KIM CARLSON: Hi everyone. And welcome to today's NCAP discussion group call on the 3rd of February at 19:00 UTC. In the interest of time, there will be no roll call. Attendance will be taken by those on Zoom. Kathy and I will update the Wiki with the names of the participants as quickly as possible. We have apologies from Ram, Russ Mundy, and Jim—Jim's here.

All calls are recorded and transcribed and these recordings and transcripts will be published on the public Wiki. As a reminder, to avoid background noise and echoing while others are speaking, please mute your phones' microphones. And with that, I will turn the call back over to you, Matt.

MATT THOMAS: Thank you very much, Kim. And good afternoon, good morning everyone wherever you are and welcome to this week's weekly NCAP discussion group. Somehow, we have already entered into February. I'm not sure where the first month of 2021 went but here we are. Today we have three or four main things that I think we're going to try and discuss during this week's call.

If you remember in our proposed study 2, we had outlined various case studies that we wanted to look at on particular strings. Over the last several weeks we've been looking at those case studies or at least doing the first initial groundwork of them looking at data from A and J root. And today we're hoping to maybe conclude those six different strings.

We will do a look at .lan which is not on any kind of special reserve list but we will also then look at .local which is on a reserve list. And hopefully, we'll continue to understand what kind of data properties we're seeing and that will help inform us in terms of our guidance that we ultimately have to form for the board's questions.

So, with that, why don't we just go ahead and get started into the overall meeting agenda? At this point in time with item number two, does anyone have an update to their SOI that they would like to declare or mention to the group?

WARREN KUMARI: I mean, I guess I do. I don't think it's a substantive change. My original SOI said, standards technical program manager. My current position is now technically director of Internet standards. I don't think that matters at all for anyone but technically it's a different title so I figured I should declare it.

MATT THOMAS: I appreciate that, Warren. It's always good to just have it out there. Thanks for that.

STEVE CROCKER: But that does mean you've been promoted, right Warren?

WARREN KUMARI: Yeah. But it's not nearly as cool as every time I join the meeting. Kathy promotes me from—that's a much more prestigious thing.

STEVE CROCKER: Well, in any case, I think we should note it and say congratulations.

WARREN KUMARI: Thanks.

MATT THOMAS: Congratulations, Warren. I'm not seeing any other hands or comments so we will conclude the update to SOIs and then move on to an update on study 2. Jim, if you wouldn't mind just spending a few minutes and refreshing the group on where things stand since last week and the progress that's been made.

JIM GALVIN: Sure. So, the package for the board has been created. There was an SSAC member who had some questions, concerns about the handling of conflict of interest statements from the original NCAP proposal.

And the fact that there was some concern that seems to have changed given the way that we had restructured the project where more of the work is done by the discussion group. So, in any case, that is ordinary with SSAC processes. That member has prepared a statement reflecting their particular position on this issue.

And that has been included in the package that is being sent to the board and a properly redacted version of that package will be distributed to this group here before it's—well, as it's sent up to the board through the SSAC board liaison for the board technical committee to give it due consideration and hopefully find in our favor to provide the contracting support that we're looking for, for this.

So, the process is moving along. Nothing substantive has changed in the proposal documentation. Just a little bit of administrative process so that we cover all bases and all of that will be visible to the discussion group and then we're just waiting for the BTC. So, I think this should be the last update. The moment that you get a copy of the report, you can assume it's been sent to the board and then we just wait. So, thanks.

MATT THOMAS: Thank you, Jim. I appreciate that. Any questions regarding study 2 or are we okay to move into the first case study on .local? Sorry. Not seeing any hands, so can we please get the slides up for the .local presentation please?

Okay. So, again, this is going to be a very similar rinse and repeat presentation that we've done on the other CORP, HOME and MAIL internal strings. This one will be focused on .local and the following one will be on .lan and that will conclude our case studies.

These two case studies I will admit are a little bit lighter on some of the slides, specifically looking at the data sensitivity analysis. Looking at SLD growth, [inaudible] growth over time and that will become evident when we start looking at the data. But there are just so many names

that my poor little Mac laptop couldn't stick a couple hundred million into memory every day and make these graphs.

I don't think they're particularly insightful anyways at that kind of volume, but just know that there are a few graphs in these that were not in—that were in the other ones but won't be in this one. So, if we can go to the next slide, please.

So, this is again starting off just looking at longitudinal trends of .local query analysis or traffic again at A and J root servers split out by IPv4, IPv6 on the left as well as each individual root.

And then on the right, we have the total amount of traffic. As you can see over time, it has climbed up to be a pretty significant amount of traffic. Recent traffic volumes are over 1.5 billion queries per day at A and J.

We experienced that same level shifting pattern that we observed on various other strings that we looked at in the mid-March, 2020 range when the whole COVID work-at-home thing. So, again, maybe this is indicative that this string is used in some kind of capacity of where you have transient devices now being used in a non-standard or expected DNS environment where those queries are now leaking out into the public DNS.

And speaking of, I saw Warren brought up SAC113, this is actually a little snippet from it and it highlights that .local was intended in 6762 for being reserved for multicast DNS use just for reference in there.

The spike is I believe sometime in mid-March to early April, sometime in there. Jothan, I don't have the exact date but I can come back and get that for you. It's roughly in that end of the first quarter of 2020. Q1 of 2020. If we can go to the next slide, please.

Again, this is the same longitudinal view of traffic but again broken out by the queue type. Here, you start to see a lot of traffic requesting for SRV records which we've observed in several of the other more popular strings that are leaking up into the DNS.

But we also do see a fair amount of solo records which is a little bit different than some of the other strings. If we can please go to the next slide. And I think this is a little bit interesting in terms of the discrepancy in terms of increased traffic versus not seeing an increased number of sources.

So, when some of the other strings like .internal, when we saw the traffic spike presumably around the COVID area, you also saw the number of unique IP addresses requesting that string also increase. Well, we don't really seem to see that in .local especially in that same timeframe.

Now, I will say that for whatever reason, the number of sources recently are increasing for that. I don't have any other explanation of why that is happening but that is definitely different than the other strings that we've been looking at in terms of that pattern of increased queries also usually associates with increased diversity of requesting sources. If we can continue to the next slide, please.

KIM CARLSON: Matt, Rod has his hand up real quick if you want to take that.

MATT THOMAS: Go ahead, Rod, please.

ROD RASMUSSEN: Yeah. It's interesting. I was just curious, since it doesn't match the pattern that we saw, pandemic pattern and other things, but I was wondering if those source IPs were maybe because of some broad use of public resolvers or large ISP infrastructure which would just bury it behind very large sources of queries.

But it just struck me but I don't know if you've done any analysis on known large public recursives or publicly available recursives or large ISP recursives as differentiating traffic from that versus other AS numbers for example.

MATT THOMAS: Good question. And I have not done any kind of analysis on that and that might be a good future experiment for us to do, is to look at a subset of the traffic coming out of known public open recursive resolvers compared to the rest to see how these graphs shift and change.

I will mention that on this graph, it does look like if you look at the graphs onto the far right, that maybe more of the growth actually is coming out of IPv6 address space.

I don't know if this also could just be a by-product of Verisign moving around its announcements and site locations of A and J. That might be something else I also need to take a look into and account for this. Steve Crocker, is your hand up? Please go ahead.

STEVE CROCKER: Yeah. I too was thinking about what might account for this phenomenon you were just talking about. It occurred to me that if you've ever seen an IP address used even once and then there's an increase, that will show up as no change in the range of IP addresses.

Whereas before, you had IP addresses that had never been used, presumably, that are then included. So, if there were a phenomenon in which there was a background of all of the IP addresses occurred a little tiny bit and then you had a massive increase, you wouldn't be able to see that phenomenon exactly in the way that this is being analyzed. So, that was just another thought trying to explore why this is behaving as it is.

MATT THOMAS: Those are good thoughts. And, yeah, I think a lot of that is maybe accounted for in just what we might all refer to as background radiation noise in the IP space out there. I will say that just even our traffic data at A and J, we have people sending us responses which makes no sense.

So, it's just the Wild West in terms of traffic out there. And in cases like that, that can also be an artificial side effect in terms of what our data is

capturing and seeing related to other non-related events. Warren, yes, please go ahead.

WARREN KUMARI: So, two things that I find interesting is the strongly [weighted to diurnal] nature of this. But what to me also seems interesting is if you overlay the change on J root and old J root addresses, I'm assuming that that doesn't accurately track what other names are doing, correct? I would think that for the distribution for .foo or something has a fairly different distribution between the real address and your old address and that that might be a very interesting signal. Does that make any sense?

MATT THOMAS: Let me see if I can rephrase it and see if I made sense of that. You're thinking that the growth or the decline of a particular string at old J root wouldn't be of interest?

WARREN KUMARI: Not so much the decline of it as the ratio of decline going to old J root for .local versus other names. So, in .local has this weird attribute that it is used for DNS discovery and the [bonjour] type stuff and many other things.

This means that queries might be more likely to be coming out of actual machines and not user-based queries. I think that that might have a different distribution or change as things move from old J root address to new J root address. Does that make any more sense?

MATT THOMAS: That does. And I will try and noodle on that a little bit more and see if we can do some future measurements around that as well.

WARREN KUMARI: Yeah. I think, I mean, an easier way to do it—easy and handwave would be to do something like you calculate an average for a string going to—and then calculate the ratio of it going to old address versus new address. And then you do it also for .local and see if that's very different.

MATT THOMAS: Okay. To see where it falls in the distribution and then look at the heads and tails of both of those to see what kinds of strings are coming out of those?

WARREN KUMARI: Yeah. And that should also help explain or show whether these are coming through normal resolvers or are being sent from something else.

MATT THOMAS: Absolute sense. I don't see any other hands at this time so why don't we continue on with the next slide? This is just again a standard distribution—a geographical distribution of where the traffic's coming out of. Almost 50% or 46% of the traffic is U.S.-based.

You have a little bit more than 10% coming from China and then you hit the long tails of countries. So, a large amount in U.S. and then still spread throughout the world. Pretty much like every other string we see on the L root. If we can go to the next slide, please.

Thank you. And the one thing that I would like to point out on this ASN distribution graph on the left is the scale of the X-axis. All of the other case studies, they usually topped out at, I think mail was 900 ASs. I think internal was 1500 or something like this.

.local is definitely spread out way more. It's almost 22,000 different ASNs we see traffic on a given day for .local queries. That being said, 50% of that traffic comes from roughly the top 10 ASNs. The ASNs on the right there if you look them up, most of them are major, large commercial ISPs or residential ISPs in the U.S. or in China. So, not a big surprise there.

And moving on to the next graph, please. Here we're taking a look at that IP growth over time and the data sensitivity of—discerning between A root, J root and old J root.

Again, here we see the same things that we've seen in other strings that the overlap between those three different catchments does have a fair amount but each specifically J root always has this much larger collection catchment than I would say A root does.

So, the graph on the right then is continuing to take a look at over the course of December, 2020, the number of unique IPs that we're seeing over time hitting those .local.

Again, as time progresses, the curve starts to flatten but ultimately, they're still up and to the right even after 30 days. If we can go to the next slide, please. So, this is going to look at the label analysis. On the left, we have the percentage of queries based off of the number of labels present in the actual query.

Almost 40% of them have three labels, only roughly 6% or 7% only had one. So, when you compare or contrast that to something like .mail, where 50% of the queries were only coming in with one label, .mail, this is very different.

I also have to wonder if there is certain interesting aspects that you would look at for certain strings in which almost all of the queries are coming in with only one label. If that tells you something particular about their association with QNAME minimization or something else about the string itself.

I think that might be an interesting well to dig and see if there's any water in that as well. But the tables on the right, the middle one is looking at the most popular second-level domains and the one on the right is looking at the most popular third-level domains rank ordered by the percent of total traffic.

Looking at the SLD list, this again is looking kind of a mix of what we saw in .corp, I believe which some of them look to be anchored under delegated TLDs. You see com and net and some other actual delegated TLDs in the list out there.

So, it makes you wonder then if that is again a byproduct of suffix search list appendage on those names or in other cases, you have things

like very specific like LORD and Fujitsu and Samsung demo where it looks like they very likely anchored those names under .local themselves.

And again, then on the right we're seeing—as we continue to expand the QNAME, the furthermore left, the first label, I should say, a lot of these again are being associated with either DNS service discovery protocols or just other general protocols that you would see out there.

Next slide, please. Actually, I wanted to end on this slide. So if we could switch back to the .lan and then we'll come back to this. I apologize, Kim, I meant to do .lan first and then do this.

All right. Well, let's see if we can continue our last case study on .lan so if we can hit the next slide, please. Here we go with our standard longitudinal view of overall traffic pattern growth. Again, on the left IPv4 four, IPv6 broken out by root.

On the right, it says the total daily amount of traffic received at A and J. What's interesting here is that you do again see that level shift in 2020 around the March timeframe but prior to that, it also seemed to be on a pretty good upward trend.

The other thing that I think is pretty interesting on the graph on the right is the decrease at the end of 2020. And this is what we'll see hopefully later in the slides, again, associated with the patch of Chromium and its reduction of random domains coming out to the Internet or to the roots for that.

Now, the other thing I'd like to note about .lan, while it's not on any special list right now, it is used in OpenWrt which is an open source project for routers, home routers that contains a wide variety of software security features all packaged together that you can install on things like your Linksys WRT54G back in the day or whatever.

But it does make use of the private TLD, private-use TLD .lan in there so it's already part of that. And if we can continue to the next slide, please. Again here, we're taking the same look at longitudinal traffic pattern broken out by queue type. Again, mostly A in quad A but we do see a decent amount of SRV records in there just like we have with the other ones.

Next slide, please. And here we are looking at the unique daily sources and this goes back to reaffirming our pattern of during the COVID bump. I would say that you see an increase in number of unique daily source IPs. But it's probably not nearly as pronounced as some of the other strings that we observed.

And then again, you see a small decline in the Chromium patch time but it seems to have rebounded, gone back up again. This will have to be another task for me to investigate because looking at the graphs on the left again, most of this seems to be coming out of IPv6 space so I need to take a look and see if Verisign changed anything again with IPv6 catchment or placement of our sites for A and J around that time period.

That being said, I would expect if you're getting .lan IPv6 increase, you would experience at the same time as .local. So, I'm wondering if maybe

this was something else in terms of how OpenWrt is now maybe more equally treating IPv4 and IPv6 addresses or something of the like.

Next slide, please. Again, looking at global distribution of the traffic, this time only a little over 20% of the traffic's coming out of the U.S. This is much more evenly spread out over the world. France, and China, and Canada accounting for pretty significant portions of this traffic as well. So, this, I think—

Yeah. Next slide please. And here we're looking at the ASN distribution. The one on the left, again, just for the X-axis you're talking about roughly 10,000 different ASNs using or sending queries for .lan QNAMEs. Again, almost 50% of this comes out of the top 10 ASNs.

The graph on the right plot, those ASNs and again, if you looked up those AS numbers, they are all mainly residential and commercial ISPs. And then the rest of them are spread out around the world and other countries ISPs.

Then continuing on to the next slide, please. Again, here we're taking a look at ASN overlap between A and J for that data sensitivity analysis. And again, we see the same pattern, J root and A root each having their own unique observation space.

While it does overlap, each one of them still seems to have its own catchment. Longitudinally over the course of the month, those curves start to flatten but again, they continued up and to the right.

I do have to admit I was not ever actually encouraged by how much any of the curves flattened in terms of all of these case studies that we've

done, which has always led me to wonder how does that impact future data analysis like if we're to rely on DITL data, is that two-day sample enough? Is it not enough based off of what we're seeing?

Or is this also just a byproduct of the Internet and just backscatter, that these are not really major sources, that they're just sources sending one query every so often and it's not making an impact overall?

Warren, I saw your hand go up. Sorry. I had the list-scrolled the wrong way. I apologize if your hand was up for a while. Steve, please go ahead.

STEVE CROCKER: I had a question about the display of the range of IP addresses. If you could go back about three slides or four slides and I can frame my—one more at least. Yeah. That's good. So, I just realized, I don't understand exactly what these graphs show, because I don't understand what the range is. Instead of having a single point on the Y-axis, you have several different points at a given point on the X-axis.

MATT THOMAS: So, there should only be one single point on the X-axis. This is for the number of unique IPs per day. It's just that this is over four years of data so you have roughly four times 365 to 1400 or 1500 points on here.

STEVE CROCKER: So, you're saying that in a short period of time, within a few days, you have a high degree of variation?

MATT THOMAS: Yes. That is the weekend effect that those lows are, I guarantee, Saturdays and Sundays.

STEVE CROCKER: I see.

MATT THOMAS: Yeah.

STEVE CROCKER: Thank you. Sorry. I just was concentrating and then I realized—I didn't realize exactly what that—go ahead.

MATT THOMAS: Thanks for the question. Warren, please, go ahead.

WARREN KUMARI: Would you continue forward a couple of slides? I think two more back to the cumulative? Next one. Yeah. So, I mean, I suspect that it is just this is growth over time. But I think also that a host which is generating .lan queries, if it gets an IP address and then move to a different IP address the same host—because DHCP or moves on to a different network, if it's mobile or something like that, will lead to some sort of increase. I don't know how one accounts for that though. I don't know if you can.

MATT THOMAS: I totally agree with what you're saying, that because the IP space is so vast and that if things are just moving around this curve never flattens because of that, right? And the only other thought I had was to maybe look at it like unique /24s over time or uniques—some kind of sub network or ASNs but then you're always somewhat limited, right?

WARREN KUMARI: Yeah. I mean, it seems like we should be able to figure out the fact that it starts—a way you could potentially account for this again is if you were to move forward three or four days, reset your counts and do again and see whether you get the same sharp increase and then it tails off.

I think that that might give you some additional info though I don't know what it would mean. So, basically if you were to move forward by a week to December 7th, restart your counts and draw the same graph and see if you get the same—it starts off fairly sharply and then becomes flatter over time. That might allow you to correct for this. [Inaudible] stat something.

STEVE CROCKER: Aren't they restarted every day? I mean the graph that we were looking at a few slides before where I was asking the question, that fluctuation in there must indicate that there is a refreshing of the count, doesn't it?

WARREN KUMARI: No, for this graph, this is a different graph to the other one. I was meaning for this particular graph.

STEVE CROCKER: I see.

WARREN KUMARI: So, this is cumulative over—however long this period is. My screen isn't—

STEVE CROCKER: I see you want something like a seven-day moving average on it.

WARREN KUMARI: Yeah. Anyway, that's a separate—like that requires somebody who actually understands math and stats, I think could something, something handwavy. And obviously in Matt's copious free time because I'm assuming doing this work takes no time at all.

MATT THOMAS: It just comes out and it goes right into Google docs. Like I snap my fingers and it happens.

STEVE CROCKER: Ask Alexa to help you out.

MATT THOMAS: So, actually I would like maybe pose a question then to the group again, looking at this cumulative graph. Like is there a concern—so my concern about this is then going again back to, is this a representative sample if you only have one day or two days or five days or whatever the duration is?

But I guess the question that ultimately is, what are you trying to compare and what is the attributes, the data that you're worried about missing and that if you had a two-day window here and you compared it to a two-day window there, are they similar in the risk factors that you're looking at, right?

Like, so if we use the two days at the beginning of the month on the top and second-level domains the same as the top end second-level domains 15 days later, right? Or is it what we're worried about is this measurement over this time period showed that the traffic was really coming from a handful of networks versus the traffic manager [inaudible] from a subsequent period of time shows that it's over this crazy diverse [inaudible], right?

STEVE CROCKER: So, that's interesting. Let me ask the question related to what I was asking before, maybe it relates. When you get the raw data, do you see how many requests are coming from the same IP address? So, you get a range of IP addresses, are you able to see for each of those IP addresses how many requests are coming so that some will generate a lot and some will generate only a few?

MATT THOMAS: Yes. We have that data. So, I could do some plots of that and show the distribution of how many requests per IP.

STEVE CROCKER: So, what I was suggesting before was that if you make up some bins, some ranges, you could plot the previous graph in a way that would show that. And then I totally agree with the discussion on this graph that the cumulative nature of this may be obscuring or not telling you what you want.

And then what you want is either instantaneous or a seven-day average or something like that but one that has a natural refresh of it and then you would see whether or not there is actually any growth or whether it's just bouncing around.

MATT THOMAS: Exactly. And I think that would be an interesting measurement especially if you've had a handwavy, true statistical measurement to compare the variance in terms of the rankings of things for this time period versus a different time period to show that variation to show how stable it would be.

STEVE CROCKER: If you have trouble understanding—and I wasn't being crystal clear about it, but if you want to chat further about what I was trying to say, I'd be happy to take it up offline.

MATT THOMAS: Sounds great. Thanks. Warren, is that a new hand or an old hand?

WARREN KUMARI: That's a new hand.

MATT THOMAS: Okay. [Inaudible].

WARREN KUMARI: So, one of your questions was, is two days a long enough time interval to look at? And I really don't think it is, largely because we do see things like diurnal and weekly type patterns to make much of the data.

And so I think that one would need a substantially longer amount of time to actually see things to figure out whether something is actually representative, right? If we did a DITL run on a weekend, that's very different to a DITL run on Wednesday.

There's also, I think a bunch of other things that we do see a number of strings which are largely dormant and then suddenly appear and get used heavily for a while. Many of them, you've hunted down and mitigated but they do still pop up and are important, like console for an example as a string that if we had checked at some point and then checked three months later, there would be a huge difference and delegating it would break a substantial amount of things. And then it got mitigated and the problem got better.

Some of what I'm wondering at this point is, based upon a bunch of different data, to me, it seems as though it's not really feasible to mitigate many of the instances that we see, right? Unless it's a specific thing where it's one or two strings which are coming out of one or two large cloud providers or ISPs, reaching out and mitigating things across many different networks is really hard.

Even mitigating it in software is interesting but it takes a long time for things to get rolled out. There's also the—we don't have full visibility and so I'm at this point kind of wondering, other than the fact that this is fascinating work, fascinating and interesting about the DNS and stuff, I'm not really sure if this is actually helping answer, can we predict what strings are going to make the Internet go boom, and is there a way to mitigate any of these if we do discover them?

So, I'm getting to the—do we need to reevaluate what it is we're trying to accomplish with this? I'm a little scared asking that both in case we do actually redo it and discover there's a bunch more work, but also because I'm concerned that if we do that you might stop generating these reports and I find them interesting.

MATT THOMAS: Thank you for that Warren. And I actually have three questions that I'd like to throw back to you because that got my brain thinking a little bit more. If we could go back one more slide, please.

WARREN KUMARI: And also while you're doing that, I'll admit my biases. I don't really think that we can mitigate strings which are deployed. And I think that determining if a string is dangerous requires semantic analysis coupled with deployment analysis. But I believe I can't prove but I believe that .nuclearreactor is a much scarier string to monkey with than .doritos.

MATT THOMAS: Absolutely. I think there's a contextual element just to the string itself, right?

WARREN KUMARI: Thank you. That's a better way than semantic. Yep.

MATT THOMAS: You're not going to delegate .embeddedinsulinpump, right? Because of that [inaudible]. I hope not. Actually, can we try and go to the slide with the Venn diagram, the three circles? Perfect. Okay. So, one of the questions that you brought up was—and [inaudible] talked about two days probably isn't enough.

But then that brought up the question of, based off of the ASM overlap analysis, how many groups or how many data collection points do we think we need? Because obviously with the overlap on A and J, each one of these has its unique catchment and so more is more, but is more better, or is the representativeness of A similar to J based off of whatever security measurement that we have posed, right?

So, I think that's something else that we need to consider is, if two days isn't enough, can you counteract that by using DITL where you have more data sources [from a catchment?]

WARREN KUMARI: I think that that comes back to contextual meaning of the string as well, I think. There are certain strings which only show up on weekends or not only but predominantly show up on weekends versus predominantly show up on weekdays.

I think that A root is also special in that people who manually do testing and look up the names from monitoring systems or from a command line default to A in their fingers, so that's a different thing. I think it also—if .nuclearreactor shows up only on C root on Thursdays, that's potentially more worrying than it showing up on—than .doritos showing up at a low level—sorry, at a much higher level spread across all letters. So, answer unclear, ask again, I think is what the magic eight ball says.

MATT THOMAS: So, then that brings up one of the other points you brought up about, there is a temporal aspect of when this analysis is done, and is that something that you think we need to potentially factor into the guidance that would come back, that these data measurements and risk assessments of applied for strings need to happen at a one point in time, at which point in time, multiple points in time? Is that something that we need to be concerned about? And how do we frame that kind of guidance?

WARREN KUMARI: I mean, I think it kind of depends on what outcome we're expecting. But in some ways, I feel that if I was a brand owner or a community representative and I wanted to apply for .mybrand or .doglovers, I would have no way of knowing if .dog or .doritos is a string that actually is going to be possible for me to apply for. I don't think there's any realistic way for me to figure that out.

I don't have, as a community person or brand owner, I don't have visibility into stuff like DITL. I've never heard of DITL. So, I think that potentially there's a—I want to apply for a name. I come up with a name. I take it to a group, possibly ICANN and say, "This is what I'm planning on applying for. Here is my deposit. Can you tell me if there's even a possibility of me getting a string?"

If I decided I wanted .corp or .local, I would go along and say, "Is this within the realms of possibility?" And there could be an initial, fairly lightweight check of, oh my goodness gracious me. No, that string is obviously squatted on by a million people or a million resolvers, give up and go home.

And then assuming that that doesn't happen, maybe then there's phase 2 where I actually pay more money and really finalize my application. And then before the name is delegated or much later in the process, there is a, we're now checking again to make sure that the string hasn't suddenly moved from, “this looks okay” into, “the world will explode. “

Of course, we do then run into my standard counter this, which is gaming, right? If I know that you're going to be applying for .doritos, I

could start artificially making that string look as though it is squatted on, not usable, whatever the correct term is at the moment. So, I think that this gets back to the, this is hard and that's why they're paying us the big bucks.

MATT THOMAS: Exactly. I just got my check yesterday.

WARREN KUMARI: Excellent.

MATT THOMAS: All excellent points, Warren. I would like to try and push through the rest of this case study so we can conclude this. And then I would like to possibly tie your last statement about the scope of this back to the last slide on the other deck for future conversations of what I think we should go on the next couple of weeks.

Can we go to the next slide please? So again, this is just a quick analysis of looking at how long the second-level domain was under .lan. It's not visible as other TLDs where you have the flat plateau between 7 and 15 characters but the fact that it suddenly drops off at 15 characters does at least suggest to me that a good portion of this traffic is probably Chromium queries.

Especially when you have almost 47 unique million names coming out in a specific day for that A and J under .lan. So, that's why a lot of the SLD

analysis fell on its face. It just didn't fit in my [Mac R session.] Next slide please.

And then this is just looking at the specific SLDs in the label length distribution. Again, most of the queries for .lan are only coming in at two labels. So, again, this is probably the random Chromium query .lan accounting for a vast majority of the traffic but the rest of them, then you can see in the middle column.

Again, we see many of these coming under already delegated TLDs. So, is that the by-product of suffix search list processing and .lan is being attached to those, or is it that they're actually being anchored under specific entities or corporations or whatnot?

To me, this looks a little bit more suffix search list like. I don't know exactly how to quantify that though. And then again on the right column, you can see that first label or the further left you go, again these are all associated with DNS service discovery and other API type protocols.

Next slide, please. Yes, this again was just a similar, you know, comparing the .lan second-level domains. On the far right, sorry, this is really messy, .lan second-level domains is on the far right, comparing it to .home, second-level domains in the middle.

Again, here, you can see that delegated TLD like structure between these. Again, this is suffix search list appendage of the string. Next slide, please. Maybe that was the end of the deck. Kim, if you could, could you switch it back to the other slide presentation?

KIM CARLSON: Yeah, that was the last slide on that deck.

MATT THOMAS: Thank you. Sorry for not telling you the order of this. I apologize. Awesome. Okay. So, I know we only have 10 minutes here. I guess we'll just probably take most of this conversation over into the next week's call to elaborate around this.

So, we've gone through six of these case studies and I think as a group now we see some of the key data attributes that we're measuring and looking at when we're looking at collision strings. And I wanted to try and taxonomize these a little bit to at least put some kind of structure around them and broken them down into three different kind of properties

The first of which I would call traffic properties. Again, and this is just looking at things like how much traffic is it getting? What kind of diverse traffic sources is it at an ASN level or /24 level?

Is it heavily skewed to come out—is all the traffic mainly coming out of one or two ASs like we saw with .internal? What's the geographical diversity? Is there something specific with the QTYPEs? And how does the traffic look over time?

Then you have other properties that look specifically at the QNAMEs and the labels, right? Again, is the traffic distributed over a large amount of names or is most of the traffic coming out of a specific set of second-level domains? How much Chromium noise or other background

radiation noise is there in it? Do the SLDs appear to be delegated TLDs suggesting this is a suffix search list appendage?

What kind of other features are in the QNAMEs? Is it all associated with DNS service discovery or other common protocols? And then also, what are the effects of the recursive resolver environment out there? Are we seeing a lot of this being impaired by QNAME minimization or other things?

And then the last, I would say, it's just maybe not as quantifiable but it goes to Warren's example of a nuclear reactor, what's the string's context, right? I think that plays an important part of this. Like Java, right? I think that was an applied for a string last time.

Obviously, it has the association with the programming language as well as coffee, other things, right? And then, there are other things that, just doing some open source intelligence gathering on that string via Google searches and whatever, can you associate that strand with things like OpenWrt projects or Kubernetes Rancher deployments and stuff like that?

And then finally, what kind of sense do you have in terms of data sensitivity? Is this something that was super specially localized to country? Like we saw with .cba in the last round that was all coming mainly out of Japan and Chiba but then the Commonwealth Bank of Australia actually thought it was queries coming out of them when it was actually not.

So, these are the high level data points that I saw. And what I hope to do tomorrow or next week is, I'm going to do a K-means clustering of

the strings based off of these different attributes. And we're going to look at delegated strings, non-delegated strings, and the mix of strings to see what kinds of strings grouped together based off of these various different measurements.

And hopefully then, that will help inform some more general insight into what we're seeing and then help form some of the additional guidance. I do want to give Jeff a moment to talk. Jeff, please go ahead.

JEFF NEUMAN: Yeah, thanks. And this is all interesting and it's good to see what's out there but I stick by the comment I still have is, so what, right? What is it that these queries—I mean, we know where they're going but we don't know for what purpose.

And so at the end of the day, the board may be faced with a balancing test of, what are we going to destroy if we delegate this string, and does that matter, right? Is it one person or one company's application and then we'll just tell them to find something else or is it something bigger, right?

And that's the part that we can't just—we could. I don't think we should just take the approach of if there's a bunch of queries and it's associated with a number of SLDs and there's a number of IP addresses that go to it, then that's proof of some harm that would be caused if we delegated the string.

That's going to be important for board members. And then board members need to make a choice, right? They need to balance, is the

harm of delegating the string greater than the potential harm of not granting the string, to basically giving priority to this previously used application that may not be configuring their system right? So, I think it's important that we have this data, but are we ever going to be able to answer the, so what question? Thanks.

MATT THOMAS: Thanks, Jeff. And I see a bunch of other hands going up. I know we are close to the top of the hour and unfortunately, I have a firm stop at 3:00 so I will drop off. And if Jim, if you could just take off to the end, if we carry over a little. Warren, please go ahead.

WARREN KUMARI: So, yeah, this is Warren. I mean, I think that there are certain things that we can tell from those two strings. Like for example, if one were to delegate .local, I think it's fairly clear that a bunch of stuff would go boom.

But I don't think that there is a way and I don't think there ever will be a way to say that this is exactly what the harm will be, especially for other strings that we don't have more understanding of like .home. And I don't think that there will ever be a way to say, "This is safe."

So, I think that what we can possibly do, yeah, I mean, Jeff's right, yes, .local [inaudible] using it as an example. I think that what the best we can do is we can say, "We can tell that delegating this string would likely be hugely impactful and would affect a lot of people, so this string is

dangerous to do something with." Kind of like we did for home, corp and mail.

But I don't think that we'll be able to do the inverse of that, which is this string does not seem as though it will be harmful. Unless there is a way to specifically tell exactly what is generating the queries, I don't think that we'll be able to get better answers.

There are a few small instances where we might be able to tell exactly, like .console as an example. We could tell what that was because it's an unusual string and the distribution was relatively small, so we could hunt it down. But there's a bunch of other things that we don't know.

And to answer Jeff's question, I mean, yes, there's a bunch of blockchain stuff emerging. Blockchain is very similar to .onion and that it's not really supposed to be using the DNS for resolution but seeing as names leak, you can't predict what will happen if the name is used.

And as for if the harm is great enough or not, I think that that depends on harm to whom, who is the responsibility of ICANN to protect. And so, is one person dying okay or is a million people each losing $1,000 okay? That's not something that's a technical question. I don't think that's anything that we can ever answer. I think that's the ICANN board has to decide what's acceptable harm and who their responsibility of care is to.

JIM GALVIN: Okay. Thanks, Warren. I did have myself in the queue at that point but I'm going to let—so, this is Jim Galvin because I know that Matt had to drop off. I'm going to let Steve jump in here and make his comment and

then I'll make my comment and bring us home. Go ahead, Steve, knowing that we're at the top of the hour here.

STEVE CROCKER: Thank you. I'll be very brief. Jeff, you phrased it in terms of having to make the case that there's sufficient harm. You could just as easily take the converse question and argue that there's no good reason to delegate a particular string. And so the threshold ought to be pretty different from what you're suggesting if you make the default that unless you can be shown that it is absolutely safe, there's no reason to delegate it. I'm sure that's not a comfortable position, but the way you phrased your challenge opens the door for exactly the opposite and then you'll find yourself in a very uncomfortable position, I would think.

JIM GALVIN: So, thank you, Steve and—

JEFF NEUMAN: Can I just quickly respond to that just [inaudible].

JIM GALVIN: Sure. Go ahead, Jeff.

JEFF NEUMAN: Yeah. And Steve, I understand that but in my mind, anyway, and it's a simple mind, it's basically ICANN rewarding those that are not using the system necessarily correctly, right? And so if there's more of a reward to

going outside the ICANN system, why would you stay within it? And I think that is extremely important.

Otherwise, these other things—and I just saw a press release today of another crypto-type thing that someone's launching or someone raised money for, right? If we can't convince people why they should be in the system, then they're just going to go outside. And why should anyone else be—

STEVE CROCKER: Actually, there's a big hole in what you're saying because they're not getting delegation. They're polluting the system, I agree with that, but they're not getting delegation. So, they're not going around and accomplishing the same thing. Obviously, this is not the time and place to have an extended version of this discussion but there is a serious discussion to have here.

JIM GALVIN: Yeah. So, in fact, let me jump on that and just offer that that's exactly the right answer, Steve. There is some discussion to be had here. Jeff, I think that you bring up some valid concerns and I will offer from my own point of view that there are lines to be drawn because as you said, Jeff, there's a balancing act that has happened here and the board has to do that balancing.

The lines that we have to draw that we have to have the discussion about is, what is it that we expect the board to balance? What I'm imagining is we're going to provide guidance on what kind of data

should be submitted as part of a package to the board to make a decision. And then we're going to provide some guidance about how to balance what it's seeing in the data.

And this is also, in my mind at least, predicated on the idea that, name collisions are here to stay and they'll always be there. So, we are going to have to have a discussion about what is harm, how to look at it, how to see it, does it matter if you see it or not, kind of thing.

These are the hard discussions to have and provide some guidance to the board to perform that balancing act. We might not be able to create a completely objective and deterministic process for a third party but at least they know what kinds of things are going to be evaluated.

And I don't know where all this is going to go. There's definitely some hard things to be figured out here because I don't think we're going to come to a final answer. And then, a whole business about mitigation and harm gets us into discussion about how much of a study 3 do we have to do or not do. And we'll get into that discussion as we get further down this analysis process.

I think with that, I'm going to have to say that we have to draw a line under this. We're at top of the hour. I want to respect people's time. Thanks everyone for coming. We're starting to get into the interesting stuff here. It was excellent to look at all these graphs, see the data.

Many thanks to Matt for doing all this work. Now is when I'm going to get into the meat of what's going on and what we're going to do with it. Plus, we are ultimately hoping to get some other data sources to volunteer to answer the same kinds of questions for us and we've got to

put that together so that we can distribute that and see if we can make that happen. So, thanks again and we'll see everyone next week. We're adjourned.

KIM CARLSON: Thanks all. Bye.

[END OF TRANSCRIPTION]