Name Collision Analysis Project Discussion Group Meeting Notes

15 May 2019 | 21:00-22:00 UTC

Attendance
Members: *Jay Daley, James Galvin, Julie Hammer, Rod Rasmussen, Ram Mohan, Chris Roosenraad, Dmitry Belyavskiy, Steve Crocker, Eric Osterweil, Anne Aikman-Scalese, Matthew Thomas*
Observers: *Jim Prendergast*
Apologies: *Jaap Akerhuis, Merike Kaeo, Matt Larson, Danny McPherson, Jeff Newman, Steve Sheng*

Decision:

Action Items from this Meeting

- Jay to document "push forward things"

Summary Notes
Call to Order
Kim called the meeting to order at 21:00.

Update to SOIs
No updates to existing members' SOIs

New Members
No new members

Final review of the Study 1 document and changes made

Jay reviewed the Study 1 document edits/changes [document will be sent to OCTO for procurement].
Task 1 has been finalized to "definition of name collision". Changes include what are in scope and subject of data studies; things in scope but will be addressed with general advice (not data studies) and what are out of scope.
Task 2 was amended to include the word "informal"; if it didn't say informal, then a public comment would be required.
Task 3 includes criteria and specific work, which as been discussed previously.
Task 4 - word "solely" was removed. Written report should provide an explanation of the issue, summarizes known (evidenced) harm of name collisions; list all the relevant previous work (using criteria listed in task 3); technical impacts of those litigations only;

important points that should be brought forward (including but not limited to questions about the data use methodology, technical gaps, competitive/opposing recommendations. Task 6 identify data sets used in past studies.

James indicated agreement with a comment in chat [from Ram] regarding
#2 - and the word “finalize”. The group should not restrictive and not be able to revisit after studies 2 and 3.

Eric he does not think the group should worry too much about “down the road” (the definition as a theory may always be held up by scrutiny).

[in chat] Rubens “*my point is not to mention DITL, but to mention that at some point studies used DITL 2012, and the datasets might be newer editions of those*”

[in chat] Steve “*Have we included the SSAC work in 2003 related to Sitefinder? The briefings in early October 2003 included some very specific effects that i think are relevant in this study*”

In response from Jim, the group does not have to be complete as part of study 1.

Steve added Sitefinder should be included (data and demonstrations of the kind of things that go wrong).

Under #3 - add Analysis of the impact of SiteFinder that meets the criteria above.

[in chat] Ram “*let’s be careful to define “informal” public consultation; ICANN has certain norms for public consultation and I would not want our informal process to be pointed out as a reason to doubt the outcome*”

Jay responded that this will be managed by OCTO, within the correct ICANN framework.

Jim stated even if the group puts in the work “informal” - public consultation has a specific meaning within ICANN. The group needs to be more deliberate in the wording.

Jay suggested “public consultation as defined by OCTO” removing the word “informal”.

Planning for Study 2

- ***Anonymization/redaction of data***

Jay reviewed, as part of the project plan, a third part will be commissioned to develop a tool that takes raw DNS data and anonymize it (and what the data will look like after). He added complete anonymization of source traffic.

Steve - what do we want accomplished - way of protecting the privacy of the data and still have it useful? He suggest treating the anonymized data as sensitive as the “real stuff”.

Jay stated, the anonymized data should not be published - but there should be a process or methodology - which is were reproducibility is introduced.

Eric noted if someone is going to come up with a set of analysis that they find useful for Study 2, at that point they figure out what the sensitive data is and how to anonymize it. One could come up with an anonymization technique after the analysis is done that preserves the fidelity. The group could say, the contractor is going to get access to sensitive data and agree on the proper way of anonymize it while having reproducibility.

Jay - the reason for this being looked at is some data providers stated they are not going to provide raw data, but will if there is a process for redacting the data. It might be useful in Study 2, for someone to take readily available data, begin to plan form of analysis and understand anonymization redaction required.

Steve - access to the original data (not necessarily by the DG or contractor) by the people who provide the data may have to provide it again under some agreed upon new anonymization process.

Rubens [via chat] - I would imagine, changing IP addresses to ASN if ASN is ISP but replacing it by <end user> for all end user ASNs

Jay - effectively locality preservation; end user query isn’t changed, we don’t need all the labels of the query - just the collision labels of a query.

Matthew - countered, there are times where the contextual awareness of all labels, and not just collision TLDs, provides avenues for additional research; important to keep all labels.

Jay - redaction of labels; is redaction of private non collision labels; but still keep standard labels that help identify the product or system that may be generating the collision.

Eric - regarding search list process, difficult to know which label to look at; not sure how many labels to remove depending on how long the list is, etc.. Might be something to figure out before knowing the analysis.

Matthew - include timestamp in the contextual awareness and where the query was destined for.

Jay - study 2 will require the contractor to do some pre-analysis of existing datasets (or datasets that are readily available) then provide some thoughts on anonymization in order to get some of the more difficult to acquire datasets.

- ***Reproducibility***

Jay - for reproducibility, at the end of the project, the group will provide for any third party researcher the ability to reproduce the research that has been undertaken by the contractor using the same datasets. This has gone to the Board and it was stated that there is no guarantee, it’s something being targeted.

Steve - challenge wanting to make it reproducible while protecting the privacy of the data.

Jay - discussed in SSAC, a panel that has specific proposals for reproducibility and those are agreed as one off terms with one off sets of contracts with those who wish to do the reproducibility.

Steve - you would want those people who has access to data to adhere to the same rules of protection the contractors are asked to adhere to.

Jay - if the data providers have a specific concern, should they be allowed to veto individual reproducibility proposals?

Steve - conflict of interest concerns. I allowing data providers to veto, could raise the question of biasing the results. Data providers wishes should be considered but not give them absolute control.

Jay - for consideration, If [we] are unable to provide reproducibility as a result of the level of protection - what impact would that have on people accepting the results of this project? Any mitigations that need to be put into place?

Eric - what is the evidence behind the exception? If the data is not transparent and someone questions it, they should offer reasons why [and not just calling it out].

Range of datasets required

Jay - open for conversation - is there a specific (range) set of datasets that should be considered?

Matthew - root data, and data available through the DNS-OARC. Root servers, the more the better.

Jay - we want to use multiple roots, as many roots as we can?

Dmitry - data from public resolvers. Important to get data from Cloudflare

Rod - open DNS, Google, etc. - and ISPs that would be willing to share data. "Large caching resolvers" covers all those. Consumer ISPs have large datasets. Customer base of the ISP will have data skewed toward either home SMB user vs enterprise level type transactions. Providers like Infoblox has customer data available because of the work it does for them. Additionally, DNS firewall products.

Matthew - longitudinal data -

Any Other Business

Jim - currently the group meets weekly. No urgent week, so not meeting during the week in Marrakech. Will not meet July 3rd. Project plan does call for a F2F meeting, day before the official start of the ICANN meeting. Plan to have a full day meeting in Montreal, in front of the ICANN meeting.

Jay - ask for agenda/discussion topics for next meeting. Will evaluate frequency of meeting.

Next Meeting

The next meeting will be on 22 May @ 21:00 UTC

Adjournment

The NCAP Discussion Group concluded its meeting without objections.

Recordings and Transcripts