KIM CARLSON: Hi all. Welcome to today's NCAP Discussion Group call on the 20th of January at 19:00 UTC. In the interest of time there will be no roll call; attendance will be taken based on those on Zoom and Kathy and I will update the wiki list with the names as quickly as possible. We have one apology from Anne Aikman Scalese. So, reminder, calls are recorded and transcribed. The recording and transcripts will be published on the public wiki. Also, as a reminder, to avoid background noise while others are speaking please mute your phones and microphones. And with that I'll turn the call over to you, Matt.

MATT THOMAS: Thanks for that, Kim. Welcome, everyone, to the NCAP Discussion Group call. Jim, I see your hand's up already. Do you want to jump in first here and say something?

JIM GALVIN: Yeah, sorry, it's part of the roll call. I realized that Patrick told us but I don't think that Kathy or Kim would have seen it. He may not make today's meeting. So, I just wanted to get his apologies in the record.

KIM CARLSON: Thanks, Jim.

MATT THOMAS: Thanks, Jim. Yes. Welcome, everyone, to the weekly NCAP discussion call. IIt's hard to believe it's Wednesday again already. Today's agenda

has been posted up on the slide. I think we might do a little slight reordering of how we go through this. So, I think instead of having the JAS Refresher after the .internal Case Review, Jeff Schmidt has generously offered to go over the efforts that they did before as a reminder. So, right after maybe item number four, we'll stick Jeff in there and then if we have any time we'll start going into the .internal Case Study Review.

With that being said, does anyone at this time have any updates to their SOI that they would like to declare or share? Not seeing any hands. So, we'll take that as a no.

Moving on to item three, the update on Study 2. Jim, would you mind just giving us an update since last week on how things have moved along with Study 2 and where we're at going forward, please?

JIM GALVIN: Sure, that sounds good, Matt. Thanks. So, we've had one significant comment inside of SSAC with respect to the proposal we need to clarify. It's suggested that we add to our revised proposal a clarification to the way statements of interest are handled with this project.

The original proposal in section 3.2 had a rather lengthy conflict of interest description, but that really was predicated on the fact that we didn't have a project manager. We were really going to do all the project management and project sponsor work ourselves. But as we've already done in Study 1 and we're going to do again in Study 2, we have a project sponsor now, and that's OCTO itself.

So, that really moves all of those issues squarely onto OCTO and so, more like any other ordinary working group PDP process and such in ICANN, we just have a statement of interest and we generally expect that the balance of interest in the group take care of things.

So, it was suggested that we should clarify that and make all that clear. So, we're putting together some text to handle all that and we'll certainly make that visible here. It's going to have to get ICANN legal review among other things, but we're going to have to work that process a bit, and expose that to you folks here too, for your review. And then we'll be able to move on with the project.

The risk is that given that next Thursday was—not tomorrow, next week—was the BTC meeting, there is the potential for not being able to make that particular board technical committee agenda, as we put this together, because of the reviews that are required to move this text forward, since no one has seen it yet. That just means that the hiring of help, the potential for that gets pushed back at least a month until the next BTC meeting. But we're going to continue our work, nonetheless. It does not slow down the analysis and work that we're doing, it just slows down the technical writer support and the beginning of the root cause analysis that we hope to accomplish. So that part, the start of that, gets delayed a little bit but we're just going to pick up and continue forward. I hope that wasn't too much, sorry. Thanks.

MATT THOMAS: That was perfect, Jim. Thanks for the update. I don't see any hands at the current time, so I assume everyone's okay with that update. Why

don't we just continue forward into our agenda, item number four, the Name Collision Outreach Efforts. I just wanted to bring this up for a quick minute or two to discuss a publication that I put up on the Verisign blog that I think is relevant to the discussion group. I pasted the link in the chatroom, if you can see it or want to read it in there. But since earlier in 2020 Verisign, specifically I've really started off a name collision outreach endeavor, looking to identify and remediate high query volume strings. I know several of the people in this discussion group were actually participants in that and if you feel like you would like to share any other insights or ad hoc stories or commentary around that, please feel free, if you do so wish.

I just wanted to, at the highest level, maybe give a little bit more technical detail in terms of what actually occurred in this program, in this initiative. Because in the blog it really just talks about going after high-affinity, high-volume strings, and I just wanted to explain a little bit more for that.

What we did with A and J data is we rank prioritized the top leaking TLD strings by two factors, the first of which is the number of queries coming out of it. But the second of which is, if you were to take a look at the cumulative distribution function of the number of ASNs for that particular string, if you looked at what percentile the top three ASNs would be at, we used that as a gating mechanism to be able to identify high-query, highly-concentrated strings at particular vendors. Based off of that, a little bit more data analysis specifically looking at the key names and the IP addresses, we could usually identify fairly accurately what kinds of systems were causing those queries to be coming out.

Once we had that information, it was really just a matter of your standard outreach initiatives.

We looked at our rolodexes of contacts either we have in the ICANN community, or various security trust groups, or other organizations to see if any of those entities that were leaking these strings had direct contacts on there. Specifically, the GNSO rolodex was very helpful for a lot of this. And then it was just a matter of simply sending a few emails to most of these people and I have to say that the community at-large is super receptive to this. I found everyone that we reached out to be very open to working with this and wanted to fix the problems. And shortly after communicating the data, we saw many of those problems disappear.

But to that end, some of those problems are a little bit more nuance to fix, some of the strings that we also identified were things that were associated with home networking equipment, things like .router or whatnot. But some of whom were specifically tied to vendors like .dlink or .zyxel or stuff like that. So those, they were still great to work with and they acknowledged the problem, they've understood the root cause of it many times. But even they have said that the deployment of fixing suffix search lists, or name appendages, or whatever the underlying cause was, in equipment-based problem is much more time consuming and a longer scale than something where it's a closed environment.

I'm sorry, just catching up on … Matt if you could expand on those categories about prefetching bug. Yeah. So Rubens' question about prefetching bugs. Several of the common causes that we identified with

the various strings were due to suffix search lists, where the string was applied in a suffix search list context and they were just being appended on all of the queries. Actually, we might have an example of some of that today in our .internal presentation if we get to that later, as well.

But I just wanted to share this outreach effort that we've been doing, and I think that it will hopefully, provide some context going forward, how outreach is very effective. But I think there are certain challenges especially when we look at some of these higher query volume strings—not even just query volume—but wider source diversity and wider causes of leaking queries for particular strings that make it more difficult to remediate or completely remove the risk with doing so.

The ones that we've gone after so far were very clear-cut. It was a square hole, and we had a box to put in it and it's the exact same size. It worked extremely well. I think as you go into the longer tale of figuring out how to address some of these name collisions, it's where we're going to start to understand and see that there's not always one simple answer or solution to this, that it might be extremely complicated. And that even if you do identify a string that has 85% of its queries coming from one entity and you get that entity to be able to remediate it, if the remaining 15 is still spread out over a super wide, diverse set of networks or are caused by a lot of different underlying software or DNS behaviors, I think it's going to become a little bit more difficult for us to say that outreach will ultimately be successful in those endeavors and that's where the risk assessment will probably change.

I don't want to take too much more time, but I thought that was useful for the group. There will be more coming out with our efforts on that

shortly. There's several other strings that we've done in the last several months that are high impact that I'm hoping to share with the group as case studies as well.

With that, Jeff Schmidt has been again super gracious and said that he would be willing to give us a refresher on JAS's efforts in terms of how they looked at name collisions in the past and how they looked at assessing risk. So, without any further questions at this point, I'd like to maybe turn it over to Jeff and let him have a little talk with the group.

JEFF SCHMIDT: Awesome. Thanks, Matt. Appreciate it. Can I share my screen here? I don't know if I can do that?

KIM CARLSON: Yeah. One moment. I'll promote you.

JEFF SCHMIDT: Okay. Thank you. So, Matt, I appreciate that. And as Matt said, what I wanted to do here was just a real quick refresher on some of the analysis work that we did now five years ago. Our final report was 3,000 and some odd pages long. The vast majority of that was what we called a horizontal study and a vertical study of DNS-OARC data at the time. And it led us to the observations and the conclusions that we came to. I thought there was some good work and some interesting approaches in there that I just didn't want to get lost in the 3,000 pages as we were all reconsidering these problems. Oh. Perfect. Thank you. There we are.

So, forgive me. This is five-year-old data and five-year-old memory. Probably the five-year-old memory is the bigger issue on my side. So, I might struggle to remember super specifics but I wanted to give you the overview version here.

And I really appreciate what Matt and Verisign are doing with the outreach. We did a lot of that too and it is helpful. Some people are more receptive than others. We reached out to over 200 firms, one way or another, when we figured something out. And we had meaningful engagement with about 20, that I'd say resulted in a direct improvement. And when I say engagement, for the record here, we never charged anybody a dollar for anything. This was all pro bono. I don't give any of the names of the firms that we engaged with, but they ranged from a department of the US government to large multinationals and everybody in between. So, you might get the gist of some of them based on what you're seeing here.

So, if you look at our report, there were two material appendices. And this report, I'll stick a link or I'm sure somebody will stick a link to it in the comments or in the chat window, but this is our public report from 2015. The horizontal study is in an effort to understand strings and query types horizontally across all name servers. This was based on DNS-OARC DITL data sets, so obviously that's again an important aspect of the data to keep in mind. It's different than say the data that Verisign is looking at which is live root server data. We're looking at these daily snapshots over, in 2015, five or six years at the time.

The horizontal study was across all queries that wound up in a top-level domain that was requested for in delegation or requested for

application during ICANN's process. And then the vertical study, which we'll talk about here in a second, was drilling in on a TLD-by-TLD basis.

So, we spent a lot of time staring at strings, as a lot of people do that do DNS research. And when you're looking at a gigantic pile of strings, it's kind of hard to pick out patterns. And so, we worked with a data science firm that we have a good relationship with, to take that pile of strings and among other things, reverse engineer out of that pile of strings, regular expressions. For all of us computer science nerds, staring at regular expressions and particularly visual examples of regular expressions are sometimes really helpful to understand what's going on underneath.

So, this whole section … And there's some description of what we did and the processes and all that. But I think this a pretty cool part. Again, unfortunately, it's buried in page 3,000 of this thing. But for the most representative regular expressions in the data set horizontally—so across all the applied for the TLDs—we reverse engineered a regular expression—again, obviously, this is something only a computer scientist can love—but then most importantly generated a visualization where you can really see things that are buried in the data. It's hard to see from this that compatibility add-ons and there's an org and there's some other features here that are worth looking at and worth trying to get an understanding of.

For each of these, we also showed some summary statistic. So, this regular expression accounted for 12% of the data set. Here it is in a breakdown by both query type protocol, and we have name server in some of the other tables as well. And then a visualization, where again

it's really hard to pick out, when you're looking at something like this, what's happening. But you can see something's worth looking into here. There's a [nouveau] there, there's an American down there. And so, given some hints like this we started digging in a little bit more.

Let me call out a couple of examples to show you. My intention isn't to go through everything here in excruciating detail. The Planet is a pattern that appears quite a bit in the data set. That is a legacy naming scheme from an ISP. We did contact them. They were medium responsive. A couple different variations.

Things that start with com. Obviously, you would expect Fritz. That's a German TiVo-like appliance or ISP or something. We did reach out to them and they were responsive. A couple different variations of Fritz.

These patterns now start to get a little bit more interesting. So, this is a pattern that starts with D-Root, and you can see it in the visualizations. There's a couple different varieties there. But now you get to really interesting things, AD root. Strings that start with AD root and have some stuff at the end account for 1.3% of all of the queue names that we saw. Compatibility add-ons, something else that sticks out here.

So then, that gave us some inspiration to start looking into what kind of software generates queries that start with AD root? What kind of software generates a query that starts with compatibility additions or add-ons or whatever? That was there. And you can see, there's other kind of hints here that did lead us to, among other things, figuring out that Microsoft active directory was a big part of what was going on. So, MSDCS, queries that start with MSDCS comprise almost 1% of data set.

These were coming from root servers, across the board, CMAL, etc. And so, it gave us a hint on what to look for.

And so, then we would Splunk through the data sets in other ways to try to understand what was going on. And at the time, I'll tell everybody here, that looking at these regular expressions, and in particular this top ten list if you will, where there were some strings that we thought were associated with active directory, that's how we found the corp issue and the associated issues and that's what got us that thread to pull on. There's a lot of these MSDCS, ADO, etc.

A couple of other things. Clearly there's a EURO starting here. There was another pattern. I think it was Triton or something like that. Sorry to be scrolly here in front of everybody. Some of these are company identifiers and company strings. Here we go. In this case, Triton EU. So that was a dot that was dropped. Triton.EU is an engineering firm that had a misconfiguration. We reached out to them, they fixed their issue, and it was located through an analysis like this. There's a couple of versions about Triton Euro M1. These are all their naming scheme, that become very, very clear when you look at an analysis like this. So, I think you guys are getting the gist there.

The vertical study, same sort of thing. And again, this is in appendix B. We've got the description. I do want to call out this GitHub. That code base is still there. We published all of our code that we used to go through the DNS-OARC data sets. So, yeah. So, that's still out there and we also wrote up a couple of pages within the DNS-OARC wiki that described what we were doing. So, we want to make sure that that's maintained in a contribution to the science.

What we looked at here was, a, there's a lengthy description. But the general gist here is we looked at SLD diversity, or second-level label diversity, and source IP or query IP diversity to understand whether querying a particular set of second-level labels was localized to a small set of sources or whether it was a larger phenomenon. And so, if you look at the entire data set, this is a cumulative distribution. So 32% of the data set at the time was random 10s.

The IN-CATS, these are Interisle categories for compatibility with the Interisle reports. We used the same classification. Invalid, random 13, random 10s, short SLD—those are our things that we added. And then you can see the cumulative distribution here, of how much of the data set is accounted for with those particular strings. And so, then we go through and do a binge scatter plot, looking at SLD diversity of the second-level labels for each one of the applied for top level labels.

So, I'll just go into corp here as an example. These second-level domains represented 93.4% of all the labels that we saw for corp or that ended in .corp in the data set, in the DNS-OARC data sets. That's the numeric description, tabular description obviously. All of these companies we reached out to. Again, we had varying levels of responsiveness, but we did reach out. Basically, every time we encountered something that we thought we could chase down and identify to an entity, we did, to the tune of 200 and some odd over the course of our engagements.

And then looking at the SLD, or the second level, the label diversity as it relates to the querying IP diversity. And we did that for every single one of the applied-for strings. Okay. So, let me pause there. I see a lot in the chat, which I have not been addressing or looking at. So let me pause.

MATT THOMAS: Jeff, that was very helpful. I have a couple questions for you, if you don't mind. In your horizontal study, was that taking a look at the full queue name or were breaking it down by label or just looking at the first label?

JEFF SCHMIDT: I'm pausing here for a second because I know we looked at both. I believe what is in this report is just the next higher-level label. I'll respond back to the list on that just to make sure. I know we looked at both and we applied this to both. My five-year memory here is a little bit fuzzy on exactly what we wound up plotting here.

MATT THOMAS: No worries. My five-minute memory is awful. The other question I had was for you, both of these studies are obviously relying on the labels to provide some additional context, but as we saw last week on .mail, where upwards of 62% or 63% of it is effectively coming in queue name minimized. Now we're only seeing the TLD string. What are your thoughts on how that impacts our analysis and how we go forward and approach this and understand the underlying risks and causes of these?

JEFF SCHMIDT: Yeah. So, queue name minimization wasn't really a thing back in 2015. And then I think we lose a lot of data when the higher-level parts of the queue name get dropped. We relied obviously on the higher-level parts of the name to tell us a lot about what was going on. If and when we lose that, that will not be helpful.

MATT THOMAS: Do you have any thoughts on what would be the most impactful secondary criteria to look at then, if the labels are impaired?

JEFF SCHMIDT: Yeah. Let me answer it a little bit differently. When we made the recommendations about corp, home, and mail, first of all obviously we didn't take that lightly. We understood the magnitude of what was going on there.

And so, the criteria that we looked at to figure "harm" if you will—and I'm doing air quotes here which you can't see—were first of all how broadly the labels were used. And it wasn't just the number of queries we saw for some unit of time or whatever. But we also considered diversity of the querying IP addresses, diversity of the querying ASs, the diversity of the receiving roots, and time diversity, to the extent that we could get it from DITL data. I think when you're trying to figure out what's going on and you don't have the higher-level labels, we're going to have to rely more on diversity of where the queries are coming from and other things.

We looked quite a bit at the published examples and other things that, for lack of a better word, told people that it was okay to use these strings. So we were certainly colored by, in the case of corp, Microsoft basically saying, "Do this." In the case of corp and home, they're effectively blessed by RC6762 or something like that, where they had that list that includes intranet and private and corp, home, LAN, etc.

Then mail doesn't have a huge diversity of second-level and higher labels. A lot of what's coming into mail are ANMX4 for the actual TLD itself. But the reason that we came up with, or a contributing factor that we came up with for that at the time, was this published example send mail configuration script in an O'Reilly book, I think it was, that seemed to be … People using that script or some derivation of that example seemed to match the traffic that we were seeing for mail. So that factored into our thinking.

If somebody, at some point, told some poor administrator that it was okay to do something, we took that as a fairly high indication of protection that we felt some obligation to look out for those poor guys and gals that have been mistakenly informed over the years.

MATT THOMAS: Thank you so much, Jeff. I apologize. I had to step away for a second. The encore of my daughter's ballet doing Frozen dancing and my son screaming is continuing. So I apologize for that. One last question for you Jeff around just general other open-source research that you utilized for looking at strings like .corp. Is there any recommendations, in terms of how you did open-source intel looking for such things as the mail configuration? Is it just literally searching GitHub and doing Google-Fu searches, or do you have any tips you'd like to share on that?

JEFF SCHMIDT: Yeah. A lot of it did start with Google-Fu and just looking around, trying to explain things. When we found somebody that was actually generating the traffic … So there was a firm—I think it was in the public

record many, many years ago—US Renal Care. They were identified by somebody other than us. Before we were engaged, somebody stood up at a public session someplace and referenced them. So we looked through the data and we found them and we reached out, and actually were engaged in a very significant conversation with the right technical folks there. And we actually linked them up with Microsoft later in the process as well.

And that was one of the ways that we … By finding a specimen in the wild, for lack of a better term … By finding a specimen in the wild, we were able to really understand what was going on. There were a couple of other instances where we were able to find a specimen in the wild. I think I mentioned during the call last week that we found an issue with the way that the US DOD had some border devices configured. We found the evidence in logs, reached out, got into a good discussion with the right people, and worked with them to understand what the underlying cause was, how this happened in the wild. Had a specimen and then worked with them to fix it.

Aside from the vendor issues, we've obviously reported the series of vendor issues to Microsoft and that's well-known at this point. I will say there were three other vendor issues that we haven't said anything about publicly but we reported and went through responsible disclosure and material things were fixed which were vendor whoopsies that we found by finding examples.

MATT THOMAS: Thanks for that, Jeff. Warren I see your hand's up, please go ahead.

WARREN KUMARI: Thank you. So this isn't so much of a question as more just a comment. It is nice to see that these are actually mitigatable, if one puts in the effort and chases down the source. What's nice is for some of these, one can actually see them on the aggregate root traffic or similar graphs. Both these and also Matt has been doing an incredibly good job at hunting down some of the worst root leakers and helping mitigate those. So I think this is more to say thank you for helping make the problem a bit less bad.

MATT THOMAS: Thanks for that, Warren. Appreciate that.

JEFF SCHMIDT: Cool. Thank you, and I'll echo that. I mean, the work, Matt that you and Verisign are doing, as evidenced in this last blog here, it is making the internet a better place. It ain't glamorous or sexy chasing down people that might not want to be chased down but the internet is better for it. So thanks.

MATT THOMAS: Thanks, Jeff. And Jim, I see your hand's up too. Please go ahead.

JIM GALVIN: Thanks, Matt. Jeff, I'm curious about one thing—an overarching question here. As Matt's been going through some new analysis for us

on some of these strings, an interesting question has surfaced. He's focused on data that he has available to him with A and J root. Your analysis was a little broader than that. You had access to more data. Do you have any thoughts about how that matters or doesn't matter? Did you do any analysis in that respect—where things appear source-wise, different root servers, that kind of thing?

Even if you didn't do anything formal, do you have any insights based on anything that you might've done in that direction? Because even Matt has demonstrated even between A and J, some of the analysis that he's done has shown that you get slightly different pictures between the two sets of root infrastructures. I'm wondering what insights you might have about that point. Thanks.

JEFF SCHMIDT: Yeah, Jim. Great question. Yes. We looked at that. A couple of things off the top of my head, and I think there's nuggets like this buried in our report here and there. But first of all, A root is certainly anomalous. A root is different than all the other ones. And we hypothesized that one, if not the reason for that is because a lot of stuff tends to have A root hardcoded in. They don't go and prefetch the hints file the way they're supposed to and spread queries around. They just have it hardcoded into query A root because somebody thought that was a great idea.

We've seen that behavior in embedded devices, low power consumption devices, devices that have part-time connectivity—things like that where there's not a full-fledged local resolver in the device but they just need to make a query every once in a while and somebody just

hacked it together and said, “Okay, I’m going to query A and we’re not going to worry about hints and this and that.” We saw that behavior and if you look at the data it’s lucky that Verisign happens to run A and J because they can see A and then something different than A. But if you look at DITL data sets, A is certainly an outlier.

We went down the rabbit hole of trying to determine a sphere of influence, which is what we tried to call it at the time, going down the rabbit hole of the AnyCast. Which root servers have which ASs nearby? And trying to come up with proximity and this, that, and the other thing. But some of the root servers have hundreds of instances and they’re peered all over the place with an ever-changing list of Ass. And then you get into not all of the BGP policies are equal. And certainly, some of them don’t prioritize, just network nearest.

And so, we actually gave up on trying to figure out any preference or topology on root severs based on BGP, just because it became untenable for us, given the constraints we had at the time. Not to say that wouldn’t be a useful endeavor but we stopped pursuing that.

MATT THOMAS: Thank you, Jeff. Yeah. Certain measurements would be nice to have but it feels like sometimes you’re organizing your Q-tip drawer, I think. It gets a little messy. That was very, very helpful for me and, I think, the group. If anyone else has any questions for Jeff, please raise your hand now. Otherwise, maybe we can spend the last 15 minutes and dig into .internal. I think we could probably get through that in the time allotted.

JEFF SCHMIDT: Cool. Thanks, everybody. Thanks, Matt.

MATT THOMAS: Thanks again, Jeff. Really appreciate it. Okay. All right. So this is going to be very similar to the .mail presentation. Very similar graphs, very similar measurements. There's a few different ones in here so I will go through this fairly fast. But obviously, raise your hand if you have any questions. Can we have the next slide, please?

So this is a case study of looking at .internal. Again, this is using A and J telemetry data from Verisign. And the first graph that we're looking at is total daily query value. The graph on the right is the aggregate of A and J combined. The ones on the left are split by A Root, J Root, and old J Root, as well as IP version.

There's clearly something that happened at the beginning of March or so in 2020. Maybe a pandemic, I don't know, caused this. But suddenly the queries for .internal seemed to suddenly grow very, very significantly compared to the historic growth rates. I think it was on December 16th, actually, A and J Root saw a spike of 1.2 billion queries for a day for .internal names.

But if we continue to go onto the next slide, again here we see that most of these queries are coming out for A and quad A. Then there's the mixture of the various other Q types that we see at other TLDs. I think the last presentation I made a couple comments, "Oh, this looks like a standard distribution of the various long-tail queue types."

Jim asked to see if we could compare apples to apples on those. So I inserted this next slide—if we can go to that please—of various delegated and undelegated strings comparing their queue types. And when I made that comment of, “Oh, this looks like the standard mixture of other queue types,” that was mainly in reference to it doesn’t look like .arpa, where .arpa is predominantly made up of pointer or PTR queries there. The regular delegated TLDs seem to have a standardized mixture. But you will notice with corp, home, local, and mail, you do seem to have a little bit more of a prevalence of SRV or service records.

So, maybe there is a little bit of a bias in some of these strings and that might be to as to why those strings are leaking in the first place. It’s because they are DNS service discovery-oriented type queries, in which they’re just auto-magically getting sent out to find these new services. And you have things like suffix search list appending to them and they’re coming out. So hopefully that gives a little bit more color, Jim, for your question last week. But if we can continue to the next slide.

Going back to just .internal, along with the increase of query volume that we saw over the course of the last four years and especially since March of 2020. We’ve also seen a pretty significant increase in the number of distinct IP addresses at A and J for .internal queries as well.

So one could argue given COVID, given the timing, that this increase in query volume and this increase in diverse IP addresses is … Maybe the likely underlying cause is transient devices are no longer in their corporate environment. Everyone took their corporate network equipment home and is starting to use that at their home and the .internals are no longer resolving in their enterprise resolution systems

but are going through their residential ISPs and coming back out through this.

So, if we talk a look then on the next graph where exactly this traffic is coming from, this one seems to be a little bit different, again, than .mail. Over 50% of it is coming from the United States and then a significant percentage, about 15 of it, is coming from Ireland. And then it quickly goes into the tail. But exactly what those two in the US and in Ireland are become a little bit more clear when we look at the autonomous systems—on the next slide—that are requesting it.

And here we can see that upwards of 85% or 86% of all the queries for .internal are coming from two autonomous systems. And actually, if you'd looked up 16509 and 14618, they're actually the same company. It's just they have two different ASs. So there is one company out there that is responsible for upwards of 80-some percent of the leakage of this particular string. And in the graph on the right is the cumulative distribution of those ASs and the percentage of traffic that they're leaking out.

And if we can go to the next slide, I think I actually put these side by side compared to what we looked at with .mail. This is definitely up and to the left, meaning that there are a fewer number of sources that are responsible for much more of the traffic. So this, to me, based off looking at prioritizing outreach and cleaning up leakage and name collision problems on the internet, this would be a good candidate to go after. It's high query volume and there's a huge percentage of the traffic at a few particular sources.

So if we continue to the next graph, here we can take a look at some of the labels specifically. The graph on the left is looking at the length of the number of labels contained within the queue name. In contrast to .mail where we saw upwards of 60% of them only had one single label, here these seem to be—I'm using air quotes here—"richer" or more full of context, in terms of the labels and the content within the query name itself.

And this is great for analysis, like Jeff just presented, because it gives you the opportunity to get more context into what's actually causing some of these queries. And it clearly shows, once you start looking at the second level domains in this middle column, given the ASs and the strings, what exactly is causing these. This even becomes more confirmed when you start looking further down into the third label, so forth, and so on.

Now, the next graph, if we actually remove all of the queries that exist for some of those popular second-level domains like EC2 and Cloud, and stuff, what does the long tail actually entail? And this is, I think, interesting because at 250 million queries a day, even if you remediate 85% of it, 15% of 200-and-some-million is still a really big number. So what's the rest of it and can we potentially look to understand the underlying cause of it?

And so the graph on the left is looking at the remaining third and second-level domains under .internal. I know it's a little bit harder to see but there are clearly some that are specific to enterprises. But there also appeared this pattern of .rancher. And I have to admit I had no idea what Rancher was. And a little bit of open-source intelligence, which is

again why I was asking Jeff about that. I came across the web page for a Kubernetes technology that is called Rancher and it suggests putting your DNS under a suffix search list for Rancher .internal. So I would assume, based off of that exact code snippet that we found on the Rancher website and the queries that we're seeing at A and J, that the Rancher software is probably very likely the root cause of many of these others.

So, I have made it an item to submit a ticket to GitHub and disclose this. But just, again, a good example of trying to figure out what is the underlying cause of these queries. If we go on to the next slide, I want to take another little segue away from .internal and look at a different string.

And this is for .svc. And so the graph on the left, again, is A and J combined traffic for .svc over time. And as of a few days ago SVC was suddenly receiving upwards of 120 some million queries per day, which is a pretty rapid increase over the last year.

I performed an analysis, which I didn't include in all of this. But it looks very similar to .internal, where 85% of the traffic ended up coming from a Washington-based, very large operating system company whose cloud infrastructure seemed to have been using some kind of Kubernetes configuration and it was leaking out these SVC queries, even though they thought their internal zone was set up to capture it and prevent them from going out, but it wasn't. And here on the right, you can see some of the code snippets of that codebase where it is explicitly using .svc.

So, again, another way to identify and look at some of the underlying causes and risks of some of these queries. If we can go to the next slide please, and the next? This, again, is looking at the data sensitivity and growth over time. This curve definitely looks different than what we saw with mail. There is, again, a discrepancy between what A and J Root sees, in that while as time progresses the curve does flatten. However, it does seem to continuously be receiving additional IPs over time. That being said, I'm not too surprised, given the amount of IP space that the underlying .internal queries are coming from. I believe they own multiple slash eights so there's a good amount of IP address range that they could call over a period of time.

If we could go to the next slide, please. This, again, is just looking at number of queries per source IP address over the course of the month. 50% of them, on the left, issued less than 10 queries during the course of the month. It's using that as a threshold to further gate and see if the curve's bent. They further flattened a little bit but nothing significant to be of concern.

If we could move to the next slide, please. Again, this is looking at SLD overlap between A and J. We're looking at data sensitivity. And, again, here we see that the catchment of A and J do differ. Each one of them has its own viewpoint of what it's seeing into .internal. And, again, as time progresses that curve starts to flatten.

And finally, on the last slide this, again, is looking at the last cumulative distribution of how many times a particular SLD was seen over the course of the month. And that is 80% of these SLDs were actually only seen once over the course of the month. So, again, gating on that we

can flatten the curve and understand this data sensitivity analysis on that. But that is the .internal analysis in all of 13 minutes. I think we have three minutes, if anyone has any questions right now. Otherwise, we can open it up to any other business for the meeting. Yes, Warren. Please go ahead.

WARREN KUMARI: One question and one observation. I believe that the Rancher issue is actually people trying to make Rancher talk to Consul, which was an issue that we'd seen before, Consul trying to do a weird DNS thing. The reason I'm mentioning any of it, though, is I think that that is all from an old version of Rancher which they stopped developing in 2018. And Rancher 2 has been out since early 2018, mid-2018. The fact that we still see evidence of this, I think, is important to note for how quickly we might be able to mitigate issues that are baked into code. That was the observation. And then the question was as part of your mitigation stuff have you reached out to Amazon about fixing the .internal leaking? And I think we all know it's Amazon from 16509, etc.

MATT THOMAS: Yes. Thanks, Warren. And I absolutely agree on your first commentary. I think this speaks volumes to when it's put into software, what runway, length of time. How long does it take for something like this to roll out? And I think if we look at other strings like around equipment manufacturers we can imagine that time scale will probably grow even by the magnitude. And to your question, yes. I've been in contact with them. They're aware of it and they are investigating and trying to fix it.

WARREN KUMARI: Okie dokie, thanks.

MATT THOMAS: Thanks, Warren. Any other questions or comments? Otherwise, I think the plan for next week is to continue with the case studies. I have prepared a .court already and I am halfway done with .home. I believe we'll probably be able to do both of those in one session. So we'll come prepared and hopefully, I'll have those decks out prior to Monday or Tuesday of next week for everyone to review. I'm not seeing any other hands. It's top of the hour, just to be respectful of everyone's time. Thank you for coming and we'll see you next week.

KIM CARLSON: Thanks, all. Bye.

MATT THOMAS: Thanks. Bye.

[END OF TRANSCRIPTION]