DNS Privacy not so private: the traffic analysis perspective

Sandra Siby[1], Marc Juarez[2], Narseo Vallina-Rodriguez[3], and Carmela Troncoso[1]

[1]EPFL, [2]imec-COSIC KU Leuven, [3]IMDEA Networks Institute

1. INTRODUCTION

The Domain Name Service (DNS) is ubiquitous in today's Internet infrastructure. Almost every connection to an Internet service is preceded by a DNS lookup. A vast majority of DNS queries are sent in plaintext. Thus, they reveal information about the connection's destination [1]. In the Web, this lack of encryption leaks information about the browsing history of users, undermining the encryption of connections that follow the DNS resolution such as HTTPS.

In order to resolve a domain name to an IP, clients send a DNS query to a *recursive resolver* – a server with caching capabilities that implements the DNS resolution protocol. Then, the recursive resolver contacts a number of *authoritative name servers*, whose main function is to hold the distributed database of domain names. The recursive resolver traverses the hierarchy of authoritative name servers in a recursive fashion until it obtains the answer for the query and sends it back to the client.

Recursive resolvers aggregate traffic from multiple clients and there is a one-to-many relationship between the recursive and authoritative servers. Hence, the privacy risk in the recursive-authoritative link is low. However, DNS traffic between the client and the recursive resolver is linked to a specific origin IP and it is exposed to a number of entities, e.g., infrastructure providers such as ISPs and ASes.

The main approach to prevent leakage of information is to encrypt the communication until, at least, the recursive resolver. Two major protocols that intend to do so are DNS-over-TLS [1] and DNS-over-HTTPS [2]. These protocols use a TLS session between the client and the recursive resolver to exchange DNS data. In DNS-over-HTTPS (DoH), DNS traffic is exchanged via an HTTPS connection.

In this work, we evaluate the effectiveness of TLS-based solutions for DNS privacy. We focus on DoH because Google[3] and Cloudflare[4] have recently launched DoH services to alleviate the privacy risks associated with DNS. Since HTTPS is essentially HTTP over TLS, we expect our analysis to also apply to DNS-over-TLS solutions.

Our goal is to determine whether it is possible to fingerprint and identify webpages from encrypted DNS traffic. We aim to identify specific webpages beyond the IP address in the IP datagrams. For instance, pages within a website which may host many pages or be behind a Content Distribution Network hosting many websites. To achieve our goal, we collect network traces containing DoH traffic and try to recognize the webpage being resolved based on traffic features such as size and order of packets. Our initial results indicate that traffic analysis is a viable tool to distinguish webpages despite the presence of encryption. Our results suggest that encryption of DNS privacy tools must be accompanied by padding to be actually considered private.

[1]`https://tools.ietf.org/html/rfc7858` [2018-05-07].

[2]`https://www.ietf.org/id/draft-ietf-doh-dns-over-https-07.txt`[2018-05-07]

[3]`https://developers.google.com/speed/public-dns/docs/dns-over-https` [2018-05-07].

[4]`https://developers.cloudflare.com/1.1.1.1/dns-over-https/` [2018-05-07]

2. EXPERIMENTAL SETUP

We set up a Raspberry Pi as a DoH client that can send DNS queries to two commercial DNS services supporting DoH (Google and Cloudflare). We run *tcpdump* to capture the network traffic when the DoH client communicates with the resolver. We assume the case of no client-side caching, thus we clear the cache between experiments.

We run two types of experiments: *(i) single-query* and *(ii) multi-query*. In the single-query experiments, the client performs a DNS query for a specified webpage and obtains the response from the resolver. We record the network traces for query/response pairs of the Alexa's top one million websites [5]. We also collect traces for the top, middle and bottom 500 webpages in the Alexa list (1,500 traces per day in total) over a period of 30 days. This allows us to study how responses change over time and whether there are differences between popular and less popular webpages.

In the multi-query experiments we aim at capturing the fact that when a page is visited (via HTTP/HTTPS), besides the page itself, there are requests to other URLS that provide resources to the page, such as images, scripts (e.g., JavaScript), fonts and styles. This means that a client will have to resolve additional DNS domains to these URLs. We consider this scenario in our multi-query experiments. In these experiments, the client sends an HTTP request to a webpage and we analyze all the DNS traffic associated with that request. We use Selenium with a headless Chrome driver to replicate this scenario. We collect traces for Alexa's top 100,000 websites. For the top 500 webpages, we also collect traces over a period of 30 days.

3. INITIAL RESULTS

We define our classification problem as follows: given a network trace with DNS queries and responses for a web-

[5]`http://s3.amazonaws.com/alexa-static/top-1m.csv.zip` [Downloaded on 2018-03-26]

page, determine if the page is in our list of pages and, if so, attribute the trace to one of the pages in that list.

Distinguishing DNS traffic. Our first question is whether it is easy for an eavesdropper to distinguish encrypted DNS traffic from other traffic. For Cloudflare, since it has distinctive IP addresses – `1.1.1.1` and `1.0.0.1` – IP-filtering is sufficient to identify DNS requests being made to this service. Google's DoH service has multiple IP addresses associated to its domain (`dns.google.com`). However, the Server Name Indication (SNI) field in the ClientHello packet during the TLS session establishment shows that the client is trying to connect to `dns.google.com`. This also facilitates the enumeration of the IP addresses of the DoH services.

Feature Extraction. We mainly consider three feature categories in our analysis: *size*, *timing*, and *ordering*. We did not observe significant differences in the TLS headers for different traces and hence we do not consider TLS metadata as a feature.

For the size category, we look at the DNS query and response sizes. In our single-query experiments this feature is a query-response size tuple. In the multi-query experiments, due to the number of DNS queries, this feature is a sequence of query and response sizes. For the timing category, we consider the inter-arrival time between subsequent queries and responses. In the multi-query experiment, where there is a sequence of timings, we look at the mean and the variance of these time differences and study its uniqueness. Third order category is only relevant to the multi-query experiment, and it refers to the pattern in which queries and responses are observed.

Our initial analysis indicates that size alone is a distinctive feature. Figure 1 shows the distribution of anonymity sets using query and response size tuples in our single-query experiments. We observe that the number of pages having similar query-response sizes is low and a considerable number of webpages have unique query/response size tuples. Unsurprisingly, in our multi-query experiments, we observe that the number of unique sizes is higher. In some cases, sites with similar sizes actually belong to the same organization – for example, google.ae, google.es, google.fr have the same sequences and all belong to Google. We also looked at how features vary with time. For example, Figure 2 shows that changes in number and size of packets over five days is scarce, and when it happens changes are minimal.

Next steps. We are working on richer classification that also takes variance in features with time into account, using RNN (recurrent neural networks). After, we will evaluate the impact that different client operating system and location may have on our attack. We are also keen on investigating the effects that caching both in the resolver and the browser has on the classification. Finally, we intend to implement some of the DNS-over-TLS solutions and compare the results with the results obtained for DoH.

4. RELATED WORK

Schulman [4] suggested that encryption alone may not be sufficient to protect users. DNS response size variations were one of the distinguishing features suggested in this work. Imana et al. [3] also studied privacy leaks in the DNS system but focused on the recursive-authoritative traffic. Prior work such as [2] tried to perform user tracking based on client-recursive traffic but considered unencrypted DNS traces (not DoH). There is an extensive body of work on website fingerprinting and TLS traffic classification. However, we have not yet come across the use of these techniques on DoH traffic.

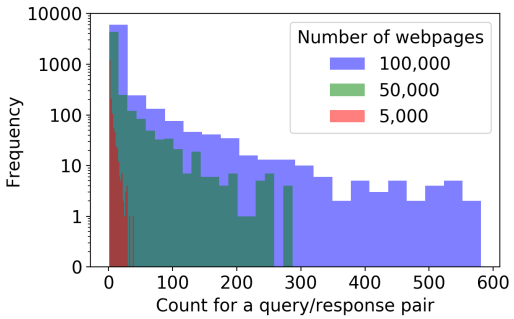

Figure 1: Distribution of anonymity sets when using query and response size pairs as a feature.

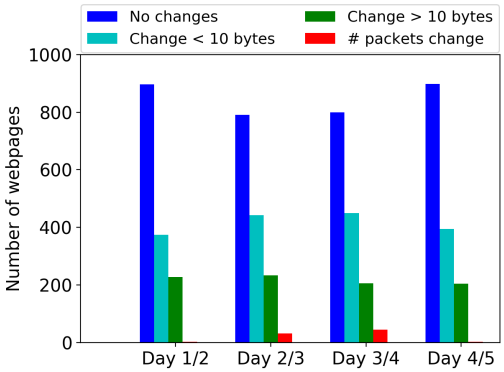

Figure 2: Changes in query and response packet sizes for 1500 webpages over 5 days. We consider changes in number and size value of the packets.

5. TALK OUTLINE

In this talk, we will present the results of our large scale study of the privacy provided by DNS-over-HTTPS solutions as deployed by Google and Cloudflare. Our analysis demonstrates that the variability in size and timing of queries enables large reduction in the anonymity sets of DNS queries. We hope that our work calls the community's attention to this problem. This is a major privacy threat for Internet users and shows that the efforts in the community to protect confidentiality of HTTP communications are flawed. We expect the discussions arising in the workshop to pave the way for the development of solutions that complement encryption and thus provide true DNS privacy.

6. REFERENCES

[1] S. Bortzmeyer. DNS privacy considerations. 2015.

[2] D. Herrmann, C. Banse, and H. Federrath. Behavior-based tracking: Exploiting characteristic patterns in DNS traffic. *Computers & Security*, 2013.

[3] B. Imana, A. Korolova, and J. S. Heidemann. Enumerating Privacy Leaks in DNS Data Collected above the Recursive. 2017.

[4] H. Shulman. Pretty bad privacy: Pitfalls of DNS encryption. In *Proceedings of the 13th Workshop on Privacy in the Electronic Society*. ACM, 2014.