NCAP Data Questions for DNS Operators

(*Outreach to begin first week of May 2021*)

As part of the Name Collision Analysis Project Study Two objectives, additional research into DNS data sources is intended to help answer the Board's questions. This additional data research focuses primarily on the concept of data sensitivity. To what extent does measuring DNS queries from specific points in the DNS hierarchy impact an analyst's ability to conduct a name collision string assessment with a high degree of confidence?

The following sections contain questions that are designed to help provide insights into data sensitivity. These questions are tailored for two specific DNS observation spaces: root servers and recursive resolvers.

Root Data

RSSAC002 data already clearly shows that individual root server operators receive different amounts of DNS query volume from differing sets of source addresses. Analysis from A and J root servers already show a large amount of traffic is unique to a specific root letter's "catchment" via its anycasting and site placement. To what end is this catchment effect applied across all of the root letters?

How do root letters compare to each other?

1. IP distribution, overlap, and geographical distribution of source IP traffic

    Expected data format from RSO:
    - RSO to collect request for NXDomain responses and summarize on a daily basis to produce a delimited text file containing the following measurements based on the free version of the MaxMind GeoIP database:
        - Date
        - Source IPv4 or IPv6 Address (potentially /24 or /48 for PII concerns)
        - Source AS Number
        - ISO 3611-2 two letter country code of source IP
        - Number of queries
        - Root letter
    - Example:
        - 2021-03-19,172.217.42.192,AS15169,US,203837,A-Root

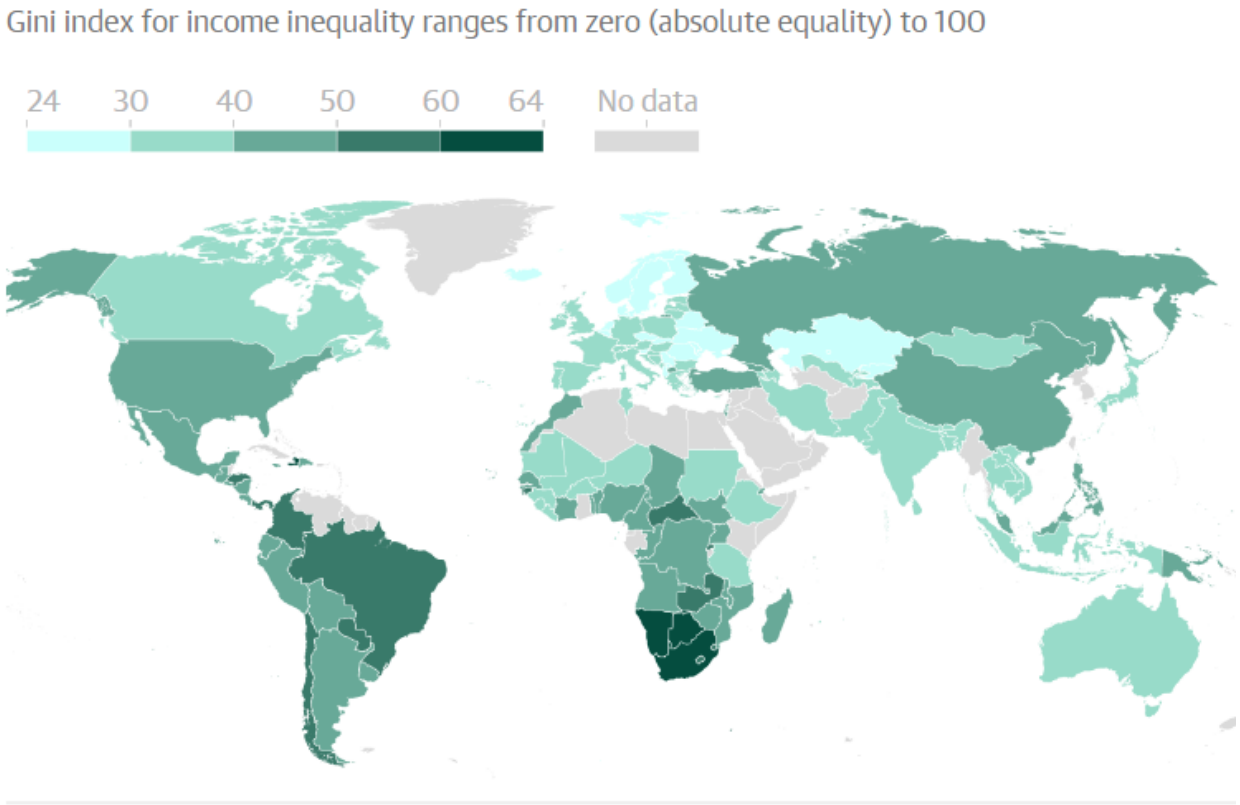
    Potential Graphs:

- Graph 1: A 13 x 13 matrix plot measuring pairwise set overlap (e.g. Jaccard or Cosine) of the 13 roots unique source IP addresses and ASNs.

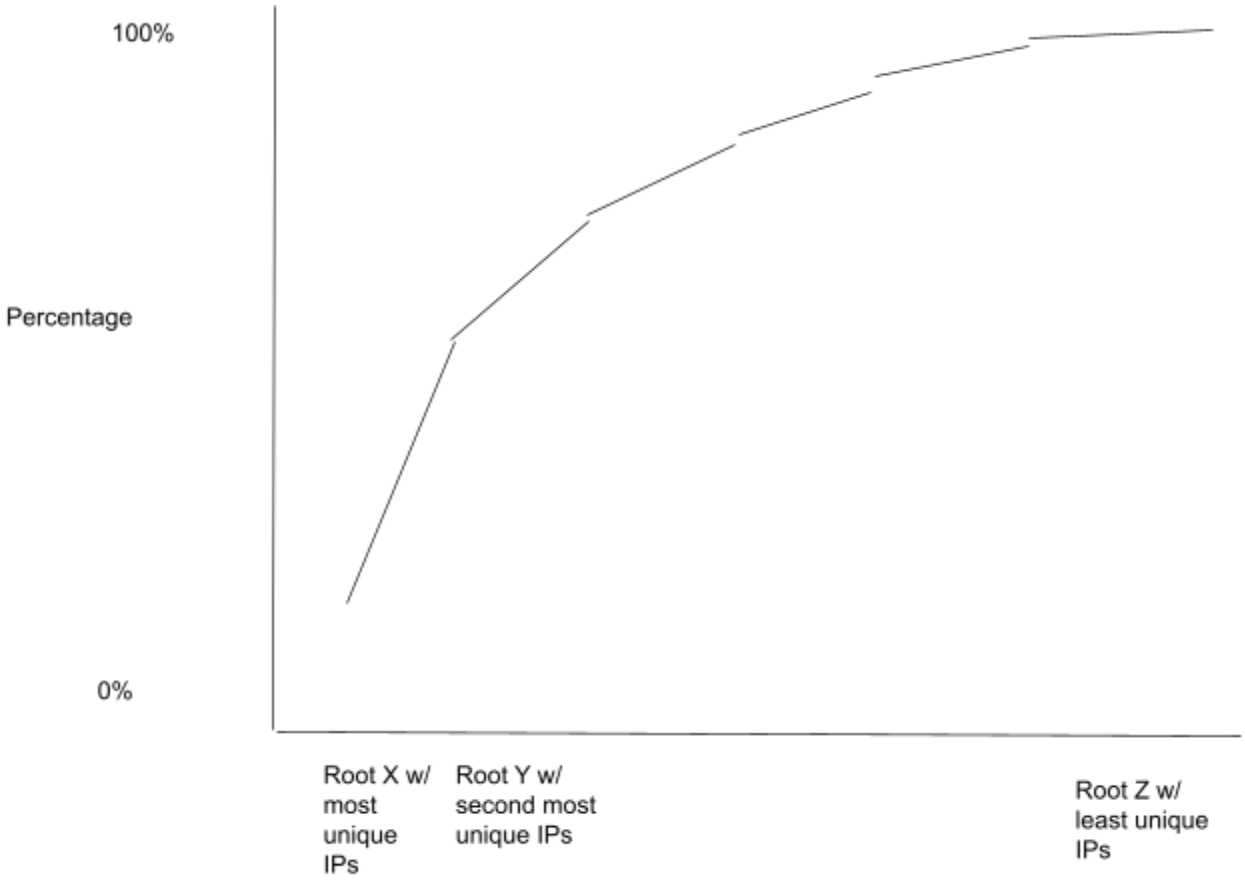
IP Address Overlap / ASN Overlap - Use Jaccard or Cosine Similarity measurements

	A Root	**B Root**	**C Root**	**D Root**	**X Root**	**J Root**
A Root	1.0	.45	.32	.39	.27	.53
B Root		1.0	.29	.59	.55	.29
C Root			1.0	.20	.45	.88
D Root				1.0	.37	.82
X Root					1.0	.76
J Root						1.0

- Some type of Geographical map shading countries by some measure of equivalence or even distribution of traffic per root letter (Gini coefficient).



- Cumulative growth curve of source IP addresses



2. · DNS query properties (label counts through qname minimization)

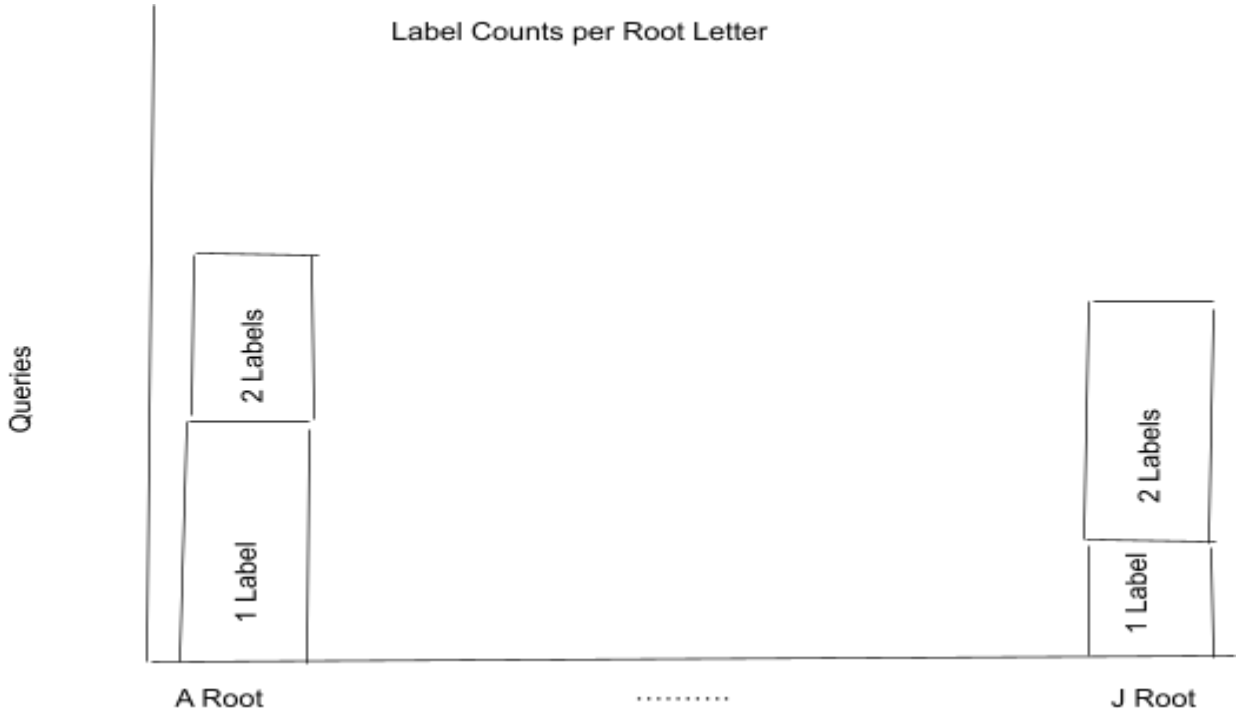
Expected data format from RSO:

- RSO to collect request for NXDomain responses and summarize on a daily basis to produce a delimited text file containing the following measurements:
  - Date
  - Number of queries containing single label
  - Number of queries containing two labels
  - Number of queries containing three labels
  - Number of queries containing four or more labels
  - Root letter

- Example:
  - 2021-03-19,100,200,300,400000,A-Root

Potential Graphs:

- A thirteen-bar graph comparing label distribution rates.



3. · Collision strings based on query volume

Expected data format from RSO:

- RSO to collect request for NXDomain responses and summarize on a daily basis to produce a delimited text file containing the following measurements based on the free version of the MaxMind GeoIP database:
  - Date
  - Rank (e.g. 1, 2, 3, 4, …… N in where N should be 10,000 (TBD))
  - String (i.e. NXDomain TLD)
  - Total number of daily queries
  - Total number of daily queries with only one label
  - Total number of daily queries with only two labels
  - Total number of daily queries with only three labels
  - Total number of daily queries with only four or more labels
  - Number of unique SLDs
  - Number of queries containing WPAD label
  - Number of queries containing ISATAP label
  - Number of queries containing DNS-SD labels (e.g. _tcp or _udp)
  - Number of unique IPv4 addresses

        - Number of unique IPv6 addresses
        - Number of unique IPv4 /24 addresses
        - Number of unique IPv6 /48 addresses
        - Number of unique ASNs
        - Number of unique Countries
        - Root letter

    Note: Calculating the top N and unique number of SLDs / IPs requires a "constellation wide" collection, summarization, and sort. This may be challenging for RSO depending on their data collection and processing capabilities. Sources like DITL may simplify this measurement since the data is already collected into one processing location.

    - Example:
        - 2021-03-19,1,HOME,3000000,200, ….. , A-Root
        - 2021-03-19,2,CORP,2999999,200, ….. , A-Root
        - 2021-03-19,3,INTERNAL,2899999,200, ….. , A-Root

    Potential Graphs:

    - Thirteen column table ranking top N collision strings by query volume.
    - Measure rank correlation (Spearman) for various levels of N (e.g. How do the rank sorted top 10, 100, 1000 strings compare between each root?)
    - Compare consistency of other collected measurements on a per string basis (e.g. IP distribution, SLD distribution, etc.). How much variance is there for the top N strings based on those measures?

Potential investigation into systemic leakage of SLDs under a TLD further focusing on "What additional signal is gained by including more root letters"?

The data collected in #3 above is appropriate to compare TLD strings between root letters. However, understanding the overlap of SLDs or even IP addresses within a particular TLD is not possible due to the data grouping and aggregation/summarization. By extending the data aggregation of #3 to be on a per TLD+SLD+IP basis would allow the following additional data sensitivity questions to be examined:

- Number of unique collision strings
    - Cumulative percentage based on number of unique strings per letter.
- Number of unique SLDs under a collision string
    - A table N rows long by 13 columns wide measuring cumulative growth.
- Number of potentially dangerous queries based on labels (e.g. wpad, isatap, etc.)
    - A table N rows long by 13 columns wide measuring cumulative growth.

Caveat: As seen in previous NCAP A and J root analysis, DNS "events", such as sudden and dramatic name leakages, occur sporadically and non-deterministically. In order to compensate for this behavior, an ideal measure should be conducted against all roots at the same time.

Ideally, the measurement would also persist longitudinally (or be conducted regularly). These constraints limit the ability to conduct these measurements to the DITL data set.

Recursive Resolver Data

Most previous name collision research has utilized root server telemetry data to quantify or study the behavior of the strings. A unique vantage point that has not been well-studied within the name collision context is recursive resolvers, specifically large open recursive resolvers. These resolvers are of interest because they have a more “direct” connection/engagement with the underlying systems requesting these non-delegated strings. From a data sensitivity perspective, one interesting area to explore is the negative cache hit rate at these resolvers.

How does traffic sent from the recursive resolver to the root system differ from the queries it answers out of its negative cache?

- Rate in which caching techniques prevent queries to root (e.g. Aggressive NSEC, NXDomain Cut, etc.)
  - Simple statistics on cache hit rates and query rate to root
- Distribution comparison of SLDs and other labels answered from cache vs. sent to root
  - Measure rank correlation (Spearman) for various levels of N leaking strings
  - Measure rank correlation (Spearman) for various levels of N leaking SLDs
- Distribution comparison of unique source IPs
  - Some form of a fractional comparison of IP diversity seen at root vs. recursive.

Caveat: The ease in which a recursive resolver operator can measure and answer these types of questions is unknown. Also, comparing different recursive resolver data is likely to be a challenge due to specific resolver implementation and deployment choices.