Understanding DNS Queries at B-Root

Jacob Ginesin, *Northeastern University* and Jelena Mirkovic, *USC/ISI*

Problem Statement

- **The Problem:** There is little known about the diversity of DNS resolver behaviors. Yet, DNS is a critical part of internet infrastructure; therefore, it's important to understand DNS behaviors and specifics.
- **The Challenge:** The complexity, diversity, and decentralized nature of DNS makes it difficult to enumerate behaviors. Additionally, there is little previous research on the subject.
- **Our Approach:** We analyzed DITL traces at B-root, collected between 2013-2022 (one day per year). We quantified malformed queries and identified dominant query senders. We also analyzed how query composition changed over time.

Overview

Dataset

- "A Day in the Life of the Internet" (DITL) DNS query dataset
- 4.4 Billion DNS Queries, collected across 10 years (2013-2022) during one specific day per year
- We sampled all queries between 12pm-1pm

Methodology

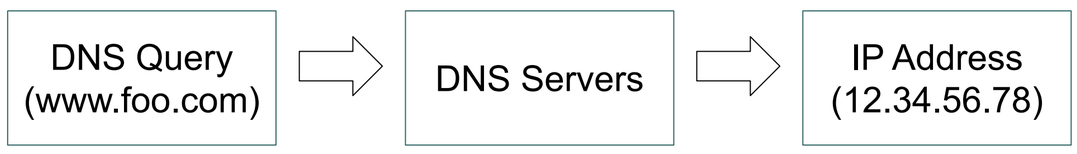
- We developed a query classification method:
  - To identify valid/expected queries
  - To identify common patterns of malformed queries
- We collected general sender statistics
- We conducted sender analysis stratified across each notable characteristic

Key Insights

- We quantified key metrics and trends in DNS traffic at B-Root, including:
  - % of chromium queries (xxxxxxxxx.)
  - QNAME minimization (see: RFC 7816)
  - Queries with & without valid TLD
  - Empty queries "."

Background - What is DNS?

- The Domain Name System (DNS) is the internet's system for mapping an alphanumeric web address (e.g., www.foo.com) to its respective IP address
- This system is invoked whenever a remote server is accessed (a URL is visited, an API request is made, or a server is SSH-ed into, etc)
- DNS is **critical internet infrastructure**



Previous Work

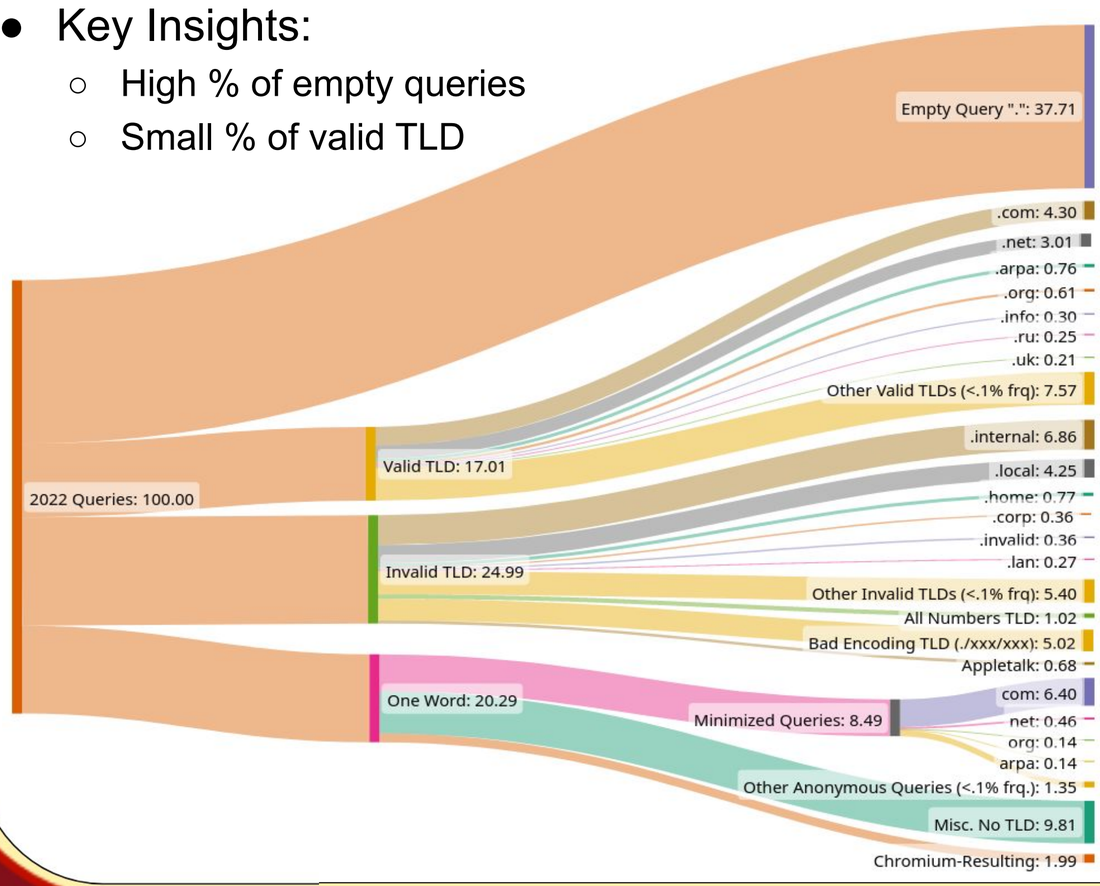
- "A Day at the Root of the Internet" - A DITL dataset analysis from 2008. Below is the taxonomy of queries (in %) across 8 root servers, based on 2008 DITL data.

Category	A	C	E	F	H	K	L	M	Total
Unused query class	0.1	0.0	0.1	0.0	0.1	0.0	0.1	0.1	0.1
A-for-A	1.6	1.9	1.2	3.6	2.7	3.8	2.6	2.7	2.7
invalid TLD	19.3	18.5	19.8	25.5	25.6	22.9	24.8	22.9	22.0
non-printable char.	0.0	0.1	0.1	0.1	0.1	0.0	0.1	0.0	0.0
queries with _	0.2	0.1	0.2	0.1	0.2	0.1	0.1	0.1	0.1
RFC 1918 PTR	0.6	0.3	0.5	0.2	0.5	0.2	0.1	0.3	0.4
identical queries	27.3	10.4	14.9	12.3	17.4	17.9	12.0	17.0	15.6
repeated queries	38.5	51.4	49.3	45.3	38.7	42.0	44.2	43.9	44.9
referral-not-cached	10.7	15.2	12.1	10.9	12.9	11.1	14.3	11.1	12.4
Valid	1.7	2.0	1.8	1.9	1.8	2.0	1.8	1.8	1.8
Valid 2006		2.3		2.1		2.5			2.1
Valid 2007		4.1		2.3		1.8		4.4	2.5

Current Results

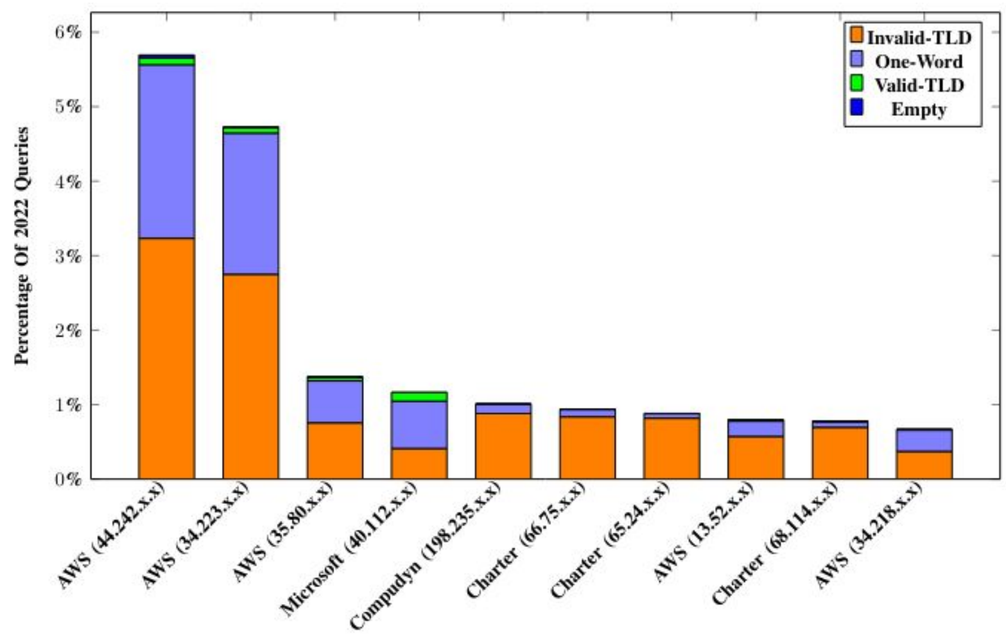
General Query Classification

- We stratified ~100 million 2022 DNS queries by query characteristic (in %):
- Key Insights:
  - High % of empty queries
  - Small % of valid TLD



Top Sender Analysis

- We analyzed the sender taxonomy for top senders to B-Root in 2022.
- Key Insights:
  - AWS alone accounts for 13.4% of B-Root DNS queries
  - High % of one-word and Invalid-TLD queries
  - Very low % of Valid TLD and empty "." queries



If interested contact ginesin.j@northeastern.edu
Work performed under REU Site program
supported by NSF grant #2051101

USC Viterbi School of Engineering
Information Sciences Institute