In rDNS We Trust: Revisiting a Common Data-Source's Reliability

Tobias Fiebig[1,2,3], Kevin Borgolte[2], Shuang Hao[4],
Christopher Kruegel[2], Giovanni Vigna[2], Anja Feldmann[3,5]

[1]TU Delft; [2]UC Santa Barbara; [3]TU Berlin;
[4]UT Dallas; [5]Max Planck Institute for Informatics;

Abstract. Reverse DNS (rDNS) is regularly used as a data source in Internet measurement research. However, existing work is polarized on its reliability, and new techniques to collect active IPv6 datasets have not yet been sufficiently evaluated. In this paper, we investigate active and passive data collection and practical use aspects of rDNS datasets. We observe that the share of non-authoritatively answerable IPv4 rDNS queries reduced since earlier studies and IPv6 rDNS has less non-authoritatively answerable queries than IPv4 rDNS. Furthermore, we compare passively collected datasets with actively collected ones, and we show that they enable observing the same effects in rDNS data. While highlighting opportunities for future research, we find no immediate challenges to the use of rDNS as active and passive data-source for Internet measurement research.

1 Introduction

The Domain Name System (DNS) is an integral part of the Internet. Forward DNS, i.e., resolving names like `www.google.com` to an IP address makes the Internet usable for end-users. Its counterpart is reverse DNS (rDNS), which allows resolving the name behind an IPv4 or IPv6 address. To resolve an IP address to a name, IANA designated two second level zones below .arpa, in-addr.arpa (IPv4) and ip6.arpa (IPv6). Below them, operators receive zones corresponding to their IP network prefixes. In the assigned zones, operators can serve pointer (PTR) resource records to point to the fully qualified domain name (FQDN) for an IP address. Example use cases of rDNS are the forward confirmation of mail servers to fight spam [1], and enriching logs for improved readability and debugging [2]. Furthermore, researchers regularly leverage rDNS to gather valuable information on networks, e.g., topologies [3, 4], the deployment state of IPv6 [5], etc.

Even though rDNS is a valuable data-source for researchers, it is not clear how rDNS is used, and whether its DNS zones are well maintained. Gao et al. report that 25.1% of all rDNS queries cannot be authoritatively answered [6], while Phokeer et al. report an increasing number of broken rDNS delegations for the APNIC region [7]. Furthermore, the reliability of new active collection techniques for IPv6 rDNS as used by Fiebig et al. [5] has not yet been investigated. Therefore, in this paper, we revisit prior research on the use of rDNS by operators and investigate the validity of active rDNS collection techniques. We make the following contributions:

– We re-visit the use of rDNS by clients and operators beyond the scope of earlier studies, e.g., Gao et al. [6], and observe that queries for IPv6 rDNS lack an authoritative answer less frequently than for IPv4 rDNS queries.

– We compare the technique by Fiebig et al. to actively obtain rDNS datasets with our passive trace datasets. We find that they are complementary and provide appropriate and meaningful datasets for future research relying on active rDNS traces.

2 Related Work

rDNS use by clients: Prior work on the use of rDNS itself is commonly part of more general approaches to understand DNS lookup patterns. The most notable are Hao et al. in 2010 [8], as well as Gao et al. in 2013 [6] and 2016 [9]. In their 2013 work, Gao et al. note that 25.1% of all PTR queries in their dataset do not receive an authoritative answer, which might be an indication of poorly maintained rDNS zones. We use the same data-provider as Gao et al. for the passive traces in our study.

Active rDNS traces use: Especially in the domain of topology discovery researchers heavily rely on DNS measurements. For example, Spring et al. [10], as well as Oliveira et al. [4] supplement network topology discovery with rDNS data. Similarly, rDNS information has seen use by security studies, such as Czyz et al., who leverage rDNS to identify dual-stack (IPv4 and IPv6) hosts, for which they then evaluate the host's security posture [11]. Note, that these studies use IPv4 rDNS, as it can be brute-force enumerated. Actively collecting global IPv6 rDNS traces is however possible by exploiting semantics of the DNS protocol to prune the search-tree of the rDNS zone when enumerating it, as demonstrated by Fiebig et al. [5], or by zone-walking via DNSSEC extensions [12]. In this paper, we use the technique by Fiebig et al. to collect active datasets for our study.

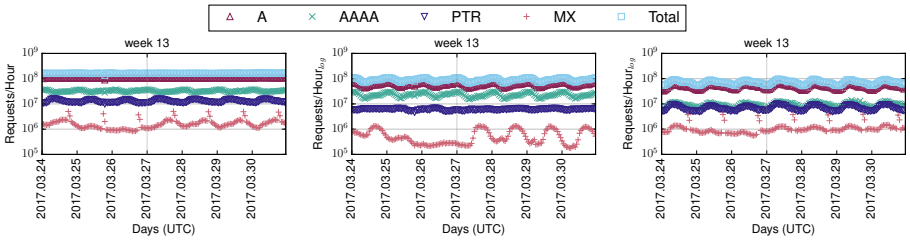
3 Passive Traces on rDNS: What Can We See?

We leverage Farsight's passive DNS dataset for data on real-world use of rDNS by clients. The dataset contains traces from DNS resolvers around the globe, providing a global overview of DNS lookup behavior [6, 9]. A full description of the collection infrastructure is out of scope for this work. The interested reader can find a comprehensive analysis in earlier publications using the dataset [6, 9]. For our study we use DNS traffic (query response pairs) observed between March 23$^{\text{rd}}$, 2017 and April 17$^{\text{th}}$, 2017 from midnight to midnight (UTC).

3.1 Biases in the Passive Traces

In a first examination of the data, we find irregular requests from a single ISP's recursive DNS resolvers (see Figure 1): There is no diurnal pattern for the total *No.* of requests/A requests, and the patterns for AAAA and PTR queries are disjoint. PTR queries are dominated by requests for names in ip6.int., the discontinued rDNS zone for IPv6 [13], belonging to addresses in an unused IPv6 range(7000::/8). Similarly, we observe DNS Service Discovery [14] PTR requests for icloud.com, dell.com, etc., in large (cumulative) volume but in the same order of magnitude of requests per second level zone. These offending requests stem from recursors belonging to a single operator.

Therefore, we split the dataset in two subsets: The ISP showing the unexpected requests pattern, and the remaining operators. Interestingly, the single operator



(a) Full Farsight dataset. (b) Only biased operator. (c) w/o biased operator.

Fig. 1. The first week of the passive trace for the three dataset splits. We visualize only the first week to enhance the readability of the figures. The outliers for MX requests (the DNS Resource Record (RR) type to denote the mailserver(s) handling mail for a domain) in Figure 1(c) stem from a Russian ISP running a daily mass-email campaign.

contributes close to half of all queries in the original dataset (see Figure 1(b)). Note, that the most likely source of the irregular requests is misconfigured Customer-Premises Equipment or an internal service. By excluding the operator, the remaining dataset appears more regular (see Figure 1(c)) and conforms to the overall volumes found in earlier studies [6, 9]. Hence, we acknowledge that there are biases in our dataset, and that there may be further biases we were unable to detect. Investigating these should be part of further work. Nevertheless, as we can control for the biases we do find, we consider our dataset admissible for the work at hand.

3.2 Dataset Overview

Next, we look at second level domains for which PTR queries are issued to distinguish between rDNS queries for IPv4 and IPv6 addresses and other use cases of PTR records. Comparatively, requests to in-addr.arpa (the IPv4 reverse zone), are two orders of magnitude more frequent than requests for ip6.arpa (the IPv6 reverse zone).

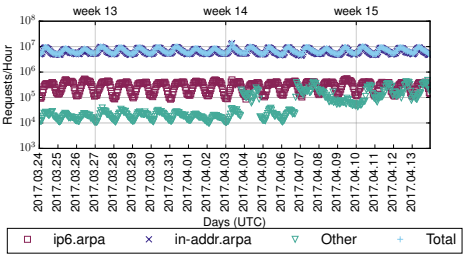
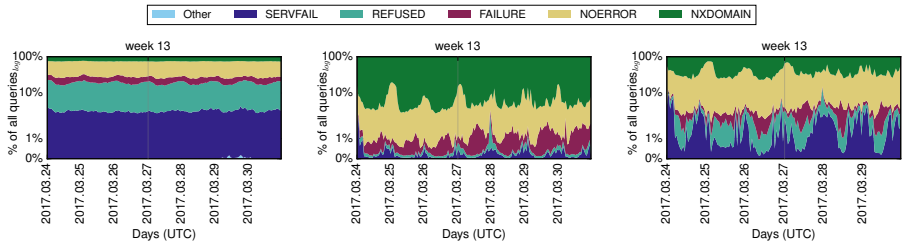


Fig. 2. Requests to ip6.arpa, in-addr.arpa and other second-level domains.

Beyond the IPv4 and IPv6 rDNS zones, we observe PTR requests to other top and second level domains. These are mostly related to DNS based service discovery (DNS-SD) by clients, which account for 77.04% of observed queries outside of .arpa (see Figure 2). Outliers in the "Other" category starting April 4$^{\text{th}}$, 2017 correspond to DNS-SD queries for services in the domain of a major news network, which leaked into the Farsight dataset through a single operator. A newly deployed model of Customer-Premises Equipment (CPE), such as a set-top box, or such a device receiving an update on or shortly before April 4$^{\text{th}}$, 2017 is the most likely source for the observed behavior. For the remainder of the paper, we focus on queries to in-addr.arpa and ip6.arpa, i.e., queries that can be clearly and safely attributed to rDNS. Next, we investigate the DNS response codes of in-addr.arpa and ip6.arpa



(a) in-addr.arpa (b) ip6.arpa (c) ip6.arpa w/o Resv.

Fig. 3. Share of response codes during the first week of our measurements. Note, that queries for reserved IPv6 addresses' rDNS accounts for over 95% of all IPv6 rDNS queries.

to determine if we still encounter the high number of queries that do not receive an authoritative answer in our data than it was observed by prior work.

3.3 DNS Response Codes in in-addr.arpa

For in-addr.arpa, 47.21% of all queries are successful, while 25.36% return NXDOMAIN, and 15.47% return REFUSED, possibly because the operators want to hide internal information which could become public from host names returned for the RRs (see Figure 3(a)). The brief increase of "Other" replies on March 29th, 2017 is due to DNS servers of a Singaporean ISP returning FORMERR for all requests. Furthermore, we find on average 3.17% of queries returning SERVFAIL, indicating that some zone delegations in in-addr.arpa are broken, or that the authoritative DNS server does not respond correctly, e.g., because the zone files/databases are inaccessible by the DNS server daemon. Another 8.77% queries result in other failures, e.g., packet loss etc., denoted as FAILURE and less than 0.02% result in FORMERR, NOTAUTH, and NOTIMP. Overall, only 12.06% of PTR requests to in-addr.arpa cannot be authoritatively answered, which stands in significant contrast to the 25.1% reported earlier by Gao et al. [6]. More important, only 3.17% of queries cannot be authoritatively answered due to broken delegations, i.e., due to a lack of care and maintenance.

3.4 DNS Response Codes in ip6.arpa

rcode	in-addr.-arpa	ip6.arpa	ip6.arpa w/o Resv.
NOERROR	47.21%	4.00%	32.30%
NXDOMAIN	25.36%	94.87%	63.87%
REFUSED	15.47%	0.14%	1.11%
FAILURE	8.77%	0.81%	1.34%
SERVFAIL	3.17%	0.18%	1.38%
FORMERR	0.01%	≤0.01%	≤0.01%
NOTAUTH	≤0.01%	-	-
NOTIMP	≤0.01%	-	-

Table 1. Distribution of rcodes for ip6.arpa and in-addr.arpa during the full measurement period.

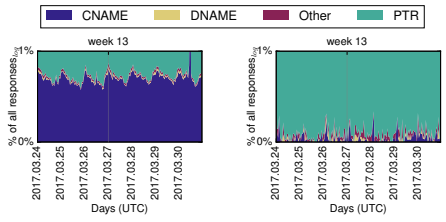
Contrary to in-addr.arpa, for ip6.arpa, only 0.99% of all requests cannot be authoritatively answered. However, we also find that just 4.00% of queries result in a NOERROR response. Instead, ip6.arpa is dominated by NXDOMAIN replies, which account for 94.87% of all responses (see Figure 3(b)). The large share of NXDOMAIN responses is caused by a small number of heavy hitter prefixes. Interestingly, these networks are exclusively local and reserved-use prefixes. This may be related to an, by now non-existent, effect observed by Wessels and Fomenkov for IPv4 in 2003 [15].

Excluding these hosts yields a more coherent picture, which we refer to as "ip6.arpa w/o Resv." (see Table 1 and Figure 3(c)). After filtering out reserved addresses, the overall response rate increases and NXDOMAIN responses account for 63.87%, and NOERROR responses correspond to 32.30%. The number of FAILURE and SERVFAIL responses does not significantly change: They remain relatively low compared to in-addr.arpa. We conjecture that SERVFAIL is less frequent for ip6.arpa than it is for in-addr.arpa because in-addr.arpa has been in use much longer. As such, it provides more time for things to go wrong, i.e., delegations and systems to become stale and to break [16]. The lower REFUSED rate for ip6.arpa may be due to less security measures being in place for IPv6 systems and infrastructure [11].

4 Passive Traces: What are rDNS Use-Cases?

We make additional observations on how operators use rDNS as landmarks to cross-compare findings from our active rDNS traces later on.

4.1 RRtypes in Successful Answers



(a) in-addr.arpa (b) ip6.arpa

Fig. 4. Share of response types.

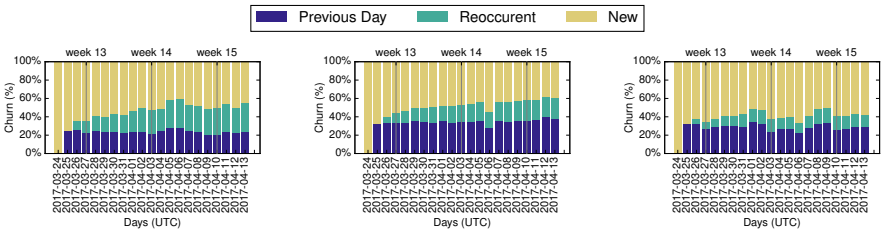
Naturally, the RRtypes of responses to rDNS queries are dominated by PTR RRs. Given that in-addr.arpa is split at octet boundaries, while IPv4 networks are not anymore, we expect a notable number of CNAME responses for in-addr.arpa, but not for ip6.arpa. Specifically, the share of CNAMEs should be higher for in-addr.arpa as they are used to delegate rDNS authority for networks that are smaller than a /24 [17]. Indeed, we find that CNAMEs account for 0.71% of all query responses in in-addr.arpa. While the share is comparatively low, it constitutes a steady baseline compared to ip6.arpa (see Figure 4). Similarly, we observe a small layer of DNAMEs—similar to CNAMEs, but for a full zone—for in-addr.arpa, but not for ip6.arpa. Other record types (A, SOA, etc., labeled "Other" in the graph) relate to additional information sent by authoritative nameservers, e.g., sending along the A record for the returned FQDN in a PTR request.

4.2 rDNS SMTP Forward Confirmation

Port	Protocol	Port	Protocol
25	SMTP	587	SMTP Submission
110	pop3	993	IMAPs
143	IMAPv4	995	pop3s
465	SMTPs	-	ICMP

Table 2. Scanned ports and protocols.

Following, we revisit the share of mail servers for which we see rDNS queries, most likely for forward confirmation, i.e., as a tool to mitigate email spam [1], where it was extremely successful. For the purpose of our study, mail servers are all systems with services listening to send and receive email. We performed simple active measurements for email servers on April 19th, 2017. We scan all hosts for which we see rDNS queries as soon as they appear in the dataset



(a) in-addr.arpa (b) ip6.arpa (/64) (c) ip6.arpa (/128)

Fig. 5. Churn for requested names in in-addr.arpa and ip6.arpa.

and we ensure that every host is only scanned once. For each host, we check if it replies to ICMP echo requests and if it at least one email related TCP port (see Table 2) is open.

Specifically, 19.98% of all addresses for which we see in-addr.arpa requests respond to ICMP echo requests, while 15.28% of all hosts for which we see ip6.arpa requests reply to ICMPv6. Hosts running email services contribute 10.05% of responding hosts in in-addr.arpa, amounting to 2.01% of all hosts for which we saw IPv4 rDNS queries. However, for ip6.arpa, 31.53% of reachable hosts, or 4.82% of all hosts, exhibit open email related ports.

Forward confirmation is commonly not performed for MUA (Mail User Agent) connections that try to relay an email. Here, the user trying to send an email is required to authenticate herself. Hence, forward confirmation should be performed mostly for: (i) spam senders, and, (ii) other email servers. However, with the increased use of blacklists, and adoption of Sender Policy Framework (SPF) and DomainKeys Identified Mail (DKIM) over the past years, spam distribution moved to using (compromised) email servers, or sending spam emails via compromised email accounts of legitimate users [18]. Although our results are a lower bound, they indicate that there are relatively more server systems among the IPv6 hosts for which we see rDNS requests than there are for IPv4 hosts for which we see rDNS requests.

4.3 Churn in Queried Reverse Names

Continuing on the notion of (more dynamic) clients and servers, we investigate the churn of queried rDNS names in our dataset, for each day, which we define as the individual shares of: (i) Names queried on the previous day as well, (ii) Names queried on any other prior, but not the previous, day, and, (iii) Names never queried before. We focus on the churn in requested names, as a heavy-hitter analysis for requesting end-hosts is not possible as this information is not included in the dataset [9] due to privacy concerns. However, if our assumption is correct, we should observe a comparatively small foundation of stable addresses, accompanied by a large amount of reoccurring and newly queried names.

Figure 5(b) and 5(c) show the churn for ip6.arpa aggregated to /64s and for full addresses, Figure 5(a) shows the churn for in-addr.arpa. In both cases, we excluded queries for private and reserved addresses. Hence, we can reason about how many reverse queries are issued for server systems (i.e., systems that commonly reoccur), and how many are issued for clients with changing addresses. We include the aggregation to /64s for IPv6 to account for IPv6 privacy extensions. During our

three week measurement period, in-addr.arpa and per-/64 aggregated requests to ip6.arpa exhibit around 50% of reoccurring records after three weeks (49.74% for in-addr.arpa and 54.12% for ip6.arpa), while for full IPv6 addresses, 43.35% of records reoccur. Over time, the share of seen names being queried for changes: On average, 24.29% of all records in in-addr.arpa reoccur on subsequent days, while in ip6.arpa 30.52% of names reoccur, rising to 35.46% when aggregating to /64s.

These results indicate that, especially for IPv6 a far higher number of IPv6 hosts for which we see rDNS queries are, indeed, not clients, or long-lived clients not using privacy extensions. Furthermore, we find that the small number of reoccurring hosts for full IPv6 addresses aligns with findings of prior work in respect to the dynamic use of /64s for IPv6 privacy extensions [19].

5 Active rDNS Measurements: What Can We Really See?

To continue our investigation of rDNS, we actively collected in-addr.arpa and ip6.arpa datasets employing and extending a rDNS collection technique we have previously published [5]. The resulting datasets allow us to estimate how many IPv6 addresses have a corresponding rDNS entry set, what portion of the rDNS space we can enumerate, and how the active dataset relates to the passive datasets.

5.1 Data Collection Infrastructure

We use a cluster of 16 machines to collect the dataset, each machine is comprised of an Intel Xeon X3450 CPU, 8GB of main memory, 300GB of hard disk storage. Each system also runs a local recursive DNS resolver (Unbound 1.4.22), against which we perform all DNS queries to benefit from caching. The cluster is orchestrated by an additional workstation that distributes jobs using GNU parallel. Lastly, there were no middle-boxes or connection-tracking routers on the path up to the default-free zone (DFZ).

5.2 Dataset and Toolchain Availability

Our toolchain is open-source, and it is documented and available at: `https://gitlab.inet.tu-berlin.de/ptr6scan/toolchain`. We provide the actively collected data to other researchers on request only, due to privacy and security concerns: The collected datasets include a significant number of server-side IPv6 addresses that are not covered by prior research, likely containing vulnerable hosts [11].

5.3 IPv6 rDNS Dataset Collection

We use our previously published enumeration technique [5] to collect our dataset. Our technique utilizes that DNS servers should respond with NXDOMAIN (DNS status code 3) only if they are queried for a non-existent name in the DNS tree which has *no* children in comparison to a name for which they know that it has children, where they should reply with NOERROR (DNS status code 0) [20]. We exploit this to prune the ip6.arpa. tree while enumerating it, thereby making an enumeration of the tree feasible, despite its size [5]. In essence, our algorithm works as follows:

– We collect seeds of IPv6 prefixes by aggregating a global routing table.

– In parallel, for each seed, starting with a target length of four nibbles:
  • If the seed is longer than the target length, we crop it accordingly and add both, the seed and the cropped seed back to the seed-set.
  • If the seed is shorter, we request all possible children (0-f). Based on the authoritative servers response we only descent subtrees with existing children up until we reached the target length, then add these items back to the seed set.
– When we went through the whole seed-set, we increase the target length by four nibbles, up until a length of 32 nibbles (128bit, a full IPv6 record) and re-do the parallel block of the algorithm.

Our technique also accounts for dynamically generated zones, slow authoritative servers, and systems that are not vulnerable to enumeration using RFC8020 [5].

We collect data from April 22^{nd}, 2017 04:07 UTC to April 25^{th}, 2017 10:15 UTC, which contains more than 10.2 million reverse records. Our dataset includes intermediate information for non-terminal records, to understand how IPv6 reverse zones are delegated and to compare that to the IPv4 datasets. Furthermore, in addition to PTR records, we also collect CNAME records.

5.4 IPv4 rDNS Dataset Collection

We extended and improved our RFC8020 based technique from prior work to support the IPv4 rDNS zone. In contrast to a brute-force approach, it allows us to investigate delegation in IPv4 rDNS:

1. We collect a view on the global routing table from RIPE RIS and Routeviews and add in-addr.arpa to the seed set.
2. We use RFC8020 based enumeration to perform a breadth-first search in the tree (instead of 16, every node now has 256 possible children).
3. When the algorithm finds a terminal node, we terminate for that branch.

Leveraging our extended technique, we collect an in-addr.arpa NXDOMAIN dataset between March 31^{st}, 2017 16:28 UTC and April 6^{th}, 2017 05:46 UTC, which spans 1.21 billion terminal records and CNAMEs.

5.5 Visible IPv4 Space: The Size of the Internet

By comparing the in-addr.arpa dataset with the global IPv4 space, we can approach the question of how well rDNS is maintained and populated by network operators. In an ideal world, we would see rDNS names, i.e., either CNAMEs or PTRs, for all allocated IPv4 addresses. Hence, the number of all active IPv4 addresses should closely model the number of IPv4 rDNS names we find. We note, that this is merely a rough indication, and a careful evaluation would first compile a dataset of all active addresses, similar to Richter et al. [21], and then look up the rDNS names for each of the IPv4 addresses in that dataset. However, within the scope of this study, we focus on an indicative numerical comparison.

With 1.21 billion PTR records in the in-addr.arpa dataset, we observe rDNS names for 28.17% of the total IPv4 address space, which numerically corresponds to the 1.2 billion active IPv4 addresses observed by Richter et al. [21] using active *and*

passive measurements. Note, that our approach may overestimate the number of hosts in this mapping (rDNS being set for whole networks, e.g., in access networks, despite not all addresses being in use), as well as underestimate it (hosts lacking rDNS entries despite being active). Nevertheless, based on our observation we at least conjecture that rDNS zones are not only regularly delegated (see Section 3), but also that network operators do indeed populate and maintain their rDNS zones. Based on our prior observation that ip6.arpa zones are less frequently involved in broken delegations or have unresponsive servers than in-addr.arpa zones, we expect to see a similar overlap of active IPv6 addresses and the ip6.arpa zone.

Visible IPv6 Space: ip6.arpa vs. CDN Dataset. For evaluating the active IPv6 space, prior work leveraging the CDN dataset forms the current state of the art base-line for investigating IPv6 adoption [19]. The CDN dataset is a dataset consisting of IP addresses that were collected from a major CDN's access logs. Researchers with access to the dataset kindly provided us with comparative aggregated values on our dataset. They reported a plain overlap between our ip6.arpa dataset with 10.2M records and their CDN dataset, with over 300M IPv6 addresses per day, of 81K hosts, out of which they identify 70K as stable, i.e., reoccurring on three subsequent days. Therefore, we conclude that our ip6.arpa dataset covers other parts of the IPv6 address space than the CDN dataset.

We assume that the root cause for this mismatch can be found in ISPs' handling of IPv6 access networks: ISPs commonly hand out /64s or /48s networks to their customers [19]. Therefore, they dynamically generate zones starting at the covering standard prefix size, i.e., /32s or /48s. This corresponds to the most commonly dynamically generated zones in the ip6.arpa dataset being /32s and /48s (see Figure 6(b)). Hence, the most likely reason for the low overlap with the CDN dataset is that the CDN dataset is client-centric, while we hardly see clients as we exclude dynamically generated zones, which are common for client networks.

Visible IPv6 Space: RFC8020 Compliance. The enumeration technique we used heavily depends on authoritative servers correctly implementing RFC8020 [20]. If a major portion of the authoritative DNS servers handling IPv6 rDNS zones does not conform to RFC8020, visibility may be limited. Therefore, we investigate how frequently rDNS servers adhere to the RFC. From the Farsight dataset, we collected all successful queries for entries in ip6.arpa, a total of 361K unique names. For each record, we determine all zone delegations up to the root (ip6.arpa) via which the leaf record can be reached, and we then query for the NS records of all intermediate zones.

Utilizing the initial leaf records, we test each of the authoritative name servers for all identified domains if they: (i) follow RFC8020; (ii) always return NXDOMAIN, even though an element in the zone tree below them exists; (iii) always return NOERROR, even though nothing exists below the queried records; (iv) do return an error (SERVFAIL, REFUSED, timeouts); and, (v) if there are any differences for this between the different authoritative servers of a domain.

We discover that 39.58% of all rDNS zones in the dataset only use authoritative servers in compliance with RFC8020, while 46.42% always return NXDOMAIN,

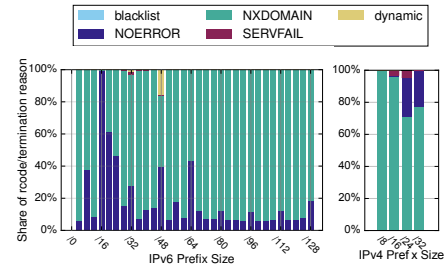
and 11.61% always return NOERROR. In turn, we will detect 46.42% of zones as having no entries at all, while 11.61% of zones will be flagged as dynamically generated due to the behavior of their authoritative servers. The remaining 2.38% are split among 0.59% of zones that return errors, and 1.79% of zones exhibiting a mix of the above conditions. Interestingly, in case of the latter, at least one nameserver is compliant with RFC8020 and can be used for enumeration, while the others always return NXDOMAIN or NOERROR.

Therefore, the likelihood that the NXDOMAIN technique is effective ranges around 40% for each *individual* zone/server. Nevertheless, upon comparing our IPv6 seed set with the delegation pattern for IPv6 rDNS, we find that the majority of top-level delegations up to /48s is covered by seeds (see Figure 6(b)). It means that we do not lose a significant number of (large) sub-trees within the rDNS tree, and instead only lose around 40% of all /64s and below, which leaves us with an estimated coverage between 16% and 40%. Furthermore, our results indicate that querying all authoritative servers of a zone during enumeration is not strictly necessary. Although it can increase the result set for some zones, the additional overhead can not be justified by the 1.79% of zones that could be enumerated additionally.

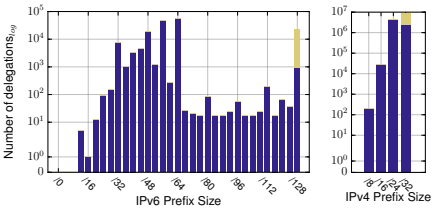
6 Comparing Active and Passive Results

6.1 CNAMEs and Delegations

In our passive dataset, we observed that CNAMEs are used to delegate rDNS authority for networks smaller than the minimum rDNS zone size. That is, smaller than a /24 network for in-addr.arpa (see Section 3). Furthermore, we find that requests to in-addr.arpa show a higher rate of SERVFAILs than requests to ip6.arpa. Correspondingly, we should find evidence of these effects in our active traces as well. Next, we look into how delegations and CNAMES occur in our active rDNS traces.



(a) Observed rcodes.



(b) Delegations (CNAMEs in yellow).

Fig. 6. rcodes and delegation for rDNS.

rDNS Zone Delegation. To investigate delegation in rDNS, we build a trie from the gathered reverse zones. Specifically, we first sort the zones by corresponding prefix size, and then add them to the trie. Sorting them before adding them to the trie ensure that we do not add a longer prefix before we add the covering shorter prefix. For each input zone, we check if a less specific prefix exists in the trie. If it exists, then we check if the authority section for the associated domain is the same. If the zone in the authority section differs, then

we encountered a delegation for the current prefix length. For terminal records, we also check if the zone reported in the authority section is a well-formed PTR zone, either under ip6.arpa or in-addr.arpa (depending on the zone we evaluate). If not, then it is a CNAME for a terminal record instead of a delegation.

Delegations within in-addr.arpa happen consistently (see Figure 6(b)): /8s are delegated to RIRs (and some Internet early-adopters who received large prefixes [21]) that are then split by the RIRs and delegated to LIRs in smaller blocks, which are further delegated to end-users and small network operators. This pattern extends down to the terminal records, where we find a high number of delegation attempts, as well as 6.2 million CNAME records. Indeed, this number corresponds to 0.51% of all 1.21 billion in-addr.arpa records are CNAMEs, close to the expected 0.71% of CNAME responses (see Section 3). Moreover, a majority of the target-zones (92.85%) that CNAMEs point to have more than one CNAME pointing to them, conforming to the designated purpose of CNAMEs in in-addr.arpa: Delegating rDNS for networks smaller than a /24, as suggested by RFC2317 [17].

For ip6.arpa, delegations mostly occur for the most commonly assigned prefix lengths, i.e., /32s, /48s, /56s, and /64s. As expected, this relates closely to the more structured addressing policies that became possible with the significantly larger address space of IPv6. In case of IPv4, a large operator may use several smaller prefixes collected from various RIRs [21], however, with IPv6, a single prefix is enough. Hence, ip6.arpa delegation happens mostly for larger prefixes.

Following IPv6 addressing best practices, we expected that most delegations occur for /48s and /56s, because /64s are the suggested maximum prefix length for a subnet and the prefix-length that should be assigned to an interface [22, 23]. We did not expect /64s to be individually delegated, as a customer with multiple subnets should receive a /48 or /56 instead. However, we find that the total number of delegations actually increases from /48s to /64s, where it peaks. We even encounter delegations for prefixes more specific than /64s, each peaking at the corresponding 4-nibble-block boundaries. Surprisingly, a high number of CNAMEs for terminal records exist, which were unexpected due to the better delegation option in ip6.arpa, with per-nibble zone boundaries.

In our dataset, 87.81% of observed IPv6 rDNS CNAMEs belong to the DHCPv6 range of a single operator, which uses them to point PTR records from a full /96 representation in the ip6.arpa zone to another zone of the form ip6.arpa-suffix.ip6.dhcp6.operator.tld. Fiebig et al. already briefly mentioned such setups [5]. Most (80.77%) of the remaining 12.19% records point to names in in-addr.arpa, to ensure coherent addressing in dual-stack scenarios. Consequently, this indicates an "IPv4 first" policy employed by operators: Operators first deploy IPv4, and then roll out IPv6 on top, leveraging CNAMEs to ensure consistency through-out the network. Yet, IPv4 remains the leading technology, even though the setup is dual-stack. Relating these numbers back to Section 3, we find that CNAMEs are slightly more common than expected, constituting 0.22% of the dataset. However, if we consider the single DHCPv6 operator as an artifact and exclude it, then we arrive at the expected low CNAME density of 0.02%, which matches the share of records of the passive trace.

SERVFAIL in the Active Traces. Finally, we observed that SERVFAILs are much more frequent for in-addr.arpa than for ip6.arpa (see Section 3). We find corroborating evidence for this in the active datasets: For in-addr.arpa, 3.40% of zones at the /16 level, and 4.87% of zones at the /24 level result in SERVFAIL (see Figure 6(a)). In contrast, for ip6.arpa, we only find a small amount of SERVFAIL for /32s and /48s, totaling 2.14% of all /32s, and 1.02% of all /48s. We attribute this to the fact that ip6.arpa has not been in use for as long as in-addr.arpa, and, in turn, had far less time to become stale and to accumulate broken delegations.

7 Conclusion

In this paper, we revisited prior results on the use of rDNS and find that rDNS zones are by now less frequently non-authoritatively answerable than observed in earlier studies [6]. We have also revisited previously presented techniques to obtain active rDNS datasets. Network behavior that we observe in the Farsight passive trace dataset are also present in the actively collected datasets, supporting the assertion that active rDNS measurement techniques produce meaningful datasets without requiring access to expensive datasets or global network vantage points. Beyond confirming prior assumptions, we find first indications for an "IPv4-first" approach by operators, i.e., operators plan and build IPv4 infrastructures first, and then deploy IPv6 later on, in their use of zone-delegations and CNAMEs for rDNS zones. These observations should be further investigated in the future. Ultimately, we find no challenges to the use of rDNS as a data-source for Internet measurement research, even though this should be closely monitored in the future. Hence, we argue that rDNS can be relied on for Internet-wide studies.

Acknowledgements We thank the anonymous reviewers and John Heidemann for their helpful feedback. We also thank David Plonka for his valuable feedback and the comparison with the CDN dataset. This material is based on research sponsored by the Defense Advanced Research Projects Agency (DARPA) under agreement number FA8750-15-2-0084, the Office of Naval Research (ONR) under grant N00014-17-1-2011 and N00014-15-1-2948, the National Science Foundation (NSF) under grant DGE-1623246 and CNS-1704253, and a Google Security, Privacy and Anti-Abuse Award to Giovanni Vigna. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. Any views, opinions, findings, recommendations, or conclusions contained or expressed herein are those of the authors, and do not necessarily reflect the position, official policies, or endorsements, either expressed or implied, the U.S. Government, DARPA, ONR, NSF, or Google.

References

[1] Cormack, G.V.: Email spam filtering: A systematic review. Foundations and Trends in Information Retrieval **1**(4) (2007)

[2] Nicholas, D., Huntington, P.: Micro-mining and segmented log file analysis: A method for enriching the data yield from internet log files. SAGE Journal of Information Science **29**(5) (2003)

[3] Zhang, M., Ruan, Y., Pai, V.S., Rexford, J.: How DNS Misnaming Distorts Internet Topology Mapping. In: Usenix Annual Technical Conference (ATC). (2006)

[4] Oliveira, R.V., Pei, D., Willinger, W., Zhang, B., Zhang, L.: In search of the elusive ground truth: Yhe Internet's AS-Level Connectivity Structure. In: Proc. ACM SIGMETRICS. Volume 36. (2008)
[5] Fiebig, T., Borgolte, K., Hao, S., Kruegel, C., Vigna, G.: Something From Nothing (There): Collecting Global IPv6 Datasets From DNS. In: Proc. Passive and Active Measurement (PAM). (2017)
[6] Gao, H., Yegneswaran, V., Chen, Y., Porras, P., Ghosh, S., Jiang, J., Duan, H.: An empirical reexamination of global DNS behavior. Proc. ACM SIGCOMM **43**(4) (2013)
[7] Phokeer, A., Aina, A., Johnson, D.: DNS Lame delegations: A case-study of public reverse DNS records in the African Region. (2016)
[8] Hao, S., Feamster, N., Pandrangi, R.: An Internet-Wide View into DNS Lookup Patterns. Technical Report, School of Computer Science, Georgia Tech (2010)
[9] Gao, H., Yegneswaran, V., Jiang, J., Chen, Y., Porras, P., Ghosh, S., Duan, H.: Reexamining DNS from a global recursive resolver perspective. IEEE/ACM Trans. Networking (TON) **24**(1) (2016) 43–57
[10] Spring, N., Mahajan, R., Wetherall, D., Anderson, T.: Measuring ISP Topologies with Rocketfuel. IEEE/ACM Trans. Networking (TON) **12**(1) (2004) 2–16
[11] Czyz, J., Luckie, M., Allman, M., Bailey, M.: Don't Forget to Lock the Back Door! A Characterization of IPv6 Network Security Policy. In: Proc. Internet Society Symposium on Network and Distributed System Security (NDSS). (2016)
[12] Borgolte, K., Hao, S., Fiebig, T., Kruegel, C., Vigna, G.: Enumerating Active IPv6 Hosts for Large-scale Security Scans via DNSSEC-signed Reverse Zones. In: Proc. IEEE Security & Privacy (S&P). (2018)
[13] Huston, G.: Deprecation of "ip6.int". RFC 4159 (Best Current Practice) (August 2005)
[14] Cheshire, S., Krochmal, M.: DNS-Based Service Discovery. RFC 6763 (Proposed Standard) (February 2013)
[15] Wessels, D., Fomenkov, M.: Wow, that's a lot of packets. In: Proc. of Passive and Active Measurement Workshop (PAM). (2003)
[16] Borgolte, K., Fiebig, T., Hao, S., Kruegel, C., Vigna, G.: Cloud Strife: Mitigating the Security Risks of Domain-Validated Certificates. In: Proc. Internet Society Symposium on Network and Distributed System Security (NDSS). (2018)
[17] Eidnes, H., de Groot, G., Vixie, P.: Classless IN-ADDR.ARPA delegation. RFC 2317 (Best Current Practice) (March 1998)
[18] Hu, X., Li, B., Zhang, Y., Zhou, C., Ma, H.: Detecting compromised email accounts from the perspective of graph topology. In: Proc. ACM Conference on Future Internet Technologies. (2016)
[19] Plonka, D., Berger, A.: Temporal and spatial classification of active IPv6 addresses. In: Proc. ACM Internet Measurement Conference. (2015)
[20] Bortzmeyer, S., Huque, S.: NXDOMAIN: There Really Is Nothing Underneath. RFC 8020 (Proposed Standard) (November 2016)
[21] Richter, P., Smaragdakis, G., Plonka, D., Berger, A.: Beyond Counting: New Perspectives on the Active IPv4 Address Space. In: Proc. ACM Internet Measurement Conference. (2016)
[22] IAB, IESG: IAB/IESG Recommendations on IPv6 Address Allocations to Sites. RFC 3177 (Informational) (September 2001) Obsoleted by RFC 6177.
[23] de Velde, G.V., Popoviciu, C., Chown, T., Bonness, O., Hahn, C.: IPv6 Unicast Address Assignment Considerations. RFC 5375 (Informational) (December 2008)

Errata

After publication of the paper, we were made aware of two minor issues in the originally published version of this article, which we corrected in this version. These issues are:

- In the in-addr.arpa part of Figure 6 a) on page 10, coloring was off. IPv4 rDNS zones returning SERVFAIL were illustrated using the beige reserved for dynamic zones in the IPv6 graphs, instead of the correct wine-red used for SERVFAIL zones.
- In Section 5.3 on page 8, we reported that we used an ip6.arpa rDNS dataset collected between March 26^{th}, 2017 01:04 UTC to March 30^{th}, 2017 10:49 UTC. However, we collected the dataset that we analyzed between April 22^{nd}, 2017 04:07 UTC and April 25^{th}, 2017 10:15 UTC. No incorrect data was reported and the results remain valid.

We thank our fellow researchers Robert Beverly from the Naval Postgraduate School and Oliver Gasser from TU Munich for pointing out these mistakes. Discovery of them would not have been possible without our publication of our collected dataset. We hence attach a plea for Open Data in the field of network measurements with this errata.