Estimation on the Root DITL Dataset

Kazunori Fujiwara, JPRS
fujiwara@jprs.co.jp

DITL dataset

- https://www.dns-oarc.net/oarc/data/ditl
  - A Day in the Life of the Internet is a large-scale data collection project initially undertaken by CAIDA and subsequently by OARC every year since 2006.
  - OARC offers access these data to researchers and OARC members through the use of a small fleet of analysis machines.

- I am grateful to OARC and the contributing data providers for the opportunity to analyze the Root DITL dataset.

Root DITL dataset

- The Root DITL dataset looks like:
  - Each Root Server Operator (RSO) collects data for a given period to the extent possible and makes the data available in the way that their own policies
    - It may not include all data for that period, and some queries are not available.
    - Some RSOs anonymize query source IP addresses.
  - Description of policy of each dataset has not been disclosed
    - Which data has been collected / omitted
    - How have the query source IP addresses been anonymized ?

Examining the Root DITL 2024/2025 dataset

- Therefore, I analyzed the status of the Root DITL 2024 / 2025 dataset.

1. Viewing the raw data by tcpdump or other tools

2. Matching with sent queries
   - In April 2024/2025, I sent queries for my private, unused domain names and "hostname.bind" queries to 26 root server addresses from my server, approximately every hour
   - Extracted queries which I sent from DITL dataset

Result: c(2024), g, k, m -root

- All my queries are included
- Query source IP addresses were non-anonymized (preserved)
  - both IPv4 and IPv6

- Remarks
  - k-root dataset is located another (ripe) directory on 2014 to 2024

Result: b, d, h -root

- All my queries were included
- Query source IP addresses were anonymized
  - /24 prefix seems to be preserved on IPv4
  - /64 prefix seems to be preserved on IPv6
  - May be 1 to 1 mapping
- The dataset made UDP checksum error
  - UDP checksum may be preserved
- No IPv4 header checksum error
- b, d and h -root seem to apply the same anonymization

Result: a, j -root

- All my queries were included
- Query source IP addresses were anonymized
  - /24 prefix seems to be preserved on IPv4
  - /96 prefix seems to be preserved on IPv6
  - May be 1 to 1 mapping
- The dataset made UDP checksum error on IPv4
  - No UDP checksum error on IPv6
- No IPv4 header checksum error
- a-root and j-root seem to apply the same anonymization

Result: f-root

- None of my queries were included
  - My "hostname.bind" queries show "NRT.cf.f.root-servers.org“
  - No data found for “cf” (Cloudflare?) nodes
- Query source IP addresses were anonymized
  - /24 prefix seems to be preserved on IPv4, the lowest 8bit is 0
  - /64 prefix seems to be preserved on IPv6, 64bit local part is 0
- The dataset made UDP checksum error
  - UDP checksum may be preserved ???
- IPv4 header checksum error
  - IPv4 header checksum may be preserved ???

Result: e-root

- Only included limited dataset (single node only?)
- None of my queries were included
- The dataset did not contain any IP addresses that I'm familiar with
- Therefore, I cannot assume the IP address anonymization status

Result: i-root

- All query source IP addresses were replaced to 10.0.0.0/8 and fe80::/16

- None of my queries were included (on 2024)

Result: L-root (2024)

- Almost of all my queries were included
- Query source IP addresses (prefix) were anonymized
  - destination (L-root) IP addresses were also changed
  - IPv6: found 9b78:: address (no prefix preserved)
  - IPv4: no prefix preserved
- No UDP checksum error
  - UDP checksum is rewritten
- No IPv4 header checksum error
  - IPv4 header checksum is rewritten

Summary of DITL 2024/2025 datasets

- Without IP address anonymization
  - c, g, k, m
- Preserved /24, /64 prefix, lower addresses were changed, 1 to 1 mapping
  - b, d, h
- Preserved /24, /96 prefix, lower addresses were changed, 1 to 1 mapping
  - a, j
- Preserved /24, /64 prefix, lower address was 0
  - f
- Query source IP addresses were unknown, prefix non-preserved
  - i, L
- Handling status cannot be estimated
  - e

DITL dataset coverage
Comparison with RSSAC002 data

- DITL-2024 dataset contains Apr. 10, 2024, 24 hours data
- DITL-2025 dataset contains Apr. 9, 2025, 24 hours data
- RSSAC002 dataset contains daily query-volume, query source IP addresses

Apr 10, 2024 DITL dataset vs RSSAC002

root	Number of IPv4 queries			IPv4 query source			Number of IPv6 queries			IPv6 query source		
	DITL	RS002	Ratio	DITL	RS002	Ratio	DITL	RS002	Ratio	DITL	RS002	Ratio
a	8.5E+09	8.0E+09	106%	12488220	10581090	118%	2.2E+09	2.1E+09	104%	1930607	1727683	112%
b	5.7E+09	5.8E+09	98%	10237642	10710458	96%	1.8E+09	1.8E+09	98%	1109043	1769990	63%
c	5.0E+09	5.0E+09	100%	9943932	9945057	100%	1.2E+09	1.2E+09	100%	1675979	1695282	99%
d	5.9E+09	5.7E+09	102%	9673281	9612759	101%	2.0E+09	1.9E+09	104%	1118411	1720877	65%
e	4.4E+07	4.5E+09	1%	68514	9735390	1%	3.1E+06	1.8E+09	0%	3595	955566	0%
f	2.1E+09	8.4E+09	25%	302825	961726	31%	1.9E+09	4.1E+09	46%	52006	143959	36%
g	3.9E+09	2.8E+09	139%	9848193			8.3E+08	5.9E+08	140%	1519067		
h	4.2E+09	4.2E+09	100%	8944853	8931635	100%	1.5E+09	1.5E+09	99%	1122243	1706651	66%
i	4.8E+09	6.7E+09	72%	5285201	4198002	126%	1.6E+09	2.3E+09	72%	1375272	1282051	107%
j	7.5E+09	7.3E+09	104%	11216843	9962041	113%	2.5E+09	2.5E+09	102%	1861871	1693263	110%
k	6.2E+09	6.4E+09	97%	9758243	9970981	98%	3.0E+09	3.0E+09	100%	1460572	1526425	96%
l	7.4E+09	7.4E+09	99%	9726955	9768881	100%	2.9E+09	2.9E+09	99%	1661197	1713938	97%
m	4.4E+09	4.4E+09	100%	9696561	9787114	99%	1.2E+09	1.2E+09	100%	1670186	1725076	97%

a, b, c, d, g, h, j, k, l, m-root: DITL dataset is complete within the limits of measurement / calculation errors

Apr 10, 2024 DITL dataset vs RSSAC002

root	Number of IPv4 queries			IPv4 query source		
	DITL	RS002	Ratio	DITL	RS002	Ratio
e	4.4E+07	4.5E+09	1%	68514	9735390	1%
f	2.1E+09	8.4E+09	25%	302825	961726	31%
i	4.8E+09	6.7E+09	72%	5285201	4198002	126%

root	Number of IPv6 queries			IPv6 query source		
	DITL	RS002	Ratio	DITL	RS002	Ratio
e	3.1E+06	1.8E+09	0%	3595	955566	0%
f	1.9E+09	4.1E+09	46%	52006	143959	36%
i	1.6E+09	2.3E+09	72%	1375272	1282051	107%

- e-root: DITL dataset is limited (1% of RSSAC002's number of queries)
- f-root: DITL dataset has 25% of RSSAC002's number of queries on IPv4
- i-root: DITL dataset has 72% of RSSAC002's number of queries.

Apr 9, 2025 DITL dataset vs RSSAC002

root	Number of IPv4 queries			IPv4 query source			Number of IPv6 queries			IPv6 query source		
	DITL	RS002	Ratio	DITL	RS002	Ratio	DITL	RS002	Ratio	DITL	RS002	Ratio
a	9.52E+09	8.5E+09	112%	8276753	8027160	103%	2.3E+09	2.2E+09	108%	1439193	1390686	103%
b	6.15E+09	6.2E+09	99%	6402393	6679950	96%	1.2E+09	1.2E+09	98%	1363756	1431773	95%
d	8.43E+09	8.4E+09	100%	6184873	6189296	100%	2.8E+09	2.7E+09	103%	1380375	1375560	100%
e	57380551	1E+10	1%	50559			4218398	3.2E+09	0%	5140		
f	2.06E+09	1E+10	20%	313768	1010552	31%	1.7E+09	4.7E+09	36%	55814	120188	46%
g	5.23E+09	1.9E+09	276%	6597748			9.7E+08	3.5E+08	277%	918792		
h	5.54E+09	5.5E+09	100%	5491428	5483965	100%	1.6E+09	1.7E+09	95%	1380070	1379532	100%
i	5.38E+09	8.2E+09	66%	4622948	4938131	94%	1.6E+09	2.3E+09	67%	910241	859123	106%
j	9.85E+09	9E+09	110%	6777642	6590291	103%	3E+09	2.8E+09	106%	1392345	1357332	103%
k	8.98E+09	9.1E+09	98%	6506488	6599113	99%	3.5E+09	3.6E+09	99%	968429	1245942	78%
m	6.07E+09	6.1E+09	99%	6385071	6484841	98%	1.4E+09	1.4E+09	100%	1103047	1383010	80%

a, b, d, h, j, k, m-root: DITL dataset is complete within the limits of measurement / calculation errors.
(I'm not sure about the difference with g-root)

Apr 9, 2025 DITL dataset vs RSSAC002

root	Number of IPv4 queries			IPv4 query source		
	DITL	RS002	Ratio	DITL	RS002	Ratio
e	57380551	1E+10	1%	50559		
f	2.06E+09	1E+10	20%	313768	1010552	31%
i	5.38E+09	8.2E+09	66%	4622948	4938131	94%

root	Number of IPv6 queries			IPv6 query source		
	DITL	RS002	Ratio	DITL	RS002	Ratio
e	4218398	3.2E+09	0%	5140		
f	1.7E+09	4.7E+09	36%	55814	120188	46%
i	1.55E+09	2.3E+09	67%	910241	859123	106%

- e-root:DITL dataset is limited (1% of RSSAC002 number of queries)
- f-root: DITL dataset has 20% of RSSAC002's number of queries on IPv4
- i-root: DITL dataset has 66% of RSSAC002's number of queries.

Summary of DITL dataset coverage

- e-root
  - DITL dataset may be limited
  - may be from single node
- f-root
  - cf nodes data may be omitted / missing
- i-root
  - some data may be omitted / missing
- a, b, c, d, g, h, j, k, l, m –root
  - DITL dataset may be complete
  - within the limits of measurement error
  - and calculation error (my own mistake)

Proposal: Request to data providers

- Please provide the following information
  - Percentage and range of data captured
  - Whether IP address anonymization is performed
    - Are /24 and /64 prefixes preserved ?
    - 1 to 1 mapping ?
    - Are any other fields being changed ?

- Representative of F-root offered me the answer
  - F-root preserves IPv4 /24, IPv6 /56 prefixes
    - lower bits are zero
  - F-root rewrites UDP checksum (on IPv4)
  - “cf” node data is not collected

Recovery of anonymized addresses ?

- If UDP checksum field is preserved,
- IPv4
  - If IP address anonymization changes the lowest 8 bits of an IPv4 address,
  - The anonymized IP address may be recovered by rewriting the changed 8 bits so that the checksum matches.
- IPv6
  - c, g, k, m-root datasets contain existing IPv6 addresses
  - Query source IPv6 address may be recovered by rewriting the existing IPv6 address with a matching /64 prefix to an IPv6 address with a matching UDP checksum.
  - IPv6 addresses not observed in c, g, k, or m-root cannot be recovered

Implemented in my tool

- My (pcap | bind9log | dnsjson) parse tool written in C
- https://github.com/kfujiwara/PcapParseC/
  - Sorry, without documentation, commit logs, …

Recovery status

- My server’s IPv4 and IPv6 addresses are recovered from b, d, h-root dataset
  - UDP checksum are preserved in b, d, h-root dataset

- My server’s IPv4 and IPv6 addresses are not recovered from a, j, f, I, L-root dataset
  - UDP checksum are rewritten in a-root, j-root dataset

Summary (1)

- To analyze data, it is necessary to understand the attributes of the data.
  - The DITL dataset is useful data, and we are grateful for the provision of the data and the analysis environment.
  - However, there is no information about whether the respective data sets anonymize IP addresses, what parts of the addresses are preserved, or whether the data is partial or complete.

Summary (2)

- I estimated the attributes of the DITL-2024/2025 dataset
  - c, g, k, m –root : No IP address anonymization, complete dataset
  - b, d, h-root: Query source address anonymized, 1 to 1 mapping, prefix preserved (/24 on IPv4, /64 on IPv6), complete dataset
  - a, j-root: Query source address anonymized, 1 to 1 mapping, prefix preserved (/24 on IPv4, /96 on IPv6), complete dataset
  - f-root: Query source address anonymized, prefix preserved (/24 on IPv4, /56 on IPv6, lower bits are 0), “cf” node data is not collected
  - e-root: limited dataset
  - I, L-root: Query source address anonymized, prefix non-preserved
- I also evaluated whether the original IP addresses can be estimated from the anonymized IP address