Institutional Privacy Risks in Sharing DNS Data

Basileal Imana
University of Southern California
Los Angeles, CA, USA
imana@usc.edu

Aleksandra Korolova
University of Southern California
Los Angeles, CA, USA
korolova@usc.edu

John Heidemann
USC/Information Science Institute
Los Angeles, CA, USA
johnh@isi.edu

ABSTRACT

The Domain Name System (DNS) is used in every website visit and e-mail transmission, so privacy is an obvious concern. In DNS, users ask recursive resolvers (or "recursives") to make queries on their behalf. Prior analysis of DNS privacy focused on privacy risks to individual end-users, mainly in traffic between users and recursives. Recursives cache and aggregate traffic for many users, factors that are commonly assumed to protect end-user privacy above the recursive. We document *institutional privacy* as a new risk posed by DNS data collected at authoritative servers, even after caching and aggregation by DNS recursives. We are the first to demonstrate this risk by looking at leaks of e-mail exchanges which show communications patterns, and leaks from accessing sensitive websites, both of which can harm an institution's public image. We define a methodology to identify queries from institutions and identify leaks. We show the current practices of prefix-preserving anonymization of IP addresses and aggregation above the recursive are not sufficient to protect institutional privacy, suggesting the need for novel approaches. We demonstrate this claim by applying our methodology to real-world traffic from DNS servers that use partial prefix-preserving anonymization. Our work prompts additional privacy considerations for institutions that run their own resolvers and authoritative server operators that log and share DNS data.

CCS CONCEPTS

• **Security and privacy**; • **Networks**;

ACM Reference Format:
Basileal Imana, Aleksandra Korolova, and John Heidemann. 2021. Institutional Privacy Risks in Sharing DNS Data. In *Applied Networking Research Workshop (ANRW '21), July 24–30, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3472305.3472324

ANRW '21, July 24–30, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8618-0/21/07.
https://doi.org/10.1145/3472305.3472324

1 INTRODUCTION

The Domain Name System (DNS) is a critical part of the Internet's infrastructure, which computers rely on to find resources or services on the Internet. It is used to map a human readable domain name, such as example.com, to an IP address, and to provide other kinds of lightweight services, such as a blacklist of IPs that send spam [31].

Since almost every activity on the Internet starts with a DNS query, collecting and analyzing DNS data is useful for research that aims to improve the stability and security of the Internet, such as understanding Internet trends [14], studying defense mechanisms against DDoS attacks [36], detecting malicious domains [10], and preventing data exfiltration [11]. Several entities make DNS data available for research and operational use. DNS-OARC [19] is one such entity that runs an initiative to collect traces from various DNS authoritative servers, and makes the data available for researchers and member institutions. Commercial services also collect and make DNS data available to help fight cybercrime [5].

As DNS queries often represent people's actions such as browsing a website or sending an email, numerous studies have been conducted to understand the privacy implications of sharing DNS data [12, 24, 26]. The initial focus of prior work has been on privacy of *individual* end-users, and protecting the *user-to-recursive* channel. Several new protocols have been proposed and deployed to improve privacy for end-users, mainly by encrypting user-to-recursive resolver traffic [21, 25, 28, 45]. With the success of DNS-over-TLS and -HTTP, the DNS privacy working group in IETF has begun looking at confidentiality between resolvers and authoritatives [32]. However, even with encryption of in-transit traffic, sharing data logged at servers remains a privacy risk.

While DNS data privacy threats to end-users are well understood, the potential impact of shared DNS data to the privacy of *institutions* has not been closely studied. To our knowledge, prior studies have also not looked at the implications for institutional privacy of prefix-preserving IP-address anonymization. Therefore, we believe there is merit in documenting institutional privacy risks that could be present in partially anonymized authoritative traffic log. We enumerate these risks, showing logs of aggregate traffic above the recursive can also reveal information about institutions.

Our first contribution is to define *institutional privacy*, a new aspect of privacy threat posed by sharing DNS data (§3),

Table 1: Institutional privacy threats we study

Event	Query	Implication to institution
Send/receive email	MX or DNSBL	Indicates relation between sender and recipient
Browse sensitive site	A or AAAA	Embarrassment or leak of demographics information

and to show that it can be a threat even when an observer only sees data above the recursive. We define institutional privacy as confidentiality of digital footprints of an institution's internal activities by its personnel. DNS is one place where activities such as sending an email or accessing a website while on a company's network could leave a trace. We show that DNS queries that originate from an institution can be used to infer such activities. Large organizations such as government entities, schools and corporations that operate their own Autonomous System (AS) and DNS resolvers for their enterprise networks are particularly susceptible to such leaks. The risk here is that the IP address of a resolver can be used to trace a query back to the AS the address belongs to.

Our second contribution is to give a methodology for *identifying DNS queries of an institution*, and *finding queries that leak information* (§4). We show that current practices of prefix-preserving anonymization [44] of IP addresses and aggregation of data above the recursive is not sufficient to protect institutional privacy by giving a methodology for identifying privacy-violating queries in anonymized DNS data. We focus on queries that leak communication patterns, specifically mail-exchange (MX) [34] and DNS blacklist (DNSBL) [31] queries; and A or AAAA queries for domains of sensitive websites, such as adult, illegal and gender-specific sites, which we identify using a website content categorization API. We show how we identify the queries, and what information leaks through the fields of queries of each type.

Our third contribution is to demonstrate the *privacy risks occur in real-world data* (§5). We study queries from more than 60 institutions (≈200 ASes since some own more than one AS) to show that many are at risk of their privacy being violated. We argue the seriousness of the risk to institutions by analyzing queries collected at a root DNS server over a short time duration (two 7-day periods), and giving concrete examples of leaks we find in the small set of ASes we study (200 out of more than 90,000 allocated ASes [4]), suggesting that a dedicated adversary can find many more.

2 DNS BACKGROUND

The Domain Name System [34] is a globally distributed database designed for mapping Internet domain names to information such as IP addresses. In addition to mapping names to IP addresses, DNS' lightweight design has prompted its use for other services such as spam defense [27].

At a high level, a DNS name resolution process involves three actors. A *stub resolver* runs on an end-user's machine and handles all queries for the operating system. The stub resolver contacts a *recursive resolver* to answer a query. A recursive resolver, often shared by many stubs, iteratively contacts one or more *authoritative servers* responsible for resolving the domain name the stub requested. A recursive resolver handles requests for multiple stubs, so it *aggregates* these requests.

We focus on DNS traffic *above* the recursive (we discuss why in §3.2). We talk about *below* and *above the recursive* to indicate traffic from the stub to recursive ("below") and the recursive to authoritative ("above"). Observations at authoritative servers (say, .com or the root, ".") see traffic from many recursives, but may be subject to query name minimization. In our analysis, we study queries to a root server. The root servers' system handles ≈125 billion queries per day (as of November 2019 [39, 40]).

To improve privacy in DNS, *Query name minimization* was standardized in 2016 [13]. Previously, iterative queries to each authoritative included the complete desired name. With query name minimization, each authoritative is asked only about one component of the full name.

3 THREAT MODEL

3.1 Institutional Privacy

We define institutional privacy as confidentiality of digital footprints of an institution's operations and activities of its personnel as a whole. Institutions may care about their privacy for many reasons. They may want to protect their reputation, their business advantage over competitors, or business secrets such as who they communicate with, upcoming layoffs, or potential mergers or acquisitions.

We demonstrate the privacy threats to institutions posed by above the recursive DNS data by studying two classes of leaks. Table 1 gives a summary of the threats. The first is identifying *when an institution sends or receives an email.* This may indicate which entities an institution is corresponding with; which, for example, could reveal a business relationship or correspondence with a controversial institution. Exchanging emails can result in two types of queries: MX queries, made by a mail server before sending an email, and DNSBL queries, which are used by the recipient's mail server to check whether the sender's IP address is blacklisted for sending spam. We discuss the details on how we identify these queries, and what leaks through the fields of the queries in §4.2.

The second risk is identifying *when privacy sensitive websites are accessed.* Accessing such sites can affect the reputation or public image of an institution (such as adult or illegal sites), or reveal employee demographics that otherwise may not be publicly known (for example, religious or

Table 2: US institutions with AS(es) that we study

Type	Examples	# of institutions
UC Schools	UCLA, UCSD, UCSF	10
Ivies/Little Ivies	Cornell, Dartmouth	24
S&P 500	Facebook, Apple	10
US Airlines	Southwest, Frontier	13
Government	DHS, US Navy	9

gender-specific sites). We identify sensitive websites by using a content categorization tool which we discuss in §4.3.

Due to possible malware infection or spamming, the presence of a DNS record is not a sure indicator that a sensitive site was accessed or a certain communication has happened. Thus, the *possibility* of malware can be used to suggest deniability. However, most enterprises would find both choices undesirable, to chose between disclosing of sensitive accesses or gaining a reputation for weak security. Furthermore, for email communication, one can check bidirectional relationships before making conclusions.

3.2 Adversary Model and Assumptions

We focus on a threat model where an adversary is *external* to the institution and is situated at an authoritative server. An adversary that can eavesdrop on traffic above the recursive can also see the same data. We choose this model because data collected high in the DNS hierarchy is generally assumed to not violate privacy due to caching and aggregation [12, 42], and is often shared for research purposes [19].

We consider an adversary whose goal is to associate the source IP address and the domain name fields of a DNS query to institutions. The adversary can do so if the way the institution operates its enterprise network meets two conditions, both typical for large institutions. The institution *must run its own recursive resolver.* and *must route traffic from its own Autonomous System (AS).* This allows an adversary to identify the IP address ranges owned by the institution, and look for queries addresses is in one of these ranges.

Many large institutions meet both of the above conditions, making them susceptible to the privacy threats we discuss in this paper. In one of the week-long datasets we study, there is at least one query from about 54% of ASes (out of 94,866 total [4]). From these ASes, we handpicked institutions that meet the conditions and represent diverse industry sectors, including public and private universities, big corporations and government/law-enforcement agencies. See Table 2 for the categories of the 66 institutions whose queries we study. Among these, universities are unique in that queries that originate from their network include activities of not just employees, but also students. When picking institutions, we excluded institutions such as ISPs and companies that provide hosting services, such as Google, since queries from their ASes may contain queries from their customers as well.

Real-world actors that fit our adversary model include: authoritative DNS server operators that are able to observe and log DNS requests, entities that have access to shared by DNS operators to support research efforts (Day-In-The-Life of the Internet initiative by DNS-OARC is an example), and government or state-level entities that have the means and resources to eavesdrop on DNS transactions [9, 12]. We identify these potential adversaries because they have the means to access DNS data originating from institutions; however, we do not claim that they do so currently except in the case of state-level actors.

4 METHODOLOGY: FINDING LEAKS

4.1 Identifying Queries of Institutions

We first show how we identify queries associated with an institution. As discussed in §3.2, our methodology for finding leaks applies to large institutions that run their own AS and their own recursive resolver for their enterprise networks.

We associate a DNS query with an institution by looking at two fields: the source IP address and the domain name. We map the source IP to an institution using publicly available IPtoASN data [17]. (This data is from *whois* data by Regional Internet Registries.) We can identify institutions even if IP addresses use partial prefix-preserving anonymization [44], provided it passes the top 24 bits, since organizations that do their own routing typically have at least a IPv4 /24 prefix.

We associate a domain name to an institution by checking the owner of the domain or visiting the associated website. We assume queries pass full domains and do not use query name minimization, whose adoption has steadily increased up to 50% since its standardization in 2016 [2, 15, 43].

4.2 Identifying Email Communication

Once we identify queries related to an institution, we search for MX and DNSBL queries that indicate email exchange.

Identifying MX Queries: We first demonstrate how an adversary can use MX queries to identify when an email is sent. We do so by exploiting the fact that, before sending an email, a sender's mail server first has to query the MX record of the recipient's domain name. Examining MX records is not new, but its use in studying institutional privacy is.

We identify all MX queries by examining query types. We filter out queries to non-existent domains by checking for NXDOMAIN response codes. We then identify the source and destination of each query using the methodology given in §4.1. The source IP of an MX query identifies the email sender, and the domain identifies the email recipient. Table 3 summarizes what information leaks through an MX query.

Identifying DNSBL Queries: We next show how an adversary can learn when e-mail is received based on DNSBL queries related to anti-spam protection services.

Anti-spam services use Domain Name System Blacklists (DNSBL) [31] to list IP addresses that have a history of sending spam. Today, many mail servers use DNSBLs to vet the sender before accepting incoming mail. When a new message arrives, the mail recipient encodes the IP address of the sender in a DNS query to a DNSBL service. The DNS reply indicates whether the address is blacklisted or not. Dozens of organizations operate DNSBLs [1].

We use regular expressions to identify the structure of DNSBL queries [31] and extract the email sender's IP address. Similar to our process for MX queries, we filter out DNSBL queries that get unsuccessful responses. We then find the institutions associated with the source and domain of each DNSBL query using our methodology given in §4.1. A DNSBL query's source IP identifies the email recipient, and the IP address in the domain identifies the sender. Table 3 summarizes what information leaks through a DNSBL query.

4.3 Identifying Sensitive Domains

Finally, we show how we identify queries that indicate access to sensitive websites using a publicly available categorization API with a pay-for-access model called Webshrinker [6]. The API, which uses a content taxonomy defined by Interactive Advertising Bureau (IAB) [3], categorizes domains based on the content hosted there. Given a domain, the API returns a list of potential categories for the domain along with a confidence score for each category. In our experiments, we pick the category with the highest score (we also explored using all categories returned and their scores, finding similar results). As shown in Table 4, Webshrinker has dedicated categories for sensitive contents such as adult, illegal, religious and gender-specific sites, which are the categories we consider. After we identify all queries of type A and AAAA from an institution (§4.1), we use Webshrinker to categorize each query, and count the number of queries per category.

5 LEAKS IN REAL-WORLD DNS DATA

We apply our methodology to real-word root DNS data anonymized using a prefix-preserving method, analyze queries from institutions listed in Table 2, and give examples of leaks.

5.1 Dataset

We use a dataset that contains DNS requests made to B-root server during the week of Jan 9-15, 2019 [7]. We verify our observations on the week of Jan 17-23, 2019 (but omit the results since they are similar). In the dataset, all source IPs are anonymized using a prefix-preserving method with Cryptopan [37]. The top-24 bits unchanged and the last 8 bits passed through a cryptographic function.

The data is currently available to researchers at no cost, but with a legal agreement that prohibits deanonymization. We carried out our work with the permission of the data owners with the agreement that we would not identify specific combinations of unique query sources and destinations that indicate relationships not already publicly known.

5.2 Communication Between Institutions

First, we find that a significant volume of email-related traffic that may leak information. Figure 1 shows that several million DNSBL and MX queries are made each day. The dip in DNSBL on 2019-01-12 and -13 correspond to lower weekend traffic. We next closely examine a subset of these queries.

MX queries: To put the millions of email-related queries into context, we break down the volumes of email-related queries by industry sectors. Figure 4A shows the number of MX queries from the institutions listed in Table 2. We cannot directly compare sectors because they are of different sizes, but we see that, due to the substantial query volume in each sector except airlines, a variety of diverse organizations are at risk of leaking information about their email traffic.

If an adversary's goal is to find incriminating evidence of association between specific institutions, an examination of second-level domains can provide more detail about who is exchanging mail with whom. We examine second-level domains owned by four institutions: U.S. Immigration and Customs Enforcement (ICE), Palantir Technologies, Planned Parenthood, and the National Rifle Association. Figure 2 shows the number of queries made to each institution's domain. Although these are all legal organizations, they are all politically active and a finding of a previously unknown association with them may raise public relations concerns.

To demonstrate that uncovering such associations is a realistic possibility, we report on a few examples of findings we made. As discussed in §5.1, for ethical reasons, we only report on institutions whose associations to others are publicly known. We find an IP address owned by U.S. Department of Justice making several MX queries to Palantir (palantir.com). This observation is inline with the known fact that Palantir works with government agencies, which has raised public relations controversy [33]. We find a query made to ICE (ice.dhs.gov) by a school board in Jefferson Parish, LA that has a history of aiding ICE with deportations [38]. These examples illustrate specific relationships that are visible in the data; we expect an adversary could find many others.

DNSBL queries: We next show DNSBL queries from a variety of institutions leak to root servers, revealing with whom they exchange emails. We give the number of DNSBL queries we find that are related to the institutions listed in Table 2. We see there are many fewer DNSBL queries that leak compared to MX queries, but there are still a fair number. Figure 4C shows counts for institutions associated with the source IP of a query (indicates email receiver) and Figure 4B shows counts for institutions associated with the IP address embedded in the domain name (indicates the email sender).

Table 3: What MX and DNSBL query fields leak

	MX query			DNSBL query		
Field	Value	Map to	Indicates	Value	Map to	Indicates
Domain	company.com	Domain owner	Recipient	1.3.0.192.bl.example.com.	192.0.3.1 ⇒ AS name	Sender
Source IP	192.0.2.1	AS name	Sender	192.0.2.1	192.0.2.1 ⇒ AS name	Recipient

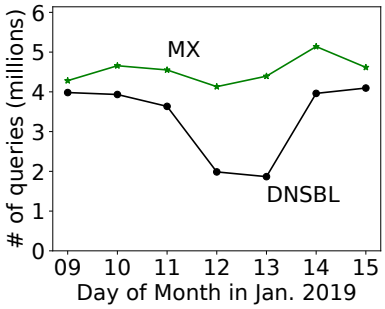

Figure 1: Total number of DNSBL and MX queries in dataset

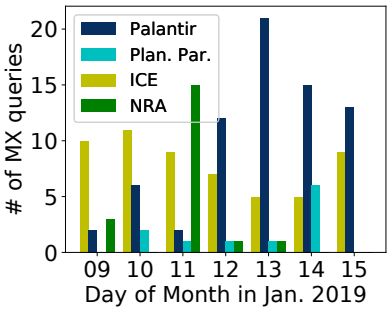

Figure 2: MX queries for specific institutions

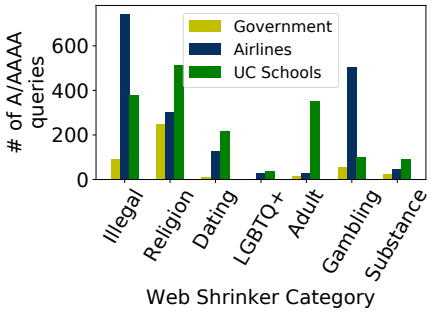

Figure 3: Query volume by category

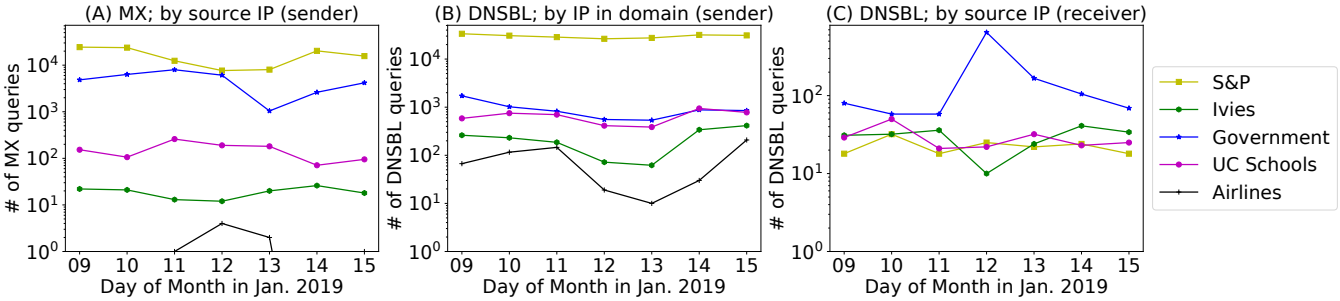

Figure 4: Leaks through MX and DNSBL (sender/receiver institutions identified using source IP or IP in domain)

Table 4: List of IAB [3] content categories studied

Category (IAB ID)	Example	Ramification
Illegal (IAB26-1)	123movies.best	embarrassing
Religion (IAB23)	jw.org	demographics
Dating (IAB14-1)	deaf.dating	demographics
LGBTQ+ (IAB14-3)	lgbt.foundation	demographics
Adult (IAB25-3)	pornhub.com	embarrassing
Gambling (IAB9-WS1)	topcasino.ml	embarrassing
Addiction (IAB7-42)	frn.rehab	embarrassing

5.3 Access to Sensitive Domains

We next demonstrate the possibility of identifying sensitive websites' access from within an institution by applying our website categorization method (§4.3). Figure 3 shows query volumes to sensitive categories made by government offices, schools and airlines. Such findings can be damaging to an institution's public image (e.g., adult and illegal categories), or leak demographic information (e.g., LGBTQ+ and religion categories). Although caching at the recursive may attenuate the true interest level for these categories (and may necessitate studying a longer duration of traffic and model caching effects to make inferences based on the query volume), it does not prevent them from leaking to the root.

These results used queries to a single root server, for just one week, so our work provides only a glimpse of the potential privacy risk for institutions from sharing such data. In practice, a determined adversary could launch a richer set of attacks beyond inferring demographics or embarrassing information. For example, an attacker may be able to analyze queries over a longer duration to establish long-term trends of sites accessed by institutions, and detect when deviations from typical patterns occur. A large-scale longitudinal study that quantifies such risks is a potential area for future work.

6 RECOMMENDATIONS

As discussed in the introduction (§1), sharing DNS data serves an important purpose in enabling Internet research to improve security and performance. However, as our work

demonstrates, such sharing poses new risks to the privacy of institutions, in addition to previously known risks to individuals. We next discuss recommendations for institutions, and for service providers that collect and share DNS data.

Actions institutions can take: The currently available best way for institutions to reduce information leakage is to run their own resolver, and deploy query name minimization [13], which mitigates the privacy leakage problem by providing domain name components only to the authoritative resolver expected to handle them. Prior work has shown an increasing adoption of query name minimization [2, 15, 43], and we hope our findings encourage a faster adoption by institutions who run their own resolvers.

LocalRoot is an alternative to minimize information leakage by caching the root zone at the recursive resolver, eliminating queries to root servers [23, 29]. This approach eliminates leakage to roots, but not to lower-level authoritatives.

Some institutions use external resolvers instead of operating their own. This approach, while it may provide greater aggregation over larger groups that share the resolver, introduces new privacy threats because all of one's data is shared with that provider. We recommend institutions who choose to follow this approach to use resolvers with non-logging policies. For resolvers that log queries, institutions should ensure the resolver's privacy policy aligns with their own privacy goals. For example, Google's public DNS removes a query's source IP from their logs but permanently stores the corresponding AS number and domain [8], which we have shown to be enough for violating institutional privacy.

Actions service providers sharing DNS data can take: As we showed in §4.1, anonymizing the last 8 bits of source IP addresses in DNS data does not effectively break the link between an institution and its queries. Partial anonymization is weaker than prefix-preserving anonymization's original proposal to anonymize all bits of the address [44]. In practice, researchers often need the head of the prefix to do geolocation or understand policy routing. Therefore, prefix-preserving anonymization can help reduce risks to individuals, but not to institutions if preserving utility of the data is desired. To protect institutions, one must either fully anonymize IP addresses and thus, significantly reduce the research value of the data, or place legal constraints on researchers (the data we use takes the latter approach). For the case of wider sharing, when legal constraints may be unenforceable, research on privacy-preserving network data sharing using rigorous approaches such as differential privacy [20] is needed.

Our work prompts additional privacy considerations for authoritative server operators that log DNS data. Best practices for anonymization and handling DNS data exist for recursive resolver operators [18]. We hope our findings will prompt use of similar guidance by root server operators.

7 RELATED WORK

End-user privacy: DNS privacy has been characterized for individuals [12, 24], with a focus on mechanisms to protect traffic below the recursive [25, 41, 45]. We document privacy threats to institutions posed by traffic above the recursive.

Multiple groups share data collected above the recursive, while relying on aggregation and caching to obscure individuals [22]. Studies by Mohaisen *et al.* [35] and Imana *et al.* [26] enumerate the types of queries that leak information in traffic above the recursive, but focus on individual privacy and do not relate the threats to institutional privacy.

The DNS privacy working group at IETF (DPRIVE) has proposed query name minimization [13] to minimize threats in traffic above the recursive. The group is also working to add confidentiality to DNS traffic between recursives and authoritatives [32]. Adding confidentiality to in-transit traffic does not prevent the privacy leaks we enumerate, which occur through sharing data logged by DNS servers.

Hardaker showed privacy leaks to root DNS servers by analyzing queries from residential households, and introduced "LocalRoot" – a root zone caching infrastructure that can eliminate the leaks [23]. The suggested approach, if adopted, would prevent the leaks to root servers shown in our results.

Institutional privacy in DNS: Concurrent with our work, Lee's undergraduate thesis identifies organizational privacy as a risk in DNS [30]. They highlight large corporations with their own ASes, and leakage of information about product development and e-mail. Our work goes well beyond their high-level threat description by identifying specific queries that leak information, showing how to find them, and using our methodology to show leaks in real-world data.

The work of [16] performed a longitudinal study of queries with EDNS0 Client Subnet (ECS) extension to study use of Google's public DNS. They identify that many mail servers are configured to use Google's resolvers and discuss how this leaks information about email communication. Although we also look at leaks related to email, we demonstrate institutional leaks to authoritatives even when ECS is not used.

8 CONCLUSION

Our work is the first to document institutional privacy as a risk to be considered when evaluating DNS privacy. Our work prompts two recommendations (detailed in §6): we suggest a quicker adoption of query name minimization in recursive resolvers, and an exploration of using more rigorous privacy-preserving approaches for sharing data for research in cases where legal constraints may be insufficient.

ACKNOWLEDGEMENTS

This work was funded in part by NSF grants CNS-1755992, CNS-1916153, CNS-1943584, and CNS-1925737. We are grateful to anonymous reviewers for their feedback.

REFERENCES

[1] Comparison of DNS blacklists. https://en.wikipedia.org/wiki/Comparison_of_DNS_blacklists. [accessed 2019-11-25].

[2] DNSThought – qname minimization. https://dnsthought.nlnetlabs.nl/. [accessed 2021-05-04].

[3] IAB tech lab content taxonomy. https://iabtechlab.com/standards/content-taxonomy/. [accessed 2021-06-25].

[4] Regional internet registries statistics. https://www-public.imtbs-tsp.eu/~maigron/RIR_Stats/RIR_Delegations/World/ASN-ByNb.html. [accessed 2019-11-10].

[5] Security information exchange protects from cybercrime - farsight security. https://www.farsightsecurity.com/solutions/security-information-exchange/. [accessed 2019-10-08].

[6] Webshrinker. https://docs.webshrinker.com. [accessed 2018-11-01].

[7] Week-long B-Root DNS requests, IMPACT ID: USC-LANDER b-root-week-message-question-20190109/rev10299. http://www.isi.edu. Provided by USC/B-Root Operations with USC/LANDER project.

[8] Your Privacy - Public DNS. https://developers.google.com/speed/public-dns/privacy. [accessed 2019-11-25].

[9] Barnes, R., Schneier, B., Jennings, C., Hardie, T., Trammell, B., Huitema, C., and Borkmann, D. Confidentiality in the face of pervasive surveillance: A threat model and problem statement. RFC 7624, Internet Request For Comments, Aug. 2015.

[10] Bilge, L., Sen, S., Balzarotti, D., Kirda, E., and Kruegel, C. Exposure: A passive DNS analysis service to detect and report malicious domains. *ACM Trans. Inf. Syst. Secur. 16*, 4 (Apr. 2014), 14:1–14:28.

[11] Born, K., and Gustafson, D. Detecting DNS tunnels using character frequency analysis. *CoRR abs/1004.4358* (2010).

[12] Bortzmeyer, S. DNS privacy considerations. RFC 7626, IETF, 2015.

[13] Bortzmeyer, S. DNS query name minimisation to improve privacy. RFC 7816, IETF, 2016.

[14] Castro, S., Wessels, D., Fomenkov, M., and Claffy, K. A day at the root of the internet. *SIGCOMM Comput. Commun. Rev. 38*, 5 (Sept. 2008), 41–46.

[15] de Vries, W. B., Scheitle, Q., Müller, M., Toorop, W., Dolmans, R., and van Rijswijk-Deij, R. A first look at QNAME minimization in the Domain Name System. In *Passive and Active Measurement* (Cham, 2019), D. Choffnes and M. Barcellos, Eds., Springer International Publishing, pp. 147–160.

[16] de Vries, W. B., Van Rijswijk-Deij, R., de Boer, P., and Pras, A. Passive observations of a large DNS service: 2.5 years in the life of Google. In *2018 Network Traffic Measurement and Analysis Conference* (June 2018), pp. 1–8.

[17] Denis, F. Free IP address to ASN database. https://iptoasn.com/. [accessed 2018-11-01].

[18] Dickinson, S., Overeinder, B., van Rijswijk-Deij, R., and Mankin, A. Recommendations for DNS Privacy Service Operators. RFC 8932, Oct. 2020.

[19] DNS-OARC. DITL traces and analysis. https://www.dns-oarc.net/oarc/data/ditl, Jan. 2018. [accessed 2019-10-08].

[20] Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography* (Berlin, Heidelberg, 2006), TCC'06, Springer-Verlag, pp. 265–284.

[21] Edmundson, A., Schmitt, P., Feamster, N., and Mankin, A. Oblivious DNS - strong privacy for DNS queries. Tech. rep., July 2018.

[22] Foremski, P., Gasser, O., and Moura, G. C. M. Dns observatory: The big picture of the dns. In *Proceedings of the ACM Internet Measurement Conference* (Amsterdam, the Netherlands, Oct. 2019), ACM, pp. 87–101.

[23] Hardaker, W. Analyzing and mitigating privacy with the DNS root service. Tech. rep., 2018.

[24] Hesselman, C., Jansen, J., Wullink, M., Vink, K., and Simon, M. A privacy framework for DNS big data applications. Tech. rep., SIDN Labs, 2014.

[25] Hoffman, P., and McManus, P. DNS queries over HTTPS (DoH). RFC 8484, Nov. 2018.

[26] Imana, B., Korolova, A., and Heidemann, J. S. Enumerating privacy leaks in DNS data collected above the recursive. In *Proceedings of the ISOC NDSS Workshop on DNS Privacy* (2017).

[27] Jung, J., and Sit, E. An empirical study of spam traffic and the use of DNS black lists. In *Proceedings of the ACM Internet Measurement Conf.* (New York, NY, USA, 2004), IMC '04, ACM, pp. 370–375.

[28] Kinnear, E., Pauly, T., Wood, C., and McManus, P. Oblivious DNS over HTTPS. Tech. rep., IETF, October 2019.

[29] Kumari, W. A., and Hoffman, P. E. Running a Root Server Local to a Resolver. RFC 8806, June 2020.

[30] Lee, S. Current practices for DNS privacy: Protection towards pervasive surveillance, 2019. Undergraduate Thesis.

[31] Levine, J. DNS blacklists and whitelists. RFC 5782, IETF, 2010.

[32] Livingood, J., Mayrhofer, A., and Overeinder, B. DNS Privacy Requirements for Exchanges between Recursive Resolvers and Authoritative Servers. Internet-Draft draft-ietf-dprive-phase2-requirements-02, Internet Engineering Task Force, Nov. 2020. Work in Progress.

[33] MacMillan, D., and Dwoskin, E. The war inside Palantir. https://www.washingtonpost.com/business/2019/08/22/war-inside-palantir-data-mining-firms-ties-ice-under-attack-by-employees/, Aug. 2019. [accessed 2019-10-23].

[34] Mockapetris, P. Domain names—concepts and facilities. RFC 1034, Internet Request For Comments, Nov. 1987.

[35] Mohaisen, A., and Ren, K. Leakage of .onion at the DNS root: Measurements, causes, and countermeasures. *IEEE/ACM Transactions on Networking 25*, 5 (Oct 2017), 3059–3072.

[36] Moura, G. C. M., Heidemann, J., Müller, M., de O. Schmidt, R., and Davids, M. When the dike breaks: Dissecting DNS defenses during DDoS. In *Proceedings of the ACM Internet Measurement Conf.* (New York, NY, USA, 2018), IMC '18, ACM, pp. 8–21.

[37] Pradkin, Y. cryptopANT - IP address anonymization library. https://ant.isi.edu/software/cryptopANT/man.html. [accessed 2019-11-09].

[38] Rainey, R. Jefferson parish a leader in deportations, but sanctuary city row worries sheriff. https://www.nola.com/news/politics/article_d3a5ee8e-d912-525f-9137-df900524c959.html, Mar. 2017. [accessed 2019-10-23].

[39] Root Operators. *http://www.root-servers.org*, Apr. 2019.

[40] RSSAC. RSSAC advisory on measurements of the root server system. Tech. Rep. RSSAC002v3, ICANN, June 2016.

[41] Schmitt, P., Edmundson, A., Mankin, A., and Feamster, N. Oblivious DNS: Practical privacy for DNS queries: Published in PoPETS 2019. In *Proceedings of the Applied Networking Research Workshop* (New York, NY, USA, 2019), ANRW '19, ACM, pp. 17–19.

[42] Spring, J. M., and Huth, C. L. The impact of passive DNS collection on end-user privacy. In *Proc. of Workshop on Securing and Trusting Internet Names* (2012).

[43] Thomas, M. Maximizing qname minimization: A new chapter in DNS protocol evolution. https://blog.verisign.com/security/maximizing-qname-minimization-a-new-chapter-in-dns-protocol-evolution/. [accessed 2021-05-04].

[44] Xu, J., Fan, J., Ammar, M., and Moon, S. B. On the design and performance of prefix-preserving ip traffic trace anonymization. In *Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement* (New York, NY, USA, 2001), IMW '01, ACM, pp. 263–266.

[45] Zhu, L., Hu, Z., Heidemann, J., Wessels, D., Mankin, A., and Somaiya, N. Connection-oriented DNS to improve privacy and security. In *2015 IEEE Symposium on Security and Privacy* (May 2015), pp. 171–186.