Mapping an Enterprise Network by Analyzing DNS Traffic

Minzhao Lyu[1,2(✉)], Hassan Habibi Gharakheili[1], Craig Russell[2], and Vijay Sivaraman[1]

[1] University of New South Wales, Sydney, Australia
{minzhao.lyu,h.habibi,vijay}@unsw.edu.au
[2] Data61, CSIRO, Sydney, Australia
craig.russell@data61.csiro.au

Abstract. Enterprise networks are becoming more complex and dynamic, making it a challenge for network administrators to fully track what is potentially exposed to cyber attack. We develop an automated method to identify and classify organizational assets via analysis of just 0.1% of the enterprise traffic volume, specifically corresponding to DNS packets. We analyze live, real-time streams of DNS traffic from two organizations (a large University and a mid-sized Government Research Institute) to: (a) highlight how DNS query and response patterns differ between recursive resolvers, authoritative name servers, web-servers, and regular clients; (b) identify key attributes that can be extracted efficiently in real-time; and (c) develop an unsupervised machine learning model that can classify enterprise assets. Application of our method to the 10 Gbps live traffic streams from the two organizations yielded results that were verified by the respective IT departments, while also revealing new knowledge, attesting to the value provided by our automated system for mapping and tracking enterprise assets.

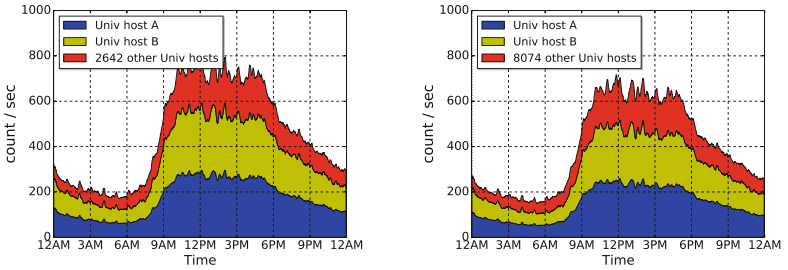
Keywords: Enterprise network · DNS analysis · Machine learning

1 Introduction

Enterprise networks are not only large in size with many thousands of connected devices, but also dynamic in nature as hosts come and go, web-servers get commissioned and decommissioned, and DNS resolvers and name servers get added and removed, to adapt to the organization's changing needs. Enterprise IT departments track such assets manually today, with records maintained in spreadsheets and configuration files (DHCP, DNS, Firewalls, etc.) – this is not only cumbersome, but also error prone and almost impossible to keep up-to-date. It is therefore not surprising that many enterprise network administrators are not fully aware of their internal assets [12], and consequently do not know the attack surface they expose to the outside world.

The problem is even more acute in university and research institute campus networks for several reasons [6]: (a) they host a wide variety of sensitive and lucrative data including intellectual property, cutting-edge research datasets, social

D. Choffnes and M. Barcellos (Eds.): PAM 2019, LNCS 11419, pp. 129–144, 2019.
https://doi.org/10.1007/978-3-030-15986-3_9



(a) Outgoing DNS queries. (b) Incoming DNS responses.

Fig. 1. University campus: outgoing queries and incoming responses, measured on 3 May 2018. (Color figure online)

security numbers, and financial information; (b) their open-access culture, decentralized departmental-level control, as well as federated access to data makes them particularly vulnerable targets for unauthorized access, unsafe Internet usage, and malware; and (c) they typically have high-speed network infrastructure that makes them an attractive target for volumetric reflection attacks.

Our aim in this paper is to develop an automated method to map internal hosts of an enterprise network by focusing only on DNS traffic which: (a) is a key signaling protocol that carries a wealth of information yet bypasses firewalls easily; (b) constitutes a tiny faction of total network traffic by volume (less than 0.1% from our measurements in two networks); and (c) is easy to capture with only a couple of flow entries (i.e mirroring UDP packets to/from port 53) in an Openflow-based SDN switch. By capturing and analyzing DNS traffic in/out of the organization, we dynamically and continually identify the DNS resolvers, DNS name-servers, (non-DNS) public-facing servers, and regular client hosts behind or not behind the NAT in the enterprise. This can let network administrators corroborate changes in host roles in their network, and also equip them with information to configure appropriate security postures for their assets, such as to protect DNS resolvers from unsolicited responses, authoritative name servers from amplification requests, and web-servers from volumetric DNS reflection attacks.

Our specific contributions are as follows. We analyze real-time live streams of DNS traffic from two organizations (a large University and a mid-sized Government Research Institute) to: (a) highlight how DNS query and response patterns differ amongst recursive resolvers, authoritative name servers, and regular hosts; (b) identify key DNS traffic attributes that can be extracted efficiently in real-time; and (c) develop an unsupervised machine learning model that can classify enterprise assets. Application of our method to the traffic streams from the two organizations yielded results that were verified by the respective IT departments while revealing new information, such as unsecured name servers that were being used by external entities to amplify DoS attacks.

2 Profiling Enterprise Hosts

In this section, we analyze the characteristics of DNS traffic collected from the border of two enterprise networks, a large University campus (*i.e.,* UNSW) and a medium-size research institute (*i.e.,* CSIRO). In both instances, the IT department of the enterprise provisioned a full mirror (both inbound and outbound) of their Internet traffic (each on a 10 Gbps interface) to our data collection system from their border routers (**outside** of the firewall), and we obtained appropriate ethics clearances for this study[1]. We extracted DNS packets from each of enterprise Internet traffic streams in real-time by configuring rules for incoming/outgoing IPv4 UDP packets for port 53 on an SDN switch (extension to IPv6 DNS packets is left for future work). The study in this paper considers the data collected over a one week period of 3–9 May 2018.

2.1 DNS Behavior of Enterprise Hosts

Enterprises typically operate two types of DNS servers: (a) **recursive resolvers** are those that act on behalf of end-hosts to resolve the network address of a URL and return the answer to the requesting end-host (recursive resolvers commonly keep a copy of positive responses in a local cache for time-to-live of the record to reduce frequent recursion), and (b) **authoritative servers** of a domain/zone are those that receive queries from anywhere on the Internet for the network address of a sub-domain within the zone for which they are authoritative (*e.g.,* `organizationXYZ.net`).

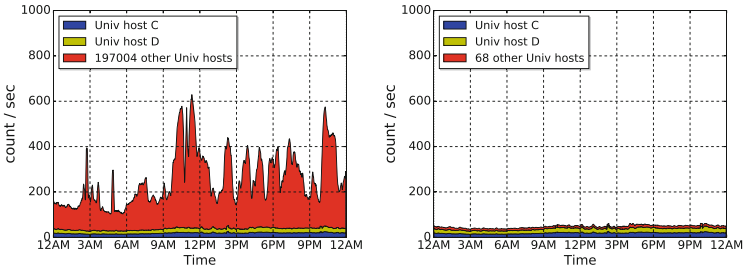
In order to better understand the DNS behavior of various hosts (and their role) inside an enterprise network, we divide the DNS dataset into two categories: (a) DNS queries from enterprise hosts that leave the network towards a server on the Internet along with DNS responses that enter the network, (b) DNS queries from external hosts that enter the network towards an enterprise host along with DNS responses that leave the network.

This analysis helps us identify important attributes related to host DNS behavior, characterizing its type/function including authoritative name server, recursive resolver, generic public-facing server (e.g web/VPN servers), or end-host inside the enterprise that may not always be fully visible to the network operators. This also enables us to capture the normal pattern of DNS activity for various hosts.

Outgoing Queries and Incoming Responses. Figure 1 shows a time trace of DNS outgoing queries and incoming responses for the university campus[2], with a moving average over 1-minute intervals on a typical weekday. The university network handles on average 353 outgoing queries and 308 incoming responses per

[1] UNSW Human Research Ethics Advisory Panel approval number HC17499, and CSIRO Data61 Ethics approval number 115/17.

[2] We omit plots for ***the research institute*** in this section due to space constraint, they are shown in Appendix 1.



(a) Incoming DNS queries. (b) Outgoing DNS responses.

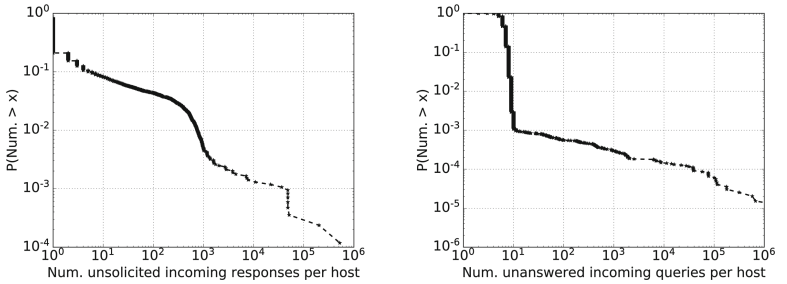
Fig. 2. University campus: incoming queries and outgoing responses, measured on 3 May 2018. (Color figure online)

second. By checking the transaction ID of queries and responses, we found that 17.28% of outgoing queries are "unanswered" (*i.e.,* 5.26M out of 30.46M) on 3 May 2018. It is also important to note that 5.24% of incoming responses to the university campus network (*i.e.,* 1.39M out of 26.59M) are "unsolicited" on the same day[3]. A similar pattern with lower number of outgoing queries and incoming responses (*i.e.,* average of 107 and 80 per second respectively) is observed in the research institute network. This network experiences approximately double the amount of unanswered queries (*i.e.,* 34.14%) and unsolicited responses (*i.e.,* 12.15%) compared to the university network.

Query per Host: We now consider individual hosts in each enterprise. Unsurprisingly, the majority of outgoing DNS queries are generated by only two hosts A and B in both networks, *i.e.,* 68% of the total in the university campus (shown by blue and yellow shades in Fig. 1(a)) and 82% of the total in the research institute – these hosts are also the major recipients of incoming DNS responses from the Internet. We have verified with the respective IT departments of the two enterprises that both hosts are the primary recursive resolvers of their organizations.

In addition to these recursive resolvers, we observe a number of hosts in both organizations, shown by red shades in Fig. 1(a), that generate DNS queries to outside of the enterprise network. The 2,642 other Univ hosts in Fig. 1(a) are either: end-hosts configured to use public DNS resolvers that make direct queries out of the enterprise network, or secondary recursive servers operating in smaller sub-networks at department-level. We found that 286 of these 2,642 University hosts actively send queries (at least once every hour) over the day and contact more than 10 Internet-based DNS servers (resolvers or name-servers). These 286 hosts display the behavior of recursive resolvers but with fairly low throughput, thus we deem them secondary resolvers. The remaining 2,356 hosts are only active for a limited interval (*i.e.,* between 5 min to 10 h) and contact a small

[3] We acknowledge that some DNS packets could have been dropped by the switches on which the span-port was configured, especially during periods of overload.



(a) Unsolicited incoming responses. (b) Unanswered incoming queries.

Fig. 3. University campus: CCDF of (a) unsolicited incoming responses and (b) unanswered incoming queries per host, measured on 3 May 2018.

number of public resolvers (*e.g.,* 8.8.8.8 or 8.8.4.4 of Google) over the day. We found that 15 of 340 hosts in the research institute display behavior of secondary resolvers.

Response per Host: Considering incoming responses (Fig. 1(b) for the university network), a larger number of "other" hosts in both organizations are observed – approximately 8 K hosts in the University and 5.8 K hosts in the research institute. Most of these "other" hosts (*i.e.,* 67%) are the destinations of unsolicited responses. To better understand the focus target of these potentially malicious responses, we analyze unsolicited incoming responses for the two enterprises. Figure 3(a) shows the CCDF of total unsolicited incoming responses per each host over a day for the university campus. Interestingly, the primary recursive resolvers in both organizations are top targets: (a) in the University campus, hosts A and B respectively are the destinations of 522 K and 201 K unsolicited incoming responses (*i.e.,* together receive 52% of total unsolicited DNS responses), and (b) in the research institute, hosts A and B respectively are the destination of 435 K and 135 K unsolicited incoming responses (*i.e.,* together receive 69% of total unsolicited DNS responses).

Incoming DNS Queries. Enterprises commonly receive DNS queries from the Internet that are addressed to their authoritative name servers.

It can be seen that two hosts of the University campus (*i.e.,* hosts C and D in Fig. 2) and one host (we name it Host C) of the research institute are the dominant contributors to outgoing DNS responses – we have verified (by reverse lookup) that these hosts are indeed the name servers of their respective organizations. Interestingly, for both organizations we observe that a large number of hosts (*i.e.,* 197K hosts of the University campus and 244K hosts of the research institute (shown by red shades in Fig. 2(a) for the university network) receive queries from the Internet, but a significant majority of them are unanswered (*i.e.,* 82.18% and 62.09% respectively) – these hosts are supposed to neither receive

Table 1. Host attributes.

	QryFracOut	numExtSrv	numExtClient	actvTimeFrac
Univ name server (host C)	0	0	0.29	0
Rsch name server (host C)	0	0	0.61	0
Univ recursive resolver (host A)	1	0.26	0	1
Rsch recursive resolver (host A)	1	0.49	0	1
Univ mixed DNS Server	0.55	0.03	0.06	1
Rsch mixed DNS Server	0.29	0.0008	0.0018	1
Univ end-host	1	0.00002	0	0.375
Rsch end-host	1	0.00003	0	0.25

nor respond to incoming DNS queries, highlighting the amount of unwanted DNS traffic that targets enterprise hosts for scanning or DoS purposes.

To better understand the target of these potentially malicious queries, we analyze unanswered incoming queries over a day for the two enterprises. Figure 3(b) is the CCDF of total incoming unanswered queries per each host for the university campus. It is seen that two hosts of the university campus receive more than a million DNS queries over a day from the Internet with no response sent back, whereas one host in the research institute has the similar behaviour. By reverse lookup, we found that the University hosts are a DHCP server and a web server that respectively received 9.4M and 4.4M unanswered queries (together contributing to 72% of red shaded area in Fig. 2(a)).

Furthermore, we analyzed the question section of unanswered incoming queries that originated from a distributed set of IP addresses. Surprisingly, in the University dataset we found that 72% of domains queried were irrelevant to its zone (*e.g.,* 47% for "`nist.gov`", 5% for"`svist21.cz`", and even 2% for "`google.com`"), and in the research institute dataset we found 84% of domains queried were irrelevant to its zone (*e.g.,* 8% for "`qq.com`", 7% for"`google.com`", and 5% for "`com`").

Considering outgoing responses (shown in Fig. 2(b) for the university network), there are 68 hosts in the campus network (shown by the red shade) and 21 hosts in the research network that respond to incoming DNS queries in addition to name servers (*i.e.,* hosts C and D). We have verified (by reverse lookup) that all hosts that generate "no Error" responses are authoritative for sub-domains of their organization zone. We also note that some hosts that reply with "Refused", "Name Error" and "Server Failure" flags to some irrelevant queries (*e.g.,* `com`) – these are secondary name servers.

2.2 Attributes

Following the insights obtained from DNS behavior of various hosts, we now identify attributes that help to automatically (a) map a given host to its function, including authoritative name server, recursive resolver, mixed DNS server (*i.e.,*

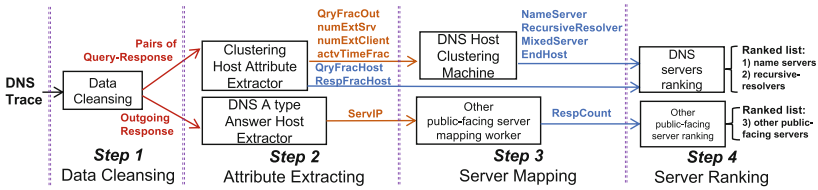


Fig. 4. Automatic classification and ranking of enterprise hosts.

both name server and recursive resolver), a (non-DNS) public-facing server, or a regular client; and (b) rank the importance of servers.

Dataset Cleansing. We first clean our dataset by removing unwanted (or malicious) records including unsolicited responses and unanswered queries. This is done by correlating the transaction ID of responses with the ID of their corresponding queries. In the cleaned dataset, incoming responses are equal in number to outgoing queries, and similarly for the number of incoming queries and outgoing responses.

Functionality Mapping. As discussed in Sect. 2.1, recursive resolvers are very active in terms of queries-out and responses-in, whereas name servers behave the opposite with high volume of queries-in and responses-out. Hence, a host attribute defined by the *query fraction of all outgoing DNS packets (**QryFraqOut**)* should distinguish recursive resolvers from name servers. As shown in Table 1, this attribute has a value close to 1 for recursive resolvers and a value close to 0 for name servers.

In addition to recursive resolvers, there are some end-hosts configured to use public resolvers (*e.g.,* 8.8.8.8 of Google) that have a non-zero fraction of DNS queries out of the enterprise network. We note that these end-hosts ask a limited number of Internet servers during their activity period whereas the recursive resolvers typically communicate with a larger number of external servers. Thus, we define a second attribute as the *fraction of total number of external servers queried (**numExtSrv**)* per individual enterprise host. As shown in Table 1, the value of this attribute for end-hosts is much smaller than for recursive resolvers. Similarly for incoming queries, we consider a third attribute as the *fraction of total number of external hosts that initiate query in (**numExtClient**)* per individual enterprise host. Indeed, this attribute has a larger value for name servers compared with other hosts, as shown in Table 1.

Lastly, to better distinguish between end-hosts and recursive resolvers (high and low profile servers), we define a fourth attribute as the *fraction of active hours for outgoing queries (**actvTimeFrac**)*. Regular clients have a smaller value of this attribute compared with recursive resolvers and mixed DNS servers, as shown in Table 1.

We note that public-facing (non-DNS) servers typically do not have DNS traffic in/out of the enterprise networks. To identify these hosts, we analyzed the answer section of *A-type* outgoing responses.

Importance Ranking. Three different attributes are used to rank the importance of name servers, recursive resolvers, and (non-DNS) public-facing servers respectively. Note that we rank mixed DNS servers within both name servers and recursive resolvers for their mixed DNS behaviour.

For recursive resolvers, we use ***QryFracHost*** defined as the *fraction of outgoing queries* sent by each host over the cleaned dataset. And for name servers, we use ***RespFracHost*** *as the fraction of outgoing responses* sent by each host. For other public-facing servers, we use ***RespCount*** as the *total number of outgoing responses that contain the IP address of a host* – external clients that access public-facing servers obtain the IP address of these hosts by querying the enterprise name servers.

3 Classifying Enterprise Hosts

In this section, we firstly develop a machine learning technique to determine if an enterprise host with a given DNS activity is a "name server", "recursive resolver", "mixed DNS server", or a "regular end-host". We then detect other public-facing (non-DNS) servers by analyzing the answer section of A-type outgoing responses. Finally, we rank the enterprise server assets by their importance.

Our proposed system (shown in Fig. 4) automatically generates lists of active servers into three categories located inside enterprise networks, with the real-time DNS data mirrored from the border switch of enterprise networks. The system first performs ***"Data cleansing"*** that aggregates DNS data into one-day granularity and removes unsolicited responses and unanswered queries (*i.e.,* step 1); then ***"Attribute extraction"*** in step 2 computes attributes required by the following algorithms; ***"Server mapping"*** in step 3 detects DNS servers and other public-facing servers; and finally ***"server ranking"*** in step 4 ranks their criticality. The output is a classification and a ranked order of criticality, which an IT manager can then use to accordingly adjust security policies.

3.1 Host Clustering Using DNS Attributes

We choose unsupervised clustering algorithms to perform the grouping and classification process because they are a better fit for datasets without ground truth labels but nevertheless exhibit a clear pattern for different groups/clusters.

Selecting Algorithm. We considered 3 common clustering algorithms, namely Hierarchical Clustering (HC), K-means and Expectation-maximization (EM). HC is more suitable for datasets with a large set of attributes and instances that

Table 2. University campus: host clusters (3 May 2018).

	Count	QryFracOut	numExtSrv	numExtClient	actvTimeFrac
Name server	42	0.0057	1e-5	0.02	0.03
Recursive resolver	14	0.99	0.06	0	0.94
Mixed DNS server	14	0.57	0.01	0.02	0.66
End-host	2195	1	2e-5	0	0.31

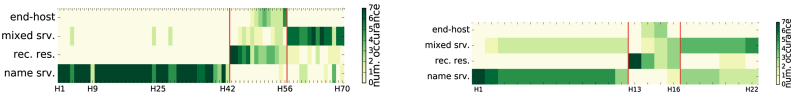
Table 3. Research institute: host clusters (3 May 2018).

	Count	QryFracOut	numExtSrv	numExtClient	actvTimeFrac
Name server	12	7e-7	5e-6	0.07	0.01
Recursive resolver	4	0.99	0.20	9e-5	1
Mixed DNS server	6	0.21	0.001	0.019	0.625
End-host	249	1	7e-4	0	0.25

have logical hierarchy (*e.g.,* genomic data). In our case however, hosts of enterprise networks do not have a logical hierarchy and the number of attributes are relatively small, therefore HC is not appropriate. K-means clustering algorithms are distance-based unsupervised machine learning techniques. By measuring the distance of attributes from each instance and their centroids, it groups datapoints into a given number of clusters by iterations of moving centroids. In our case there is a significant distance variation of attributes for hosts within each cluster (*e.g.,* highly active name servers or recursive resolvers versus low active ones) which may lead to mis-clustering.

The EM algorithm is a suitable fit in our case since it uses the probability of an instance belonging to a cluster regardless of its absolute distance. It establishes initial centroids using a K-means algorithm, starts with an initial probability distribution following a Gaussian model and iterates to achieve convergence. This mechanism, without using absolute distance during iteration, decreases the chance of biased results due to extreme outliers. Hence, we choose an EM clustering algorithm for *"DNS Host Clustering Machine"*.

Number of Clusters. Choosing the appropriate number of clusters is the key step in clustering algorithms. As discussed earlier, we have chosen four clusters based on our observation of various types of servers. One way to validate the number of clusters is with the "elbow" method. The idea of the elbow method is to run k-means clustering on the dataset for a range of k values (say, k from 1 to 9) that calculates the sum of squared errors (SSE) for each value of k. The error decreases as k increases; this is because as the number of clusters increases, the SSE becomes smaller so the distortion also gets smaller. The goal of the elbow method is to choose an optimal k around which the SSE decreases abruptly (*i.e.,*



(a) Univesity campus. (b) Research institute.

Fig. 5. Hosts clustering results across one week.

ranging from 3 to 5 in our results, hence, k = 4 clusters seems a reasonable value for both the university and the research institute).

Clustering Results. We tuned the number of iterations and type of covariance for our clustering machine to maximize the performance in both enterprises. Tables 2 and 3 show the number of hosts identified in each cluster based on data from 3 May 2018. We also see the average value of various attributes within each cluster. For the cluster of name servers, *QryFracOut* approaches 0 in both organizations, highlighting the fact that almost all outgoing DNS packets from these hosts are responses rather than queries, which matches with the expected behavior. Having a high number of external clients served also indicates the activity of these hosts – in the University campus and research institute respectively 42 and 12 name servers collectively serve 84% (*i.e.,* $42 \times 2\%$ and $12 \times 7\%$) of external hosts.

Considering recursive resolvers in Tables 2 and 3, the average *QryFracOut* is close to 1 for both organizations as expected. It is seen that some of these hosts also answer incoming queries (from external hosts) possibly due to their misconfiguration. However, the number of external clients served by these hosts is very small (*i.e.,* less than 10 per recursive resolver) leading to an average fraction near 0. Also, looking at the number of external servers queried (*i.e., numExtSrv*), the average value of this attribute for recursive resolvers is reasonably high, *i.e.,* 14 and 4 hosts in the University and the research network respectively contribute to 84% and 80% of total *numExtSrv* – this is also expected since they commonly communicate with public resolvers or authoritative name servers on the Internet.

Hosts clustered as mixed DNS servers in both organizations have a moderate value of the *QryFracOut* attribute (*i.e.,* 0.57 and 0.21 for the University and the research network respectively) depending on their varying level of inbound/outbound DNS activity. Also, in terms of external clients and servers communicated with, the mixed servers lie between name servers and recursive resolvers. Lastly, regular end-hosts generate only outbound DNS queries (*i.e., QryFracOut* equals to 1), contact a small number of external resolvers, and are active for shorter duration of time over a day (*i.e., actvTimeFrac* less than 0.5).

Interpreting the Output of Clustering. Our clustering algorithm also generates a confidence level as an output. This can be used as a measure of reliability for our classifier. If adequate information is not provided by attributes of an instance then the algorithm will decide its cluster with a low confidence level.

The average confidence level of the result clustering is 97.61% for both organizations, with more than 99% of instances classified with a confidence-level of more than 85%. This indicates the strength of our host-level attributes, enabling the algorithm to cluster them with a very high confidence-level.

Server Clusters Across a Week. We now check the performance of our clustering algorithm over a week. Figure 5 shows a heat map for clusters of servers. Columns list server hosts that were identified in Tables 2 and 3 (*i.e.,* 70 hosts in the University network and 22 hosts in the research network). Rows display the cluster into which each server is classified. The color of each cell depicts the number of days (over a week) that each host is identified as the corresponding cluster – dark cells depict a high number of occurrences (approaching 7), while bright cells represent a low occurrence closer to 0.

In the University network we identified 42 name servers, shown by H1 to H42 in Fig. 5(a); the majority of which are repeatedly classified as a name server over a week, thus represented by dark cells at their intersections with the bottom row, highlighting the strong signature of their profile as a name server.

Among 14 recursive resolvers of the university campus, shown by H43 to H56 in Fig. 5(a); two of them (*i.e.,* hosts A and B in Fig. 1) are consistently classified as recursive resolver, and the rest are classified as either mixed DNS server or even end hosts (due to their varying activity). Lastly, 14 mixed servers, shown by H57 to H70 in Fig. 5(a), are classified consistently though their behavior sometimes is closer to a resolver or a name server.

Our results from the Research Institute network are fairly similar – Fig. 5(b) shows that hosts H1-H12 are consistently classified as name servers, while hosts H13-H16 are recursive resolvers and H17-H22 are mixes servers.

3.2 Server Ranking

Our system discovered 5097 public-facing (non-DNS) servers in the University, and 6102 at the Research Institute. However, only top 368 and 271 of these servers respectively appeared in the answer section of more than 100 outgoing DNS responses over a day. Additionally, 6 top ranked DNS servers, in each organization, contribute to more than 90% of outgoing queries and responses. Servers ranking provides network operators with the visibility into the criticality of their internal assets.

3.3 IT Verification

IT departments of both organizations were able to verify the top ranked DNS resolvers, name-servers, and non-DNS public-facing servers found, as they are directly configured and controlled by IT departments of the two organizations,(*e.g.,* major name-servers and web-servers). Additionally, we revealed unknown servers configured by departments of the two enterprises (we verified their functionality by reverse DNS lookup and their IP range allocated by IT

departments). Interestingly, 3 of the name-servers our method identified were implicated in a DNS amplification attack soon after, and IT was able to confirm that these were managed by affiliated entities (such as retail stores that lease space and Internet connectivity from the University) - this clearly points to the use of our system in identifying and classifying assets whose security posture the network operators themselves may not have direct control over.

3.4 Clustering of End-Hosts: NATed or Not?

Lastly, to draw more insights we further applied our clustering algorithm (using the same attributes introduced in Sect. 2.2) to IP address of end-hosts, determining whether they are behind a NAT gateway or not (*i.e.,* two clusters: NATed and not-NATed). In both networks, all WiFi clients are behind NAT gateways. Additionally, some specific departments of the two enterprises use NAT for their wired clients too. We verified our end-host clustering by reverse lookup for each enterprise network. Each NATed IP address has a corresponding domain name in specific forms configured by IT departments. For example the University campus wireless NAT gateways are associated with domain-names as "`SSID-pat-pool-a-b-c-d.gw.univ-primay-domain`", where "`a.b.c.d`" is the public IP address of the NAT gateway, and `SSID` is the WiFi SSID for the University campus network (we will disclose `SSID` and `univ-primay-domain` when this paper is de-anonymized). Similarly, in the Research institute NAT gateways use names in form of "`c-d.pool.rsch-primary-domain`" where "`c.d`" is the last two octets of the public IP address of the NAT gateway in the Research institute. On 3rd May, our end-host clustering shows that 292 and 19 of end-hosts IP addresses are indeed NATed in the University campus and the Research institute respectively – we verified their corresponding domain names configured by their IT departments.

We note that the two clusters of end-hosts are distinguished primarily by one attribute *actvTimeFrac* – a NATed IP address (representing a group of end-hosts) is expected to have a longer duration of DNS activity compared to a not-NATed IP address (representing a single end-host)[4]. We observe that some IPs with domain-names of NAT gateways are incorrectly classified as not-NATed end-hosts. This is because their daily DNS activity was fairly low, *i.e.,* less than an hour. On the other hand, not-NATed end-hosts with long duration of DNS activity (*i.e.,* almost the whole day) were misclassified. Verifying end-hosts classified as NATed, 84.3% of them in the University campus and 86% in the Research institute have corresponding domain-names as for NAT gateways allocated by IT departments. For end-hosts classified as not-NATed, 80.7% and 90.0% in the respective two organizations do not map to any organizational domain-names.

[4] We omit CCDF plots due to space constraint, they are shown in Appendix 2.

Looking into the performance of end-hosts clustering across a week, we note that 78.3% end-hosts in the University campus are consistently labeled as NATed over 7 days[5]. However, for the research institute, only 32.0% of NATed IPs are consistent across the entire week – 34.5% of IPs were absent on some days and the remaining 33.4% were misclassified as not-NATed for their low activity (*e.g.,* only active 2 h during a day).

4 Related Work

DNS traffic has been analyzed for various purposes, ranging from measuring performance (effect of Time-to-Live of DNS records) [3,7,13] to identifying malicious domains [2,8,9] and the security of DNS [5,10,11,14]. In this paper we have profiled the pattern of DNS traffic for individual hosts of two enterprise networks to map network assets to their function and thereby identify their relative importance for efficient monitoring and security.

Considering studies related to malicious domains, [8] inspects DNS traffic close to top-level domain servers to detect abnormal activity and PREDATOR [9] derives domain reputation using registration features to enable early detection of potentially malicious DNS domains without capturing traffic. From a security viewpoint, the authors of [5] study the adoption of DNSSEC [1], highlighting that only 1% of domains have implemented this secure protocol due to difficulties in the registration process and operational challenges; [10,11] focus on authoritative name servers used as reflectors in DNS amplification attacks; some researchers [14] have reported that the amplification factor of DNSSEC is quite high (*i.e.,* up to 44 to 55) whereas this measure is 6 to 12 for regular DNS servers.

DNS data can be collected from different locations (such as from log files of recursive resolvers [4,7] or authoritative name servers) or with different granularity (such as query/response logs or aggregated records). Datasets used in [5,10,11] contain DNS traffic for top level domains such as `.com`, and `.net`. We collect our data at the edge of an enterprise network, specifically outside the firewall at the point of interconnect with the external Internet. We note that while using data from resolver logs can provide detailed information about end hosts and their query types/patterns, this approach limits visibility and may not be comprehensive enough to accurately establish patterns related to the assets of the entire network.

5 Conclusion

Enterprise network administrators find it challenging to track their assets and their network behavior. We have developed an automated method to map internal hosts of an enterprise network by focusing only on DNS traffic which carries a wealth of information, constitutes a tiny faction of total network traffic and is

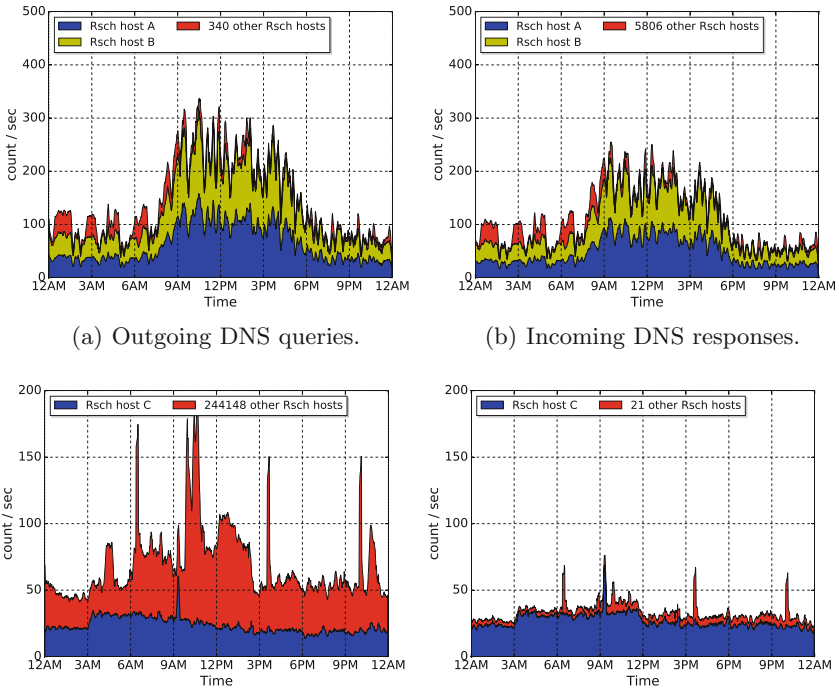
[5] We omit consistency plots due to space constraint, they are shown in Appendix 2.

easy to capture. By analyzing real-time live streams of DNS traffic from two organizations we highlighted how DNS query and response patterns differ amongst recursive resolvers, authoritative name servers, and regular hosts. We then identified key DNS traffic attributes that can be extracted efficiently in real-time. Lastly, we developed an unsupervised machine learning model that can classify enterprise assets, and we further applied our technique to infer the type of an enterprise end-host (NATed or not-NATed). Our results have been verified with IT departments of the two organizations while revealing unknown knowledges.

Acknowledgements. This work was completed in collaboration with the Australian Defence Science and Technology Group.

Appendix 1. DNS Behavior of Hosts (Research Institute)

(see Figs. 6 and 7).



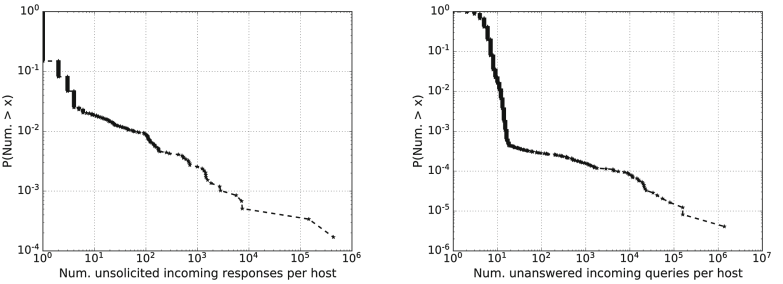
(a) Outgoing DNS queries.

(b) Incoming DNS responses.

(c) Incoming DNS queries.

(d) Outgoing DNS responses.

Fig. 6. Research institute: outgoing queries, incoming responses, incoming queries and outgoing responses, measured on 3 May 2018.

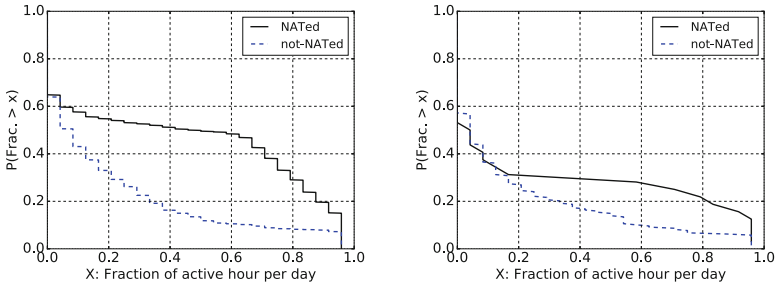


(a) Unsolicited incoming responses. (b) Unanswered incoming queries.

Fig. 7. Research institute: CCDF of (a) unsolicited incoming responses and (b) unanswered incoming queries per host, measured on 3 May 2018.

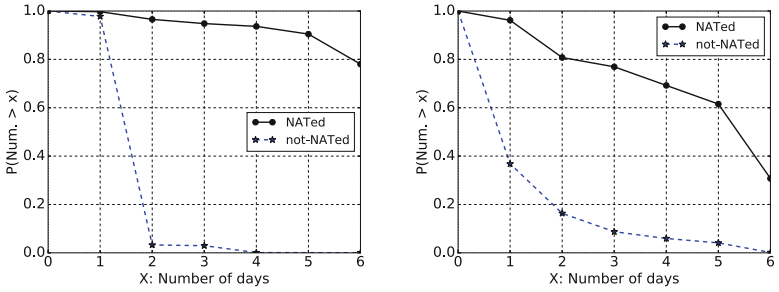
Appendix 2. NATed vs. not-NATed End-Hosts

(see Figs. 8 and 9).



(a) University campus. (b) Research institute.

Fig. 8. CCDF: fraction of active hour per day for end-host IP addresses with/without domain names.



(a) University campus. (b) Research institute.

Fig. 9. CCDF: Consistency of end-hosts clustering across a week.

References

1. DNS Security Introduction and Requirements (2018). https://www.ietf.org/rfc/rfc4033.txt. Accessed 28 May 2018
2. Ahmed, J., Gharakheili, H.H., Russell, C., Sivaraman, V.: Real-time detection of DNS exfiltration and tunneling from enterprise networks. In: Proceedings of IFIP/IEEE IM, Washington DC, USA, April 2019
3. Almeida, M., Finamore, A., Perino, D., Vallina-Rodriguez, N., Varvello, M.: Dissecting DNS stakeholders in mobile networks. In: Proceedings of ACM CoNEXT, Incheon, Republic of Korea, December 2017
4. Choi, H., Lee, H.: Identifying botnets by capturing group activities in DNS traffic. Comput. Netw. **56**(1), 20–33 (2012)
5. Chung, T., et al.: Understanding the role of registrars in DNSSEC deployment. In: Proceedings of ACM IMC, London, UK, November 2017
6. Deloitte: Elevating cybersecurity on the higher education leadership agenda (2018). https://www2.deloitte.com/insights/us/en/industry/public-sector/cybersecurity-on-higher-education-leadership-agenda.html
7. Gao, H., et al.: Reexamining DNS From a global recursive resolver perspective. IEEE/ACM Trans. Netw. **24**(1), 43–57 (2016)
8. Hao, S., Feamster, N., Pandrangi, R.: Monitoring the initial DNS behavior of malicious domains. In: Proceedings of ACM IMC, Berlin, Germany, November 2011
9. Hao, S., Kantchelian, A., Miller, B., Paxson, V., Feamster, N.: PREDATOR: proactive recognition and elimination of domain abuse at time-of-registration. In: Proceedings of ACM CCS, October 2016
10. MacFarland, D.C., Shue, C.A., Kalafut, A.J.: Characterizing optimal DNS amplification attacks and effective mitigation. In: Proceedings of PAM, New York, NY, USA, March 2015
11. MacFarland, D.C., Shue, C.A., Kalafut, A.J.: The best bang for the byte: characterizing the potential of DNS amplification attacks. Comput. Netw. **116**(C), 12–21 (2017)
12. Marshall, S.: CANDID: classifying assets in networks by determining importance and dependencies. Technical report, Electrical Engineering and Computer Sciences, University of California at Berkeley, May 2013
13. Müller, M., Moura, G.C.M., de O. Schmidt, R., Heidemann, J.: Recursives in the wild: engineering authoritative DNS servers. In: Proceedings of ACM IMC, London, UK, November 2017
14. van Rijswijk-Deij, R., Sperotto, A., Pras, A.: DNSSEC and its potential for DDoS attacks: a comprehensive measurement study. In: Proceedings of ACM IMC, Vancouver, BC, Canada, November 2014