ICANN NCAP Meeting #41 (2021-02-10) Report Recap

Introduction. This meeting wrapped up the data analysis performed on collision strings of interest with A & J root servers. Matt Thomas provided an overall of the analysis performed to date, suggesting machine learning techniques may be able to help cluster collision strings and benign strings based on analytical features used thus far. Initial discussion centered around the performance of the clustering shown in Matt's analysis. The second half of the meeting evolved into ruminations on how the group can get to reasonable answers to board questions from here.

Data Analysis Wrap-up. Matt categorized the data analysis into three groups: Traffic Properties (e.g. query volume), Qname and Labels (e.g. distribution of queries over SLDs), and Other Attributes. These formed the basis of 32 distinct features as inputs to a clustering analysis. A brief overview of the clustering algorithms and parameters was presented followed by questions about the performance and behavior of the clustering. One particularly pertinent question was whether the analysis could be performed on delegated and undelegated data to compare, especially for those names that went from undelegated to delegated. One answer is that an analysis of historic data could be performed using the DNS-OARC DITL data.

> *[Editor's Note: DITL data is pcap-based query data from nearly all roots and a few other participating DNS operators. It captures data for a 48-hour period every year. Data goes back a number of years, but there have been data availability issues from time to time. Furthermore, DITL data must be analyzed on DNS-OARC systems. Perhaps the*

> *biggest drawback, noted in the discussion is that DITL data covers a very short period of time each year.*

Board Questions. The remainder of the meeting was spent reviewing the ICANN board questions. The group struggled to come up with answers or even clear ways to arrive at answers to the board questions. The group felt the data analysis was interesting and useful, but it is unclear how to go from the analysis to recommendations. It was suggested that collision mitigation can be most easily performed when a collision is perpetrated by a single vendor or operator such as the case with Chromium.

> *[Editor's Note: Mitigation may not always be so easy even when a specific operator or vendor is identified. For example, there may be a well known consumer device behind a collision that once deployed cannot be easily recalled or fixed. It may also be the case that a collision occurs after delegation, which was never discussed, but may be just as severe.]*

One idea the group has been considering if the analysis performed here could be requested from other sources. This might help verify the data and signals that can be made available. There has also been questions that have come up about the definition of a name collision. Thus ar the standard definition from Study 1 has been rigidly adhere to, but there may be reasons to tweak to that definition going forward. The specifics about these changes was not discussed. It was also suggested that board questions do not necessarily be answered exactly as opposed, but the answers must provide guidance to the board.

Raised in the last few minutes of the Q&A is the notion of “risk” and “harm”. Risk may always be present, but seems very unclear what the actual harm is and how to predict it. The meeting ended with the notion that collision distribution may be biased by the group’s examination of query volume versus what the actual risk and harm may be.

References

ICANN NCAP Meeting #41 page

DNS-OARC DITL Traces and Analysis