Reexamining DNS from a Global Recursive Resolver Perspective

Hongyu Gao, Vinod Yegneswaran, Jian Jiang, Yan Chen, *Member, IEEE,* Phillip Porras, Shalini Ghosh, Haixin Duan

***Abstract*—The performance and operational characteristics of the DNS protocol are of deep interest to the research and network operations community. In this paper, we present measurement results from a unique dataset containing more than 26 billion DNS query-response pairs collected from more than 600 globally distributed recursive DNS resolvers. We use this dataset to reaffirm findings in published work and notice some significant differences that could be attributed both to the evolving nature of DNS traffic and to our differing perspective. For example, we find that although characteristics of DNS traffic vary greatly across networks, the resolvers within an organization tend to exhibit similar behavior. We further find that more than 50% of DNS queries issued to root servers do not return successful answers, and that the primary cause of lookup failures at root servers is malformed queries with invalid TLDs. Finally, we observe that the number of DNSSEC-enabled domains has increased sharply and that over 24% of second-level domains have IPv6 authorative servers.**

***Index Terms*—DNS, Measurement, Malicious Domain Detection**

I. Introduction

THE Domain Name System (DNS) protocol plays a cardinal role in the operation of the Internet by enabling the bi-directional association of domain names with IP addresses. It is implemented as a hierarchical system with a few trusted root servers that distribute the responsibility of updating the name-to-IP-address mapping to hundreds of millions of authoritative nameservers that correspond to each domain. DNS as a protocol has steadily evolved since its initial specification [22]–[25] as has the mix of applications that find new and innovative ways of using it. Most applications today and future Internet architectures (such as Named Data Networks and Software-Defined Networks) depend on DNS. It is also increasingly abused by malware authors, both as an effective redirection mechanism for obfuscating location of their servers [15] and as a covert channel for command and control [13, 26].

Hongyu Gao is with Google, 265 N Rengstorff Ave #12, Mountain View, CA 94043 (e-mail: hygao@u.northwestern.edu).

Vinod Yegneswaran, Phillip Porras and Shalini Ghosh are with SRI International, Computer Science Laboratory, 333 Ravenswood Ave, Menlo Park, CA, USA 94025 (e-mail: vinod@csl.sri.com; porras@csl.sri.com; shalini@csl.sri.com).

Jian Jiang is with the Institute for Network Sciences and Cyberspace, Tsinghua University, China and the School of Information, UC Berkeley (e-mail: jiangjian@berkeley.edu).

Yan Chen is with the Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL 60208 USA, and the State key Laboratory of Networking and Switching Technology (Beijing University of Posts and Telecommunications), China. Yan Chen is the corresponding author of this paper (e-mail: ychen@northwestern.edu).

Haixin Duan is with the Institute for Network Sciences and Cyberspace, Tsinghua University, China (e-mail: duanhx@tsinghua.edu.cn).

Given its crucial importance for the Internet's functioning, DNS has been the subject of many measurement studies during the last decade. Prior measurement studies have scrutinized the behavior of DNS caches [18], characterized global DNS activity from the perspective of root servers [10, 11] and evaluated the effectiveness of DNS in the context of content-delivery networks [30]. The first study of global DNS activity was by Danzig et al., which uncovered the prevalence of many bugs in popular DNS implementations [12]. More recently, this problem was revisited by Brownlee et al., who measured the prevalence of bogus DNS traffic at the F-root nameserver finding that some of the same problems persist: 60-85% of observed queries were repeated queries from the same host and more than 14% of requests involved queries that violated the DNS specification. Jung et al., measured that a significant portion of DNS lookups (more than 23%) receive no answer and that they account for more than half of all DNS packets in the wide-area due to persistent retransmissions.

Several of these studies were conducted more than a decade ago and often from a small number of vantage points. Collaboration between the Internet research and operations community has evolved significantly since these foundational studies and we now have access to a new and unique data source, the Internet Systems Consortium (ISC)'s Secure Information Exchange (SIE) [14], which enables researchers to monitor DNS activity from hundreds of operational networks in real-time. One of the driving forces behind such data sharing has been its untapped potential for rapidly identifying malware domains. In particular, domain registrations and DNS access patterns could be an effective means for tracking cyber-criminal behavior and several recent studies have explored the application of machine-learning techniques to automatically identify malicious domains [3, 9, 31].

In this paper we report on findings from a global and multidimensional analysis of DNS activity, as observed from a large set of widely distributed and operational DNS resolvers. Specifically, we analyze two weeks of data from more than 600 resolvers comprising more than 26 billion queries and responses. First, we systematically dissect this data, present high-level characteristics of observed traffic behavior and identify invariant characteristics across resolvers. Second, we use this dataset to critically reexamine the validity of certain prior measurement studies, in the context of this more global perspective and modern traffic characteristics. domain groups. We make the following key findings:

- We find that resolvers from different /24 subnets have different profiles, including query/response counts; unanswered

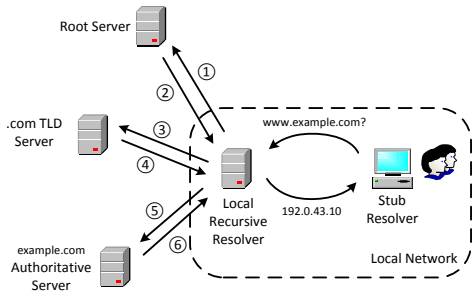


Fig. 1. An illustration of the DNS resolution process for www.example.com

Fig. 2. The geo-location of the DNS resolvers that contribute to the data.

query rates; unsolicited response rates; query type distributions; and query success-to-failure ratios.

- In comparison with prior measurement results, A queries continue to dominate, AAAA queries have sharply increased and other query types depict a decrease in popularity.
- We find that although root servers are always available (*i.e.* , have no unanswered queries), more than 15.1% of the queries sent by recursive DNS resolvers are unanswered.
- We explored the cause of DNS query with negative answer (queries that do not return "NOERROR"). We identify DNSBL as having a much higher failure ratio than do other query types.
- We find that invalid TLD (Top Level Domain) is the primary cause of query with negative answer at root servers, and that the percentage of invalid TLD has increased in comparison with the results from prior measurements. However, A-*for*-A queries have decreased in popularity, and almost disappeared in our data.
- We find that 12.0% of traffic to root severs and 8.0% to other servers are *truly* repeated queries. We further identify the possible causes including concurrent query, CNAME chain sanitization, premature retransmission, and implementation quirks.
- We find that the number of DNSSEC-enabled domains has increased sharply compared to prior reports.
- We find that 24.1% of SLDs (Second Level Domains) have IPv6 authoritative servers.

II. Background and Dataset

DNS Protocol. The Domain Name System (DNS) is a distributed, hierarchical naming system that translates between domain names and IP addresses. Client end hosts (also called stub resolvers) simply contact a recursive resolver that implements the hierarchical resolution process of iterating through nameservers to perform the translation. In the example shown in Figure 1, the stub resolver queries the local recursive resolver for the IP address of `www.example.com`. The recursive resolver usually resides within the local network of the client's organization and is managed by the organization's administrator. However, clients can also choose to contact recursive resolvers located outside their local network (e.g., OpenDNS resolvers and Google public DNS resolvers). Assuming an empty cache, the recursive resolver starts by querying the root server for the IP address of www.example.com. The root server responds with a referral to the `.com` TLD server. The recursive resolver then queries the `.com` TLD server, and in response is provided with a referral to the authoritative server for `example.com`, which hosts the name-to-address mapping. Finally, the recursive resolver contacts the authoritative server of `example.com` to obtain the corresponding IP address.

Data. Our data is collected from a high-volume passive DNS source at the Security Information Exchange (SIE) [14]. This provides a near real-time data feed from multiple hundreds of DNS recursive resolvers distributed over the Internet. These resolvers represent large ISPs, universities, as well as public DNS service providers located in North America and Europe, suggesting a wide diversity in the user population behind these resolvers. We plot the geo-locations of the DNS resolvers in Figure 2 using the MaxMind geolocation database [21].

Due to privacy concerns, the data-collection sensor is deployed "above" the recursive resolvers and records all DNS queries and responses between the recursive resolvers and the remote DNS servers. The sensor does not collect traffic between client stub resolvers and recursive resolvers. As a result, the identity of client endhosts that sit behind the recursive resolvers are not available.

Previous SIE data analysis has shown that 93% of the domain labels immediately under the .edu TLD have a resource record in the SIE data in a two-week observation period [32]. The DNS servers that generate responses are dispersed in 70.7% of the /8 CIDR blocks and 69.2% routable ASes [32]. We collected all DNS traffic in the raw SIE channel for two weeks from December 9, 2012 to December 22, 2012. In total, our dataset contains about 26 billion DNS queries and responses.

Note that our dataset, although the most diverse to-date, still has a geographic bias, because the monitored resolvers are exclusively located in U.S. and Europe (Figure 2). Hence, we focus our study on macroscopic characteristics and temporal behaviors, which we believe do not have strong correlations with geographic location.

Local and Root Perspective. Since our data is collected from local recursive DNS resolvers, it naturally enables studying DNS behavior from the perspective of the local resolvers. On the other hand, 13 root servers of vital importance sit atop the DNS hierarchy. Due to their importance, multiple prior works have analyzed DNS protocol behavior from the perspective of the root servers [10, 11, 34].

We attempt to analyze our DNS data from the root perspective as well. As described in Section II, if a client-side nameserver restarts with empty cache, or the TTL expires for a TLD namesever entry, the recursive resolution process starts

by querying the root servers and obtaining a referral to an authoritative TLD nameserver. Although our data is collected from local recursive DNS resolvers, the availability of the response nameserver's IP address enables us to isolate the DNS traffic to and from root servers. Given the volume and diversity of our dataset, we believe that the subset of DNS queries and responses is a representative sample of DNS traffic that root servers experience. In this paper, we analyze the DNS traffic characteristics from both the local perspective (*i.e.* , using the full dataset) and the root perspective (*i.e.* , using only traffic to and from root servers), whenever applicable.

III. DNS Traffic Characteristics

In this section, we analyze the characteristics of the collected DNS traffic from various perspectives.

A. High-level Characteristics

Our data includes traffic from 628 distinct DNS resolvers including 10 IPv6 resolvers. Not surprisingly, we find significant variance in the volume of DNS queries that they generate. The most active resolver generates more than 70M queries per day, which translates to an average of more than 800 queries per second. In contrast, 407 resolvers generate fewer than 10,000 queries during the two week measurement period.

This observed range shows that the query volume of DNS resolvers has a heavily skewed distribution. A small fraction of deployed DNS resolvers are serving the majority of the DNS queries. This observation is consistent with that of prior measurement studies by Pang et al. [30] in 2004 and Osterweil et al. [27] in 2012. Interestingly, the vast majority of inactive resolvers belong to a European educational institution (354 resolvers) and a US educational institution (49 resolvers). We subsequently learned that DNS experiments are conducted at these institutions, and speculate that ongoing DNS experiments may be the reason behind the large number of inactive DNS resolvers. Nonetheless, the amount of traffic generated by the inactive resolvers is negligible and should not remarkably affect our measurement results.

We further agglomerate IP addresses into /24, /16 and /8 subnets, respectively. We also put all resolvers with IPv6 addresses into one group. Our monitored DNS resolvers span 71 distinct /24 subnets, 33 distinct /16 subnets, and 22 distinct /8 subnets. This further validates that our data is collected from vantage points distributed widely across the IPv4 address space.

1) Organizations: We use /24 subnets to group DNS resolvers into organizations and bin all resolvers with IPv6 addresses into a special group. Although large organizations may have /16 or /8 subnets, we find /24 subnets to be a good way to group DNS resolvers as it provides sufficient abstraction and enables capturing the difference between different subnets within large organizations.

We identify the 20 top /24 subnets in our data with the highest traffic volume. By using `whois` lookups to determine the organization of the /24 subnets, we identified six commercial US ISPs; one US educational institute; two commercial European ISPs; one European educational institute; and a public DNS service provider. Many organizations deploy DNS resolvers in multiple /24 subnets as shown in Table I. Due to privacy concerns, we use the location (US or EU) and type (commercial, EDU or public) to denote the organizations. The bulk of the data is collected from US ISP A, which serves a large population and contributes a large number of resolvers.

Organization	Resolver #	Traffic %
US ISP A (subnet 1)	40	32.6%
US ISP A (subnet 2)	34	22.7%
US ISP A (subnet 3)	10	17.4%
Public DNS Service	4	11.7%
US ISP B	2	2.0%
US ISP C (subnet 1)	2	1.6%
US ISP C (subnet 2)	2	1.5%
US ISP D (subnet 1)	8	1.1%
US ISP D (subnet 2)	8	1.0%
US ISP E (subnet 1)	2	1.0%
EU ISP A	8	1.0%
US ISP C (subnet 3)	2	1.0%
US ISP F	4	0.8%
US ISP D (subnet 3)	8	0.7%
EU EDU (subnet 1)	11	0.6%
US ISP C (subnet 4)	1	0.5%
US EDU	50	0.4%
US ISP E (subnet 2)	1	0.4%
EU EDU (subnet 2)	2	0.2%
EU ISP B	2	0.2%
IPv6	10	0.6%

TABLE I
THE PERCENTAGE OF TRAFFIC GENERATED FROM THE TOP 20 /24 SUBNETS WITH IPV4 RESOLVERS, AND THE AGGREGATE TRAFFIC GENERATED BY IPV6 RESOLVERS.

2) DNS Data Type: In normal operation, each DNS query is associated with a response. However, cases exist when a DNS query is not answered or a DNS response is received without a matching query, either due to misconfiguration, backscatter from attack traffic or packet loss. Hence, we group DNS traffic in our data into three categories: query-response pairs, unanswered queries and unsolicited responses. More than 83.3% of the entries in our data are query-response pairs, 14.9% are unanswered queries and 1.8% are unsolicited responses. The percentage of abnormal cases, including both unanswered queries and unsolicited responses, is 16.7%, which seems anomalous and is worthy of deeper investigation. An obvious consideration is packet loss in the data collection infrastructure. We find that three subnets deviate significantly from others with drastically lower percentage of query-response pairs and higher percentage of unanswered queries. They belong to two organizations– the public DNS service and the European educational institute. In addition, the public DNS service is the only organization that suffers from a high percentage of unsolicited answers (15.2%).

Finally, we recompute the numbers for the percentage of query-response pairs, unanswered queries, and unsolicited responses after excluding the two anomalous organizations. We find these numbers to be 88.6%, 11.3%, and 0.03% respectively. The low percentage of unsolicited responses also indicates that packet loss may not be a detrimental issue in the SIE data collection infrastructure outside of these two providers.

3) Server-side Traffic Distribution: Besides the traffic distribution across monitored resolvers, we are also interested in where the traffic goes. By counting the destination IPs of DNS queries, we identify nearly 1.38 million distinct DNS authoritative servers, including 17,874 (1.3%) IPv6 servers.

Organization	Traffic %
Akamai	11.6%
.com/.net TLD	6.1%
Amazon	4.7%
Google	4.4%
cox.net	3.2%
Apple	2.3%
Mcafee	2.1%
dynect.net	2.0%
Root	1.8%
iana.org	1.5%
others	60.3%

TABLE II
SERVER-SIDE DISTRIBUTION OF DNS TRAFFIC, AGGREGATED BY ORGANIZATIONS.

Perspective	Year	A	AAAA	PTR	MX
Local	2012	66.2%	13.4%	11.1%	2.3%
Local	2001	60.4% - 61.5%	N/A	24% - 31%	2.7% - 6.8%
Root	2012	57.5%	26.6%	4.8%	0.2%
Root	2008	60%	15%	8.4%	3.5
Root	2002	55.5%	4.7%	19.9%	4.6%

TABLE III
DISTRIBUTION OF DNS LOOKUPS BY POPULAR QUERY TYPES. THE TABLE OMITS THE PERCENTAGE OF OTHER QUERY TYPES. THE PERCENTAGES FOR YEARS 2001, 2002 AND 2008 ARE FROM [18], [34] AND [11], RESPECTIVELY.

The authoritative servers span 30,129 distinct ASes, 229 distinct /8 subnets, 23,614 distinct /16 subnets and 378,298 distinct /24 subnets.

The distribution of DNS traffic across the 1.38 million DNS servers follows a heavy-tailed distribution. While the 500 busiest servers attracted nearly half (49.9%) of all DNS queries; 94.1% of DNS servers received less than 10,000 queries during the two week measurement period.

We further group the DNS servers into organizations, for which we first correlate the addresses with DNS records in our dataset to identify their domain names, then empirically label the domain names with organizations (*e.g.* , we label addresses with `*.apple.com` and `*.mac.com` domain names as *apple*). Table II presents the 10 organizations received highest DNS traffic volume in our dataset, which together absorbed 39.7% of all DNS queries.

B. Query Type Breakdown

The DNS protocol supports a variety of query types for different purposes. To summarize the most popular types, an A query translates a domain name into IPv4 addresses, a AAAA query translates a domain name into IPv6 addresses, an MX query translates a domain name into mail exchange hostnames, and a PTR query translates an IP address back to domain names. We examine how popular each query type is in the real world, and measure how this distribution has changed over time. Table III shows the distribution of the four popular types of DNS queries in real-world traffic. Because we do not have access to legacy DNS traffic, we quote the numbers reported by Jung et al. [18], Wessels et al. [34] and Castro et al. [11] in the row of year 2001, 2002 and 2008, respectively. Jung et al. collected their data from local resolvers at MIT and KAIST. On the other hand, Wessels et al. and Castro et al. reported the distribution observed from only root servers.

From the Perspective of Local DNS Resolvers. After more than ten years, the A query remains the most dominant DNS query type in US and Europe, accounting for about

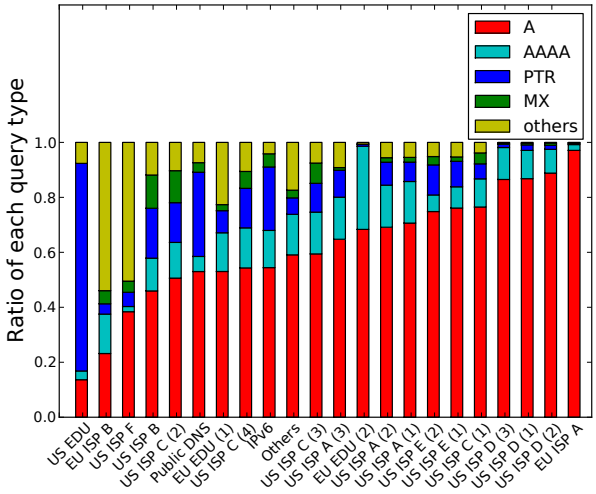


Fig. 3. The query type breakdown in different organizations from local perspective.

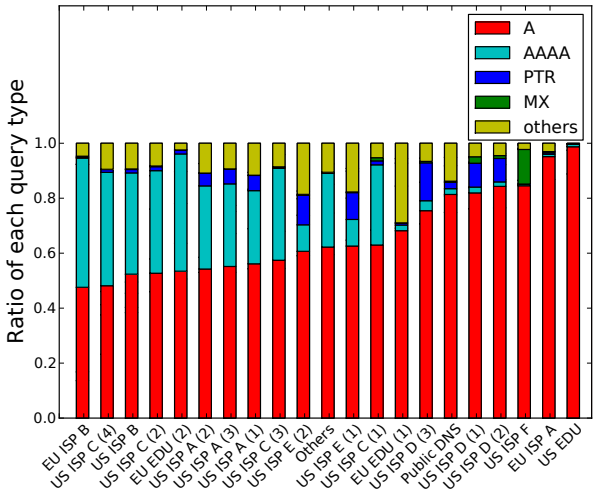


Fig. 4. The query type breakdown in different organizations from root perspective.

66.2% of total queries. This percentage remains stable with a slight increase after 10 years. With wider deployment of IPv6 protocol, the volume of AAAA queries (13.4%) has risen sharply. This query type did not exist 10 years ago. Meanwhile, the percentage of PTR queries has decreased from 24-31% to 11.1% and MX queries have decreased from 2.7-6.8% to 2.3%. While the absolute number of queries has also grown significantly in the past 10 years the growth of other query types is not comparable to that of AAAA queries.

From the Perspective of Root DNS Servers. We observe a similar trend with local perspective. The percentage of A query remains steadily high at root servers. The percentage of AAAA query has increased with time, while the percentage of PTR and MX query has decreased. However, the change is more drastic from the root perspective than from the local perspective. At root, the percentage of AAAA queries has increased by 466% from 2002 to 2012. In contrast, the percentages of PTR query and MX queries have shrunk by 76% and 94% respectively in the same time period.

Query Types in Different Organizations. Figures 3 and 4 plot the distribution of query types across different organizations. Figure 3 analyzes all the traffic at local resolvers,

Qtype	A	PTR	DNSBL	AAAA
Root perspective	66.0%	9.1%	0.2%	5.8%
Local perspective	50.9%	28.2%	7.2%	4.5%

TABLE IV
FOUR QUERY TYPES CAUSING THE LARGEST NUMBER OF NEGATIVE ANSWERS FROM THE ROOT AND LOCAL PERSPECTIVE, RESPECTIVELY.

Perspective	Invalid TLD	A-*for*-A	Private IP	Non-printable char
Root	53.5%	0.4%	0.1%	3.2%
Local	1.2%	<0.1%	0.8%	0.2%

TABLE V
THE PERCENTAGE OF *invalid TLD*, A-*for*-A, PRIVATE IP IN PTR QUERY AND NON-PRINTABLE CHARACTERS FROM THE ROOT AND LOCAL PERSPECTIVE, RESPECTIVELY.

whereas Figure 4 only analyzes queries at root servers. We observe that different organizations have different characteristics. In addition, the organization profiles from local perspective is highly diverse. The organization profiles from root perspective are more consistent. We find several organizations that exhibit drastically different patterns in comparison to others:

1) The US EDU subnet issues 75.5% of all PTR queries. However, the same US EDU subnet almost exclusively issues A queries (98.7%) to root servers.
2) The EU ISP A subnet issues almost exclusively A queries (97.1%). The same EU ISP A subnet issues almost only A queries (95.1%)to root servers as well.

C. DNS Query Success Rates

Next, we study the question of how many modern DNS queries return successful answers. We reuse the categorization method adopted by Jung et al. in [18]. In particular, DNS queries with successful answers are those having "NOERROR" as the return code in the response. We further divide the remaining queries into two categories: queries without response, and queries returning negative answers. Our definition of negative answer broadly includes all responses whose return code is *not* "NOERROR".

From the local perspective, the aggregated ratios of DNS queries with successful answers, negative answers and no answers are 66.9%, 18.0% and 15.1%, respectively. The overall ratios are similar to the result from ten years ago, when Jung et al. reported that the percentages of answers with successful, negative answer and unanswered queries were 64.3%, 11.1% and 23.5% respectively in their MIT trace [18]. This suggests that many of the contributors to DNS queries with negative answers and no answers, persist from a decade ago. From the root perspective, the ratio of unanswered query is 0, meaning that every query issued to the root servers is answered. It implies that root servers were always available during the measurement period. However, the percentage of successful answers returned by root servers, (*i.e.* , referrals to nameservers that should know the queried hostnames), is significantly lower than that of other servers. 54.0% of the queries issued to root servers return negative answers. In comparison, in 2000 only about 2% of lookups to root return negative answers [18]. The sharply increased percentage of query with negative response at root servers may result from the high ratio of invalid traffic reaching them, as reported by multiple previous measurement studies at root servers [10, 11]. We further investigate the cause of failed query in §III-D.

D. Causes of Queries with Negative Answers

We first identify which query types cause the most negative answers. Table IV shows the top four types with their respective percentages. We find that A queries cause the vast majority of negative answers, in viewing from both the root and the local perspective. In comparison, in 2000 the dominant query type resulting in negative answers was the PTR query type [18]. Due to the shrinking percentage of PTR queries in our traffic, A queries have now become the dominant contributor to negative query responses. At the root servers, negative answers caused by PTR queries and DNSBL are much less common when compared with the local perspective.

Different query types also have differing ratios of negative answers. The ratio of DNSBL query with negative answers to the total number of DNSBL queries is 73.9%, which is significantly higher than any other query types due to the nature of blacklist lookup: most of lookups do not hit the blacklist, in which case an 'NXDomain' response is returned. We further analyze DNSBL in §III-D5. Among the other three types, the ratio of PTR query with negative answers to the total number of PTR queries is 46.5%, which is higher than corresponding ratios for the A (14.8%) and AAAA (6.5%) query types.

Independent from query types, prior research has identified problematic query names that evoke negative answers [10, 11, 18, 34], including invalid TLDs, A-for-A, non-printable characters and private IP address in 'PTR' query. We investigate them in detail in §III-D1, §III-D2, §III-D3 and §III-D4, and present their respective percentage in Table V. Note that the columns are *not* mutually exclusive.

1) Invalid TLD: Invalid TLD denotes the case when the queried hostname does not have a valid TLD. This may be caused by either user typos or client-side application implementation bugs. Because the queried names do not exist, such queries will result in NXDomain as the response. Table V presents that 1.2% of the traffic from the local perspective contains an invalid TLD. However, 53.5% of the queries seen by root servers contain invalid TLDs. This observation, although highly skewed, seems reasonable, because queries with invalid TLDs terminate the recursive resolution process at root servers, in the absence of valid TLD servers. Recall that from the root perspective, the total percentage of queries with negative answers is 54.0% (§III-C). It means that *invalid TLD* has become the primary contributor to negative answers at root servers.

Multiple prior studies have investigated the prevalence of *invalid TLD* domains at root servers [10, 11, 34]. The percentage of *invalid TLD* domains reported in 2001, 2003 and 2008 were 20%, 19.53% and 22.0%, respectively, which is stable. Surprisingly, it has sharply increased to 53.5% in 2012, from the perspective of our dataset. In addition, the resolvers issuing *invalid TLD* queries were wide spread in all the major organizations that we monitor. Note that the above comparison

TLD	Count (M)	%
local	68.4	21.9%
no_dot	52.2	16.7%
belkin	51.2	16.4%
corp	9.6	3.1%
lan	2.9	0.96%
home	2.3	0.74%
localdomain	1.7	0.54%
loc	1.5	0.48%
internal	1.4	0.45%
pvt	1.2	0.39%

TABLE VI
LIST OF 10 MOST FREQUENTLY QUERIED INVALID TLDS WITH COUNTS IN MILLIONS AND PERCENTAGE.

only applies to root servers. The percentage of *invalid TLD* is low from the local perspective.

We summarize the most common invalid TLDs in Table VI. For each TLD, the table shows its count in million as well as its percentage among all invalid TLDs. We observe that a large number of invalid domains do not contain any dot. We put such domains in a special "no_dot" group, which is the second most popular form of invalid TLDs. Together with "local" and "belkin" these three invalid TLDs are far more popular than the other ones. ".local" is a pseudo-TLD that a computer running Mac OS X uses to identify itself if it is not assigned a domain name. Similarly, queries with ".lan," ".home," ".localdomain," ".loc," and ".internal" are likely used by other programs under certain circumstances. Nevertheless, these queries are meant to stay local, and should not leak out to the Internet. "Belkin" is a famous brand that manufactures electronic device. We suspect that queries with ".belkin" are generated by the device under the same brand due to misconfiguration. These are likely good candidates to be suppressed by local implementations. Although we have identified several likely causes of frequently appearing invalid TLDs, user typos can also result in invalid TLDs. In our data, the count of invalid TLDs exhibit a long-tailed distribution. More than 500,000 other invalid TLDs are used much less frequently.

2) A-for-A Query: A-*for*-A query denotes the case that the queried "hostname" is already an IP address. Because an IP address is also represented as a dot-separated string, the IP address A.B.C.D will be interpreted as having the TLD "D". Thus, A-*for*-A queries are a special case of *invalid TLD* queries.

In comparison with multiple prior works [10, 11, 34], we observe an interesting trend. The percentage of A-*for*-A seen by root servers reported in 2001, 2003 and 2008 was 12-18%, 7.03% and 2.7%, respectively. The decreasing trend continues in our data collected in 2012, where A-*for*-A only contributes 0.4% of the traffic. It indicates that most buggy implementations that caused this problem have been fixed. From the local perspective, the percentage of A-*for*-A is also negligible (<0.1%).

3) Private IP Address in PTR: RFC 1918 defines several networks that can be used internally but cannot be routed on the Internet: 10.0.0.0/8, 172.16.0.0/12 and 192.168.0.0/16. In theory, PTR queries to these IP addresses should be handled by the DNS administrators locally without leaking to the global Internet.

Qtype	A	PTR	SOA	AAAA
Ratio across types	42.3%	17.8%	14.5%	14.0%
Ratio within type	9.5%	23.9%	87.6%	15.5%

TABLE VII
THE FOUR TYPES WITH LARGEST NUMBER OF UNANSWERED QUERIES. THE FIRST ROW SHOWS THE PERCENTAGE IN UNANSWERED QUERIES. THE SECOND ROW SHOWS THE PERCENTAGE OF UNANSWERED QUERY WITHIN THE CORRESPONDING TYPE.

However, this is not the case in the real-world deployment. Previous studies revealed that at root servers, 7% and 1.61% of the queries are PTR queries with private IP addresses in a 2001 trace [10] and a 2002 trace [34], respectively. In our 2012 trace, only 0.1% of the queries issued to root servers have private IP addresses. This decreasing trend shows that more DNS administrators handle PTR queries with private IP addresses properly now. Viewing from the local perspective, 0.8% of the queries are PTR queries with private IP address.

4) Non-printable Characters in Query Name: According to RFC 1035, domain names should only contain alphanumeric characters and hyphens, separated by dots. We follow the name in [11, 34] and refer to characters outside this scope as non-printable characters. From the root perspective, 3.2% of the query names contain non-printable characters. In comparison, this number in 2002 [34] and 2008 [11] is 1.94% and 0.1%, respectively.

Viewing from the local perspective, 0.2% of the queries contain non-printable characters and 0.9% of the failed queries contain non-printable characters.

5) DNS Blacklists: DNS blacklist (DNSBL) is a popular method used by site administrators to vet domains for spam, malware etc. Although DNSBL utilizes the DNS protocol, it does not translate between hostnames and IP addresses. Rather, site administrators use it to determine whether the target hostname is blacklisted, by crafting the target hostname into a special URL under the blacklist provider's domain and issuing an A query. When the query reaches the blacklist provider's authoritative nameserver, the nameserver will send a response according to its own format. In popular DNSBL designs, the return code will be NXDomain (domain not exist) if the target hostname does not hit the blacklist. In particular, 73.9% of DNSBL queries return NXDomains, which gives DNSBL queries significantly higher failure odds than other query types.

The usage of DNS blacklists has been reported in [17]. DNS blacklists lookups accounted for 0.4% and 14% of lookups in their December 2000 trace and Feburary 2004 trace, respectively. In 2012, DNSBL queries account for 1.7% of the lookups. The percentage is lower than year 2004, but higher than year 2000.

E. Unanswered Queries

As described in § III-C, root servers do not incur any unanswered queries from our networks. Every query issued to the root servers is answered. All the unanswered queries are caused by other servers. In addition, the ratio of unanswered queries differs drastically across different organizations.

We further measure the correlation between unanswered query with different query types, and show the results in

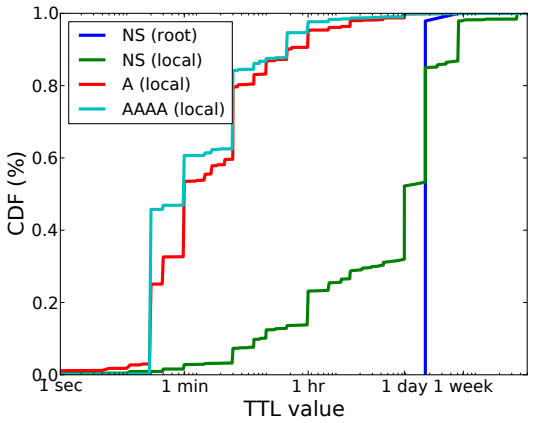


Fig. 5. The cumulative distribution of TTLs of NS record returned by root servers, and three record types, A, AAAA and NS, returned by other servers.

Table VII. Due to the dominance of the A query, it also represents the largest portion of unanswered queries (42.3%). However, the probability of an A query without an answer is the lowest among the four listed types (9.5%). Interestingly, 87.6% of SOA queries do not get an answer.

F. TTL Distribution

The time-to-live (TTL) field in the DNS responses informs the resolver how long it should cache the results. Figure 5 shows the cumulative distribution of TTL values of three distinct record types in our DNS data: A, AAAA and NS. Root servers very rarely reply with A or AAAA records in answer section, so we only plot NS record TTLs returned by root servers. In particular, A and AAAA record provides a direct mapping from a hostname to an IPv4 address and an IPv6 address, respectively and the NS record provides a reference to the authoritative nameserver that should know the queried hostname when the nameserver being queried does not know the IP address of the queried hostname. We observe that NS records have much larger TTL values than A and AAAA records. This result is consistent with the result reported by Jung et al. from ten years ago [18], except that AAAA record did not exist back then. Given that AAAA and A records play a similar role, which is to translate domain names to IP addresses, it is reasonable to observe that AAAA records and A records shares similar TTL distributions. On the other hand, the longer TTL value of NS records is the key reason that keeps the load of DNS servers residing higher in the hierarchy manageable. Only 1.8% of the queries are issued to root servers in our trace, because in most cases the client-side nameserver knows the authoritative nameserver using the cached NS records. If NS records have a much shorter TTL, the client-side nameserver will need to query the root servers much more frequently. We also observe that the TTL of NS records returned by root servers is extremely regular: almost all records have TTL of two days.

We further compare the TTL of A and NS records in 2012 and that in 2000 as reported in [18]. The TTL of NS records roughly remains stable. However, the TTL of A records in 2012 is much smaller. In 2000, only about 20% of A records have TTL less than one hour. About 20% of A records have TTL larger than one day. In 2012, about 90% of A records have TTL less than one hour and almost 0% of A records have TTL larger than one day. This difference shows the wide deployment of CDN and other services that leverage short TTLs, which inevitably poses more pressure on the DNS infrastructure.

Hosting Domain	# of Hosted Domains	%
domaincontrol.com	4,481,303	12.7%
worldnic.com	762,973	2.2%
hostgator.com	718,416	2.0%
bluehost.com	545,446	1.6%
rzone.de	471,652	1.3%
1and1.com	456,083	1.3%
ovh.net	450,216	1.3%
yahoo.com	449,760	1.3%
websitewelcome.com	415,534	1.2%
dreamhost.com	411,659	1.2%

TABLE VIII
LIST OF 10 MOST POPULAR DNS HOSTING DOMAINS.

G. DNS Hosting

In DNS, a domain has one or more NS records indicating its delegation names, *i.e.* , the names of its authoritative servers. A common practice for domain owners is to outsource name resolution to DNS hosting service providers such as Godaddy by pointing the NS records to names under corresponding hosting domains (*e.g.* , Godaddy uses `domaincontrol.com` as its hosting domain).

In our dataset, we observe nearly 35.3 million distinct SLDs. Only a small fraction (2.1%) of these domains have in-domain NS names. The rest either host on different yet self-operated hosting domains, or most likely outsource name resolution to DNS hosting service providers. Most domains have NS names under 1 (85.0%) or 2 (10.2%) hosting domains. The market share of hosting domains also follows a heavy-tailed distribution; the top 100 hosting domains serve name resolution for 50.0% of the domains. Table VIII lists 10 of the most popular hosting domains observed in our dataset.

H. DNSSEC Deployment

The original design of DNS did not include data integrity protection. Attackers can fool end-users by forging DNS responses in a number of ways [7, 19]. The Domain Name System Security Extensions (DNSSEC) attempts to integrate Public Key Infrastructure (PKI) into DNS by assigning hierarchically delegated public keys to DNS domains and using digital signatures to provide data integrity and origin authentication. DNSSEC is backwards compatible with the existing DNS. It is implemented by adding more DNS data types such as DNSKEY, DS, RRSIG, NSEC and NSEC3, and a few flag bits in DNS packets [4]–[6]. The Internet coordination organization ICANN is now promoting the DNSSEC deployment. As of the last day of our dataset, 107 out of 317 TLDs have deployed DNSSEC [16].

In our dataset, we observe 151,179 domains with DNSSEC-related records, including 107 TLDs (all signed TLDs at that time), 135,924 SLDs, and 15,148 lower level domains. In comparing with measurements by Osterweil et al. from 2005 to 2008 [28], who observed 871 signed "production" domains, the number of signed domains has increased by nearly 170

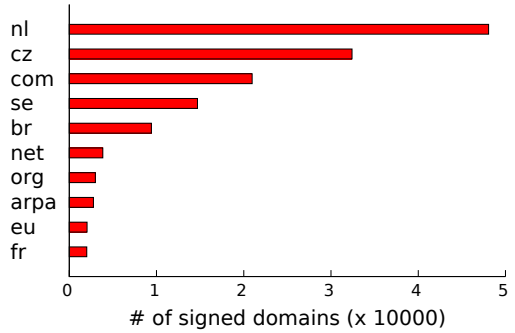


Fig. 6. The number of signed domains breakdown in TLDs.

times.[1] The enormous increase demonstrates great progress of DNSSEC deployment in recent years.

Figure 6 breaks down the number of signed domains in TLDs, which again shows a highly skewed distribution. The top 10 TLDs in Figure 6 contribute 92.1% of all signed domains in our dataset. Some country code TLDs (notably `.nl`, `.cz`, `.se` and `.br`) contribute significantly more signed domains than others, demonstrating their extraordinary efforts in deploying DNSSEC.

I. IPv6 Deployment

Besides DNSSEC, IPv6 is another major change currently happening in DNS. We have seen in §III-A that the volume of IPv6 DNS traffic and the number of IPv6 resolvers are quite small. However, the deployment of IPv6 could still be considered progressful from the server-side perspective. From our dataset, we observe near a quarter of (24.1%, 8.5 million out of 35.3 million) of SLDs having IPv6 authoritative servers, an astonishing percentage comparing to the numbers reported in §III-A. This is mainly because popular DNS hosting providers are positive with adopting IPv6. Twenty-eight out of the top 100 hosting domains in our dataset have IPv6 authoritative servers, which make their customers resolvable from IPv6. The high percentage of IPv6 enabled domains suggests that ISPs and other DNS resolution service providers such as GoogleDNS and OpenDNS should support IPv6 in the iterative process of DNS resolution.

J. Repeated DNS Queries

Multiple previous studies of root DNS servers have revealed that over 56%-85% of queries observed at root servers are repeated [11, 34]. These studies further identified that misconfigured or abusive clients mainly caused these astonishingly high numbers. Ideally, a "normal" resolver should not issue many, if any, repeated queries to authoritative servers because of the effect of caching. However, our dataset shows that this is not the case—a considerable portion of DNS queries from "normal" resolvers could still be considered repeated. In this section, we analyze the prevalence and explore potential reasons behind the repeated query behavior of resolvers in more detail.

[1]Osterweil et al. empirically excluded signed domains that were likely to be deployed only for testing. We do not apply such division as only few domains are seemingly testing domains from our empirical inspection.

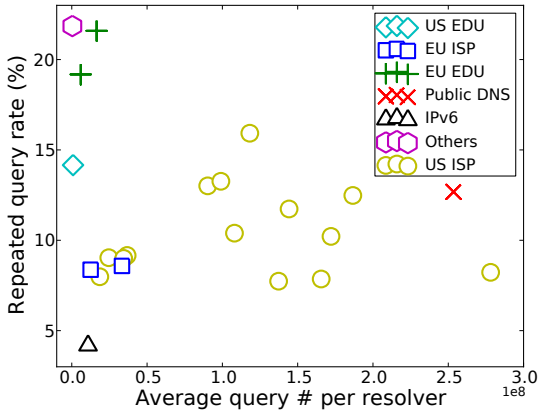


Fig. 7. The repeated query ratio and average number of queries per resolver for major organizations.

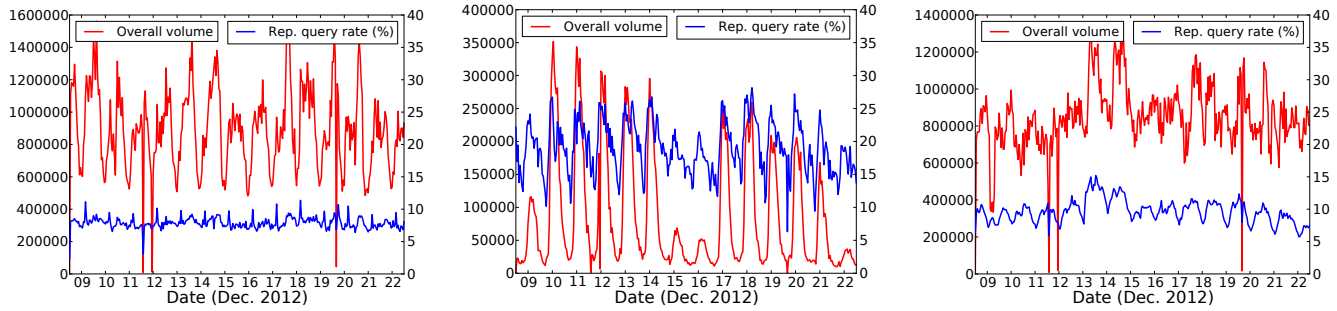
1) Simulation Methodology: For our analysis, we simulate an infinite resolver cache while replaying the captured DNS traffic. If the query returns an A, AAAA, or PTR record, the resolver knows the IP address of the queried domain or the domain for the queried IP address. It should not issue a query for the same domain or IP address before the TTL expires. If it issues such a query, we count it as a repeated query.

We find that from the perspective of our resolvers, the percentage of repeated queries that is issued to the root servers and other authoritative severs is 12.0% and 8.8% respectively. The ratio of repeated query varies significantly across different organizations. We plot the hourly repeated query ratio, in addition to the overall query volume for major organizations in Figure 7. Although some organizational subnets have a repeated query rate of 20% or higher, their traffic volume is low. The repeated query rates for the largest subnets lie between 10% and 15%.

2) Hourly Plot of Repeated Query Ratio: The three resolvers shown in Figure 8 exhibit very different characteristics. The university resolver (Figures 8(b)) has the highest repeated query rate. Meanwhile, it also exhibits a strong positive correlation (p-value < 0.001) between the repeated query rate and the query volume (*i.e.* , the repeated query rate rises when the query volume rises). In addition, its overall query volume shows a clear diurnal pattern and weekly pattern. The traffic peaks appear during business hours of each day. Much more DNS traffic occurs during weekdays and less traffic during weekends.

The commercial ISP resolver (Figure 8(a)) has a repeated query rate that varies between 5% and 10% during most of the days. The overall query traffic also exhibits a strong diurnal pattern, *i.e.* , the traffic volume rises during night time and falls during day time. It reflects the typical network usage of a residential network. However, we do not observe strong weekly pattern. In addition, a strong positive correlation (p-value < 0.001) between the overall query volume and the repeated query rate exists.

The public DNS resolver (Figure 8(c)) has fluctuating repeated query rate ranging from 5% to 15%. Because its users span different time zones, we naturally observe neither diurnal patterns nor weekly patterns from its overall query volume. Although hard to observe from the plot, statistical



(a) US ISP A (subnet 3) (b) EU EDU (subnet 1) (c) Public DNS Service

Fig. 8. The hourly repeated query ratio and overall DNS query volume for typical resolvers

tests also indicate a strong positive correlation (p-value $<$ 0.001) between its overall query volume and its repeated query rate.

While many resolvers exhibit strong positive correlation between the query volume and the repeated query rate, it is not always the case. The former suggests that cache eviction plays an important role in the volume of repeated queries. The higher the query volume is, the higher the repeated query rate will be. We further observe that resolvers within a /24 subnet show high homogeneity (*i.e.* , either all of them or none of them exhibit strong correlation, with very few exceptions). We omit these graphs due to space considerations. This reflects on the administrative policies within /24 subnets, the choice and configuration of network and DNS software.

3) Possible Causes: To further understand the cause of these repeated queries, we perform additional analysis to separate repeated queries that were issued in close proximity (the remainder could be attributed to cache eviction). We find that over 75% of repeated queries (across all resolvers) are due to related queries issued in close temporal proximity and the remainder are likely due to cache evictions at the resolver. We investigate two popular resolver implementations BIND (9.9.2-P1) and Unbound (1.4.16), as well as the behaviors of OpenDNS and GoogleDNS, from which we distill a few possible implementation-related factors that cause repeated queries in close temporal proximity.

• **CNAME Chain Sanitization.** When a response includes multiple records forming a CNAME chain, both BIND and Unbound issue extra queries to verify the trustworthiness of the chain. This is an intentional security enhancement to counter the Kaminsky attack [19], which could cause repeated queries and increased response times. Nearly 20% of A and AAAA queries in our dataset were eventually responded to with CNAME answers, which makes CNAME chain sanitization contribute to about 40% of all repeated queries in our simulation.

• **Concurrent Overlapping Queries.** A resolver could issue repeated queries if it receives two *overlapping* queries in close proximity. Two queries are considered overlapping if they belong to *either* of the two cases: (i) They request identical name; or (ii) Some parts of their delegation chain or CNAME chain are identical. If the identical segment is missing in the cache, the resolver will send two identical requests, which will be counted as repeated query. Implementing birthday attack protection [33] can help mitigate this effect. We observe that both BIND and Unbound have implemented birthday attack protection, but interestingly GoogleDNS and OpenDNS do not strictly suppress identical queries. When probing GoogleDNS and OpenDNS with identical queries, we observed repeated queries from same resolver instances (IP addresses), although the numbers of repeated queries are less than our identical probes. One possible explanation is that Google and OpenDNS implement birthday attack protection per thread, therefore identical queries processed by mutiple threads on same instance could still trigger repeated queries.

• **Premature Retransmissions.** We found that Unbound takes an arguably aggressive retransmission strategy, waiting for only one round-trip time before it retransmits the request. BIND is more conservative and has a minimum retransmission timeout of 800 ms. In our local experiments, we observed that Unbound issued several times more repeated queries than did BIND due to its premature retransmission timer.

• **Resolver Quirks.** Resolvers might also have some implementation quirks (or bugs) that could trigger repeated queries in some cases. We have found that, in certain cases, BIND will resolve expired NS names twice before replying to client queries, resulting in repeated queries and increased response times. Given the complexity of the name resolution process, we suspect similar vagaries could lurk in resolver implementations.

In Table IX, we summarize comparisons made in this paper with five prior studies and highlight our results. Note that this table only includes a subset of our measurement results that are directly comparable with results in prior work.

IV. Related Work

DNS Measurement Studies. Many prior studies have measured the performance of the DNS infrastructure. Multiple measurement studies conducted at root servers reported that a large percentage of traffic at root servers is invalid [1, 10, 11, 34]. In particular, Brownlee et al. discovered that 60%-85% of queries to the F-root server are repeated [10]. Castro et al. analyzed traffic collected from multiple root servers and reported that 98% of the traffic is invalid [11]. Castro et al. confirmed in a later study that a low fraction of busy clients (0.55%) generate the most invalid traffic at root servers [1]. We cross-compare some of these same results from the perspective of a globally distributed resolver set to assess the persistence

Compared Feature	Prior Work	Result Summary
Root perspective		
Private IP address in PTR Query	[10, 34]	The ratio in 2001, 2002 and 2012 is 7%, 1.61% and 0.1%, respectively.
Invalid TLDs	[10, 11, 34]	The ratio is steady 20% in 2001, 2002 and 2008, but rises to 50% in 2012.
A-*for*-A	[10, 11, 34]	The ratio is 12%, 7.03%, 2.7% and <0.1% in 2001, 2002, 2008 and 2012.
Non-printable character in query name	[11, 34]	The ratio drops from 1.94% (year 2002) to 0.1% (year 2008), but rises to 3.2% in 2012
DNS query type breakdown	[11, 34]	The ratio of AAAA queries increases from 4.7% (year 2002), to 15% (year 2008) to 26.6% (year 2012)
Queries with negative answer	[18]	14.7%, 27.3% and 54% in Jan 2000, Dec 2000 and 2012, respectively.
Local perspective		
DNS query type breakdown	[18]	There are no AAAA queries in 2000, but 13.4% of all queries are AAAA queries in 2012.
Queries with negative answer	[18]	The ratio rises from 11.1% (year 2000) to 18.0%.
Queries with no answer	[18]	The ratio drops from 23.5% (year 2000) to 15.1% (year 2012).
Cause of negative answer	[18]	PTR and A queries cause the most negative answer in 2000 and 2012, respectively.
TTL distribution	[18]	TTLs of A records in 2012 are much smaller than those in 2000.
DNSBL query ratio	[17]	0.4%, 14.0% and 1.7% in year 2000, 2004 and 2012, respectively.

TABLE IX
SUMMARY OF COMPARISONS WITH PRIOR MEASUREMENT STUDIES

of such problems. Our vantage point provides a different perspective and greater opportunity for understanding the root cause of certain phenomena.

Jung et al. analyzed SMTP traffic with DNS blacklist lookups [17]. In this work, we compare the DNS blacklist usage in 2012 with their reported findings. Ager et al. used active probing techniques to compare local DNS resolvers with public DNS services like GoogleDNS and OpenDNS in terms of latency, returned address, and so on, by actively issuing DNS queries from more than 50 commercial ISPs [2]. Otto et al. studied the impact of using public DNS resolvers instead of local resolvers on the network latency of CDN content fetching [29]. Liang et al. measured and compared the latency of root and TLD servers from various vantage points [20]. Our measurement study has a different goal from theirs. In particular, we study the performance of recursive DNS resolvers. We do not cover the client perceived DNS performance in our study.

DNS Performance Studies. Jung et al. characterized the DNS traffic obtained from two university sniffers and evaluated the effect of different TTL values with trace driven simulations [18]. Pang et al. measured DNS server responsiveness from the vantage points inside a large content distribution network [30] finding that a significant fraction of LDNS resolvers do not honor TTLs. Wessels et al. measured how the cache policy of different DNS software affects the number of DNS queries by trace driven simulations [35]. Bhatti et al. conducted experiments to reduce the TTL of A records on university DNS resolvers and found a low increase in DNS traffic [8]. While the focus of this paper is on the broad high-level characteristics of DNS data such as the overall distribution of query types and failures, prevalence of repeated queries etc., revisiting the implications of caching and DNS performance in greater depth in the context of the SIE dataset is future work.

V. CONCLUSIONS

In this paper, we conduct a comprehensive measurement study with more than 26 billion DNS query-response pairs collected from 600+ global DNS resolvers. Besides reaffirming some findings in published work, our results reveal some significant differences. We witness the demise of A-for-A queries and a significant rise in AAAA queries. We also find that queries for invalid TLDs are responsible for more than 99% of queries with negative answer observed at root servers and that TTLs of A records become much smaller than a decade ago. In Table IX, we summarize comparisons made in this paper with five prior studies and highlight our results. Note that this table only includes a subset of our measurement results that are directly comparable with results in prior work.

Our findings can help implementation, deployment, and configuration of DNS software, websites, and other applications. First, because of the increase of AAAA queries for IPv6 addresses, websites should take IPv6 support into account. DNS providers and ISPs should also consider to support IPv6 for DNS resolution because of the high percentage of IPv6 resolvable domains. The high failure ratio of PTR queries implies that some DNS administrators pay less attention to configuring reverse mappings from IP addresses to domain names. The high rate of invalid TLD queries to root servers suggests that client-side implementations should differentiate local names (used only in Intranets) with global domain names. Our analysis of repeated queries reveals that complementary security enhancements of resolvers could have non-negligible effects on DNS resolution, suggesting that more evaluations should be conducted before the wide adoption of such features. Our analysis also reveals several possible optimizations to suppress unnecessary queries.

Furthermore, we propose a novel approach that isolates malicious domain groups from temporal correlation in DNS queries, using a few known malicious domains as anchors. On average, this approach achieves more than 96% detection accuracy while producing more than 50 previously unknown malicious domains for every known malicious anchor domain.

ACKNOWLEDGMENT

We are grateful to Paul Vixie and the SIE contributors for providing us access to this extraordinary data source. We would also like to thank Yunsheng Cao for his help with some of our early analysis scripts. This material is based upon work supported in part by the National Science Foundation under grant nos. CNS-0831300 and CNS-1314956, Army Research Office under Cyber-TA Grant no. W911NF-06-1-0316, the National Basic Research Program (973 Program) of China under grant no. 2009CB320505, the National Natural Science Foundation of China under grant no. 61472215, and the Open Foundation of State key Laboratory of Networking and Switching Technology (Beijing University of Posts and Telecommunications) (SKLNST-2013-1-17). Any opinions,

findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

REFERENCES

[1] Understanding and preparing for DNS evolution. In *Traffic Monitoring and Analysis*, volume 6003 of *Lecture Notes in Computer Science*. 2010.

[2] B. Ager, W. Mühlbauer, G. Smaragdakis, and S. Uhlig. Comparing DNS resolvers in the wild. In *Proceedings of the 10th ACM SIGCOMM Internet Measurement Conference*, 2010.

[3] M. Antonakakis, R. Perdisci, W. Lee, N. Vasiloglou, and D. Dagon. Detecting malware domains at the upper DNS hierarchy. In *Proceedings of the USENIX Security Symposium*, 2011.

[4] R. Arends, R. Austein, M. Larson, D. Massey, and S. Rose. DNS Security Introduction and Requirement. RFC 4033, 2005.

[5] R. Arends, R. Austein, M. Larson, D. Massey, and S. Rose. Protocol Modifications for the DNS Security Extensions. RFC 4035, 2005.

[6] R. Arends, R. Austein, M. Larson, D. Massey, and S. Rose. Resource Records for the DNS Security Extensions. RFC 4034, 2005.

[7] S. M. Bellovin. Using the Domain Name System for System Break-ins. In *the 5th conference on USENIX UNIX Security Symposium - Volume 5 (Security 1995)*, pages 18–18, Berkeley, CA, USA, 1995. USENIX Association.

[8] S. Bhatti and R. Atkinson. Reducing DNS caching. In *Computer Communications Workshops*, april 2011.

[9] L. Bilge, E. Kirda, C. Kruegel, and M. Balduzzi. EXPOSURE : Finding malicious domains using passive DNS analysis. In *18th Annual Network and Distributed System Security Symposium*, San Diego, 02 2011.

[10] N. Brownlee, k. claffy, and E. Nemeth. DNS measurements at a root server. In *IEEE Global Telecommunications Conference (GLOBECOM)*, Nov 2001.

[11] S. Castro, D. Wessels, M. Fomenkov, and K. Claffy. A day at the root of the internet. *SIGCOMM Comput. Commun. Rev.*, 38(5):41–46, Sept. 2008.

[12] P. B. Danzig, K. Obraczka, and A. Kumar. An analysis of wide-area name server traffic: a study of the internet domain name system. In *Proceedings of the ACM SIGCOMM Conference*, 1992.

[13] C. J. Dietrich. Feederbot - a bot using DNS as carrier for its C&C. http://blog.cj2s.de/archives/28-Feederbot-a-bot-using-DNS-as-carrier-for-its-CC.html, 2011.

[14] Farsight Security, Inc. Farsight Security Archive. https://archive.farsightsecurity.com/.

[15] T. Holz, C. Gorecki, K. Rieck, and F. C. Freiling. Measuring and detecting fast-flux service networks. In *Proceedings of Network and Distributed Security Symposium*, 2008.

[16] ICANN. TLD DNSSEC Report. http://stats.research.icann.org/dns/tld_report/.

[17] J. Jung and E. Sit. An empirical study of spam traffic and the use of DNS blacklists. In *Proceedings of the 4th ACM SIGCOMM Internet Measurement Conference*, 2004.

[18] J. Jung, E. Sit, H. Balakrishnan, and R. Morris. DNS performance and the effectiveness of caching. *IEEE/ACM Transactions on Networking*, 10(5):589–603, Oct. 2002.

[19] D. Kaminsky. It is the end of the cache as we know it. BlackHat USA, 2008.

[20] J. Liang, J. Jiang, H. Duan, K. Li, and J. Wu. Measuring query latency of top level DNS servers. In *Proceedings of Passive and Active Measurement Conference*, 2013.

[21] MaxMind, Inc. http://www.maxmind.com/.

[22] P. Mockapetris. Domain Names—Concepts and Facilities, RFC 1034. http://www.ietf.org/rfc/rfc1034.txt.

[23] P. Mockapetris. Domain Names—Concepts and Facilities, RFC 882. http://www.ietf.org/rfc/rfc882.txt.

[24] P. Mockapetris. Domain Names—Implementation and Specification, RFC 1035. http://www.ietf.org/rfc/rfc1035.txt.

[25] P. Mockapetris. Domain Names—Implementation and Specification, RFC 883. http://www.ietf.org/rfc/rfc883.txt.

[26] C. Mullaney. Morto worm sets a (DNS) record. http://www.symantec.com/connect/blogs/morto-worm-sets-dns-record, 2011.

[27] E. Osterweil, D. McPherson, S. DiBenedetto, C. Papadopoulos, and D. Massey. Behavior of DNS top talkers, a .com/.net view. In *Proceedings of Passive and Active Measurement Conference*. 2012.

[28] E. Osterweil, M. Ryan, D. Massey, and L. Zhang. Quantifying the operational status of the DNSSEC deployment. In *the 8th ACM SIGCOMM Internet Measurement Conference*, pages 231–242. ACM, 2008.

[29] J. S. Otto, M. A. Sánchez, J. P. Rula, and F. E. Bustamante. Content delivery and the natural evolution of DNS: remote dns trends, performance issues and alternative solutions. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference*, 2012.

[30] J. Pang, J. Hendricks, A. Akella, R. De Prisco, B. Maggs, and S. Seshan. Availability, usage, and deployment characteristics of the domain name system. In *Proceedings of the 4th ACM SIGCOMM Internet Measurement Conference*, 2004.

[31] K. Sato, keisuke Ishibashi, T. Toyono, and N. Miyake. Extending black domain name list by using co-occurrence relation between DNS queries. In *Proceedings of LEET*, 2010.

[32] J. Spring, L. Metcalf, and E. Stoner. Correlating domain registrations and DNS first activity in general and for malware. In *Securing and Trusting Internet Names*, 2011.

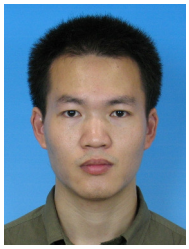
[33] J. Stewart. DNS cache poisoning–the next generation, 2003.

[34] D. Wessels and M. Fomenkov. Wow, That's a lot of packets. In *Passive and Active Network Measurement Workshop (PAM)*, 2003.

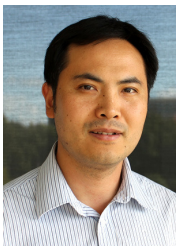
[35] D. Wessels, M. Fomenkov, N. Brownlee, and k. claffy. Measurements and laboratory simulations of the upper DNS hierarchy. In *Passive and Active Network Measurement Workshop*. 2004.

Hongyu Gao Hongyu Gao received his B.S. degree in computer science from Peking University, Beijing, China, in 2008. He received his Ph.D. degree in computer science from Northwestern University, Evanston, USA, in 2013. His research interests span the areas of online social network security and spam detection. He is now a software engineer at Google.

Vinod Yegneswaran Vinod Yegneswaran in a Senior Computer Scientist at SRI International pursuing advanced research in network and systems security. He obtained his A.B. in Computer Science from the University of California at Berkeley and his Ph.D. in Computer Science from the University of Wisconsin, Madison. He has served on several NSF panels and program committees of security and networking conferences, including the IEEE security and privacy symposium. His current research interests include DNS traffic analysis, SDN security and anti-censorship technologies.



Jian Jiang Jian Jiang received his B.E. and M.E. degrees in computer science from Xi'an Jiaotong University, Xi'an, China, in 2002 and 2005 respectively, and Ph.D. degree in computer science from Tsinghua University, Beijing, China, in 2014. He is now a research specialist at UC Berkeley. His research interests include security analysis of network protocols and systems, and network measurement.



Yan Chen Dr. Yan Chen is a Professor in the Department of Electrical Engineering and Computer Science at Northwestern University, Evanston, IL. He got his Ph.D. in Computer Science at University of California at Berkeley in 2003. His research interests include network security, measurement and diagnosis for large scale networks and distributed systems. He won the Department of Energy (DoE) Early CAREER award in 2005, the Department of Defense (DoD) Young Investigator Award in 2007, and the Best Paper nomination in ACM SIGCOMM 2010. Based on Google Scholar, his papers have been cited for over 7,000 times and his h-index is 34.

Phillip Porras Phillip Porras holds an M.S. in Computer Science from the University of California, Santa Barbara. He is an SRI Fellow, and a program director of the Internet Security Group in SRI's Computer Science Laboratory. Phillip continues to publish and conduct technology development on numerous topics including intrusion detection and alarm correlation, privacy, malware analytics, active and software defined networks, and wireless security. He has participated on numerous program committees and editorial boards, participates on multiple commercial company technical advisory boards, and holds 12 U.S. patents.

Shalini Ghosh Shalini Ghosh received her Ph.D. in Computer Engg. from UT Austin in 2005, and is currently a Senior Computer Scientist in the Computer Science Laboratory at SRI International. At present she is a Visiting Scientist at Google Research. Her main area of research interest is applications of Machine Learning and Data Mining for Natural Language Understanding and Dependable Computing. She is also interested in probabilistic relational models, e.g., Markov Logic Networks. In 2013, she and her co-authors received the Best Paper Award from the 19th IEEE PRDC conference. She has served on the program committees of many conferences (e.g. KDD, NAACL), review-boards of journals, NSF panels, and was invited to be a guest lecturer in an EECS course at UC Berkeley.

Haixin Duan Haixin Duan is a professor of Tsinghua University in Beijing and a visiting scholar at ICSI (International Computer Science Institute) in Berkeley. He got his Ph.D degree from Tsinghua University and B.E. and M.E degrees from Harbin Institute of Technology on computer science. Professor Duan's research interests include network security and network measurement.