New Approaches to Safe Analysis of Long-Term DNS Data

John Heidemann and Wes Hardaker
USC/Information Sciences Institute

Empirical data is important to evaluate DNS and Internet naming. Traces can drive experiments, or be altered to explore "what-if" scenarios. Long-term data analysis can is needed to identify trends in DNS use.

Day-In-The-Life-of-the-Internet (DITL) is the most widely available DNS data source today [3, 6]. From it's original framing of measuring the Internet as a whole, DITL has adapted to DNS measurements that occur once or twice a year for about 48 hours, typically with data from most of the DNS root operators from their respective DNS root servers. DITL data is available through DNS-OARC to its members. Some researchers have access to their own sources of DNS data, but usually such data is considered proprietary and is not available to other researchers, or it is only available with anonymization [1].

DNS data is often limited because of privacy concerns. Complete DNS data from an individual will indicate what websites that user accesses and thus poses a privacy risk. Nearly all users request data through a shared recursive-resolver (or "recursive"), and a recursive resolver then caches the results of queries from multiple users. Recursives thus providing some anonymity via caching, aggregation, and deniability. However, even with this aggregation, DNS above the recursive can still reveal access patterns of institutions [7, 8], and unique queries can easily leak information. (If I am the only person who looks up privatedomain.example.com, that query will fingerprint me from .com's perspective and potentially even from the DNS root if DNS Minimization [2] is not used.)

Our goal is to provide research access to DNS data to support experimentation and long-term analysis, and to do so while minimizing privacy risks.

Proposed Safe Access to Long-term DNS data: We see three tiers of access to DNS data: (1) curated datasets, (2) controlled access to specialized data, (3) internal analysis with controlled output. These tiers employ less anonymization and greater supervision, with the goal of balancing research risks and benefits. Our thinking here is guided by the Belmont principles [9], and their extension to to network data analysis as described in the Menlo Report [5].

We recognize that nearly any data sharing poses some risk, so we plan to combine any data sharing with a legal agreement with researchers restricting them from redistributing the data and to not attempt to deanonymize the data or identify individuals. This agreement formalizes our expectations, but its success requires good faith on behalf of the researchers. We therefore augment it with technical methods in each of the tiers.

The first tier is curated datasets. We identify events we believe are of interest (such as a denial-of-service attack or deployment of a protocol change), then anonymize the data around that event. We plan to use Cryptopan to anonymize IPv4, v6, and MAC addresses [10], and we have the option to anonymize DNS query names and anonymize or drop queries that appear to be sensitive.

The second tier is to provide controlled access to specialized data. Some researchers have very specific research needs, where a slice through the data may pass some fields without anonymization while other sensitive fields are omitted. For example, a study might pass clear IP addresses and IP-packet-level TTLs, while discarding query names. Even if IP addresses are sensitive, use of the data is not sensitive if the researcher agrees not to deanonymize the data they have (as required by the legal agreement), and if we strip other sensitive information (query names). These datasets are more labor intensive to generate, since we must customize the dataset to each researcher, but they can enable a class of research impossible with general curated data.

Finally, some researchers may require direct access to raw data. For example, studies of DNS data sensitivity [8] cannot be done on anonymized data. We can accommodate such research by providing a subset of the data to researchers on our computing facilities, but forbidding copying data out. When the research has concluded, the researcher can provide any resulting graphs for expert analysis, and we will manually examine the research results to insure private information does not leak. This approach is the most labor intensive and so will be justified only when the first two tiers cannot support a high-value line of research.

Status and Next Steps: We plan to explore these methods of data sharing through the DIINER project at USC/ISI, initially using b.root-servers.net data and potentially using other data sources where possible. b.root-servers.net already participates in DITL (tier one), providing partial IP-address anonymization, and we provide anonymized, curated datasets through IMPACT [4]. Through DIINER we plan to expand to consider all three tiers of data access, and to work with specific researchers to show the results.

This work is partially supported by the project "DNS, Identity, and Internet Naming for Experimentation and Research (DIINER)" (NSF CNS-1925737).

REFERENCES

[1] M. Allman. Case connection zone DNS transactions. website https://www.icir.org/mallman/data/ccz-dns-logs/ReleaseNotes, Mar. 2019.

[2] S. Bortzmeyer. DNS query name minimisation to improve privacy. RFC 7816, Internet Request For Comments, Mar. 2016.

[3] S. Castro, D. Wessels, M. Fomenkov, and K. Claffy. A day at the root of the Internet. *ACM Computer Communication Review*, 38(5):41–46, Oct. 2008.

[4] DHS IMPACT. The IMPACT portal. Website https://impactcybertrust.org/, 2016.

[5] D. Dittrich and E. K. (editors). The Menlo report: Ethical principles guiding information and communication technology research. Technical report, United States Department of Homeland Security, Sept. 2011.

[6] DNS-OARC. Day In The Life of the internet (DITL) 2014. https://www.dns-oarc.net/oarc/data/ditl, Apr. 2014.

[7] W. Hardaker. Analyzing and mitigating privacy with the DNS root service. In *Proceedings of the ISOC NDSS Workshop on DNS Privacy*, San Diego, California, USA, Feb. 2018. The Internet Society.

[8] B. Imana, A. Korolova, and J. Heidemann. Enumerating privacy leaks in DNS data collected above the recursive. In *Proceedings of the ISOC NDSS Workshop on DNS Privacy*, San Diego, California, USA, Feb. 2018. The Internet Society.

[9] Nat'l Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The Belmont report: Ethical principles and guidelines for the protection of human subjects of research, Apr. 1979.

[10] J. Xu, J. Fan, M. H. Ammar, and S. B. Moon. Prefix-preserving IP address anonymization: Measurement-based security evaluation and a new cryptography-based scheme. In *Proc.10th IEEE Intl. Conf. on Network Protocols*, pages 280–289. IEEE, Nov. 2002.

DINR'20, 2020, Marina del Rey, CA, USA
2020.