Enumerating Privacy Leaks in DNS Data Collected above the Recursive

(Short paper)

Basileal Imana*, Aleksandra Korolova* and John Heidemann†*
*University of Southern California, {imana, korolova}@usc.edu
†USC/Information Science Institute, johnh@isi.edu

***Abstract*—As with any information system consisting of data derived from people's actions, DNS data is vulnerable to privacy risks. In DNS, users make queries through recursive resolvers to authoritative servers. Data collected below (or in) the recursive resolver directly exposes users, so most prior DNS data sharing focuses on queries *above* the recursive resolver. Data collected above a recursive resolver has largely been seen as posing a minimal privacy risk since recursive resolvers typically aggregate traffic for many users, thereby hiding their identity and mixing their traffic. Although this assumption is widely made, to our knowledge it has not been verified. In this paper we re-examine this assumption for DNS traffic above the recursive resolver. First, we show that two kinds of information appear in query names above the recursive resolver: IP addresses and sensitive domain names, such as those pertaining to health, politics, or personal or lifestyle information. Second, we examine how often these classes of potentially sensitive names appear in Root DNS traffic, using 48 hours of B-Root data from April 2017.**

I. Introduction

The Domain Name System (DNS [1]) is a critical part of the Internet's infrastructure. People use DNS to translate a human readable domain name such as www.example.com to an IP address, as well as to resolve services and look up other information. Since almost every activity on the Internet starts with a DNS lookup, understanding DNS performance is important, and DNS can reveal information about the Internet and its users. As a result, DNS traffic has been collected from different locations in the DNS ecosystem. Analysis of this traffic has been used for many purposes, including understanding Internet trends [2], detecting security threats [3], preventing information leakage [4], and detecting Internet-wide activity [5].

As with any information system consisting of data derived from people's actions, DNS data can risk the privacy of its users [6]. Most work resolving a DNS query is done by a *recursive resolver* (shown in Figure 1 and defined in Section II). Previous work on mitigating the privacy risks of DNS focused on data collected below or on the recursive resolvers, due to greater end-user privacy implications [7], [8]. A general consensus has emerged, as described by Spring and Huth [9], that DNS data collected *above* a recursive leaks little private information because the recursive resolver aggregates and mixes traffic from many users, hiding the real end-users in the process. In Bortzmeyer's review of DNS privacy [6], he suggests that aggregation from the recursive may not *guarantee* privacy since some domain names pass through and leak private information.

The main contribution of this paper is to re-examine the question of DNS data privacy above the recursive resolver. We show that *DNS data above the recursive resolver can leak private information*. We identify two classes of information that leak through DNS query names: IP addresses and sensitive domain names. These may constitute personally-identifiable information (PII) in some cases.

We enumerate and analyze these potential privacy leaks. The first case is *when the domain name of a DNS query itself contains sensitive information*. Examples are an individual's name inside a domain name of a host, or an IP address assigned by an ISP that can be associated with a specific customer for the duration of a lease. In this work, we study IP addresses as one class of information that can potentially contribute to privacy leaks and use root DNS data to answer the questions:

> What type of queries contain contain an IP addresses inside the domain name and how often do such queries appear? And what are the privacy implications of each type of query?

The second case is *when there is not enough aggregation at a recursive resolver*. In this case, there may be a single or a few users sharing a recursive resolver. This makes feasible privacy attacks that would otherwise be possible only for an adversary that can observe data below the recursive. While we do not directly address the problem of measuring how much aggregation there is at recursive resolvers, we show the prevalence of queries to sensitive domain names that can be used to profile users when there is insufficient aggregation at the resolvers. More specifically, we ask:

> How common are sensitive domain names such as those pertaining to gender, health, religion, ethnicity and lifestyle?

Through these studies, our work will help understand the privacy considerations that need to be taken into account when sharing DNS data collected above the recursive.

II. DNS Overview

The Domain Name System [1], [10] is a globally distributed database designed for mapping domain names to information as IP addresses. A DNS *client* sends a query for a given

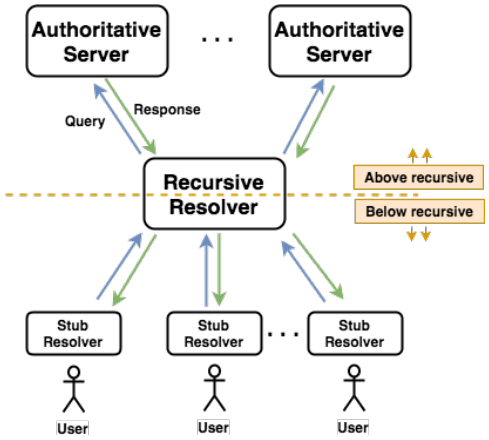

Fig. 1. Actors in the Domain Name System

kind of information (the *query type*) and a *query name* to a DNS *server*. The server responds with a *resource record* answering the query type and name, or an error. In addition to mapping names to IP addresses, DNS resolves other types of queries, and its lightweight design has prompted its use in spam defenses [11].

At a high level, three actors are involved in domain name resolution process (Figure 1). A stub resolver running on a client's machine sends a query to a recursive resolver. An end-user accesses the DNS through a stub-resolver, typically running directly on their computer. This stub contacts a recursive resolver; the recursive resolver then handles the name resolution on behalf of the stub by iteratively contacting a number of authoritative servers relevant to the domain name the stub requested. Each authoritative server is responsible for a given part of the DNS hierarchy. The recursive resolver generates requests to one or more authoritative servers to answer a new query from stub. Recursive resolvers often cache resource records and return responses from this cache without contacting authoritatives. Resource records can be held for a time period specified by the Time to Live (TTL) field in the response from the authoritative servers.

When a recursive resolver handles requests for multiple users, it *aggregates* these requests. Some requests are handled from its cache and do not require additional queries to authoritatives. Requests that do go to authoritatives are placed from the recursive, providing a level of indirection from stubs and users. The authoritative server sees only the recursive server's IP address, not those of stubs.

We talk about *below* and *above the recursive* to indicate DNS traffic from the stub to recursive ("below") and the recursive to authoritative server ("above").

III. Threat Model

A. Potential Adversaries

When considering threats to user privacy, we assume a passive adversary that has access to DNS traffic or logs above a recursive resolver, typically by observing traffic in the network, or on an outgoing network connection of a recursive server or at the incoming connection of one or more authoritative servers.

We assume that the adversary is interested in learning characteristics about a specific targeted user or group of people. The adversary may supplement network traffic with additional knowledge about the domain names or IP addresses associated with the targeted users or groups.

An adversary may also be partially active, perhaps causing users to query specific domain names by sending them e-mail with embedded URLs, as is commonly done with 1-pixel web beacons used in mail tracking [12]. If the adversary controls the URL, they can use it to direct DNS traffic to authoritative DNS servers under their control.

B. Privacy and Personally-Identifiable Information in the Context of DNS

Our goal is to protect user privacy, specifically information about what websites or Internet services a user accesses. In addition, we wish to conceal the user's IP address, since that can be used in some cases to track or geolocate the user.

Related to privacy is *Personally-Identifiable Information* (PII). PII is a legal term, with an interpretation that varies depending on jurisdiction. NIST's Special Publication 800-122 [13] defines PII as "any information that can be used to distinguish or trace an individual's identity" and "any other information that is linked or linkable to an individual".

European Union's General Data Protection Regulation [14] defines personal data as "any information relating to an identified or identifiable natural person". Such identifiable information includes names, identification numbers and online identifiers [14].

In the context of DNS, the source IP and query name might sometimes be considered to contain PII. From the above PII definitions, we believe these fields can be regarded as privacy sensitive when they contain PII such as a person's name, or information that can be linked to a person through readily available external knowledge. Above the recursive resolver, the source IP is not usually associated with an individual. However, IP addresses can appear in the query name as well.

Whether or not a user's IP address is PII varies. NIST identifies an IP address as not a PII on its own, but classifies it as linked PII since it can identify an individual when combined with external information. In 2016, Court of Justice of the European Union also ruled that a dynamic IP address can constitute personal data if it can be used to identify an individual with the help of external information [15]. The external information needed to associate IPs to individuals are typically held only by ISPs and not externally available.

IV. Privacy Leaks Above the Recursive

We next look at how the query name and source IP can leak private information, giving examples of each. We then turn to aggregation by the recursive resolver and threats from query injection.

A. Sensitive Information in Query Names

The query name of a DNS query itself may contain sensitive information, either PII or linked PII. While some of these names will not appear above the recursive, due to caching, and the original of all will be obscured by the recursive, they represent a powerful leak of information because some will pass through the recursive. The query may leak private information regardless of whether the resolver is shared among many users or not.

1) IP addresses: We first consider IP addresses in the query name, describing how they leak and then considering when they are sensitive.

Examples of queries where the domain name contains an IP address include:

- Reverse-DNS queries (rDNS) such as 0.2.0.192.in-addr.arpa (typically for a PTR record).
- DNS-based queries for IP reputation, such as DNS-based IP blacklisting (DNSBL) with 0.2.0.192.sbl.spamhaus.org.
- Queries to domain names that are assigned to ISP or cloud customers, e.g., 192-0-2-0.dedicated.static.sonic.net and 0.2.0.192.rst4.r.skype.net. These domains are often pre-allocated and assigned to particular Customer Provided Equipment (CPE) in people's homes.

Not all IP addresses are equally privacy sensitive. In order to understand the privacy implication of IP addresses inside domain names, we analyze each of the above categories separately.

Reverse-DNS queries are used for a number of reasons such as logging IP addresses of website visitors, checking legitimacy of email origins, and troubleshooting a network. Often these identify infrastructure, such as mail servers. However, they also can identify CPE. CPE often arises when it has been compromised and is used for scanning or spam, the basis of DNS backscatter for ssh [5]. Given the variety of common use cases, rDNS queries leak little information about individuals and pose limited privacy risks.

IP reputation queries, on the other hand, are mainly used to filter spam emails that originate from illegitimate mail servers [11]. While they do not identify individuals, an IP address that appears frequently in IP reputation queries might suggest it is spamming (again, a feature used in DNS backscatter [5]).

Domain names with the IP address embedded in them are often assigned to customer-provided equipment by ISPs, web hosts, and cloud service providers. Sometimes these CPE names include keywords that can be used to differentiate between statically and dynamically assigned IP addresses. While CPE names cannot be directly mapped to individuals without external information, CPE names that indicate static IPs suggest that a DNS name might be associated with an individual for the long-term. On the other hand, the privacy implication of dynamic IP addresses depends on how often they change, and the availability of the ISP's data that maps IPs to individuals. Padmanabhan *et al.* [16] showed that stability of dynamic addresses greatly varies across countries, with some countries intentionally changing them regularly to promote privacy; however, for most North American ISPs, even dynamic addresses are stable for weeks.

2) Trackable names: A domain name may also contain a unique name that is associated with a single individual or small well-defined group of people. For instance, an individual may own a domain last-name.example.com and use it to host their own e-mail or other services. In such scenario, a pattern of DNS queries to last-name.example.com may indicate when the individual performs certain online activities.

This type of privacy attack is feasible particularly for an adversary who has a prior knowledge about association between a domain name and an individual or group of people. The adversary can then look for the queries that contain the unique name in the DNS data and use the rate at which the queries appear to track and learn certain behaviors of the targeted individual or group of people.

Fortunately, while domain names may be identifying with external information (for example, clintonemail.com was Hillary Clinton's private server), knowledge about the association of domain to individual requires external information, since there are many Clintons, and many possible variations that Hillary Clinton might use. Another risk this kind of domain name poses is that it is discoverable by person with only partial external information (say, knowledge that Hillary Clinton has some private e-mail domain somewhere).

3) Sensitive Names: Some names may be sensitive because they imply information about groups of individuals that may be sensitive, such as names that pertain to gender, ethnicity or lifestyle that may be persecuted. Examples of sensitive domain names that may leak such information include those that pertain to addiction (such as aa.org, Alcoholics Anonymous), religion (for example, jw.org, jewishjournal.com), sexual preference (for example, gaycities.com), ethnicity (for example, irishcentral.com), or lifestyle choices (for example, veggieboards.com).

These names benefit from the aggregation typical to a recursive resolver, since mixing and caching obscures who specifically is associated with each category.

B. Source IP and Insufficient Aggregation

We generally assume the source IP of the recursive resolver does not directly divulge personal information, since most recursives are operated by organizations or ISPs.

However, they do leak information when there is insufficient aggregation. A workplace of 100 people may have only a few people working at night, making nighttime queries more easily identifiable. Other organizations may be small (a handful of people), limiting the population among which queries are aggregated.

Another risk of (insufficient) aggregation are some recent proposals that everyone run a *personal* recursive resolver. Such approaches have been suggested to guarantee individuals benefit from DNSSEC processing [17], or to reduce risks of cache poisoning [18]. While enhancing security, these

TABLE I
DNS DATASET USED IN THIS PAPER

dataset	duration	queries approx. total	sampled and filtered
B-ditl-2017	48 hours	$5.7x10^9$	1,085,703

approaches may eliminate the privacy benefits of aggregation (as Schomp et al. observe [18]).

C. Query Injection

We observed earlier that an adversary can inject queries that can pierce through the recursive resolver (Section III-A). Such an adversary is quite powerful. One way an adversary can achieve this is by causing a user to query a non-existent domain. This will ensure the DNS query will penetrate through the resolver's cache. One possible way to defeat this attack is to refuse to resolve names in e-mail messages, however with HTML-formatted mail, the collateral damage of this defense may be high.

A similar query injection technique is used by Netalyzr, a network measurement and debugging service developed at ICSI and UC Berkeley [19]. For debugging, the tool uses nonce DNS names (for example, 369839a0-32153-dcf252d3-821e-46e1-b706.netalyzr.icsi.berkeley.edu.) to ensure queries are not blocked by caches and to identify specific queriers. While nonces accomplish their goal, they also allow a third party that can observe DNS queries to identify specific sessions.

V. Data Analysis and Results

A full analysis of DNS data privacy is underway; this short paper presents some early results about query names in the categories given in Section IV.

A. Dataset

We use B-root's 2017 DITL (Day-In-The-Life of the Internet) data. The dataset (Table I) contains approximately 5.7 billion queries and responses collected for 49 hours starting at 2017-04-11t11:00 UTC. We deterministically sample the dataset over the time period, taking the first hundred thousand messages from approximately every hour. The data is divided into packet capture files, each of fixed size of $\approx$ 2GB. At the B-Root traffic rate, each file lasts between 90 and 120 seconds. We take the first 100K from every 50^{th} file in the dataset. After sampling, we keep only queries with successful responses (i.e., response code NOERROR).

B. IP Addresses in Domain Names

DNS queries whose domain name contains an IP address have distinct patterns which we used to categorize them into different types listed in Section IV-A1. Specifically, we classified the IP addresses into four categories: rDNS, DNSBL, CPE and unclassified, by using keywords and regular expressions specific to the structure of each type of query. For example, rDNS queries end with either in-addr.arpa or ip6.arpa; DNSBL queries contain subdomains such as .dnsbl., .bl. and .sbl. following the IP address; and queries to domain names assigned to CPEs or web hosts often contain an IP address where a hyphen is used in place of a dot to separate the octets. Example queries of each type are listed in Section IV-A1.

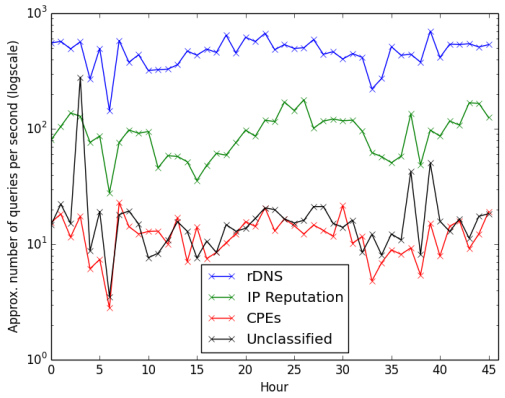

Fig. 2. IPv4 addresses in domain names

Figure 2 shows approximate number of queries per second for IPv4 addresses and how they vary across the two days spanned by the dataset. We use the duration of each observation period to approximate the query rates. Table II shows total counts and percentages relative to each group and relative to whole traffic for both IPv4 and IPv6.

The experiments show that the less privacy sensitive rDNS queries contribute to the largest percentage of IP addresses embedded in domain names of DNS queries. DNSBL and CPE categories account for a smaller but notable fraction. While IP addresses in domain names are not sensitive for majority of the queries in the dataset, there are non-insignificant number of queries which contain IP addresses that might potentially leak private information.

Analysis over time (Figure 2) shows that individual observations show considerable variation, suggesting that multiple samples at different times are needed to establish accurate long-term trends.

C. Sensitive Domain Names

There are many potential sensitive domain names. To categorize them systematically, we used the Alexa top sites by category list of domains [20], which defines 17 main categories. Of these, we select five categories: gender, ethnicity, religion, lifestyle and health, and count how many of them appear.

Alexa top sites breaks down each category into a number of nested subcategories. For each of the five categories, we select all second level subcategories. For each subcategory, we take all domains that are publicly available (up to 50 domains) and combine them into one list per category. To check whether a domain name from the lists appears in the dataset, we compare the suffix of query names in the dataset against each name in the list. For this experiment, we further filtered the queries

TABLE II
CATEGORIZED COUNT AND PERCENTAGES OF IPV4 AND IPV6 ADDRESSES

	IPv4			IPv6		
all queries	1,085,703	(100%)		1,085,703	(100%)	
query names without IP	1,042,857	(96.1%)		1,084,840	(99.9%)	
query names with IP	42,846	(3.9%)	[100%]	863	(0.08%)	[100%]
rDNS	33,715	(3.11%)	[78.7%]	852	(0.078%)	[98.7%]
DNSBL	6,820	(0.63%)	[15.9%]	6	(0.0006%)	[0.07%]
CPEs	891	(0.08%)	[2.1%]	0	(0.00%)	[0.00%]
Unclassified	1420	(0.13%)	[3.3%]	5	(0.0005%)	[0.06%]

TABLE III
COUNT OF ALEXA TOP SITES WE USED BY CATEGORY

Category	Subcategories	Domains
ethnicity	30	859
religion	62	2,158
lifestyle	7	265
health	37	1,621
gender	36	1,126

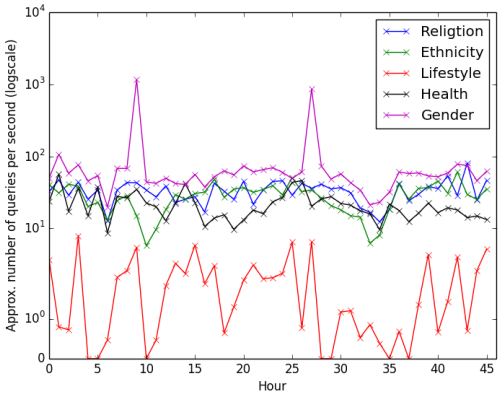

Fig. 3. Sensitive domains from five categories

in the dataset to include only A and AAAA queries, as these types of queries are used to translate a domain name into an IP address, and are observed when a user attempts to visit a particular website.

Table III lists the number of categorized Alexa top site domains we used in our experiment. Table IV and Figure 3 show how many and how often such domains appear in the dataset. All categories combined account for approximately 1.2% of the queries. The small percentage is due to the frequency of domain names in DNS traffic coming from a Zipfian distribution with a long tail, with generic popular websites contributing to the majority of the traffic [21].

TABLE IV
COUNT AND PERCENTAGES OF SENSITIVE DOMAIN NAMES

all queries	1,085,703	(100%)	
queries different from A or AAAA	251,261	(23.1%)	
queries of type A or AAAA	834,442	(76.9%)	
not sensitive names	821,690	(75.7%)	
sensitive names	12,752	(1.2%)	[100%]
Ethnicity	2030	(0.19%)	15.9%]
Religion	2437	(0.22%)	[19.1%]
Lifestyle	141	(0.01%)	[1.1%]
Gender	6559	(0.6%)	[51.4%]
Health	1585	(0.15%)	[12.4%]

VI. RELATED WORK

Several areas of prior work are relevant to our work on DNS privacy.

Understanding privacy risks: Our work draws on prior work on understanding privacy risk of DNS data [6], [22]. Kang *et al.* [23] provide an overview of challenges related to DNS privacy and summarize the above and other related prior work. We extend their work to enumerate different types of information that leaks through query names, and we use root DNS data to examine how often such information is observed.

Spring and Huth [9] formalized the probability of one being able to reconstruct a user's DNS behavior from observations above the recursive, concluding that it is very unlikely. They suggest the difficulty depends on the degree of aggregation, but do not quantify that degree in practice. We also show that sensitive names and IP addresses leak through a recursive, even when aggregation attenuates the signal.

Other groups have looked at different ways DNS may disclose privacy of end-users. Kintis *et al.* [24] analyzed privacy leaks resulting from use of the client-subnet extension to DNS, something now widely deployed by content delivery networks. Guha and Francis [25] showed the feasibility of DNS-based location surveillance when dynamic DNS is used to change IP to host mappings. Finally, Krishnan and Monrose [26] looked at the privacy implications of DNS prefetching. Overall, our work is more general and identifies different classes of potentially sensitive information that leak through query names.

Quantifying privacy leaks: Prior work on quantifying privacy leaks focused on data below the recursive. For example, Könings *et al.* [27] looked at privacy leakage through multicast DNS queries originating from devices on a university network. Herrmann *et al.* [28] evaluated behavior based tracking through DNS using queries observed by two recursive resolvers of a university network. DNS below the recursive poses greater risks that has been better studied; our study focuses on privacy leaks of DNS data above the recursive.

Mitigating privacy leaks of DNS: Prior work on improving the privacy of DNS has taken many approaches. Zhao *et al.* [29], [8] designed new protocols to improve the privacy of

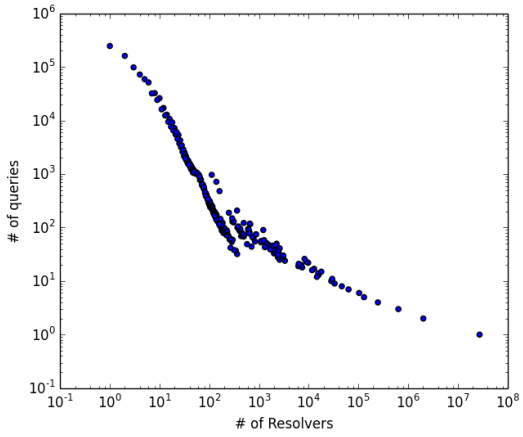

Fig. 4. Distribution of total number of queries across recursive resolvers

DNS data below or in recursive resolvers. Lu and Tsudik [30] proposed a privacy-preserving protocol with particular focus on privacy of domain name owners of DNS queries. Zhu *et al.* [7] suggested use of TLS to protect DNS privacy; this approach was later standardized by the IETF and is now seeing deployment [31]. The standardization work focuses on privacy between the stub and the recursive resolver, although they also evaluate performance between recursive and authoritative resolvers.

Rather than protecting individual queries, Bortzmeyer's query minimization approach seeks to limit information disclosed to authoritative servers to reduce privacy leakage [32].

While these works seek to conceal or minimize queries, our work instead examines what observed queries may leak. Our work applies if protocols such as TLS and query minimization are not used (as typical today), or if the operator of the DNS recursive or authoritative servers observe queries.

VII. Future Work

There are two main paths for future work: measuring how much aggregation there is at recursive resolvers in the wild and developing techniques for quantifying privacy leaks of root DNS data at a large scale.

The DNS protocol presents a unique challenge in measuring aggregation, with multi-level caching at browsers and resolvers, and because network address translation (NAT) devices affect the traffic seen by authoritative servers. Furthermore, there are diverse DNS clients, some of whom send more queries than would be expected, as observed in the 1990s [33] and continuing today. This diversity makes it difficult to determine aggregation levels solely from traffic arriving at an authoritative server. We are planning to explore this question through controlled experiments.

VIII. Conclusion

This paper re-examined the widely held assumption that DNS data collected above the recursive does not leak privacy. We identified and studied two types of information that appear in query names that pose some privacy concerns: IP addresses and sensitive domain names. We showed that, while a majority of the query names are not sensitive, significant numbers of both types of these names may expose some private data of end-users. Our work is a new attempt towards quantifying privacy leaks of root DNS data at a larger scale.

References

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

[2] S. Castro, D. Wessels, M. Fomenkov, and K. Claffy, "A day at the root of the internet," *SIGCOMM Comput. Commun. Rev.*, vol. 38, pp. 41–46, Sept. 2008.

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

[4] K. Born and D. Gustafson, "Detecting DNS tunnels using character frequency analysis," *CoRR*, vol. abs/1004.4358, 2010.

[5] K. Fukuda, J. Heidemann, and A. Qadeer, "Detecting malicious activity with DNS backscatter over time," *ACM/IEEE Transactions on Networking*, vol. 25, pp. 3203–3218, Aug. 2017.

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

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

[8] F. Zhao, Y. Hori, and K. Sakurai, "Two-servers PIR based DNS query scheme with privacy-preserving," in *The 2007 International Conference on Intelligent Pervasive Computing (IPC 2007)*, pp. 299–302, Oct 2007.

[9] J. M. Spring and C. L. Huth, "The impact of passive DNS collection on end-user privacy," in *Securing and Trusting Internet Names*, 2012.

[10] P. Mockapetris, "Domain names - implementation and specification," RFC 1035, IETF, 1987.

[11] J. Jung and E. Sit, "An empirical study of spam traffic and the use of DNS black lists," in *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement*, IMC '04, (New York, NY, USA), pp. 370–375, ACM, 2004.

[12] S. Englehardt, J. Han, and A. Narayanan, "I never signed up for this! privacy implications of email tracking," in *Privacy Enhancing Technologies Symposium (PETS)*, 2018.

[13] K. S. Erika McCallister, Tim Grance, "Guide to protecting the confidentiality of personally identifiable information," Tech. Rep. 800-122, DIANE Publishing, April 2010.

[14] European Comission, "EU general data protection regulation." https://www.privacy-regulation.eu/en/index.htm, 2016. [accessed 2017-11-30].

[15] "CJEU judgement." http://curia.europa.eu/juris/document/document.jsf?docid=184668&doclang=EN&cid=1095511, 2016. [accessed 2017-11-30].

[16] R. Padmanabhan, A. Dhamdhere, E. Aben, k. claffy, and N. Spring, "Reasons dynamic addresses change," in *Proceedings of the 2016 Internet Measurement Conference*, IMC '16, (New York, NY, USA), pp. 183–198, ACM, 2016.

[17] NLnetLabs, "Dnssec-trigger." web page http://www.nlnetlabs.nl/projects/dnssec-trigger/, May 2014.

[18] K. Schomp, M. Allman, and M. Rabinovich, "DNS resolvers considered harmful," in *ACM Workshop on Hot Topics in Networks*, (Los Angeles, Californa, USA), ACM, Oct. 2014.

[19] C. Kreibich, N. Weaver, B. Nechaev, and V. Paxson, "Netalyzr: Illuminating the edge network," in *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, IMC '10, (New York, NY, USA), pp. 246–259, ACM, 2010.

[20] "Alexa top sites by category." https://www.alexa.com/topsites/category, 2017. [accessed 2017-11-30].

[21] J. Jung, E. Sit, H. Balakrishnan, and R. Morris, "DNS performance and the effectiveness of caching," *IEEE/ACM Trans. Netw.*, vol. 10, pp. 589–603, Oct. 2002.

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

[23] A. R. Kang, J. Spaulding, and A. Mohaisen, "Domain name system security and privacy: Old problems and new challenges," *CoRR*, vol. abs/1606.07080, 2016.

[24] P. Kintis, Y. Nadji, D. Dagon, M. Farrell, and M. Antonakakis, "Understanding the privacy implications of ECS," in *Proceedings of the 13th International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment - Volume 9721*, DIMVA 2016, pp. 343–353, 2016.

[25] S. Guha and P. Francis, "Identity trail: Covert surveillance using DNS," *Privacy Enhancing Technologies: 7th International Symposium, PET 2007 Ottawa, Canada, June 20-22, 2007 Revised Selected Papers*, pp. 153–166, 2007.

[26] S. Krishnan and F. Monrose, "DNS prefetching and its privacy implications: When good things go bad," in *Proceedings of the 3rd USENIX Conference on Large-scale Exploits and Emergent Threats: Botnets, Spyware, Worms, and More*, LEET'10, (Berkeley, CA, USA), pp. 10–10, USENIX Association, 2010.

[27] B. Knings, C. Bachmaier, F. Schaub, and M. Weber, "Device names in the wild: Investigating privacy risks of zero configuration networking," in *2013 IEEE 14th International Conference on Mobile Data Management*, vol. 2, pp. 51–56, June 2013.

[28] D. Herrmann, C. Banse, and H. Federrath, "Behavior-based tracking: Exploiting characteristic patterns in DNS traffic," *Computers & Security*, vol. 39, no. Part A, pp. 17 – 33, 2013. 27th IFIP International Information Security Conference.

[29] F. Zhao, Y. Hori, and K. Sakurai, "Analysis of privacy disclosure in DNS query," in *2007 International Conference on Multimedia and Ubiquitous Engineering (MUE'07)*, pp. 952–957, April 2007.

[30] Y. Lu and G. Tsudik, "Towards plugging privacy leaks in domain name system," *CoRR*, vol. abs/0910.2472, 2009.

[31] "DNS privacy implementation status." https://dnsprivacy.org/wiki/display/DP/DNS+Privacy+Implementation+Status, 2017. [accessed 2017-11-15].

[32] S. Bortzmeyer, "DNS privacy considerations," RFC 7816, IETF, 2016.

[33] P. B. Danzig, K. Obraczka, and A. Kumar, "An analysis of wide-area name server traffic: A study of the Domain Name System," in *ACM SIGCOMM Conference*, pp. 281–292, Jan. 1992.

[34] M. A. Rajab, F. Monrose, A. Terzis, and N. Provos, "Peeking through the cloud: DNS-based estimation and its applications," in *Applied Cryptography and Network Security: 6th International Conference, ACNS 2008, New York, NY, USA*, pp. 21–38, 2008.

[35] D. W. Kim and J. Zhang, "Deriving and measuring DNS-based fingerprints," *Journal of Information Security and Applications*, vol. 36, no. Supplement C, pp. 32 – 42, 2017.

[36] H. Guerid, K. Mittig, and A. Serhrouchni, "Privacy-preserving domain-flux botnet detection in a large scale network," in *2013 Fifth International Conference on Communication Systems and Networks (COMSNETS)*, pp. 1–9, Jan 2013.

[37] H. Federrath, K.-P. Fuchs, D. Herrmann, and C. Piosecny, "Privacy-preserving DNS: Analysis of broadcast, range queries and mix-based protection methods," in *Proceedings of the 16th European Conference on Research in Computer Security*, ESORICS'11, pp. 665–683, 2011.

[38] H. Shulman, "Pretty bad privacy: Pitfalls of DNS encryption," in *Proceedings of the 13th Workshop on Privacy in the Electronic Society*, WPES '14, (New York, NY, USA), pp. 191–200, ACM, 2014.

[39] E. Jung, "A data-driven decision making with big data analysis on DNS log," *Information Science and Applications 2017: ICISA 2017*, pp. 426–432, 2017.

[40] S. Castillo-Perez and J. Garcia-Alfaro, "Evaluation of two privacy-preserving protocols for the DNS," in *2009 Sixth International Conference on Information Technology: New Generations*, pp. 411–416, April 2009.

[41] "The European Union data privacy directive." http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:HTML, 1996. [accessed 2017-11-30].