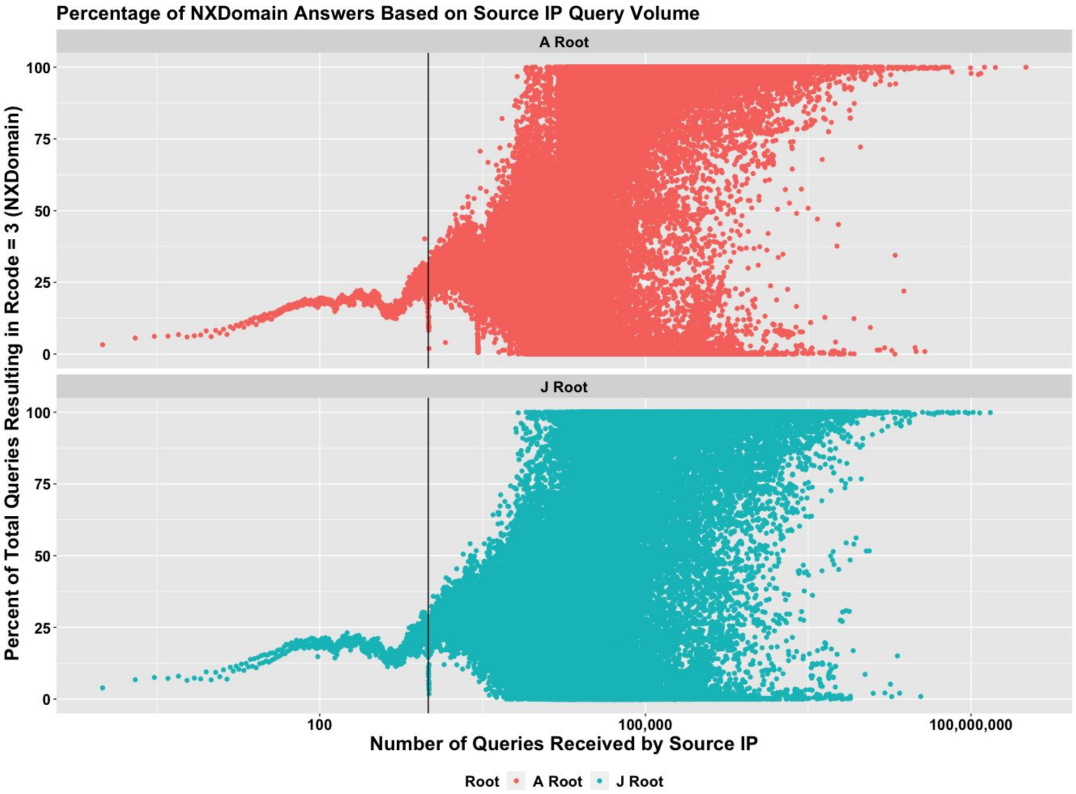
Perspective Study Update

Additional Measurements and New Data

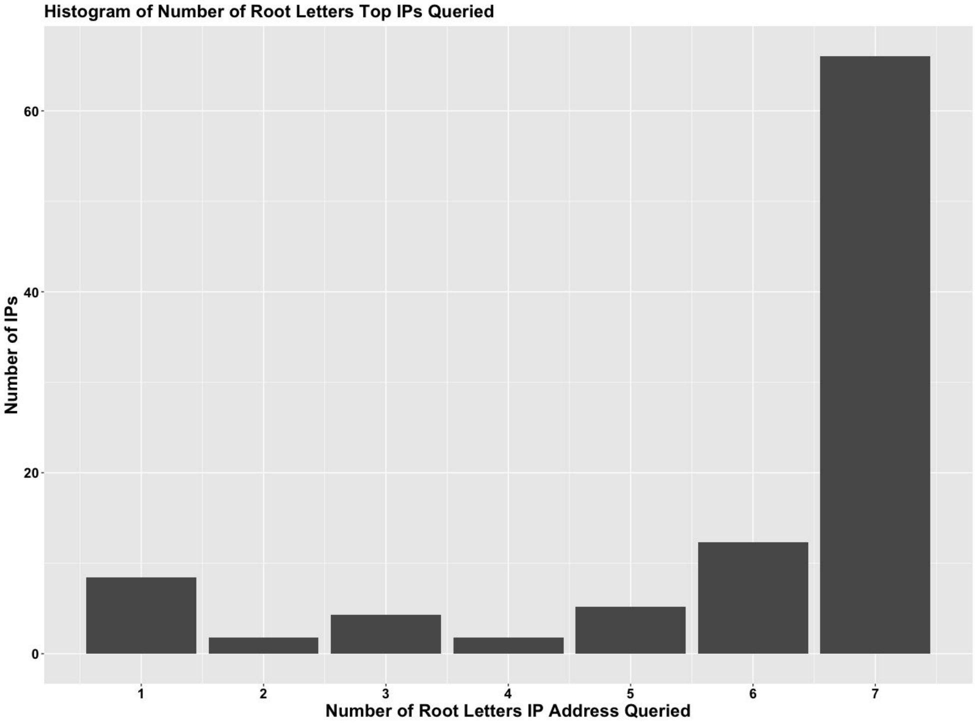
1. High level metadata about ASN distribution of RSS data
2. Use new threshold in addition to total traffic percentage
    a. Regenerate figures and statistics of RSI overlap, geographic, and geospatial distribution
3. Extend TLD overlap between A and J from top 1K to all
    a. Rank comparison using CDM of query volume and source diversity
    b. Measure Jaccard of TLD overlap at various top-N lengths
4. Additional RR data
    a. Rank correlation of PRR and RR to A and J using all TLDs instead of just top 1K
    b. Distribution of rank difference between PRR and RR to RSIs

New threshold for similarity of RSIs



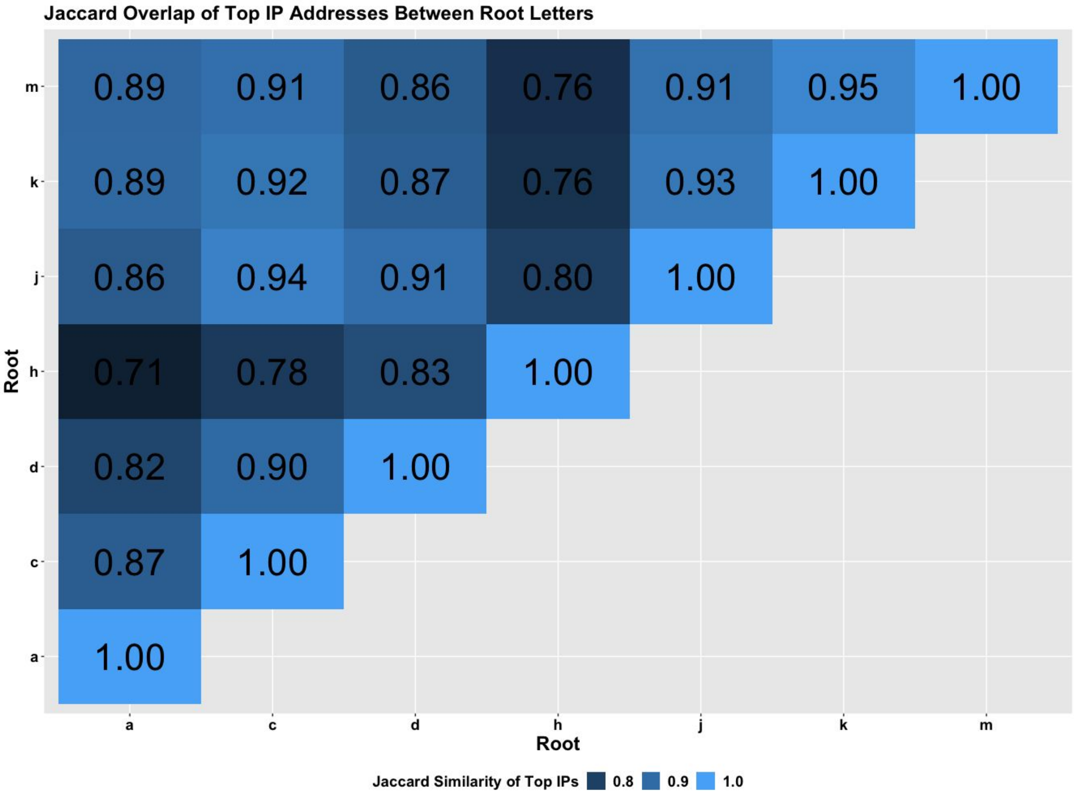
- Previous version used IPs that accounted for 90% of total traffic.
- Appendix 2 showed the behavior of IPs issuing low volume of queries were not behaviorally the same.
- Similarity analysis was done again using a threshold of 1K queries

New threshold for similarity of RSIs



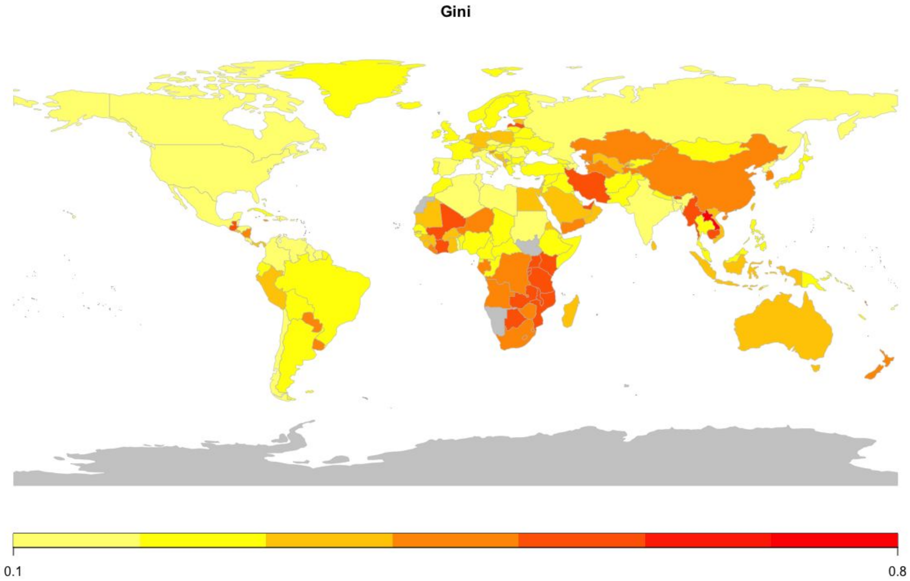
- Figure indicates 66.1% of these IP addresses are seen by all RSIs and 78.1% are seen at 6 or more RSIs
- Analysis on the top 115K IPs and found that 89% of those IPs are seen by all seven of the RSIs.

New threshold for similarity of RSIs



- On average 86% of the IP addresses are observed at any two roots.
- 96% of the top 115K IPs are observed at any two roots.

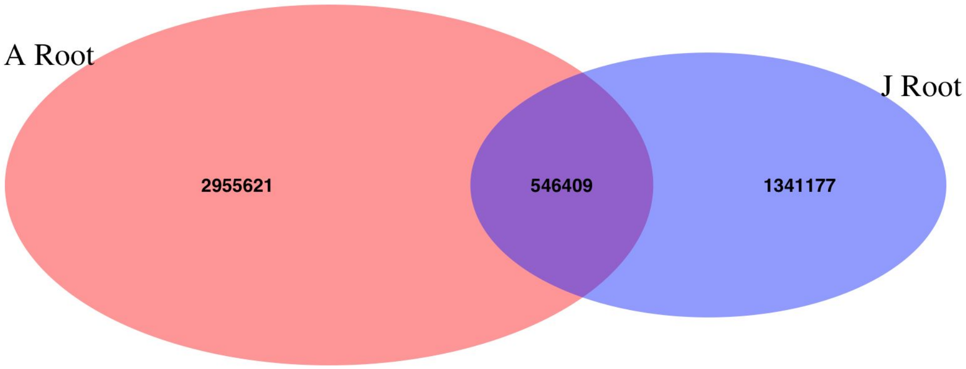
New threshold for similarity of RSIs



- Geographical and Geospatial measurements updated accordingly.

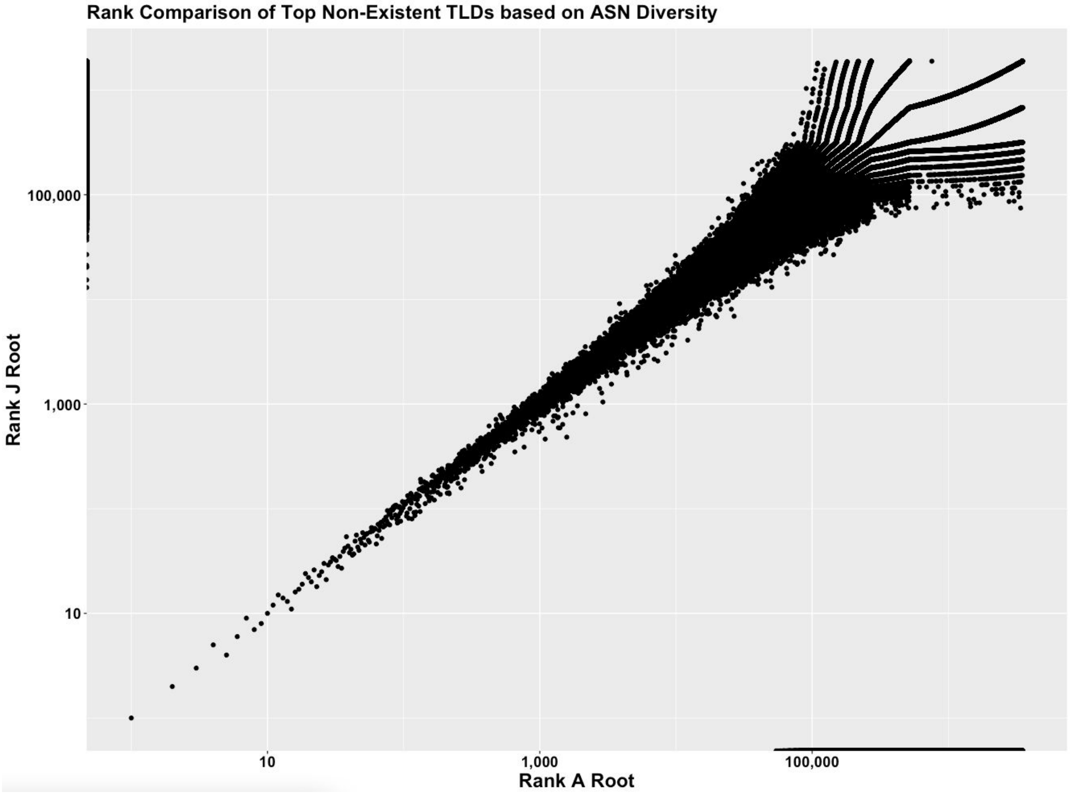
- No significant change in measurements or findings

Extend TLD overlap between A and J from top 1K to all



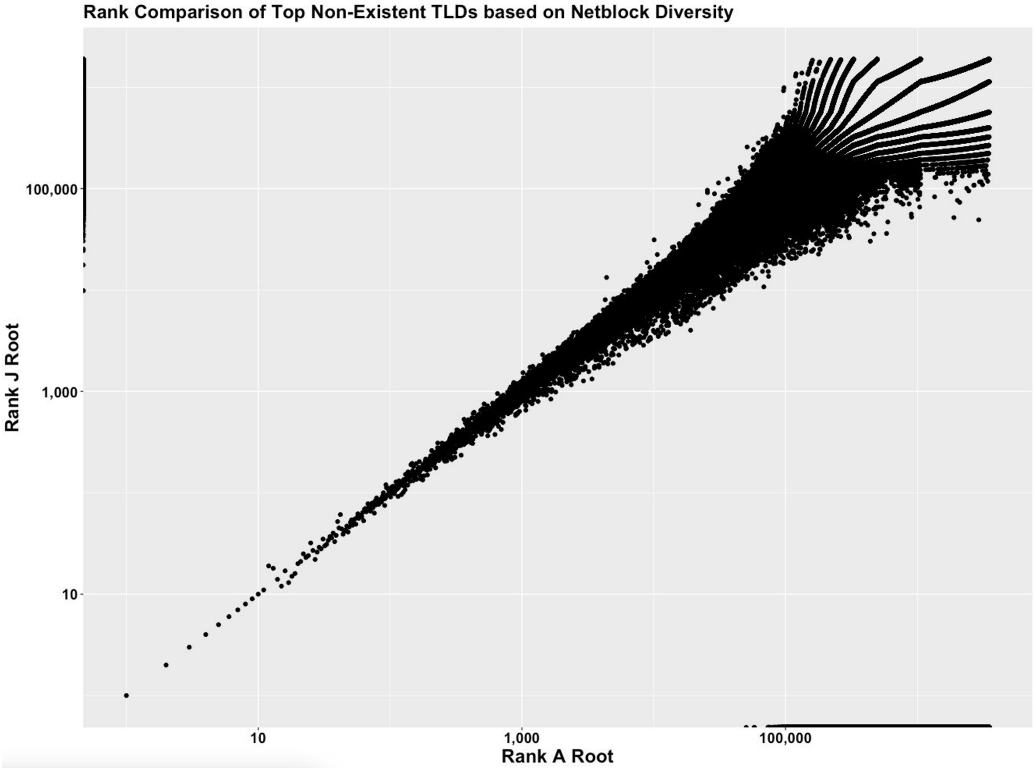
- The entire set of non-existent TLDs were compared at A and J RSIs using the 2020 DITL data matching the regular expression [a-z0-9]{3,63}
- This resulted in 13.9 billion unique non-existent TLDs.
- To remove Chromium queries, a minimum of five queries was required

Extend TLD overlap between A and J from top 1K to all



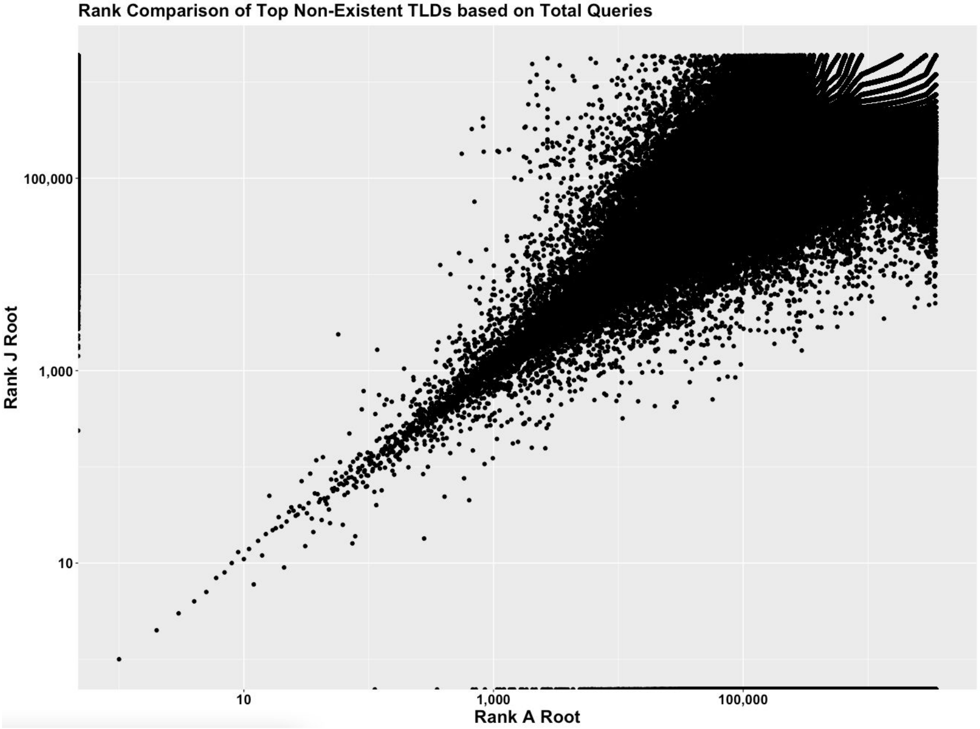
- If a TLD was observed at one RSI but not at the other RSI, a rank value of zero was assigned to that TLD.
- Dots at x=0 or y=0 mean that particular TLD was not seen by the other RSI.
- Very strong rank correlation for the non-existent TLDs up to approximately rank 10K

Extend TLD overlap between A and J from top 1K to all



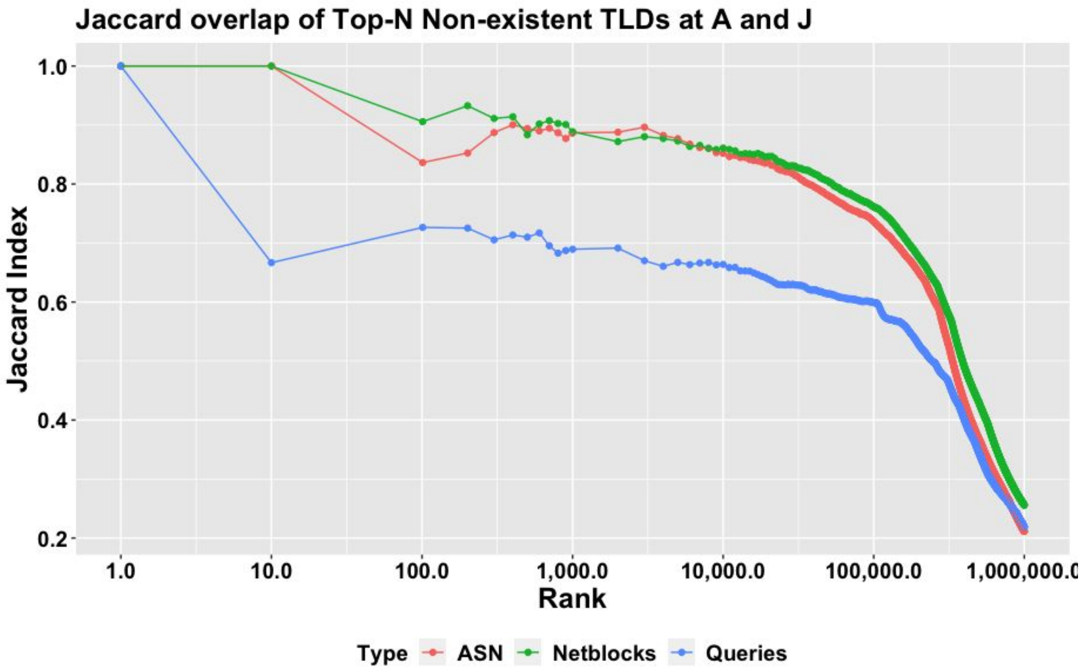
- Non-existent TLD strings observed at only one RSI became more frequent at rank levels above 100K
-

Extend TLD overlap between A and J from top 1K to all



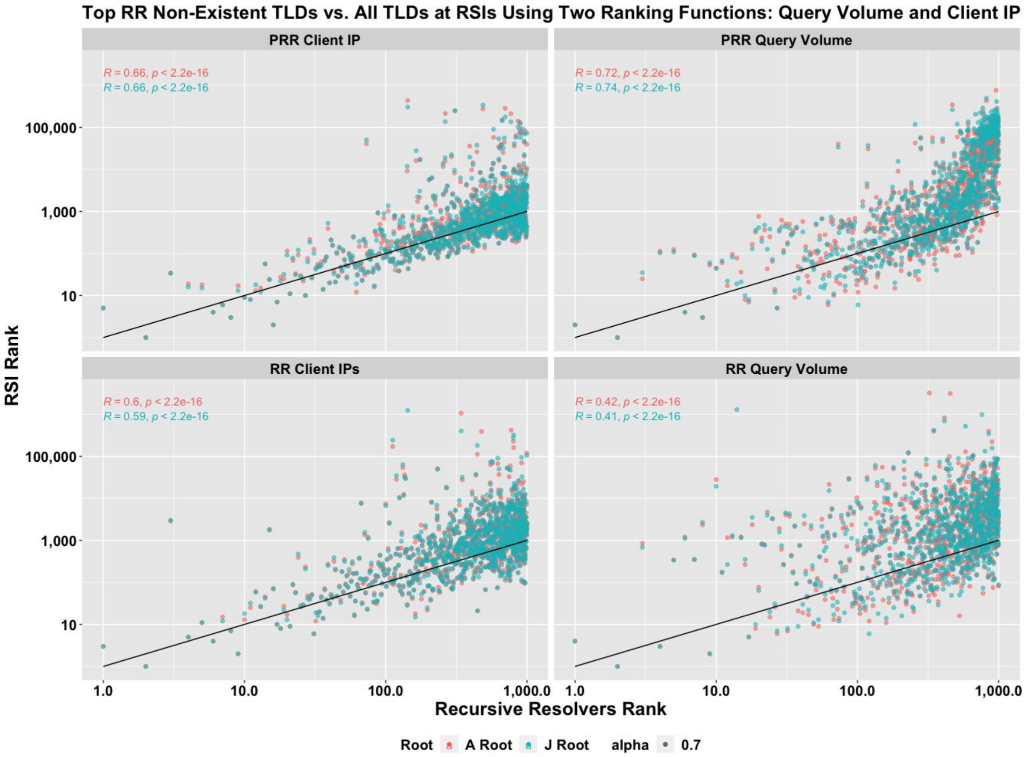
- Query volume also displays a strong correlation for the top non-existent TLDs up to rank 1,000 and non-existent TLD strings only observed at one RSI become more common after that level.

Extend TLD overlap between A and J from top 1K to all



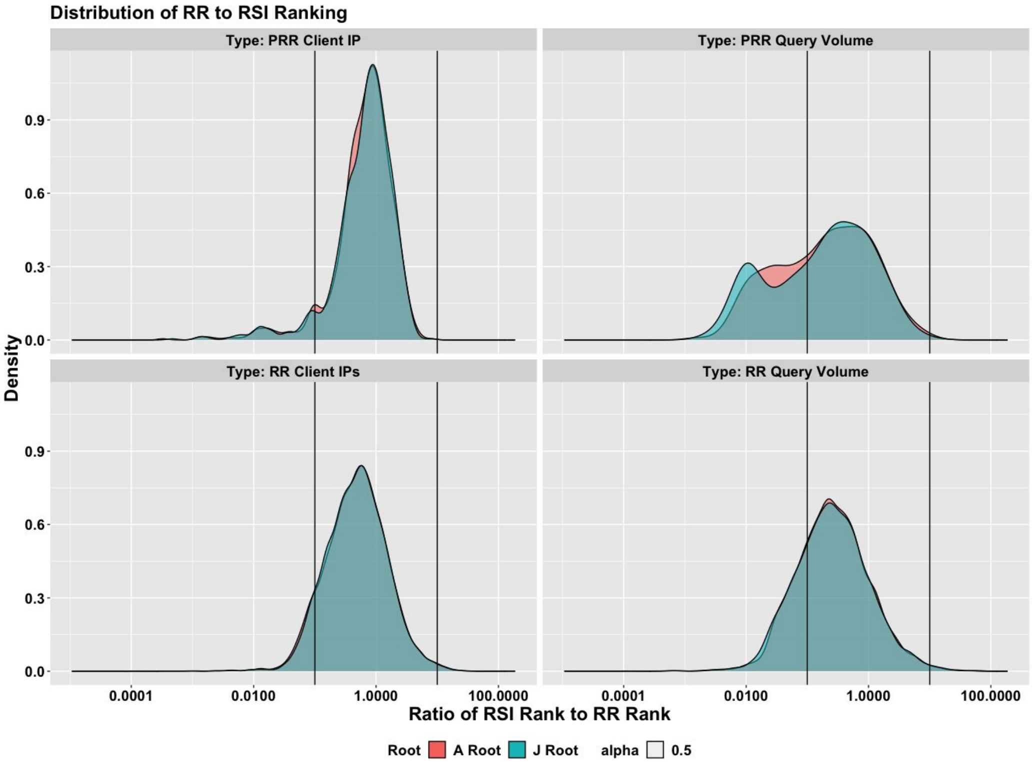
- How similar top-N lists are at various rank depths, Figure 18 shows the Jaccard value of the set similarity between A and J using the three CDM ranking functions.
- Network diversity measurements of netblock and ASNs show roughly 90% overlap until rank level 10K, which the overlap begins to degrade due to the TLDs being observed by just one of the RSI.
- Query volume measurements show 70% overlap until rank level 1K.

Additional RR data



- Query volume and distinct IP addresses, first 100 top non-existent TLD strings roughly correlate between the PRR/RR and RSI.
- However, higher ranking non-existent TLDs exhibit huge discrepancies (several orders of magnitude) between the PRR/RR and RSI ranking.
- From a name collision perspective, this suggests that even if a non-existent TLD has a very high rank based on RSI data, that measurement may not reflect the entire name collision impact posed by that string.

Additional RR data



- Figure shows the distribution of the ratio of rank at the PRR and RR to the rank at RSI.
  - x-rank divided by y-rank, in which an equal ranking would equal 1.
- Most TLDs exhibit +/- 1 magnitude difference.
- Subset of the top 1-K PRR and RR non-existent TLDs that exhibit differences of more than 3+ orders of magnitude.
  - Showing that the top-N at a given PRR or RR can be significantly different than how an RSI may quantify that string.