Enabling Network Security Through Active DNS Datasets

Athanasios Kountouras[1], Panagiotis Kintis[2], Chaz Lever[2], Yizheng Chen[2], Yacin Nadji[2], David Dagon[1], Manos Antonakakis[1], and Rodney Joffe[3]

[1] Georgia Institute of Technology, School of Electrical and Computer Engineering,
{kountouras,manos}@gatech.edu, dagon@sudo.sh
[2] Georgia Institute of Technology, School of Computer Science,
{kintis,chazlever,yzchen,yacin}@gatech.edu
[3] Neustar, rjoffe@centergate.com

Abstract. Most modern cyber crime leverages the Domain Name System (DNS) to attain high levels of network agility and make detection of Internet abuse challenging. The majority of malware, which represent a key component of illicit Internet operations, are programmed to locate the IP address of their command-and-control (C&C) server through DNS lookups. To make the malicious infrastructure both agile and resilient, malware authors often use sophisticated communication methods that utilize DNS (i.e., domain generation algorithms) for their campaigns. In general, Internet miscreants make extensive use of *short-lived disposable domains* to promote a large variety of threats and support their criminal network operations.

To effectively combat Internet abuse, the security community needs access to freely available and open datasets. Such datasets will enable the development of new algorithms that can enable the early detection, tracking, and overall lifetime of modern Internet threats. To that end, we have created a system, Thales, that actively queries and collects records for massive amounts of domain names from various seeds. These seeds are collected from multiple public sources and, therefore, free of privacy concerns. The results of this effort will be opened and made freely available to the research community. With three case studies we demonstrate the detection merit that the collected active DNS datasets contain. We show that (i) more than 75% of the domain names in public black lists (PBLs) appear in our datasets several weeks (and some cases months) in advance, (ii) existing DNS research can be implemented using only active DNS, and (iii) malicious campaigns can be identified with the signal provided by active DNS.

1 Introduction

The Domain Name System (DNS) is a fundamental component of the Internet. Most network communication on the Internet starts with a DNS lookup that maps a domain name to a corresponding set of IP addresses. Cyber criminals frequently leverage DNS to provide high levels of network *agility* for their illicit

operations. For example, most malware relies on DNS to locate its command-and-control (C&C) servers. Such servers are used to send commands from the attacker, exfiltrate secret information, and send malware updates.

DNS abuse is an enduring, if not permanent, feature of the Internet, which might at best be managed through various policies, remediation technologies and defenses. Traditionally, network operators have relied on *static blacklists* to detect and block DNS queries to malware domains. Unfortunately, static blacklists, which are often manually compiled, cannot keep pace with the quantity of network agility of modern threats. This results in blacklists that are incomplete and become outdated quickly.

To overcome the limitations of static blacklists, new analytical systems have been proposed [12,15,13,14,26,29] to shorten the response time necessary to react to new threats and secure networks. Those systems rely on the efficient collection and presentation of passive DNS datasets. However, such datasets are difficult to find, challenging to collect, and often require restrictive legal agreements. These obstacles make further innovation difficult and are an impediment to repeatability of research.

The lack of open and freely available DNS datasets puts the security community at a disadvantage because they lack access to datasets describing a critical component used by adversaries on the Internet. Clearly, the security community is in need of open, freely available DNS datasets than can help increase the situational awareness around modern threats. This is illustrated by the fact that most modern threats rely on DNS for their illicit activities.

This paper provides a solution aimed at filling this gap. We introduce the concept of active DNS and discuss a new large scale system, Thales, which is able to systematically query and collect large volumes of active DNS data. The output of this system is a distilled dataset that can be easily used by the security community. Thales has been reliably active for more than six months and collected many terabytes of DNS data, while causing only a handful of abuse complaints. Access to this dataset is currently available to the community from the following project website: `http://www.activednsproject.org/`.

In summary, our paper makes the following contributions:

1. We present a system, Thales, that can reliably query, collect, and distill active DNS datasets. Due to the public nature of our seed data, our active DNS datasets do not contain any potentially sensitive information that preclude their use by the security community. Thales has been collecting active DNS data for more than six months with almost zero down time (only three days). During this time, the system has generated more than a terabyte of unprocessed DNS PCAPs along with tens of gigabytes of de-duplicated DNS records per day. Thus, the active DNS datasets represent a significant portion of the world's daily DNS delegation hierarchy.

2. We provide in-depth comparison between the newly collected active DNS datasets and passive DNS collected from a large university network. We show that the active DNS datasets provide greater breadth (i.e., reaches out to a larger portion of the IPv4, IPv6, and DNS space). Conversely, passive DNS

yields a denser graph between the queried domain names and the remaining IP and DNS infrastructure.

3. We practically explore how active DNS can be used to improve the security of modern networks through several case studies. We show that the active DNS datasets can be use for early detection of financial and other Internet threats. Our analysis shows that more than 75% of malicious domain names appear in the active DNS datasets months before they get listed in a public blacklist. We demonstrate how active DNS can be used to implement and extend existing DNS related research, specifically, by implementing an algorithm used to detected potential domain ownership changes. Finally, we show how active DNS can be used as a signal to identify malicious campaigns on the Internet.

2 Active DNS Data Collection

With this section we introduce Thales. We will begin by discussing the network and system infrastructure necessary to systematically and reliably collect the active DNS datasets. Then, we will discuss the details of the domain names that compile the daily seed for Thales. The section will be concluded by discussing the long term measurement behind the collected active DNS datasets.

2.1 Infrastructure

The reliable collection of DNS data is far from easy. Thales was designed to retain high levels of availability, efficiency and scalability. The goal of Thales is clear; the generation of active DNS datasets that will provide systematic snapshots of the DNS infrastructure, several times per day. These datasets will enable the security community to construct a timeline of the evolution of threats in the broader Internet.

Our system, Thales, is composed of two main modules as seen in Figure 1: (a) the traffic generator and (b) the data collector. The first is responsible for generating large numbers of DNS queries using a list of seed domain names as an input to the system. The second module is responsible for collecting the network traffic and guiding these raw DNS datasets for further processing (i.e., data deduplication).

Traffic Generation In order to achieve high availability, redundant systems are used to generate traffic. Linux containers (LXC) [7] are setup across several physical systems, creating a DNS scanning cluster of 30 LXC containers. Each LXC contains its own local recursive software [4] and is assigned a job, where a subset of the overall daily seed domain names will have to be resolved by a particular container. High efficiency is achieved by increasing the rate of DNS

---

[4] We used the Unbound (`https://www.unbound.net/`) recursive software in every LXC container.
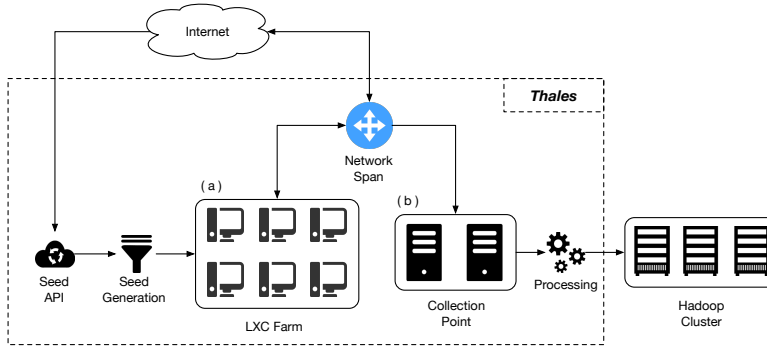
Fig. 1: The Seed API is responsible for collecting the seed domains from various sources and the Seed Generation reduces them to a list of unique domains. The LXC Farm corresponds to the query generator which is connected to the internet through a Network Span. That in turn is sending traffic to the Collection Point from where data is being reduced and stored for long term on our Hadoop Cluster.

resolution requests (a.k.a. queries per second) that can be handled by the recursive in the LXC container. However, just increasing the resources of the LXC container will not suffice for the container to handle a large enough number of DNS requests. This is because the local recursive in the LXC is bounded by the maximum number of ports that can be used for UDP sockets. This means that the number of requests that can be sent by a host have to be limited to the number of available concurrent ports that the local recursive (in the LXC container) can handle.

At any given point in time, a container could theoretically handle up to 64,512 ($2^{15} - 1024$) sockets per IP address – and therefore 64,512 UDP query packets in transit. The LXC containers support custom network interfaces, which support assigning a different IP address to each container. More specifically, we use 30 contiguous IPs out of an assigned IP block of 63 available addresses (/26). Thus, they are able to send and receive up to $30 \times 64,512 \approx 2^{21}$ simultaneous DNS resolution requests from the infrastructure. These results are achieved by deploying the containers on two physical systems. Each of these two systems has 64 processing cores and 164GB of RAM. It is worth pointing out that using LXC containers allows us to scale the infrastructure horizontally by simply adding more systems to our scanning cluster.

Data Collection The requests submitted by Thales are collected at two vantage points. The first one is on the LXC container that has submitted the resolution request for a given domain name, whereas the second one is at the SPAN of a switch that routes traffic for all our containers. As mentioned earlier, we are utilizing several IP addresses from several local virtual LANs (VLAN). These

VLANs have been "trunked" to a single 1Gbit interface on a host that collects all port 53 UDP traffic. We are collecting traffic at both points for redundancy and verification of correctness for the daily active DNS datasets.

```
{
    "date": "20160303",
    "qname": "0746jiaoyou.com.",
    "qtype": 1,
    "rdata": "61.151.239.202",
    "ttl": 3600,
    "authority_ips": "58.216.26.232,120.52.19.142",
    "count": 5,
    "hours": 32778,
    "sensor": "active-dns"
}
```

Fig. 2: A sample record from our dataset that shows the data fields that are stored. The `authority ips` field represents the authoritative nameservers that replied for this domain name and the `hours` variable captures the hour of the day that this record was seen in a 24 bit integer.

Capturing network traffic results (on average) in a massive 1.67TB of raw data in *packet capture* format (pcap). This data is transferred in a local Hadoop cluster composed of 22 data nodes. The Hadoop cluster is responsible for parsing the pcap files, deduplicating the resource records (RRs) and converting the RRs into meaningful DNS tuples of following format: `(date, QNAME, QTYPE, RDATA, TTL, authorities, count)` as seen in Figure 2. Deduplication is a critical step, since many responses we collect remain the same throughout a day. Thus, after removing duplicate RRs, we are left (on average) with approximately 85GB of data per day. Detailed measurements for both daily raw and deduplicated RRs will be discussed in Section 2.3.

2.2 Domain Seed

Before Thales can begin scanning the domain name system, it has to be provided with a list of domain names that will act as candidates for resolutions. We will refer to these domain names as the *seed* for Thales. The seed is an aggregation of publicly accessible sources of domain names and URLs that we have been collecting for several years. These include but are not limited to Public Blacklists, the Alexa list, the Common Crawl project, and various Top Level Domain (TLD) zone files.

More specifically, we are using the zone files that are published daily by the administrators of the zones for com, net, biz and org. In Figure 3 we present the number of domains obtained by each zone file. Because of the relative number of small daily changes, compared to the size of the zone files, the daily changes are not that apparent in Figure 3. We note that the number of domains obtained by
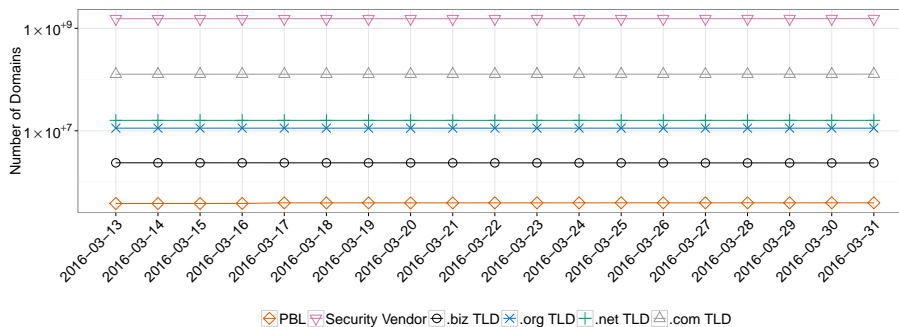
Fig. 3: Number of domains over time per seed input. The security vendor list contains about 1.5 billion domains and from the TLDs com is obviously the largest one with about 127 million domains.

zone files changes as new domains get registered and old ones expire (and get removed from the zone). In Thales we input these zone files that we collect daily to our domain seed. This way our seed includes the current state of each zone every day.

We also add the entire Alexa [3] list of popular domains to the domain seed. This provides us with a large number of domains that would most likely be queried in a network by users.

In order to capture domains that might not be available in one of the zone files, we built a crawler that collects and parses domains seen in the Common-Crawl dataset [4]. The Common-Crawl dataset is an open repository of web crawl data that offers large volumes of crawled pages to anyone. We used components (i.e., URLs, HTML code) from the common crawl dataset to extract only the domains of the pages visited. Due to the size of even the Common-Crawl "metadata section" from the common crawl, we are still using the data published for last September 2015 and will start updating that list regularly. Because the common crawl data is published in monthly releases, the domain list that we extract from it and use in our seed list remains the same between updates.

A different list of data that we utilize in our domain seed is a feed of interesting domains that have been provided to us by a security company. This feed provides us with domains that have been observed to engage in forms of potentially malicious Internet activity. Because the feed provides us with new domain names constantly, we gather all new information and append it to the already existing list of interesting domains. We push the updated list to our collection infrastructure daily. The feed provides us with tens of thousands of new domains each day, making this list one of the fastest growing lists we use.

Finally we use a collection of public blacklist data in order to provide our data with interesting hand curated domains that originate from malicious activity.
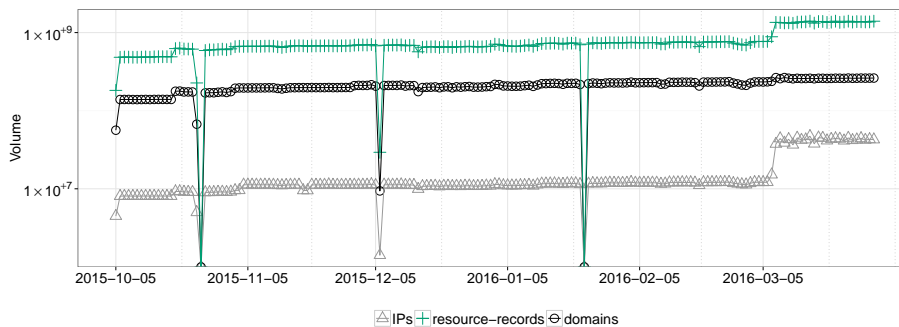
Fig. 4: Volumes of IPs, resource records and domains observed with Thales. March 7th was the day when we started querying for the QTYPEs: SOA, AAAA, TXT and MX. There have been two full outages on October 25, 2015 and January 23, 2016. On December 6, 2015 we had an outage that lasted for most of the day but we were able to recover the system later in the day.

More specifically the public blacklists we employ are: Abuse.ch [2], Malware DL [9], Blackhole DNS [8], sagadc [10], hphosts [6], SANS [11] and itmate [1]. We aggregate these lists daily and we input them into our domain seed by replacing the old list.

2.3   Measurements

Thales has been collecting data for a little less than six months. For the purpose of this paper we are focusing on analyzing all data in this section and then limit in depth analysis to the last 12 days of March (the last full week forth) for more specific measurements, unless a different window is explicitly stated. Over six months, Thales identified approximately 10,714,784 unique IP addresses, 199,110,841 unique domain names and 662,319,389 unique RRs per day. Figure 4 shows the distribution of IP addresses, domain names and RRs on average per day from October 5th to March 3rd 2016.

During these months, we experienced two outages. The first was when the system was initially setup because of an update which was not rolled out correctly and caused the system to go off-line. Therefore, there is no data available for October 25, 2015, and policy has been updated to avoid future interruption since then. On January 23, 2016, our campus data center was undergoing maintenance for the cooling infrastructure, which caused a temporary shutdown of all our systems. Such cases can now be mitigated by Thales. We have made the system portable, which gives us the ability to move it to another location within a day's prior notice. Also on December 6, 2015 early in the day we had a hardware failure on our system that was detected early in the morning. We were able to recover the system and perform a check of the system by the same afternoon.
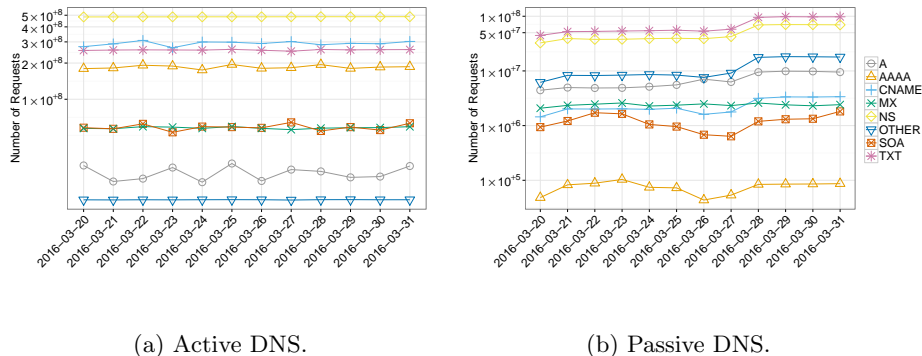
7

(a) Active DNS.   (b) Passive DNS.

Fig. 5: The distribution of different query types (QTYPE) in the active (left) and passive (right) DNS datasets. The active DNS dataset is almost sustaining the same volume of records per day, whereas the passive DNS dataset is fluctuating more over time. Note the growth after March 28, when the Spring Break was over and the Institute was operating at full capacity again.

After the system check, we immediately restarted the collection, but there was not enough time in the day to go through the entirety of data in our seed list. This is depicted by the significant dip in the data. This incident was not a full outage since we were able to collect some data for the day.

3 Comparing Active And Passive DNS Datasets

Passive DNS has been an invaluable weapon in the community's arsenal for research combatting malware, botnets, and malicious actors [12,13,14,28,22]. Passive DNS, though, is rare, difficult to obtain, and often comes with restrictive legal clauses (i.e., Non Disclosure Agreements). At the same time, laws and regulations against personal identifiable information (PII), the significant financial cost of the passive collection, and storage infrastructure are some of several reasons **that make passive DNS cumbersome**. The primary goal for the active DNS dataset is to reduce the barrier for (repeatable) security research on DNS. In this section, we show how active DNS relates and contrasts to passive DNS. We will see that, while not a true replacement for passive DNS, Thales is able to create active DNS datasets that in many cases contain an order of magnitude more domain names and IP addresses.

3.1 Datasets

We first discuss how we obtain our passive DNS datasets. Our passive DNS dataset consists of traffic collected at our university network. The collection point is both *below* and *above the recursive*. This means that we collect the responses on the both paths; (1) between the (anonymized) clients and the local

recursives and (2) between the local recursives and the upper layers of the DNS hierarchy (i.e., name servers, top level domains, etc.). For the active and passive DNS comparison, we decided to utilize datasets collected during the entire month of March 2016.

Figures 6 and 5 show eight detailed plots of the distribution of records in both our active and passive DNS datasets. Note that all plots are log-scale for the y-axis. As we can see, the active DNS dataset does not fluctuate a lot, compared to the passive DNS one. This is primarily an artifact of the collection technique, since the daily changes in the domain name seed we are using is minimal. On the other hand, the passive DNS dataset, is is primarily driven by the behavior of the users on the local network, which may fluctuate on weekends, holidays, and during certain periods such as exams. This also explains the sudden increase in traffic for passive DNS, since our campus network experienced a reduction in traffic from March 21st until March 25th during spring break. Therefore, Figure 6c shows an increase to more than double the unique resource records (RRs) identified per day after Monday, March 28th, when the spring break ended. Table 1 shows a breakdown of the datasets over the last 12 days of March, in much greater detail.

It is worth noting that Thales is able to generate an order of magnitude more unique domain names, IP addresses and RDATA in the active DNS dataset (see Figure 6, subfigures a to e), in comparison to the passive DNS data collected in a large university. This means that in actual DNS records, the active DNS dataset is more than comparable to the passive DNS that someone can collect in a large university. Now, as we can see from Figure 6, (f), active DNS is not able to create as dense graphs of resource records, as someone would expect to find in passive DNS data. This is somewhat to be expected, as in active DNS, Thales is scanning all possible domain names that can be seen in our public sources. This inevitably will include domain names that are rare, and in the context of a graph compiled by RRs, they will form islands. While not necessarily bad, we would advise researchers to take cautionary sanity steps when they utilize the active DNS data for spectral processes.

The diversity of the different query record types (QTYPEs) we are able to identify, in the two different datasets compared can been seen in Figures 5a and 5b. Although there is a big difference regarding the volume of the records available, on average the visibility is very similar, since we are collecting the most popular QTYPEs when querying for the active DNS datasets.

4   Case Studies

To this point, we exposed several of the data properties from the active DNS datasets. In this section, we demonstrate the security value of these new active DNS datasets. We should clarify that our goal is not to claim as a contribution any of the following abuse detection processes. All of them have been discussed by previous work in the field. Rather, our goal is to practically demonstrate,
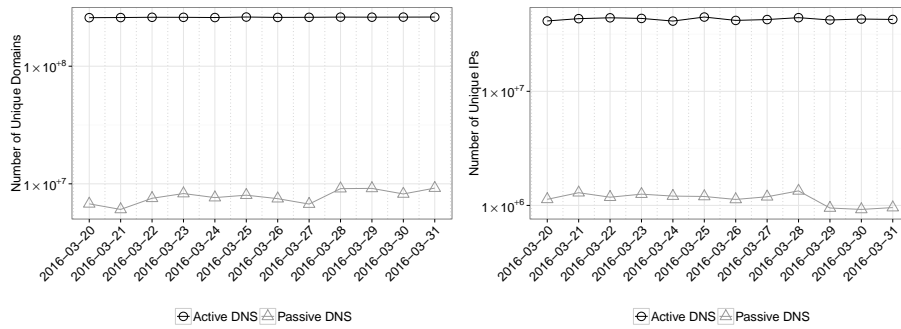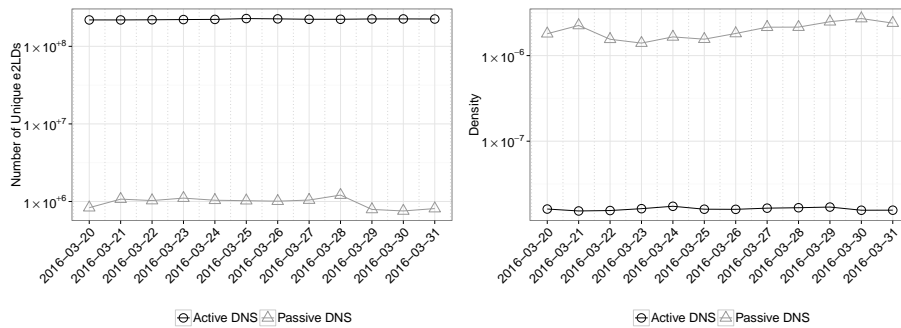
(a) Unique domain names per day.

(b) Unique IP addresses per day.
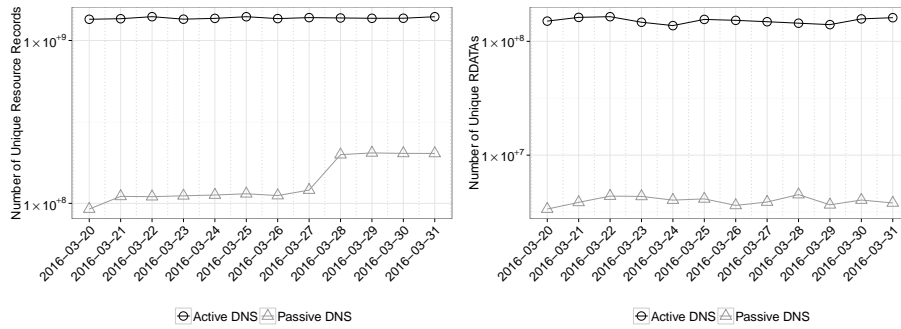
(c) Unique resource records (RR) per day.

(d) Unique responses (RDATA) per day.

(e) Unique effective second level domain names per day.

(f) The density of the Resource Records graph in the active and passive DNS dataset.

Fig. 6: The distribution of different records in our active and passive DNS datasets. The plots show that Thales is able to generate orders of magnitude more data than the passive DNS collection engine (Figures a to e) and much more diverse (Figure f).

10

	Domains		IPv4/IPv6		RDATA		RR		e2LD	
Date	Active	Passive	Active	Passive	Active	Passive	Active	Passive	Active	Passive
3/20	258,702	6,759	41,360	1,130	150,629	3,356	1,350,118	92,218	219,009	831
3/21	259,305	6,056	43,333	1,292	162,366	3,845	1,360,660	110,379	219,009	1,072
3/22	260,676	7,535	44,090	1,180	164,685	4,364	1,400,427	109,896	219,985	1,028
3/23	260,420	8,267	43,538	1,255	147,190	4,338	1,352,019	111,247	221,466	1,105
3/24	259,389	7,635	41,273	1,206	137,491	4,024	1,367,554	112,513	222,464	1,037
3/25	261,883	8,008	44,769	1,197	155,830	4,125	1,399,724	114,518	228,119	1,024
3/26	260,011	7,479	41,830	1,127	152,918	3,616	1,362,978	111,646	226,030	1,009
3/27	260,506	6,727	42,556	1,190	148,728	3,871	1,382,096	120,624	223,313	1,043
3/28	261,551	9,100	44,216	1,340	144,365	4,499	1,375,399	199,023	223,345	1,208
3/29	261,171	9,145	42,189	948	140,225	3,658	1,369,100	204,017	225,513	789
3/30	261,513	8,200	42,992	921	157,477	4,030	1,370,090	202,702	225,642	754
3/31	261,766	9,195	42,651	956	161,387	3,798	1,399,218	202,511	225,128	809

Table 1: Number of data points collected over the last 12 days of March 2016. Values are in thousands ($\times 10^3$).

	Aggregate ($\times 10^3$)		Mean		Median	
QTYPE	Active	Passive	Active	Passive	Active	Passive
A	3,082,960	813,485	256,913,375.92	67,790,485.33	257,181,439.5	54,989,441.0
AAAA	292,278	81,992	24,356,555.67	6,832,692.33	23,918,026.5	5,920,971.5
CNAME	174,881	136,901	14,573,484.5	11,408,450.0	14,582,732.0	8,495,216.5
MX	2,222,465	908	185,205,470.67	75,690.83	184,075,003.5	83,309.0
NS	5,822,874	586,695	485,239,507	48,891,296.25	485,117,732.0	39,316,201.5
SOA	3,498,172	28,162	291,514,366.5	2,346,885.75	291,172,940.5	2,022,850.0
TXT	701,689	14,499	58,474,102.67	1,208,253.83	58,304,209.5	1,205,094.5
Other	694,067	28,655	57,838,938.5	2,387,929.75	57,693,964	2,380,550

Table 2: The distribution of QTYPEs for the active and passive DNS in our datasets.

using the actual active DNS datasets, the security merit that active DNS data can offer to the research and operational communities.

4.1 Enhancing Public Blacklists

Due to the nature of Thales we can make use of the collected data in ways that can reveal abuse signal about domains before they are identified as actual malicious use. Blacklisted domains, for example, are an interesting category of candidate indicators of abuse that can be registered, set-up, and pointed to an IP location well before they are actually used in malicious activities. Thus, active DNS could be used as a potential source of raw datasets that can be used for timely domain abuse detection.

As we have already discussed, alongside the active DNS data collection, we were also able to gather a plethora of public domain name blacklists. As expected, domain names in these blacklists also appeared in the active DNS traces we collected using the active DNS project. For all domain names seen in both the public blacklists and active DNS data, we identified two important dates. The first denotes the first day the domain name was probed by Thales. This behavior is driven by the addition of the domain in our seed list that can be caused by a change in any of the zone files collected daily from the top level domain authorities. The second important date we identified is the first day one of the many blacklists we collect (on a daily basis) actually listed this domain name as part of a particular abusive activity.

We compared the first seen dates of blacklisted domains and the first seen date of a domain resolved by Thales and we plotted the results in a cumulative distribution function (CDF) that depicts the time difference in days between a resolution in our passive or active DNS data and the appearance of the domain in a public blacklist. Negative values represent the number of domains that have first appeared in our active or passive DNS data before getting eventually blacklisted. On the other hand, positive values represent domains that had been blacklisted before they had a resolution in our data.

It is worth pointing out that not all the public domain names blacklists were used as a seed domain source for Thales, rather the ones that are described in Section 2.2. That is, we should expect a fair amount of both positive and negative values in these CDFs. Positive values indicate that a domain name was first seen in a blacklist and then in either the active or passive DNS data that we present in Figure 7, while negative values indicate that the domain was first seen in DNS before being blacklisted.

Thales resolves domains that came in part from zonefiles for major top-level domains. It queries any domain registered in that zone within a day after it was registered and added in the zonefile. This creates a temporal history of the DNS activity capable of describing the IP infrastructure history that supported the domain name, before blacklisting, at the time, and after it was blacklisted. This is a new property that active DNS datasets will freely offer to the security community, and it is a property that is rarely seen in passive DNS data. The reason for this behaviour that active DNS exhibits compared to passive
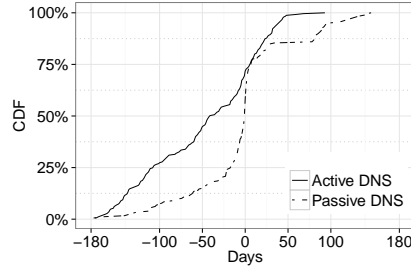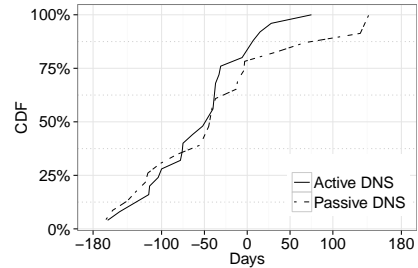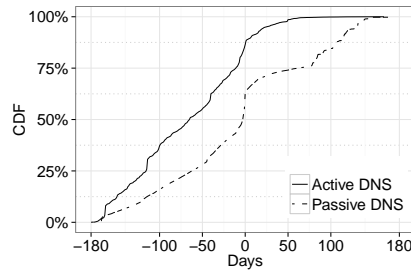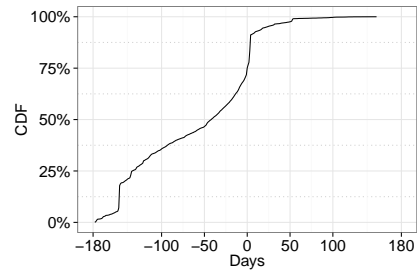
(a) Zeus malware.
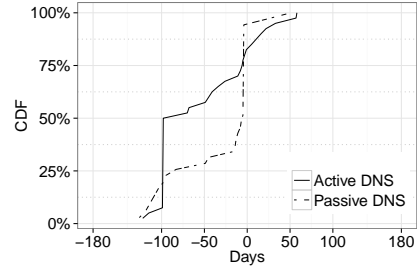
(b) Spam.
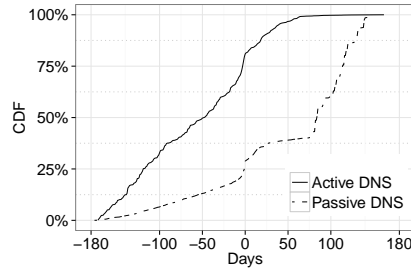
(c) Phishing.

(d) Exploit.

(e) Difference of days from the first time a domain name was seen in active and passive DNS before it appeared in a PBL.

(f) The difference between the first date a blacklisted domain was seen in active DNS versus the passive DNS dataset, for the domains that were seen before they were blacklisted. Approximately 70% of the 17,000 domains that exist in both datasets and were blacklisted, were first seen in active DNS.

Fig. 7: Cumulative distribution of the first seen date in active and passive DNS, subtracting the first seen date of the same domain in a PBL for Zeus, Spam, Phishing, and Exploit domains.

13

DNS is simple; infections get remediated and hosts are mobile, thus making it hard for the network operator to passively observe the network evolution of the infrastructure that supports a domain. Thus, Thales should be able to offer a strong signal augmenting existing passive DNS data to which researchers and network operators have access.

Figure 7 shows the CDF plots for different classes of malicious domain names (Figures 7a to 7d). The values plotted include the domains in our active and passive DNS datasets that have been blacklisted. Several instances of these domains are found in our dataset long before they are blacklisted; for example 50% of domain names associated with spam were queried approximately 2.5 months before they were blacklisted. On the other hand, we do not have the same visibility for ephemeral types of attacks, like phishing and exploit kits. In the latter two cases, approximately 75% of the domain names are queried by Thales at least one day earlier, with the 50% mark being at around 50 days earlier.

In total 42,000 domain names have been blacklisted and also appeared in our active DNS dataset. From this set, 30% were queried and data have been collected for approximately 100 days before the blacklisting instance (Figure 7(e)). For 75% of the blacklisted domain names, we have collected data for more than a week before they appeared on a PBL. Considering that PBLs have been used as ground truth for various security systems [26,30,23,21], we are planning to utilize this data over time to model the behavior of these domains and identify the threats long before current systems, or even before they are utilized by the adversaries.

On the other hand, we were able to identify 20,000 domain names in the passive DNS dataset that also appear in blacklists. The dashed line in Figure 7 plots represents these domain names. Approximately 50% of the domain names that are blacklisted appear in the passive DNS data feed, with only 25% revealing themselves 50 days earlier than the blacklisting event, as shown in Figure 7e. In this case, there are only 20,000 domain names that have been blacklisted and the visibility that we have is approximately 15% for the 100 days mark. About 50% of all the domain names were seen roughly two days before they were blacklisted. This clearly supports our claim about the merit of active DNS datasets, and how well they complement existing passive DNS repositories. The early linkage between domain names and IP infrastructure witnessed by the active DNS data will be able to enrich the signal that passive DNS data contains, potentially making local DNS modeling efforts easier for researchers and operators.

In most cases, the active DNS dataset contains domain names far before they appear in either the passive DNS or the blacklist dataset. Note that the intersection between active and passive DNS records that have been blacklisted is approximately 19,000. This is almost half of the domains in the active DNS dataset and 95% of the domain names in the passive DNS dataset. Passive DNS seems to show better results in early days for the spam domain names case (Figure 7b), but active DNS catches up very fast (within 15 days) and then loses the advantage again at the time of the blacklisting events (0 point in the plot).

Lastly, Figure 7f depicts the difference between the day a blacklisted domain name was first seen in our active DNS dataset and the day it was seen in our passive DNS dataset. This includes only the domain names that were seen before the PBLs included them. Approximately 17,000 domain names have been found in both active and passive DNS before they were blacklisted. The vast majority of them were first resolved by Thales, at least one day before it was visited by a system in our university. Approximately 40% of the domain names were already being resolved by Thales for more than 100 days before they appeared in the passive DNS dataset.

4.2   Enhancing The Detection Of Domain's Residual Trust Change

On the Internet, domain names serve as trust anchors for numerous systems and services, and for many, ownership of a domain is enough to prove one's identity. Work by Lever et. al [25] discussed the problems caused by the use of domains as trust anchors and showed that residual trust, implicitly inherited by domains after an ownership change, is a root cause of many seemingly disparate security problems. Therefore, identifying changes in ownership, due to expiration or some other cause, is an important problem in protecting against the abuse of residual trust. WHOIS [19] is typically used to discover more information about the owner of a particular domain, and thus, it would a appear to be a natural fit for creating a remedy to this problem. However, collecting WHOIS at scale is outside the grasp of most organizations due to rate limiting imposed on automated collection of WHOIS records. To make matters worse, these limits frequently vary by registrar, further adding to the complexity of collecting WHOIS data at scale. To circumvent this problem, Lever et al., proposed Alembic, a lightweight algorithm for locating potential ownership changes that relies solely on passive DNS. This algorithm relied upon three different components: changes in infrastructure, changes in lookup volume distribution, and change in SOA records.

While passive DNS is much easier collect, it is also very sparse, and this results in two limitations with respect to Alembic. Scores can only be computed for domains observed in passive DNS and that have sufficient historical resolutions. Active DNS can help improve upon these limitations. First, Figure 6e shows that active DNS captures many more effective second level domains than passive DNS. Given that the passive DNS dataset used for comparison was generated from a large university network, this result is particularly important. It demonstrates that even large networks have difficulty matching the breadth of domains that can be collected using active DNS querying. Next, active DNS querying can consistently gather specified DNS record types over time. In particular, Figure 5b and Figure 5a show that active DNS results in substantially more SOA records than passive DNS each day. Since one of the key components of the Alembic scoring is SOA records, active DNS should be able to enhance the performance of the Alembic scoring algorithm. While active DNS provides many benefits, it is important to note that the one component Active DNS cannot enhance is the lookup volume distribution of domains. This component is

derived by user behavior observed in passive DNS, and therefore, there is no analog in the active DNS dataset.

To evaluate whether Alembic could work using only active DNS, we implemented a modified version of the algorithm that excluded lookup volume distribution as a component and used a fixed window size of two weeks. Then we computed scores for March 27, 2016 using our modified algorithm. In total, this resulted in 63,332,836 domains with non-zero scores, where larger scores indicate a higher confidence in an ownership change. The distribution of those scores can be seen in Figure 8. The majority fell in the range between 0.4 and 0.5, and further inspection re-
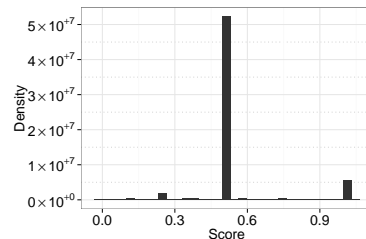

Fig. 8: Histogram showing the distribution of Alembic scores for March 27, 2016.

vealed that the SOA component contributed the most to these scores. In short, most of the scores in this range were a result of changes in the SOA record for the domains. Since we saw very little change in hosting infrastructure, it is possible these scores could simply be the result of minor changes within the SOA record. The next largest range was between 0.9 and 1.0 and consisted of 5,652,910 domains. According to the algorithm, domains with a score in this range are most likely to have undergone a change in ownership. 5,625,397 (99.5%) of these domains had a score of 1.0, indicating that both infrastructure and SOA records had undergone complete changes. Indeed, we found 10,885 of these domains on a public service's list [5] of expired domains for March 27, 2016. The remainder of these domains provide interesting cases for further study.

Our modified version of the Alembic algorithm, originally proposed by Lever et al., provides an interesting example of how active DNS can be used to enhance or extend existing research. Without active DNS, deploying an algorithm like Alembic would require access to a large scale passive DNS dataset (e.g., university, enterprise, Internet service provider). However, using openly available active DNS data, as offered by this research, can help remove the barriers to using or deploying existing DNS research.

4.3 Tracking Malicious Domain Names In Non-routable IP Space

Bogons are private, reserved, or otherwise unallocated network blocks [32,18,34]. Bogons should be boring since by definition they should not be hosting anything in the context of the global Internet. But occasionally, a domain name, like `messisux.bix`, resolves to a bogon like `0.0.0.0` despite the fact this IP can not host anything. The presence of a domain name, however, indicates a service that should be globally reachable exists. These "nonsense" resolutions are at times caused by misconfigurations, brand protection services, and occasionally,

16

Operation Hangover	CopyKittens
`alertmymailsnotify[dot]com`	`alhadath[dot]mobi`
`cloudone-opsource[dot]com`	`big-windowss[dot]com`
`download-mgrwin[dot]com`	`cacheupdate14[dot]com`
`necessaries-documentation[dot]com`	`fbstatic-akamaihd[dot]com`
`newsfairprocessing[dot]com`	`fbstatic-a[dot]space`
`onestop-shops[dot]com`	`fbstatic-a[dot]xyz`
`servicesloginmail-process[dot]com`	`gmailtagmanager[dot]com`
`servicesprocessing[dot]com`	`haaretz[dot]link`
`websourceing[dot]com`	`haaretz-news[dot]com`
`worldvoicetrip[dot]com`	`mswordupdate15[dot]com`
	`mswordupdate16[dot]com`
	`mswordupdate17[dot]com`
	`patch7-windows[dot]com`
	`patch8-windows[dot]com`
	`patchthiswindows[dot]com`
	`walla[dot]link`
	`wethearservice[dot]com`
	`wheatherserviceapi[dot]info`
	`windowkernel[dot]com`
	`windows-drive20[dot]com`
	`windowskernel14[dot]com`
	`windows-my50[dot]com`
	`windowsupup[dot]com`

Table 3: Operation Hangover and CopyKittens Attack Group Infrastructure and Domain Names.

malicious actors. To investigate further, we don our threat researcher hats and analyze domain names that resolved to bogon IP space during our analysis. Here we focus on malicious infrastructure as it is a primary interest of the security community. However, we also note that active DNS data that resolves to bogons would be useful in other contexts such as identifying potential trademark infringements.

We identified two known malicious campaigns in the subset of bogon data: "Operation Hangover" and "CopyKittens." The former is infrastructure of a cyber espionage threat targeting government, military, and private sector networks with some ties to India [17]. Domain names seen in active DNS data for this threat are shown on the left hand side in Table 3. The latter is infrastructure for threats targeting "high ranking diplomats at Israel's Ministry of Foreign Affairs and some well-known Israeli academic researchers specializing in Middle East Studies" [33] and its active DNS domains are shown on the right column in Table 3.

These are useful indicators despite the fact these attacks are known and likely inactive. Neutered, yet unidentified, infections are likely still operating in networks today, which should lead to incidence responses and damage assess-

ments. For example, knowing the specific internal machine that was infected with targeted malware is useful even after an attack has taken place. An end-user machine on a company's corporate network has different implications than a locked down server in a data center, or the CEO's personal laptop. Interestingly, some targeted threats do resolve to bogon space, while active, to reduce their network footprint [27]. This suggests signal for malicious detection in active DNS's non-routable IPs.

5 Related Work

The collection of passive DNS data has been proposed by Weimer et al. [35] over a decade ago as a method that network operators could use to investigate security events in their environments. Zdrnja et al. [36] was the first to discuss how passive DNS data can be used for spotting security incidents using domain names. Notos [12] and Exposure [15] used the idea of building passive DNS reputation by statistically modeling various properties of the successfully resolved passive DNS traffic. Plonka et al. [29] introduced Treetop, a scalable way to manage a growing collection of passive DNS data and at the same time correlate zone and network properties. Since then, several researchers were able to use proprietary passive DNS data to build systems that can detect abuse in the Internet [13,14,26,16,31,24]. Clearly, passive DNS is considered to be a very valuable tool that network operators and security researchers use in the fight against Internet abuse. As already discussed, our active DNS project can provide researchers open access to DNS datasets, comparable to the very useful passive DNS, but without any concerns on personally identifiable information (PII) or other legal barriers to repeatable DNS research.

There have been many commercial and nation efforts to create passive DNS repositories. The costs for the commercial offerings [5] often pose a barrier for researchers and network operators. Now, some of the national efforts are hindered by DNS policy, and thus have yet to be widely adapted by the community. Perhaps the most successful has been `passiveDNS.cn`, which was quickly dismissed as an unreliable source of DNS information. The reason behind this development is very simple. The Chinese operators [6] passively collected DNS records that have been already censored by their egress sensors. In our project, we do not censor the views of the recursive DNS servers that Thales uses to resolve the seed domain names on a daily basis.

With the respect of active scanning efforts, most of the efforts have been conducted from the side of the industry. In the last year, however, new work surfaced from the academic community [20] that provides the ability to researchers to scan the entire IPv4 space and use the results for open security research. This is the work that is closest to the proposed system. The key difference, however, is that Censys was not designed to scan the domain name space, rather, IPv4. Thus,

---

[5] For example, `https://www.farsightsecurity.com/`

[6] `http://www1.cnnic.cn/ScientificResearch/LeadingEdge/fymly1/`

while researchers could find some DNS logs into this great public project, our work both complements Censys and also is designed to deal with DNS scanning.

6 Conclusion

DNS is vital to the operation of the Internet. Users, systems, and services rely on its operation for most network communication—often without even realizing it. Malware is no different. It makes use of DNS to locate C&C servers and provide network agility. Despite all its uses, it is incredibly difficult to gain access to large, open, and freely available DNS datasets, and even when possible, such data is often encumbered with privacy regulations or access restrictions. This severely limits the pool of security researchers than can leverage DNS in their work. Furthermore, it limits the repeatability of existing DNS based research. Clearly, there is a need in the research community for access to large, open, and freely available DNS data. To that end, this work built a new system, Thales, to query and collect massive quantities of DNS data starting from publicly available lists of domains (e.g., zone files, Alexa, Common Crawl, etc.). We are releasing the resulting active DNS data from this system to the public, and since this data is derived from public sources, it can be easily incorporated into new or existing research without having to worry about privacy regulations or access restrictions.

To prove its merit, we provide an in-depth comparison between active DNS and a passive DNS dataset collected on a large university network. This analysis showed that active DNS data provides a greater breadth of coverage (i.e., greater quantity and greater variety of records), but passive DNS data provides a denser, more tightly connected graph. Due to these differences, we provided case studies demonstrating how active DNS can be used to facilitate new research or even re-implement existing DNS related research. It is our sincere hope that by opening up active DNS to the security community we can spur more and better research around DNS.

7 Acknowledgment

References

1. I.T. Mate List. `http://vurldissect.co.uk/daily.asp/`, 2016.

2. Abuse.ch domain blacklist. `http://www.abuse.ch/`, 2016.

3. Actionable analytics. `https://www.alexa.com`, 2016.

4. Common Crawl. `https://commoncrawl.org/`, 2016.

5. Domain Graveyard. `http://domaingraveyard.com/`, 2016.

6. Hphosts feed. `http://hosts-file.net/?s=Download`, 2016.

7. LinuxContainers.org. `https://linuxcontainers.org/`, 2016.

8. Malc0de Database. `http://malc0de.com/bl/BOOT`, 2016.

9. Malware Domain List. `https://www.malwaredomainlist.com/`, 2016.

10. Sagadc.org list. `http://dns-bh.sagadc.org/`, 2016.

11. SANS ISC Feeds. `https://isc.sans.edu/feeds/`, 2016.

12. M. Antonakakis, D. Dagon, X. Luo, R. Perdisci, W. Lee, and J. Bellmor. A Centralized Monitoring Infrastructure for Improving DNS Security. In *Recent Advances in Intrusion Detection*, pages 18–37. Springer, 2010.

13. M. Antonakakis, R. Perdisci, W. Lee, N. Vasiloglou, and D. Dagon. Detecting Malware Domains in the Upper DNS Hierarchy. In *Proceedings of the 20th USENIX Conference on Security (USENIX Security)*, August 2011.

14. M. Antonakakis, R. Perdisci, Y. Nadji, N. Vasiloglou, S. Abu-Nimeh, W. Lee, and D. Dagon. From throw-away traffic to bots: Detecting the rise of dga-based malware. In *Proceedings of the 21st USENIX Conference on Security Symposium*, Security'12, pages 24–24, Berkeley, CA, USA, 2012. USENIX Association.

15. L. Bilge, E. Kirda, C. Kruegel, and M. Balduzzi. EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis. In *Proceedings of NDSS*, 2011.

16. Y. Chen, M. Antonakakis, R. Perdisci, Y. Nadji, D. Dagon, and W. Lee. DNS Noise: Measuring the Pervasiveness of Disposable Domains in Modern DNS Traffic. In *Dependable Systems and Networks (DSN), 2014 44th Annual IEEE/IFIP International Conference on*, pages 598–609, June 2014.

17. B. Coat. Snake In The Grass: Python-based Malware Used For Targeted Attacks. `https://www2.bluecoat.com/security-blog/2014-06-10/snake-grass-python-based-malware-used-targeted-attacks`, 2014.

18. M. Cotton and L. Vegoda. Special Use IPv4 Addresses. RFC 5735 (Best Current Practice), Jan. 2010. Obsoleted by RFC 6890, updated by RFC 6598.

19. L. Daigle. WHOIS Protocol Specification. RFC 3912 (Draft Standard), Sept. 2004.

20. Z. Durumeric, D. Adrian, A. Mirian, M. Bailey, and J. A. Halderman. A search engine backed by Internet-wide scanning. In *Proceedings of the 22nd ACM Conference on Computer and Communications Security*, Oct. 2015.

21. M. Felegyhazi, C. Kreibich, and V. Paxson. On the Potential of Proactive Domain Blacklisting. In *Proceedings of the 3rd USENIX Conference on Large-scale Exploits and Emergent Threats: Botnets, Spyware, Worms, and More (LEET)*, April 2010.

22. T. Holz, C. Gorecki, K. Rieck, and F. C. Freiling. Measuring and detecting fast-flux service networks. In *NDSS*, 2008.

23. K. Ishibashi, T. Toyono, H. Hasegawa, and H. Yoshino. Extending black domain name list by using co-occurrence relation between dns queries. *IEICE transactions on communications*, 95(3):794–802, 2012.

24. S. Krishnan and F. Monrose. An empirical study of the performance, security and privacy implications of domain name prefetching. In *Dependable Systems Networks (DSN), 2011 IEEE/IFIP 41st International Conference on*, pages 61–72, June 2011.

25. C. Lever, R. Walls, Y. Nadji, D. Dagon, P. McDaniel, and M. Antonakakis. Domain-Z: 28 Registrations Later Measuring the Exploitation of Residual Trust in Domains. In *37th IEEE International Symposium on Security and privacy*, May 2016.

26. J. Ma, Saul, L. K, S. Savage, and G. M. Voelker. Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, June 2009.

27. Mandiant. APT1. Technical report, 2013. `http://intelreport.mandiant.com/Mandiant_APT1_Report.pdf`.

28. Y. Nadji, M. Antonakakis, R. Perdisci, and W. Lee. Connected colors: unveiling the structure of criminal networks. In *Research in attacks, intrusions, and defenses*, pages 390–410. Springer, 2013.

29. D. Plonka and P. Barford. Context-aware Clustering of DNS Query Traffic. In *Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement*, IMC '08, pages 217–230, New York, NY, USA, 2008. ACM.

30. P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta. Phishnet: predictive blacklisting to detect phishing attacks. In *INFOCOM, 2010 Proceedings IEEE*, pages 1–5. IEEE, 2010.

31. B. Rahbarinia, R. Perdisci, and M. Antonakakis. Segugio: Efficient Behavior-Based Tracking of Malware-Control Domains in Large ISP Networks. In *Dependable Systems and Networks (DSN), 2015 45th Annual IEEE/IFIP International Conference on*, pages 403–414, June 2015.

32. Y. Rekhter, B. Moskowitz, D. Karrenberg, G. J. de Groot, and E. Lear. Address Allocation for Private Internets. RFC 1918 (Best Current Practice), Feb. 1996. Updated by RFC 6761.

33. M. L. . C. C. Security. CopyKittens Attack Group. `https://eforensicsmag.com/copykittens/`, 2015.

34. J. Weil, V. Kuarsingh, C. Donley, C. Liljenstolpe, and M. Azinger. IANA-Reserved IPv4 Prefix for Shared Address Space. RFC 6598 (Best Current Practice), Apr. 2012.

35. F. Weimer. Passive DNS Replication. In *In Proceedings of the 17th FIRST Conference on Computer Security Incident Handling*, June 2005.

36. B. Zdrnja, N. Brownlee, and D. Wessels. Passive monitoring of DNS anomalies. In *Proceedings of DIMVA Conference*, 2007.