MONTREAL – DAAR Improvements
Wednesday, November 6, 2019 – 13:30 to 14:30 EDT
ICANN66 | Montréal, Canada

SAMANEH TAJALIZADEHKHOOB: Good afternoon, everybody. Thanks for participating in the DAAR Improvement session. It's a pity that I have to sit because there was no microphone for walking, but luckily I can see you all. I'm Samaneh . This is the second time I'm presenting the DAAR projects in an ICANN meeting.

This session will include less details about the project itself, assuming that the audience more or less know what the project is about, and will focus more on what other improvements that were made since the initial version and based on the feedback received. I will just briefly touch the definitions.

Just a brief introduction. DAAR is a project that uses two main sources of data: the zone files, which we get from CZDS and other sources, and the third party reputation block list. The system takes data from both sources, overlaps the data and creates a reputation metrics. The data sets that DAAR has been collecting starts from January 2018, and the collection is ongoing.

There are, on average, 1,200 gTLDs included in the DAAR data, and on average, 195 million domain names included. The system studies four types of security threats: phishing, spam, malware and botnet command and control domains. And the data collection frequency

depends on the source. For the zone files, it's once a day, and for the reputation feeds -- the RBLs -- depends on the frequency of the source of the RBL provider.

Why the project is different from some of the already existing metrics work in this area is that it includes historical data, it collects data on most of the gTLDs out there, it has monthly stats and it tries to adopt the continuous evaluation methodology, both for the methodology itself and the data feeds added or removed.

This slide lists some of the example metrics that the DAAR system can produce. The metrics include counts of the domains that are listening zone files, or basically, the size of the zone per gTLD. Also, the domains that are newly listed each month in gTLDs. Also, all kinds of percentages, which are normalized by size, percentage of overall security threats, but also, percentages per type. These are just some examples. And of course, time series metrics, because it goes back in time.

Here are some example analytics that the system currently produces. What is important to note is that everything that the DAAR system produces at the moment is anonymized and aggregated, so there is no indication of names in the DAAR report and analytics.

This is for instance, how the security threats, that the system collects, are distributed. So at the moment, around 80% of the data that we collect is spam. This is for September 2019. But from previous report,

I know that the stats are more or less the same. And around 12% of the data is phishing, followed by botnet C&C and malware domains.

This is another example analytic that the system can produce. The plot that you're seeing is showing the distribution of gTLDs and the percentage of security threats that they cover per threat.

So for instance, if you look at spam line, which is the dashed blue line, you see that around [AUDIO BREAK] percent of the spam is derived by less than five percent of gTLDs, or is concentrated in less than 5% gTLDs, which is very different from the case of malware, for instance, almost 70% is produced by 2% of the gTLDs. So what this plot is showing in general, is that there is a lot of concentration of threats within minimum amount of gTLDs, which influences actions or says that actions can be very influential.

Another example of analytics that the system can provide is time series analysis, going back in time to produce normalized metrics. So, what you see in this plot, which is plotted only for spam, basically, each circle is a gTLD, and the size of the circle shows the amount of absolute security threat count each TLD has in their network. And the x-axis shows the count of the size of the TLD zone in log scale and the y-axis shows percentage of security threat.

This is for month January, 2019. And what you see is, if you look at the line over here which contains TLDs more or less same size metric, you see that there are differences in concentration of spam. For instance, the same size TLD has 30% of domain used for some, versus the one

which is here, is the same size but has less than 5%. The question that follows up on this is why this happens. What are the differences? One can use this plot to go over time. So I'm just showing the examples of how can the trend be seen if I go over time.

This was a summary of what has been done, what the system can do up to now. Of course, we have had sessions starting from one in Kobe, which we discussed DAAR, and we've received feedback from you guys and the rest of the community, who might not be here. And I also presented the project in different venues such as IDS and other venues. And we've got feedback and hence came the improvements.

This slide shows a summary of the different requests that we received for the DAAR project to move on different directions. Community asked us for more transparency on the DAAR data progress and on the methodology itself. We've had requests on not reporting totals, so not reporting total security threat. Because the threat space is so much bias by spam because 80% of our data is on spam, but reporting it's per threat so it's distinguishable.

We've got requests to consistently use the term 'abuse' versus 'security threat'. We all know that currently there are ongoing discussions within the community about the definition of spam, so we are watching for that. We've got request on adding info on time to leave of the domains that are used for security threats, adding ccTLDs to the DAAR system, adding registrar metrics, publishing DAAR detail on anonymized data as well as distinguishing between maliciously registered domains and compromised ones.

Now, of course, there are limits to the changes we could and we were willing to make up to now. So we thought about every chapter of the request and we went through some processes to investigate how feasible some of them are and we add up some of the changes. One of the most important things that has happened in the past six-seven months is that we started to share the DAAR data with the registries so each registry can receive their own data.

This is an aggregate metric per type, so they can see what are the threat accounts per threat type. The data can be pulled through the MoSAPI API, which registries typically have access to. And this is how the data can be queried. I'm not going to spend a lot of time on this. What we know as of until September 2019, there were 21 registry operators that were pulling this data. If there are questions about how the data can be pulled or how it can be used, feel free to contact Gustavo Lozano from our DNS department, or one of us.

The next change that has happened in DAAR is that, for those of you who were in this discussion in Kobe or have seen the DAAR monthly reports, you know that the current stats are produced based on a snapshot date.

So, until now, the security threat metric that the system was using was, for instance, numbers on 31st of September 2019. So, the last day of each month. We've received feedback that aggregates are better than point-in-time data because of possible outliers. We agreed and we are going to change that from now on.

This slide is showing, basically, on a more granular level, on the x axis, you see days within a month, in this case, I'm plotting September 2019. And on the x axis, you see some of domains in zone. So basically, the plot shows the daily changes of sizes within zone files. On the right plot you see the same thing, but in terms of security threats, so the daily changes of the number of security threats within a month across all the gTLDs. Both plots have upward trends.

The takeaway from this slide was that, what is a good aggregate to use per month since we are now moving from a snapshot statistics to monthly. If you look at the mean and the median of each of the plots, of the threat plot and the total zone file, we see that they are more or less very similar.

So we decided just in case of future outliers, to use median, which is a better metric for capturing extreme cases. We've also worked with parts of the community, such as the registry stakeholder group, on discussing ideas and exchanging feedback towards producing newer metrics and analytics that are useful for part of the community.

I'm personally [inaudible] more into what kind of other factors are out there that could drive security threats within gTLDS. So all of us know that a very obvious example driver for security threat is size. Size can be used as an explanatory for concentrations of security threat domains, but it can also be used as an attack surface for cyber criminals.

Now, the question was what other factors are out there that we can capture them to explain why there are more concentrations of threats within certain networks in comparison to others. We know that in practice, there are some real world variables out there that we might not be able to capture all of them.

What is important is that, in reality, operators are dealing with several different aspects. The market is very heterogeneous, and size is not the only driver. So it's important to take all of these factors into account, such as the policies that the registries have, as well as their geographical locations, as well all other factors that might influence the overall security threat level.

This is what I've tried to do to model the threat concentrations using a statistical model. This is the details of the model I used the generalized linear model and I used median of security threat counts as a dependent variable. So, this very complicated-looking table includes four models for each type of threat. Each column is a model and includes four independent variables which are models and dependent variable.

So what the model is trying to do is study the effect of each of these variables on the dependent variable, which is the security threat concentration. If we look at the -- I don't know if the red line is readable. No? I wish it was. If you look at the first box over here, it is for the size of the gTLDs. So basically, all of these four numbers shows the amount of influence that size has on deriving threat counts.

The positive sign shows positive relation, which means the amount is 2.3 positive correlation between size of domains in zone and median concentration of phishing in a zone file. What is important in this model, in general, are the signs and the statistical significance which are shown by the stars. So basically, this line is showing that size is a statistical significant factor in explaining threads within gTLDs.

Now, the numbers do not matter that much. What does matter is that it is positive and is significant. For some of the other factors, it's not the case. For instance, if we look at Spec 13 TLD -- so basically gTLDs that have Spec 13 -- you see that only in case of spam there is significant relations. So basically, what this line says is that there is a positive statistically significant relation between TLDs that have Spec 13 and those that have a lot of concentration of spam in their networks.

I also looked at restriction code, which basically says how restricted the gTLDs are. And again, what we see here is that in terms of phishing, we see the more restricted the gTLDs, the less abuses concentrated in their networks. And last but not least, we see that TLD type, basically, in terms of new gTLDs, phishing, spam and botnet show significant concentration of abuse.

Moving on. This is a detailed breakdown of one of the factors that was included in the statistical model analysis in the previous slide. So basically, what you see here in the left plot is the distribution of gTLDs in terms of their restriction types. These are the colors, and per threat. So basically, the left plot shows that in terms of botnet command and

control servers, the majority of threats are located within generic gTLDs, but there is around 15% of them that are located in generic restricted gTLDs.

For the rest of the threats, its majority generic gTLDs. And for spam, as we also observed in the previous model, you see that there is around 5% of brand gTLDs that have spam in their networks. The right plot shows the absolute numbers. For instance, in the case of spam, this 5% translates to around 50,000 domain names that are located in brand gTLDs for September 2019.

This slide shows the same metric, but over time. So how did the concentration of abuse changed over time within different restriction types, for phishing, command and control. And the general trend that the plot shows is that it's either constant or going down. It doesn't mean that it's going down within these times, it could also mean that it is getting harder to be detected, restrictions put on WHOIS data.

The next topic, which was requested by the community, was more transparency on the DAAR progress and the project. Around a month and a half ago, we created a mailing list for our own communications about the DAAR project but also communities' feedback and to facilitate discussion around this project.

The mailing list is called *DNS Abuse Measurement*. It's not only intended to only be used for DAAR purpose, but it can also be used for that, and other possible future or already existing DNS measurements

with ICANN Org and ICANN Community. So we encourage you to become a member, if you are not already part of it.

JOHN CRAIN: And I'm just going to point out that we ourselves have not started using that as a distribution mechanism yet, but we intend to, directly after this meeting. This is when we actually plan to launch it, but people found out about it a little bit before. So this will be one place where, if you want to come to talk to us about any of our measurement projects and you want to do that in an open and transparent method, you can do that. It's also going to be archived, etc. It's a mailing list so feel free to join up.

SAMANEH TAJALIZADEHKHOOB: The next action that happens since a month ago is that the project is now ready to take ccTLDs into its analytics. So this started by many ccTLDs asking us whether it's possible to join the project. The initial problem was that we obviously have no access to most of the ccTLD zone files due to most of them being restricted. So as of today, we are ready to have a mechanism for ccTLDs to include their zone files in the DAAR system and to pull their statistics and aggregated metrics from the DAAR system through the MoSAPI API.

The pulley mechanism will be similar to those of gTLDs, and I have been told that ccTLDs either have access to MoSAPI or can get access. In case you guys have more questions about that, you can always come forward and ask us. So we encourage those TLDs that are in the

room or are listening to this presentation later, to come forward and participate. The requests can be made to this email address, which is very hard to read, for some strange reason.

JOHN CRAIN: It's globalsupport@icann.org. 'Globalsupport' being all one word. And if you're on the measurements mailing list, we'll send it there. So join the measurements list and then you'll get this.

SAMANEH TAJALIZADEHKHOOB: Moving forward. So the next improvement cycle that is currently scheduled for DAAR is to have DAAR v2, which we hope will have certain features that are currently improved. That includes more blacklist feeds, adding evaluation mechanism to adding or removing feeds, not only by having more but also removing those that we conclude are not accurate enough.

Also, to have registrar metrics. I explained this before in several sessions, that as of now, it's hard for us to develop metrics for registrars due to lack of WHOIS access. We have to query millions of domains per day for having accurate metrics, and it's important for the methodology to be replicable and reproducible. Hence, we do not want to use the ICANN-only data. And the last item is to include and study other factors that can affect security threat concentrations within registrar's and registries, such as registration policies and others.

I haven't talked much about the monthly reports. The project currently produces a report every month. The report goes back in time starting from January 2018, and includes, but are not limited to, some of that analytics that I showed in this presentation. Why you should go and read the report? It's because I did not show the detail of the methodology and the data collection.

One of the very important takeaways from this project is, at least for us, and the goals of the project is that the community are able to produce this system based on the methodology that we outline And that requires you guys to know the detailed methodology and the data collection. I haven't talked about what kind of feeds we use and the names of the feeds. There are certainly more analytics included in the report and will be included from the newer versions. And I already talked about access to the data.

Also important is that the changes that I included in this presentation are most probably going to be adapted on the reports starting from January 2020. So the new reports will have the changes included. With that, I would like to conclude the presentation and I'm happy to receive questions if there are any.

STEVE CONTE: I have a number of questions from the remote participants. So please slot me in with the live questions too.

KURT PRITZ: This is Kurt Pritz. I have three questions, the first of which will demonstrate my ignorance. So, in the pie chart, it showed that three quarters or more of the threats were spam, and then there were the other three. So those other three, botnets, malware, they're often distributed by spam. So are those sort of double counted as a spam and a malware? Or, if it's a malware delivered by a spam. is it not included in the spam count?

SAMANEH TAJALIZADEHKHOOB: I think they are counted in this case, due to its being hard to distinguish at the moment. We are over-counting and not under-counting.

KURT PRITZ: Okay. In your chart that showed that there was a correlation between new gTLDs and the amount of security threats. To me, that's not meaningful because there's such a wide range of new TLDs. If you think of the legacy TLDs here, there's ones that are less restricted and cheaper, and much more restricted. And so we all understand that, but my concern is, when you put up numbers like that, people wave their hands and say, "Ah, new gTLDs, we understand they're bad."

So I would caution against publishing that data that way, if you agree. So it's a comment, I guess, not a question. And my third thing is, on one of the graphs, it showed sort of a downward trend in abuse. And what would be helpful, if it's possible, and I don't know if it is, is to offer some sort of explanation. So is that a statistically meaningful

decrease? And if so, why do we think that is? Or is it seasonal, and not meaningful?

So, the numbers are informative, but they give rise to questions right away like, "Oh, are we doing better? And why? Or are we not really doing better the way we've displayed the range there?

SAMANEH TAJALIZADEHKHOOB: Thank you for your remarks and your question. About the second point that you made on the statistical model relation between new gTLDs and security, you're absolutely right. And this is also part of the feedback that we've received previously from the community, and also we've discussed with the registry stakeholder group.

And that is why we went further and include other variables, such as how restricted the zone is. And, we intend to do that more. So I understand that that gives a bit of skewed image, but that is not going to be the only statistic that we provide, for sure.

About the third point that you've made on the trend on the plot, I think you're talking about the time series. What we know is that the overall abuse trend is also going downwards. What is hard to say, and that is why most of the monthly records do not have absolute interpretations, is that due to the effect of GDPR, it's hard to say whether the abuse is really decreasing or we are unable to detect it. It's getting harder to detect. Wider RBLs, that is the current explanation. I hope that answered your question.

KURT PRITZ: Just to reiterate the point, even where you've identified generic TLDs, there's even a wide range of generic. I work with a TLD that has its own security measures in place that are part of the ICANN contract, so we would be labeled as generic and we're just one of hundreds and hundreds of generic TLDs that have a wide range of protections or security measures. And so again, I just want to caution against making the categories too broad, even categorized as generic. Thanks.

SAMANEH TAJALIZADEHKHOOB: Valid point. I'm just using this discussion to receive feedback Do you have suggestions for other additional factors that can be measured to distinguish even more?

KURT PRITZ: Heck, no. But thanks for the question, and I'll think about that. And we should think about how we want to portray this, so the reports meaningful and we can act on it and not say, "Okay, all generic the TLDs this…" rather than a specific corrective action we could take that would be really meaningful.

JOHN CRAIN: Over the last few months we've actually been (or I should say Samenah, because she does all the work) has actually been playing with different classifications and we've been talking to some of the

registries and others who actually understand -- we don't actually even understand what some of the classifications mean.

Some of them are in very old documents. We're not looking to make up classifications, we're looking for classifications that are out there and exist, and there are a few. We're not sure how meaningful there are. So there's probably going to be some discussion about what kind of classifications we're going to use in this data.

Because clearly, older versus newer brand versus not -- a lot of this may not be as meaningful as we think it is, especially as we start looking at the data, we're like, "Oh, maybe it's not as clear cut as we thought it was. Maybe we need a different classification process."

SAMANEH TAJALIZADEHKHOOB: Yeah, we would love to hear your feedback on that through the mailing list.

RICK WILHELM: Thanks, Rick Wilhelm, VeriSign. Just sort of a follow up on -- actually, a couple of follow ups. On the topic of classifications -- and some of this is for the benefit of folks who haven't been wallowing in this for a while -- the challenge with with classifications, of course, as Kurt was saying, the way that DAAR is constructed, any particular TLD that ends up in a particular classification, has their numbers rolled in, lumped in, clubbed together -- depending on one's terminology -- with other TLDs in that in that group.

And so then, you get painted with that brush, whether it's a good paint or a bad paint. And for some TLDs, that's a good paint. And for some TLDs, that's not such a good paint. This is different than in other types of things where you look at a classification and one's TLD might be displayed on its own in a separate table or a separate thing where you were listed and you're sort of like in a different class and then yours are listed -- your numbers, your individual TLD numbers are listed next to other TLDs of that same type, and then they're like for like, or kind of compared apples and oranges -- here, as Kurt was sort of saying, when they're classified together, that particular TLD's numbers are clubbed in together.

And that's why it's sort of challenging. And as John and Samaneh know, and the registry group that's been working with John and Samaneh on this, have been wrestling with this challenge of categorization, because we've not yet found sort of the -- So that's one thing, just to sort of amplify what Kurt was saying.

Regarding the lines down and to the left -- and that's on the slide 24 distribution restriction types over threats over time -- I'm not a security researcher and I don't live in it all day, and so while I can understand the linkage of GDPR making it challenging to determine the source of or attribution or whatever you might call it, like the source of the security threat, I'm not quite getting the linkage between GDPR restrictions that kicked in, viz a viz, the temp spec, and the detection of phishing, malware or other things.

So I understand that once a security researcher says, "Ah, there's a phishing or malware problem. Now we've got to go find out and do notification--" or some other sorts of thing, but I don't understand how GDPR and WHOIS data availability gets in the way of a security researcher saying, "Ah, there's a problem." And then I've got one other question that I think is easier. But let me sort of ask that question first.

JOHN CRAIN: Okay, so that's the message we're getting from the RBL provider. So I'm just going to do this, and ask Caro from Spamhaus, who explained this quite well in IDS, I think, but it was too much science for me.

UNKNOWN SPEAKER: When you look at the availability of WHOIS data, people usually think of being used for attribution, like what you explained. However, there's another very valuable part where the data comes in, which is not so much attribution, but is more correlation. So if you can determine that you're looking at one bad domain -- and for whatever reason you thought it was bad -- having access to the full or most of the WHOIS information, allows you to correlate that with other domains out there.

At some point, you'll probably, in a very reliable way, be able to say, "Okay, these other domains are related in such a way that they are

also bad." At that point, you can actually go ahead of the threat and you can say, "Okay, I haven't seen anything for this yet, but I can just say they're the same actor, they're being used in the same way, they're being set up in the same way."

Lots of different correlations you can make. And having the data available allows you to do that. Now that the data is not available anymore, at least not at the same way it used to be, that becomes a lot harder.

JOHN CRAIN: Thank you. What was your second question, or follow up?

RICK WILHELM: Okay. Thanks, I'll have to sort of chew on that, but that's helpful. The first one's an easy one, is there going to be a preview of DAAR v2 available? Because you said it's coming in January. And so, presumably, that's coming over right around the corner.

SAMANEH TAJALIZADEHKHOOB: Actually, what I meant that is going to change in January is the monthly report, not the DAAR v2. I think we will start working on it next year, or we are currently developing ideas. Yeah, it will be later.

JOHN CRAIN: Yeah, there are certain elements that we can easily change, like going from point in time to average. And things like that will be changing.

There's a whole discussion we want to have with you about what DAAR -- we don't know what DAAR v2 point whatever it is, zero or 1.9 AB, whatever we call it, is yet, because we want to have that discussion with the community to see what are the best options to put in there and what do we want to look like?

RICK WILHELM: Gotcha. So, December is more 1.1, not stick you with a version number. And then, lastly, ccTLDs, do you know yet how those are going to be represented? You talked about ccTLDs coming in, is there also a process for ccTLD withdrawal or egress, and such?

SAMANEH TAJALIZADEHKHOOB: Yes. About the first part of your question, and we've discussed this before, we would like to have this discussion with the community of how to represent the statistics in the monthly reports. What we know from now, at least it's obvious from now, is that once they join, they can receive their own aggregated numbers through MoSAPI, and they can also opt out whenever they would like to.

JOHN CRAIN: Yeah, I mean, we're not in a contractual situation with the ccTLDs, so this is purely voluntary on their parts. We've had a lot of interest, I must say. As you're well aware, and I think you pointed out as well, it's apples and pears. We have the majority of the generic space, we don't

actually have everything. We have the most, every now and again we lose zone files and stuff.

And we're not going to have that in the ccTLDs, so we need to figure out how we report on the ccTLD space and whether it's completely separate and there's some verbiage in there that makes it very clear that it's not the same. It's not straightforward. So we're going to have to have that discussion. We're not currently planning to include it into the monthly reports until we've figured that problem out.

RICK WILHELM: Okay, thanks. The registry stakeholder group would probably like to have a discussion with you about that. And while Donna's here, she's certainly not convened the stakeholder group to offer a coherent opinion about that other than that -- because generics don't have the ability to opt out. Thanks.

BERTRAND DE LA CHAPELLE: I am Bertrand de La Chapelle, Executive Director of the Internet and Jurisdiction Policy Network. Just two quick questions. The first one, in one of the slides you demonstrated the correlation between the size of the registry of the TLD, how much did you factor in the legacy value of VeriSign, in particular?

Because if I understand correctly, the bigger the TLD zone, the larger the number of problems. The fact is that a lot of things are happening, probably in dot com or dot net because it has been there for so long

that all the actors are using this. Have you tried to do this correlation by excluding the very large legacy zones, such as dot com and dot net?

And the second question -- I was about to ask exactly the same question on the correlation with the detection and the impact of GDPR. I was very interested by the answer, and when I look at the line there, in the whole debate about DNS abuse, I find it striking that we are here saying that one of the impacts of GDPR, for better or worse, is that we have less capacity to detect. If we agree on this, is there any possibility to insert in the debate about the new system, the possibility to have the correlation function without revealing the data?

Because technically, there is absolutely no obstacle to do sort of reverse WHOIS that doesn't reveal who it is, but that says those registrants have actually registered the same series of sites. Is it something that is taken into account in the debate or not? And apologies for not having followed the EPDP as closely as I probably should.

JOHN CRAIN: This isn't the EPDP meeting, but I do believe that this has been part of discussions in various places. There's been discussions about hashing data, etc., but you know, that's kind of out of scope for here but it's an interesting question.

BERTRAND DE LA CHAPELLE: I know it is, but the point is that sometimes there are precisely discussions that are taking place in different parts of ICANN. And I think here, there's a very clear connection between one process and the other one.

SAMANEH TAJALIZADEHKHOOB: So, about the first part of your question, if I understood it correctly, did you mean that we take the effect of the age of theTLD into account, excluding the ones that are very old or very big?

BERTRAND DE LA CHAPELLE: No, it's not a matter of that. If you take the new TLDs, for instance, and you want to verify that there is indeed a correlation between the size -- disproportionately within the size, to nonlinear proportion -- is there an excessive weight of the legacy TLD and the enormous zone of dot com and dot net because of the sheer numbers, and habit of doing abuses on these, and if you make the calculation on the other TLDs, do you see the same correlation? That was my question.

SAMANEH TAJALIZADEHKHOOB: Yeah, I understand it. In this case, yes. If I were going to model one-to-one and I would only model size and its effect on threat concentration, then your point was right. But what is happening in this model, and I tried to explain it in the slide before, also, the specific technique that I used to model takes into account the variances between instances in a model. In this case, it's gTLDs.

So, although there are certain gTLDs included in this model that are bigger than the other, but the fact that size matters [inaudible] what you just said. The big ones are not biasing the rest. It's just an overall effect. Actually, if I want to be very precise, it's not the bigger you are, the more abuse you are. It is, the bigger you are there are more chances or there are more attack surfers out there.

STEVE CONTE: We have three questions online, if that's okay.

RICK WILHELM: I think one of the challenges here is that these numbers aren't -- when we're talking about size, we should be talking pro rata or threat per registered domain name. And so, that doesn't always come across in some of the numbers. In some of them it does, in some of the stuff that we saw here today on these charts, it doesn't. It doesn't always come across as clearly as it might. That's all. I think it's pro rata -- maybe that's not the right term. Proportionality, I think is probably a better term to think about it -- and that doesn't always come across when we were looking at some of these numbers today.

SAMANEH TAJALIZADEHKHOOB: The model includes per type, per registry. But your point is correct. The absolute numbers do not matter. What matters is the difference between proportions. So yeah, the number by itself doesn't say much.

STEVE CONTE: Actually, I have four questions from online. Much of these were typed in during the presentation, so I leave it to you to determine the context of the question. First three questions are from Matthias Pfeiffer from Berlin. First question is, is it possible to provide all the raw data, like DNS-OARC does?

SAMANEH TAJALIZADEHKHOOB: No.

STEVE CONTE: Okay, second question, also from Mathias. Regarding "other factors measuring security threats", is the domain registration price a valuable data which DAAR consider as an important factor in the future?

JOHN CRAIN: She's looking at me, so that must be political one or something. We don't have that data, is one thing. I think adding new data points is something we're going to discuss with the community. I know pricing is hard to get, and sensitive. They've been other studies on that. So it's not in our current, even thoughts of version 2.0, but there'll be a discussion.

STEVE CONTE: Third question is, how many ccTLDs exactly are asked to join the DAAR measurement?

SAMANEH TAJALIZADEHKHOOB: I think by now, we have maybe around 20 ccTLDs that are currently giving their zones to [inaudible], our contractor for the DAAR project.

JOHN CRAIN: I'm going to say 42. We don't have exact numbers. The people who didn't laugh at 42 are not the engineers. So, just this week, I've had at least a dozen people come up to me and say, "Yes, we really want to do this." There have been a couple of dozen who've wanted to come in this way and wanted to make sure they do directly into the DAAR system.

I'm guessing we might have three, maybe four dozen, in the next few weeks. But it's purely a guess. And hopefully, as more join, we'll get a denser and denser data set. But I suspect there may be some registries out there that just never joined. I don't think we'll ever get to 100% but I'll be very happy to be proven wrong.

STEVE CONTE: Thank you. And I have one final question from online, from John McCormac from Hoster Stats dot com. He asks, "Could DAAR integrate pricing data with abuse data to identify gTLDs and ccTLDs that have more issues than others?

JOHN CRAIN: I think that's the same question as earlier. We've not -- I mean, could -- requires us to get the data etc. And whether or not we do, is a much broader discussion. I mean, purely as an engineer, I'd say yippee, great. If we've got solid data, do it. But it's a broader discussion than just whether or not the data is available. Who knows?

In theory, if you've got a good data set, and people agree it's valuable and it's not contentious, you could put any -- it's a database, right? It's just like with a categorization. If you have a clear categorization list, you can use it. The only question is that people agree that you should use it and it's the right categorization. So, it's going to be the same with questions about pricing and things like that.

STEVE CONTE: I have one more new question, so please put me back in queue.

SAMANEH TAJALIZADEHKHOOB: Are there any questions from the room? Yes?

KRISTOF TUYTELEERS: Good afternoon, Kristof Tuyteleers, DNS Belgium. I have a question [inaudible]. Why are you going to discuss new data points with the community but are you not discussing the current data sources with the community?

JOHN CRAIN: I think we are, but not in specific -- what we want to discuss is the methodology. What you've not seen here, because it's not ready yet, is Samaneh (because she's way smarter than me) has been working on a methodology for measuring these kind of lists and how you make choices, and we will be publishing that in the coming months.

We mentioned that we want to have a methodology for picking and for choosing and removing RBLs, because there might be reasons to remove particular ones. But we want to do it via methodology and not via, for example, just opinions. We want to have some science behind it. And Samaneh's already working on that. And I hope we'll see that published probably early next year, and it will be part of the discussion.

If you look at the reviews we had of the system, where we had a couple of independent reviewers, they actually pointed out that they thought we needed more data around specific threat types. And we're also looking at how we include those.

PETER JANSSEN: Peter Janssen, .eu registry, so one of those CC's that might be considering something in the future. And I apologize if the information is already out there and I missed it, but are you collecting information even if the cc is not providing his own [inaudible]?

JOHN CRAIN: We're collecting some information, because it's just in the RBL feeds. So collecting. But analyzing, no. We're not going to include ccTLDs without their expressed permission.

PETER JANSSEN: Okay, so if I tell you that I'm interested in the data but I can't give you the zone file, can you do something interesting for me or not?

JOHN CRAIN: Possibly. We have to have -- that's a discussion, right? So there is additional data sampling or data [inaudible] we want to do around what's actually in the zone versus what's in the RBL at a specific point in time. For that, we need the list of names and a time, which is not necessarily a zone file.

And there's a couple of ccTLDs that said, "Well, we'd rather just give you a list of names." That's engineering, on our side. Because at the moment, the system -- or I should say on a contractor side -- at the moment, the system ingests two data formats.

PETER JANSSEN: Yeah, I get that. But if there is something which is considered to be sensitive by some parties, it's the names, it's not what is the rest in the zone file. I mean, the name servers are not that important (some of them) for us, in any case. But something in between, which would be -- you have a point in time where you say, we have so many names that we see floating around which might be considered malicious in some

way. If we tell you at that moment in time, we had 3.748 million domain names, wouldn't that be enough for you?

JOHN CRAIN: For today, yes. Because today we're only counting, but we're hoping to become much more in advance. So what I'd say as a generic answer to that is, if members of the community think it's useful research for us to do, and we can do it -- because you know, the science -- then let's have that discussion.

And it doesn't necessarily have to be DAAR. DAAR is a tool. And that's why we have a measurement mailing list and not a DAAR mailing list, because if you want us to do other things, come talk to us. Actually, talk to probably Samaneh , because it'll go over my head.

PETER JANSSEN: That's what I'm doing now, right?

JOHN CRAIN: Okay. So yeah, we can discuss that and see exactly what it is you want.

PETER JANSSEN: Okay. Thanks.

STEVE CONTE: I have another question from Matthias, online. The question is, what's the problem to get pricing data of gTLDs?

SAMANEH TAJALIZADEHKHOOB: Basically, as a researcher before joining ICANN, I did studies on that. The first thing, there are already some studies done out there, that shows the relation between price and abuse concentration. Second, to answer Matthias' question, is that the data is very dynamic and since the space is really heterogeneous, one needs to find in the first place and crowd several websites on a short time frame to get that data. That is currently not the intention of DAAR and it's not planned to be.

Personally, I think it is important to look at the relation between pricing and security threat, but it is not what we are planning to do for now. If there are further questions on techniques, of how to do it, personally I can help, but it's not related to this session. I would like to thank you all and conclude the session. Thank you.

[END OF TRANSCRIPTION]