion this poster intends to open.

4 CONCLUSION

We created a new active measurement infrastructure to study Internet evolution with large-scale, DNS-based multi-protocol measurements. We thereby follow the ambitious goal to cover a large domain name space with longitudinal measurements. The goal of this poster is to foster collaboration with other researchers to analyze the dataset. As an overview, we have created a web page showing current statistics and further information about our studies: netray.io.

Acknowledgements. Funded by the Excellence Initiative of the German federal and state governments.

REFERENCES

[1] [n. d.]. https://github.com/zmap. ([n. d.]).

[2] [n. d.]. https://censys.io/. ([n. d.]).

[3] [n. d.]. https://www.rabbitmq.com. ([n. d.]).

[4] [n. d.]. AWS IP Address Ranges. https://docs.aws.amazon.com/general/latest/gr/aws-ip-ranges.html. ([n. d.]).

[5] [n. d.]. Google IP address ranges. https://kx.cloudingenium.com/cloud/google-cloud/google-ip-address-ranges/. ([n. d.]).

[6] [n. d.]. Microsoft Azure Datacenter IP Ranges. https://www.microsoft.com/en-us/download/details.aspx?id=41653. ([n. d.]).

[7] Zakir Durumeric, David Adrian, Ariana Mirian, Michael Bailey, and J. Alex Halderman. 2015. A Search Engine Backed by Internet-Wide Scanning. In *22nd ACM Conference on Computer and Communications Security*.

[8] Zakir Durumeric, Eric Wustrow, and J. Alex Halderman. 2013. ZMap: Fast Internet-wide Scanning and Its Security Applica tions. In *Proceedings of USENIX Conference on Security*.

[9] Google. 2018. WebPagetest CDN domains. github.com/WPO-Foundation/webpagetest/blob/master/agent/wpthook/cdn.h. (2018).

[10] Jan Rüth, Christian Bormann, and Oliver Hohlfeld. 2017. Large-Scale Scanning of TCP’s Initial Window. In *ACM Internet Measurement Conference*.

[11] Jan Rüth, Ingmar Poese, Christoph Dietzel, and Oliver Hohlfeld. 2018. A First Look at QUIC in the Wild. In *Passive and Active Measurement Conference*.

[12] Quirin Scheitle, Taejoong Chung, Jens Hiller, Oliver Gasser, Johannes Naab, Roland van Rijswijk-Deij, Oliver Hohlfeld, Ralph Holz, Dave Choffnes, Alan Mislove, and Georg Carle. 2018. A First Look at Certification Authority Authorization (CAA). *ACM SIGCOMM Computer Communication Review* 48, 2 (May 2018), 10–23.

[13] Roland van Rijswijk-Deij, Mattijs Jonker, Anna Sperotto, and Aiko Pras. 2016. A High-Performance, Scalable Infrastructure for Active DNS Measurements. *IEEE JSAC* 34, 7 (2016), 149–160.

[14] Torsten Zimmermann, Jan Rüth, Benedikt Wolters, and Oliver Hohlfeld. 2017. How HTTP/2 Pushes the Web: An Empirical Study of HTTP/2 Server Push. In *IFIP Networking Conference*.