KOBE – DAAR Update
Wednesday, March 13, 2019 – 08:45 to 10:15 JST
ICANN64 | Kobe, Japan

JOHN CRAIN: One of our reporting tools and normally if you come to these events, you have to listen to me [drone] on but I've got a new employee and she's going to tell you all about what we're up to now. So, why don't you introduce yourself and take it from here?

SAMANEH TAJALIZADEHKHOOB: Hi, everybody. This is Samaneh and today I'm going to talk about the update on the Domain Abuse Activity Reporting tool.

So, this is my background. I started since October 2018 as a SAR specialist and working with John Crain. I'm the DAAR project owner and I do research on other topics related to DNS abuse. I work remotely with John. I'm located in the Netherlands and I have research background in web security, Internet measurements, banking security policy analysis and similar topics.

Okay. We all know that domains are increasingly used for malicious purposes or abuse online. Therefore, there is a growing need for proactive measures by the operators and

registries and registrars. We know that the measures that are currently taken are heterogeneous, meaning that there are some operators that are taking a lot of measures and are doing good in terms of fighting abuse and there are those that are doing less and/or know very little from the abuse concentration in their own networks and in comparison to the networks of their peers.

When you look closely into what is the reason for that, why there is such a gap between how operators perform in terms of abuse handling and abuse monitoring, you often see that – well, the profit margins are thin, so basically this is not a priority. There is no unified methodology that providers can use to monitor abuse or to fight abuse. In short, there is not enough incentives to do that.

So ICANN came up with this project, Domain Abuse Activity Reporting. Just to give you a heads up, because most of you have heard about the project before, I'm going to do a very short introductory about the project itself and we'll focus more on the update of what has happened since Barcelona. The DAAR project is a system for reporting domain name registration and abuse within registries and registrars. That is the original idea. Something like DAAR already exist in Academia, for instance. There is a lot of research on that and other industry players also have different reputation list or reputation metrics. How is DAAR different from the current existing systems is that it studies all

the gTLD registries and in the future hopefully registrars that we can collect data on. Because the data collection has been going on already for a year and a half, it allows for historical search and comparison and analysis.

It employs a large set of abuse feeds and studies multiple threats. And one of the most important points is that it takes scientific reproducible approach, meaning that everything that has been carried out as a methodology in DAAR in principle should be reproducible by any industry player or researcher out there. So, that's that.

The data can be used to study malicious registration behavior, to report on threat activity level on gTLD for now on TLD level, for historical security threat concentrations and should assist the operational security community to an academic research for policy analysis and decision making. And hopefully help operators for their anti-abuse programs, terms of services, etc.

So, a bit of introduction on the methodology and the data that is used in DAAR. Basically, we have three sets of data employed: the DNS zone data, the WHOIS data, and the open source or commercial abuse feeds or reputation feeds. I will use these two terms interchangeably during the presentation.

This is in summary how the data is collected and merged with each other. Number one is the zone file which is collected either

from the centralized zone data system that is developed by ICANN or is collected individually depending on the TLD.

So when we get a list zone files, depending on registry or registrar level metric, we collect WHOIS data and that is the registrar ID from the WHOIS data and we overlap that with block list or the reputation data feeds, number three here. When you merge all these data sets, the result is basically a reputation metric which expresses how much abuse is concentrated in a network of gTLD operator for now.

More into the reputation feed stats that are used in DAAR. So, for now we are collecting data on these four reputation types, phishing, malware, spam, botnet command and control. It's important to understand that DAAR is not identifying all abuse types, only about four. We are basically looking at the domain names that we have from the zone files that are associated with the names in the reputation – in the abuse feeds.

How did we select the data sources? The data sources, first of all, they should match the threat classification that we use, to the four types that I showed in the previous slide. They should have positive reputation in the academic literature or studies and among the security communities and it should also be broadly adopted by the security community for commercial security system, network operators, or messaging providers, etc.

Is DAAR itself an abuse list service? No. It's just a collection of abuse feeds that is compared with a set of domains in the zone file. We are collecting the same abuse data that any person basically can collect. Some of the feeds are free, so they only require registration and some of the feeds are paid services. But in principle, anybody can pay and get the feeds so that the study would be reproducible. There has been also academic research using these feeds, so most of these feeds have been over and over evaluated in different studies.

This is a summary of the lists. They are different feeds within each list but I don't list the details here. For more information, you can always visit the DAAR page on the Methodology paper and look at the details of the abuse feeds.

Okay. So, we move on to the DAAR monthly reports. To give you a short summary of what has happened so far is that starting from February 2019, we started publishing monthly report based on the DAAR data I just outlined for you. These monthly reports, first of all, you can find them on the DAAR webpage, so this is the red webpage I'm highlighting here. This is how the page looks like. On February 4 we published the first of the monthly reports and later, end of February, we published the monthly report for the whole 2018 and two months from 2019. You should be able to find all the reports on the webpage.

There is also a Context document followed by the reports which briefly explain what is the report, what's the goal of the report, and provide some kind of context for those who do not want to go through the whole Methodology paper but want to quickly understand what is the report aimed for.

Here I'm showing some examples of the analytics that can be extracted from the data that we are collecting. So it's important to have in mind that these are not the only basically metrics that can be extracted from the data; these are examples. Hopefully with your help and input we will be able to improve and provide more and more examples that can also be useful for you as community.

What you see here in the plot is an overall distribution of the abuse data we are using. Basically, more than 89% of the data is spam. As you can see – and this is the data from January 2019, so it's a snapshot. Then the percentage is followed by phishing, malware, and botnet C & C domains.

Here you can see the total number of domains identified as security threats over time. So, basically this allows for time series analysis. You can see that within new gTLDs, the trend of security domains account has been more or less very stable and getting closer to each other.

Is there any question so far? Yes?

UNIDENTIFIED FEMALE: A gentle reminder, kindly provide your name and affiliation please. Thank you.

RICK WILHELM: Rick Wilhelm, Verisign. I thought in slide four it said that it was increasing. This shows that it's flat.

SAMANEH TAJALIZADEHKHOOB: You mean the general –

RICK WILHELM: Yes. Slide four said that abuse is increasing. This line appears to show that it's flat.

SAMANEH TAJALIZADEHKHOOB: Yeah. I know what you are referring to. I talked about the general abuse trends. Of course, for the data we are collecting, it is flat but what is also important to have in mind is that it's becoming harder and harder for the reputation feeds provider to detect abuse domains. For instance, if before a provider was able to detect one domain for the reputation feed and find the others based on WHOIS data, doing the search is becoming harder and attackers are becoming more advanced in hiding domains. So, the flat trend that you see in this plot does not

necessary say that the abuse is becoming more flat, it can also be an indication that it's harder for the reputation feeds providers to detect the abuse that is going on.

JOHN CRAIN: This is numbers on what's on the reputation feeds. There are other data sets out there. There's a recent document out by SURBL and Spamhaus, I believe, that talks about their ability to detect and how those numbers are actually going down. That's another area that we actually want to go and study out this. Every time you look at data, you find more things to study.

We often get asked the question, have we seen an increase in reported names in DAAR since May 25, and the answer is no. We've not seen an increase in reported names, which is this flat that you're seeing. Maybe we're asking the wrong question.

SAMANEH TAJALIZADEHKHOOB: Thank you, John.

JOE WEIN: Maybe I should comment on that. Joe Wein from SURBL. We found for domains which we track via registrant information, numbers of registrant that we've been able to track have roughly gone down two-thirds since GDPR has come into effect.

SAMANEH TAJALIZADEHKHOOB: Yes. Thank you, Joe. Is there any other question? Yes?

TIM CHEN: Hi. Tim Chen from DomainTools. Do you pull anything from WHOIS besides the registrar ID which I think the field you mentioned? And my second question is, are you doing that pulling yourself or are you relying on a service to do it for you?

SAMANEH TAJALIZADEHKHOOB: Can you repeat the question?

TIM CHEN: Sure. You mentioned in WHOIS you pulled, I believe, is the registrar ID field when you said this data coming from WHOIS in your chart. And my first question is are you pulling other fields or is that the only field you take from WHOIS?

SAMANEH TAJALIZADEHKHOOB: Yeah. Thank you for the question. For now we are not doing the registrar level metrics. I will explain further why is that, but if we were going to do it, we were only going to extract the registrar ID field.

JOHN CRAIN: And you're talking about pulling the data, right? So to be technically correct, when you do a WHOIS query, you don't get to choose which data you get back. It's how you process the data when it comes back. We're all using a third party to do the WHOIS collection. We have statistics on registrars so we have some of that data, but it's completely not of the quality I would like due to the fact that pulling WHOIS data at that kind of level – for example, if you talk about 180,000 roughly or 190,000 names in the zones, if we were to do a 190 … I'm sorry, 190 million. If we would do the 190 million WHOIS queries a day against your systems, you would probably get upset at us.

[DICK MOLANDON]: Hi. My name is [Dick Molandon]. This is not a question about the statistical data that you've got. This is a question about how much is this is costing OCTO? I mean how much of the budget feed is going towards the feeds, for example, this external party you just mentioned now to do the WHOIS data checking, etc.?

JOHN CRAIN: I don't have the exact numbers but one of the review committees has asked for those numbers and we've gone after them to pull them out. I hate estimating these kind of things but it's less than 100,000 in feeds I think. But please don't [take that thing]. We've been asked for those numbers in another process

and I've gone to Finance, etc. and said, "Hey, can we disclose these and be [what are they]?"

SAMANEH TAJALIZADEHKHOOB: Okay. This is the distribution of the resolved domains that are within the zone files. As we all know, the majority of domains are located in the legacy gTLDs and around 12% are in new gTLDs. When we look at the abuse percentage, we see that more than 50% is located in the new gTLDs and around 48% in the legacy gTLDs at least. When looking at these plots, it's important not to get a biased conclusion from the plot and basically pay attention to the fact that what does it mean when we move in the report from up to down? What is the logic of the report?

So if we dig more into the detail of this 52% abuse in new gTLDs, we see here in this plot that actually around 90% of this 52% is derived by less than 50 providers. So basically, within the new gTLDs this is the area that if they have more active abuse fight, we will able to reduce abuse significantly. Now, the abuse might move to other places but at least this is an action point that we can conclude from the plots.

When we look at this part of the plot, so the 48%, we see that as we all expect within the legacy TLDs, around 90% of the abuse is derived by something around four of the gTLD operators, which

is expected because in the gTLD space the heterogeneity is way less so they are really big providers that derive the whole space.

Is there any question up to now? No? Okay.

Here you can see the distribution of TLD types over different abuse types. So what you see on the bars is the percentage of abuse domains over different threats. As you can see, some threat types are only concentrated in certain TLD types. For instance, botnet C&Cs are majority located in the legacy TLDs, whereas for phishing and malware, it's more equally distributed.

In this slide you see two plots. In this plot on the left, on the X axis you see the number of domains that are resolved in gTLD and here you see the number of domains that are abused. And within legacy and new gTLDs, the trend is upwards, meaning that what the plot is showing you – and this is also an example of how visualization can be biased if it is not controlled for certain variable. And I put it purposely also in this report to show that it's important to take certain characteristics of the provider into account if you want reliable abuse metrics.

So here you see that for instance this provider is probably the worst one in terms of abuse because it has the maximum abuse count. And by the way, this is on log scale. But when you normalize this, meaning that when I take the size of the provider into account, you see that the plot totally changes. So in order to

give an overview on the X axis you see the same, so the number of domains that are resolved within the TLD space and on the Y axis instead of seeing raw count, you see a percentage of abuse and that is the number of abuse domains divided by the size of the provider. So you are actually now taking the size of the provider into account. This is needless to say but of course the more domains and operator has, the more attack surface is there. So it's important to control the operator size.

When doing that, the circle size here also shows the absolute number of abuse domains, you see changes a lot. For instance, if you follow this line over here, you see that providers with the same size – so this is same size providers – are doing very different in terms of fighting abuse. There is this provider that is doing [inaudible] very good and there is this provider that actually has the highest abuse percentage. So I would say this is the space that we would like to focus on to know why such difference exists and what the community can do to basically decrease such difference.

UNIDENTIFIED FEMALE: I have a comment online from Alan Woods from Donuts. Noting one of your initial statements, can we be a bit clear that registry operators don't need an incentive to deal with abuse? And whilst abuse management is an expensive aspect of being an

operator, it is something that a responsible actor shall of course undertake for the benefit of the DNS and ultimately for the protection of the Internet user. It is remiss to invoke profit margins and is rather an inflammatory thing to suggest that registry operators generally are more interested in making money than keeping a safe space. I have no doubt that there are very small percentage out there that perhaps would be so cynical but can we please not generalize on this as it helps nobody.

SAMANEH TAJALIZADEHKHOOB: Good question.

JOHN CRAIN: I'd say thank you, Alan, for that and so noted. I don't think it was meant to be a generalization or maybe it was a bad generalization and we would take that into note.

SAMANEH TAJALIZADEHKHOOB: Yeah. And what is actually important and I try to focus on that was that initially I also pointed out that the space is very heterogeneous, meaning that there are providers that are taking very good measures and there are providers that are not. I outlined what are the possible reasons and the data suggest the

same. So you see here that the data suggest the same. We would like to help to basically reduce such difference.

UNIDENTIFIED FEMALE: I have a question.

SAMANEH TAJALIZADEHKHOOB: Yes?

UNIDENTIFIED FEMALE: Hi. So will you be providing the community with more details on the providers? I recall that the statistical analysis of DNS abuse in new gTLD's report that ICANN had published a while ago, the SADAG report, use many of the similar source data providers and provided more details out of registrar and registry level. Is that something that you're planning on doing in the future?

SAMANEH TAJALIZADEHKHOOB: I'm actually one of the authors of that report but back then I was in Academia so I'm quite familiar with the work. In short, we are planning to provide data to the operators but I will go more into the details of that once I get to the next steps of the project.

JOHN CRAIN: As of the question of whether or not we're going to start adding names to the monthly report which is what I think you're actually asking, that's not going to be my decision. I think that's a community discussion. We have no plans at this current time.

UNIDENTIFIED MALE: Hi, this is [inaudible] from Neustar. In the previous slide, could you talk a bit about your use of log scale versus linear scale here?

SAMANEH TAJALIZADEHKHOOB: Yeah. The reason why I used log scale was the numbers are big and it's hard to visualize purely. So basically what would happen is that because the differences are a lot, you would see a bunch of dots here and it's hard to visualize. It's a technique in statistics and data analysis to basically help the audience to see the visualization. That's it. It's not used in the actual metric. It's just for the purpose of visualization.

UNIDENTIFIED MALE: Sure.

SAMANEH TAJALIZADEHKHOOB: Any other question? Here you're seeing that trend over time for how abuse evolved within new gTLD. Basically, what

you see as a summary is that not much is changing. No matter what abuse type, the trend is quite stable.

These were some examples of the visualizations on analytics that can be out of the data. The current project status is that we put the DAAR project for review, the DAAR methodology for review. We received reviews and comments. We published our responses to DAAR comments on February 1. We published a series of the monthly reports of the 2019 and 2018 ones. This has already been a starting point for constructive discussions with registries and industry members. It's really important to have in mind that this is work in progress. What we really would like to have is your feedback, what are the things that you think should be added that can be useful for you as the community, registries and registrars. Basically, that's it. Yes?

REGINA LEVY: This is Reg Levy from the Tucows family of registrars. I'd like to know more about specifically what each data point represents in terms of "abuse." What is it you're specifically looking at? What types and levels and whether, for example, one thing stays up for four days. Does that count as four data points? Is it one or etc.?

SAMANEH TAJALIZADEHKHOOB: Sorry, I'm not sure if I fully understood the question. Could you repeat it again?

REGINA LEVY: What constitutes abuse?

SAMANEH TAJALIZADEHKHOOB: Basically, what we count as one abuse data point is unique domain that is seen in each of the abuse feeds. I didn't explain this because it was already mentioned in the previous DAAR presentations. I didn't go into the details but one abuse data point is one unique domain per day per data sheet. Does that answer your question?

REGINA LEVY: Partially. So you indicated that it's one domain that appears one time in all of the abuse reporting feeds or in simply one?

SAMANEH TAJALIZADEHKHOOB: It's sum of all. I mean depending on what metric you're talking about because there were different metrics. But depending on the metric, we aggregate per feed per time unit. It can be a day or month, and then it's a collection of different feeds.

REGINA LEVY: Thank you.

JOHN CRAIN: So if you took a specific data point, it may be one abuse in the overall count but it may actually appear individually on different things counted separately. So if something could be a phish and a malware – in fact, sometimes they are – they'd be indicated as both. In the overall count, it would just be one particular point in time. But we're open to discussing if there are different ways of measuring.

For example, we have data of about length of time on the list but we've not cut the data that way for publication yet. We'd like to know how to cut this data that's more useful. For example, at the moment we're only doing new versus generic, but there are different ways of defining different types of TLDs. I mean we've made up all kinds of different acronyms over the years for different types of TLDs. Is that useful? Is that not useful? I think that's a good discussion to have. We have the data, it's supposed to help inform some of the discussions so if we need to cut it in different ways, let's have that discussion and figure out if we can do that.

REGINA LEVY: I think initially it might have been extremely important to distinguish between old legacy gTLDs and new gTLDs but I think it's starting to become more interesting to see gTLDs versus ccs, for example. The main thrust of my initial question though is I want to make sure that as the abuse reporting aggregators that you guys are using might update what they consider abuse that ICANN continue to stay away from dealing with content.

JOHN CRAIN: Absolutely.

[WATSON SINGH]: [Watson Singh] from Afilias. Does it study how long the appeal is going to stay there? You say you don't double count that if they're going to be [inaudible] and then come back again. Did you count them? Also did you have study of how long is the average staying time?

SAMANEH TAJALIZADEHKHOOB: I will answer the second part of the question first. We currently do not have the data on reoccurrence. So the time, we don't have it. But it's part of our future work. Like I said, depending on the time you are looking at, for instance, if a domain appears today on the list and goes off and appears three

days later, it's counted twice if the metric is per day. Does that answer your question?

[WATSON SINGH]: So you're basically saying that 24 hours is the time that you use to cut off?

SAMANEH TAJALIZADEHKHOOB: Yes.

[WATSON SINGH]: Okay.

JOHN CRAIN: And the reason for that is access to some file data. We use CZDS at 24-hour granularity.

SAMANEH TAJALIZADEHKHOOB: Yes.

ROD RASMUSSEN: For clarification, if I'm seeing a monthly report and it appears multiple times during the month, is that one or one for each time of the month?

JOHN CRAIN: Let's be very clear. The monthly reports are point in time in a month. They are not an aggregate of the statistics over a month. We do say that in the documents, etc. but I know people don't always read all the nitty-gritty, but they are point in time with a specific date on the report of which date that is. They are not an aggregate of monthly data.

ROD RASMUSSEN: Being one of those persons that didn't read the nitty-gritty and the way it's labeled, it would be helpful to maybe make that list nitty-gritty because that assumption –

JOHN CRAIN: We will use bold fonts and spell it out more clearly.

SAMANEH TAJALIZADEHKHOOB: That's a good question actually.

JOHN CRAIN: But actually, in all seriousness, if you see other things like that where you're missing things – if we're having these misunderstandings and it's in the report, then it's our amiss for not making it clear enough. So, point them out to us. We will use bold fonts and things like that.

ROD RASMUSSEN: This is Rod Rasmussen by the way. It would be really useful actually to have this – when you're asking about, "How can we make this data better?" a point in time in a month is interesting. Hopefully it's representative of what the rest of the month looks like because if it falls on a Friday versus a Sunday, there's actually big differences in spam volumes depending on what day of the week we're talking about, right? So I'd actually rather see taking data aggregate across the month either averaged out or conglomerated together because it's like, "How many domains appeared in the month of January we're reporting on?" That's actually should be a much higher number but it wouldn't be 31 times as many necessarily in this particular report is. I think that's actually a far more interesting phenomenon to watch and it gets rid of bias especially if you think about February being 28 days, you get this bias. It's going to look really flat between March and February on the same day and we've got same kinds of activities going on. Thanks.

SAMANEH TAJALIZADEHKHOOB: If I can make a comment on that, I would say this is a valid point. But both of the methods have their downsides. Looking at a point in time and also aggregate per month both come with their downsides. But we are happy to provide both so then the community can use. Yes?

UNIDENTIFIED MALE: My name is [inaudible] from .global and we run around abuse monitoring system. We had the chance to compare our data with the DAAR project. That was quite interesting obviously. Well, the findings, that was kind of scary for me is that we discovered that domain name that was deleted some six or eight months ago was still appearing in the different reputation feeds. So the datasets that we discussed with ICANN were actually referring to a set of numbers that were non-existing domain names. I think that if you want to improve the DAAR system and the statistic about that, you should actually at some point check whether the domain name is existing, maybe resolving and so on. Whether it's difficult in the numbers you have but it gives a false picture for also the registry operator. And it also signals that those registrars that had actually done their job, they haven't been paid for that.

SAMANEH TAJALIZADEHKHOOB: It's a very valid point and it's in our future work to do.

JOHN CRAIN: Yeah. That was a discussion I think the day before yesterday and we've actually been looking at this already. It's on our work list. We've not put dates on it yet.

We went through the reviews. There are a lot of recommendations in those reviews and many of them we're actually going to implement are in the comments. So we're going to set up a timeline and we're going to start doing those but that's actually one of them. It's not that we're going to change the dataset. It's going to be a different view on the data.

So, one is what's in the reputation feeds and as number that you're measuring which is a very good one and we will also start measuring what says out of those data feeds, what are actually in the zone?

UNIDENTIFIED MALE: Just a small comment on that one. I'll also take that back to the different reputation feed providers. They should also improve their system because I don't think it's not for their reputation that they have data that is all domain name that is not existing being reported. Thank you.

ROD RASMUSSEN: I was going to raise a point that exactly addresses this. This is about fidelity in the block list in the various reputation feeds. As a network operator, a suspended domain is actually something I do want to know about. Even if it has been deleted six months ago, for the purposes of detecting malware on my network

trying to reach out to a CNC, for example, I need to have that information. Which ends up being is it's being published because people are going to take that information and put that into a system to do that particular task. That's not appropriate for reputation of what's actually in the zone. Because you're pulling zone files so you know what's actually resolving or not, actually it would be interesting to pull the status of a domain if it's been suspended or not. You can actually take a look at registry status and see potentially what to do [inaudible].

JOHN CRAIN: That requires a WHOIS lookup.

ROD RASMUSSEN: WHOIS lookup, yeah.

JOHN CRAIN: Or an RDAP lookup.

ROD RASMUSSEN: But only the ones that are appearing on the feeds, not the entire thing. Anyway, it's important for people to realize that it's not that the reputation feeds are reporting "bad data," it's reporting data and you have to understand what the actual impetus for that is. Sometimes people apply that properly with their

networks to protect themselves. And sometimes they don't because, "Hey, I've got a big list of things I'm just going to shove it into a firewall or whatever." So these are things to keep in mind.

But it would be good I think feedback to the reputation providers, "I have done this for many years giving this feedback. Give us better fidelity and what actually this is and what I should be doing with it."

JOHN CRAIN: And actually comparing the two datasets – if somebody is interested in the data, it's actually going to be interesting so we're all going to do that because we want to see what that difference is. We've taken that on board and we're taking that one on.

SAMANEH TAJALIZADEHKHOOB: Yes, go ahead.

RICK WILHELM: Rick Wilhelm, Verisign. Getting back to the point in time thing, I will have to admit that like Rod, I did not notice the point in time thing. They're staring me in the face on page 4 but I now apprehended and [inaudible]. And I understand why that the

feeds that come from the providers are point in time and whatnot. But the notion that the DAAR report would be point in time just absolutely baffles me from the kind of document that DAAR would be producing and where this document is going to go and the kinds of people that read it and then the actions that they're going to take because of it, the notion that it would be point in time is not understandable.

I understand your point about the fact that yeah, there's pros and cons of point in time versus aggregate. If you do an aggregate, it has downsides too. For the purposes in the kind that wants a document like this gets into the wind and starts blowing about the community and numbers as we saw in blog posts and things like that, once it gets in the air, the point in time document with its snapshot freeze frame approach has a potential to be much more misleading, more misleading than an aggregated number which while it might have its issues presents I think a more accurate picture of what's going on over the course of a month. Thank you.

SAMANEH TAJALIZADEHKHOOB: I think that's a valid point. I agree with it. Have in mind that within the same document, we provide also historical so you see point in time of other months, but it's a valid point and we will take it into account.

RICK WILHELM: Yeah, point in time with other months which is a series of random days. We know from our work in registry reporting – I'm deeply involved in our registry reporting that numbers within a month go all over the map for various reasons. Like Rod said, behavior, day of week matters a lot and such like that. Not to mention the fact that even hour of the day obviously in a global world. Thank you.

SAMANEH TAJALIZADEHKHOOB: Yes?

RICHARD ROBERTO: Hi, Richard Roberto from Google. Forgive me because I apparently also didn't read the nitty-gritty so it may be in there somewhere, but it seems like that the data that's being collected has been collected without a follow-up research that goes into products like this where you decide, "We've got this body of data. What kind of problems can we solve with it?" And then trying to figure out what the utility of that data is will help change the kind of reporting we want to see out of the body of data, and it seems like that maybe it hasn't been done. I'm wondering, "Is that true? Is there a plan to go back and try and solve that?"

SAMANEH TAJALIZADEHKHOOB: Yeah. What is the truth is that the data collection process is also in progress so we are improving as we receive feedback. We change how we are collecting with the third party. That's what we are working on. So, yes, the way that it's collected in the beginning is not exactly how we are planning to use it now or in the future but we are improving on that as well.

RICHARD ROBERTO: Is there a process that is ongoing?

SAMANEH TAJALIZADEHKHOOB: Yeah, it's ongoing.

RICHARD ROBERTO: And transparent too? Who is a part of that process?

JOHN CRAIN: At the moment, the only mechanism we have is publishing and coming to meetings like this. I would love to see some kind of group with those skillsets. We've got ongoing discussions about this. Once we have our timelines for what improvements we intend to do, we will be publishing that. But if you have ideas on the technical level of how to improve things then yes, please. It

would be nice as an industry – we were having these discussions more, not just at the policy level but also at the technical level.

SAMANEH TAJALIZADEHKHOOB: Yes?

PETER VAN ROSTE: Good morning. My name is Peter Van Roste from CENTR. Picking up on John's call for feedback and input. We had a really interesting discussion in one of our working groups, the Security Working Group, on this so I'm happy to take a few of those discussions offline.

But one recurring question – or let's call it a concern – is the selection of these feeds. What are the criteria? And that relates a bit to the overall question. Is there a governance model part of the decision? Thanks.

JOHN CRAIN: If by governance model you mean things like the ability to remove names and things like that, yes, it is. We published the generics of that. One of the things that we realized talking about nitty-gritty in documents, etc. is that the process of how we select and more importantly how we deselect needs writing up more clearly. And we actually need to work a little bit on that

process. It's another one of the things on our timeline to do. Fairly high priority one. For example, one of the reports indicated that we should add more malware data, that's great. And of course to do that, we're starting to go for our process that we use. We going to better document it and then we'll publish that. But yes, governance in that term is part of that.

SAMANEH TAJALIZADEHKHOOB: Okay, as we already discussed some of these points, here is a summary of the project's next steps in terms of methodology data and results. So we are planning to improve the system as a whole based on the comments and reviews received. We are [logged on], just pointed out we would like to develop a process for systematically reviewing feeds, documents that we are using, and this means evaluating them in terms of false positives, governance models, etc. We would like somewhere in the future to also be able to distinguish between maliciously registered domains and compromised because they are under different entities to be mitigated, and we understand that point and we are also keen to work on that.

In terms of data, we want to add more malware feeds. We want to have discussions about sharing the current data with registries if they are interested to receive that data, viewing their own data.

And in terms of results, we would like to develop new metrics. These metrics can be from all kinds. For instance, currently we are looking at the data from perspective of new gTLD point of view. We would like to also add some of the other properties of TLDs like if they are brand, etc. and some of the other things that we just discussed today.

JOHN CRAIN: On publishing data, the data that's used in the reports – obviously we have that on a monthly basis – our plan is to give the contracted parties access to that. We have a system called MoSAPI so I'm told by our registry services folks. That is a development and should be available around the Bangkok IDS GDD Conference. So you should be able to pull not the feed list because licensing doesn't allow us to pass through the lists but the data that we're using for the reports like how many names did we count, how many were phish, how many were malware, how many were different types – that will all be available to you on a daily basis. Or if we can improve our granularity, whatever granularity we have it on.

RICK WILHELM: There's a question earlier about cost and just related to – on slide 33 – there's some things on Next Steps. On these things obviously from a software perspective, before this stuff gets

memorialized into code, which of course cost money, some of these are more difficult thing especially the third bullet on the slide – distinguish between maliciously registered domains and compromised. When I saw that, one of my eyebrows sort of raised because that's a hard problem, I think, at least to me. Maybe not to other folks in this room but it is to me. Before that gets put into code which cost money and then run and then results come out, and then people start to [inaudible] about them, how are we going to get an eye on that "algorithm" right to reviewing in advance?

JOHN CRAIN: Any large cost items have to go through the budgeting process. For example, if that was to be a high-cost effort, we're not going to be doing it this year because we're going to have to budget, etc., we're going to have to do all the standard stuff that we have to do for budgeting. That one specifically might not be something we can just easily do. We might because there's a lot of smarter people than me in this room so maybe somebody has an idea how to do that easily, but like you, I don't. ICANN has budgeting processes. I actually don't have budget sign off so I have to go through all the approval processes, etc. to do that kind of thing.

SAMANEH TAJALIZADEHKHOOB: I would like to add to what John said. Like you mentioned, this is a very difficult problem. Currently I'm a bearer of a project within SIDN .nl registry and university [inaudible] that are funding students, especially students to work on this problem. So, it is on our list of future things to-do steps but we might also discuss and wait for the results of the research to have more before we put that in quote and contract with parties. Yes?

ROD RASMUSSEN: Having done that for about 10 years, it's kind of a 90/10 problem, maybe 80/20 problem or 90% of it is actually fairly easy to create rules around. The other 10%, 20%, maybe 30% is really hard. I would encourage that you get the providers to actually do that to the extent that can … because that actually helps the rest of the ecosystem out. Just looking as a consumer of data feeds, if I actually can have fidelity – again that word – around whether or not the domain itself is fully involved in whatever the activity may be or it's been compromised, I'd take different security measures around it. I may block e-mail but I may allow resolution depending on what it is, things like that. This is a scenario where I think ICANN could actually help from kind of a leadership role of pushing back on the data providers and say, "Can you do this?"

JOHN CRAIN: There is an ongoing conversation with data providers – and I see Joe put his hand up because I was going to ask Joe when he could he have that done by.

JOE WEIN: We do actually already have a data feed for correct websites. So that is already partly available at SURBL.

Joe Wein of SURBL. So we are one of the data providers and we already have a correct data feed that is available for this project.

SAMANEH TAJALIZADEHKHOOB: Thank you.

RICK WILHELM: One follow-up to try and tie to something – a comment that Reg made earlier. We need to be a little bit concerned I think about delving too close to content here when it gets to this sort of thing. And I'm not sure where that line is. I'm not an attorney, I'm not deep into the bylaws and what but I can feel the ground getting squishy underneath my feet here. We're big on content neutral.

JOHN CRAIN: We don't do any of these without talking to our lawyers and executives, etc. Those are decisions that are – above my pay grade is probably the wrong terminology to use but there are controls in place for that. For example, there are datasets that we would love to use but we can't because of contractual things. So every time we do something or we change something, we have to go through legal review and budget review and all those kinds of things. I think everybody in this community is very aware and sensitive to the issue of staying away from content.

SAMANEH TAJALIZADEHKHOOB: Yes?

TIM CHEN: I want to step out and just briefly from interrogating the inputs to this model to say I really appreciate that you're doing this. I think it's really valuable work and I hope you're able to continue it. I will say that having accurate data is incredibly important for any analysis, so I'm very supportive of that. Everything has been said here.

I want to point out that there's a lot in this deck and also I was reading the website about how this is a dataset for the community and you had a question a prior slide, "What can we do to make this valuable to the community?" Quite frankly, I

think it's limited right now and so I want to encourage you to spend as much time thinking about transparency for the broader ICANN community as we're spending time here talking about transparency for the inputs to the model. I'll give a couple of examples and I'll close.

The analysis begs a question, who are the 50 TLDs that are 90% of the abuse in that one? It seems like there's some resistance to publishing that. I'm not sure why. Secondly, you said, John, a minute ago you might publish some of these data to the contracted parties. I think this is all publicly available open source data. You make that point in the website for the most part except for the paid services, so I'm not sure why that data can't also be published broadly. So I just want to encourage that thinking. There's a lot of work to do. [It's hard to] belabor here but I hope there's a balance here and think about the community more broadly than just what certain parties care about because enormous value can provide more granularity certainly in security use cases that I'm familiar with.

JOHN CRAIN: Absolutely.

SAMANEH TAJALIZADEHKHOOB: Thank you for your input. I'll move on. There are also challenges ahead for the things we want to do. These are some of the challenges and we discussed others. For instance, we would like to have registrar level metrics and ccTLD level metrics. The problem with registrar level metrics is WHOIS data – John already briefly pointed that out. Because we want to use the publicly available WHOIS data and we don't want to use other sources, we need to query the database on daily basis. This is not possible currently with the amount of domains we have, it's hard to scale. So we would need to find other methods to collect this data and we are working on this, but we would like to also know your opinion and discuss this within the community.

The same goes for TLD metrics. In order to have metrics for ccTLDs, we would need the zone files of older ccTLDs which we don't have access to. We have had discussions with [inaudible] about this but this is also an ongoing discussion. So if there are operators that would like to give in their zone files, we are more than happy to include that in the data.

We also talked about remediation metrics but that is also another work that is very hard to do because we need to look at the up time or take down times of domains and it's complicated and we need more data points and measurements to measure domains over time but it is part of our thinking process.

So, where do we want to go from here? We would like to have open discussions with you like actually what we just have now or after this session. Also feel free to e-mail us any time to daar@icann.org. This is a channel to discuss your concerns or communicate it with us. Like I said, if you're a ccTLD and would like to input your zones into the DAAR system, feel free to contact us.

That's it. We will have a session on DAAR on IDS Symposium in Bangkok and hopefully in the future. I'm happy to receive your questions if there are any.

REGINA LEVY: This is Reg Levy from Tucows. The EPDP discussed the requirements for the office of the CTO for data aggregation and doing stuff with it. Research – thank you. That's the word I'm looking for. Good morning, everyone. Are you affiliated at all with that? Are you under the office of the CTO's research portion?

SAMANEH TAJALIZADEHKHOOB: Yeah. We are.

JOHN CRAIN: Yes, we are part of OCTO.

REGINA LEVY: Thanks.

SAMANEH TAJALIZADEHKHOOB: Any other questions? Yes?

JAY DALEY: Hi, Jay Daley. Can I suggest that as part of your next step you add in something about formalizing the engagement with the community, whether that's some form of standing/steering group or something? Because clearly there's a lot of detail that needs to be discussed with the community here and it would be useful to have people who can regularly contribute through that rather than just sort of casual meetings.

JOHN CRAIN: Yeah. It just occurred to me that we kept saying we have an e-mail and we should probably have a mailing list rather than an e-mail. Let's have an offline discussion about that if you have ideas or other people have ideas about … I think ICANN meeting is useful but it's three times a year and I think you're right. We need a better format for ongoing discussions. So I'm happy to talk about that. We have lots of different ways of doing that in ICANN. I don't know myself which one is most appropriate.

ALAN WOODS: Alan Woods for the record. Apologies for having the remote commentary earlier but I was in a class. A quick question for you there just as I came in, Reg, what you're talking about being part of OCTO, the DAAR specifically. It's very clear you do not actually process personal data at all. Is there any concept or necessity can you think of in the future where personal data, something that you would actually deem to get into?

JOHN CRAIN: I can't discount it because we don't know the future, but what I will say is if we ever get to that stage, we will be having very deep discussions about use of data and whether it follows the rules of the various regulations around the world. The problem is with researchers – "Yeah, we'll take all your data," and we often confuse wants with need. So if we ever get to that situation where we have a project that we believe needs that, we're going to have to argue the use case just like everybody should do when they use data.

SAMANEH TAJALIZADEHKHOOB: Okay, if there is no other question then I would like to conclude the session. Thank you all for participating.

UNIDENTIFIED FEMALE: Thank you, everyone. The slide material –

[END OF TRANSCRIPTION]